# Development of breast cancer risk prediction models using the UK Biobank dataset

A thesis submitted in fulfilment of the requirements for the degree
Doctor of Philosophy
in the Faculty of Biology, Medicine and Health

**2021**

**Kawthar Al-ajmi**
School of Health Sciences

## Table of Contents

**Total word count: 51,535 (Excluding references)**

## List of tables

## List of figures

## List of appendices

## Abbreviation

| | |
|---|---|
| BC | Breast cancer |
| A1/2 | Allele |
| AUC | Area under a receiver operating characteristic curve |
| BCAC | Breast Cancer Association Consortium |
| BMI | Body mass index |
| BP | Base-pair |
| *BRCA1/2* | DNA repair associated 1 and 2 |
| CDSR | Cochrane Database of Systematic Reviews |
| CI | Confidence interval |
| CUP | Continuous Update Project |
| CBC | contralateral breast cancer |
| CRUK | Cancer research UK |
| EAF | Effect allele frequency |
| ER | Oestrogen receptor |
| FN | False negative |
| FP | False positive |
| GWAS | Genome-wide association studies |
| Gy | Gray unit |
| hCG | Human chorionic gonadotropin |
| HER2 | Human epidermal growth factor receptor 2 |
| HR | Hazard ratio |
| HRT | Hormonal replacement therapy |
| ICD10 | International Classification of Diseases |
| IGF | Insulin like growth factor 1 |
| KMCC | Korean Multi-center Cancer Cohort |
| LCL | Lower confidence level |
| MAF | Minor allele frequency |
| MD | Mammographic density |
| MRI | Magnetic resonance imaging |
| MMP | Matrix metalloproteinase |
| MR | Mendelian randomisation |
| MWS | Million Women Study |
| NCC | National Cancer Centre cohort |
| NCI | National cancer institute |
| NHS | Nurses' health study |
| NHW | Non-Hispanic white |
| NPV | Negative predictive value |
| O/E | Expected/observed |
| OC | Oral contraceptives |
| OR | Odds ratio |
| PAF | Population attributable fraction |
| PH | Proportional hazard |
| PPV | Positive predictive value |
| PR | Progesterone receptor |
| PRS | Polygenic risk scores |
| QC | Quality control |
| ROC | Area under a receiver operating characteristic curve |
| RR | Relative risk |
| SNPs | Single nucleotide polymorphisms |
| TN | True negative |
| TP | True positive |

| | |
|---|---|
| UCL | Upper confidence level |
| WCRF | World Cancer Research Fund |
| WHI | Women Health Initiative study |
| WHR | Waist to hip ratio |

## Research contribution

a) Publications included in this thesis

1- **Kawthar Al-ajmi,** Artitatya Lophatananon, Martin Yuille, William Ollier, Kenneth R Muir. Review of non-clinical risk models to aid prevention of breast cancer (**published**: 3rd September 2018, Journal: Cancer causes & controls)

- **Candidate's role:** Data collection, literature review, results interpretation, writing original draft, critical revision and editing of the manuscript.

2- **Kawthar Al-ajmi,** Artitatya Lophatananon, Kenneth R Muir. Anthropometric and reproductive factors and breast cancer risk in the UK Biobank female cohort (**published**: 26th July 2018, Journal: PLOS ONE)

- **Candidate's role:** Data acquisition, data analysis and results interpretation, results validation, writing original draft, critical revision and editing of the manuscript.

3- **Kawthar Al-ajmi,** Artitatya Lophatananon, Kenneth R Muir: Association of non-genetic factors with breast cancer risk in genetically predisposed groups of women in the UK Biobank cohort (**published** :24th April 2020: JAMA Open network journal)

- **Candidate's role:** Data acquisition, data analysis and results interpretation, results validation, writing original draft, critical revision and editing of the manuscript.

4- **Kawthar Al-ajmi,** Artitatya Lophatananon, Kenneth R Muir. Development and assessment of breast cancer risk prediction models based on the UK Biobank female cohort (**written and to be submitted**)

- **Candidate's role:** Data acquisition, data analysis and results interpretation, results validation, writing original draft, critical revision and editing of the manuscript.

5- **Kawthar Al-ajmi,** Artitatya Lophatananon, Kenneth R Muir. Testing the developed breast cancer prediction model in the community, Pilot study. (**Under development**)

- **Candidate's role:** Data acquisition, data analysis and results interpretation, results validation, writing original draft, critical revision and editing of the manuscript.

## Summary

Personalised breast cancer risk prediction based mainly on the modifiable factors can help to increase awareness about breast cancer risk factors. It can also encourage females to adhere to a healthier lifestyle and can be used as an educational tool for the public. The project aimed to develop user-friendly breast cancer models. A web-based portal was also developed for the public.

Chapter 1 **Introduction**: Describes the epidemiology of breast cancer with a focus on the UK burden. Screening program was discussed, and risk prediction models were introduced at this chapter. Later discussed the most popular models used to predict BC regardless the type of the included risk factors. Moreover, model's performance and applications were discussed. The UK biobank prospective cohort was introduced. Finally, thesis aims were listed at the end of this chapter

Chapter 2 **Literature Review**: A literature review of breast cancer risk factors with up-to-date evidence is presented in this chapter. The focus was on studies published after 2010 with some consideration of older studies when needed. Mechanism and evidences of each factor were discussed in this chapter. Moreover, the availability of these risk factors variables in the UK Biobank was checked to end up with the most significant risk factors for the model.

Chapter 3 **Methodology**: This chapter presented the elaborated methodology of each chapter and paper presented in this thesis. The start was with a general introduction of the UK biobank, its characteristics, and power calculation. Later, detailed sections on 1) defining BC cases and controls 2) defining the menopausal status of the participants used in this project. Then a detailed methodology of each chapter was presented.

Chapter 4 **Review**: Is a published paper describing details about the existing and most common breast cancer models, types, performance criteria (calibration and validation) and

their applications were discussed in this chapter. Moreover, a review of non-clinical risk models of breast cancer is described in this chapter. Fourteen models identified and included in the review as non-clinical models and their performances measures have been reported. All the models were well calibrated with modest discrimination power. The discrimination C-statistics ranged from 0.56 to 0.89. The chapter identified a need for more accurate breast cancer models for UK females.

Chapter 5 **Risk factors of breast cancer among UK Biobank females:** Is a published paper describing the association between anthropometric and reproductive factors and breast cancer risk among UK females. The direction and magnitude of the risk varied based on the menopausal status. In pre-menopausal, being older, taller, with low waist to hip ratio, low BMI, first degree family history of BC, early menarche age, nulliparous, late age at first live birth, high reproductive interval index, and long contraceptive use duration were all significantly associated with an increased BC risk. In post-menopausal, getting older, being taller, having high BMI, first degree BC family history, nulliparous, late age at first live birth, and high reproductive interval index were all significantly associated with an increased risk of BC.

Chapter 6 **Effects of lifestyle on breast cancer risk across genetically defined risk groups:** Is a published paper describes contribution of non-genetic factors (lifestyle habits) to breast cancer risk in genetically predisposed groups based on polygenic risk scores. This chapter explores how risk of breast cancer can be reduced by adhering to healthier lifestyle options such as (more exercise, maintain a healthy weight (BMI <25 kg/m$^2$), low alcohol intake (No or < three times a week alcohol intake frequency), and no contraceptive avoiding hormonal replacement therapy (no or HRT used for < 5 years) even in females with higher genetic risk or predisposition.

Chapter 7 **Risk prediction models development**: Describes the development and validation of four breast cancer risk prediction models, two epidemiological models – no clinical nor genetic risk factors were included - pre-and post-menopausal models. Two combined models based on epidemiological and genetic factors: pre-and post-menopausal with polygenic risk scores as a risk score. Internal validation was performed on the four models; however, the external validation was only assessed for the two epidemiological models due to a lack of data with genetic data for validation. Finally, a brief description of the (RiskWomen) website developed by our group was given.

Chapter 8 **Conclusions and future work:** Describes the keys findings of this project and lists possible limitations and strengths. Moreover, discussion on its possible applications and its implication for the public use is also presented. Later, recommendations and future work are also presented together with final conclusions.

# Abstract

**Aim**: The work presented in this thesis is based on the following aims; 1) to systematically review non-clinical/non-genetic breast cancer risk prediction models, 2) to review the published risk factors of BC (reproductive, anthropometric, lifestyle and dietary) to take as a base of the model development, 3) to assess the BC risk factors using the UKBiobank prospective cohort, 4) to explore the effects of adherence to "healthier lifestyles" in groups based on different genetic predispositions, 5) and to develop BC risk prediction models (epidemiological and genetic models).

**Methods**: For aim 1, a PRISMA approach was employed to carry out the systematic review. For aim 2, the literature was reviewed and summary of evidences was presented. For aim3, the UKB data was analysed using the ***glm*** model to derive relative risk and 95% confidence intervals. For aim 4, the hazard ratios of different lifestyle categories were calculated based on the tertile groups of genetic predisposition score (using 305 SNPs). For aim 5, backward stepwise logistic and bootstrap regression approaches were used to derive the best fitting (epidemiological and genetic).

**Results**: For aim 1, 14 epidemiological (non-clinical and non-genetic) models were identified. All of the models were well calibrated but had poor or moderate ability to discriminate in internal validation analyses. However, external validation was also missing for most of the models. Additionally, generalisability is also problematic as some variables are specific for some populations.

For aim 2, a list of modifiable risk factors (physical activity, alcohol, smoking, BMI, OC use ,HRT, and diet), partially modifiable risk factors (age at first birth, null-parity, and breastfeeding), and non-modifiable risk factors (age, genetic factors, family history of breast cancer, early menarche age, late menopause age, benign breast disease, breast density, height , abortion, and radiation) were summarised and evaluated.

For aim 3, the following risk factors: age, height, low BMI, low waist to hip ratio, first degree family history of BC, early menarche age, null-parity, late age at first live birth, high reproductive interval index, and long duration use of contraceptive were all significantly associated with an increased BC among pre-menopausal females. While among post-menopausal, age, height, high BMI, first degree BC family history, null-parity, late age at first live birth, and high reproductive interval index were all significantly associated with an increased risk of BC.

For aim 4, our analysis showed potential BC risk modifications as a consequence of selected modifiable lifestyle factors (more exercise, healthy weight, low alcohol intake, no contraceptive or no or limited HRT use). The results were significant regardless of whether women had higher genetic risk.

For aim 5, two epidemiological models based on menopausal status (pre- and post-menopaused models) were developed together with a computation of the absolute 5 years risk. Later, the discriminatory power of the models was significantly improved by adding a PRS as a risk score for breast cancer in the extended genetic models.

**Conclusions:** The work presented in this dissertation can be used for a) increasing the public awareness regarding the possible risk factors of BC, b) encouraging females to change their lifestyle into a healthier style to reduce their BC risk, c) using the models as an educational

tool for the community and primary care as a strategy for cancer education and prevention, d) encouraging females at higher BC risk to attend the screening invitation.

## Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of any other university or other institute of learning.

## Copyright statement

# Acknowledgment

Whenever anyone asks me what I liked the most about my PhD experience, my answer always will be ‘my supervisors’. Dr Artitaya Lophatananon and Prof Kenneth Muir are the most helpful and supportive supervisors you will ever meet. I had a very good PhD experience because of them. Having meetings with supervisors are usually considered an unpleasant experience, however, I always felt so good after our meetings and as member of a family; for that feeling I am so grateful.

By the end of the PhD, I was lucky to meet Dr Krisztina Mekli, who helped me in the analysis. She was so supportive and helpful whenever I needed her help. She is genuinely kind, warm-hearted and smart. Prof William Ollier, thank you for astounding editing and insightful feedback at short notice. I would like to sincerely to thank the Omani government for funding me through my PhD scholarship. Moreover, I would like to express my gratitude to UK Biobank and the ATP cohort participants and staff. Without them, this work would not have been possible.

Special thanks go to my Manchester friends: Siham Al-hadhrami, Hafidha AL-hattali, Asma Al-hosni, Widad AL-rawahi, Aisha Al-rahbi and Aisha AL-sulimani. I am thankful for the debates, dinners, long chats, sleep-over nights, hilarious photos, editing advice, rides to the airport, and all the special memories that you were part of.

My much-loved family, Abduraof, Hajer, Hawraa, Khalil, Einas, Fatma, Ahmed, Khaloud, Batool and Ali. Thank you for always trusting in me. Thank you for the love, prayers, and continued support.

Acknowledging the support my father **Ibrahim**, mother **Safiya** and Uncle **Ahmed** has been the most difficult to write about as no words can describe how thankful and grateful, I am for them. They gave me the chance to follow my dream, supported me emotionally and financially, and they always believed in me.

Lastly, I am so thankful for the love of my husband **Hassan** and my daughter **Ascia** whom I dedicate this thesis to. I married Hassan in the middle of the PhD journey, and it was the best decision ever. Hassan made the journey more enjoyable and fun with his endless support and laughter. He and Ascia made this experience a very happy journey and for that I am thankful.

To **IBRAHIM & SAFIYA** my amazing parents, **HASSAN** my wonderful husband, and **ASCIA** and **SAWSAN** my brilliant daughters to whom I owe every success in my life.

# Chapter 1 : Introduction

### 1.1 Breast cancer epidemiology

The global burden of breast cancer (BC) is measured by the incidence, mortality, and economic costs of the disease. Globally, BC is the commonest cancer among females (25% of all diagnosed female cancers worldwide [1]) and was the second leading cause of cancer death among females (626,679) in 2018 [2, 3].

Figure 1.1 and figure 1.2, show that incidence rates are almost four times higher among high income regions (94.2 in Australia and New Zealand) when compared to low income countries (25.9 south central Asia). However, BC incidence rates are increasing globally. Not surprisingly, mortality in cases is lower in developed countries when compared to developing countries (figure 2). Several reasons explain the difference in incidence and mortality rates between developed and developing countries. Availability of early detection facilities e.g. mammogram and magnetic resonance imaging (MRI), use of chemoprevention (tamoxifen and endocrine therapy), accessibility to health care system, population awareness of the disease and its symptoms and risk factors, socio-economic status are just some of the reasons behind the mortality rate differences [4, 5]. Factors such as behavioural (alcohol intake, smoking, sedentary lifestyle, intake of contraceptive hormones and hormonal replacement therapy, low parity rate, low breastfeeding rate), ethnicity, higher socio-economic status are linked with higher incidence rate in developed countries [6].

BC is the most common cancer among females in the UK, where 15% of newly developed cases are BC cases [7, 8]. Moreover, UK has the highest age-standardised incidence and mortality rates of BC in the world; two in every 1000 women aged 50 and above are likely to develop BC annually [9]. It is estimated that 41,760 females will die from the disease each year (accounting for 14% of female cancer deaths) [10], making it the second leading cause of cancer deaths among females [1, 11]. BC is more prevalent among white Caucasian females more than Black or Asian females living in the UK.

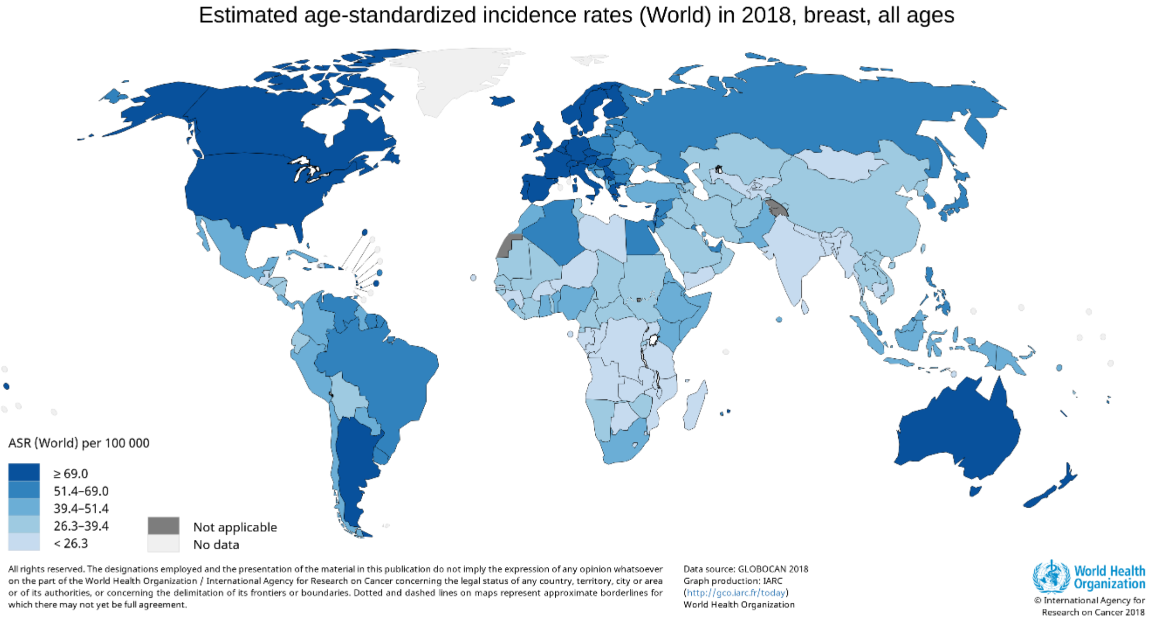

Estimated age-standardized incidence rates (World) in 2018, breast, all ages
ASR (World) per 100 000
≥ 69.0
51.4–69.0
39.4–51.4
26.3–39.4
< 26.3
Not applicable
No data
All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.
Data source: GLOBOCAN 2018
Graph production: IARC
(http://gco.iarc.fr/today)
World Health Organization
World Health Organization
© International Agency for Research on Cancer 2018

Figure 1.1: Breast cancer age standardised incidence rates from 2018 world statistics presented by the international agency for research on cancer. Available from: http://gco.iarc.fr/today/home

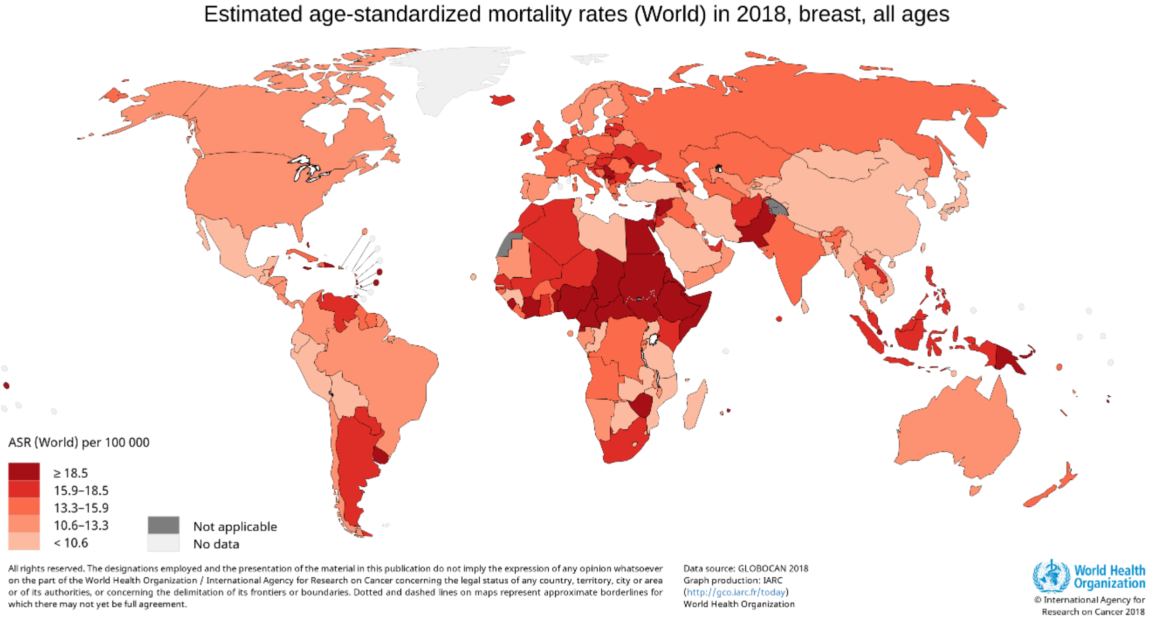

Estimated age-standardized mortality rates (World) in 2018, breast, all ages
ASR (World) per 100 000
≥ 18.5
15.9–18.5
13.3–15.9
10.6–13.3
< 10.6
Not applicable
No data
All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.
Data source: GLOBOCAN 2018
Graph production: IARC
(http://gco.iarc.fr/today)
World Health Organization
World Health Organization
© International Agency for Research on Cancer 2018

Figure 1.2: Breast cancer age standardised mortality rates from 2018 world statistics presented by the international agency for research on cancer. Available from: http://gco.iarc.fr/today/home

## 1.2 Breast cancer screening programme

In the UK, the BC national screening programme (mammography) was started in 1988 [12]. Females aged 50-70 years are invited for screening every three years. In 2009, as part of an extension trial, females aged 47-73 years were also randomly selected and included. Even though the incidence rate is increasing, nonetheless a drop in mortality rate has occurred in the UK since the early 1990s [13] as a result of screening programme initiative. The primary goal of the National Screening Programme is to minimise the BC-related death rate as much as possible. As a consequence, more BC cases have been identified in early pre-clinical stages leading to more effective treatment. This has led to a reduction in BC-related death rates. However, mammography has also been associated with some harm, such as over-diagnosis, false positive results, false negative results, radiation-induced cancer, unnecessary breast biopsies [3] and increased discomfort and anxiety caused by the screening [14]. Over and mis-diagnosis are now regarded as being the main drawbacks of routine mammography screening. Overdiagnosis is defined as being the detection of a tumour by screening but which lacks the potential to progress to a symptomatic cancer. Such 'tumours' can even regress or the patient can eventually die from other causes before BC progresses to a clinical stage [15]. This presents a significant challenge for screening as it is hard to distinguish between life-threatening tumours and over diagnosed tumours. As a consequence, both tumours are treated equally and the 'over diagnosed' patients experience unnecessary treatment complications without any clear and obvious benefit [15]. This represents a significant ethical dilemma and impacts on both the patients and the health care organisation. BC treatments themselves have the potential to cause other diseases, such as cardiovascular diseases; an increased frequency of cardiovascular disease has been observed in women treated with radiotherapy [16]. Furthermore adjuvant based treatment could also be cardiotoxic as well [17]. As consequence, overdiagnosis can lead to increased mortality rates due to other causes.

The generation of false positive results in mammography screening, often results in increased anxiety and distress due to concern related to the possibility of a BC diagnosis. It was estimated that the mental well-being of women given any equivocal/positive screening result will have a negative affect for at least 3 years after the false screening [18]. Additionally, such false positive mammography reported will usually generate unnecessary invasive breast biopsies and surgical procedures. In the UK, around 2.3% of females with a false positive mammography screen had a lumpectomy performed [14]. In contrast, false negative results can lead to delayed medical care and not benefiting from early detection [19].

Regarding the issues relating to radiation induced cancer, it has been estimated that the levels of ionizing radiation used in mammography have increased by sixfold between 1980s and the present day due to the use of more powerful imaging techniques [20]. According to a recent study [21], in a cohort of 100,000 females having annual screenings from ages 40 to 55 and biannually until age of 74 years (with a dose of 3. mGy per screening), this will stimulate the development of 86 incident cases of BC. Furthermore, this study estimated that 11 deaths in this cohort could be attributed to radiation-induced BC.

Risk prediction models can help to identify females at risk based on their personalised risk estimation. This approach helps stratify females according to their risk group and helps in prioritising the high-risk females within the screening program. Additionally, it can be used to encourage females to adopt a healthier lifestyle as a prevention strategy for BC. Moreover, a lot of females do not respond to the BC screening invitation letter and using the risk prediction model can help to increase their awareness about BC risk.

### 1.3 Breast cancer risk model

A risk prediction model is an individualised statistical method to estimate the probability of developing certain medical diseases based on specific risk factors in currently healthy individuals within a specific period of time. The model contains an algorithm that combines general risk factors for developing the disease and generates an individualised risk score. The baseline risk of the disease is usually estimated using a prospective cohort study. This cohort represents the population at risk whose risk factor values are zero. The individualised risk score is a score derived from a set of risk factor values multiplied by the beta weights associated with these factors.

The main statistical tests used to assess the risk factors and their beta weights are logistic regression and Cox proportional hazards regression. The variables in the model can be a combination of behavioural, environmental, psychological, and genetic measures of the individual. Risk prediction models can predict both the individualised risk and the population risk. The latter is obtained by using average values of risk factors calculated from the population. To ensure the reliability and generalizability of the risk model prediction, validation is needed using an independent sample from the same or different population. The performance of each prediction model may vary according to the population used. One model can be very accurate in a high-risk population and less accurate with low-risk population and vice versa [22].

The first risk prediction model developed in 1976 was the Framingham Coronary model. The model was used to predict risk of developing heart disease by Kannel et al. [23]. Since then, number of risk models has grown gradually for many chronic diseases such as (cancer, heart diseases, stroke, diabetes, osteoporosis, and bronchitis & emphysema). Clinicians use such risk prediction models to aid decisions on: planning treatment and prevention strategies, designing interventions, identifying high risk individuals, estimating disease’s burden in a population/s and in assisting in producing the benefit-risk indices [24].

### 1.4 Types and applications of breast cancer risk prediction models

Different models have been established to assess the likelihood of developing BC depending on the assessed factors in each model. Usually, BC and ovarian cancer risk prediction models are developed in parallel to assess the individualised cancer risk for arbitrary females. These models must be distinctive and specific so as not to be confused with the prognostic models which predict the likelihood of cancer recurrence or mortality.

In general, BC risk prediction models can be divided into three groups. The first type of models estimate the risk of developing BC; the second models estimate the risk of having the high-risk mutations such as BRCA1 and BRCA2; the third models estimate the risk of both [25]. Combination of risk factors are used in these models e.g. environmental, hormonal, epidemiological, dietary, genetic, and clinical risk factors. Family history is also often included in these models. Some models assess BC family history only, whilst others also include ovarian cancer family history. For cancer prevention purposes, risk prediction models should consist of modifiable factors such as lifestyle, hormonal factors, and diet.

Risk prediction models can be used as a preliminary approach to prioritize individuals on screening programmes, genetic counseling and testing, and can be used to advance the BC research [25]. At an individualized level it could help behavioral change and encourage users to adopt healthier lifestyle. Models can be used in a clinical setting if they can be accessed via the internet. The most widely used risk prediction models by clinicians are listed in table 1.1 [26].

In the Gail model, the risk factors included are current age, ethnicity, age at first life birth, age at menarche, the number of previous breast biopsies, the number of BC first-degree female relatives, and history of atypical hyperplasia. This model does not include genetic information, extended family history or ovarian family history.  This model is not applicable for predicting BC in BRCA mutation carriers [22, 26, 27].

Other risk models have incorporated genetic aspects of BC and therefore they have applications among individuals with a familial risk pattern. These models are called genetic risk prediction models and they include the family history of BC and ovarian cancer along with other risk factors obtained from the epidemiological studies.

The first genetic model used was the Claus model; it had an assumption of 1 autosomal dominant gene with age-dependent penetrance and no BRCA genes assumptions included at that time. Subsequently, an extended version of Claus model was developed by adding the ovarian cancer cases still without BRCA mutations assumptions.

The BRCAPRO model is based on BRCA1/2 genes assumptions and family history of BRCA1/2 associated cancers. BRCAPRO is effective for multiple ethnicities. This model can be adjusted to include relatives with mastectomies and it has been recently modified for estimating contralateral BC (CBC) penetrance [26].

The IBIS model was established using BRCA1/2 gene assumptions and adjusted for the residual effects of a third dominantly inherited common gene. Further to genetic factors, this model includes body mass index, age at menarche, age at menopause, parity, age at first childbirth, and benign breast diseases. A disadvantage of this model is that it is only applicable to healthy females only.

The BOADICEA model (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) applies BRCA genetic susceptibility and additional polygenic factors, family history of BRCA1/2 associated cancers, mutation screening result, human epidermal growth factor receptor2 (HER2), progesterone receptor (PR), oestrogen receptor (ER), and basal cytokeratin (CK) expression (CK5/6 and CK14) [26, 28].

The “Your disease risk” model was developed by Harvard University specifically for American women. BRCA1/2 status is included along with family history of breast and other cancers. Modifiable and non-modifiable risk factors are also included. A disadvantage of this model is that is more accurately predictive in females aged 40 and above.

According to the validation test (Area Under the Curve-AUC), the IBIS model performed the best among the 5 models described above while the other models underestimated the BC risk. A recent study compared the Gail and IBIS models regarding the 10-year BC rates. This found that IBIS had a better calibration and discrimination ability with AUC of 0.7 and AUC of 0.63 in the Gail model. Another small study among Ashkenazi high-risk BC women found that both IBIS and BOADICEA overestimated the BC risk. Nevertheless, BOADICEA was better calibrated (O/E ratio 0.80, 95% confidence interval (CI) 0.54–0.93) than IBIS (O/E ratio 0.52, 95% CI 0.32–0.87). In addition, an Australian  large cohort study showed that the BOADICEA model was well calibrated (O/E ratio 0.92, 95% CI 0.76–1.10) and with good discrimination (AUC 0.7) [26].

Table 1.1: The common risk prediction models used by clinicians

| Model [26] | Cancer | Genetic | Non-genetic | Tumour pathology | BRCA1/2 status | Associated tumours | Web link |
|---|---|---|---|---|---|---|---|
| Gail (1989) | BC | None | Yes | Yes | No | No | www.cancer.gov/bcrisktool |
| eCLAUS (1993) | BC | 1 gene | No | No | No | No | www.cyrillicsoftware.com |
| BRCAPRO | BC, OC | BRCA1/2 | No | Yes | Yes | Yes | http://bcb.dfci.harvard.edu/bayesmendel/brcapro.php |
| IBIS (1993) | BC, OC | BRCA1/2 + single moderately penetrant gene | Yes | No | Yes | Yes | www.ems-trials.org/riskevaluator |
| BOADICEA (2008) | BC, OC | BRCA1/2 + polygenic component | No | Yes | Yes | Yes | http://ccge.medschl.cam.ac.uk/boadicea/boadicea-web-application |
| Your disease risk (2000) | BC | No | Yes | No | Yes | Yes | http://www.yourdiseaserisk.wustl.edu/ |

BC= Breast cancer, OC= Ovarian Cancer

## 1.5 Calibration and discrimination of risk models

In the development of the risk prediction model, a calibration process is required. Model calibration refers to the level of agreement between predicted and observed results. A graphical evaluation of calibration is often illustrated by applying predictions on the x-axis and outcome on the y-axis where a perfect fit is represented by a 45°line. The calibration plot is a simple scatter plot in case of linear regression, whereas for binary outcomes the x-axis contains only 0 and 1 values. Additionally, the E/O statistic (expected/observed) measures the agreement between the observed and the expected values. An E/O statistic close to 1 indicates good calibration, whereas less than 1 means underestimation and more than one means overestimation [29, 30].

Another important assessment is model discrimination where it measures how well the model is able to discriminate between affected and un-affected subjects. The AUC (area under the receiver operating characteristic curve (ROC)), and the C statistic are used to test the discrimination power of the model [29]. In practice, the C statistic is the most common test used to assess model discrimination. It measures how efficient the model is at discriminating between females who are affected or not affected by BC. A C statistic of 0.5 indicates no discrimination between individuals who develop the condition and those who do not. In contrast a C statistic of 1 implies perfect discrimination [22, 26]. Models to be used for prioritizing people for screening or effective clinical decision making must have good discrimination power [24].

## 1.6 Application of breast cancer risk prediction models

The first risk prediction model for disease was developed in 1976; this was the Framingham Coronary model used for predicting the risk of developing heart disease [23]. Since then, the number of disease risk models has grown gradually for many chronic diseases, such as cancers, heart diseases, stroke, diabetes, osteoporosis, and bronchitis & emphysema to name but a few.

These models can help inform individuals to adopt healthier ways of living their lives and prevent or slow down disease onset [31]. Additionally, these prediction models have the potential to guide clinicians, health care system, and policy makers in taking decisions on: designing intervention trials to prevent disease, planning treatment and preventions strategies, identifying high risk individuals, estimating the burden of a disease in a population/s and in assisting in producing  benefit-risk indices [24]. The first of the above-mentioned applications is that of designing and identifying the eligibility criteria for screening and intervention trials.

An example of disease prevention planning is the Gail Breast Cancer Risk Assessment Model [32]  [33]. This was designed to initialize a randomized, placebo-controlled trial of the chemo-preventive impacts of tamoxifen in females with a high risk of developing BC.

Another application has been increased ability to identify the high-risk individuals  who may benefit from screening and preventive interventions such as tamoxifen chemo-prevention. An example of this application was when the US Food and Drug Administration  used the 5-year BC risk cut-off of 1.67% or more, as a basis for advising/recomending females aged 35 years or older  to embark on tamoxifen chemo-prevention [24].

Risk prediction models can help in developing ‘benefit-risk indices’ and their greater understanding. For example, BC prevention trials showed a reduction of 49% in invasive BC in females at high risk after undergoing tamoxifen chemo-prevention. However, in endometrial cancer, pulmonary embolism, and stroke occurred more in females taking tamoxifen treatment compared to females not taking this treatment. The Gail model [34], facilitated the development of an ‘ index of benefit-risk’ which could be used to evaluate the benefits of  tamoxifen in reducing the risk of BC and its harm / adverse effects in other diseases. The overall conclusion was that in older females using a risk of 1.67% is not justified and a much higher risk is needed before recommending  tamoxifen treatment.

Prediction models can also help in estimating the population burden of a targeted outcome and the cost of the interventions potentially used. An example of this was again provided by using the Gail model [32] and the benefit risk index by Freedman *et al* [35], to estimate the number of females who would be eligible and benefit from tamoxifen chemo-prevention in the USA.

The most common use of risk prediction models is to help clinicians decide, what are the best screening and interventions approaches to be used in their patients? At the present time, genetic susceptibility risk models are now also used for patients with a family history of particular inherited diseases.

However, many models looking at similar outcomes or similar targeted populations have been developed. With the increasing number of models, clinicians need to be able to decide which model is the most appropriate for use on their patients. As a minimum, they need to know how well the model predicts the outcome, how good the model is in predicting if it is applied to someone of different heritage to the population the model was originally developed for. Lastly, they need to know which risk prediction model works best out of the group of modles available. Different performance measures are available for comparison such as: discrimination, calibration, accuracy, dispersion, precision, and utility. All models should be evaluated based on its ultimate use [24].

These prediction models can either target the whole population or be directed at just high-risk individuals. Most of the risk-prediction models developed lack good discrimination and accuracy even though they are well calibrated. They cannot accurately discriminate between individuals who will develop the disease from those who will not develop it. Thus it is important to build models targeting the whole population for prevention strategies rather than just be restricted to high-risk individuals; if not they could miss substantial number of subjects with the disease [24].

BC risk prediction models in particular has increased over the past three decades. In the last decade, BC prediction models have been improved with better discrimination. Even though, they had modest discrimination power and acceptable calibration, that does not indicate that these models are useless. One of the main aims for developing BC prediction models is to develop risk-based screening programs. Females in Europe are invited every two to three years for mammogram radiation which is considered as one of the risk factors for BC [36, 37]. Using BC models can identify and prioritise females who need the mammogram screening and reduce BC risk. Nevertheless, official screening programs do not use the BC models due to high uncertainty level in its discrimination power. One possible explanation of this modest discrimination power is the nature of BC and its low incidence rate [38]. In general population, the probability of developing BC is low (even among high risk females) which might lead to low discrimination power of BC model compared to discrimination power of a common disease like cardiovascular diseases. The best ROC value of BC prediction model and was reported in 2017 by [37] with ROC of 0.71.

In this project, the team is developing a personalised risk prediction BC model incorporating epidemiological risk factors and genetic factors using the UKBiobank prospective cohort. The UK Biobank project is a population-based prospective cohort with extensive phenotypic data and genetic data collected on about 500,000 individuals from across the United Kingdom (22 assessment centres across the UK). Individual recruitment was throughout the UK to ensure heterogeneity of socioeconomic and ethnic background with a mixture of urban and rural recruitment. The recruited subjects aged between 39 and 71 at recruitment. This research resource is exceptional in its size and scope with a numerous variation of health-related data for each participant. These data including lifestyle questionnaires, biological measurements, brain and body imaging, blood and urine biomarkers, and genome-wide genotyping data. Moreover, follow up information is also available by getting the permission to link to the health records. UK biobank was established to allow comprehensive investigation of non-genetic and genetic determinants of the diseases (outcomes) among middle and old aged people [39-41].

The UK biobank population ensured a comprehensive distribution across all exposures to ensure the detection of reliable associations between exposures and the interested health outcome/s. UK biobank is available as open research resource for all researcher around the world without the need to collaborate with UK based institute. As a result, UK biobank was considered to be used in our project as it was the best and most convenient open data source for the UK population and because it was open for all researchers. Application process was easy and was through University of Manchester. Especially after getting the approval for using the data within the time limit of the Ph.D. project.

## 1.7 Thesis aims

1- To systematically review published BC risk prediction models which were based on modifiable risk factors (No genetic nor clinical model was to be included in the review).

2- To review the published/established risk factors of BC.

3- To assess the reproductive, lifestyle, anthropometric and diet risk factors associated with developing BC among UK females. These results provide key risk factors for inclusion in the risk prediction development work.

4- To investigate the effect of lifestyle in different genetic predisposition and their effect on the BC risk.

5- To develop two BC epidemiological based risk prediction models based on menopausal status (pre-menopausal and post-menopausal). Internally and externally validating the two models using a valid independent cohort (these models are intended to incorporate only risk factors and for general educational purpose).

6- To explore genetic predisposition using the UK Biobank female cohort and to incorporate genetic factors (PRS) to the models develop in aim number 4 and assess the model’s performance (these models are intended for targeted screening application).

# Chapter 2 : Literature review

## 2.1 Introduction

This scoping review was conducted to assess risk factors associated with BC. The review focused on dietary, lifestyle, anthropometric, genetic, environmental, and reproductive factors. The review covered evidence from different types of study including observational, experimental, and meta-analysis.  Evidence was examined against criteria to provide summary level information. The relevance of each reviewed factor to the prospective cohort of the UK biobank is presented. This chapter aims to introduce BC risk factors and their potential to be included in all analyses in this thesis.

## 2.2 Methods

Relevant studies were retrieved using combination of controlled keywords and vocabulary using various search engines: PubMed, Google Scholar, and Science Direct. The search terms comprised of: (breast) AND (cancer or neoplasm or tumor) AND (risk or risk factor or factors); combined with each of the following strings: (UK one million woman study* OR BCAC study* OR PROCAS study*OR UK* OR "UK BIOBANK" OR World Cancer Research Fund OR WCRF updates, "CRUK" OR Cancer research UK OR cancer consortium OR systematic* OR meta* or review ) OR (prospective* OR cohort* OR retrospective* OR case control OR trial* OR experiment* MR* OR mendelian randomisation) OR (age OR parity OR [first AND {pregnan* OR child OR birth}]) OR menarche OR menopause OR menopausal OR hormonal OR HRT OR hormonal replacement therapy OR breastfeed* OR contraceptive* OR oral* OR (reproductive AND [risk OR life]) OR abortion OR miscarriage OR family history OR familial OR (genetic AND [risk OR predisposition]) OR hereditary OR (breast AND [susceptibility OR alleles OR SNPs]) OR (modif* OR non-modi*) OR (prediction* OR model* OR hazard* OR index* OR assessment* OR tool* ) OR (lifestyle OR diet OR anthropometric OR physical OR activity* OR alcohol OR drink* OR smok* height OR weight OR BMI OR body mass index) OR (rad* OR mammogram OR MD OR breast density OR screening) OR (benign OR invasive OR cancerous OR breast disease).

Terms related to BC recurrence or prognosis were excluded as these terms were not in the scope of reviewed areas. The included studies were MR studies as the robust evidence for causality, systematic reviews, meta-analysis, prospective studies and case-control studies. Relevant references appeared in some of the selected studies and were also manually reviewed. The search results were further refined to only include: peer reviewed articles, articles with full lists of references, articles with a publication date, articles with full text, and a clear methodology. Articles were included if they met the above criteria.

### 2.2.1 Study selection

The scoping review consisted of the evidence gathered from cohort studies, case-control studies, nested case-control studies, experimental studies, systematic reviews and meta-analyses amd MR studies. All reviewed papers were published in English language. The papers included spanned from 1st January 2010 to 31st July 2020. This time limitation was specified to obtain most recent up-to-date studies as breast cancer has been well researched over the past decades with many results from large well-designed study. For some factors, if they were limited studies within this period, evidences established prior to 1st January 2010 was included as appropriate. The review process prioritised MR studies, prospective cohort studies and meta-analyses when the level of significance of the risk factors had been evalutaed. However, a study's limitations was taken into consideration and each study was assessed on its own merit. The quality of the papers was reviewed by the research team. The assessment criteria were:

- Studies reporting a clear methodology and results sections
- Number of subjects, when, how and where the subjects were recruited was reported (both cases and controls)

- Use of appropriate statistical analysis

### 2.2.2 Level of significance assessment for risk factors

Level of significance was evaluated based on the Harvard report [42] to categorise significant, probable and possible risk factors of BC. This categorisation is not an indication of causal relationship but rather an estimation of magnitude of the association between the exposure and the outcome of interest. There are other criteria available including the recently published IACR 2020 criteria [43] and the WRCF criteria [44]. The degree of confidence was defined by the Harvard team [42] as follows:

- Significant - an established association between outcome and exposure where bias [systematic error], chance, confounders [misrepresentation of an association by unmeasured factor/s] are eliminated with significant levels of confidence.
- Probable - an association exists between the outcome and the exposure where bias, chance, confounders cannot be eliminated with sufficient confidence – inconsistent results found with different studies.
- Possible - inconclusive or insufficient evidence of an association between the outcome and the exposure – studies with unsatisfactory quality or statistical power to confirm the association).

Studies included in this scoping review were MR, systematic review, meta-analysis, prospective and case-control studies. The effect estimate of the observational studies might be affected by bias, confounders, and reverse causation. However, MR studies still might be affected with confounders but less likely than the observational studies. The level of evidence was therefore assessed by hierarchical evidence from MR study and/or clinical trial study followed by systematic review meta-analysis, followed by cohort study and lastly case-control study. Besides, each evidence was assessed individually before making any conclusion.

## 2.3 Results

The published risk factors for BC available in the literature were classified into three groups: modifiable, partially modifiable, and non-modifiable risk factors.

### 2.3.1 Modifiable risk factors

#### Physical activity

**Evidence:** Results from a meta-analysis conducted in 2019 by the WCRF (22,900 pre- and 103,000 post-menopausal BC cases) reported a statistically significant inverse association between vigorous physical activity (reported either as MET-hour/week or minutes/day) and both pre- and post-menopausal BC when comparing highest to lowest level [45]. Relative risks for the pre- 0.79 (95% CI 0.69–0.91) and for the post- 0.86 (95% CI 0.78–0.94) were observed. Another recent prospective [46] study published in 2020 (47,456 pre- and 126,704 post-menopausal females in UK Biobank) reported a reduced risk of BC among both pre- (RR 0.75; 95% CI 0.60–0.93) and post- (RR 0.87; 95% CI 0.78–0.98) after adjusting for adiposity. Moreover, a systematic review of 19 cohort studies and 29 case-control studies [47] suggested an inverse association between physical activity and BC with stronger evidence among post-menopausal BC (risk reduction ranging from 20% to 80%) compared to pre-menopausal BC (evidence was weak and judged to be indecisive) . A recent MR study in 2020 confirmed an inverse association between ER (Estrogen receptor) + BC (Breast Cancer) and physical activity (one-unit increase in accelerometer-measured physical activity was associated with 49% reduced risk of ER+ BC, OR=0.51 (95% CI, 0.27-0.98)) [48]. More studies [49-51] also supported an inverse association between physical activity in postmenopausal BC whilst the association with premenopausal BC was not clear. The evidence on physical activity was consistent amongst all types of studies: MR study, meta-analysis, prospective study, and systematic review. Thus physical activity can be classified as significant risk factor.

**Level of evidence:** Evidence: Decrease Risk among postmenopausal - Inconsistent results in pre-menopause group– significant risk factor.

**Availability in UK biobank:** Yes

### Alcohol

**Evidence**: The IARC (International Agency for Research on Cancer) has classified alcohol as a cause of BC [52], with a 7-10% risk increase in each 10g consumed daily (approximately 1 drink/day) [53, 54]. A review of 53 cohort and case-control studies [55] reported a risk of 1.32 (95% CI, 1.19-1.45) for an alcohol intake of 35-44 g/day and a risk of 1.46 (95% CI, 1.33-1.61) for an alcohol intake of ⩾45 g /day. Additionally, they concluded that 4% of female BC patients from developed countries were attributed to alcohol consumption. Assessing alcohol intake with menopausal status and alcohol type was carried out by the Petri research team [56]. The authors concluded that >27 drinks/ week increased BC risk among premenopausal females regardless the type of alcohol. While amongst postmenopausal women a consumption of >6 drinks/ week increased BC risk. The Million Women Study revealed the strongest BC risk among females consuming at least 15 drinks/ week (RR 1.29; 95% CI 1.23–1.35) [57]. More recent studies, prospective studies [58], meta-analysis [59], and a review [60], supported an established association between lifetime alcohol consumption and BC. Nevertheless, the findings from MR study based on UKBiobank data published in 2019, concluded that there was no casual association between BC and alcohol intake (OR=0.96 with 95%CI, 0.77-1.18) [61]. In order to accept the final conclusion of the MR study of no causal relationship between the alcohol intake and the risk of BC, I evaluated the MR study and found several limitations either observed or as reported in the paper. The observed limitations were: the full list of the genetic instruments used to assess the causality in breast cancer was not reported, the value of the genetic instrument's variance was not reported, the authors did not discuss the MR assumptions and did not show an evidence of no violation. Moreover, there was no stratification in the analysis based on the menopausal status. In addition to what I observed, the authors also self-reported several limitations on this MR study. This study reported the lack of statistical power despite the very large sample size of 322,193 individuals of UK biobank. This lack of power was caused by the relatively small number of events observed. Another limitation was that the individuals included were better educated, more affluent, healthier than the average UK population at the same age range and had a lower alcohol consumption [49].

The alcohol association with BC risk was inconsistent between above mentioned MR study and other reviewed prospective and retrospective studies. Evidence from the MR study using data from the UKBiobank did not show causality of alcohol and breast cancer. However, observational results reported a significant association. As a result, this risk factor was identified only as a probable risk factor.

**Level of evidence:** Increases Risk – inconsistent results - The effect varies depending on menopausal status, diet, BMI, benign disease history, and breast density- Probable risk factor.

**Availability in UK biobank:** Yes

**Smoking**

**Evidence:** The updated report of IARC 2004 stated that smoking was positively associated with BC [62, 63]. In 2009, the Canadian Expert panel on tobacco and BC concluded that active smoking was associated with BC [64]. Other evidence reported from the American Association of Surgeons [65, 66] and the Canadian Expert Panel [64] also concluded a positive association between BC and smoking. Other studies (a prospective study [67] and a case-control [68]) found an association between smoking and hormone receptor positive BC but no association with triple negative BC [67, 68]. Moreover, a recent meta-analysis of approximately 40,000 cases from 11 prospective studies reported that smoking increases the risk of BC mortality with an RR of 1.10 (95% CI, 1.04–1.16) [69].

However, early childhood or before the first pregnancy exposure to tobacco did not appear to increase the risk of BC [70]. A meta-analysis of 11 studies [71] and a prospective cohort study [72] concluded that there was no association between BC and smoking before the first birth. A study using the UK biobank cohort reported a non-significant association between BC and smoking with RR of 0.92 (95%CI, 0.81-1.06) [73]. A retrospective study conducted in 2017 also concluded no clinical significance on tumour characteristics and ER-, HER2- and early BC [62].

The IARC report, Canadian expert report, American association of surgeons, all reported a positive association. However, other studies such as the UKBiobank based study [73] concluded no association; this might be caused by the healthier effect of the UKBiobank participants recruited. Other studies including meta-analyses [71] and prospective studies [72] also showed no association; but this negative association was between BC risk and smoking before the first birth not the general smoking effect. The inconsistency between studies on the effect of smoking on BC risk qualifies it as being regarded as a probable risk factor.

**Level of evidence:** Evidence: Increase Risk - Inconsistent results - Slight association with active smoking (long, early, and heavy consumption) – Probable risk factor.

**Availability in UK biobank:** Yes

**BMI**

**Evidence**: The 2019 WCRF report confirmed an inverse association of BMI and pre-menopausal BC with a RR 0.94 (95% CI 0.91–0.98) per 5 kg/m$^2$ [45]. Additionally, a positive association of BMI with post-menopausal BC was reported with a RR of 1.12 (95% CI 1.10–1.15) per 5 kg/m2. The results from a prospective cohort from

UK Biobank showed a RR of pre-menopausal females of 0.90 (95% CI 0.96-0.99) and for post-menopausal, a RR of 1.02 (95% CI 1.01-1.03) [74]. A Norwegian prospective study of 1663 BC cases and 99,717 controls suggested a decreased risk of BC among overweight and obese females who had no family history of BC. Nevertheless, any protective effect disappeared in females with a BC family history (both in overweight and obese pre-menopausal females) [75]. Even though, obesity in pre-menopausal women is a protective factor still obesity is associated with poor prognosis and increased BC mortality [76]. Moreover, BMI was tested using the MR approach and found no causal relationship between BMI and BC [77] among hormone receptor-positive and negative BC. However, another MR study in 2019 confirmed that a genetically high level of plasma HDL is associated with an increase of BC risk (OR=1.08 (95% CI, 1.04-1.13)) [78]. Almost all evidence (MR studies, WCRF report, and prospective studies) concluded there was a positive association between BMI and BC risk depending on the indivuduals menopausal status. This level of agreement suggests it as being a significant risk factor.

**Level of evidence:** Evidence: Increase Risk - Well established factor – increased risk among postmenopausal and decreased risk among pre-menopausal – Significant risk factor

**Availability in UK biobank:** Yes

**Hormonal factors**

**OC use**

**Evidence:** A recent prospective study of a Caucasian population [79] with 11,517 BC cases reported a RR of 1.20 (95% CI, 1.14 to 1.26) with BC risk among all current users of hormonal contraceptives. The duration of OC use affects the magnitude of the BC risk. The risk increased from RR of 1.09 (95% CI, 0.96 to 1.23) for females with OC use less than a year to RR of 1.38 (95% CI, 1.26 to 1.51) for females with OC use more than 10 years [79]. Another large prospective study of 116,608 female participants reported that current use of OC is associated with higher risk of BC with RR of 1.33 (95%CI, 1.03-1.73) while past use was not significantly associated with higher risk of BC [80]. Re-analysis of the collaborative Group studies (54 studies) provided the most comprehensive evidences on OC and BC risk. They showed higher risk of BC with a RR of 1.24 (95% CI, 1.15-1.33) among females who were currently using OC [81].

Collaborative analysis of 54 studies (cohort and case-control) reported almost no difference among non-users and females who had stopped using combined OC for 10 years or more (RR 1.01 (95% CI, 0.96-1.05) [81]. Duration of OC use, dose and type of hormones and age at first use, showed no significant effect on BC risk [9]. UK Biobank  results did not confirm the association between BC and OC use even with menopausal stratification , (RR of 1.26 (95% CI, 0.95-1.67) among pre-menopausal and a RR of 1.124 (95% CI, 0.99-1.27) among post-menopausal women [74].

The evidence in some studies [79] [81] [9] showed positive association while the UKBiobank study [74] concluded no association existed even after menopausal stratification. As a result, oc use was considered to be a probable risk factor.

**Level of evidence:** Increase Risk - Not a conclusive factor – Probable risk factor

**Availability in UK biobank:** Yes

**HRT**

**Evidence :** Evidence from the collaborative Group [82] of 51 epidemiological studies (52,705 BC cases and 108,411 controls) reported a RR of 1.35 (95% CI 1.21–1.49) among females who used HRT for 5 years or more. Results from the Million Women [83] Study showed higher risk of BC among current users of HRT with a RR 1.66 (95% CI 1.58–1.75). A more recent study based on UK biobank data reported a higher risk of post-menopausal BC among users of HRT with a RR of 1.14 (95%CI, 1.04-1.26) [74].

Overall, using oestrogen-progestagen combined HRT was found to have little advantage compared to using oestrogen-only HRT among non-hysterectomised females. However, 5 years use of either type of HRTs resulted in 5-6 extra cancer cases per 1000 females and 15-19 extra cases per 1000 among HRT use of 10 years. The extra endometrium cancer cases were mainly amongst oestrogen only HRT while the extra BC cases were mostly among oestrogen-progestagen HRT [84]. However, not all studies supported this association. In a randomised trial (188 with HRT and 190 of no HRT) with 10.8 years of follow up, the results showed no association between BC risk and HRT use with HR of 1.3 (95% CI, 0.9-1.9) [85]. It is noted that this study reported OR to only one decimal place. Whilst inconsistent results were reported in the reviewed studies; many studies confirmed the association and others failed to do so. Using HRT was considered to be a probable risk factor for BC.

**Level of evidence:** Evidence: Increase Risk - In both oestrogen only and combined oestrogen/progesterone preparations – Probable risk factor.

**Availability in UK biobank:** Yes

**Diet**

**Evidence**: The WCRF report of the Continuous Update Project (CUP), the largest source for cancer prevention based on nutrition, diet, and physical activity, reported as follows.

: Limited suggestive evidence (not significant but was consistent) that consuming non-starchy vegetables can decrease the risk of ER- breast cancer

: Consuming food containing carotenoids, and food high in calcium decrease the risk of both pre- and post-menopausal BC [36].

: Dairy product decreased the risk among pre-menopausal BC only.

For non-starchy diet, the 2017 CUP meta-analysis of 12 studies [36] reported a RR 0.98 (95%CI, 0.93-1.02), while 2013 pooling projects of 20 studies [86] reported a RR 0.99 (95%CI, 0.95-1.04), and CUP additional analysis of 25 studies (2017) reported RR of 0.97 (95%CI, 0.91-1.02) [36]. However, CUP 2017 and the pooling project 2013 reported a RR 0.79 ((95%CI, 0.63-0.98) and a RR 0.82 (95%CI, 0.74-0.90), respectively when ER- breast cancer risk was assessed. Total carotenoid level was associated inversely with BC with a RR of 0.82 (95%CI, 0.71-0.96) [36]. Nevertheless, this inverse association of beta-carotene and BC risk was not confirmed by a MR study conducted in 2019 [87]. Moreover, dairy products were assessed by CUP 2017 using 7 studies and showed a RR of 0.95 (95%CI, 0.92-0.99) and another meta-analysis reported a RR of 0.97 (95%CI, 0.63-0.99) with pre-menopausal BC risk [88]. Nevertheless, no significant association was observed between dairy products and post-menopausal BC. Furthermore, meta-analysis of 5 studies (2980 BC cases) showed a 13% decrease in pre-menopausal BC risk per 300 milligram pf dietary calcium per day (RR 0.87 (95%CI, 0.76-0.99)). While meta-analysis of 6 studies (10,137) showed a 4% decrease in post-menopausal BC risk per 300 milligram pf dietary calcium per day (RR 0.96 (95%CI, 0.94-0.99)) [36].

Additionally, red meat consumption reported to increase BC in the pre- and post-menopausal BC [89]. Meta-analysis of prospective studies (N=6) concluded high risk of BC with processed meat (RR=1.09 (95%CI, 1.03-1.16) and suggested no significant association with unprocessed meat (RR=1.06 (95% CI, 0.99-1.14) [90]. Another prospective study 2018 based on UKBiobank confirmed the BC risk with processed meat (HR=1.21 (95% CI, 1.08-1.35) but not with unprocessed red meat (HR=0.99 (95% CI, 0.88-1.12) in the highest tertile consumer with (>9 g/day) [91]. Meta-analysis of ten studies (2009) reported relative risk of 1.57 (95% CI, 1.23-1.99) among case-control studies (N=7) and a RR of 1.11 (95% CI, 0.94-1.31) among cohort studies (N=3) [92]. Furthermore, using EPIC prospective study with 7119 BC cases concluded modest increase in BC risk with a HR of 1.10 (95% CI, 1.00-1.20) when comparing high versus low consumption of processed meat [93]. A case-control study (2011) including 2,386 BC cases and 1,703 controls confirmed the positive association between well-done red meat and BC risk with an OR= 1.5 (95% CI,1.3-1.9) [94]. It is noted that this study reported ORs to only one digit. Further research is needed to assess the relationship between the specific types of food and BC risk. Adopting healthier diet can help in reducing the risk of BC.

**Level of evidence:** Decrease Risk - Inconsistent results – Probable risk factor.

**Availability in UK biobank:** Yes

### 2.3.2 Partially modifiable risk factors

**Childbearing related factors**

**Age at first birth**

**Evidence**: Age of first child is considered as a determinant of BC incidence [95]. Risk of dying from BC was significantly decreased for females who had first child at age 20 - 24 (RR 0.88; 95% CI 0.78-0.99) and at age 25 -29 (RR 0.80; 95% CI 0.70-0.91) compared to females who had their first child at age 20 and below [96]. The findings from the UK Biobank cohort study confirmed that increasing age of having first child increases breast cancer risk among pre-menopausal females [74]. Females who had their first full birth at age 25-29 had RR of 1.88 (95% CI, 1.04-3.42) and females who had their first full birth at age ≥30 had RR of 1.94 (95% CI, 1.06-3.54) compared to females who had their first full birth at age <20 among UK biobank cohort. The same conclusion was reported by these two meta-analyses with a higher risk of BC with older age of first birth [81, 97]. All studies supported an evidence base which confirmed the protective effect of early age of first birth and lower BC risk.

**Level of evidence:** Decrease Risk - Well established factor – Probable risk factor.

**Availability in UK biobank:** Yes

### Parity

**Evidence**: High parity level is associated with low BC risk among females [74] diagnosed after the age of 45 years according to results from the UK biobank study. The RRs were reported according to menopausal status with a RR of 0.76 (95%CI, 0.64-0.91) among pre-menopausal females and a RR of 0.82 (95%CI, 0.73 -0.93) among post-menopausal females. It was also suggested that the number of children increases, the risk of BC decreases. Females with at least one full-term pregnancy have a 25% reduction in BC risk compared to nulliparous females [98]. A meta-analysis from Nordic countries (three cohort and five case-control studies with 10,703 total participants and 5,568 BC cases) showed that females who gave birth at age younger than 20 years had a 30% lower BC risk compared to females who gave birth after the age of 35 years [99]. Another findings from the UK biobank study concluded that parity was strongly associated with ductal carcinoma in situ (DCIS) (RR=0.40 (95%CI, 0.21-0.79)) [100]. Moreover, a recent prospective study based on the US Nurses Health studies confirmed that parity was inversely associated with the risk of ER+ breast cancer (HR = 0.82 (95% CI, 0.77–0.88) but no association with ER- breast cancer (HR=0.98 (95% CI, 0.84–1.13)[101]. Evidence [74] [100] based on UKBiobank , meta-analysis based on Nordic populations and the US Nurses prospective studies, all confirmed an inverse association between having children and low risk of BC even correcting for menopausal stratification.

Level of evidence: Increase Risk - Well established factor – Probable risk factor.

Availability in UK biobank: Yes

**Breastfeeding**

**Evidence**: The effect of the breastfeeding on BC is still inconclusive, however one of the social recommendations of WRCF/AICR for cancer prevention was breastfeeding [102]. A recent systematic meta-analysis [103, 104] investigated adherence to WRCF/IACR recommendations on cancer prevention and mortality. They confirmed that adhering to the 2007 WCRF/AICR recommendation (including breastfeeding) lowers the risk of BC. A meta-analysis (27 studies involving 13,907 BC cases) [105] concluded that breastfeeding, especially for long duration was inversely associated with BC risk. The RR was 0.61 (95% CI, 0.44–0.85) when comparing ever breastfed with never breastfed females, and RR was 0.47 (95% CI, 0.37–0.60) when comparing long to short duration of breastfeeding. Another meta-analysis (47 epidemiological prospective and case-control studies) showed a 4.3% reduction in BC risk for every 12 months of breastfeeding [106]. However, systematic review of 24 studies concluded that 13 out 24 studies reported a reduced risk of BC with breastfeeding [107]. Breastfeeding is considered to be protective factor for BC and the WRCF recommended breastfeeding for cancer prevention. The follow up studies [103, 104] confirmed that adhering to this recommendation was shown to lower the risk of BC.

**Level of evidence:** Decrease Risk - Not conclusive - As the breastfeeding period increase the risk decrease – Probable risk factor.

**Availability in UK biobank:** No

### 2.3.3 Non-modifiable risk factors

**Age**

**Evidence**: Aging is the most important known risk factor of BC after gender [108]. BC incidence increases with age and can reach its peak at the age of menopause and later it decreases gradually until becoming constant [109]. In the UK (2010 to 2012), about 80% of cases were over the 50s and about 24% were 75 years old or more

[110]. In the recent prospective study of UK biobank cohort, age proved to be significantly associated with pre- (RR 1.46 (95% CI, 1.02-1.07 ) and post-menopausal (RR 1.03 (95% CI, 1.02-1.04) BC [74]. Even though BC incidence increases with age, BC occurs in younger females and it appears to be more aggressive with larger sizes, more advanced stage, affected lymph nodes and with weaker survival rates [111]. All evidence from reported studies concluded a positive association between the BC risk and increasing age.

**Level of evidence:** Increase Risk - Well established factor – significant risk factor.

**Availability in UK biobank:** Yes

### Hereditary risk factors

#### Genetic factors

**Evidence:** BC susceptibility genes with risk alleles are grouped into three categories depending on their risk and frequency. High-penetrance gene variants: high risk of more than 4 but very rare with minor allele frequency (MAF) <0.005. Second are moderate-penetrance gene variants: their risk between 2 to 4 and they are rare with MAF of 0.005-0.01. The last group are low-penetrance gene variants: their risk contribution is less than 1.5 but they are common with MAF >0.05 [112].

Multiple genetic risk factors contribute to the BC development, however approximately 20% of the hereditary BC are caused by BRCA1 and BRCA2 gene mutations and a further 5-10% are attributed to mutations in other rare susceptibility genes such as TP53, STK11, PTEN, ATM and CHEK2. Furthermore, low-risk common variants associated with breast cancer in excess of 90 loci may contribute to a further 23% of the heritability [113]. One study reported that 55%–65% of BRCA1 mutation carriers and 45% of BRCA2 carriers developed BC by the age of 70 [114]. Moreover, a prospective study showed that the risk of cumulative BC by age 80 was 72% among BCRA1 carriers and 69% risk among BRCA2 carriers [115].

Another genetic factor is mutation in TP53 gene. By the age of 30, females with a mutation in TP53 have 30% risk of BC and about 18 to 60 fold-risk for developing BC at age of < 45 years compared to the general population [116]. Less than 1% of familial BC is caused by this mutation. Additionally, matrix metalloproteinase (MMP-2 c-735-T) gene polymorphisms are associated with the risk of BC at young age by 1.64-fold with OR=1.64; 95% CI, 1.01–2.70 [117]. These mutations are classified as high-penetrance low frequency mutations.

Mutations in moderate-penetrance (low frequency) gene variants such as in *CHEK2, PALB2, BRIP1, ATM, CHD1* are also associated with higher risk of BC [118]. Additionally, GWAS (genome wide association studies) have identified so far more than 180 loci (low-penetrance high frequency) associated significantly with BC risk and collectively accounted for 18% of BC heritability [112, 119, 120].

High penetrant mutations explain only about 20% of the familial BC [121] and the moderate-penetrant variants explains about 5% of the familial BC [122] and low penetrant explains about 18% of the familial BC [123]. A large GWAS BC study (2019) consisted of ten prospective studies analysed 94,075 BC cases and 75,017 controls and developed the PRS to predict BC. The study identified 313 SNPs. The study reported that for overall disease per 1 standard deviation was 1.61 (95%CI: 1.57–1.65) with area under receiver-operator curve (AUC) = 0.63 (95%CI: 0.63–0.65) [120]. The study also reported that women in the top centile of the PRSs, have a lifetime risk of overall breast cancer of 32.6%. Furthermore, compared with women in the middle quintile, those in the highest 1% of risk had 4.37- and 2.78-fold risks, and those in the lowest 1% of risk had 0.16- and 0.27-fold risks, of developing ER-positive and ER-negative disease, respectively. The SNPs considered in this thesis were a combination of high, moderate, and low-penetrant genes. There is an agreement between all types of studies on the strong association between BC risk and BC risk genetic mutations.

**Level of evidence:** Increase Risk - Well established factor – Significant risk factor**.**

**Availability in UK biobank:** Yes

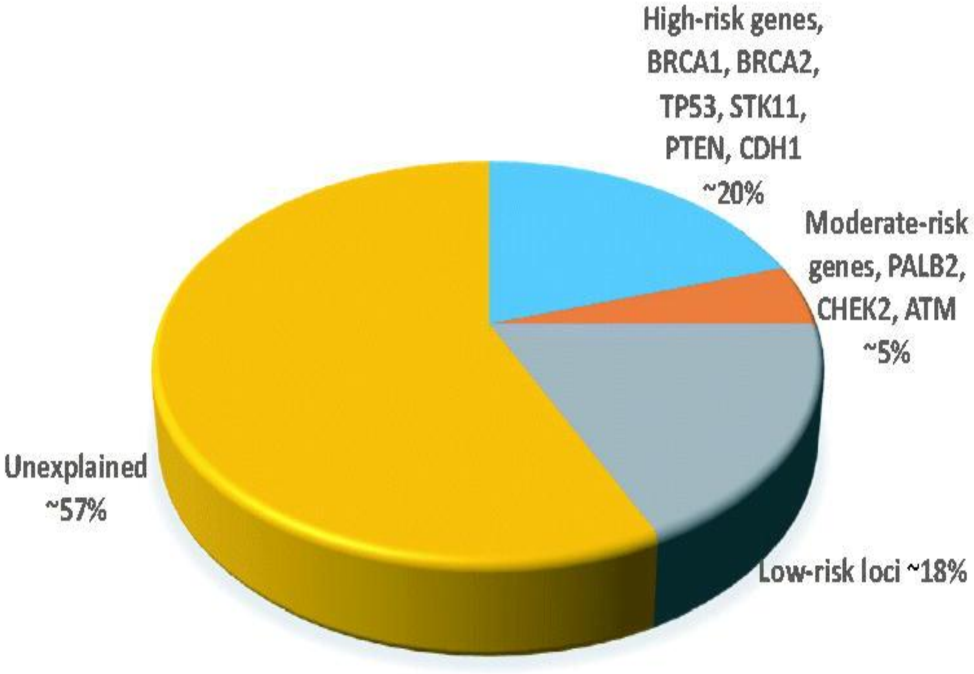


Figure 2.1: The distribution of familial BC risk explained by the currently known susceptibility genes [112].

### Family History of breast cancer

**Evidence:** Family history of BC is a well-established risk factor for BC [74, 124, 125]. A female with either mother or sister with BC has 2-3 times increased BC risk [74, 108]. It has been reported that females with BC family history (two or more developed BC younger than 50 years or three at any age) have 11 times chance to develop BC even if they had no BRCA mutations [126]. Regardless of menopause status, the estimated risks were higher in females who reported only their sibling(s) (Pre RR (1.82 (95% CI,1.21-2.76) and post RR (1.61 (95% CI, 1.34-1.94)) affected with BC as compared to females who reported only their mother (Pre RR (1.72 (95%

CI,1.36-2.18) and post RR (1.57 (95% CI, 1.35-1.82)) affected with BC [74]. All evidence supported the strong association between the BC family history and the risk of BC.

**Level of evidence:** Increase Risk - Well established (depends on type, number, and age of relative/s when developed the disease) - Especially with ovarian cancer – Significant risk factor.

**Availability in UK biobank:** Yes

**Hormonal factors:**

**Early menarche Age**

**Evidence:**

Evidence from a meta-analysis of 117 epidemiological studies including (118,964 BC cases and 306,091 controls) showed an elevated risk of BC for every year younger at menarche with RR of 1.05 (95% CI, 1.04-1.06) [127]. A recent UK biobank study showed, early age at first menarche onset was associated with increased risk of pre-menopausal BC [74] with RR  1.23 (95% CI, 1.04-1.45). Moreover, pre-menopausal BC was reduced by 7% and post-menopausal BC by 3% for each year of delay in menarche after the age of 12 [128]. All evidence (prospective, retrospective, and meta-analysis studies) confirmed a positive association between early menarche age and BC risk.

**Level of evidence:** Increase Risk - Well established factor - More effective than late menopause age – Probable risk factor.

**Availability in UK biobank:** Yes

**Late menopause age**

**Evidence:** Many studies with different designs concluded the positive association between BC and late menopause age (menopause age over 50 years) [108, 109, 129, 130]. Later age at menopause onset associates with increasing BC risk by 3% for each year of menopause delay [82, 127]. Meta-analysis of 117 studies [127] showed BC increased risk of RR=1.03 (95% CI, 1.02–1.03) for each year delay in menopause. All evidence (prospective, retrospective, and meta-analysis studies) confirmed the positive association between late menopause age and BC risk.

**Level of evidence:** Increase Risk - Well established factor - Probable risk factor.

**Availability in UK biobank:** Yes

**Breast related factors:**

**Benign Breast Disease (BBD)**

**Evidence**: Breast proliferative disease without atypia and with atypia showed an increased risk of BC [131-134]. According to multi-centre prospective study (615 cases and 624 controls), results from nested case-control analysis suggested that OR of BBD without atypia was 1.45 (95% CI 1.10–1.90) and BBD with atypia was 5.27 (95% CI 2.29–12.15) compared to normal pathology of the breast [131]. It has been suggested a RR of 1.5 to 1.6 among females with benign breast diseases compared to females of the general population [135]. The RR of the non-proliferative disease was 1.27 (95% CI 1.15-1.41) while proliferative changes without atypia had a RR of 1.88 (95% CI 1.66-2.12), and the RR of atypical hyperplasia was 4.24 (95% CI 3.26-5.41). Females with severe atypical epithelial hyperplasia have 4 to 5 times higher risk to develop BC more than females without these changes [9]. This risk might increase to 9 folds if the female has a family history of BC with the breast proliferative changes. Females with complex fibro-adenomas, palpable cysts, sclerosis adenosis, duct papillomas, and moderate or florid epithelial hyperplasia have a slightly higher risk of BC (1.5-3.0 times more) than females without these conditions [9]. All evidence supported the positive association between BC and benign breast diseases with proliferative or non-proliferative disease.

**Level of evidence:** Increase Risk - Well established factor – significant risk factor.

**Availability in UK biobank:** No

### Breast Density

**Evidences:** High mammographic density (MD) is a well-established risk factor of BC [136-140] . In a systematic meta-analysis of (14,000 BC cases and 226,000 controls) from 42 studies, they concluded the positive association between MD and risk of BC [141]. About 16% to 32% of BC cases are related to high MD primarily among premenopausal women [142]. The pooled RRs of 5% to 24% MD was 1.79 (95% CI 1.48-2.16), RR of 25% to 49% MD was 2.11 (95% CI 1.70-2.63), RR of 50% to 74% MD was 2.92 (95% CI 2.49-3.42), and RR of >74% MD was 4.64 (95% CI 3.64-5.91) relative to <5% group among incidence studies [141]. High MD is an important risk factor alongside with age, carrying high penetrance genes (*BRCA1/2*), and the presence of atypia on a breast biopsy [142]. All evidence supported a positive association between BC and high MD.

**Level of evidence:** Increase Risk - Well established factor – significant risk factor.

**Availability in UK biobank:** No

### Height

**Evidence**: According to the updated report by WRCF 2018, height was reported to have a strong association with both pre- (RR=1.06 (95% CI, 1.02-1.11)) and post-menopausal (RR=1.09 (95% CI, 1.07-1.11)) BC [143]. The findings from MR causal studies between height and BC supported role of height as a causal factor. The study carried out in 2018 [144] reported HR of 1.09 per 10 cm increase in height, with (95% CI,1.02 to 1.17). Another MR study in 2015 reported OR of 1.22 (95% CI, 1.13 to 1.32) in the first consortium (46325 cases and 42482 control) and 1.21 (95% CI, 1.05 to 1.39) in the second consortium (16003 cases and 41335 control) [145]. Results from large cohort study (the EPIC cohort) [146] reported a positive association between height and post-menopausal BC (RR 1.10 with 95% CI 1.05–

1.16). Furthermore, a meta-analysis of 159 prospective studies showed a pooled BC RR of 1.17 (95% CI = 1.15 - 1.19) per 10cm increase in height [147, 148]. Another pooled analysis of prospective studies also suggested positive association among post-menopausal females (RR=1.07 with 95% CI: 1.03, 1.12) [149]. The UK biobank study showed a RR of 1.18 (95% CI = 1.04 - 1.34) per 10cm increase in height among pre-menopausal and a RR of 1.23 (95% CI = 1.14 - 1.33) per 10cm increase in height among post-menopausal. Not all prospective studies confirmed the positive association. A BC register-based cohort study with 13,572 participants concluded no statistical evidence of association between height and BC risk (OR=1.06 (95% CI, 0.98 to 1.14) [150]. Evidence from case-control studies was inconsistent.

Evidence from MR studies [144, 145], the WRCF report [143] and prospective and meta-analysis studies, all confirmed a relationship between height and BC. However, register-based cohort and case-control studies failed to confirm this association. As the positive association was confirmed by stronger studies, height was considered as being a significant risk factor.

**Level of evidence:** Increase Risk - Well established factor - More effective among postmenopausal – Significant risk factor.

**Availability in UK biobank:** Yes

**Abortion**

**Evidence**: Induced abortions and spontaneous miscarriages showed inconsistent results [74, 151, 152]. A pooled analysis of 53 prospective and retrospective studies [152] concluded that spontaneous (RR=0·98 (95% CI, 0·92–1·04) was not associated with the risk of BC while induced abortion had reduced risk (RR=0·93 (95% CI, 0·89–0·96)). More recent meta-analysis based only on prospective studies (Fifteen prospective studies) showed significant association between spontaneous (RR=1.02 ((95%CI, 0.95-1.09)) and induced abortion (RR=1.00 (95%CI, 0.94-1.05)) with BC

[153]. Recently in 2020, another meta-analysis (six cohort and eight case-control studies) was conducted and concluded no significant association between BC and abortion even among nulliparous females ((RR = 1.02 (95%CI, 0.94–1.12) [154]. Most studies concluded no association between abortion (spontaneous or induced) and BC risk. However, a positive association was shown by few studies. As consequence, I categorised it as possible risk factor.

**Level of evidence:** Increase Risk - Inconsistent results or insufficient evidence – Possible risk factor.

**Availability in UK biobank:** Yes

**Radiation**

**Evidence**: Ionising radiation is a well-established risk factor of breast cancer [155-157]. The WCRF reported in 2017 that exposure to the ionising radiation even from the medical treatments can increase the BC risk even with low doses [36]. The relative risk  ranges from 1.1 to 2.7 at 1 Gy for exposed females before the age of 40 [151]. Whether it is single exposure or multiple exposures, it results in an equal total radiation dose [158]. This confirmed it as being a risk factor in all reviewed studies.

**Level of evidence:** Increase Risk - Well established factor – Significant risk factor

**Availability in UK biobank:** No

## 2.4 Discussion

Factors associated with breast cancer have been reviewed using four different search engines. To select relevant articles, further checks were applied to ensure quality and validity.  Evidence of each risk factor was described, and level of significance was evaluated.

The level of evidence was categorised into three levels based on the Harvard report [42]. The first level is ‘significant’ (bias, chance, and confounders were eliminated with significant confidence), the second level is ‘probable’ (bias, chance, and confounders were eliminated with sufficient confidence), and the third level is ‘possible’ (inconclusive or insufficient evidence of an association between the outcome and the exposure). Risk factors that showed significant evidence were genetic factors, family history of BC, radiation, height, physical activity, and BMI. Probable risk factors were age, smoking, OC use, HRT use, diet, early menarche age, late menopause age, benign breast diseases, breast density, null-parity, age at first birth, alcohol, and breastfeeding. Lastly, spontaneous miscarriage or induced abortion was a possible risk factor.

The studies included in this scoping review were MR studies, systematic review meta-analysis, prospective studies, and case-control studies. Bias, confounders, and reverse causation could affect the results of these studies except from results of MR studies.

Starting with case-control studies which were frequently reported in the literature more than other type of epidemiological studies. This was because of the long latency period in cancer study and the low cancer incidence rate [159]. Case-control studies can lead to important scientific findings with less effort, money, and time compared to prospective studies. However, they are more susceptible to bias [160] and could affect the true association between the exposure and breast cancer risk. Addressing the confounding issue in the design phase (either by matching or restriction) or in the analytical phase (stratification) of the case-control study can strengthen the final findings/results [159].

Evidence from cohort studies included in the review are more reliable and could identify possible causal relationship between exposure and outcome [161]. This is due to the nature of the prospective cohort studies. Subjects are recruited based on presence and absence of the exposure/s and have been followed for period of time to see whether they develop the relevant outcome. Another advantage includes the ability to study multiple exposures and multiple outcomes at the same cohort [162]. In addition, the combined effect of multiple exposures (such as BMI and diet on BC risk) can be estimated [162]. Even with this study type, confounder, selection bias and loss of follow-up still exist. Similarly, biases can be controlled in the design phase (randomisation) or in the analytical phase.

Meta-analysis is a formal quantitative analysis of previous studies sharing the same research question and population. Meta-analysis is a subset of systematic reviews. It can lead to more precise estimate of the exposure on the interested outcome [163, 164]. In the evidence hierarchy, meta-analysis is considered the best evidence-based study. Many BC risk factors discussed in this chapter were drawn from meta-analysis studies and they contributed to more precise risk estimates. Additionally, systematic reviews were also included in this review. The systematic review approach gathers experimental evidence from previous studies to address a specific research question. No quantitative estimate is calculated by this type of research [163]. Systematic approaches are used to minimise bias to provide more reliable scientific findings[165]. Systematic review does not have to be restricted to meta-analysis when it is not possible or valid, however many systematic reviews do contain meta-analyses.

Finally, MR study aims to assess the putative causal relationships between the modifiable exposures and the outcome through using a genetic instrument (SNPs associated with the exposure) [166, 167]. MR study problems arise from observational studies which include unmeasured confounders, limited genetic instruments, and reverse causation [166]. MR study uses genetic variants to distinguish correlation from causation in observational data. The reliability of a MR investigation depends on the validity of the genetic variants as instrumental variables (IVs) [168]. With the availability of results from genetic association studies, the use of genetic instruments for inferring causality in observational epidemiology has become increasingly popular [169].

In this review, evidence of each BC factor from different types of studies were reviewed and evaluated. BC risk factors have been reported in other large UK studies for example the one million women and EPIC-Norfolk study, there was no report on breast cancer risk factor from the UKBiobank study. In this thesis, these risk factors were investigated using the UKBiobank dataset to compare and contrast the findings.  Furthermore, as there was no UK based risk prediction model, the results from this literature review provided suggestive risk factors for development of the BC risk models. The factors reviewed in this chapter informed a priori list of factors to be studied in this thesis.

In conclusion, BC factors that increase BC risk are increasing age, clear family history of BC or ovarian cancer or other benign breast diseases, presence of BC predisposition genes, high breast density, radiation exposure, height, early menarche and late menopause, no children, old first birth , high BMI among post-menopausal and low BMI among pre-menopausal and finally alcohol consumption. Other BC factors that might increase BC risk including sedentary lifestyle, lack of breastfeeding, smoking, using OC and HRT hormones and unhealthy diet. A BC factor classified as a ‘probable factor’ is the induced or spontaneous abortion although it is not yet confirmed.

# Chapter 3 : Methodology

This chapter describes in detail the methods used in chapter 4 (review of non-clinical risk models to aid prevention of breast cancer), chapter 5 (risk of breast cancer in the UK Biobank female cohort and its relationship to anthropometric and reproductive factors), chapter 6 (association of non-genetic factors with breast cancer risk in genetically predisposed groups of women in the UK Biobank cohort) and chapter 7 (development and assessment of breast cancer risk prediction models based on the UK Biobank female cohort).

## 3.1 Methodology of chapter4 (reviewing breast cancer models: first paper methodology)

Existing publications on non-clinical and non-genetic BC models were reviewed to summarise risk factors incorporated into these models, identify the populations that models were derived from and determine model calibration, model validation, and model utility. A systematic literature review was carried out using the following databases to search for relevant publications with no-restriction of publication starting date through to 31th July 2016): PubMed (https://www.ncbi.nlm.nih.gov/pubmed/); ScienceDirect (http://www.sciencedirect.com/); the Cochrane Database of Systematic Reviews (CDSR) (http://www.cochranelibrary.com/). The MeSH (Medical Subject Headings) search terms were "assessment tool, assessment model, risk prediction model, predictive model, prediction score, risk index, breast cancer, breast neoplasm, breast index, Harvard model, Rosner and Colditz model, and Gail model". A **P**referred **R**eporting **I**tems for **S**ystematic **R**eviews and **M**eta-**A**nalyses (PRISMA) approach was applied for selecting the articles to be included in the review [170, 171]. The search criteria were specified to include only if study satisfied the following criteria 1- peer reviewed 2- reference available 3- publication date 4- full text was available. The search term included the following words "breast cancer / neoplasm and prediction). After reviewing all articles, only 316 articles were valid for further review.

The identification step (initial search) resulted in 316 articles. Next was a screening step which aimed to (a) exclude any articles with duplication from the identification step (b) further screen against the risk prediction of BC non re-occurrence and non-mortality of BC). The screening process reduced the number of potential articles down to 61. These articles were further censored for their eligibility (the third step). The eligibility criteria included: any articles relating to breast cancer risk models that incorporated either non-clinical or non-genetic factors, being published in the English language, were full articles, reported methodology in full detail, models that contained variables which were considered to be modifiable and/or self-reported by the respondents. The exclusion criteria were: models were based on genetic risk factors, models with just clinical risk factors, models published with abstract only, models with male BC, female with BC prevalent cases and not incident cases, models with single risk factor investigation such as mammography or genetic profile , and models assessing BC outcome using an invasive techniques such as biopsies. The final filtering step suggested 14 studies that were eligible for inclusion in the review. Further to the MeSH search, the literature search was extended to include publications relating to systematic reviews and meta-analyses. This search strategy yielded no additional relevant publications.

Methodology for chapter 5, 6 and 7 related to data analysis with different aspects using the UKBiobank cohort. The UK based-datasets that can be used to investigate BC risk factors, develop BC risk prediction model were the One Million Women study (http://www.millionwomenstudy.org), the European Prospective Investigation of Cancer (EPIC)-Norfolk (https://www.epic-norfolk.org.uk/about-epic-norfolk/) and the UKBiobank (https://www.ukbiobank.ac.uk/). I did apply for data access for both the One Million women study and the EPIC-Norfolk. The process to gain access of both datasets took a significant longer that I expected and by the time I were to receive the dataset, I would have entered the second half of my 3rd year therefore the most approachable and ready dataset was the UKBiobank.

The following relates to the study population and case-control identification applied to these three chapters.

### 3.1.1 Study population

#### 3.1.1.1 Description of the UK Biobank Project

The UK Biobank is a national health project established by the Wellcome Trust medical charity, Department of Health, Medical Research Council, Scottish Government, and the Northwest Regional Development Agency. The project is ongoing (at the time of the submission of this thesis) and aims to improve the diagnosis, treatment, and prevention of serious diseases such as cancer, diabetes, stroke, heart disease, osteoporosis, arthritis, eye diseases, dementia and depression. A total of 502,650 participants (males and females) aged between 39 to 71 years were enrolled in the study between 2006 and 2010. There are 22 UK assessment facilities across England, Wales, and Scotland. Participants continue to be longitudinally followed up for capture of subsequent health events. Participants gave the UK Biobank written informed consent to use their data and samples for health-related research purposes. Ethics approval was obtained from the North West (Haydock) research ethics committee (reference: 11/NW/0382). Each participant provided biological samples (blood, saliva and urine). Data on health outcome were obtained from (a) self-reported which was verified by study health nurses, (b) from HES (Hospital Episode Statistics), (c) from Primary Care linkage record and (d) from the death registry. Data on demographic, lifestyle, detailed physical / physiological measurements were collected by questionnaire and physical examinations. Many participants completed additional detailed questionnaires on work history, diet, and cognitive function. Anonymised data are available to researchers across the world [40, 172]. More details can be found at http://www.ukbiobank.ac.uk/.

The UK Biobank is an open research source for any researchers around the world including those funded by industry and academia. The initial funding of UK Biobank was £62 million, later an additional £6 million was invested for extra baseline measurements (saliva sampling and eye measures). A further £25M funding was obtained during 2011-2016. As the project grows more baseline assessments were carried out in a large sub-sample such as MRI, DEXA bone scanning, and a wrist activity monitor for 20,000 participants. Data on genotyping was available in all the UKBiobank participants. Data on biochemistry and imaging data are also available for research. The UKBiobank have also been utilised by other research groups to study BC such as cancer epidemiology group at Oxford University, American research group at Albert Einstein College in USA, Epidemiology group at Bristol University in UK, and Genomic Medicine center group in Massachusetts USA etc.

#### 3.1.1.2 Access to the UK Biobank dataset

All users must apply to access the UKBiobank data for a specific research purpose. Data used in chapter 5, 6 and 7 were obtained from the UKBiobank (application number 5791-Development and validation of risk prediction model for breast and ovarian cancers). The application was reviewed and approved by the scientific committee. Once approved, a key file contained a unique code for file download and for appropriate file conversion was sent by the UKB team. The instructions of extracting, accessing, decoding and converting the dataset are explained in details by the UKBiobank as shown in appendix 1 (https://biobank.ndph.ox.ac.uk/~bbdatan/Accessing_UKB_data_v2.3.pdf). The key file was a 64-digit code used to assure more confidentiality. Once downloaded, all files were saved in a secured drive hosted by the University of Manchester. The dataset contained a unique identification number for each participant specific to each application. The UKBiobnak team informed us of any updates relating to withdrawn participants which were subsequently excluded in the data quality control process. The genotype dataset was distributed to each institution that hosted one or more UKBiobank applications. Each application can link to the genotype data using a bridging file. Each research group accessed the genotype data via CSF, a High-Performance Computing (HPC) cluster.

### 3.1.2 Defining breast cancer cases and controls

The UK Biobank database contained a record of all cancers including their subtype, occurring either before or after participant enrolment. Outcome data was derived from data linkage from HES (provided as code derived from International Classification of Diseases (ICD10, ICD9)), data from self-reported and death registry. Details of codes used to identify BC are summarised in table 3.1. Furthermore, the Stata codes used to define breast cancer cases and controls were provided in appendix 3.

#### 3.1.2.1 Breast cancer cases

BC was defined as a malignant neoplasm of the breast (breast cancer BC) in females. There were four sources to identify BC cases (ICD10, ICD9, self-reported, and death registry). All deceased participants were excluded due to their inability to develop the desired outcome. BC cases were characterised as ‘incident’ or ‘prevalent’ using ‘age or date when they first attended the centre’ and ‘age when first reported BC cancer’. The incident cases were identified from the date of attending the assessment centre. Any female who developed BC after that date was considered as being an incident case (no lag period of two years). This was to (1) follow UKBiobank identification method of incident cases – they identified incident cases post recruitment to the biobank study [173] (2) maximise the BC incident cases. Cases were defined by ICD10 and ICD9, if their ‘attending age’ was greater than ‘cancer diagnosis age’ - this was then considered as a prevalent case. Subjects were considered to be incident cases if their ‘attending age’ was less than their ‘cancer diagnosis age’. For self-reported cases, the same criteria were applied. Age when first attended the assessment centre was compared with the interpolated age of the participant when cancer was first diagnosed (reported by the participant herself). Thus, if the BC cases appeared as an incident in any of these three identification sources, then the cases were deemed as being incident cases. In addition, prevalent cases were defined only if they had been identified as prevalent by any of the three sources. BC cases that failed to be identified as an incident or prevalent were further excluded.

### 3.1.2.2 Breast cancer controls

Female participants were defined as controls if they had no record of any cancer, *in-situ* carcinoma, an undefined neoplasm and still alive at the time of data extraction.

Table 3.1: Identification of cases and controls of UK Biobank cohort

| **Categories** | **ICD10 codes** | **ICD9 codes** | **Self-reported cancer’s codes** |
|---|---|---|---|
| **Breast cancer cases** | | | |
| Incident: | Codes start with C50 and its subclasses, C501, C502, C503, C504, C505, C506, C507, C508, and C509 | Codes start with 174 and its subclasses 1741, 1742, 1743, 1744, 1745, 1746, 1747, 1748, and 1749 | 1002 code only |
| Prevalent: | | | |
| **Subjects excluded from the study** | | | |
| 1- Other cancers | Codes start with C except codes for BC | Codes start with 1 or 20 except codes for BC | All other codes except 1002 code |
| 2- Breast In situ carcinoma | Codes of D050, D051, D057, D059 | 2330 code only | - |
| 3- Other in situ carcinoma | Codes start with D0 except codes for breast in situ carcinoma | Codes start with 230 or 231 or 232 or 233 or 234 except codes for breast in situ carcinoma | - |

| Categories | ICD10 codes | ICD9 codes | Self-reported cancer's codes |
|---|---|---|---|
| 4- Neoplasm of unknown nature or behavior | Codes start with D37 or D38 or D39 or D40 or D41 or D42 or D43 or D44 or D45 or D46 or D47 or D48 | Codes start with 235 or 236 or 237 or 238 or 239 | - |
| 5- Dead | Yes or no code, any participant identified as deceased was removed from the analysis | | |
| **Controls included in the analysis** | | | |
| All controls | Remaining participants without any code or with codes other that the codes in the above groups. | | |

### 3.1.3 Defining pre- and post-menopausal status in UK biobank females

All analyses performed in chapter 5, 6 and 7 were stratified by menopause group. The following described the method to define menopause status. All Stata codes used to define menopause status was described in appendix 3.

#### 3.1.3.1 Pre-menopause

The classification criteria used for assigning pre-menopausal status were;

1. Females aged ≤ 55 years old (this age cut-point was based on the NHS's definition of menopause age in the UK which is between 40 to 55 years [174]).

2. Females with menarche age ≥ 7 years old (the UK ranges of menarche age is 7 to 20 years [175]).

3. Females who reported still having their period and had not undergone either hysterectomy or bilateral oophorectomy.

#### 3.1.3.2 Post-menopause

Post-menopausal females were defined using the following criteria.

1. Females who did not report a history of hysterectomy or bilateral oophorectomy and reported no longer having periods.

2. Females with menopause age ≥ 40 years old.

In total, 57,712 pre-menopausal females and 138,554 post-menopausal females were identified. These criteria were employed to minimise inclusion of both pre-mature and the medically induced pre- or post-menopausal women. Following stratification by Caucasian heritage, these numbers were reduced to 43,975 pre-menopausal and 114,721 post-menopausal females.  It is to be noted that study numbers in each chapter may vary slightly due to different censor times effecting the number of withdrawn participants, deceased participants and further case identification.

## 3.2 Methodology of chapter 5 (risk of breast cancer in the UK Biobank Female cohort and its relationship to anthropometric and reproductive factors: second paper methodology).

### 3.2.1 Study population and study design

Data from women within the UK Biobank longitudinal cohort study were used (details were described previously in 3.1.1). The study design was a nested case-control study. BC case and female control identification is described in section 3.1.2.  Participants were considered as a case if they developed the outcome of interest (breast cancer) after they enrolled the study. The total number of pre-menopausal BC incident cases was 618 with 57,089 controls and total number of post-menopausal BC incident cases was 1,757 with 112,757 controls. The exposures were defined prior to the disease or outcome development based on the baseline data collection.

### 3.2.2 Anthropometric and reproductive factors

Summary of anthropometric and reproductive factors and coding approach is shown in table 3.2. Furthermore, Stata codes for defining the anthropometric variables can be found in appendix 3. An assessment of the variables is included in supplementary materials (appendix 4).

Table 3.2: classification of the variables included in the analysis

| Variable | Groups | Coding |
|---|---|---|
| Menopausal status | Pre-menopausal | Pre- menopausal: reported as pre-menopausal & no history of hysterectomy or bilateral oophorectomy & their age is ≤55 and menarche age ≥7 years old (to maximise the number of the real pre-menopausal females - so any female reported as pre- and their age > 55 or had menarche age < 7 and did not had hysterectomy nor oophorectomy will be removed – most probably this female is miscategorised). |
| | Post-menopausal | Post-menopausal: reported as post-menopausal & no history of hysterectomy or bilateral oophorectomy (only natural menopause) & their menopause age is ≥ 40years old ((to maximise the number of the real post-menopausal females – so any female reported as post- and their menopause age < 40 and did not had hysterectomy nor oophorectomy will be removed - most probably this female is miscategorised). |
| Menarche age | (>13) | This variable was divided into two groups based on value ranges derived from literature [114]. |
| | (≤13)) | |
| Age at first birth | (<20) | This variable was divided into four groups based on literature ranges [156]. |
| | (20-24) | |
| | (25-29) | |
| | (≥30) | |
| BMI | Healthy (18.5 – 24.9) | BMI variable was divided into three groups based on WHO classification [157]. Underweight group has very low in number and therefore was excluded. |
| | Overweight (25-29.9) | |
| | Obese (≥30) | |
| Waist to hip ratio (WHR) | Low (≤0.80) | WHR variable was calculated by dividing the waist over the hip measurements of the participants. Then later was divided into three groups based on WHO classification [158]. |
| | Moderate (0.81-0.85) | |
| | High (>0.85) | |
| Reproductive interval index-years | Low (≤12) | Reproductive interval index was the difference between the age at first birth and age at menarche. The index was divided into four groups based on the IQR (Inter Quartile Range) of the reproductive interval index values among the controls only. |
| | Moderate (12-16) | |
| | High (>16) | |
| | No children | |
| Deprivation score | Calculated by UK biobank team | Deprivation score was calculated and was made available by the UK Biobank. The score was based on the prior national census output areas [159]. The score evaluated four aspects: 1) unemployment, 2) houses without an owned car, 3) non-house ownership, 4) overcrowding in one house [160]. |
| Height | Below mean<br>(< 156.10 cm) | Height measures were grouped into three groups based on the mean height values of the control group. |
| | Within mean ± SD (156.10-168.75cm) | |

| Variable | Groups | Coding |
|---|---|---|
| | Above mean<br>(>168.75 cm) | |

### 3.2.3 Data analysis

All analyses were stratified by menopausal status: pre- and post-menopausal (previously described in the methodology section of chapter 4 section 3.1.3). To compute BC incidence within the cohort, the Stata *stptime* command was used to obtain the overall person-time of observation and disease incidence rate. To calculate the time variable as in year for each participant, the endpoint (either the date of cancer diagnosis or the end of the follow-up - January 1st, 2016) was subtracted with the date of study enrolment. Incidence rates were estimated for the whole cohort and pre- and post-menopausal separately. Moreover, population attributable fractions (PAF) were calculated to estimate how much risk could be eliminated by controlling that risk factor. PAF was calculated using the STATA *punaf* command [176] where the fraction was estimated compared to whole cohort and compared to the most significant subgroup associated with the BC.

To assess associations between exposures and BC risk in the cohort, relative risk (RR) and 95% confident intervals (95% C.I.) was computed using a binomial generalised linear regression model. Regression analyses were performed for each independent variable and were adjusted for age, family history of BC in first degree relatives, and deprivation score.

All statistical analysis was performed using Stata MP 14.1 software for Windows [177]. Results with 95% confident intervals not including 1 were considered as being statistically significant.

### 3.2.3.1 Study power

Study power was calculated for prospective cohort study design for pre and post menopause group. Study power was computed using PS (Power and Sample Size Calculation) program, an interactive program for performing power and sample size calculations [178].

### Pre-menopause group

The study comprised 618 BC subjects and approximately 56856 control subjects (ratio of case per control equal to 1:92). The prevalence of exposure among controls was set at 5%, 10%, 15% and 20%. Results of true relative risks in exposed subjects relative to unexposed subjects with study power of 90% and 80% are shown in table 3.3. The Type I error probability associated with this test of the null hypothesis that this relative risk equals 1 is 0.05. A continuity-corrected chi-squared statistic or Fisher's exact test was applied to evaluate this null hypothesis. Results from table 3.3 shows true detectable relative risk with different prevalence of exposure. However, it is noted that even though the study was adequately powered to detect large effects for rarer exposures (<5%), more modest effects (< 30%) may be missed.

Table 3.3: Detectable relative risk for pre-menopause group with difference prevalence of exposure in controls with study power of 90% and 80%

| Study power | Prevalence of exposure in controls | Detectable relative risk |
|---|---|---|
| 90% | 5% | ≤0.480 or ≥1.648 |
| | 10% | ≤0.628 or ≥1.428 |
| | 15% | ≤0.700 or ≥1.332 |
| | 20% | ≤0.745 or ≥1.276 |
| 80% | 5% | ≤0.529 or ≥1.544 |
| | 10% | ≤0.669 or ≥1.363 |
| | 15% | ≤0.735 or ≥1.284 |
| | 20% | ≤0.776 or ≥1.236 |

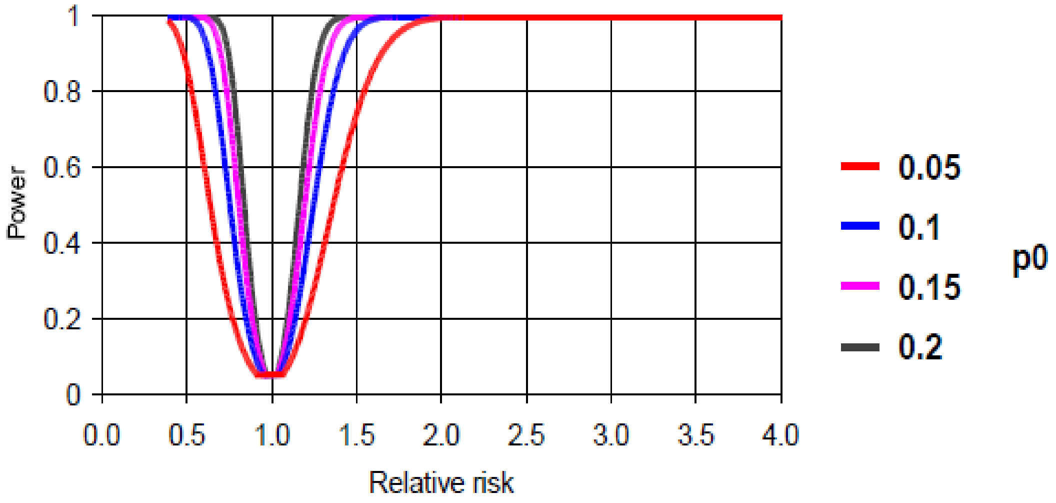


Figure 3.1: Study power for pre-menopause group with different detectable relative risks given probability of exposure among controls 0.05, 0.10,0.15 and 0.20 with a sample size of 618 cases and 51856 controls (α =0.05).

**Post menopause group**

The study was performed with 1757 experimental subjects and approximately 112797 control subjects (ratio of case per control equal to 1:62). The prevalence of exposure among controls was set at 5%,10%, 15% and 20%. Results of true relative risks in exposed subjects relative to unexposed subjects with study power of 90% and 80% are shown in table 3.4. The Type I error probability associated with this test of the null hypothesis, that this relative risk equals 1, is 0.05. A continuity-corrected chi-squared statistic or Fisher's exact test was applied to evaluate this null hypothesis. Results from table 3.4 shows different of true detectable relative risk with different prevalence of exposure. However, it is worth to mention even though the study was adequately powered to detect large effects for rarer exposures (<5%), more modest effects (< 30%) could be missed.

Table 3.4: : Detectable relative risk for post-menopause group with difference prevalence of exposure in controls with study power of 90% and 80%

| **Study power** | **Prevalence of exposure in controls** | **Detectable relative risk** |
|---|---|---|
| 90% | 5% | ≤0.677 or ≥1.367 |
| | 10% | ≤0.773 or ≥1.246 |
| | 15% | ≤0 .818 or ≥1.193 |

| Study power | Prevalence of exposure in controls | Detectable relative risk |
|---|---|---|
| | 20% | ≤0.847 or ≥1.161 |
| 80% | 5% | ≤0.714 or ≥1.312 |
| | 10% | ≤0 .801 or ≥1.210 |
| | 15% | ≤0 841 or ≥1.165 |
| | 20% | ≤0.866 or ≥1.138 |

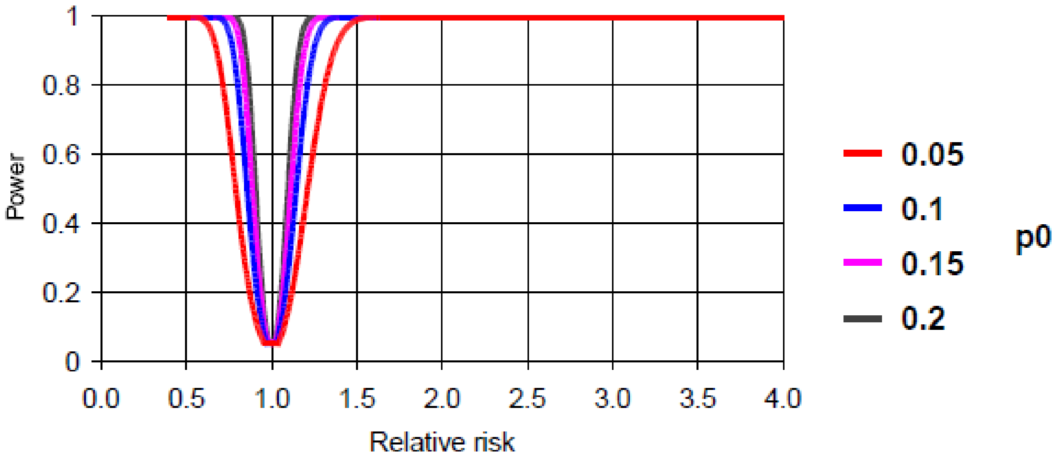


Figure 3.2: Study power for post-menopause group with different detectable relative risks given probability of exposure among controls 0.05, 0.10,0.15 and 0.20 with a sample size of 1757 cases and 112448 controls (α =0.05)

## 3.3 Methodology of chapter 6 (Assessing non-genetic modifiable risk factors with BC risk in genetically predisposed females: third paper methodology)

### 3.3.1 Study population and study design

Data from women within the UK Biobank longitudinal cohort study were used (details was described previously in 3.1.1). The data set for this analysis was last updated on March 31, 2019. The study design was a nested case-control study. BC case and female control identification is described in 3.1.2. Participants were considered as a case if they develop the outcome of interest (breast cancer) after they enrolled in the study. The exposures were defined prior to the disease or outcome development based on the baseline data collection.

### 3.3.2 Defining breast cancer modifiable risk factors

Cancer Research UK [179] has reported risk factors for BC development as being either modifiable or non-modifiable. Based on their published list of factors associated with BC, five modifiable factors were identified: weight, alcohol intake, physical activity, oral contraceptive use, and hormonal replacement therapy (HRT) intake for more than 5 years. A scoring system based on the presence or absence of these 5 factors to derive favourable lifestyle, intermediate lifestyle, and unfavourable lifestyle was developed. This approach was adopted from previous studies on coronary heart disease[180] and dementia [181]. The details of the five factors and score definition is presented in table 3.5. Eligible participants were stratified into three categories: favourable lifestyle (4 healthy factors present), intermediate lifestyle (2 or 3 healthy factors present), and unfavourable lifestyle (1 healthy factor present).

Table 3.5: Criteria for healthy lifestyle classification

| **Healthy lifestyles criteria** | **UK Biobank cohort** | **Codes** |
|---|---|---|
| Healthy weight | Healthy: BMI <25 kg/m$^2$<br>Unhealthy: BMI ≥ 25 kg/m2 | Healthy: 1<br>Unhealthy: 0 |
| Regular physical activity | Healthy: At least ≥ once per week<br>Unhealthy: No physical activity at all | Healthy: 1<br>Unhealthy: 0 |
| No/limited alcohol intake | Healthy: No alcohol intake or used for < three times/week<br>Unhealthy: Used alcohol ≥ three times /week | Healthy: 1<br>Unhealthy: 0 |
| No contraceptive intake | Healthy: No OC use<br>Unhealthy: Used OC | Healthy: 1<br>Unhealthy: 0 |
| No/limited HRT intake | Healthy: No HRT use or used HRT < 5 years<br>Unhealthy: Used HRT for ≥ 5years | Healthy: 1<br>Unhealthy: 0 |
| Classifications | | |
| Favourable lifestyle | Presence of 4-5 healthy lifestyle factors | Sum: At least 4 |
| Intermediate lifestyle | Presence of 2-3 healthy lifestyle factors | Sum: 2 or 3 |
| Unfavourable lifestyle | Presence of only one healthy lifestyle factor or none | Sum: 1 or 0 |

### 3.3.3 Data analysis

Relative risks (RRs) and 95% CIs of the basic risk factors were computed with an adjustment for age and family history using a binomial generalized linear regression model. Cox proportional hazards regression was used to assess the hazard ratios (HRs) of the lifestyles and BC risk. First, HRs were computed for each genetic stratum with the low genetic risk group as a reference group and for each lifestyle (favourable, intermediate, and unfavourable) stratum with the favourable category as a reference group. Second, HRs in each lifestyle stratum were calculated within each genetic risk group. All analyses were adjusted for age and family history. The Cox proportional hazards regression model assumption for each analysis was tested. The test was to assess a non-zero slope in a generalized linear regression of the scaled Schoenfeld residuals on functions of time. A non-zero slope is an indication of a violation of the proportional hazard assumption. All the test results suggested no violation of proportional hazards assumption as shown in table 6.3 in chapter 6. A 2-sided P-value <0.05 was considered significant. The Stata *Ltable* [182] command was used to compute a 10-year cumulative BC incidence for each lifestyle category within each genetic risk stratum. Results presented in graphic bar charts were generated using Microsoft Excel 2016 (Microsoft Corp) [183]. All analyses were performed using Stata/MP software version 14 (StataCorp LLC) [184].

#### 3.3.3.1 Preparation SNPs for the polygenic risk scores (PRS)

The PRS are scores calculated by adding the weighted risk alleles by their effect sizes which are derived from GWAS findings [185]. Our PRS was computed from the published SNPs list of that of the Mavaddat group and their coefficient values [120]. Mavaddat and Colleagues developed this PRS in a dataset comprised 94,075 case subjects and 75,017 control subjects of European ancestry from 69 studies. The best performing PRSs were validated in an independent test set comprising 11,428 case subjects and 18,323 control subjects from 10 prospective studies and 190,040 women from UK Biobank (3,215 incident breast cancers). For the best PRSs (313 SNPs), the odds ratio for overall disease per 1 standard deviation in ten prospective studies was 1.61 (95%CI: 1.57–1.65) with area under receiver-operator curve (AUC) ¼ 0.630 (95%CI: 0.628–0.651). The lifetime risk of overall breast cancer in the top centile of the PRSs was 32.6%. Compared with women in the middle quintile, those in the highest 1% of risk had 4.37- and 2.78-fold risks, and those in the lowest 1% of risk had 0.16- and 0.27-fold risks, of developing ER-positive and ER-negative disease, respectively. This justifies the use of Mavaddat list of SNPs.

To compute PRS, the UK Biobank high-density genome-wide SNP data set was available for 488 377 all participants (males and females). The SNP data were from individuals who were included on the basis of being female (matched genetic and self-reported sex) and their genetic ethnic grouping (white). During the quality control process, individuals with missingness (>2%), outliers for heterozygosity (fraction of non-missing markers) by removing individuals who deviate ±3 SD from the samples' heterozygosity rate mean based on guidance [186], and duplicates, as well as those who were biologically related, were excluded.

The PRS for BC was constructed using the 313 SNPs previously determined to confer BC risk by the hard threshold approach used by Mavaddat *et al* [120]. Of these 313 SNPs, 306 were present in the UK Biobank data set; however, one SNP (rs10764337) was triallelic and therefore excluded. The final number of SNPs used for PRS construction was therefore 305 (details are presented in table S2 in the Supplement of chapter 6). Forty of 305 SNPs had been directly genotyped and successfully passed the marker test applied by UK Biobank. The remaining 265 SNPs were imputed SNPs. The quality of the imputation was estimated using the information scores which is a number between 0 and 1 where 0 indicates complete uncertainty and 1 indicates complete certainty. The lowest information score was 0.86. This value equals the one reported by Mavaddat and colleagues [120] for the UKB and is above the 0.85 threshold for higher-quality variants [187], hence filtering was not performed. Linkage disequilibrium was assessed, and no $r^2$ value between any two SNPs reached 0.9. Plink open source software version 1.90 was used to carry out the quality control processes [188].

### 3.3.3.2 Linkage disequilibrium

The reason for not employing a more stringent $r^2$ threshold was to adhere to the method which had been developed by Mavaddat and colleagues [120] and yielded a reliable and validated PRS score. The authors stated in the paper (page 26): 'SNPs were sorted by p value and filtered on LD, such that uncorrelated SNPs (correlation $r^2 < 0.9$) with lowest p value for association with overall breast cancer in the training set were retained (more rigorous pruning, for example at $r^2 < 0.2$, would have removed from consideration informative SNPs from regions with multiple correlated signals)'.

However, a pruning based on an $r^2$=0.2 threshold using the indep-pairwise function in the Plink software was performed.

### 3.3.3.3 Calculating PRS

Individual participant PRS was calculated in three steps as follows: 1) adding the number of risk alleles of each SNP 2) multiplying it by the previously published effect [122] and 3) summation of each value from second step to derive raw PRS.

The raw PRS was standardised by dividing each raw PRS by the SD of the PRS derived from the control group. No transformation of the PRS data was required since the standardised scores were normally distributed (Figure 2 in the Supplement of chapter 6). A tertile genetic risk classification using standardised PRS values from controls was generated. Each participant was then assigned to a genetic risk group: low (1st tertile up to 33.33%), intermediate (2nd tertile between 33.34% and 66.67%), and high (3rd tertile from 66.68% to 100%).

### 3.3.3.4 Study power

**Study power**

Study power was calculated for the prospective cohort study design for pre and post menopause group**.** Study power was computed using PS (Power and Sample Size

Calculation) program, an interactive program for performing power and sample size calculations [178].

### Post menopause group

The study was conducted with 2728 experimental subjects and approximately 88489 control subjects (ratio of case per control equal to 1:32). The prevalence of exposure among controls was set at 5%, 10%, 15% and 20%. Results of true relative risks in exposed subjects relative to unexposed subjects with study power of 90% and 80% are shown in table 3.6. The Type I error probability associated with this test of the null hypothesis that this relative risk equals 1 is 0.05. A continuity-corrected chi-squared statistic or Fisher's exact test was applied to evaluate this null hypothesis. Results summarised in table 3.6 show different of true detectable relative risk with different prevalence of exposure.

Table 3.6: Detectable relative risk for post-menopause group with difference prevalence of exposure in controls with study power of 90% and 80%

| Study power | Prevalence of exposure in controls | Detectable relative risk |
|---|---|---|
| 90% | 5% | ≥0.736 or ≥1.293 |
| | 10% | ≥0.816 or ≥1.197 |
| | 15% | ≥0 .853 or ≥1.155 |
| | 20% | ≥0.875 or ≥1.129 |
| 80% | 5% | ≥0.768 or ≥1.250 |
| | 10% | ≥0 .839 or ≥1.169 |
| | 15% | ≥0 871 or ≥1.133 |
| | 20% | ≥0 .892 or ≥1.111 |

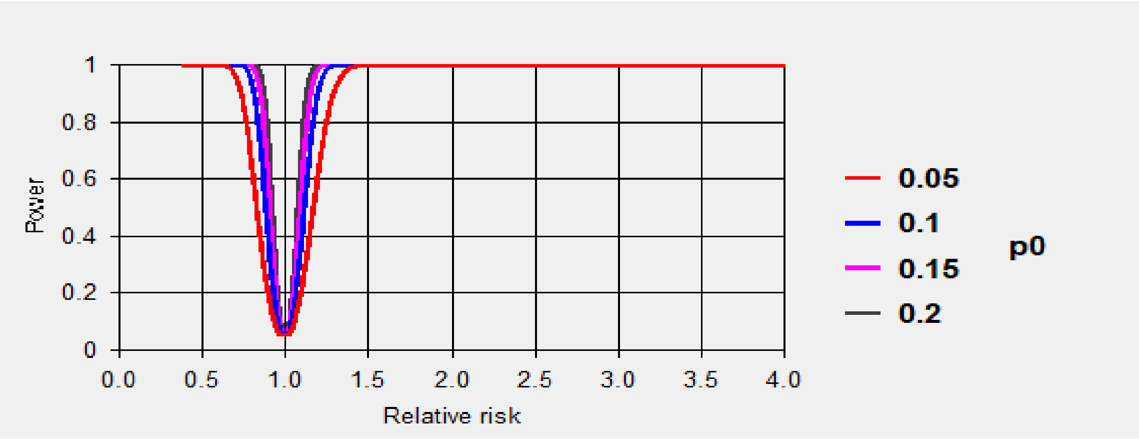


Figure 3.3: Study power for post-menopause group with different detectable relative risks given probability of exposure among controls 0.05, 0.10,0.15 and 0.20 with a sample size of 2728 cases and 88489 controls (α =0.05)

## 3.4 Methodology of chapter 7 (Development and assessment of breast cancer risk prediction models based on the UK Biobank female cohort: fourth paper methodology)

### 3.4.1 Study population

a) Model development-training data

Data from the UK Biobank longitudinal cohort were used to develop these models. BC and control identification was described extensively in previous section of the methodology (section 3.2.1).

b) Model validation- testing data

An independent Canadian Caucasian cohort (The Alberta's Tomorrow Project [ATP]) was used to validate the epidemiological models. The ATP is a longitudinal population study with 55,000 participants (aged between 32 to 71 years). The study started in year 2000 with a planned follow-up period of 50 years. Blood and urine samples have been collected for 30,000 participants and more than 215,000 surveys have been completed by participants. The study main goal is to improve health and to generate risk reduction strategies in the future.

Each participant signed a consent form and the study followed Declaration of Helsinki guidelines and all procedure involving human subjects were ethically approved by the former Alberta Cancer Board’s Research Ethics Committee and the University of Calgary Conjoint Health Research Ethics Board and the Alberta Cancer Research Ethics Committee.

#### 3.4.1.1 Defining cases and controls in the ATP cohort

The outcome of interest (BC) was identified using the Alberta Cancer Registry (ACR) database provided by the ATP team. Three variables were used to identify BC cases (ACR cancer site specific, ACR cancer site aggregate, and ACR ICD_O topography). Subsequently, the incident cases were defined by age at cancer diagnosis greater than enrollment age. All BC incident cases were included in the analysis along with the controls (alive subjects without any history of BC, other cancers, carcinoma in situ or unknown neoplasm). Furthermore, all eligible subjects were categorised according to their menopausal status. Although the ATP study collected data on biological sample, genotype data was not available at the time of the data analysis hence the ATP data was use only to validate epidemiological BC risk prediction model.

### 3.4.2 Data analysis

All the statistical analysis were carried out using Stata MP 14.1 software for Windows [177]. Genetic data were analysed using PLINK 1.9 and PLINK 2.0 [188] to derive PRS (details were described in section 3.3.3.1). The selected risk factors to incorporate into prediction models were derived from assessing the association between all the available risk factors listed in table 3.8 and risk of BC in the development dataset. The Stata codes for defining all used variables for model development are provided in appendix 3. Relative risks (RR) and 95% confident intervals (95% C.I.) were computed using the logistic binomial generalised linear regression model. The RR was calculated to report the ratio of the rate of the event occurring in the exposed group versus the rate of the event occurring in the non-exposed group. RR can be directly determined in a cohort study. When the outcome event is common (incidence of 10% or more), it is often more desirable to estimate an RR. All variables were explored for distribution and missingness. The distribution of continuous variables was explored by plotting histogram. If the variable was dichotomous or category, the chi –square test to explore equally distribution between groups was performed. For missingness, the command “misstable summarize <var>” was used. Number of missingness in each variable is shown in appendix 1. The missing percentage of the variables used in the model is low as shown in appendix 1 with the highest percentage being 2.3% in (menarche age). As a result, CCA (complete case analysis) might be considered to be used as it is the most common approach for large size studies with a small proportion of missing data. It would, however, affect the precision of the estimates (larger standard errors) especially with multivariate data analysis [189]. It was recommended to perform either single or multiple imputation if the missing percentage is more than 10% to avoid biased results [190]. Similarly, no imputation was performed. For variables (1- age the first birth), (2-reproductive index), and (3-number of pregnancy terminations) the missing proportion was high due to the fact that a lot of the female participants did not get pregnant and so they left this option blank as it was not valid in their case. Another reason to avoid imputation of these variables was they were not included in the developing the BC prediction models.

Once candidate variables were identified, I applied the logistic regression model to further develop and test the models. Logistic regression is a widely used statistical model that allows for such multivariate analysis and modeling of a binary dependent variable. The multivariate analysis estimates coefficients (log odds) for each predictor included in the final model and adjusts them with respect to the other predictors in the model. The coefficients quantify the contribution of each predictor to the outcome risk estimation. A further advantage of logistic regression model is that it does not require the errors/residuals to be normally distributed and the variance of the residuals does not need to be constant.

In this analysis, I did not assess any interaction terms due to lack of study power, rather the analysis was stratified by the menopausal status as BC risk factors. Bootstrap regression of 100 simulations and stepwise regression were used to identify significant factors to fit the model with the highest prediction power in each menopausal status. Once the selected predictors were identified, Multi-collinearity testing for all variables in the models was carried out to ensure an absence of closely related between variables. The analysis produced values of the variance inflation factor (VIF), tolerance and $R^2$. VIF quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to assess accurately the contribution of predictors to a model. It has been suggested that values of VIF that exceed 10 are often regarded as indicating multicollinearity. VIF value close to 1 indicates that there is no correlation among the $j^{th}$ predictor and the remaining predictor variables. Another direct measure of multicollinearity is tolerance. Tolerance is defined as the amount of variability of the selected independent variable not explained by the other independent variables. Tolerance value ranges from 0 to 1. Tolerance was computed by subtract 1 with $R^2$. It has been suggested that if the value of tolerance is less than 0.2 or 0.1 and, simultaneously, the value of VIF 10 and above, then the multicollinearity is problematic [191]. In this analysis *collin* command was used to assess the collinearity between included factors. *Collin* command reports both VIF and tolerance estimates. All included risk factors proved no collinearity with VIF less than 10 and tolerance more than 0.1.

Models were then fitted using multivariate logistic regression to assess their performance. Various tests to verify model fit (post estimation) included model specification using *linktest* and assessment of linearity in the logit (assess whether log odds of the outcome is linearly associated with the covariates) using *lowess graph*.

Table 3.7: All the variables assessed to build the breast cancer risk prediction model

| | Variable | Factors | Groups | Coding |
|---|---|---|---|---|
| 1 | Menopausal status | Group stratification | Pre-menopausal | Pre- menopausal: reported as pre-menopausal & no history of hysterectomy or bilateral oophorectomy & their age is ≤55 and menarche age ≥7 years old (to maximise the number of the real pre-menopausal females - so any female reported as pre- and their age > 55 or had menarche age < 7 and did not had hysterectomy nor oophorectomy will be removed – most probably this female is miscategorised). |

|  | Variable | Factors | Groups | Coding |
|---|---|---|---|---|
|  |  |  | Post-menopausal | Post-menopausal: reported as post-menopausal & no history of hysterectomy or bilateral oophorectomy (only natural menopause) & their menopause age is ≥ 40years old (to maximise the number of the real post-menopausal females – so any female reported as post- and their menopause age < 40 and did not had hysterectomy nor oophorectomy will be removed - most probably this female is miscategorised). |
| 2 | Menarche age | Reproductive | (>13) | This variable was divided into two groups based on literature ranges [127]. |
|  |  |  | (≤13)) |  |
| 3 | Age at first birth | Reproductive | (<20) old | This variable was divided into four groups based on literature ranges [192]. |
|  |  |  | (20-24) old |  |
|  |  |  | (25-29) old |  |
|  |  |  | (≥30) old |  |
| 4 | BMI | Anthropometric | Healthy (18.5 – 24.9) | This variable was divided into three groups based on WHO classification [193]. Underweight group were very low in number and would not be enough for the association calculations. |
|  |  |  | Overweight (25-29.9) |  |
|  |  |  | Obese (≥30) |  |
| 5 | Waist to hip ratio (WHR) | Anthropometric | Low (≤0.80) | This variable was calculated by dividing the waist over the hip measurements of the participants. Then later was divided into three groups based on WHO classification [194]. |
|  |  |  | Moderate (0.81-0.85) |  |
|  |  |  | High (>0.85) |  |
| 6 | Reproductive interval index-years | Reproductive | Low (≤12) | Reproductive interval index is the difference between the age at first birth and age at menarche. The index was divided into four groups based on the IQR (InterQuartile Range) of the reproductive interval index values among the controls only. |
|  |  |  | Moderate (12-16) |  |
|  |  |  | High (>16) |  |
|  |  |  | No children |  |
| 7 | Deprivation score | Covariate | Calculated by UK biobank team | The score was calculated for all the participants before participating in UK Biobank. The score was based on the prior national census output areas [195]. The score evaluates four aspects: 1) unemployment, 2) houses without an owned car, 3) non-house ownership, 4) overcrowding in one house [196]. Data were provided by the UK biobank team. |
| 8 | Height | Anthropometric | Below mean (< 156.10 cm) | The height was grouped into three groups based on the mean of the control group. |
|  |  |  | Within mean ± SD (156.10-168.75cm) |  |

|  | Variable | Factors | Groups | Coding |
|---|---|---|---|---|
|  |  |  | Above mean (>168.75 cm) |  |
| 10 | Family history | Covariate | No history | Family history of breast cancer was grouped into three categories as shown. 1) females with no family history of BC, 2) females with either a mother or sister had/has BC, 3) females with both mother and sister had/have BC. |
|  |  |  | Mother or Sibling history of BC |  |
|  |  |  | Mother and Sibling history of BC |  |
| 11 | Moderate physical activity |  | High | This information available from UK biobank team as: In a typical WEEK, on how many days did you do 10 minutes or more of moderate physical activities like carrying light loads, cycling at normal pace?<br>Five and more days: High<br>Four or three: Moderate<br>Two or one: Low<br>Zero: Never |
|  |  |  | Low |  |
|  |  |  | Never |  |
| 12 | HRT users |  | No | This information available from UK biobank team as (Ever used hormone-replacement therapy (HRT): yes/no) |
|  |  |  | Yes |  |
| 13 | OC users |  | No | This information available from UK biobank team as (Ever taken oral contraceptive pill: yes/no) |
|  |  |  | Yes |  |
| 14 | Smoking status |  | Never | This information available from UK biobank team as (smoking status: never, previous, current). |
|  |  |  | Previous |  |
|  |  |  | Current |  |
| 15 | Alcohol drinking status |  | Never | This information available from UK biobank team as (alcohol drinker status: never, previous, current). |
|  |  |  | Previous |  |
|  |  |  | Current |  |
| 16 | Beef intake |  | Within average | This information available from UK biobank team as (How often do you eat beef? The groups were allocated based on the control group consumption |
|  |  |  | Above average |  |
| 17 | Raw vegetables intake |  | Yes | Coded as yes or no |
|  |  |  | No |  |
| 18 | Processed meat intake |  | Within average | This information available from UK biobank team as (How often do you eat processed meats (such as bacon, ham, sausages, meat pies, kebabs, burgers, chicken nuggets) The groups were allocated based on the control group consumption |
|  |  |  | Above average |  |
| 19 | Parity |  | No | This information was extracted from (number of live births provided by UK biobank team). |
|  |  |  | Yes |  |
| 20 |  |  | No |  |

| | Variable | Factors | Groups | Coding |
|---|---|---|---|---|
| | Mammogram history | | Yes | This information available from UK biobank team as (Ever had breast cancer screening / mammogram: yes/no) |
| 21 | Age | | Continuous variable | No further coding was needed |
| | Age at first birth (year) | | Continuous variable | No further coding was needed |
| 22 | Number of Pregnancy termination | | Continuous variable | No further coding was needed |
| 23 | Contraceptive use duration (year) | | Continuous variable | This variable was calculated by subtracting the (Age when last used oral contraceptive pill) from (Age started oral contraceptive pill) variables provided by the UK biobank team. |
| 24 | HRT duration (year) | | Continuous variable | This variable was calculated by subtracting the (Age last used hormone-replacement therapy (HRT)) from (Age started hormone-replacement therapy (HRT)) variables provided by the UK biobank team. |
| 25 | Body mass index | | Continuous variable | No further coding was needed |
| 26 | Number of live births | | Continuous variable | No further coding was needed |
| 27 | Hip circumference | | Continuous variable | No further coding was needed |
| 28 | Waist circumference | | Continuous variable | No further coding was needed |
| 29 | Standing height | | Continuous variable | No further coding was needed |
| 30 | Sitting height | | Continuous variable | No further coding was needed |

#### 3.4.2.1 Study power

Previously, a well-used "rule of thumb" for sample size to ensure at least 10 events per candidate variable (EPV) where "candidate" indicates a predictor in the development data set was considered, before any variable was selected for inclusion into the final model. However, the 10 EPV did not fully account for the magnitude of predictor effects, the overall outcome risk, the distribution of predictors, and the number of events for each category of categorical predictors.

Riley *et al* [197] proposed a method to derive sample size (n) and number of events (E) in the model development data set. The following three criteria should be met: (i) small optimism in predictor effect estimates as defined by a global shrinkage factor of ≥0.9, (ii) small absolute difference of ≤ 0.05 in the model's apparent and adjusted Nagelkerke's R$^2$ (R$^2$ CS_adj), and (iii) precise estimation of the overall risk or rate in the population (or similarly, precise estimation of the model intercept when predictors are mean centred). The values of n and E (and subsequently EPP) that meet all three criteria provide the minimum values required for model development.

To adjust for overfitting during model development (and thereby improve the model's predictive performance), statistical methods for penalisation of predictor effect estimates were used. The concept was based on the global shrinkage factor (a measure of overfitting).

A global shrinkage factor (S) referred to as a uniform shrinkage factor. Consider a logistic regression model has been fitted using standard maximum likelihood estimation (i.e., traditional and un-penalised estimation). Subsequently, S can be estimated (e.g., using bootstrapping or via a closed-form solution) and applied to the estimated predictor effects. A desired shrinkage factor is 0.9.

The calculation steps were carried out to derive sample size that met all three criteria.

**Criterion (i): calculating sample size to ensure a shrinkage factor ≥ 0.9**

The formula below was used to calculate sample size require. Base on the above approach, to develop a new logistic regression model based on up to 20 candidate predictor parameters (p) with an anticipated R$^2$ CS_adj of at least 0.1, then to target an expected shrinkage ($S_{VH}$) of 0.9, a sample size of 1698 is needed.

$$n = \frac{p}{(S_{VH} - 1)\ \ln\left(1 - \frac{R^2_{\mathrm{CS\_adj}}}{S_{VH}}\right)}.$$

$= 20 \div [(0.9 - 1)\ln\left(1 - \frac{0.1}{0.9}\right)]$ =1698

Next, the calculated sample size was translated to the number of events (E) and event per predictor parameter (EPP).

For binary outcomes, $E = n\phi$ and EPP = $n\phi/p$, where $\phi$ is the overall outcome proportion in the target population (i.e., the overall prevalence for models,). 1698 subjects were needed based on an $R^2$CS_adj of 0.1 and expected shrinkage of 0.9, then if the intended setting has $\phi$ of 0.1 (i.e., overall outcome risk is 10%), the required $E$ = 1698 × 0.1 = 169.8. With 20 predictor parameters, the required EPP = (1698 × 0.1) ÷ 20= 8.5

**Criterion (ii): ensuring a small absolute difference in the apparent and adjusted $R^2$ Nagelkerke**

This step was undertaken to ensure a small absolute difference ($\delta$) between the model's apparent and adjusted proportion of variance explained. The formula below was used.

- $R^2_{CSadj}$ value, it can range between 0 and 1, and so a small difference (i.e. ≤ 0.05). In this calculation, $R^2$ was set at 0.01.
- It has been recommended that $\delta$ is ≤ 0.05, such that the optimism is Nagelkerke's percentage of variation explained is ≤ 5%.
- For an outcome proportion of 5%, the max ($R^2_{CS\_app}$) is 0.33.

$$S_{VH} \geq \frac{R^2_{\mathrm{CS}_{\mathrm{adj}}}}{R^2_{\mathrm{CS}_{\mathrm{adj}}} + \delta \max\left(R^2_{\mathrm{CS}_{\mathrm{app}}}\right)}.$$

= 0.1 ÷ [0.1 + (0.05 × 0.33)] = 0.858.

Therefore, $S_{VH}$ must be at least 0.86 to meet criterion (ii) which was lower than the recommended value of at least 0.90 suggesting criteria II was met.

**Criterion (iii): ensure precise estimate of overall risk (model intercept)**

This step was carried out by calculate the margin of error in outcome proportion estimates ($\phi$) for a null model (i.e. no predictors included). For a binary outcome, an approximate

95% confidence interval for the overall outcome proportion is

$$n = \left(\frac{1.96}{\delta}\right)^2 \hat{\phi}(1 - \hat{\phi}) \, .$$

It was recommended that a more stringent margin of error ≤ 0.05 with the outcome proportion of 0.5, therefore the number of required samples was.

$n = (1.96 \div 0.05)^2 \times 0.5(1 - 0.5) = 384.2$

Therefore, 385 participants were adequate to ensure precise estimation of the overall risk in the population of interest.

In summary, the UKBiobank sample size provided a sufficient sample size to build a risk prediction model for up to 20 predictors with a desired shrinkage factor of 0.9.

### 3.4.2.2 Model performance

Model internal validation was performed using cross validation of 10 fold to assess model discrimination and calibration. For each k-fold in the dataset, the model was built on (k – 1) fold and subsequently was tested to check the performance for k$^{th}$ fold. The command for performing the 10-fold cross validation is called "*cvAUROC*" [198]. Model discrimination was estimated using sensitivity and specificity to calculate the AUC. The AUCs were plotted using the receiver operating characteristics curves (ROC). Calibration of the models was estimated by comparing the expected against the observed event. Hosmer-Lemeshow goodness of fit test was used to assess the model calibration.

External validation including discrimination and calibration for epidemiological model was assessed using the testing dataset (the ATP cohort).

### 3.4.2.3 Absolute 5-years risk calculation

The 5-years absolute risk (AR) of breast cancer for pre- and post-menopausal UK females was estimated. The AR calculation was described previously [199]. Below describes steps applied to obtain absolute 5-year risk for each individual.

Step1: The risk was calculated by estimating the risk component of the risk factors. The risk component was derived by multiplying the RR of each risk factor associated with that individual (r= RR1 x RR2 x RR3 x RR4 ….. x RRn; see Table S1).

Step2: Derive values to compute baseline hazard rate. First value is the constant value which was calculated by the following formula:

$$\text{Constant value} = \frac{\text{\% of BC in Pre} - \text{menopausal (age } 0 - 54)}{\text{\% of females in pre} - \text{menopausal status (age } 0 - 54)}$$

(See supplementary materials table S1 and table S2). The numbers to compute constant value were obtained from the Office for National Statistics (ONS) (https://www.ons.gov.uk/).

Second value is BC age-specific rates. The rates were estimated from ONS census for every 5-year age until age 54. The BC age-specific rate was calculated by the following formula:

$$\text{BC age} - \text{specific rate} = \frac{\text{number of BC cases at specific age}}{\text{Total female population at specific age}}$$

See appendix table S1 for example.

Third value is the attributable risk (AR). AR was calculated using the formula:

$$\text{AR} = \frac{\text{Incidence among exposed} \ - \ \text{Incidence among unexposed}}{\text{Incidence among exposed}}$$

The fourth value is BC age specific mortality rate. The BC mortality rate from all causes of death except BC death at specific age was estimated by subtracting the female death from BC from female death from all causes.

Step3: The baseline hazard rate was obtained using values obtained from step2 with the following formula:

$$\text{BC incidence base hazard rate} = \left[\frac{\text{constant for pre} - \text{menopausal } \times \text{ BC rate by age}}{100000}\right] \times [1 - \text{AR}]$$

Step4: The 5-year absolute risk for BC was calculated as percentage using values derived from step1-4 =

$$5 - \text{year absolute risk} = \left[\frac{\text{RR of all risk factors } \times \text{ BC baseline hazard rate}}{(\text{RR of all risk factors } \times \text{ BC baseline hazard rate }) + \text{ BC mortality rate}}\right]$$

$$\times [1 - \ \exp[-(\text{RR of all risk factors } \times \text{ BC baseline hazard rate } + \text{ BC mortality rate}]$$

This 5-year absolute risk can be compared with BC age specific rate obtained from ONS (supplementary materials, table S1 chapter 7).

### 3.4.3 Journal linkage

The aim of this thesis is to construct a BC risk prediction model designed for the UK females specifically. In order to achieve this aim I divided the work into four steps. Starting with reviewing the existing epidemiological risk prediction models and risk factors incorporated into the models. The product of this review was formatted as published article number 1 (chapter 4) and I used the reviewed risk factor list as a knowledge background for the model development. The reviewed risk factors were grouped as reproductive, anthropometric, dietary and lifestyle factors. Anthropometric and reproductive risk factors were tested using data from the UKBiobank cohort.  Any significant risk factors were selected for model development work. The results of anthropometric and reproductive risk factors were published as an article number 2 (chapter 5). However, other risk factors (dietary and lifestyle) were also analysed as explained in chapter 7. At the beginning of year 2019, the UKBiobank team released the genetic data and I decided to use it in the model development to improve the utility and discrimination power. The decision was taken to incorporate the polygenic risk scores of 305 preselected single-nucleotide variations of BC into the prediction model. The availability of genetic data provided an opportunity to further explore the association of non-genetic factors of BC in genetically predisposed groups (based on their PRS). Results of these analyses were published as publication no 3 (chapter 6). The findings of chapter 4-6 were applied to the development of BC risk prediction as described in chapter 7.

# Chapter 4 : Review of non-clinical risk models to aid prevention of breast cancer

Publication number 1

## Review of non-clinical risk models to aid prevention of breast cancer

This chapter is presented as a journal article:

**Al-ajmi, K**., et al., Review of non-clinical risk models to aid prevention of breast cancer. Cancer Causes & Control, 2018. **29**(10): p. 967-986.

**Review of non-clinical risk models to aid prevention of breast cancer**

Kawthar Alajmi, Artitaya Lophatananon, Martin Yuille, William Ollier, Kenneth R Muir
Centre for Epidemiology, Division of Population Health, Health Services Research and Primary Care, Faculty

Key words: Assessment risk tool, calibration, discrimination, risk factors, risk prediction, concordance, and E/O statistics.

## 4.1 Abstract

A disease risk model is a statistical method which assesses the probability that an individual will develop one or more diseases within a stated period of time. Such models take into account the presence or absence of specific epidemiological risk factors associated with the disease and thereby potentially identifies individuals at higher risk. Such models are currently used clinically to identify people at higher risk, including identifying women who are at increased risk of developing breast cancer. Many genetic and non-genetic breast cancer risk models have been developed previously. Existed non-genetic/non-clinical models for breast cancer that incorporate modifiable risk factors were evaluated. This review focuses on risk models that can be used by women themselves in the community in the absence of clinical risk factors characterization. The inclusion of modifiable factors in these models means that they can be used to improve primary prevention and health education pertinent for breast cancer.

Literature searches were conducted using PubMed, ScienceDirect and the Cochrane Database of Systematic Reviews. Fourteen studies were eligible for review with sample sizes ranging from 654 to 248,407 participants. All models reviewed had acceptable calibration measures, with expected/observed (E/O) ratios ranging from 0.79 to 1.17. However, discrimination measures were variable across studies with concordance statistics (C-statistics) ranging from 0.56 to 0.89. It was concluded that breast cancer risk models that include modifiable risk factors have been well calibrated but have less ability to discriminate. The latter may be a consequence of the omission of some significant risk factors in the

models or from applying models to studies with limited sample sizes. More importantly, external validation is missing for most of the models. Generalization across models is also problematic as some variables may not be considered applicable to some populations and each model performance is conditioned by particular population characteristics. In conclusion it is clear that there is still a need to develop a more reliable model for estimating breast cancer risk which has e good calibration, ability to accurately discriminate high risk and with better generalisability across populations.

## 4.2 Introduction

Breast cancer is the most common cancer among females in high, middle and low-income countries and it accounts for 23% of all new female cancers globally [13, 200]. While there has been a significant reduction in mortality, incidence rates have continued to rise [201]. Breast cancer incidence rates are high in North America, Australia, New Zealand, and Western and Northern Europe. It has intermediate levels of incidence in South America, Northern Africa, and the Caribbean but is  lower in Asia and sub-Saharan Africa [200]

Early detection of breast cancer improves prognosis and increases survival. Mammographic imaging is the best method available for early detection [202] contributing substantially in reducing the deaths caused by breast cancer  [203]. Unfortunately, mammography mass screening still leads to some levels of over-diagnosis and over-treatment [204]. As yet routine mammography screening is not readily available globally, particularly in some developing countries [205, 206]. This is supported by the observations that for every million adult women there are only four mammogram screening machines in Sudan has four mammogram machines, whereas Mexico has 37 and Canada has 72 [207]. Under these circumstances, it is clearly more appropriate to prioritise access to mammographic screening or other targeted interventions (such as tamoxifen chemoprevention) for higher-risk individuals who could be identified using a sensitive and specific risk prediction model [24]. Such risk prediction models are individualized statistical methods to estimate the probability of developing certain medical diseases. This is based on specific risk factors in currently healthy individuals within a defined period of time [27].  Such  prediction models have a number of potential uses such as:  planning intervention trials; designing population prevention policies; improving clinical decision-making; assisting in creating benefit/risk indices; and estimating the burden cost of disease in population [24].

A general case can also be made for using risk models for certain diseases. For example, their use can allow the application of risk-reducing interventions that may actually prevent the disease in question. If their application can be based on use of existing health records this will avoid increasing levels of anxiety in at least low to moderate risk individuals. The National Cancer Institute of the USA (NCI) has confirmed that the application of "risk prediction" approaches has an extraordinary chance of enhancing " The Nation's Investment in Cancer Research" [208]. This provides an explanation for the rapid increase in the number of models now being reported in the literature [209, 210]. It is clear that not all developed models are valid or can be widely used across populations. The minimum performance measures required for a useful and robust risk prediction model in clinical decision making are discrimination and calibration. [211].

It was recognised that risk models are increasingly now being used as part of a "triage" assessment for mammography and/or for receipt of other more personalized medical care. There is a growing interest in applying risk prediction models as educational tools.

The models developed can differ significantly with regard to; the specific risk factors that are included; the statistical methodology used to estimate, validate and calibrate risk; in the study design used; and in the populations investigated to assess the models. These differences make it essential that any assessment of model usefulness takes into account both their internal and external validity. Here, the focus was on the reliability, discriminatory accuracy and generalizability of breast cancer risk models that exclude clinical (any variable which needs physician input e.g. presence of atypical hyperplasia) and any genetic risk factors. Accurate assessment of risk using easily acquired data is essential as a first stage of tackling the rising burden of breast disease globally. Well validated models with high predictive power are preferable although this is not the case for all models. The usability of any model is dependent on the purpose the model will be used for and its' target populations [212]. Furthermore, it has been suggested that adapting existing predictive models to the local circumstances of a new population rather than developing a new model for each time is a better

This review focuses on breast cancer risk predicting models that incorporated modifiable risk factors and/or factors that can be self-reported. Such models could be applied as an educational tool and potentially used to advise at risk individuals on appropriate behavioral changes.

## 4.3 Methods

### 4.3.1 Databases

The following databases were searched for all related publications (up to July 2016): PubMed (https://www.ncbi.nlm.nih.gov/pubmed/); ScienceDirect (http://www.sciencedirect.com/); the Cochrane Database of Systematic Reviews (CDSR) (http://www.cochranelibrary.com/). Terms used for the search were “assessment tool, assessment model, risk prediction model, predictive model, prediction score, risk index, breast cancer, breast neoplasm, breast index, Harvard model, Rosner and Colditz model, and Gail model”. Risk models were retrieved based on any study design, study population or types of risk factors.

A Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach was applied for selecting reviewed articles [170, 171]. A total of 61 genetic and non-genetic breast cancer risk models were identified and then filtered to include only risk models with non-clinical factors (Figure 4.1). These models contain variables which are considered to be modifiable and/or self-reported by the respondents. For this review, 14 studies were eventually considered to be eligible. No literature reviews were found on breast cancer risk models solely focusing on epidemiological risk factors although all the selected reviews summarised generic composite risk models. The literature search was extended to include publications relating to systematic reviews and meta-analyses; this did not reveal any appropriate publications.

### 4.3.2 Confidence in risk factors

Details relating to the degree of confidence in variables used as risk factors in the risk models were taken from the Harvard report [42]. This categorisation is not an indication of causal relationship rather an estimation for the association magnitude and direction between the exposure and the interested outcome. The degree of confidence was categorised as either:

- Significant (an established association between outcome and exposure where chance, bias [systematic error], confounders [misrepresentation of an association by unmeasured factor/s] are eliminated with significant confidence)
- Probable (an association exists between the outcome and the exposure where chance, bias, confounders cannot be eliminated with sufficient confidence – inconsistent results found with different studies)
- Possible (inconclusive or insufficient evidence of an association between the outcome and the exposure)

## 4.4 Results

### 4.4.1 Potential risk factors included in breast cancer non-clinical models

The variables used in the fourteen models reviewed and the degree of confidence (significant, probable, or possible) in those variables as breast cancer risk factors for breast cancer are summarised in table 4.1.

Age, age at first birth, age at menarche, family history of breast cancer, self-reported history of biopsies were the most common variables used amongst the fourteen models selected. These variables are considered as significant risk factors for developing breast cancer [42]. Other additional variables were observed in less models. These included ethnicity (Jewish - significant), significant hormonal replacement therapy, diet (some probable and others possible), physical activity (possible), height (significant), weight (probable- for pre-menopausal women and significant for post-menopausal women). Among pre-menopausal females, weight is considered to be a protective factor [75]. In contrast amongst post-menopausal women weight is considered to be a risk factor [213-215] as is parity, oral contraceptive pill use ( significant), pregnancy history, timing and type of menopause ( significant), menstrual regularity (possible), menstrual duration and gestation period (probable), smoking (possible), mammogram screening (probable) and age of onset of breast cancer in a relative ( significant).

The largest number of significant factors included in a model (n=10 variables) was seen in the study reported by Colditz and Rosner [25]. This was followed by studies by reported by Park [216], Novotny [217], and Rosner [218]. The number of the significant, probable, and possible variables in the models were evaluated to compare their performance based on the type and number of the variable included.

Figure 4.1: Identification of eligible risk models using PRISMA flowchart

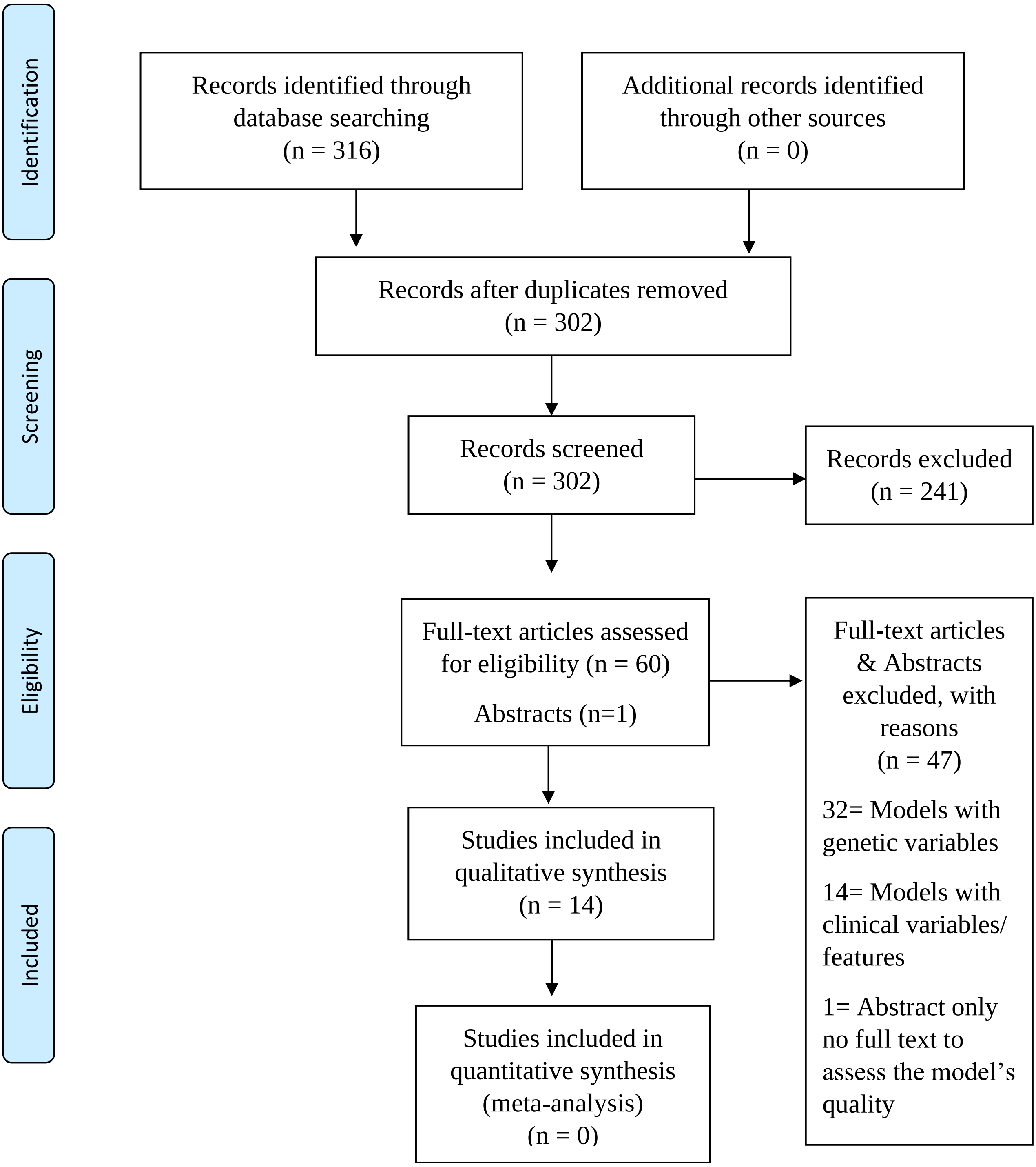

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 316)
Additional records identified through other sources (n = 0)
Records after duplicates removed (n = 302)
Records screened (n = 302)
Records excluded (n = 241)
Full-text articles assessed for eligibility (n = 60)
Abstracts (n=1)
Full-text articles & Abstracts excluded, with reasons (n = 47)
32= Models with genetic variables
14= Models with clinical variables/ features
1= Abstract only no full text to assess the model's quality
Studies included in qualitative synthesis (n = 14)
Studies included in quantitative synthesis (meta-analysis) (n = 0)

Table 4.1: Breast cancer risk factors included in the 14 models

| **Name of model** | Gail 1989[219] | Rosner 1994[220] | Rosner 1996[218] | Colditz 2000[95] | Ueda 2003[221] | Boyle 2004[222] | Lee 2004[223] | Novotny 2006[217] | Gail 2007[224] | Matsuno 2011[225] | Banegas 2012[226] | Pfeiffer (2013) [227] | Park 2013[216] | Lee 2015[228] | Effect | Level of evidence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Basic characteristics** | | | | | | | | | | | | | | | | |
| **Age** | Yes | Yes | Yes | Yes | | Yes | Yes | Yes | Yes | | Yes | | | Yes | Increased risk | Significant |
| **Ethnicity** | | | | | | | | | | Yes | | | | | Jewish increased risk | Significant |
| **Height** | | | | Yes | | | | | | | | | | | Increased risk | Significant |
| **Weight** | | | | Yes | | | | | | | | | | | Increased risk in postmenopausal | Probable |
| **BMI** | | | | Yes | Yes | Yes | | | | | | Yes | Yes | Yes | | Probable |
| **Alcohol intake** | | | | Yes | | Yes | Yes | | | | | Yes | Yes | | Increased risk | Probable |
| **Smoking** | | | | | | | Yes | | | | | | Yes | | Increased risk | Possible |
| **Physical activity** | | | | | | Yes | | | | | | | Yes | Yes | Decreased risk | Possible |
| **Diet** | | | | | | Yes | | | | | | | | | Decreased risk | Probable |
| **Hormonal and reproductive factors** | | | | | | | | | | | | | | | | |
| **Age at menarche** | Yes | Yes | Yes | Yes | Yes | Yes | | Yes | Yes | Yes | Yes | | Yes | Yes | Increased risk | Significant |

| **Name of model** | Gail 1989[219] | Rosner 1994[220] | Rosner 1996[218] | Colditz 2000[95] | Ueda 2003[221] | Boyle 2004[222] | Lee 2004[223] | Novotny 2006[217] | Gail 2007[224] | Matsuno 2011[225] | Banegas 2012[226] | Pfeiffer (2013) [227] | Park 2013[216] | Lee 2015[228] | Effect | Level of evidence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Age at first live birth** | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Increases risk | Significant |
| **Age at subsequent birth** |  | Yes | Yes |  |  |  |  |  |  |  |  |  |  |  | Increases risk | Significant |
| **Age at menopause** |  | Yes | Yes | Yes |  |  |  |  |  |  |  | Yes | Yes | Yes | Increased risk | Significant |
| **Hormone replacement therapy use** |  |  |  | Yes |  | Yes |  |  |  |  |  | Yes | Yes |  | Increases risk | Significant |
| **Oral contraceptive use** |  |  |  | Yes |  |  |  | Yes |  |  |  |  | Yes |  | Increases risk | Significant |
| **Breast feeding** |  |  |  |  |  |  | Yes |  |  |  |  |  | Yes |  | Decreases risk | Probable |
| **Pregnancy** |  |  |  |  |  |  |  |  |  |  |  |  | Yes |  | Decreases risk | Possible |
| **Parity** |  |  |  | Yes |  |  |  |  |  |  |  | Yes |  |  | Decreases risk | Significant |
| **Children number** |  |  |  | Yes |  |  |  |  |  |  |  |  |  | Yes | Decreases risk | Possible |

| **Name of model** | Gail 1989[219] | Rosner 1994[220] | Rosner 1996[218] | Colditz 2000[95] | Ueda 2003[221] | Boyle 2004[222] | Lee 2004[223] | Novotny 2006[217] | Gail 2007[224] | Matsuno 2011[225] | Banegas 2012[226] | Pfeiffer (2013) [227] | Park 2013[216] | Lee 2015[228] | Effect | Level of evidence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Menopause type** | | | | Yes | | | | | | | | | | | Surgical menopause reduces risk | Possible |
| **Menstrual regularity** | | | | | | | Yes | | | | | | | | Menstrual regularity and duration – inconsistent results | Possible |
| **Menstrual duration** | | | | | | | Yes | | | | | | | Yes | | Possible |
| **Menopausal status** | | | | | | | | | | | | Yes | | Yes | Post-menopause increases risk | Possible |
| **Gestation period** | | | | | | | | | | | | | | Yes | Increases risk | Possible |
| **Family history of breast and/or ovarian cancer or diseases** | | | | | | | | | | | | | | | | |
| **Family history of breast cancer** | Yes | | | Yes | Yes | Yes | Yes | Yes | | Yes | | Yes | Yes | Yes | Increases risk | Significant |
| **First-degree relatives with breast cancer** | Yes | | | Yes | | | | | Yes | | Yes | | | | Increases risk | Significant |
| **Age of onset of breast** | | | | | | Yes | | | | | | | | | Increases risk | Probable |

| **Name of model** | Gail 1989[219] | Rosner 1994[220] | Rosner 1996[218] | Colditz 2000[95] | Ueda 2003[221] | Boyle 2004[222] | Lee 2004[223] | Novotny 2006[217] | Gail 2007[224] | Matsuno 2011[225] | Banegas 2012[226] | Pfeiffer (2013) [227] | Park 2013[216] | Lee 2015[228] | Effect | Level of evidence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **cancer in a relative** | | | | | | | | | | | | | | | | |
| **Benign breast disease** | | | | Yes | | | | Yes | | | | Yes | | | Increases risk | Probable |
| **History of breast biopsies** | Yes | | Yes | | | | | Yes | Yes | Yes | Yes | | Yes | | Increases risk | Significant |
| **Mammogram** | | | | | | | | | | | | | | Yes | Increases risk | Probable |
| **Summary of risk factors included in each model** | | | | | | | | | | | | | | | | |
| **Significant factors** | 5 | 5 | 6 | 10 | 3 | 5 | 3 | 6 | 3 | 5 | 5 | 5 | 7 | 5 | Max of 10 and min of 3 factors | |
| **Probable factors** | 0 | 0 | 0 | 4 | 1 | 3 | 2 | 1 | 0 | 1 | 0 | 3 | 3 | 2 | Max of 4 and min of 0 factors | |
| **Possible factors** | 0 | 0 | 0 | 2 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 1 | 3 | 5 | Max of 5 and min of 0 factors | |
| **Total factors** | 5 | 5 | 6 | 16 | 4 | 8 | 8 | 7 | 3 | 5 | 5 | 9 | 13 | 12 | Max of 16 and min of 3 factors | |

### 4.4.2 Evaluation measures of the risk models

The most important measures used to assess the performance of the models were considered to be:

- Calibration (reliability): The E/O statistic measures the calibration performance of the predictive model. Calibration involves comparing the expected versus observed numbers of the event using goodness-of-fit or chi square statistics. A well-calibrated model will have a number close to 1 indicating little difference between the E and O events. If the E/O statistic is below 1.0 then the event incidence is underestimated, while if the E/O ratio is above 1.0 then incidence is overestimated [29, 30].

- Discrimination (precision): The C statistic (Concordance statistic) measures the discrimination performance of the predictive model and corresponds to the area under a receiver operating characteristic curve. This statistic measures how efficiently the model is able to discriminate affected individuals from un-affected individuals. A C-statistic of 0.5 indicates no discrimination between individuals who go on to develop the condition and those who do not. In contrast a C-statistic of 1 implies perfect discrimination [26, 229]. Good discrimination is important for screening individuals and for effective clinical decision making [24].

- Accuracy: is tested by measuring of ‘sensitivity’, ‘specificity’, ‘positive predictive value’ (PPV) and ‘negative predictive value’ (NPV). All of these terms are defined in table 4.2. These measures indicate how well the model is able to categorise specific individuals into their real group (i.e. 100% certain to be affected or unaffected). Accuracy is equally important for both individual categorisation and for clinical decision making. Nevertheless, even with good specificity or sensitivity, low positive predictive values may be found in rare diseases [24] as the predictive values also depend on disease prevalence. With high prevalence, PPV will increase while NPV will decrease [230].

- Utility: this evaluates the ease with which the target groups (public, clinicians, patients, policy makers) can submit the data required by the model. Utility evaluation assesses lay understanding of risk, risk perception, results interpretation, level of satisfaction and worry [231]. This evaluation usually uses surveys or interviews [30].

Table 4.2: Formulas used to calculate the accuracy of the model

| Term | Definition | Equation |
|---|---|---|
| Sensitivity | Probability of a test will indicate 'positive' among those with the disease | (TP) / (TP+FN) |
| Specificity | Probability of a test will indicate 'negative' among those without the disease | (TN) / (TN+FP) |
| Positive predictive value | Probability of a patient having disease when test is positive | (TP) / (TP + FP) |
| Negative predictive value | Probability of a patient not having disease when test is negative | (TN) / (FN + TN) |

TP: true positive; TN: true negative; FP: false positive; FN: false negative •

Calibration and discrimination were the most common measures used to assess the breast cancer risk models under review. Internal calibration was performed in just three of the 14 models with values ranging from 0.92 to 1.08. These calibration values represented a good estimate of the affected cases using these models. For external calibration, six of the 14 models used an independent cohort. Rosner [218] and Pfeiffer [227] reported the highest with E/O values of 1.00 and followed by Colditz [42] with an E/O of 1.01.

The C-Statistic values measuring internal discrimination ranged across studies from 0.61 to 0.65. The Park [216] model achieved the best outcome (C-Statistic = 0.64). Additionally, Park [216] showed the highest value with a C-Statistic of 0.89 when applied to subjects recruited from the NCC (National Cancer Centre) screening program. The lowest C-Statistic (0.56) was observed in the Gail model) [224]. Overall, this demonstrates that the models have better calibration than discrimination. Accuracy was only evaluated in the Lee model

[228]. Sensitivity, specificity, and overall accuracy were calculated. The values indicate low accuracy with values ranging from 0.55 to 0.66.

In qualitative research relating to the impact and utility [232] of the Harvard Cancer Risk Index (HCRI) [42], nine focus groups (six female, three male) showed good overall satisfaction with HCRI. Participants appreciated both the detailed explanation and the updated inclusion of risk factors. On the other hand, some participants criticized the absence of what they considered to be important factors (e.g. environmental factors and poverty). Some participants believed that some of the factors on which subjects had been assessed might cause anxiety. It is also noted, however, that the case has been made that such anxiety provides motivation for action to mitigate risk [233].

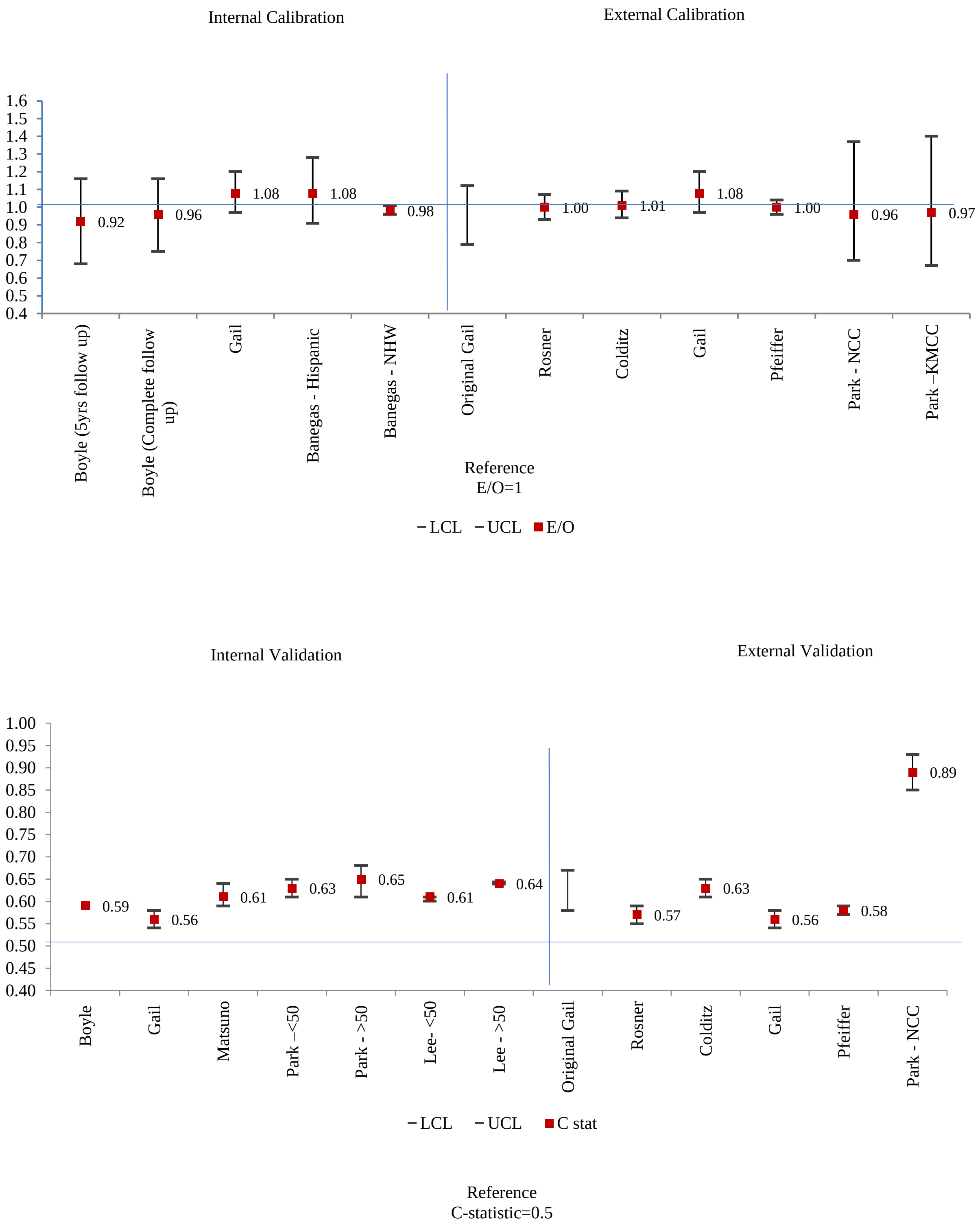
Table 4.3: Summary of the evaluation measures of the risk models

| Model | Calibration | | | Discrimination | | | Accuracy | Utility |
|---|---|---|---|---|---|---|---|---|
| | Derived model | Internal | External | Derived model | Internal | External | Sensitivity, Specificity, PPV, NPV | |
| **Gail 1989** | | | 0.79-1.12 | | | 0.58-0.67 | | |
| **Rosner 1994** | – | – | – | – | – | | – | – |
| **Rosner 1996** | | – | 1.00(0.93–1.07)1 | – | | 0.57 (0.55–0.59)1 | | |
| **Colditz 2000** | – | – | 1.01 (0.94–1.09)1 | – | | 0.64 (0.62–0.66)1 | – | Good[2] |
| **Ueda 2003** | – | – | – | – | – | | – | – |
| **Boyle 2004*** | | a)0.96 (0.75-1.16) cohort1<br>b)0.92(0.68-1.16) cohort2 | – | 0.59 | | | – | – |
| **Lee 2004** | – | – | – | – | – | | – | – |
| **Novotny 2006** | – | – | – | – | – | | – | – |
| **Gail 2007** | – | 1.08 (0.97–1.20) | 0.93 (0.97–1.20)3 | – | | 0.56 (0.54–0.58)3 | – | – |
| **Matsuno 2011** | 1.17 (0.99-1.38) | | | | 0.614 (0.59- 0.64) | | – | – |
| **Banegas 2012**** | – | a)1.08 (0.91–1.28); Hispanic<br>b)0.98 (0.96–1.01); NHW | – | – | – | – | – | – |
| **Pfeiffer 2013** | | | 1.00 (0.96-1.04) | | | 0.58 (0.57-0.59) | | |

| Model | Calibration | | | Discrimination | | | Accuracy | Utility |
|---|---|---|---|---|---|---|---|---|
| | Derived model | Internal | External | Derived model | Internal | External | Sensitivity, Specificity, PPV, NPV | |
| **Park 2013*** ** | – | – | a)0.97(0.67-1.40); KMCC<br>b)0.96 (0.70–1.37); NCC | – | a) 0.63 (0.61–0.65) <50 years (KMCC)<br>b) 0.65 (, 0.61–0.68) ≥50 years (KMCC) | a) 0.61(0.49-0.72); KMCC<br>b)0.89(0.85-0.93); NCC | – | – |
| **Lee 2015** | | | | | Overall: 0.62<br>(0.620-0.623)<br>Under 50: 0.61<br>(0.60-0.61)<br>Above 50: 0.64<br>(0.63-0.64) | | a) Sensitivity:<br>Overall: 0.55 (0.54-0.56)<br><50: 0.61 (0.60-0.62)<br>>50:0.59 (0.59-0.60)<br>b) Specificity:<br>Overall: 0.66 (0.65-0.67)<br>> 50: 0.58 (0.57-0.59)<br>< 50:0.64 (0.63-0.65)<br>c)Accuracy<br>Overall: 0.60 (0.60-0.61)<br>>50:0.59 (0.59-0.60)<br><50:0.61 (0.61-0.62) | - |

Table 3 legend: *Boyle 2004 used two cohorts for calibration (1-Cohort with complete follow up and 2- cohort with 5 years of follow up at most); **Banegas 2012 used two cohorts for calibration (1- Hispanic and 2-Non-Hispanic White (NHW)); ***Park 2013 used two cohorts for calibration and discrimination, using two Korean cohorts: 1- the Korean Multicentre Cancer Cohort (KMCC) and 2-National Cancer Centre (NCC) cohort); [1] [234]; [2] [235]; [3] [27].

Figure 4.2: Calibration[i] and discrimination performances of the 14 breast cancer risk models



### 4.4.3 Overview of current models

All the models described (except for Lee *et al* 2004) [223] are extended versions of either the Gail model or the Rosner and Colditz model. The Gail model developed in 1989 [219] was the first risk model for breast cancer and included the following variables: age, menarche age, age at first birth, breast cancer history in first-degree relatives, history of breast biopsies and history of atypical hyperplasia. The range of calibration of the Gail modified models was E/O= (0.93-1.17) and the discrimination range was C-Statistics= (0.56-0.65). This indicates that these models are well calibrated, although discrimination could be improved.

Ueda *et al*, (2003) [221] modified the Gail model by including age at menarche, age at first delivery, family history of breast cancer and BMI in post-menopausal women, as risk factors in his model for Japanese women. However, as with the original Gail model, no validation was performed. In the Boyle model [222], more factors were included such as alcohol intake, onset age of diagnosis in relatives, one of two diet scores and BMI and HRT. This results in calibration with E/O close to unity and less acceptable discrimination of C-stat= 0.59. The Novotny model [217] added the number of previous breast biopsies performed on a woman and her history of benign breast disease. However, no validation assessment was performed for this model. Newer models [224, 226, 236] included the number of benign biopsies. This resulted in acceptable calibration but less acceptable discrimination (Gail 2007: E/O=0.93; C-stat= 0.56; Matsuno: E/O=1.17, C-statistic= 0.614; and Banegas E/O= 1.08). Park et al (2013) [216] included menopausal status, , number of pregnancies, duration of breastfeeding, oral contraceptive usage, exercise, smoking, drinking, and number of breast examinations as risk factors. This model has an E/O= 0.965; C-stat = 0.64. However, the C-statistic reported from the external validation cohort was high compared to the original C-statistic. They reported a C-statistic of 0.89 using the NCC cohort. This discrepancy was claimed to be caused by the population characteristics (participants were 30 years and above, recruited from cancer screening program, from a teaching hospital in an urban area) by the Korean group [216]. In the same year Pfeiffer et al (2013) developed a model where parity was considered as a factor and had E/O of 1.00 and a C-statistic of 0.58. The later Gail model published in 2007 used logistic regression to derive relative risks. These estimates are then combined with attributable risks and cancer registry incidence data to obtain estimates of the baseline hazards [224].

The Rosner and Colditz model of 1994 [220] was based on a cohort study of more than 91,000 women. The model used Poisson regression (rather than logistic regression as in the Gail model). The variables were: age, age at all births, menopause age, and menarche age. This model was not validated. A new version in 1996 [218] included one modification (current age was excluded) and gave an E/O=1.00 and a C-statistic= 0.57. In 2000, Colditz et al [42] modified the model with risk factors for: benign breast disease, use of postmenopausal hormones, type of menopause, weight, height and alcohol intake. This model gave an E/O= 1.01; C-statistic = 0.64.

Lee *et al* 2004 [223] used two control groups: a "hospitalised" group and a nurses and teachers group. The risk factors in the hospitalized controls were: family history, menstrual regularity, total menstrual duration, age at first full-term pregnancy, and duration of breastfeeding. The risk factors in the nurses/teachers control group were: age, menstrual regularity, alcohol drinking status, and smoking status. This model was not based on Gail or Rosner and Colditz. Hosmer-Lemeshow goodness of fit was used to assess model fit which had a P-value=0.301 in (hospital controls) and P-value=0.871 in (nurse/teacher controls). No calibration or discrimination measures were reported.

Lee [228] used three evaluation techniques to assess the discrimination and the accuracy of their model: support vector machine, artificial neural network, and Bayesian network. Of the three, support vector machine showed the best values among the Korean cohort. However, accuracy and discrimination were less acceptable in this model.

In summary, calibration performance is similar between models (Modified Gail and modified Rosner, Colditz), yet modified Gail models showed better discrimination performance with the C-statistic of the Park model being 0.89.

Table 4.4: Characteristic summary of the reviewed breast cancer risk models

| Author/ Model | Study design | Participants | Ethnicity | Outcome | Statistical method | Effect estimates | Sample size | Risk factors considered in the models | Age target | Stratification |
|---|---|---|---|---|---|---|---|---|---|---|
| Gail 1989 | Case-control | White American females from the Breast Cancer Detection Demonstration Project (BCDDP). | American - Caucasian | Invasive breast cancer + in situ carcinoma | unconditional logistic regression | Relative risk | 2,852 cases<br>3,146 controls | Age at menarche, age at first live birth, number of previous biopsies, and number of first-degree relatives with breast cancer. | Any age | None |
| Rosner 1994 | Cohort | Registered nurses | American - Caucasian | Invasive breast cancer | Poisson regression | Cumulative incidence | 2,341 cases, 91,523 controls | Age, age at all births, menopause age, menarche age | 30-55 years | Number of births |
| Rosner 1996 | Cohort | Registered nurses | American - Caucasian | Invasive breast cancer | Poisson regression | Relative risk | 2,249 cases, 89,132 controls | Menarche age, first live birth age, subsequent births age, menopause age | Any age | None |
| Colditz 2000 | Cohort | General women | American - Caucasian | Invasive breast cancer | Poisson regression | Cumulative incidence | 1,761cases,<br>56,759 controls | Benign breast disease, use of HRT, weight, height, menopausal type, and alcohol intake | Women aged 30–55 years | None |
| Ueda 2003 | Case- control | General women | Japanese - Asian | Invasive breast cancer | Conditional logistic regression | Relative risk | 376 cases<br>430 controls | Menarche, first birth age, family history, and BMI in post-menopausal women. | Any age | Menopausal status |

| **Author/ Model** | **Study design** | **Participants** | **Ethnicity** | **Outcome** | **Statistical method** | **Effect estimates** | **Sample size** | **Risk factors considered in the models** | **Age target** | **Stratification** |
|---|---|---|---|---|---|---|---|---|---|---|
| Boyle 2004 | Case–control | General women | Italian - Caucasian | Invasive breast cancer | Conditional logistic regression | Absolute + relative risk | 2569 cases<br>2588 controls | Menarche age, first birth age, alcohol intake, family history, age of diagnosis in relatives, and one of the two diet scores. BMI and HRT were included only for women older > 50. | 23–74 years(cases)<br>20–74 years (controls) | Age (< 50 & > 50) |
| Lee 2004 | Case–control | 1-General women<br>2-Well educated (nurse/teacher) | Korean - Asian | Invasive breast cancer | Hosmer-Lemeshow goodness of fit | Probability | 384 cases<br>270 controls | With hospitalized controls: family history, menstrual regularity, total menstrual duration, first full-term pregnancy age, breastfeeding duration while with nurse/teacher controls: age, menstrual regularity, drinking status, smoking status | Age at least 20 years | None |
| Novotny 2006 | Case–control | General women | Czeck females - Caucasian | Invasive breast cancer | Unconditional Logistic regression | Relative risk | 4598 matched pairs | Age at birth of first child, family history of breast cancer, No. of previous breast biopsy, menarche age, parity, history of benign breast disease | Age matched | None |
| Gail 2007 | Case- control | General women | African American | Invasive breast cancer | Conditional logistic regression | Absolute + relative risk | 1607cases<br>1647 controls | Menarche age, No. of affected mother or sisters, No. of benign biopsy. | 35 – 64 years | Age (< 50 & > 50) |
| Matsuno 2011 | Case- control | General women | Asian and Pacific Islander American | Invasive breast cancer | Conditional logistic regression | Absolute + relative + attributable<br>risks | 589 cases<br>952 controls | Menarche age, age at first live birth, No. of biopsies, family history, ethnicity | Any age | Ethnicity |

| **Author/ Model** | **Study design** | **Participants** | **Ethnicity** | **Outcome** | **Statistical method** | **Effect estimates** | **Sample size** | **Risk factors considered in the models** | **Age target** | **Stratification** |
|---|---|---|---|---|---|---|---|---|---|---|
| Banegas 2012 | Longitudinal study | General women | Hispanic | Invasive breast cancer | Cox proportional hazards regression | Relative risk | 6,353 cases<br>128,976 controls | Age, age at first live birth, menarche age, No. of first-degree relatives with breast cancer, No. of breast biopsies. | Postmenopausal participants aged ≥ 50 | None |
| Pfeiffer 2013 | Prospective study | White over 50 yrs. old | Whit e& non-Hispanic Caucasian | Invasive breast cancer | Cox proportional hazards regression | Relative and attributable<br>risks | 7,695 cases<br>240,712 controls | BMI, oestrogen, and progestin MHT use, other MHT use, parity, age at first birth, premenopausal, age at menopause, benign breast diseases, family history of breast or ovarian cancer, and alcohol consumption. | 50 and above | None |
| Park 2013 | Case- control | General women | Korean - Asian | Invasive breast cancer | Unconditional Logistic regression | Absolute risk | 3,789 cases<br>3,789 controls | Family history, menarche age, menopausal status, menopause age, pregnancy, first full-term pregnancy age, No. of pregnancies, breastfeeding duration, OC usage, HRT, exercise, BMI, smoking, drinking, No. of breast examinations. | Any age | Age (< 50 & > 50) |
| Lee 2015 | Case- control | General women | Asian | Invasive breast cancer | Conditional logistic regression | | 2,291 cases and 2,283 controls | First full-term pregnancy age, children No., menarche age, BMI, family history, menopausal status, regular mammography, exercises, oestrogen exposure duration, gestation period, menopause age | Any age | Age (< 50 & > 50) |

### 4.5 Discussion

There is increasing interest among clinicians, researchers, and the public in the use of risk models. This makes it important to fully evaluate model development and application. Each risk model should be assessed before it can be recommended for any clinical application. Performance assessment should involve the use of an independent population [237] separate from the population used to build the model. Breast cancer risk models that include non-genetic and non-clinical risk factors were reviewed and any model included clinical risk factors were excluded. By using PubMed, ScienceDirect, Cochrane library and other research engines, 14 models met these criteria. The most recent model examined was developed in 2015 [228]. Most models were based on two earlier risk models developed over 20 years ago - the Gail model [219] and the Rosner and Colditz model [220]. The modified versions of these two original models varied in the risk factors included and the estimation methods used. In 2012, there were two literature reviews published which analysed breast cancer risk prediction models [10, 48], however the review focuses particular on modifiable risk factors and/or self-reported factors and have updated the models published after 2012 [216, 227, 228].

Most models with modifiable risk factors included report acceptable calibration, with E/O close to 1 but less acceptable discrimination with C-statistic close to 0.5. Calibration and validation were improved when more significant factors were included. A possible explanation for less acceptable discrimination performance could be the inclusion of weaker evidence-based factors (probable and possible risk factors). All the models had combinations of probable and possible factors with no single model restricted to the inclusion of the significant factors. The high discriminatory power in the Korean model KoBCRAT especially among NCC cohort could potentially explain by several factors. Firstly, the model used 7 significant BC risk factors (5 out of them were modifiable factors). Secondly, the population characteristics (participants were ≥30 years, recruited from cancer screening program, from a teaching hospital in an urban area). While the other Korean KMCC were aged from 15 to 85 years and from rural. Thirdly, participants were better educated and compared to the other validation groups so they might be more aware of BC risk factors. Fourthly, the model divided participants according to age (<50 and ≥50 years) which could result in improvement of model discrimination between two groups as risk factors varies between pre- and post-menopausal females.

Various factors affect model performance. Inclusion of less significant factors is likely to occur in studies with small sample sizes [210, 229]. Some important clinical risk factors were not included and this may affect the model's final performance [238]. Breast cancer heterogeneity may also contribute to poor performance as different cancer types may have different risk factors [210]. Most of the models included in this review did not stratify breast cancer into its subtypes during model development. Rosner and Colditz however evaluated the model's performance based on breast cancer subtypes (ER±, PR± or HR2±) and concluded that risk factors vary according to the subtypes [239, 240]. Finally, even when strong risk factors are included in a model, significant increases in C-statistic have not been seen [241].

Model performance statistics were affected by the criteria used to stratify the analysis. Four models were stratified by age (below 50 and above 50). One model [221] was stratified by menopausal status and one model [236] was stratified by ethnicity. Another study (Rosner [220] was stratified by number of births. Breast cancer risk models could be improved if appropriate factors were used to stratify the population. For example, pre-menopausal and post-menopausal females have different risk factors in breast cancer development. The models that applied menopausal status have some limitation in that this may not be applicable to women who have had hysterectomy. For example, 30% of US women have a hysterectomy and the likelihood of oophorectomy varies by age at hysterectomy.[18] Hence completion of risk assessment outside of a clinical setting is problematic as women may be challenged to define their menopausal status. Even though the overall performance of these models appears to be moderate in differentiating between cases and non-cases, they may still serve as a good educational tool as part of cancer prevention. Utility evaluation assesses the public's knowledge of breast cancer risk factors rather well and could be used to promote cancer risk reduction actions.

A significant limitation in the development of risk models is the absence of consensus standards for defining and classifying a model's performance. For example, what is the level of good or acceptable calibration or measures of discrimination? What are acceptable measures of specificity and sensitivity in diagnostic / prognostic / preventive models? How close to unity should calibration and discrimination be for a model to be considered valid? What is the utility cut-off in each type of model? All of these questions are hard to answer without global agreement. However, this lack of consensus is understandable as these values vary depending on the type of the model type (diagnostic, prognostic, preventive), goal (clinical tool, educational tool, screening tool), targeted audience (public, high risk patients, patients visiting the clinic) and the disease itself and its types or subtypes (such as breast cancer, familial breast cancer, lobular/ductal/invasive/in situ carcinoma breast cancer). This suggests that the closer value of E/O and C-statistics to 1, the better model performance. Such a pragmatic attitude permits us to begin to focus on improving the availability of effective risk reduction actions.

Furthermore, some of the models reviewed cannot be applied to some of the populations as the risk factors may vary between different populations. For example, alcohol consumption would not be applicable to Muslim women. It is recommended that researchers develop a more reliable and valid breast cancer risk model which has good calibration, accuracy, discrimination, and utility where both internal and external validation indicates that it can be reliable for general use. In order to improve our models the following should be considered: 1) the model type (diagnostic, prognostic, preventive), goal (clinical tool, educational tool, screening tool), targeted audience (public, high risk patient), 2) inclusion of significant risk factors while incorporating the clinical and/or genetic risk factors where possible, 3) dividing the model into disease subtypes, age and menopausal status, 4) ensuring that a model is developed that can be validated externally.

## Acknowledgements

The research team is very grateful to Ministry of Higher Education in Oman for supporting the author with her PhD studies without their financial support it would not be possible to publish this paper.

**Appendix 1: Models reviewed in this review**

| Title | Size of study | Population | First author | Reference |
|---|---|---|---|---|
| **Included in this review** | | | | |
| Projecting individualized probabilities of developing breast cancer for white females who are being examined annually | 2,852 cases<br>3,146 controls | Caucasian | Gail 1989 | [219] |
| Reproductive risk factors in a prospective study of breast cancer: the Nurses' Health Study. | 2,341 cases, 91,523 controls | Caucasian | Rosner 1994 | [220] |
| Nurses' health study: log-incidence mathematical model of breast cancer incidence. | 2,249 cases, 89,132 controls | Caucasian | Rosner 1996 | [218] |
| Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study. | 1,761cases,<br>56,759 controls | Caucasian | Colditz | [95] |
| Estimation of individualized probabilities of developing breast cancer for Japanese women. | 376 cases<br>430 controls | Asian | Ueda | [221] |
| Contribution of three components to individual cancer risk predicting breast cancer risk in Italy. | 2569 cases<br>2588 controls | Caucasian | Boyle | [222] |
| Determining the Main Risk Factors and High-risk Groups of Breast Cancer Using a Predictive Model for Breast Cancer Risk Assessment in South Korea. | 384 cases<br>270 controls | Asian | Lee | [223] |

**Appendix 1: Models reviewed in this review**

| Title | Size of study | Population | First author | Reference |
|---|---|---|---|---|
| Breast cancer risk assessment in the Czech female population–an adjustment of the original Gail model. | 4598 matched pairs | Caucasian | Novotny | [217] |
| Projecting individualized absolute invasive breast cancer risk in African American women. | 1607cases<br>1647 controls | African | Gail | [224] |
| Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. | 589 cases<br>952 controls | Asian | Matsuno | [225] |
| Evaluating breast cancer risk projections for Hispanic women. | 6,353 cases<br>128,976 controls | Hispanic | Banegas | [226] |
| Risk Prediction for Breast, Endometrial, and Ovarian Cancer in White Women Aged 50 y or Older: Derivation and Validation from Population-Based Cohort Studies | 42,821 cases<br>114,931 controls | White, non-Hispanic women aged 50+ | Pfeiffer | [242] |
| Korean risk assessment model for breast cancer risk prediction. | 3,789 cases<br>3,789 controls | Asian | Park | [216] |
| Computational Discrimination of Breast Cancer for Korean Women Based on Epidemiologic Data Only. | 2,291 cases and 2,283 controls | Asian | Lee | [228] |

**Reference excluded in this review**

[32, 243-289]

Appendix 2: Epidemiological factors used in building other risk models and were not included in our reviewed models

| Reference number | Risk factors |
|---|---|
| [290] | Prostate cancer history |
| [291] | Ovarian cancer history |
| [292] | Onset of the affected family members with BC |
| [293] | Duration of natural pre-menopausal period |
| [252] | Prior breast procedure |
| [294] | History of hypertension, diabetes, high cholesterol, cardiovascular diseases, rheumatoid arthritis, osteoporosis, use of aspirin and calcium supplements. |

| **Reference number** | **Risk factors** |
|---|---|
| [248] | Education, occupation, family income, history of severe stress |

# Chapter 5 : Risk of breast cancer in the UK Biobank Female cohort and its relationship to anthropometric and reproductive factors

Publication number 2

## Risk of breast cancer in the UK Biobank Female cohort and its relationship to anthropometric and reproductive factors

This chapter is presented as a journal article:

**Al-ajmi, K.,** et al., Risk of breast cancer in the UK Biobank female cohort and its relationship to anthropometric and reproductive factors. Plos One, 2018. **13**(7).

## Risk of breast cancer in the UK biobank female cohort and its relationship to anthropometric and reproductive factors

Kawthar Alajmi[1], Artitaya Lophatananon[1], William Ollier[1], Kenneth R Muir[1].

[1] Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom

### 5.1 Abstract

**Background:** The incidence of breast cancer in the UK is one of the highest in the world. Breast cancer is a multifactorial condition where many risk factors can contribute to its development. Anthropometric and reproductive factors have been reported as being established risk factors for breast cancer. This study explores and measures the contribution of anthropometric and reproductive factors in UK females developing breast cancer in a large longitudinal cohort. **Methods:** Data from the UK Biobank prospective study of 273,467 UK females (mean follow up of 6.9 years, up to 2016) were analysed. Incident cases of breast cancer were based on ICD10, ICD9 classification and self-reported data. Relative risks (RRs) and 95% confidence intervals (CIs) for each factor were adjusted for age, family history of breast cancer and deprivation score. The analyses were stratified by the menopausal status of individuals. **Results:** Over the 9 years period of follow up the total number of breast cancer cases was 14,231 with 3,378 (23.7%) incident cases and 10,853 (76.3%) prevalent cases; breast cancer incidence rate was 2.09 per 1000 person-years. In pre-menopausal women, increase in age, being taller, having low BMI, low waist, low waist to hip ratio, family history of breast cancer in first degree relatives, early age at menarche, nulliparous, late age at first live birth, high reproductive interval index, and long contraceptive use duration were all significantly associated with an increased breast cancer risk.

In post-menopausal women, getting older, being taller, having high BMI, large waist, large hip circumference, breast cancer family history in first degree relatives, nulliparous, late age at first live birth, and high reproductive interval index were all significantly associated with an increased risk of breast cancer development. The population attributable fraction (PAF) suggested that an early first live birth, lower reproductive interval index and increase number of children can contribute to BC risk reduction up to 50%. **Conclusions:** This study utilises the large sample size, statistical power and longitudinal study design of UK Biobank to collect broad risk exposure data and confirms and extend associations between anthropometric and reproductive factors and the risk of breast cancer development in UK women. Furthermore, it has enabled us to calculate the attributable fraction of risk contributed by each risk factor. From this it was deduced that UK females can reduce their lifetime risk of breast cancer by controlling their weight, reassessing individual approaches to the timing of childbirth and options for contraception and consider early screening for women with family history in the first degree relative

## 5.2 Introduction

Breast cancer (BC) is the most common cancer in females, globally accounting for 23% of all new female cancers [40, 127, 172, 200]. In the UK, BC accounts for 15% of all newly diagnosed cancer cases in the population regardless of gender [110]. Global variations in BC incidences arise mainly from the availability of early detection and treatment facilities; however other factors may also affect this variation. Factors such as population structure (age, ethnicity, and race), life expectancy, environment, lifestyle, prevalence of risk factors, health insurance status, availability of new treatments, and pathology can enhance this variation [295]. Several risk factors have been reported in the literature. Reproductive risk factors including, early age at menarche, late menopause age, late age at first birth, low parity, hormonal replacement therapy usage, contraceptive use, hysterectomy and bilateral oophorectomy have all been identified as conferring risk for developing BC [192, 296]. Another major factor for increasing BC incidences is the accumulated effect of anthropometric factors. Increased height, weight, hip circumference, waist circumference, body mass index (BMI), and waist to hip ratio (WHR) have been reported as increasing BC risk depending on the menopausal status of women [297]. Given the unique opportunity the UK biobank [40] project offers for assessing a wide range of disease risk factors in a large longitudinal cohort, the effect of anthropometric and reproductive factors on BC risk was measured.

## 5.3 Materials and methods

### 5.3.1 Study population and study design

UK Biobank is a national-based health project that aims to improve the diagnosis, treatment, and prevention of diseases such as cancers, diabetes, stroke, heart disease, osteoporosis, arthritis, eye diseases, dementia and depression [40]. A total of 502,650 participants aged between 39 to 71 years were enrolled in the study between 2006 and 2010 and they continue to be clinically followed up. Details can be found at http://www.ukbiobank.ac.uk/. In addition to the collection of biological samples (blood, saliva and urine), health, demographic and anthropometric data were collected in 22 UK assessment facilities across England, Wales and Scotland. Detailed physical / physiological measurements were further supported by the administration of questionnaires and eye examination. Many participants completed additional detailed questionnaires on work history, diet, and cognitive function. Anonymised data are now available to researchers across the world [40, 172]. Our study acquired data on the female cohort (273,467 female participants) from UK Biobank. The UK Biobank female cohort had a mean follow up time of 6.9 years (at 2016). Data on exposures were defined prior to the development of BC in cases or prior to the first assessment date in controls.

### 5.3.2 Defining breast cancer cases and controls

BC was defined as a malignant neoplasm of the breast. The UK Biobank database contained record of all cancers including their subtype occurring either before or after participant enrollment using the International Classification of Diseases (ICD10, ICD9) and their self-reported data. Details of codes used to identify BC cases are summarized in Table 5.1.

Table 5.1: Codes used to identify breast cancer cases and controls

| Categories | Frequency (%) | ICD10 codes | ICD9 codes | Self-reported cancer's codes | Self-reported non-cancer diseases |
|---|---|---|---|---|---|
| **Breast cancer cases** | | | | | |
| Incident: | 3,378 (1.24%) | Codes start with C50 and its subclasses , C501, C502, C503, C504, C505, C506, C507, C508, and C509 | Codes start with 174 and its subclasses 1741, 1742, 1743, 1744, 1745, 1746, 1747, 1748, and 1749 | 1002 code only | - |
| Prevalent: | 10,853 (3.97%) | | | | |
| **Subjects excluded from the study** | | | | | |
| 6- Other cancers | 23,540 (8.61%) | Codes start with C except codes for BC | Codes start with 1 or 20 except codes for BC | All other codes except 1002 code | - |
| 7- Breast In situ carcinoma | 636 (0.23%) | Codes of D050, D051, D057, D059 | 2330 code only | - | - |

| Categories | Frequency (%) | ICD10 codes | ICD9 codes | Self-reported cancer's codes | Self-reported non-cancer diseases |
|---|---|---|---|---|---|
| 8- Other in situ carcinoma | 2,463 (0.90%) | Codes start with D0 except codes for breast in situ carcinoma | Codes start with 230 or 231 or 232 or 233 or 234 except codes for breast in situ carcinoma | - | - |
| 9- Neoplasm of unknown nature or behavior | 121 (0.04%) | Codes start with D37 or D38 or D39 or D40 or D41 or D42 or D43 or D44 or D45 or D46 or D47 or D48 | Codes start with 235 or 236 or 237 or 238 or 239 | - | - |
| **Controls** | | | | | |
| All controls | 232,476 (85.01%) | Remaining codes or subjects with no code assigned | Remaining codes or subjects with no code assigned | - | All of Self-reported of non-cancerous diseases or no code assigned |
| **Total** | 273,476 (100%) | | | | |

#### 5.3.2.1 Breast cancer cases

In the database, each participant had 9 follow-up time point records for ICD10, 11 follow-up time point records for ICD9 and 9 follow-up time point records for self-reported status of cancer. The case-control groups were identified by utilising all these three data sources. The codes for BC are presented in table 5.1. Cases were characterised as incident or prevalent using 'age or date when they attended the center' and 'age when first reported BC cancer'. With cases defined by ICD10 and ICD9, if their 'attending age' was greater than 'cancer diagnosis age' then this was considered as a prevalent case. Subjects were considered to be incident cases if their 'attending age' was less than their 'cancer diagnosis age'. For self-reported cases, the same criteria were applied. Age when first attended the assessment centre was compared with the interpolated age of the participant when cancer was first diagnosed. To combine and classify the type of cases from 3 different sources, the following criteria were applied:

1. If the BC cases appeared as being incident using any of these three identification methods then the cases were deemed to be incident cases.

2. Prevalent cases were defined using combination of rules a) only if the participant has been identified as a prevalent case by any of the three methods and b) none of these methods define the same participant as being an incident case.

In total, there were 14,231 BC cases with 3,378 being incident cases and 10,853 prevalent cases.

#### 5.3.2.2 Controls

Female participants were defined as controls if they had no record of cancer, *in-situ* carcinoma or an undefined neoplasm (232,476 controls).

#### 5.3.2.3 Exclusion criteria

In the case group, 10,853 (3.97%) prevalent BC cases were excluded. In the control group, participants were excluded due to following reasons; other type of cancers (23,540), breast *in situ* carcinoma (636), other *in situ* carcinomas (2,463) and unknown neoplasm (121).

### 5.3.3 Exposures

Reproductive variables included menarche age, menopause age, menopausal status, parity (yes/no), number of children, age at first live birth, pregnancy history, pregnancy termination and number of terminations, reproductive interval index (difference between menarche age and age at first birth), history of oral contraceptive (OC) use and its duration, and history of hormonal replacement therapy (HRT) use and its duration. Anthropometric variables included BMI, waist and hip circumferences, waist to hip ratio (WHR) and height (sitting and standing).

## 5.4 Statistical analysis

To assess associations between exposures and BC risk in the cohort, relative risk (RR) and 95% confident intervals (95% C.I.) were computed using a binomial generalised linear regression model. Regression analyses were performed for each independent variable and were adjusted for age, family history of BC in first degree relatives, and deprivation score. The independent variables list and description are presented in Table 5.2.

Table 5.2: classification of the variables included in the analysis

| | Variable | Groups | Coding |
|---|---|---|---|
| 1 | Menopausal status | Pre-menopausal | Pre- menopausal: reported as pre-menopausal & no history of hysterectomy or bilateral oophorectomy & their age is ≤55 and menarche age ≥7 years old (to maximise the number of the real pre-menopausal females - so any female reported as pre- and their age > 55 or had menarche age < 7 and did not had hysterectomy nor oophorectomy will be removed – most probably this female is miscategorised). |
| | | Post-menopausal | Post-menopausal: reported as post-menopausal & no history of hysterectomy or bilateral oophorectomy (only natural menopause) & their menopause age is ≥ 40years old (to maximise the number of the real post-menopausal females – so any female reported as post- and their menopause age < 40 and did not had hysterectomy nor oophorectomy will be removed - most probably this female is miscategorised). |
| 2 | Menarche age | (>13) | This variable was divided into two groups based on literature ranges [127]. |
| | | (≤13)) | |
| 3 | Age at first birth | (<20) | This variable was divided into four groups based on literature ranges [192]. |
| | | (20-24) | |
| | | (25-29) | |
| | | (≥30) | |
| 4 | BMI | Healthy (18.5 – 24.9) | This variable was divided into three groups based on WHO classification [193]. Underweight group were very low in number and would not be enough for the association calculations. |
| | | Overweight (25-29.9) | |
| | | Obese (≥30) | |
| 5 | Waist to hip ratio (WHR) | Low (≤0.80) | This variable was calculated by dividing the waist over the hip measurements of the participants. Then later was divided into three groups based on WHO classification [194]. |
| | | Moderate (0.81-0.85) | |
| | | High (>0.85) | |
| 6 | Reproductive interval index- years | Low (≤12) | Reproductive interval index is the difference between the age at first birth and age at menarche. The index was divided into four groups based on the IQR (Interquartile range) of the reproductive interval index values among the controls only. |
| | | Moderate (12-16) | |
| | | High (>16) | |
| | | No children | |
| 7 | Deprivation score | Calculated by UK biobank team | The score was calculated for all the participants before participating in UK Biobank. The score was based on the prior national census output areas [195]. The score evaluates four aspects: 1) unemployment, 2) houses without an owned car, 3) non-house ownership, 4) overcrowding in one house [196]. Data were provided by the UK biobank team. |
| 8 | Height | Below mean<br>(< 156.10 cm) | |

|  | Variable | Groups | Coding |
|---|---|---|---|
|  |  | Within mean ± SD (156.10-168.75cm) | The height was grouped into three groups based on the mean of the control group. |
|  |  | Above mean (>168.75 cm) |  |

All analyses were stratified by menopausal status: pre- and post-menopausal. The criteria for pre-menopausal were females aged ≤ 55 years old (according to the NHS the menopause age in the UK is between 40 to 55 years [174]) who reported that they still had periods and did not report a history of hysterectomy or bilateral oophorectomy, and menarche age ≥ 7 years old (the menarche age in the UK ranges from 7 to 20 years [175]). Post-menopausal females were defined as those who reported no longer having periods and did not report a history of hysterectomy or bilateral oophorectomy and their age at menopause was reported to be more than 40 years old. These criteria were employed to minimise inclusion of both pre-mature and the medically induced pre- or post-menopausal women. After further application of criteria, 61,903 participants were in pre-menopause group and 133,704 participants were in post-menopause group.

To compute BC incidence within the cohort, the Stata *stptime* command was used to obtain the overall person-time of observation and disease incidence rate. To calculate time for each participant, the endpoint was subtracted (either the date of cancer diagnosis or the end of the follow-up - January 1st, 2016) with the date of study enrolment. Incidence rates were estimated for the whole cohort and pre- and post-menopausal separately. Moreover, population attributable fractions (PAF) were calculated using the *punaf* command [176] where the fraction was estimated compared to whole cohort and compared to the most significant subgroup associated with the BC. This was done to estimate how much risk could be eliminated by controlling that risk factor in both groups.

All statistical analysis was performed using Stata MP 14.1 software for Windows [177]. Results with 95% confident intervals not including 1 were considered as being statistically significant.

## 5.5 Results

The UK biobank female cohort consisted of 273,476 female participants with a mean age of 56.3 years (SD ±8.00). The follow up time was 9.8 years up to January 2016 where the database was frozen for this analysis. The total number of BC cases was 14,231 with 3,378 (1.24%) incident cases and 10,853 (3.97%) prevalent cases. The total number of controls was 232,476 (85.01%). The remaining participants were either females with other cancer 23,540 (8.61%) or with breast *in situ* carcinoma 636 (0.23%), or other *in situ* carcinoma 2,463 (0.90%) or unknown neoplasm 121 (0.04%). A total of 3,162 (93.60%) of incident cases were identified by ICD10 and the rest 216 (6.40%) were identified by self-reporting. All the BC cases identified by ICD9 were solely prevalent cases. When further applying criteria for menopause status, the total number of pre-menopausal females was 61,903 (31.65 %) and post-menopausal was 133,704 (68.35 %). Out of the total pre-menopausal females, 618 (1.07%) were incident cases and 57,089 (98.93%) were controls. For post-menopausal females, 1,757 (1.53%) were incident cases and 112,757 (98.47%) were controls (Figure *5.1*). The BC incidence rate of the whole cohort was 2.09 per 1000 person-years. The pre-menopause BC incidence rate was 1.55 per 1000 person-years and the post-menopause BC incidence rate was 2.24 per 1000 person-years. The incidence rate ratio between the pre- and post-menopausal females is 1.45 with 95% CI 1.32 - 1.59.

Comparisons of mean values of age, deprivation score, anthropometric and reproductive variables (all continuous variables) of the participants conditioned on the menopausal status are summarised in 5.3. In both pre- and post-menopause groups, cases were older than controls and the mean age differences were statistically significant (*Student's t-test p-values<0.05*). Results using the Townsend deprivation score showed that case's mean score were significantly lower than control mean score in both pre- and post-menopause females (Student's *t-test* p-values < 0.05).

For anthropometric variables, in the pre-menopausal group, the mean values of standing and sitting height in cases were higher as compared to controls (*Student's t-test p-values<0.05*). On the other hand, mean values of BMI, waist circumference and waist to hip ratio were significantly lower in cases as compared with controls (*Student's t-test p-values<0.05*). In the post-menopause case group, the mean values of standing and sitting height, BMI, waist circumference, and hip circumferences were higher when compared with controls (*Student's t-test p-values<0.05*).

Analysis of reproductive factors in pre-menopause case group, showed higher mean values of age at first birth, reproductive interval index, and contraceptive use duration as compared with controls (Student's *t-test p values <0.005*). In addition, among the post-menopausal group, mean values of menopause age and duration of HRT use were significantly higher in cases compared with controls. In contrast, mean values of number of live births were lower in cases as compared to controls in post-menopausal females.

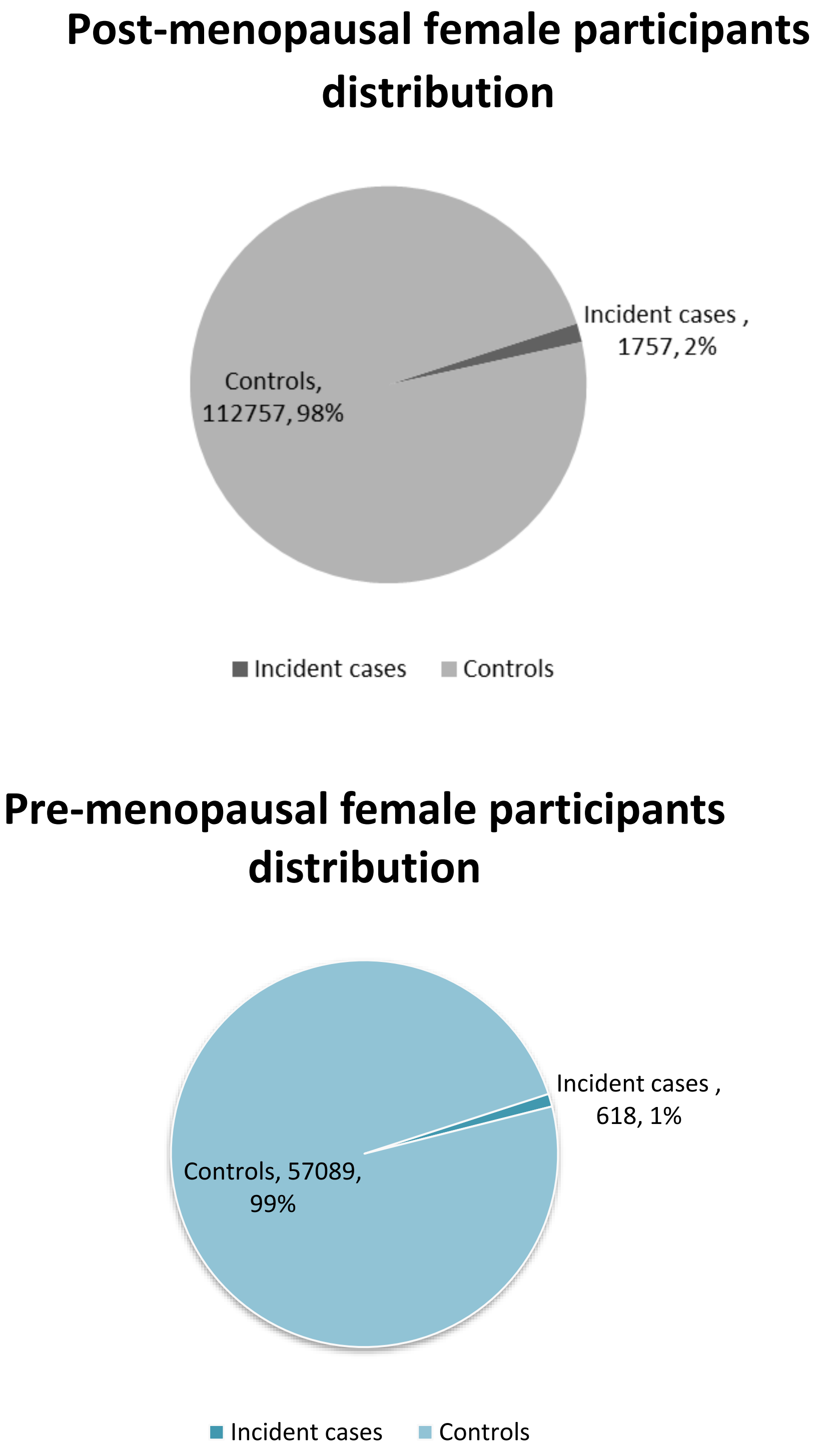


Figure 5.1: UK biobank data distribution based on menopausal status

Table 5.3: Mean comparisons between cases and controls in pre- and post-menopause status

| | Pre-menopausal | | | | Post-menopausal | | | |
|---|---|---|---|---|---|---|---|---|
| **Variables** | **No. (cases/controls)** | **Case's mean** | **Control's mean** | **P-value*** | **No. (cases/controls)** | **Case's mean** | **Control's mean** | **P-value*** |
| Age (year) | (618/ 57,089) | 46.43 | 45.83 | <0.001 | (1,757 /112,757) | 60.67 | 59.76 | <0.001 |
| Deprivation score | (618/56,999) | -1.49 | -1.09 | 0.0014 | (1,755 /112,639) | -1.72 | -1.48 | 0.006 |
| Body shape measures | | | | | | | | |
| BMI (kg/m$^2$) | (612/ 56,847) | 25.95 | 26.43 | 0.0263 | (1,750/112,270) | 27.45 | 27.01 | 0.0004 |
| Waist Circumference (cm) | (613 /56,890) | 80.97 | 82.23 | 0.0122 | (1,752/112,426) | 86.03 | 84.78 | <0.001 |
| Hip Circumference (cm) | (613 /56,889) | 102.16 | 102.51 | 0.4075 | (1,752/112,423) | 104.32 | 103.12 | <0.001 |
| Waist to Hip ratio | (613 /56,883) | 0.79 | 0.80 | 0.0008 | (1,752/112,416) | 0.82 | 0.82 | 0.1144 |
| Standing Height (cm) | (612 /56,896) | 164.70 | 164.04 | 0.0107 | (1,751/112,391) | 162.61 | 161.91 | <0.001 |
| Sitting height (cm) | (603 /56,406) | 87.86 | 87.54 | 0.0305 | (1,724/111,654) | 86.36 | 86.03 | 0.0003 |
| Reproductive factors measures | | | | | | | | |
| Menarche age (year) | (605 /55,286) | 12.95 | 13.05 | 0.1051 | (1,727/110,214) | 12.93 | 12.98 | 0. 1783 |
| Menopause age (year) | N/A | | | | (1,757/112,757) | 50.85 | 50.58 | 0.0065 |
| Number of live births | (618 /57,053) | 1.49 | 1.57 | 0.0945 | (1,754/112,685) | 1.77 | 1.88 | 0.0001 |
| Age at first birth (year) | (336 /33,071) | 27.70 | 27.03 | 0.0152 | (1,171/79,421) | 25.46 | 25.30 | 0.2311 |
| Number of Pregnancy termination | (221 / 20,149) | 0.61 | 0.69 | 0.1270 | (529/34,166) | 0.47 | 0.52 | 0.1399 |
| Reproductive interval index (year) | (521/47,237) | 14.66 | 13.93 | 0.0113 | (1,483 /96,718) | 12.50 | 12.29 | 0.1310 |
| Contraceptive use duration (year) | (519/ 50,012 | 11.62 | 9.99 | <0.001 | (1,610/ 102,760) | 7.51 | 7.68 | 0.3859 |
| HRT duration (year) | (609/56,210) | 0.05 | 0.03 | 0.2001 | (1,553/ 102,786) | 2.25 | 1.92 | 0.0005 |
| Total | 618 / 57,089 | | | | 1,757 / 112,757 | | | |

**Student's t-test*

Relative risks (RRs) of the key characteristics and anthropometric measures of pre- and post-menopausal females are illustrated in Table *5.4*.4. For both pre-and post-menopausal females, age as a continuous variable showed a slight increased risk of developing BC (RR=1.05, 95%CI; 1.02-1.07) and RR=1.03, 95%CI; 1.02-1.04, respectively). Results of Townsend deprivation score showed a decreased risk of BC associated with increased deprivation score (more deprived) among both pre- (RR=0.96, 95%CI; 0.94-0.99) and post-menopausal (RR=0.97, 95%CI; 0.96-0.99) females.

Table 5.4: Relative risk of key characteristics and anthropometric factors in pre- and post- menopausal females

| Menopausal status | Pre-menopausal | | | | | Post-menopausal | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **Number of cases/controls** | **RR** | **P-value** | **LCL** | **UCL** | **Number of cases/controls** | **RR** | **P-value** | **LCL** | **UCL** |
| Age in years (Continuous) * | 618/ 57,089 | 1.046 | <0.001 | 1.024 | 1.069 | 1,757 /112,757 | 1.033 | <0.001 | 1.024 | 1.042 |
| Deprivation score (Continuous) ** | 618/56,999 | 0.962 | 0.004 | 0.937 | 0.988 | 1,755 /112,639 | 0.973 | 0.001 | 0.957 | 0.990 |
| Family history *** | | | | | | | | | | |
| No | 520/ 51,547 | Ref | | | | 1,458/ 99,998 | Ref | | | |
| Yes | 97/ 5,326 | 1.770 | <0.001 | 1.427 | 2.194 | 290/ 12,367 | 1.582 | <0.001 | 1.397 | 1.792 |
| Mother BC history *** | | | | | | | | | | |
| No mother BC history | 532/ 51,750 | Ref | | | | 1,529/ 102,184 | Ref | | | |
| Mother BC history | 78/ 4,360 | 1.724 | <0.001 | 1.362 | 2.181 | 192/ 8,145 | 1.569 | <0.001 | 1.353 | 1.820 |
| Sibling BC history *** | | | | | | | | | | |
| No sibling BC history | 579/ 54,125 | Ref | | | | 1,553 / 103,570 | Ref | | | |
| Sibling BC history | 23/ 1,108 | 1.823 | 0.004 | 1.206 | 2.756 | 120/ 4,782 | 1.613 | <0.001 | 1.343 | 1.938 |
| Family history- Combined*** | | | | | | | | | | |
| No family history at all | 520/ 51,547 | Ref | | | | 1,458/ 99,998 | Ref | | | |
| Mother or Sister BC history | 93/ 5,184 | 1.756 | <0.001 | 1.408 | 2.190 | 268/11,807 | 1.540 | <0.001 | 1.351 | 1.754 |
| Mother and Sister BC history | 4/142 | 2.592 | 0.054 | 0.982 | 6.837 | 22/560 | 2.594 | <0.001 | 1.717 | 3.920 |
| BMI in kg/m$^2$ (Continuous) | 612/ 56,847 | 0.983 | 0.041 | 0.968 | 0.999 | 1,750/112,270 | 1.018 | <0.001 | 1.009 | 1.027 |
| BMI – categorical | | | | | | | | | | |
| BMI - Healthy (18.5-24.9) | 326/26,983 | Ref | | | | 626/44,215 | Ref | | | |
| BMI - Overweight (25-29.9) | 186/18,319 | 0.839 | 0.055 | 0.701 | 1.004 | 681/42,624 | 1.102 | 0.078 | 0.989 | 1.228 |
| BMI - Obese (>=30) | 100/11,545 | 0.733 | 0.007 | 0.586 | 0.918 | 443/25,431 | 1.241 | 0.001 | 1.098 | 1.401 |
| Waist Circumference in cm (Continuous) | 613 /56,890 | 0.992 | 0.020 | 0.985 | 0.999 | 1,752/112,426 | 1.008 | <0.001 | 1.004 | 1.012 |
| Hip Circumference in cm (Continuous) | 613 /56,889 | 0.997 | 0.518 | 0.990 | 1.005 | 1,752/112,423 | 1.012 | <0.001 | 1.007 | 1.016 |
| Waist to Hip (Continuous) | 613 /56,883 | 0.131 | 0.001 | 0.038 | 0.446 | 1,752/112,416 | 1.520 | 0.226 | 0.772 | 2.994 |
| Waist to Hip – categorical | | | | | | | | | | |
| Waist to Hip - Low (<=0.80) | 362/30,170 | Ref | | | | 678/45,184 | Ref | | | |
| Waist to Hip - Moderate (0.81-0.85) | 139/13,993 | 0.829 | 0.060 | 0.682 | 1.008 | 475/30,741 | 1.010 | 0.869 | 0.898 | 1.135 |
| Waist to Hip - High (>0.85) | 112/12,720 | 0.744 | 0.006 | 0.602 | 0.920 | 599/36,491 | 1.073 | 0.213 | 0.961 | 1.198 |

| Menopausal status | Pre-menopausal | | | | | Post-menopausal | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **Number of cases/controls** | **RR** | **P-value** | **LCL** | **UCL** | **Number of cases/controls** | **RR** | **P-value** | **LCL** | **UCL** |
| Sitting Height in cm (Continuous) | 603 /56,406 | 1.023 | 0.041 | 1.001 | 1.046 | 1,724/111,654 | 1.032 | <0.001 | 1.019 | 1.046 |
| Standing Height in cm (Continuous) | 612 /56,896 | 1.017 | 0.010 | 1.004 | 1.030 | 1,751/112,391 | 1.021 | <0.001 | 1.013 | 1.029 |
| Standing Height in cm – categorical | | | | | | | | | | |
| Below mean ± SD (150.20-156.06 cm) | 57/6,447 | Ref | | | | 285/21,259 | Ref | | | |
| Within mean ± SD (159.21-165.71 cm) | 388/37,137 | 1.181 | 0.243 | 0.893 | 1.562 | 1,153/ 75,173 | 1.168 | 0.019 | 1.025 | 1.330 |
| Above mean ± SD (169.02-175.00 cm) | 167/13,314 | 1.429 | 0.021 | 1.057 | 1.933 | 313/ 15,964 | 1.533 | <0.001 | 1.305 | 1.802 |

All adjusted for age + Family history of BC + deprivation score; *adjusted for deprivation score only; ** no adjustment, ***Adjusted for age + deprivation score

Family history of BC is a well-defined risk factor for BC. The strength of this risk factor varies according to the number and relationship of the affected family members. Females who reported having had a family history of BC were at increased risk for developing BC in both pre- and post-menopausal females with (RR=1.77, 95%CI; 1.43-2.19) and (RR=1.58, 95%CI; 1.40-1.79), respectively. Both pre- and post-menopause subjects with their siblings affected with BC were at increased risk of 82% (pre-menopause RR=1.82, 95%CI; 1.21-2.76) and 61% (post-menopause (RR=1.61, 95%CI; 1.34-1.94) respectively. Similar results were also seen in subjects who reported only their mother affected with BC, RR=1.72 (95%CI; 1.36-2.18) in pre- and (RR=1.57, 95%CI; 1.35-1.94) in post-menopausal women. All of these significant associations were stronger among pre-menopausal compared to post-menopausal women. In the post-menopause group, subjects with both mother and sibling affected with BC were almost at three-fold increase BC risk (RR =2.59, 95%CI; 1.72-3.92). Despite a similar relative risk estimate, no association was reported in pre-menopause group (RR =2.59, 95%CI; 0.98-6.84).

For anthropometric exposures treated as being continuous variables, increasing BMI (RR=0.98, 95%CI; 0.97-1.00), waist circumference (RR=0.99, 95%CI; 0.99-1.00), and waist to hip ratio (RR=0.13, 95%CI; 0.04-0.45) were associated with reduced BC risk among the pre-menopause group. The WHR as a categorical variable (low ≤0.80 as reference group, moderate (0.81-0.85) and high >0.85) showed significant risk reduction only in the high WHR group (RR=0.74 with 95%CI; 0.60-0.92). BMI as a categorical variable showed that obese women with a BMI ≥30 had 26.7 % decreased BC risk compared to women with normal range BMI (RR=0.73 with 95%CI; 0.59-0.92). For height, per 1 cm of increased height (cm), BC risk was increased by 2% (RR=1.02, 95%CI; 1.00-1.03). Height as a categorical variable showed that women in the tallest group (height ranges from 168.8 to 199 cm) had their BC risk increased by 43% compared to shorter females with height ranges from 152.20 to 156.06 cm.

In post-menopausal women, increasing BMI (RR=1.02, 95%CI; 1.00-1.03), waist circumference (RR=1.01, 95%CI; 1.00-1.01), hip circumference (RR=1.01, 95%CI; 1.01-1.02), standing height (RR=1.02, 95%CI; 1.01-1.03) and sitting height (RR=1.03, 95%CI; 1.02-1.05) were associated with a slight increased risk of BC. BMI as a categorical variable showed that obese subjects had 24.1% increased risk for BC (RR=1.24, 95%CI; 1.10-1.40) when compared to the normal BMI group. For height treated as a categorical variable, results suggested that the tallest group (height ranges from 168.8 to 199 cm, mean=172.0) were at 53% increased risk of BC (RR=1.53, 95%CI; 1.31-1.80) when compared to the reference group (height ranges from 100 to 156 cm, mean=153.1).

### 5.5.1 Reproductive factors and breast cancer

RRs for the reproductive factors and BC risk are presented in Table 5.5. For the pre-menopause group, menarche age as continuous variable showed a slight risk reduction (RR=0.95, 95%CI; 0.90-1.00). When menarche age was grouped into >13 years old (as a reference group) versus ≤13 years old, a moderate increased risk was observed (RR=1.23, 95% CI; 1.04-1.45). For the post-menopause group, age at menarche did not show any significant association with BC risk (confidence interval value included 1).

Table 5.5: Relative risks of the reproductive factors based on the menopausal status

| **Menopausal status** | | **Pre-menopausal** | | | | | **Post-menopausal** | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **No. cases/controls** | **RR** | **P-value** | **LCL** | **UCL** | **No. cases/controls** | **RR** | **P-value** | **LCL** | **UCL** |
| Menarche age in years (Continuous)* | 605 /55,286 | 0.948 | 0.042 | 0.900 | 0.998 | 1,727/110,214) | 0.987 | 0.388 | 0.958 | 1.017 |
| Menarche age – categorical | | | | | | | | | | |
| Menarche age (>13) | 198/ 20,785 | Ref | | | | 625 /40,534 | Ref | | | |
| Menarche age (≤13) | 407/ 34,501 | 1.228 | 0.017 | 1.037 | 1.454 | 1,102/69,680 | 1.029 | 0.569 | 0.933 | 1.134 |
| Menopause age in years (Continuous)* | Not applicable | | | | | 1,757/112,757 | 1.006 | 0.284 | 0.995 | 1.018 |
| Parity | | | | | | | | | | |
| No | 188/15,024 | Ref | | | | 326/18,855 | Ref | | | |
| Yes | 430/42,029 | 0.764 | 0.002 | 0.643 | 0.908 | 1,428/93,830 | 0.821 | 0.001 | 0.728 | 0.926 |
| Number of births (Continuous) | 618 /57,053 | 0.925 | 0.024 | 0.864 | 0.990 | 1,754/112,685 | 0.899 | <0.001 | 0.863 | 0.937 |
| First live birth age in years (Continuous) | 336 /33,071 | 1.022 | 0.055 | 1.000 | 1.045 | 1,171/79,421 | 1.010 | 0.142 | 0.997 | 1.023 |
| First live birth age – categorical | | | | | | | | | | |
| First live birth age (<20) | 12/2,422 | Ref | | | | 97/7,330 | Ref | | | |
| First live birth age (20-24) | 74/7,873 | 1.719 | 0.082 | 0.933 | 3.168 | 369/27,992 | 0.966 | 0.763 | 0.773 | 1.207 |
| First live birth age (25-29) | 138/12,625 | 1.882 | 0.038 | 1.036 | 3.417 | 492/31,181 | 1.091 | 0.435 | 0.876 | 1.360 |
| First live birth age (≥30) | 112/10,151 | 1.938 | 0.031 | 1.062 | 3.539 | 186/12,918 | 1.055 | 0.669 | 0.825 | 1.350 |
| pregnancy termination | | | | | | | | | | |
| No | 117/9,544 | Ref | | | | 321/ 19,771 | Ref | | | |
| Yes | 104/10,605 | 0.835 | 0.181 | 0.641 | 1.088 | 208/14,395 | 0.981 | 0.834 | 0.823 | 1.171 |
| Pregnancy termination number (Continuous) | 221 / 20,149 | 0.898 | 0.232 | 0.753 | 1.071 | 529/34,166) | 0.973 | 0.673 | 0.858 | 1.104 |
| Reproductive interval index in years (Continuous) | 521/47,237 | 1.003 | 0.002 | 1.001 | 1.005 | 1,483 /96,718 | 1.003 | <0.001 | 1.001 | 1.004 |
| Reproductive interval index – categorical | | | | | | | | | | |
| Low index (≤12) | 109/12,673 | Ref | | | | 585/41,334 | Ref | | | |
| Moderate index (12.01-16) | 98/9,499 | 1.146 | 0.329 | 0.872 | 1.506 | 359/22,601 | 1.128 | 0.073 | 0.989 | 1.287 |
| High index (>16.01) | 126/10,041 | 1.421 | 0.008 | 1.098 | 1.838 | 213/13,928 | 1.130 | 0.128 | 0.965 | 1.323 |
| No children | 188/15,024 | 1.530 | <0.001 | 1.208 | 1.937 | 326/18,855 | 1.333 | <0.001 | 1.163 | 1.528 |

| Menopausal status | | Pre-menopausal | | | | | Post-menopausal | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Variables** | **No. cases/controls** | **RR** | **P-value** | **LCL** | **UCL** | **No. cases/controls** | **RR** | **P-value** | **LCL** | **UCL** |
| Contraceptive use | | | | | | | | | | |
| No | 53/ 6,297 | Ref | | | | 366/23,896 | Ref | | | |
| Yes | 565/50,646 | 1.261 | 0.106 | 0.952 | 1.670 | 1,389/88,638 | 1.124 | 0.053 | 0.998 | 1.265 |
| Contraceptive duration in years (Continuous) | 519/ 50,012 | 1.024 | <0.001 | 1.013 | 1.034 | 1,610/ 102,760 | 1.003 | 0.319 | 0.997 | 1.010 |
| HRT use | | | | | | | | | | |
| No | 599/ 55,336 | Ref | | | | 943 /65,669 | Ref | | | |
| Yes | 18/1,565 | 0.945 | 0.813 | 0.590 | 1.513 | 811/46,830 | 1.141 | 0.006 | 1.038 | 1.255 |
| HRT duration in years (Continuous) | 609/56,210 | 1.063 | 0.298 | 0.947 | 1.193 | 1,553/ 102,786 | 1.013 | 0.054 | 1.000 | 1.025 |
| Mammogram history | | | | | | | | | | |
| No | 359 /37,546 | Ref | | | | 50/5,408 | Ref | | | |
| Yes | 285/19,341 | 1.190 | 0.054 | 0.997 | 1.420 | 1,706/107,289 | 1.260 | 0.120 | 0.942 | 1.686 |

All adjusted for age, family history of BC and deprivation score, * adjusted more for BMI.

Parous women were at reduced BC risk in both pre- (RR=0.76, 95% CI; 0.64-0.91) and post-menopausal women (RR=0.82, 95% CI; 0.73-0.93) when compared to nulliparous women. The 'number of children' when treated as a continuous variable showed moderate decreased BC risk (pre-menopause group RR=0.93, 95% CI; 0.86-0.99 and post-menopause group RR=0.90, 95% CI; 0.86-0.94). In contrast, increasing maternal age at live birth showed slight increased BC risk in both pre- (RR=1.02, 95% CI; 1.00-1.05) and post-menopausal women (RR=1.01, 95% CI; 1.00-1.00). Further analysis was carried out in parous women to explore the association of age at live birth and BC risk. Age at first live birth as categorical variable (< 20 years old as the reference group, 20-24, 25-29, and ≥30 years old) showed that among pre-menopausal females , BC risk was almost double when they reported having had their first child at age ≥30years old and at age 25-29 years as compared to women who reported having their first baby at age <20 years old (RR 1.94; 95% CI, 1.06-3.54 and RR=1.88 with 95% CI; 1.04-3.42, respectively). This effect was not seen in post-menopausal females (all 95% CI values included 1). Both pregnancy termination history (ever versus none) and number of terminations were not significantly associated with BC development in both pre- and post-menopausal females (all 95% CI values included 1).

The Reproductive Interval Index (the difference between age at first child and the age of menarche) based on the interquartile range of the control group (low as reference group, moderate, high, and no children) only showed statistically significant increased risk in 'high' (RR=1.42, 95% CI; 1.10-1.84)) and 'no children' groups (RR=1.53, 95% CI; 1.21-1.94) in pre-menopausal females. In post-menopausal group, only females reporting no children showed an increased risk of BC by 13% (RR=1.33, 95% CI; 1.16-1.53) when compared to the low index group.

History of oral contraceptive (OC) pills used showed no association with BC risk in both pre- and post-menopause groups. Within the OC use group, however, OC duration showed a slight increased BC risk in pre-menopause women (RR=1.02, 95% CI; 1.01-1.03) but not in post-menopausal women. Hormone replacement therapy (HRT) was not associated with risk of BC in pre-menopause UK females. In the post-menopause group, women who reported using HRT were at moderate significant increased risk (RR=1.14, 95%CI; 1.04-1.26).

Women in both pre- and post-menopausal groups who reported having had mammograms were at increased risk of BC (RR= 1.19, 95% CI 1.00-1.42 and RR= 1.26, 95% CI 0.94-1.69, respectively).

PAF was calculated for the modifiable risk factors only based on the menopause status (Table 5.6). Two fractions were estimated; the PAF among the studied population and the PAF among the sub-population (the exposed significant group) to evaluate how many cases could be avoided if a particular factor was eliminated. Among pre-menopausal females, these modifiable factors were the strongest in reducing the BC risk. Giving birth at age <30 can eliminate about 44.6% of the BC cases in general population, and about 48.4% among females who had first children at age ≥30years old and about 46.9% of cases among females who had first children at age 25-29. Followed by low reproductive interval index with about 34.6% of BC cases can be eliminated among null-parous females and about 29.6% of BC cases can be eliminated among females with high index (>16.01). Being parous can eliminate only 9.2% of the cases without taking into consideration the number of children they gave birth to. Finally, having BMI ≥30 and WTH >0.85 can eliminate 70% and 66.2% of the cases among pre-menopausal women, respectively.

Table 5.6: Population Attributable Fraction (PAF) among modifiable breast cancer risk factors according to the menopausal status

| | **Pre-menopausal** | | **Post-menopausal** | |
|---|---|---|---|---|
| | **PAF in population** | **PAF in subpopulation group** | **PAF in population** | **PAF in subpopulation group** |
| BMI | | | | |
| BMI - Healthy (18.5-24.9) | Ref | | | |
| BMI - Obese (>=30) | -0.091 | -0.707 | 0.083 | 0.194 |
| Waist to Hip ratio | | | | |
| Waist to Hip - Low (<=0.80) | Ref | | | |

| | Pre-menopausal | | Post-menopausal | |
|---|---|---|---|---|
| | **PAF in population** | **PAF in subpopulation group** | **PAF in population** | **PAF in subpopulation group** |
| Waist to Hip - High (>0.85) | -0.080 | -0.662 | *NS* | *NS* |
| Parity (Yes/No) | | | | |
| Yes | Ref | | | |
| No | 0.072 | 0.092 | 0.033 | 0.179 |
| Number of births | | | | |
| None | Ref | | | |
| More than one child | 0.088 | 0.247 | 0.046 | 0.211 |
| First live birth age | | | | |
| First live birth age (<20) | Ref | | | |
| First live birth age (25-29) | *NS* | 0.469 | *NS* | *NS* |
| First live birth age (≥30) | 0.446 | 0.484 | *NS* | *NS* |
| Reproductive interval index | | | | |
| Low index (≤12) | Ref | | | |
| High index (>16.01) | 0.149 | 0.296 | *NS* | *NS* |
| No children | 0.223 | 0.346 | 0.089 | 0.250 |
| HRT use (No /Yes) | | | | |
| No | Ref | | | |
| Yes | *NS* | *NS* | 0.058 | 0.125 |

Among post-menopausal women; based on their population attributable fractions, reducing BMI to under 30 could result in an 8.3% reduction in BC cases in the general population and a 19.4% reduction in BC among obese females; being parous can eliminate 17.9% among null parous females; having more than one child can eliminate 21.1% % among females with <1 child; not using HRT can eliminate 12.5% of cancer cases among users. The most effective preventative factors identified were giving birth at earlier age, having more than one child, reducing the reproductive interval index, and reducing weight.

## 5.6 Discussion

This study explores the effect of anthropometric and reproductive factors on risk of developing BC in the UK Biobank female cohort. The BC incidence rate in the pre-menopause group was 1.55 per 1000 person-years and 2.24 per 1000 person-years in the post-menopause group. McPherson *et al* reported a similar finding that in every 1000 UK women over 50 years old, two females will be diagnosed with BC [9].

Findings from previous studies suggested that differences in risk factors and incidences of BC were based on the menopausal status [110, 127, 298]. Some of the risk factors were common across pre- and post-menopause groups while other factors showed different effects. Therefore, all the analyses were stratified by menopausal status. A detailed critical analysis of our results is provided below with a comparison to the available literature related to BC in the UK.

**Age:** For both pre- and post-menopausal groups, age is associated with increasing risk of developing BC. Age is a well-established risk factor for BC [3]. BC incidence increases with age during the reproductive years by the double in every 10 years up until the menopause [110, 298]. A potential explanation could be cells becoming more susceptible to environmental carcinogens and modification in the biological ageing which stimulates or allows tumour growth and metastasis [299].

**Family history:** Family history of BC is also a well-established risk factor. Our findings suggested that females with a first degree relative (sibling or mother) affected with BC were at high risk of developing BC. Regardless of menopause status, the estimated risks were higher in females who reported only their sibling(s) affected with BC as compared to females who reported only their mother affected with BC. The estimated risks were even higher when both mother and sister were affected with BC. Evidence of family history of BC in the first degree relatives and BC risk has been well documented by many studies with different study designs [9, 300]. The variation of reported estimated risks was due to family history nature such as affected age, number and type of the affected family members [301, 302]. It is known that BRCA1 and BRCA2 gene mutations are responsible for this strong association for cases diagnosed at young age [303, 304]. The stronger effect of family history among pre-menopausal females in this study suggested a component of familial BC [302]. Possible explanations to higher estimated risks observed in subjects with sibling affected include recall bias. With self-reported data, maternal history is more likely to be incomplete as compared to the sibling history. Another possibility is the confounder effect such as parity; mothers of subjects were obviously parous while sisters could be either parous or nulliparous. It is known that parity is a protective factor against BC hence if subject’s sisters were null-parous; one would expect to observe higher risk. Sisters are more likely to share the same or similar environmental factors than mother and a daughter. Finally, multiple family relatives having an early onset or bilateral cancer increases the risk even more [298].

**Deprivation score:** Deprivation score data was available for the dataset. Our result suggested that the most deprived females appeared to have lower BC risk compared to least deprived females in the UK Biobank cohort. Our cohort appeared to be mainly from least deprived districts like Bristol (8.8%), Leeds (8.9%), Newcastle (7.4%), and Nottingham (6.8%). Most deprived districts included Stockport (0.76%), Manchester (2.7%), and Birmingham (4.9) contributed less in this cohort. This sampling distribution could have an effect on the association direction between deprivation and BC.

**Variables related to body size**: Inverse associations were observed with BMI, waist and hip circumference and waist to hip ratio in the pre-menopausal group. While among post-menopausal females, increased risks were reported. A Norwegian prospective study suggested a decreased risk of BC among overweight and obese females who had no family history of BC. Nevertheless once a female has a family history, that protection effect disappeared in both overweight and obese pre-menopausal females [75]. Our results however suggested that risk was reduced even when family history of BC was present among pre-menopausal females. One study reported an estimation of 3% risk increase in BC for every 1 kg/m$^2$ in post-menopausal females [213], while another study reported that weight gains of 5-12 kg increases the post-menopausal BC risk by 50% and modest weight loss (5-10%) can decrease BC risk by 25-40% [214]. Furthermore, overweight and obesity are associated with poor prognosis and increased BC mortality [76].

BMI is a modifiable factor and reducing BMI could contribute to a reduction in the BC risk, the PAF in pre-menopause cancers due to BMI is 10.0% when compared to the reference (BMI more than 25) and the PAF was 5.1% in post menopause women when compared to the reference (BMI less than 25 [305]. Our study confirmed elimination of 8.3% of BC if females reduced their BMI lower than 30 among general population but if obese females (BMI≥30) reduced their BMI to normal BMI range, a 19.4% of BC risk will be eliminated among post-menopaused females. Another way to assess central adiposity among individuals is by measuring WHR (waist to hip ratio). A systematic review on the relationship of WHR and BC concluded that 24% risk reduction was associated with small WHR in post-menopausal females. In contrast among pre-menopausal the effect was very little [306]. Another review suggested the same conclusion; pre-menopausal BC is not associated with WHR however, 1.4 to 5.4 times of BC risk was proven among post-menopausal females [307]. Our study showed BC risk reduction was associated with increased WHR up to 25.6% in pre-menopausal females but failed to prove any association with post-menopausal females. The findings on height and BC risk supported adult height being associated with BC risk in both pre- and post-menopausal groups. The EPIC cohort study [146] reported a positive association between height and post-menopausal BC (RR 1.10 with 95% CI 1.05–1.16). Furthermore, a meta-analysis of 159 prospective studies showed a pooled BC RR of 1.17 (95% CI = 1.15 - 1.19) per 10cm increase in height [147, 148]. Another pooled analysis also suggested positive association among post-menopausal females (RR=1.07 with 95% CI: 1.03, 1.12) [149]. No association was reported in pre-menopausal females (RR 1.02 with 95% CI: 0.96, 1.10). Not all prospective studies confirmed the positive association. A register-based cohort study with 13,572 participants concluded no statistical evidence of association between height and BC risk [150]. Evidence from case-control studies was inconsistent. Our study showed a RR of 1.18 (95% CI = 1.04 - 1.34) per 10cm increase in height among pre-menopausal and a RR of 1.23 (95% CI = 1.14

- 1.33) per 10cm increase in height among post-menopausal. All the results mentioned previously were for standing height; sitting height was examined and found a BC risk association with sitting height. Taller sitting height is associated with 25.5% BC risk increase per each 10 cm increase in pre- (RR= 1.26, 95% CI 1.01-1.57) and 37.0% in post-menopausal (RR= 1.37, 95% CI 1.21-1.57) per 10 cm increase.

The relationship between height and BC suggests a protective effect among females with short stature rather than a continuous increased risk with the increasing of female's height. One possible explanation is that short females would be exposed to lower levels of insulin like growth factor 1 (IGF 1) throughout childhood and adolescence. IGF-1 is considered to be a strong mitogen for BC cells and IGF-1 receptors are expressed in breast tumour tissues 10 folds higher than normal breast tissues [308, 309].

**Reproductive factors**: Our findings suggested protective effect of factors related to childbearing and having more children among pre- and post-menopausal females. Risk factors in pre-menopausal females were early menarche age (less than 13 years old), late age at first live birth (more than 25 years of age), high reproductive interval index, and increased duration of OC used were considered as risk factors for BC in pre-menopausal females. Factors such as nulliparous, high reproductive interval index and increased duration of OC used were risk factors in post-menopausal females.

Increased production of steroid hormone starts around the time of menarche and decreases significantly near the menopause [127] . Hormones produced by the ovary directly affect the breast function and development. Studies showed long period of hormonal exposure increases the risk to develop BC. Late menarche and early menopause are known to be protective factors as the period of hormonal exposure is reduced. Lengthening the reproductive years by an early menarche of one year has a stronger effect than delaying the menopause by one year [127]. The strength of menarche age and menopause age on BC development can be affected by BMI [310, 311]. The association between the BC and menopause age can be weaker among post-menopausal females with high BMI as seen in the meta-analysis [127]. Our results showed an evidence of BC risk reduction by late age of menarche but not by early age the menopause age as the previous studies even when BMI was adjusted for in the analysis. A meta-analysis of 120,000 BC cases and 300,000 controls done by a collaborative research group confirmed the existing association between early menarche and developing risk of BC. Extra risk is associated with lengthening female's reproductive years by one year during menarche rather than lengthening one year at menopause [127]. The RR associated with early menarche was 1.05 (95% CI 1.04–1.06) and the RR associated with late menopause was 1.03(95% CI 1.03–1.03) [127].

Childbearing in a known protective factor against BC although other factors might help confound this protection, such as breast feeding [106]. Combination of both factors can help protect females even more. Unfortunately, there were no data available on breastfeeding in our cohort and unable to assess this effect. In the case of parity, our results showed a significant evidence of risk reduction among both pre- and post- menopausal females with a stronger effect among pre-menopausal. Likewise, as the number of children increases, the protective effect increases. Our results suggested an elimination of 9.2% among pre- and 17.9% among post- BC risk associated with being a parous female while other study reported a lower yet an affective risk reduction of 13.3% for the same factor [312]. As the number of children increases, the attributed risk reduction increases accordingly with reduction of 5.2% among pre- and 5.4% among post-menopausal females [305]. Nevertheless, our results suggested a higher reduction among pre- (8.8%) and a lower reduction percentage among post- menopausal women (4.6%).

Termination of pregnancy, whether induced or natural did not appear to affect the BC risk. Thus, younger age at childbirth is a protective factor against BC and this was observed among pre-menopausal females with p values <0.05. Studies showed early pregnancy causes permanent morphological changes to the breast and makes it more resistant to carcinogenic changes [192]. Our study supported the elimination of 44.6% of BC risk if females in general had their first child in their 20s rather than ≥30 years old among pre-menopaused females. This reduction can reach up to 48.4% among females who had their first child at age of ≥30 if they had their first child in their twenties. Furthermore, the reproductive interval variable (duration between the menarche and first child) was explored and the results supported evidence reported in the literature that as the duration increases the risk also increases. Long term hormonal exposure has been confirmed to be a risk for BC [298]. Our study showed a BC reduction of 14.9% in pre-menopausal women if they have reproductive interval of < 16 and this reduction can reach up to 29.6% if those females with reproductive interval of ≥16 had interval of 12 or less among pre-menopausal females.

Mammogram history suggested borderline significant increased risk in pre-menopausal women and no association in post-menopausal women. The mammogram itself *per se* is not a risk factor for BC but women who reported having had a mammogram were more likely to be diagnosed. Mammogram screening is proved to reduce the BC mortality by 29% among females aged between 50 – 69 years [313].

**Hormone use :** Oral contraceptive use is known to be a risk factor of BC and this risk rises with longer duration of use [314]. It has been proposed that using OC can activate breast tumours which are already present. Oestrogen is recognized as enhancing tumour growth, and with OC and later HRT use these hormones promotes the tumour growth even more [314]. Our findings suggested a positive association between BC and OC duration amongst pre- menopausal females only. Moreover, HRT users showed 14.1% more risk for developing BC among our cohort. Extensive evidence showed an increase in BC incidence in current HRT users and that risk returns to normal soon after use terminates. Combined oestrogen-progesterone therapy revealed higher risk compared to oestrogen only preparations including results from the Women Health Initiative study (WHI). Recent results from WHI found both oestrogen only and combined formulations convey greater risk for BC if the females started the HRT in less than 5 years after the menopause compared to longer gap [82, 310, 315-317]. The study also carried out further analysis of HRT. Their results showed attenuated BC risk among obese females which is driven by hormonal adiposity of the breast. Endogenous oestrogen rises with the increase of the BMI among HRT non-users which increases the breast adiposity [310]. Another major study carried out in the UK (Million Women Study) identified that BC risk is associated with current use of HRT and the risk is considerably greater among combined oestrogen- progesterone users than other types of HRT [83]. According to our analysis stopping HRT can reduce the risk by 5.8% and by 12.5% risk among HRT users. The Million Woman Study estimated this figure to be 4.6% [312] and a more recent study has put this figure higher at 14.5% [305].

A summary for the significant factors associated with development of BC among UK females is presented in Table *5.7*.

Table 5.7: Summary of the significant factors associated with breast cancer among both pre- and post-menopausal females in the UK

| **Variable** | **Pre-menopausal (effect size) *** | **Post-menopausal (effect size) *** | **Conclusion** | **Modifiable** |
|---|---|---|---|---|
| **Non-modifiable** | | | | |
| Age (continuous) | Risk (5%) | Risk (3%) | Getting older – more risk | No |
| BC family history (categorical) | Risk (77%) | Risk (58%) | Family history – more risk | No |
| Deprivation score (continuous) | Protective (3.8%) | Protective (2.7%) | More deprived – less risk | Yes |
| Sitting Height in cm (continuous) | Risk (2.3%) | Risk (3.2%) | Taller - more risk | No |
| Standing Height in cm (continuous) | Risk (1.7%) | Risk (2.1%) | Taller - more risk | No |
| Standing Height in cm (categorical) | Risk (42.9%) | Risk (53.3%) | Taller - more risk | No |
| Menarche age in years (continuous) | Protective (5.2%) | - | Older – less risk | No |
| Menarche age (categorical) | Protective (22.8%) | - | Older – less risk | No |
| **Modifiable** | | | | |
| BMI (continuous) | Protective (1.7%) | Risk (1.8%) | High BMI – less risk in pre- and more risk in post | Yes |
| BMI (categorical) | Protective (26.7%) | Risk (24.1%) | High BMI – less risk in pre- and more risk in post | Yes |

| Variable | Pre-menopausal (effect size) * | Post-menopausal (effect size) * | Conclusion | Modifiable |
|---|---|---|---|---|
| Waist Circumference in cm (continuous) | Protective (0.8%) | Risk (0.8%) | High waist circumference – less risk in pre- and more risk in post | Yes |
| Hip Circumference in cm (continuous) | - | Risk (1.2%) | High hip circumference – more risk in post | Yes |
| Waist to Hip (continuous) | Protective (86.9%) | - | High ratio – less risk in pre | Yes |
| Waist to Hip (categorical) | Protective (25.6%) | - | High ratio – less risk in pre | Yes |
| Contraceptive duration in years (continuous) | Risk (2.4%) | - | Larger interval – more risk in pre | Yes |
| HRT use (categorical) | - | Risk (14.1%) | HRT use – more risk I post | Yes |
| **Partially modifiable** | | | | |
| Parity (categorical) | Protective (23.6%) | Protective (17.9%) | More children – less risk | Yes |
| Number of births (continuous) | Protective (7.5%) | Protective (10.1%) | More children – less risk | Yes |
| First live birth age (categorical) | Risk (93.8%) | - | Older – more risk | Yes |
| Reproductive interval index in years (continuous) | Risk (0.3%) | Risk (0.3%) | Larger interval – more risk | Yes |
| Reproductive interval index (categorical) | Risk (53%) | Risk (33%) | Larger interval – more risk | Yes |

*****risk increases or decreases for each unit change in these variables

In conclusion, an analysis was carried out to confirm risk and protective factors and BC risk in the UK Biobank female cohort. The findings suggest that protective factors in women included reducing BMI, waist circumference, waist to hip ratio, increasing the numbers of births, having birth at an early age, minimising the use of oral contraceptive and HRT and their durations. Most of our findings are in keeping with evidence reported from the other UK large cohort studies such as the One Million Women and EPIC studies. Evidence from this large study can be further used in translational research such as prevention programmes. Our study has some strengths and limitations. The strengths of this study are large nation-wide prospective population-based cohort with a follow up time of 9 years and a sizable number of incident cases (UK Biobank). Furthermore, to our knowledge, this is the first study investigating the effect of the anthropometric and reproductive factors with BC risk among the UK Biobank female cohort. The results of this study can be used to inform BC prevention strategies and be used to educate the public and form a basis for building risk prediction models for BC for the UK population. Additionally, reproductive interval index is a new measure and only reported by our study using UK data. Estimation of d the general PAF and the PAF of the subgroups for BC in the UK Biobank female cohort is novel. The attributable risks calculated for the modifiable factors can be translated into action to reduce BC incidence.

A possible weakness is the lack of some information such as breastfeeding history, ovarian cancer family history, BC onset of the family members and BC type. As this is a volunteer-based study, the females who participated in the UK Biobank study are likely to be better educated and less deprived compared to the whole population.

**Acknowledgment**

We would like to express our appreciation to UK biobank participants and staffs for providing the research community with this valuable source of data. This work used data from the application number 5791. We are also greatly thankful to Ministry of higher education in Oman for their financial support to the first. We would like to acknowledge funding for ICEP ("This work was also supported by CRUK [grant number C18281/A19169]")

**S**1 Table: Codes used to identify breast cancer cases and controls

| **Categories** | **Frequency (%)** | **ICD10 codes** | **ICD9 codes** | **Self-reported cancer’s codes** | **Self-reported non-cancer diseases** |
|---|---|---|---|---|---|
| **Breast cancer cases** | | | | | |
| Incident: | 3,378 (1.24%) | Codes start with C50 and its subclasses , C501, C502, C503, C504, C505, C506, C507, C508, and C509 | Codes start with 174 and its subclasses 1741, 1742, 1743, 1744, 1745, 1746, 1747, 1748, and 1749 | 1002 code only | - |
| Prevalent: | 10,853 (3.97%) | | | | |
| **Subjects excluded from the study** | | | | | |
| 10- Other cancers | 23,540 (8.61%) | Codes start with C except codes for BC | Codes start with 1 or 20 except codes for BC | All other codes except 1002 code | - |
| 11- Breast In situ carcinoma | 636 (0.23%) | Codes of D050, D051, D057, D059 | 2330 code only | - | - |
| 12- Other in situ carcinoma | 2,463 (0.90%) | Codes start with D0 except codes for breast in situ carcinoma | Codes start with 230 or 231 or 232 or 233 or 234 except codes for breast in situ carcinoma | - | - |
| 13- Neoplasm of unknown nature or behavior | 121 (0.04%) | Codes start with D37 or D38 or D39 or D40 or D41 or D42 or D43 or D44 or D45 or D46 or D47 or D48 | Codes start with 235 or 236 or 237 or 238 or 239 | - | - |
| **Controls** | | | | | |
| All controls | 232,476 (85.01%) | Remaining codes or subjects with no code assigned | Remaining codes or subjects with no code assigned | - | All of Self-reported of non-cancerous diseases or no code assigned |
| **Total** | 273,476 (100%) | | | | |

S2 Table: Classification of the variables included in the analysis

| | Variable | Groups | Coding |
|---|---|---|---|
| 1 | Menopausal status | Pre-menopausal | Pre- menopausal: reported as pre-menopausal & no history of hysterectomy or bilateral oophorectomy & their age is ≤55 and menarche age ≥7 years old (to maximise the number of the real pre-menopausal females - so any female reported as pre- and their age > 55 or had menarche age < 7 and did not had hysterectomy nor oophorectomy will be removed – most probably this female is miscategorised). |
| | | Post-menopausal | Post-menopausal: reported as post-menopausal & no history of hysterectomy or bilateral oophorectomy (only natural menopause) & their menopause age is ≥ 40years old ( to maximise the number of the real post-menopausal females – so any female reported as post- and their menopause age < 40 and did not had hysterectomy nor oophorectomy will be removed - most probably this female is miscategorised). |
| 2 | Menarche age | (>13) | This variable was divided into two groups based on literature ranges [127]. |
| | | (≤13)) | |
| 3 | Age at first birth | (<20) | This variable was divided into four groups based on literature ranges [192]. |
| | | (20-24) | |
| | | (25-29) | |
| | | (≥30) | |
| 4 | BMI | Healthy (18.5 – 24.9) | This variable was divided into three groups based on WHO classification [193]. Underweight group were very low in number and would not be enough for the association calculations. |
| | | Overweight (25-29.9) | |
| | | Obese (≥30) | |
| 5 | Waist to hip ratio (WHR) | Low (≤0.80) | This variable was calculated by dividing the waist over the hip measurements of the participants. Then later was divided into three groups based on WHO classification [194]. |
| | | Moderate (0.81-0.85) | |
| | | High (>0.85) | |
| 6 | Reproductive interval index- years | Low (≤12) | Reproductive interval index is the difference between the age at first birth and age at menarche. The index was divided into four groups based on the IQR (InterQuartile Range) of the reproductive interval index values among the controls only. |
| | | Moderate (12-16) | |
| | | High (>16) | |
| | | No children | |
| 7 | Deprivation score | Calculated by UK biobank team | The score was calculated for all the participants before participating in UK Biobank. The score was based on the prior national census output areas [195]. The score evaluates four aspects: 1) unemployment, 2) houses without an owned car, 3) non-house ownership, 4) overcrowding in one house [196]. Data were provided by the UK biobank team. |
| 8 | Height | Below mean (< 156.10 cm) | The height was grouped into three groups based on the mean of the control group. |
| | | Within mean ± SD (156.10-168.75cm) | |
| | | Above mean (>168.75 cm) | |

S3 Table: Summary of the significant factors associated with breast cancer among both pre- and post-menopausal females in the UK

| Variable | Pre-menopausal (effect size) * | Post-menopausal (effect size) * | Conclusion | Modifiable |
|---|---|---|---|---|
| **Non-modifiable** | | | | |
| Age (continuous) | Risk (5%) | Risk (3%) | Getting older – more risk | No |
| BC family history (categorical) | Risk (77%) | Risk (58%) | Family history – more risk | No |
| Deprivation score (continuous) | Protective (3.8%) | Protective (2.7%) | More deprived – less risk | Yes |
| Sitting Height in cm (continuous) | Risk (2.3%) | Risk (3.2%) | Taller - more risk | No |
| Standing Height in cm (continuous) | Risk (1.7%) | Risk (2.1%) | Taller - more risk | No |
| Standing Height in cm (categorical) | Risk (42.9%) | Risk (53.3%) | Taller - more risk | No |
| Menarche age in years (continuous) | Protective (5.2%) | - | Older – less risk | No |
| Menarche age (categorical) | Protective (22.8%) | - | Older – less risk | No |
| **Modifiable** | | | | |
| BMI (continuous) | Protective (1.7%) | Risk (1.8%) | High BMI – less risk in pre- and more risk in post | Yes |
| BMI (categorical) | Protective (26.7%) | Risk (24.1%) | High BMI – less risk in pre- and more risk in post | Yes |
| Waist to Hip (continuous) | Protective (86.9%) | - | High ratio – less risk in pre | Yes |
| Waist to Hip (categorical) | Protective (25.6%) | - | High ratio – less risk in pre | Yes |
| Contraceptive duration in years (continuous) | Risk (2.4%) | - | Larger interval – more risk in pre | Yes |
| HRT use (categorical) | - | Risk (14.1%) | HRT use – more risk I post | Yes |
| **Partially modifiable** | | | | |
| Parity (categorical) | Protective (23.6%) | Protective (17.9%) | More children – less risk | Yes |
| Number of births (continuous) | Protective (7.5%) | Protective (10.1%) | More children – less risk | Yes |
| First live birth age (categorical) | Risk (93.8%) | - | Older – more risk | Yes |
| Reproductive interval index in years (continuous) | Risk (0.3%) | Risk (0.3%) | Larger interval – more risk | Yes |
| Reproductive interval index (categorical) | Risk (53%) | Risk (33%) | Larger interval – more risk | Yes |

# Chapter 6：Association of non-genetic factors with breast cancer risk in genetically predisposed groups of women in the UK Biobank cohort

Publication number 3

## Association of non-genetic factors with breast cancer risk in genetically predisposed groups of women in the UK Biobank cohort

This chapter is presented as a journal article:

**Al Ajmi K,** Lophatananon A, Mekli K, Ollier W, Muir KR. Association of Nongenetic Factors with Breast Cancer Risk in Genetically Predisposed Groups of Women in the UK Biobank Cohort. JAMA Netw Open. 2020;3(4): e203760.

doi:10.1001/jamanetworkopen.2020.3760

## Association of non-genetic factors with breast cancer risk in genetically predisposed groups of women in the UK Biobank cohort

Kawthar Al Ajmi[1] ·Artitaya Lophatananon[1] Krisztina Mekli[2] William Ollier[1,3] Kenneth R. Muir[1]

[1] Division of Population Health, Health Services Research and Primary Care, Faculty of Biology, Medicine and Health, Centre for Epidemiology, The University of Manchester, Manchester M139 PL, UK
[2] Cathie Marsh Institute for Social Research, School of Social Sciences, Faculty of Humanities, The University of Manchester, Manchester M139 PL, UK
[3] School of Healthcare Science, Faculty of Science and Engineering, Manchester Metropolitan University, John Dalton Building Manchester M1 5GD UK

### 6.1 Abstract

**Importance:** The association between non-inherited factors, including lifestyle factors, and the risk of breast cancer (BC) in women and the association between BC and genetic makeup are only partly characterized. A study using data on current genetic stratification may help in the characterization.

**Objective:** To examine the association between healthier lifestyle habits and BC risk in genetically predisposed groups.

**Design, setting, and participants:** Data from UK Biobank, a prospective cohort comprising 2728 patients with BC and 88 489 women without BC, were analysed. The data set used for the analysis was closed on March 31, 2019. The analysis was restricted to postmenopausal white women. Classification of healthy lifestyle was based on Cancer Research UK guidance (healthy weight, regular exercise, no use of hormone replacement therapy for more than 5 years, no oral contraceptive hormones use, and alcohol intake frequency <3 times/wk). Three groups were established: favourable (4 healthy factors), intermediate (2-3 healthy factors), and unfavourable (1 healthy factor). The genetic contribution was estimated using the polygenic risk scores of 305 preselected single-nucleotide variations. Polygenic risk scores were categorized into 3 tertiles (low, intermediate, and high).

**Main outcomes and measures**: Cox proportional hazards regression was used to assess the hazard ratios (HRs) of the lifestyles and polygenic risk scores associated with a malignant neoplasm of the breast.

**Results:** Mean (SD) age of the 2728 women with BC was 60.1 (5.5) years, and mean age of the 88 489 women serving as controls was 59.4 (4.9) years. The median follow-up time for the cohort was 10 years (maximum 13 years) (interquartile range, 9.44-10.82 years). Women with BC had a higher body mass index (relative risk [RR], 1.14; 95% CI, 1.05-1.23), performed less exercise (RR, 1.12; 95% CI, 1.01-1.25), used hormonal replacement therapy for longer than 5 years (RR, 1.23; 95% CI, 1.13-1.34), used more oral contraceptives (RR, 1.02; 95% CI, 0.93-1.12), and had greater alcohol intake (RR, 1.11; 95% CI, 1.03-1.19) compared with the controls. Overall, 20 657 women (23.3%) followed a favourable lifestyle, 60 195 women (68.0%) followed an intermediate lifestyle, and 7637 women (8.6%) followed an unfavourable lifestyle. The RR of the highest genetic risk group was 2.55 (95% CI, 2.28-2.84), and the RR of the most unfavourable lifestyle category was 1.44 (95% CI, 1.25-1.65). The association of lifestyle and BC within genetic subgroups showed lower HRs among women following a favourable lifestyle compared with intermediate and unfavourable lifestyles among all of the genetic groups: women with an unfavourable lifestyle had a higher risk of BC in the low genetic group (HR, 1.63; 95% CI, 1.13-2.34), intermediate genetic group (HR, 1.94; 95% CI, 1.46-2.58), and high genetic group (HR, 1.39; 95% CI, 1.11-1.74) compared with the reference group of favourable lifestyle. Intermediate lifestyle was also associated with a higher risk of BC among the low genetic group (HR, 1.40; 95% CI, 1.09-1.80) and the intermediate genetic group (HR, 1.37; 95% CI, 1.12-1.68).

**Conclusions and relevance:** In this cohort study of data on women in the UK Biobank, a healthier lifestyle with more exercise, healthy weight, low alcohol intake, no oral contraceptive use, and no or limited hormonal replacement therapy use appeared to be

associated with a reduced level of risk for BC, even if the women were at higher genetic risk for BC.

## 6.2 Introduction

Breast cancer (BC) is the most common cancer in women as well as the second most common cause of cancer-related death in women [9, 318]. In the UK it is estimated that more than 55,000 new cases of BC occur annually [318]. Both genetic and lifestyle factors play crucial roles in the complex mechanism of BC. Evidence supporting the genetic component of BC is seen with highly penetrant rare gene variants, such as in the BRCA1 and BRCA2 genes.

These particular variants ad other rare variants, however, account for just a small proportion (<5%) of overall BC cases and for 15% to 20% in familial BC cases [319]. Genome-wide association studies have identified number of single-nucleotide variations (SNVs) associated with risk for BC development, although these SNVs individually contribute only a small genetic proportion or are in genes exhibiting medium to low penetrance. The cumulative genetic contribution of all known and unknown BC associated genetic variants is known as the additive genetic effect. A polygenic (PRS) is a construct which combines known genetic variants for a trait and can contribute to estimating the risk of a trait. Genetic data availability allowed us to estimate PRS in a substantial proportion of all BC cases (88%) [120, 320, 321]. The application of genetic risk stratification to individuals as a clinical tool for aiding BC screening is now on the horizon [320]. Mavaddat et al [322] showed that women at the top 5% of the PRS can develop BC at age 37 years, while those in the lowest 20% of the PRS will likely never develop BC.

Some lifestyle and behavioural factors can play an important role in and contribute to the risk of BC [83, 315, 323-326]. Our Study investigating BC risk changes by adhering to healthier lifestyle even among predisposed females. This approach is not yet been investigated yet by any research group. However, two studies [180, 181] investigated the effect of adhering to healthier lifestyle in women exhibiting different PRSs with dementia and coronary heart disease. Both concluded that following healthier lifestyle is associated with lower risk of the studied outcome even if the subject had high PRS.

Whereas inherited genetic risk for disease is not modifiable, this factor is not the case for most known nongenetic risk factors. The central hypothesis examined in this study is that, regardless of a person's PRS, overall BC risk can be reduced by following a favourable lifestyle.

### 6.3 Methods

Data from women within the UK Biobank longitudinal cohort study were used. The data set for the analysis was closed on March 31, 2019. The UK Biobank is a national cohort including 502 650 men and women aged between 39 and 71 years. Cases were enrolled between 2006 and 2010 and continue to be longitudinally followed up for capture of subsequent health events. Participants gave the UK Biobank written informed consent to use their data and samples for health-related research purposes. Ethics approval for use of UK Biobank data was obtained from the North West-Haydock research ethics committee. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline for cohort studies.

In this analysis, the inclusion criteria to select study participants were (1) British women who were white (age, 40-71 years), (2) postmenopausal women who did not report a history of hysterectomy or bilateral oophorectomy and reported no longer menstruating, and (3) women with a menopause age of 40 years or older. Deceased participants were excluded from our analysis. Of the UK Biobank cohort of 273 402 female participants, 114 723 women (42.0%) fulfilled our inclusion criteria.

The study outcome was defined as women with a malignant neoplasm of the breast. Cases and controls were identified according to the criteria summarized in Figure 1 in the Supplement. Three coding systems was used to identify cases  with BC and those serving as controls: International Statistical Classification of Diseases and Related Health Problems, Tenth Revision; International Classification of Diseases, Ninth Revision; and self-reported (eTable 1 in the Supplement). If cases with breast cancer appeared to have an incident case of BC according to any of these 3 coding systems, they were deemed incident cases (age at cancer diagnosis was older than age when they attended the assessment center of the UK Biobank study). Cases were considered prevalent only if they were defined as such according to any of the 3 coding systems, which was applicable only if none of the 3 approaches had described the BC case as being an incident case. A total of 2728 postmenopausal women with incident cases of BC were eligible for the analysis. Controls were defined as women without a history of any cancer, carcinoma in situ, or unknown neoplasm. The final number of controls selected by menopausal status and our set criteria was 88 489. Figure 1 in the Supplement illustrates the number of study participants in the case and control selection process.

Cancer Research UK [179] has reported risk factors for BC development as being either modifiable or non-modifiable. Based on their list, 5 modifiable factors were identified: weight, alcohol intake, physical activity, oral contraceptive use, and hormonal replacement therapy (HRT) intake for more than 5 years. A scoring system based on the presence or absence of these 5 factors was developed to derive favourable lifestyle, intermediate lifestyle, and unfavourable lifestyle. This approach was also used in other studies for example coronary heart disease study [180] and dementia study [181]. I applied equal weight to all factors and assigned a score based on response. The total score was summated. The details of the 5 factors and score definitions are presented in Table 6.1. Eligible participants were stratified into 3 categories: favourable lifestyle (4 healthy factors present), intermediate lifestyle (2 or 3 healthy factors present), and unfavourable lifestyle (1 healthy factor present).

A PRS was computed based on the Mavaddat score [120] using the UK Biobank high-density genome-wide SNV data set available for 488 377 of their participants. The SNV data were used from individuals who were included on the basis of being female (matched genetic and self-reported sex) and their genetic ethnic grouping (white). During the quality control process, individuals with missingness (>2%), outliers for heterozygosity by removing individuals who deviate ±3 SD from the samples' heterozygosity rate mean based on guidance [186], and duplicates, as well as those who were biologically related, were excluded.

The PRS for BC was constructed using the 313 SNVs previously determined to contribute some risk by the hard threshold approach used by Mavaddat et al [120], Of these 313 SNVs, 306 were present in the UK Biobank data set; however, SNV rs10764337 was triallelic and excluded. The final number of SNVs used for PRS construction was therefore 305, and their details are presented in Table 2 in the Supplement. Forty of 305 SNVs had been directly genotyped and successfully passed the marker test applied by UK Biobank. The remaining 265 SNVs had been imputed. The quality of the imputation was estimated using the information scores available, which is a number between 0 and 1 where 0 indicates complete uncertainty and 1 indicates complete certainty. The lowest information score was 0.86. Linkage disequilibrium was assessed, and no r2 value between any 2 SNVs reached 0.9. Plink open source software version 1.90 was used to carry out the quality control processes [188].

Individual participant PRS was created by adding the number of risk alleles at each SNV and then multiplying the sum by the effect size as the previously published effect size [120]. The raw PRS was standardised by dividing each raw PRS by the SD of the PRS derived from the control group. No transformation to the PRS data was applied because the scores were normally distributed (Figure 2 in the Supplement). A tertile genetic risk classification using standardised PRS values from controls was generated. Each participant was then assigned to a genetic risk group: low (1st tertile up to 33.33%), intermediate (2nd tertile between 33.34% and 66.67%), and high (3rd tertile from 66.68% to 100%).

Relative risks (RRs) and 95% CIs of the basic risk factors were computed with an adjustment for age and family history using a binomial generalized linear regression model. Cox proportional hazards regression was applied to assess the hazard ratios (HRs) of the lifestyles and BC risk. First HRs for each genetic stratum was computed with the low genetic risk group as a reference group and for each lifestyle (favourable, intermediate, and unfavourable) stratum with the favourable category as a reference group. The HRs in each lifestyle stratum were calculated within each genetic risk group. All analyses were adjusted for age and family history. The Cox proportional hazards regression model assumption for each analysis was tested. A 2-sided P value <.05 was considered significant. The *Ltable* [182] command was used to compute a 10-year cumulative BC incidence for each lifestyle category within each genetic risk stratum. Results presented in graphic bar charts were generated using Microsoft Excel 2016 (Microsoft Corp) [183]. All analyses were performed using Stata/MP software version 14 (StataCorp LLC) [184].

Table 6.1: Criteria for healthy lifestyle classification

| Healthy lifestyles criteria | UK Biobank cohort | Codes |
|---|---|---|
| Healthy weight | Healthy: BMI <25 kg/m2<br>Unhealthy: BMI ≥ 25 kg/m2 | Healthy: 1<br>Unhealthy: 0 |
| Regular physical activity | Healthy: At least ≥ once per week<br>Unhealthy: No physical activity at all | Healthy: 1<br>Unhealthy: 0 |
| No/limited alcohol intake | Healthy: No alcohol intake or used for < three times/week<br>Unhealthy: Used alcohol ≥ three times /week | Healthy: 1<br>Unhealthy: 0 |
| No contraceptive intake | Healthy: No OC use<br>Unhealthy: Used OC | Healthy: 1<br>Unhealthy: 0 |
| No/limited HRT intake | Healthy: No HRT use or used HRT < 5 years<br>Unhealthy: Used HRT for ≥ 5years | Healthy: 1<br>Unhealthy: 0 |
| Classifications | | |
| Favourable lifestyle | Presence of 4-5 healthy lifestyle factors | Sum: At least 4 |
| Intermediate lifestyle | Presence of 2-3 healthy lifestyle factors | Sum: 2 or 3 |
| Unfavourable lifestyle | Presence of only one healthy lifestyle factor or none | Sum: 1 or 0 |

### 6.3.1 Statistical analysis

Relative risks (RR) and 95% confident intervals (95% CI) of the basic risk factors were computed with an adjustment for age and family history using a binomial generalised linear regression model. Cox's proportional-hazard regression was applied to assess the hazard ratios (HRs) of the lifestyles and breast cancer risk. First HRs for each genetic stratum was computed with the low genetic risk group as a reference group and for each of the lifestyles (favourable, intermediate, and unfavourable) stratum with the favourable category as a reference group. The HRs in each lifestyle stratum was calculated within each genetic risk group. All analyses were adjusted for age and family history. The proportional hazard assumption for each analysis was tested. *Ltable* [182] command was used to compute a 10-year cumulative BC incidence for each lifestyle category within each genetic risk stratum. Results presented in graphic bar charts were generating using Microsoft Excel [183]. All analyses were performed using Stata 14 MP software [184].

## 6.4 Results

The median follow-up time for the cohort was 10 years (maximum, 13 years) (interquartile range, 9.44-10.82 years). The total number of the incident BC cases was 2728, and the total number of controls was 88 489. The mean (SD) age of the cases was 60.1 (5.5) years and for controls was 59.4 (4.9) years. The mean (SD) body mass index (BMI) measures (calculated as weight in kilograms divided by height in meters squared) were 27.3 (5.0) for cases and 26.9 (4.9) for controls. In addition, cases used more HRT (30.4%) compared with controls (25.2%). Furthermore, women with BC more often reported no regular physical activity (13.3%) compared with controls (12.0%).

Table 6.2 presents the distribution of the general characteristics and estimated RR results. A

1-year increase in age was associated with a 2.3% increase in BC development risk. Having 1 female first-degree family member (either mother or sister) with BC was associated with a 48.6% increase in BC risk, while having both mother and sister affected was associated with a doubling of the risk of BC compared with women without a family history of BC. An unhealthy weight (BMI ≥ 25) was associated with a 13.9% increased risk of BC (RR, 1.14; 95% CI, 1.05-1.23). Participants who reported that they did not have regular physical activity were had a 12.2% increased risk of BC (RR, 1.12; 95% CI, 1.01-1.25), and alcohol intake frequency 3 or more times per week was associated with an increased BC risk of 10.7% (RR, 1.11; 95% CI, 1.03-1.19). Use of HRT for 5 or more years was associated with an increased BC risk of 22.9% (RR, 1.23; 95% CI, 1.13-1.34). History of oral contraceptive use did not show any association with BC risk among women in the UK Biobank (RR, 1.02; 95% CI, 0.93-1.12); however, this factor was retained as part of lifestyle classification. Overall, 20 657 women (23.3%) followed a favourable lifestyle, 60 195 women (68.0%) followed an intermediate lifestyle, and 7637 women (8.6%) followed an unfavourable lifestyle. Intermediate and unfavourable lifestyles were both associated with higher risk of BC compared with the favourable lifestyle (intermediate: RR, 1.25; 95% CI, 1.13-1.37; unfavourable: RR, 1.44; 95% CI, 1.25-1.65).

The mean standardised PRS of the cases was 26.26 (range, 21.63-29.40), which is higher than the mean standardised PRS of the control group (25.807; range, 21.119-29.941). This difference was examined using a t test, and a significant difference between the mean score was apparent between cases and controls (P < .001). Moreover, the estimated HR for overall BC among postmenopausal women per unit increase in PRS was 1.55 (95% CI, 1.48-1.61). Analysis of the PRS tertile groups indicated a gradient of increased BC risk across tertiles (for second tertile vs first tertile, P < .001; for third tertile vs first tertile, P < .001). Women in the higher genetic risk group (3rd tertile) were at significantly higher risk of BC (RR, 2.55; 95% CI, 2.28-2.84) compared with women in the low genetic risk group after adjusting for age and family history. Similarly, women in the intermediate risk group showed a moderate increased risk (RR, 1.49; 95% CI, 1.32-1.68) compared with those in the low genetic risk group.

Table 6.2: Relative Risks RR for basic characteristics, lifestyles and genetic categories

| **Risk factors** | **Frequency (%)** | | **RR** | **95% LCL** | **95% UCL** |
|---|---|---|---|---|---|
| | **Cases** | **Controls** | | | |
| **Age *** | 2,728 (2.99%) | 88,489 (97.01) | 1.023 | 1.016 | 1.030 |
| **Family history **** | | | | | |
| No family history | 2,276 (83.80) | 78,408 (88.84) | Ref | | - |
| Mother or Sister BC history | 412 (15.17) | 9,405 (10.66) | 1.486 | 1.341 | 1.647 |
| Mother and Sister BC history | 28 (1.03) | 440 (0.50) | 2.099 | 1.462 | 3.012 |

| Risk factors | Frequency (%) | | RR | 95% LCL | 95% UCL |
|---|---|---|---|---|---|
| | Cases | Controls | | | |
| **Weight** | | | | | |
| Healthy | 995 (36.55) | 35,537 (40.25) | Ref | | |
| Unhealthy | 1,727 (63.45) | 52,749 (59.75) | 1.139 | 1.054 | 1.230 |
| **Regular physical activity** | | | | | |
| At least once a week | 2,329 (86.74) | 76,466 (88.00) | Ref | | |
| No physical activity | 356 (13.26) | 10,423 (12.00) | 1.122 | 1.005 | 1.252 |
| **Alcohol intake frequency** | | | | | |
| No intake or < three times a week | 1,566 (57.40) | 52,892 (59.80) | Ref | | |
| Used alcohol ≥ three times a week | 1,162 (42.60) | 35,557 (40.20) | 1.107 | 1.027 | 1.193 |
| **Contraceptive intake** | | | | | |
| No | 561 (20.58) | 17,240 (19.50) | Ref | | |
| Yes | 2,165 (79.42) | 71,149 (80.50) | 1.017 | 0.925 | 1.119 |
| **HRT intake** | | | | | |
| No | 1,895 (69.64) | 66,093 (74.82) | Ref | | |
| Yes | 826 (30.36) | 22,244 (25.18) | 1.229 | 1.131 | 1.335 |
| **Healthy lifestyle score** | | | | | |
| Favourable lifestyle | 530 (19.43) | 20,657 (23.34) | Ref | | |
| Intermediate lifestyle | 1,909 (69.98) | 60,195 (68.03) | 1.245 | 1.132 | 1.369 |
| Unfavourable lifestyle | 289 (10.59) | 7,637 (8.63) | 1.436 | 1.246 | 1.654 |
| **PRS as a category** | | | | | |
| Low | 440 (19.67) | 24,297 (33.70) | Ref | | |
| Intermediate | 655 (29.28) | 23,983 (33.27) | 1.486 | 1.318 | 1.675 |
| High | 1,142 (51.05) | 23,814 (33.03) | 2.545 | 2.282 | 2.838 |

*No adjustment, **Adjusted for age only

Results of estimated HRs for lifestyle and BC risk in each genetic risk group are presented in Table 6.3. The results of Cox proportional hazards regression model assumption testing in the low, intermediate, and high genetic risk groups suggested no statistically significant violation of Cox proportional hazards regression model assumption. In the low genetic risk group, significantly increased HRs were observed in both the unfavourable lifestyle (HR, 1.63; 95% CI, 1.13-2.34) and intermediate lifestyle (HR, 1.40; 95% CI, 1.09-1.80) groups compared with the favourable lifestyle group. In the intermediate genetic risk group, significantly increased HRs were shown in the unfavourable (HR, 1.94; 95% CI, 1.46-2.58) and intermediate (HR, 1.37; 95% CI, 1.12-1.68) lifestyle groups. In the higher genetic risk strata, a significant HR was observed in the unfavourable lifestyle group (HR, 1.39; 95% CI, 1.11-1.74) compared with favourable lifestyle. All of the above results suggest that, within the same genetic risk group, adhering to a less healthy lifestyle (intermediate and unfavourable lifestyle) is associated with an increased risk of BC. Figure 1 shows a forest plot of HRs according to genetic risk group and lifestyle categories.

The results of the 10-year cumulative incidence rate of BC in all genetic risk groups suggest incremental rates of increase from favourable to intermediate to unfavourable (Figure 6.2) lifestyle. A favourable lifestyle had the lowest 10-year cumulative BC incidence rate across all genetic risk groups (low, 3%; intermediate, 5%; and high, 9%). Similar findings in the 10-year cumulative BC incidence rate were observed for an unfavourable lifestyle across the genetic risk groups (low, 5%; intermediate,
9%; and high, 12%).

Table 6.3: Hazard ratios of breast cancer based on lifestyles stratified by the genetic risk group

| **Genetic risk group** | **Healthy lifestyle score*** | **Frequency (%)** | | **HR** | **95% LCL** | **95% UCL** |
|---|---|---|---|---|---|---|
| | | **Cases** | **Controls** | | | |
| **Low genetic risk group** | Favorable lifestyle | 75 (17.05) | 5,550 (22.84) | Ref | | |
| | Intermediate lifestyle | 317 (72.05) | 16,540 (68.07) | 1.401 | 1.090 | 1.802 |
| | Unfavorable lifestyle | 48 (10.91) | 2,204 (9.08) | 1.630 | 1.135 | 2.342 |
| | PH assumption P- value | 0.989 | | | | |
| | P for trend | 0.004 | | | | |
| **Intermediate genetic risk** | Favorable lifestyle | 117 (17.86) | 5,582 (23.27) | Ref | | |
| | Intermediate lifestyle | 458 (69.92) | 16,336 (68.11) | 1.372 | 1.119 | 1.682 |
| | Unfavorable lifestyle | 80 (12.21) | 2,065 (8.61) | 1.945 | 1.463 | 2.587 |
| | PH assumption P- value | 0.084 | | | | |
| | P for trend | 0.000 | | | | |
| **High genetic risk** | Favorable lifestyle | 236 (20.67) | 5,571 (23.39) | Ref | | |
| | Intermediate lifestyle | 792 (69.35) | 16,278 (68.35) | 1.130 | 0.977 | 1.307 |
| | Unfavorable lifestyle | 114 (9.98) | 1,965 (8.25) | 1.391 | 1.112 | 1.740 |
| | PH assumption P- value | 0.693 | | | | |
| | P for trend | 0.007 | | | | |

*Adjusted for age and family history of BC

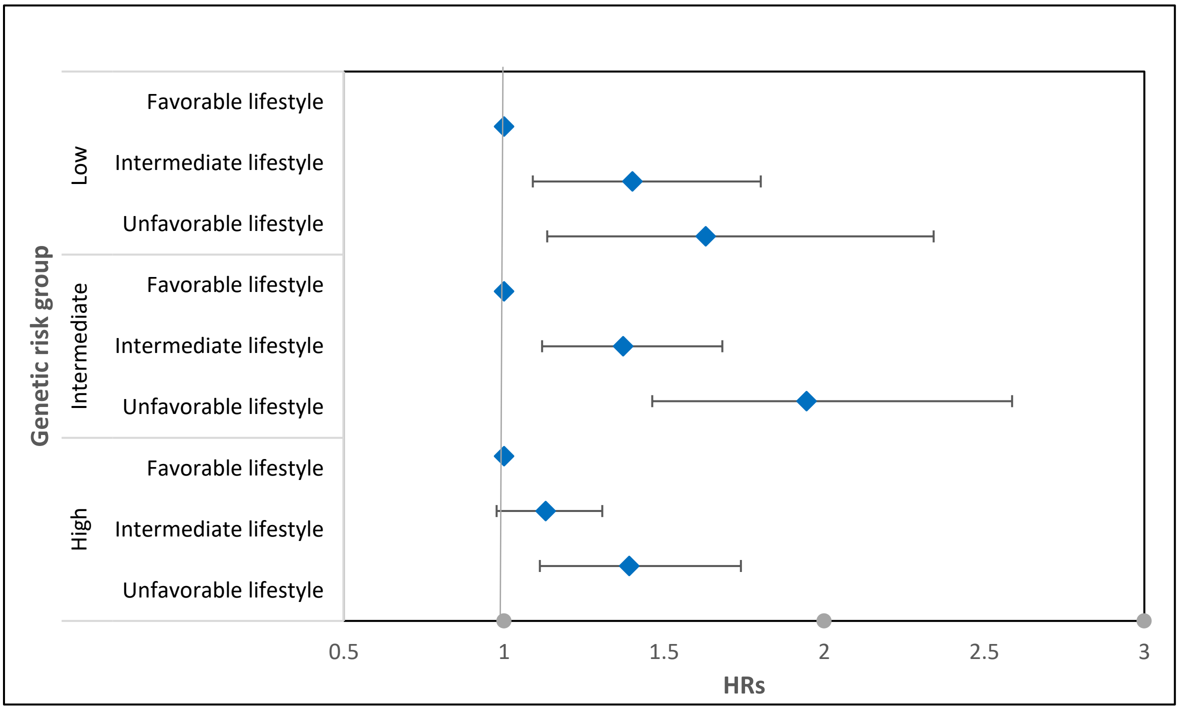


Figure 6.1: Error plot of the HR and 95% CI of breast cancer based on the lifestyle and genetic factors. The HR of each genetic group was stratified based on the three lifestyles (favourable, intermediate, and unfavourable). With favourable lifestyle as the reference group in the three genetic groups.
Legend**: ⌂:** Reference, ⌂: Not significant, ⌂: Significant

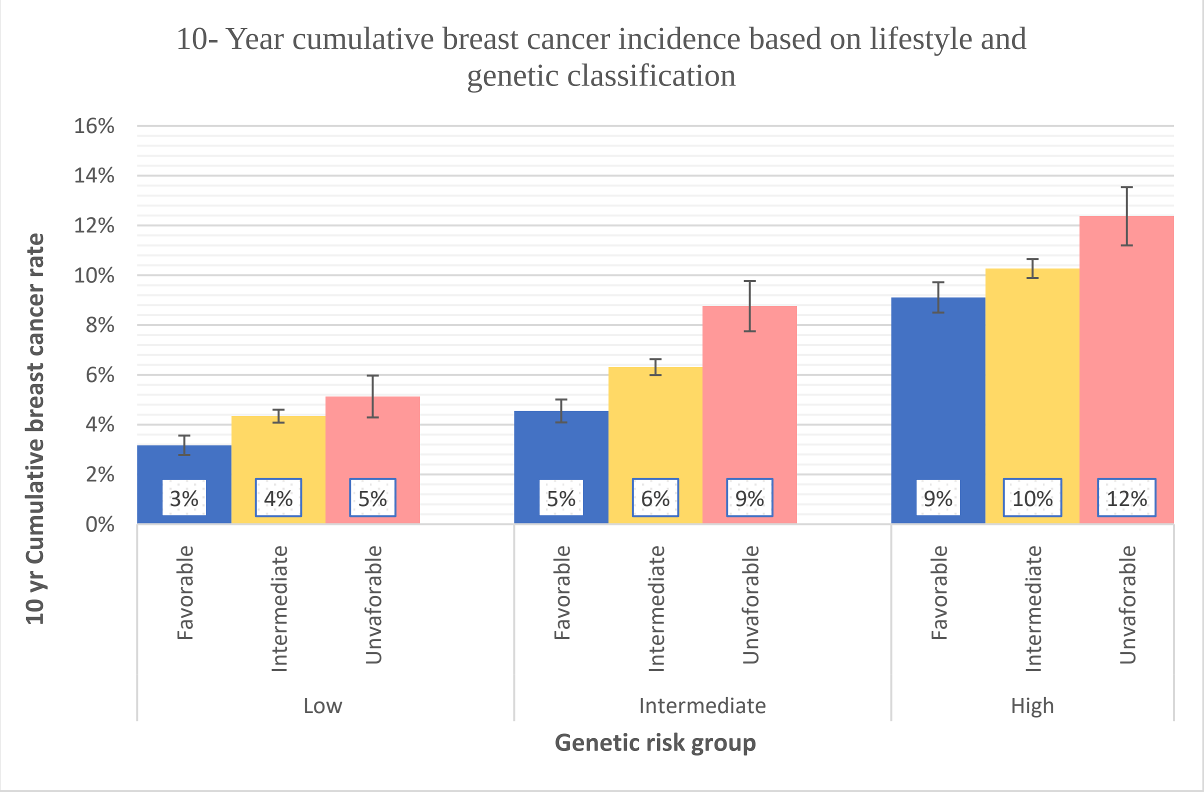


Figure 6.2: 10-year cumulative breast cancer incidence rate of UK Biobank post-menopausal females classified according to genetic and lifestyles factors. The error plot at the top represents the average rate with the maximum and minimum incidence rate.

## 6.5 Discussion

It has been estimated that BC could be prevented in 23% of cases in the UK [318]. Thus, it is important to understand the contribution of modifiable risk factors to BC and how they affect or add to the inherited genetic factors. This study therefore investigated the association between genetic and lifestyle factors with BC risk and tested the hypothesis that BC risk in postmenopausal women can be modified or reduced by improving lifestyle habits, even for the highest genetic risk group. It was opted to investigate our hypothesis only in postmenopausal women because of the high proportion of BC incidence and prevalence in this group [179, 327]. Furthermore, BC in premenopausal women is usually a more aggressive disease, likely caused by high penetrance genes, [328-331], resulting in a less-favorable prognosis [332].

This study used genetic and lifestyle data generated by UK Biobank, a longitudinal study of the contribution of genetic, environmental, and lifestyle risk factors in disease. Participants were grouped by their level of polygenetic risk for BC using the SNV data available within the UK Biobank database. The 305 SNVs included in the PRS were mainly common variants with limited contribution to BC risk. Aggregated effect sizes of these SNVs were used to develop a standardised PRS.

Although many risk/protective factors contribute to BC development [333], 5 robust modifiable risk factors were selected, recognized previously by Cancer Research UK as being associated with BC in white females [323, 334-336]. The frequencies of these modifiable risk factors are high in women in the UK, and if they can be modified can potentially reduce BC incidence. The prevalence of these 5 modifiable risk factors in the UK Biobank female cohort were as follows: 63.4% exhibiting unhealthy weight in BC cases vs 59.8% in controls, 13.3% of BC cases having no regular exercise vs 12.0% of controls, 42.6% of BC cases with regular alcohol intake vs 40.2% of controls, 79.4% of BC cases who used oral contraceptives vs 80.5% of controls, and 30.4% of BC cases who received HRT vs 25.2% of controls.

The findings from other large cohorts, including the Million Women Study and the Breast Cancer Association Consortium, have indicated that BC risk increase is associated with unhealthy weight [323, 324, 337], no or limited exercise [324, 325], level of alcohol intake [324-326], use of oral contraceptives [323, 324], and use of HRT [83, 315, 323, 324, 338]. The Cancer Research UK suggested that the relative contributions of these factors to BC development are 2% for HRT, 8% for obesity, 8% for alcohol intake, and less than 1% for use of oral contraceptives [318]. The results from this study echoed the risk factors published by the Cancer Research UK in that maintaining a healthy weight is associated with reduced BC risk by 13.9%, participating in regular exercise is associated with reduced BC risk by 12.2%, maintaining alcohol intake at less than 3 times a week is associated with reduced BC risk by 10.7%, and avoiding HRT use is associated with reduced BC risk by 22.9%. Even though the numbers vary between our study and the CRUK but they are in the same direction and concluded similar findings for HRT use, obesity and alcohol intake. For the HRT the CRUK reported contribution of 2% (all population) when reporting the PAF and ours was 22.9% (post-menopausal females only). This difference in the figures because the CRUK is reporting the PAF while I am reporting the risk increase. These results support the selection of these modifiable lifestyle risk factors for BC, with the exception of oral contraceptive use. Thus, further studies are needed to investigate whether there is a causal association between new risk factors and BC using, for example, a mendelian randomization approach. The weighted approach for each factor should also been explored to quantify the risk estimates and guide prevention measures.

Even though oral contraceptive use has been suggested previously to be associated with BC, this risk factor did not show any association in our study. Possible explanations for this observation could be that it did not take into account other related factors that could be associated with the results, including the type of oral contraceptive used [339], the duration of use [340], and age at the time when the drugs were stopped [341]. Furthermore, women who have had human chorionic gonadotropin injections as part of infertility or weight loss treatments showed a lower risk of BC [342]. All of these factors may have implications in BC risk. For example, if women stopped oral contraceptive use for more than 10 years before their enrollment in the UK Biobank study, their BC risk will be reduced or returned to the same risk of women who never used oral contraceptives [341].

Exhibiting 2 or 3 of these healthy lifestyle factors (intermediate lifestyle) was associated with increased risk of BC by 24.5% compared with an increase of 43.6% in women who adhered to none or 1 of these factors (unfavorable lifestyle). Our findings suggest that women may be able to alter or reduce their risk of developing BC by following healthier lifestyles.  Further analysis demonstrated that a high PRS was associated with higher risk of BC. This level of increased risk is in line with other published findings [120]. The HRs derived from our analysis were generated by including only postmenopausal women. In contrast, the study by Mavaddat et al [120] reported HRs that were derived from both premenopausal and postmenopausal women.

The beneficial risk-reducing association of adhering to healthy lifestyles across all genetic risk stratification groups supports our hypothesis that BC risk reduction is seen regardless of the effect size of the PRS. Also an association between 10-year cumulative BC incidence rate and both lifestyle and genetic factors was found when assessed together. This increase suggests that BC incidence may be reduced by following favorable lifestyles even in women with high genetic risk.

This study suggests that the lifestyle followed by women may contribute to reducing the incidence of BC in those who have an increased genetic predisposition for this condition. Similar approaches have been used to investigate complex risk factors associated with dementia [181] and coronary heart disease. [180]. Both studies came to a conclusion similar to ours. In the dementia study, by adhering to favorable lifestyles (no current smoking, moderate alcohol intake, healthy diet, and regular exercise), the level of dementia was reduced. Similarly, in coronary artery disease, no smoking, no obesity, healthy diet, and regular exercise were associated with a reduction in the extent of coronary heart disease in participants, and this result was also observed in cases within the highest PRS group.

### Strength and limitations

A strength of our study is that a large sample size was analyzed and the selection of participants was spread across the UK. Furthermore, the quality and comprehensive nature of the phenotypic exposures assessed by UK Biobank were robust and of high standards. Our use of a prospective study design allowed exposure assessment before BC development in the cohort. However, the study has some limitations. The PRS used was restricted to white women and therefore presents a limitation on its generalizability to a wider range of racial/ethnic groups. Additional validation of these PRSs in other populations is needed to further understand its utility in genetic risk stratification. Our analysis was restricted to postmenopausal women; therefore, these results cannot be applied to premenopausal women. Moreover, lag measurement period of two years between the recruitment date and incidence of BC was not taken into consideration in this analysis. That was done based on two reasons: 1) I applied the same criteria used by the  UKBiobank- identification method of incident cases (incident cases were identified post recruitment to the biobank study [173] 2) to maximise BC incident cases as the UKB is relatively a young cohort and cancer incidence was lower than general population.  However, this limitation could lead to reverse causation.

Additionally, deceased females were removed from the analysis and that could lead to biased results of the interested estimate. In epidemiological studies, identification of an appropriate comparison group to the case group is a very important aspect.  The control group should be representative of the population from which cases were drawn.  An important principle of both observational and interventional study designs is to ensure unbiased comparisons between a group with and a group without exposure to a condition of interest. The impact of biased comparisons are results which either under- or overestimate the risk. In this analysis, I excluded deceased subjects from both cases and controls (survival cohort). Deceased controls may represent generally poorer health in the population they were drawn from thus including deceased controls may introduce bias to the estimated risk particularly if risk factors are associated with cancer such as smoking , alcohol consumption etc. The bias could affect the risk towards the null. In 2017, a systematic review on the use of deceased controls in epidemiologic research reported general agreement on the reviewed studies on the use of deceased controls, which is likely to introduce bias, and should be avoided in any scenarios [343] . Furthermore, the UKbiobank is a cohort study design and at the time of submission this piece of work, the study was still ongoing therefore at any time point, controls may become cases. Deceased controls however was eliminated on changing status to become cases and therefore could potentially introduce bias. As the analyses also excluded deceased cases, this could also potentially introduce survival bias and effect the estimated risks, particularly when case fatality was greater with higher exposure [344].

Moreover, the scoring approach could be improved by assigned a weighted score for healthy lifestyle factors by using β coefficients of each lifestyle factor derived from the Cox proportional hazards regression model. Another study limitation is that this study did not look for a formal interaction between lifestyle and genetic factors due to insufficient study power. Preliminary data analysis suggested no significant interaction between different lifestyles and genetic risk groups, therefore the two variables were considered as independent in the analysis.

However, the benefits reported herein for healthy lifestyle factors may also be seen in younger women. In addition, our analysis did not investigate the various known pathologic-based subtypes of BC, including ER positive and negative, PR positive and negative, and ERBB2 (formerly HER2 or Her2/neu) positive and negative.

Conclusions

The results of this study suggest that promotion of healthy lifestyles through adequate levels of exercise, healthy weight, no or limited alcohol intake, and avoidance of hormonal replacement therapy should be encouraged to reduce the risk of BC. Following a healthy lifestyle appears to be associated with a reduced level of BC risk in all 3 genetic risk strata, further illustrating the importance of lifestyle factors in common diseases with a genetic predisposition, such as BC.

## Acknowledgements

We would like to express our gratitude to UK Biobank staff and participants for this valuable source of data. The application number of this work is 5791. Professor Kenneth Muir and Dr Artitaya Lophatananon had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. KM and AL are supported by the NIHR Manchester Biomedical Research Centre and by the CRUK Integrated Cancer Epidemiology Programme ICEP [grant number C18281/A19169]).

# Appendix

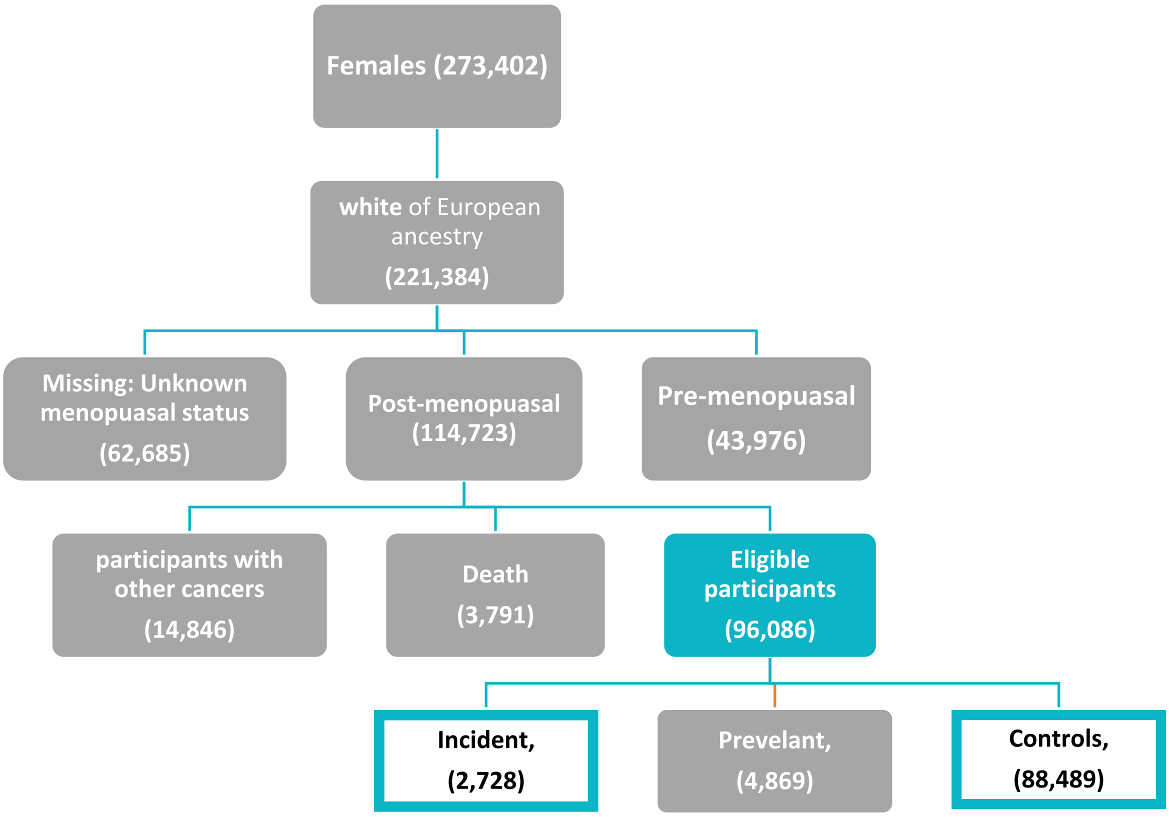


Figure S1: Number of participants in each filter step.

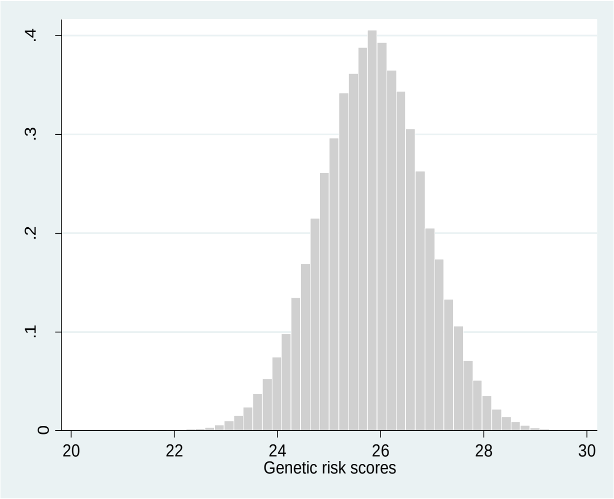


Figure S2: Distribution of standardised polygenic risk scores of the 305 selected SNPs among UK Biobank post-menopaused females

# Chapter 7 : Development and assessment of breast cancer risk prediction models based on the UK Biobank female cohort

Publication number 4

## Development and assessment of breast cancer risk prediction models based on the UK Biobank female cohort

This chapter is presented as journal article:

**Al-ajmi, K.,** et al., Development and assessment of breast cancer risk prediction models based on the UK Biobank female cohort. Manuscript prepared for submission.

Development and assessment of breast cancer risk prediction models based on the UK Biobank female cohort

Kawthar Al-ajmi[1], Artitaya Lophatananon[1], Krisztina Mekli[2], William Ollier[1,3], Jennifer Vena [4], Grace Shen-Tu[5,6], Jian-Yi Xu[7,] Kenneth R. Muir[1]

[1]Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK M13 9PT
[2]Cathie Marsh Institute for Social Research, School of Social Sciences, Faculty of Humanities, The University of Manchester, Manchester M139 PL, UK
[3] Faculty of science and Engineering, Manchester metropolitan University M1 5GD
[4] Alberta's Tomorrow Project, Cancer Control Alberta, Alberta Health Services, Richmond Road Diagnostic and Treatment Centre, 1820 Richmond Road SW, Calgary, AB, Canada, T2T 5C7.
[5] University of Alberta, Edmonton, Alberta, Canada; the Research Institute, [5] Hospital for Sick Children, University of Toronto, Canada.
[6] Cancer Measurement, Outcomes, Research and Evaluation, Cancer Control Alberta, Alberta Health Services, Calgary, AB, Canada

**Key words:** Risk prediction model, breast cancer, UK Biobank.

## 7.1 Abstract

**Background:** Risk prediction models which estimate the probability of breast cancer (BC) in women are becoming an increasingly important tool for focusing early screening and delivering programmes designed to reduce exposure to modifiable risk factors. Risk models for breast cancer were developed based on epidemiological risk factors and epidemiological factors plus genetic factor stratified by the menopausal status.

**Methods:** Data from the UK Biobank longitudinal cohort were used which provided a total of 3,565 incident BC cases and 126,815 matched controls. Relative risks were estimated to compare the risk of each female with female controls matched for age and menopausal status. Then, backward stepwise regression and bootstrap regression were performed to identify the best fit model. An absolute 5-year risk of breast cancer was estimated using national incidence and mortality statistics of the UK. Model's performance was evaluated using the E/O (Estimated to Observed) statistics, Hosmer-Lemeshow goodness of fit test and AUC (Area under curve). Furthermore, a polygenic risk score of 305 selected SNPs was added to the original models. At the end, the Alberta Tomorrow Project (ATP) cohort was applied to externally validate our models.

**Results:** Among pre-menopausal females, age, family history, age at menarche, height, physical activity and BMI were significant. However, among post-menopausal females, age, family history, height, number of live births, HRT use, alcohol intake and BMI were significant. The discriminatory power for pre- and post-menopausal epidemiological models were 0.584 and 0.580, respectively. Their discriminatory powers were improved to 0.665 and 0.648 respectively after including the polygenic risk score. The external validation results of epidemiological models suggest our model was well calibrated.

**Conclusion:** Many of the BC risk factors which make a significant contribution to the models were modifiable. Thus, these models could be used to increase awareness and potentially help inform behavioral changes. Moreover, these models could be used to identify high-risk females to increase regular medical surveillance using the extended genetic model.

## 7.2 Introduction

Breast cancer (BC) is the most common female cancer and is the second most common cause of cancer death amongst females. In the UK, 15% of newly developed cancer cases are BC cases [7, 8]. Several risk prediction models have been developed to estimate the likelihood of developing BC as summarised in systematic review articles [22, 27, 345]. The included variables can be demographic, epidemiological, genetic, and/or clinical factors. The available BC models are derived principally from either genetic or non-genetic risk factors. The majority of these models are however, not specifically designed to be used by the public and do not focus on modifiable factors which could have an impact on prevention and health promotion.

Most risk prediction models which are based on non-clinical and non-genetic risk factors (similar to our proposed model) were based on the original Gail model [219] or Colditz and Rosner [346] model [333]. Overall, these models demonstrated a good calibration (E/O near 1) and fair discrimination with C-statistics ranging from 0.61 to 0.65 (internal validation) which is closer to 0.5.

This study aims to develop an individualised risk prediction model for BC focusing on the modifiable risk factors using the UK Biobank data (UKB). Models were developed based on the menopausal status (pre- and post-menopausal models). The risk of BC increases as the female ages. Early menarche and late menopause are associated with increased risk of BC [347]. Different risk factors, prognosis, and mortality rates have been previously reported in pre-and post-menopausal BC [348]. Therefore, the aim was to develop different BC risk models based on the menopausal status. The main goal of the model is to enable the use in cancer education and prevention.

## 7.3 Materials and methods

### 7.3.1 Study population and study design

c) Model development-training data

Data from the UK Biobank longitudinal cohort were used to develop these models. UK Biobank is a national study of 502,650 subjects aged between 39 to 71 years at time of recruitment. Subjects were enrolled between 2006 and 2010 and are still followed for health outcomes. More details can be found at http://www.ukbiobank.ac.uk/.

d) Model validation- testing data

An independent Canadian Caucasian cohort (The Alberta's Tomorrow Project [ATP]) was used to validate the epidemiological models externally. This longitudinal population study of 55,000 participants (aged between 32 to 71 years) was initiated in the year 2000 with a planned follow-up period of 50 years.

### 7.3.2 Defining cases and controls in the UK Biobank

Our outcome is the malignant neoplasm of breast (breast cancer BC) in females. The cases and controls were identified according to the criteria mentioned in Table 7.1. There are four sources to identify the BC cases (ICD10, ICD9, self-reported, and death registry). Deceased subjects have no chance to develop the disease (no chance to be an incident case) so they were excluded from the analysis. If any BC cases appeared in the death registry variables, then they were excluded from the analysis. If the BC cases appeared as an incident in any of these four identification sources then the cases were deemed as incident cases. In addition, prevalent cases were defined only if they have been identified as prevalent by any of the four sources. This is only applicable if none of the three methods had incident cases in their classification. BC cases failed to be identified as an incident or prevalent will be excluded. Also subjects who had informed the UKB of their withdrawal to the study were excluded. In total, there are 15,920 BC cases in UK Biobank with incident cases (6,125) and prevalent cases (9,795). However, the analysis was restricted among Caucasian only and ended up with 5,044 incident cases and 7,854 prevalent cases. Yet, incident cases only were used in the analysis, which was later divided according to the menopausal status (cases reduced to 3,565 after removing subjects without menopausal status information).

Table 7.1: Identification of cases and controls of UK Biobank cohort used in the analysis

| **Sources** | **Breast cancer cases** | | **Controls** | **Subjects excluded from the control group** | | |
|---|---|---|---|---|---|---|
| | **Incident** | **Prevalent** | **Controls used in the analysis** | **Participants with cancer history** | | |
| | | | | **Alive participants with cancer** | **Death registry** | |
| **ICD10** | Codes start with C50 and its subclasses, C501, C502, C503, C504, C505, C506, C507, C508, and C509 | | Participants without any other cancer codes from the three sources | Participants with any cancer code (Other than breast cancer) from the three sources either (breast in situ, other in situ, other cancers, and neoplasm of unknown nature or behaviour | All dead females were excluded from the analysis regardless the cause | Total |
| **ICD9** | Codes start with 174 and its subclasses 1741, 1742, 1743, 1744, 1745, 1746, 1747, 1748, and 1749 | | | | | |
| **Self-reported cancers** | 1002 | | | | | |
| **Pre-menopausal** | 837 | 250 | 38,328 | 4,172 | 388 | 43,975 |
| **Post–menopausal** | 2,728 | 4,869 | 88,487 | 14,846 | 3,791 | 114,721 |
| **Used in the analysis** | **3,565** | **-** | **126,815** | **-** | **-** | |

*Caucasian participants only, **No menopausal status information

### 7.3.3 Defining cases and controls in the ATP cohort

The outcome of interest (BC) was identified using the Alberta Cancer Registry (ACR) database provided by the ATP team. Three variables were used to identify the BC cases (ACR cancer site specific, ACR cancer site aggregate, and ACR ICD_O topography). Subsequently, the incident cases were identified if age at cancer diagnosis was greater than enrollment age. Any BC incident cases were included in the analysis along with the controls (alive subjects without any history of BC (prevalent cases), other cancers, carcinoma in situ).

### 7.3.4 Defining pre- and post-menopausal status

The classification criteria used for assigning pre-menopausal status were: females aged ≤ 55 years old (this age cut-point was based on the NHS's definition of menopause age in the UK which is between 40 to 55 years [174]), with menarche age ≥ 7 years old (the UK ranges of menarche age is 7 to 20 years [175]), who reported still having their period and did not undergone neither hysterectomy nor bilateral oophorectomy. Post-menopausal females were defined as those who did not report a history of hysterectomy or bilateral oophorectomy and reported no longer having periods, and their menopause age ≥ 40 years old. In total, 57,712 pre-menopausal females and 138,554 post-menopausal females were identified. Following stratification by Caucasian heritage, these numbers were reduced to 43,975 pre-menopausal and 114,721 post-menopausal females. For the ATP, the same definition was applied, and the total numbers are 10,112 for the pre-menopausal group and 15,096 for the post-menopausal group.

### 7.3.5 Generating the polygenic risk score (PRS)

A PRS was generated from the high-density single nucleotide polymorphism (SNP) genotyping data for the UK Biobank dataset of 488,377 participants. Analysed samples were Caucasian females (matched self-reported sex confirmed with genetic sex and Caucasian status based on genetic ethnic grouping). A further quality control step removed individuals with >2% ‘missingness’, those who were outliers for heterozygosity and either duplicates or highly related participants.

The PRS was constructed using 305 SNPs as determined by Mavaddat et al, 2019 [120]. Their details are presented in table S1. Forty of these 305 SNPs were directly genotyped and successfully passed the QC steps applied by UK Biobank (https://www.biorxiv.org/content/10.1101/166298v1). The remaining 265 SNPs were statistically imputed. The quality of the imputation was estimated, and the lowest information score was 0.86. Linkage disequilibrium (LD) was assessed and no $r^2$ value between two SNPs reached 0.9. Plink 1.90 open source software was used to carry out the QC process [188].

The PRS was calculated by adding up the number of risk alleles at each SNP and then multiplying the sum by the effect size as estimated by Mavaddat et al, 2019 [120]. The PRS was standardised by dividing each raw PRS by the standard deviation of the PRS derived from the control group. This standardised PRS was then used in fitting both pre- and post-menopausal models.

### 7.3.6 Statistical analysis

All the statistical analysis were carried out using Stata MP 14.1 software for Windows [177]. Genetic data were analysed using PLINK 1.9 and PLINK 2.0 [188].

**Model development**

The risk factors selected to incorporate into prediction models were done by assessing the association between all the available risk factors listed in table 7.3 with BC in the development dataset. Relative risks (RR) and 95% confident intervals (95% C.I.) were computed using the binomial generalised linear regression model. This regression was used to deal with the binary outcome and flexible linear regression. It allows the response variable (dependent and independent) to have any error distribution other than normal distribution.

The analysis was stratified by the menopausal status as BC risk factors. Bootstrap regression (non-parametric approach were no distribution assumptions are required) of 100 simulations and backward stepwise regression (used in case of collinearity between variables as backward regression will keep variable while forward stepwise will firmly remove both variables) were used to identify significant factors to fit the model with the highest prediction power in each menopausal status. Once the selected predictors were identified, multicollinearity test for all variables in the models was run to ensure an absence of multicollinearity (see results in tableS2). Models were then fitted using multivariate logistic regression to assess their performances. Various tests were applied to verify model fit (post estimation) including model specification using *linktest* and assessment of linearity in the logit (assess whether log odds of the outcome is linearly associated with the covariates) using *lowess* graph.

**Model performance**

Model internal validation was performed using cross validation of 10 folds to assess model discrimination and calibration. For each k-fold in the dataset, the model was built on k – 1 fold and subsequently was tested to check the performance for k$^{th}$ fold. The command for performing the 10-fold cross validation is called "*cvAUROC*" [198]. Model discrimination was estimated using sensitivity and specificity to calculate the AUC. The AUCs were plotted using the receiver operating characteristics curves (ROC). Calibration of the models was estimated by comparing the expected against the observed event. Hosmer-Lemeshow goodness of fit test was used to assess the model calibration.

External validation was assessed using the testing dataset (the ATP cohort) and both discrimination and calibration were tested.

### 7.3.7 Absolute 5-years risk calculation:

The 5-years absolute risk (AR) of breast cancer for pre- and post-menopaused UK females was estimated. The AR calculation was described previously [199]. Here a demonstration of an example of calculation for the pre-menopaused females. Steps below described the AR calculation.

Step1: The risk was calculated by estimating the risk component of the risk factors. The risk component was derived by multiplying the RR of each risk factor associated with that individual (r= RR1 x RR2 x RR3 x RR4 ….. x RRn; see Table S1).

Step2: Derive values to compute baseline hazard rate. First value is the constant value which was calculated by the following formula:

$$constant\ value = \frac{\%\ of\ BC\ in\ Pre{-}menopausal\ (age\ 0{-}54)}{\%\ of\ females\ in\ pre{-}menopausal\ status\ (age\ 0{-}54)}$$

(See supplementary materials table S1 and table S2). The numbers to compute constant value are obtained from the Office for National Statistics (ONS) (https://www.ons.gov.uk/). In our example, the pre-menopaused constant was 0.439.

Second value is BC age-specific rates. There were estimated from ONS census for every 5-year age until age 54. The BC age-specific rate was calculated by the following formula:

$$BC\ age - specific\ rate = \frac{number\ of\ BC\ cases\ at\ specific\ age}{Total\ female\ population\ at\ specific\ age}$$

See appendix table S1 for example.

Third value is the attributable risk (AR). AR was calculated using the formula:

$$AR = \frac{Incidence\ among\ exposed\ -\ Incidence\ among\ unexposed}{Incidence\ among\ exposed}$$

Where exposed in this scenario is being pre-menopaused. The pre-menopaused AR value in our example is -1.178.

Fourth value is BC age specific mortality rate. The BC mortality rate was estimated from all causes of death except BC death at specific age. This was done by subtracting the female death from BC from female death from all causes; see appendix.

Step3: The baseline hazard rate was obtained using values obtained from step2 with the following formula:

$$BC\ incidence\ base\ hazard\ rate = \left[\frac{constant\ for\ pre - menopausal\ \times\ BC\ rate\ by\ age}{100000}\right] \times [1 - AR]$$

Step4: The 5-year absolute risk for BC among pre-menopaused was calculated as percentage using values derived from step1-4 =

$$5 - year\ absolute\ risk = \left[\frac{RR\ of\ all\ risk\ factors\ \times\ BC\ baseline\ hazard\ rate}{(RR\ of\ all\ risk\ factors\ \times\ BC\ baseline\ hazard\ rate\ ) +\ BC\ mortality\ rate}\right]$$

$$\times [1 - \, exp[-(RR \; of \; all \, risk \; factors \; \times \; BC \; baseline \; hazard \; rate$$
$$+ \; BC \; mortality \, rate]$$

This 5-year absolute risk can be compared with BC age specific rate obtained from ONS (supplementary materials, table S1 and table S2). In our example, the 5-year absolute risk was calculated for the epidemiological model (pre-and post) only.

## 7.4 Results

### 7.4.1 Model development- Variable selection and evaluation

**Stepwise regression:** In the pre-menopausal epidemiological model, final model included age, family history of BC, moderate physical activity, menarche age, BMI and standing height. In the post-menopausal model, the final selected variables were age, family history of BC, number of live births, BMI, height, HRT use and alcohol intake.

**Bootstrap regression:** In the pre-menopausal model, variables selected were age, family history of BC, moderate physical activity, BMI and standing height (menarche age was dropped using this method). In the post-menopausal model, the final variables were age, family history of BC, number of live births, HRT use, height, BMI and alcohol drinking status.

For pre-menopausal model, stepwise regression defined 6 variables while bootstrap regression defined 5 variables. Further both models were tested for model performance (as presented below in model validation), and our result suggested variables from stepwise regression fitted the model better; hence 6 variables were incorporated in our final model. Whilst the variables derived from bootstrap and stepwise regression in the post-menopausal model yielded the same factors. Not all significant risk factors were included in the models (such as moderate activity, beef consumption, menopause age, number of live births, smoking, alcohol intake, hip and waist circumferences in post-menopausal model) as they didn’t give the best model fit and estimation.

Our model post estimation tests suggested that fitted models were well specified. The link function *_hatsq* P-value>0.05 indicated that the fitted model did not omit further relevant variables. The variables included in the models showed absence of multicollinearity (results in the supplementary table S2). Our results of both the tolerance and the *Variance Inflation Factor* (VIF) values were 1 or very close to 1, suggesting all of the variables in the models were completely uncorrelated with each other.

### 7.4.2 Model validation

a) Internal validation

Both calibration and discrimination of the models were tested separately. The results are presented in figure 1 and 2 for pre-model and figure 3 and 4 for post-model. Pre-menopausal model shows a good calibration. The Hosmer-Lemeshow goodness of fit test (chi-square p-value >0.05) and its plot suggested our model was well calibrated. Our model suggested a discrimination power of 57.9%. Similarly, post-menopausal model showed good calibration and discrimination power of 58.0%.

The discrimination power was considerably improved by adding the PRS into the models which resulted in pre-AUC of 0.665 and post- AUC of 0.648.

b) External validation

The ATP cohort (Canadian population) sample size included in the external validation exercise consisted of incident cases of 709 and controls of 32,238. There were 10,112 pre-menopaused females and 15,096 post-menopaused females. The results were 125 incident cases and 9,837 controls among pre-menopausal females and 332 incident cases and 14,110 controls among post-menopausal females.

Calibration and discrimination results are presented in fig 5 and fig 6 for pre and post epidemiological model, respectively. The pre-menopausal model, the Hosmer-Lemeshow goodness of fit test suggested model was not well calibrated (*Chi-square p-value* <0.05) and model discrimination (AUC 0.686) yielded much higher value as compared to the development model (AUC 0.584). For post-menopausal model, the calibration plot suggested model was well calibrated (*Chi-square p-value*>0.05). The discrimination power (AUC 0.585) was similar to the development model (0.587).

### 7.4.3 Extended models by adding genetic risk score to epidemiological models

As a step for extension the epidemiological model for BC risk prediction, standardised PRS was estimated and added to both models (pre- and post-menopausal). First, the performance of PRS alone was tested to predict BC in pre- and post-menopausal groups. The discrimination power for genetic pre-model was 65.17% and for genetic post-model was 62.59%. Next, the PRS was fitted and epidemiological risk factors into the pre and post model. The AUC of extended pre model was 0.665 and the AUC of the extended post- model was 0.648. The extended genetic models of both pre- and post-menopausal showed better discrimination and calibration performance (see figure 7.1, 7.2, 7.3 and figure 7.4).

Table 7.2: Internal and external validation results of all the six models

| **Models developed by our group** | **Internal validation** | | **External validation with ATP cohort** | |
|---|---|---|---|---|
| | **P value of the HL goodness of fitness** | **Concordance statistics (95% CI)** | **P value of the HL goodness of fitness** | **Concordance statistics (95% CI)** |
| ***Epidemiological models*** | | | | |
| Pre-Epidemiological-Model (No genetics) | 0.523 | 0.587 (0.567 - 0.607) | 0.343 | 0.686 (0.625 - 0.747) |
| Post- Epidemiological – Model (No genetics) | 0.534 | 0.580 (0.568 - 0.591) | 0.513 | 0.585 (0.537 - 0.632) |
| ***Genetic models*** | | | | |
| Pre-Gen-Model (genetics only) | 0.345 | 0.652 (0.630 - 0.673) | - | - |
| Post-Gen-Model (genetics only) | 0.467 | 0.626 (0.614 - 0.638) | - | - |
| ***Extended genetic models*** | | | | |
| Pre-Extended Genetic-Model (epi + genetic factors) | 0.344 | 0.665 (0.643 - 0.686) | - | - |
| Post- Extended Genetic-Model (epi + genetic factors) | 0.377 | 0.647 (0.635 - 0.658) | - | - |

Table 7.3: Relative risks of the significant risk factors in the regression model by the menopausal status in the UKB data.

| Relative risk | Frequency (%) | | Pre-menopausal | 95% CI | | Pre-model | Frequency (%) | | Post-menopausal | 95% CI | | Post- model |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Controls | Cases | | | | | Controls | Cases | | | | |
| *Unmodifiable* | | | | | | | | | | | | |
| **Age** | 38,329 (97.86) | 837 (2.14) | 1.031 | 1.011 | 1.050 | Included | 88,489 (97.01) | 2,728 (2.99) | 1.023 | 1.016 | 1.031 | Included |
| **Family history** | | | | | | Included | | | | | | Included |
| **No history** | 34,520 (90.33) | 701 (84.05) | Ref | | | | 78,408 (88.84) | 2,276 (83.80) | Ref | | | |
| **Mother or Sibling history of BC** | 3,593 (9.40) | 128 (15.35) | 1.754 | 1.449 | 2.125 | | 9,405 (10.66) | 412 (15.17) | 1.509 | 1.356 | 1.680 | |
| **Mother and Sibling history of BC** | 102 (0.27) | 5 (0.60) | 2.414 | 0.981 | 5.942 | | 440 (0.50) | 28 (1.03) | 2.192 | 1.493 | 3.219 | |
| **Menarche age** | | | | | | Included | | | | | | |
| **>13 years old** | 14,092 (37.88) | 283 (34.51) | Ref | | | | 31,690 (36.54) | 966 (35.99) | Ref | | | |
| **≤ 13 years old** | 23,108 (62.12) | 537 (65.49) | 1.157 | 1.001 | 1.339 | | 55,031 (63.46) | 1,718 (64.01) | 1.024 | 0.945 | 1.110 | |
| **Height** | | | | | | | | | | | | |
| **Below the mean** | 439 (1.15) | 11 (1.32) | Ref | | | | 2,251 (2.55) | 54 (1.98) | Ref | | | |
| **Within the mean** | 36,372 (95.01) | 785 (93.90) | 0.861 | 0.472 | 1.573 | | 84,634 (95.79) | 2,599 (95.45) | 1.280 | 0.975 | 1.681 | |
| **Above the mean** | 1,473 (3.85) | 40 (4.78) | 1.084 | 0.551 | 2.130 | | 1,471 (1.66) | 70 (2.57) | 1.984 | 1.383 | 2.846 | |
| **Standing height** | 38,284 (97.86) | 836 (2.14) | 1.019 | 1.008 | 1.030 | Included | 88,356 (97.01) | 2,723 (2.99) | 1.016 | 1.001 | 1.022 | Included |
| **Menopause age** | - | - | - | - | - | | 88,489 (97.01) | 2,728 (2.99) | 1.019 | 1.001 | 1.029 | |
| **Polygenic risk scores** | 31,555 (97.88) | 685 (2.12) | 1.735 | 1.607 | 1.873 | Included | 72,094 (96.99) | 2,237(3.01) | 1.584 | 1.518 | 1.653 | Included |
| *Modifiable* | | | | | | | | | | | | |
| **Moderate activity** | | | | | | Included | | | | | | |
| **High** | 18,102 (47.70) | 367 (44.16) | Ref | | | | 42,314 (48.70) | 1,258 (46.85) | Ref | | | |
| **Low** | 15,617 (41.15) | 340 (40.91) | 0.743 | 0.603 | 0.916 | | 36,081 (41.53) | 1,128 (42.01) | 0.888 | 0.780 | 1.011 | |

| **Relative risk** | **Frequency (%)** | | **Pre-menopausal** | **95% CI** | | **Pre-model** | **Frequency (%)** | | **Post-menopausal** | **95% CI** | | **Post- model** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Controls** | **Cases** | | | | | **Controls** | **Cases** | | | | |
| **Never** | 4,232 (11.15) | 124 (14.92) | 0.692 | 0.563 | 0.851 | | 8,494 (9.78) | 299 (11.14) | 0.845 | 0.743 | 0.960 | |
| **Number of live births** | 38,313 (97.86) | 837 (2.14) | 0.956 | 0.902 | 1.014 | | 88,452 (97.01) | 2,725 (2.99) | 0.953 | 0.922 | 0.986 | Included |
| **Hip circumference as continuous** | 38,276 (97.86) | 836 (2.14) | 0.996 | 0.989 | 1.002 | | 88,362 (97.01) | 2,724 (2.99) | 1.010 | 1.007 | 1.014 | |
| **Waist circumference as continuous** | 38,276 (97.86) | 835 (2.14) | 0.995 | 0.989 | 1.000 | | 88,366 (97.01) | 2,724 (2.99) | 1.009 | 1.006 | 1.012 | |
| **Waist circumference** | | | | | | | | | | | | |
| **Low** | 20,778 (54.29) | 465 (55.69) | Ref | | | | 37,013 (41.89) | 1,062 (38.99) | Ref | | | |
| **Moderate** | 9,270 (24.22) | 207 (24.79) | 0.998 | 0.845 | 1.177 | | 24,196 (27.38) | 736 (27.02) | 1.060 | 0.964 | 1.166 | |
| **High** | 8,224 (21.49) | 163 (19.52) | 0.886 | 0.740 | 1.060 | | 27,159 (30.73) | 926 (33.99) | 1.188 | 1.087 | 1.300 | |
| **Body mass index as continuous** | 38,259 (97.86) | 835 (2.14) | 0.984 | 0.971 | 0.998 | Included | 88,286 (97.01) | 2,722 (2.99) | 1.018 | 1.010 | 1.025 | Included |
| **Body mass index** | | | | | | | | | | | | |
| **Healthy** | 17,952 (46.93) | 422 (50.54) | Ref | | | | 34,911 (39.55) | 974 (35.80) | Ref | | | |
| **Overweight** | 12,346 (32.27) | 265 (31.74) | 0.913 | 0.782 | 1.066 | | 33,474 (37.92) | 1,055 (38.77) | 1.130 | 1.034 | 1.234 | |
| **Obese** | 7,628 (19.94) | 143 (17.13) | 0.797 | 0.658 | 0.966 | | 19,275 (21.84) | 672 (24.70) | 1.250 | 1.131 | 1.381 | |
| **HRT users** | | | | | | | | | | | | Included |
| **No** | 37,213 (97.25) | 810 (93.77) | Ref | | | | 52,181 (59.07) | 1,484 (54.54) | Ref | Ref | | |
| **Yes** | 1,054 (2.75) | 27 (3.23) | 1.177 | 0.798 | 1.735 | | 36,156 (40.93) | 1,237 (45.46) | 1.203 | 1.114 | 1.299 | |
| **OC users** | | | | | | | | | | | | |
| **No** | 3,185 (8.32) | 66 (7.89) | Ref | | | | 17,240 (19.50) | 561 (20.58) | Ref | Ref | | |
| **Yes** | 35,090 (91.68) | 771 (92.11) | 1.060 | 0.823 | 1.367 | | 71,149 (80.50) | 2,165 (79.42) | 0.935 | 0.851 | 1.028 | |
| **Smoking status** | | | | | | | | | | | | |

| Relative risk | Frequency (%) | | Pre-menopausal | 95% CI | | Pre-model | Frequency (%) | | Post-menopausal | 95% CI | | Post- model |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Controls | Cases | | | | | Controls | Cases | | | | |
| Never | 24,939 (65.17) | 551 (65.99) | Ref | | | | 52,687 (59.72) | 1,554 (57.07) | Ref | | | |
| Previous | 9,631 (25.17) | 210 (25.15) | 0.987 | 0.841 | 1.159 | | 29,045 (32.92) | 956 (35.11) | 1.116 | 1.028 | 1.211 | |
| Current | 3,700 (9.67) | 74 (8.86) | 0.905 | 0.708 | 1.157 | | 6,491 (7.36) | 213 (7.82) | 1.113 | 0.962 | 1.287 | |
| Alcohol drinking status | | | | | | | | | | | | |
| Never | 1,007 (2.63) | 18 (2.15) | Ref | | | | 3,906 (4.42) | 93 (3.41) | Ref | | | |
| Previous | 999 (2.61) | 20 (2.39) | 1.120 | 0.589 | 2.130 | | 2,897 (3.28) | 91 (3.34) | 1.319 | 0.984 | 1.768 | |
| Current | 36,302 (94.76) | 799 (95.46) | 1.231 | 0.769 | 1.973 | | 81,626 (92.31) | 2,543 (93.25) | 1.309 | 1.061 | 1.613 | |
| Beef intake | | | | | | | | | | | | |
| Within average | 34,592 (90.47) | 742 (88.97) | Ref | | | | 78,896 (89.38) | 2,391 (87.94) | Ref | | | |
| Above average | 3,645 (9.53) | 92 (11.03) | 1.177 | 0.945 | 1.465 | | 9,372 (10.62) | 328 (12.06) | 1.155 | 1.027 | 1.299 | |
| Raw vegetables intake | | | | | | | | | | | | |
| Yes | 34,089 (88.94) | 748 (89.37) | Ref | | | | 79,014 (89.29) | 2,422 (88.78) | Ref | | | |
| No | 4,240 (11.06) | 89 (10.63) | 1.045 | 0.837 | 1.305 | | 9,475 (10.71) | 306 (11.22) | 0.949 | 0.841 | 1.071 | |
| Processed meat intake | | | | | | | | | | | | |
| Within average | 37,563 (98.11) | 821 (98.09) | | | | | 86,989 (98.40) | 2,681 (98.35) | Ref | | | |
| Above average | 724 (1.89) | 16 (1.91) | 1.011 | 0.613 | 1.667 | | 1,415 (1.60) | 45 (1.65) | 1.032 | 0.765 | 1.392 | |

Highlight: indicates significant association

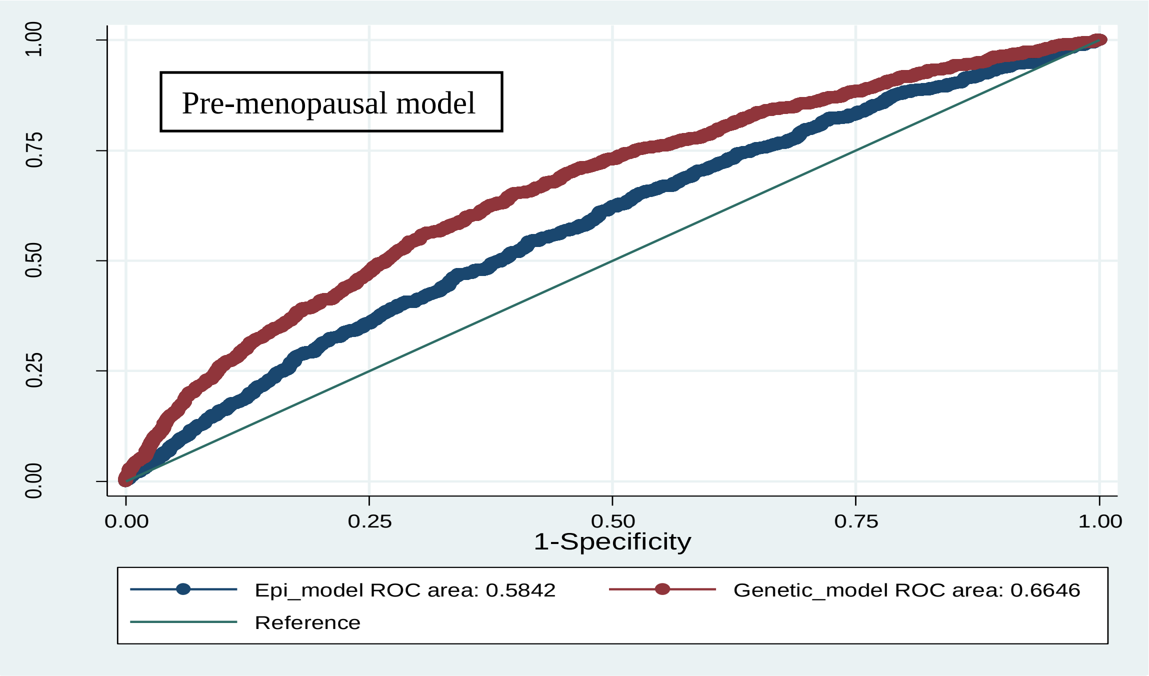


Figure 7.1: AUC curves of the pre- menopausal epidemiological model (blue) and the extended genetic pre-menopausal model (red) compared to each other.

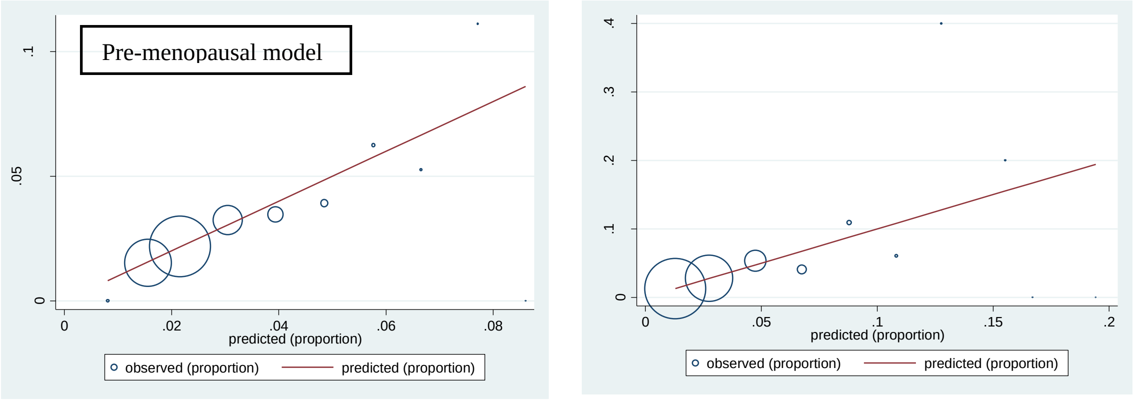


Figure 7.2: Calibration plots of pre- menopausal epidemiological model (left) and the extended genetic pre-menopausal model (right).

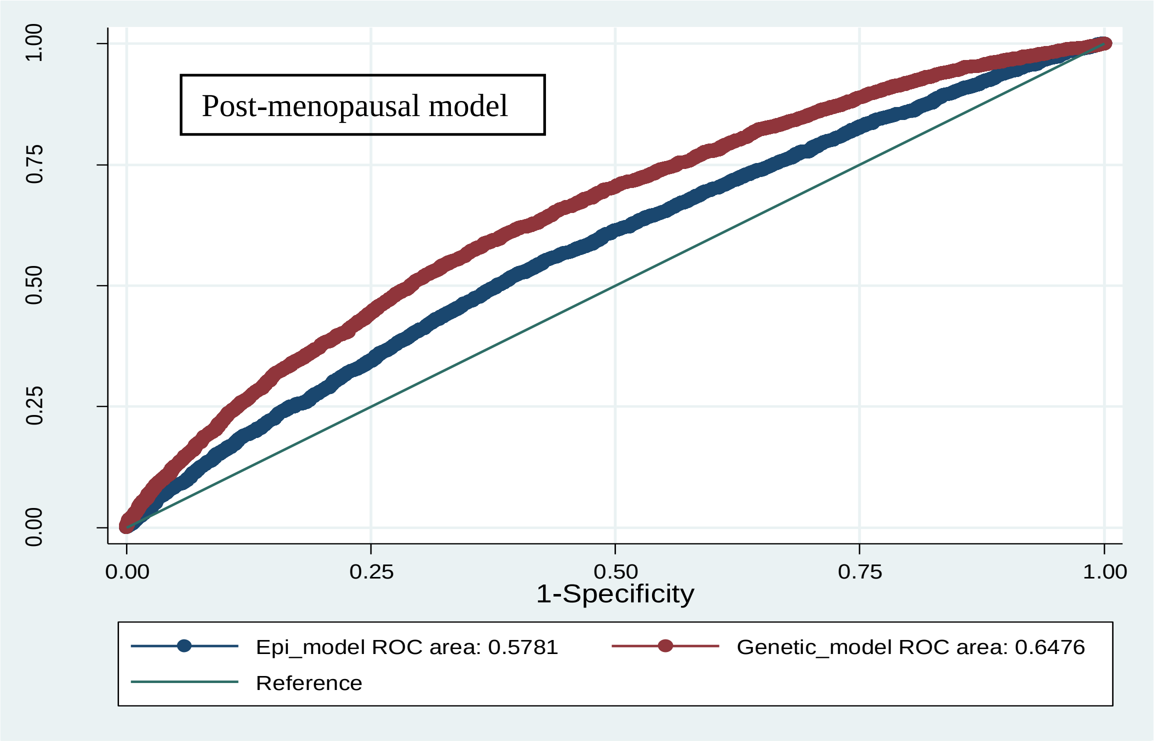


Figure 7.3: AUC curves of the post-menopausal epidemiological model (blue) and the extended genetic post-menopausal model (red) compared to each other.

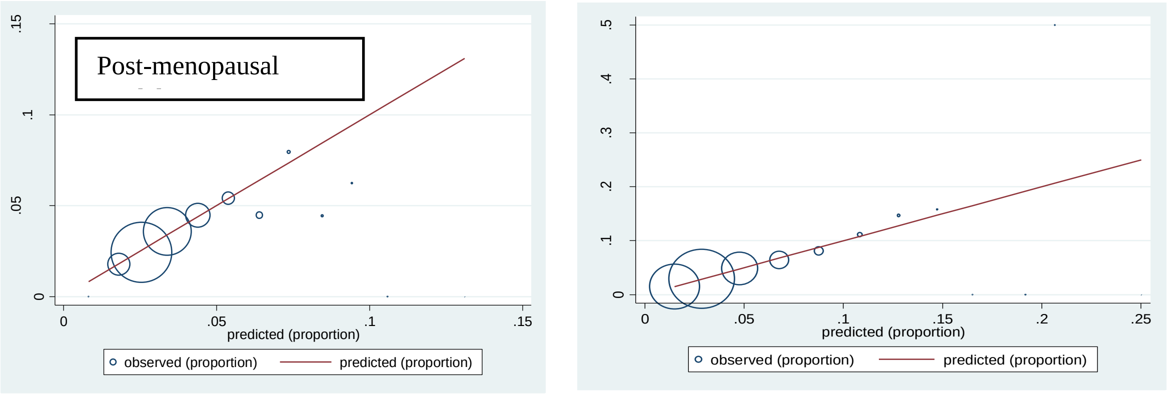


Figure 7.4: Calibration plots of post-menopausal epidemiological model (left) and the combined post-menopausal model (right).

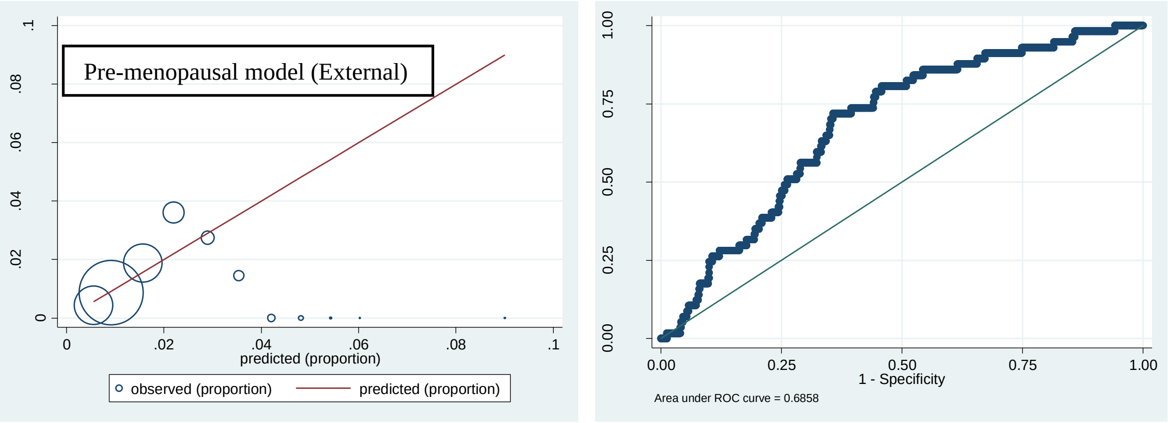


Figure 7.5: Results of the external validation, calibration curve and the AUC curve of the Canadian ATP cohort using the pre-menopausal BC subjects and controls.

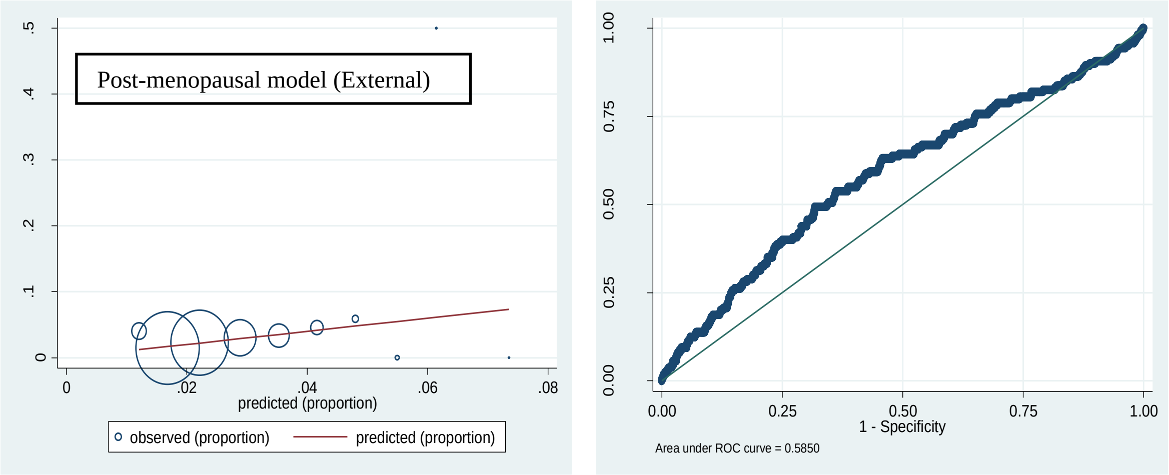
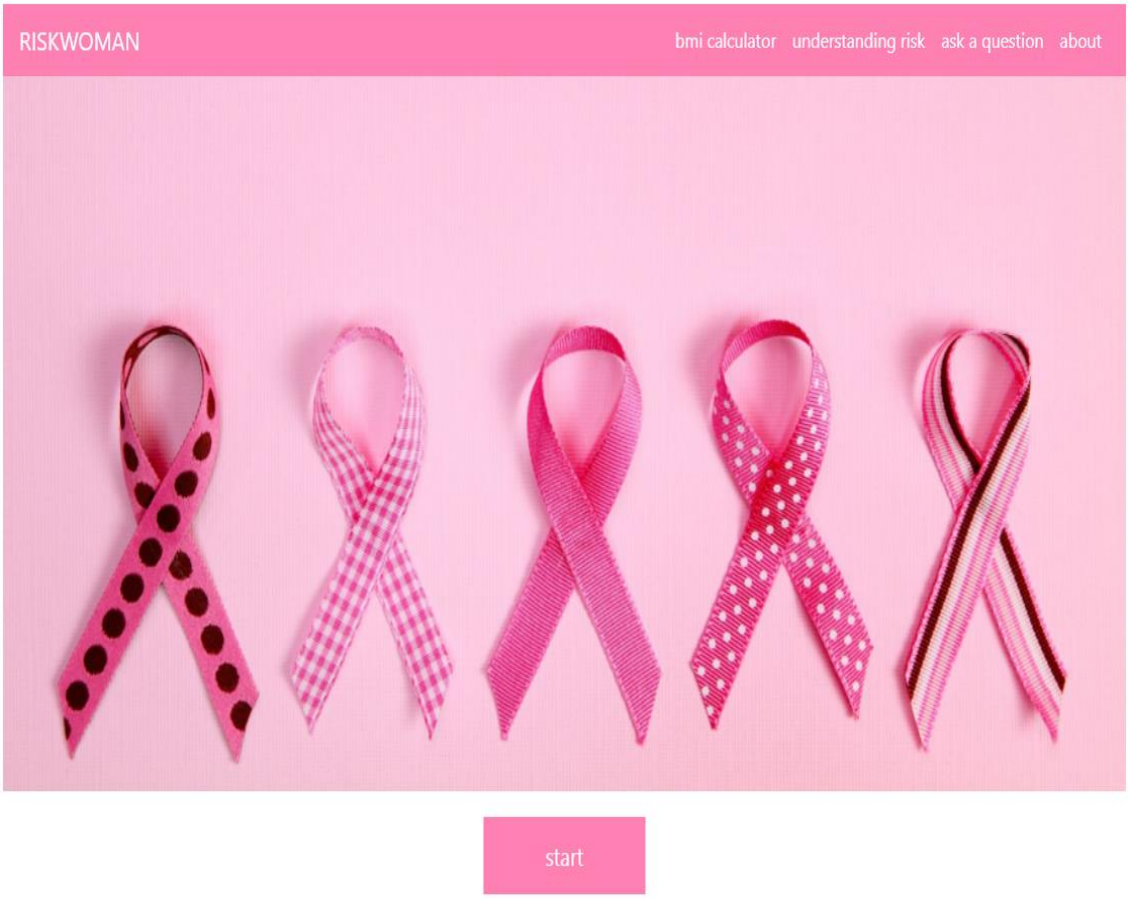


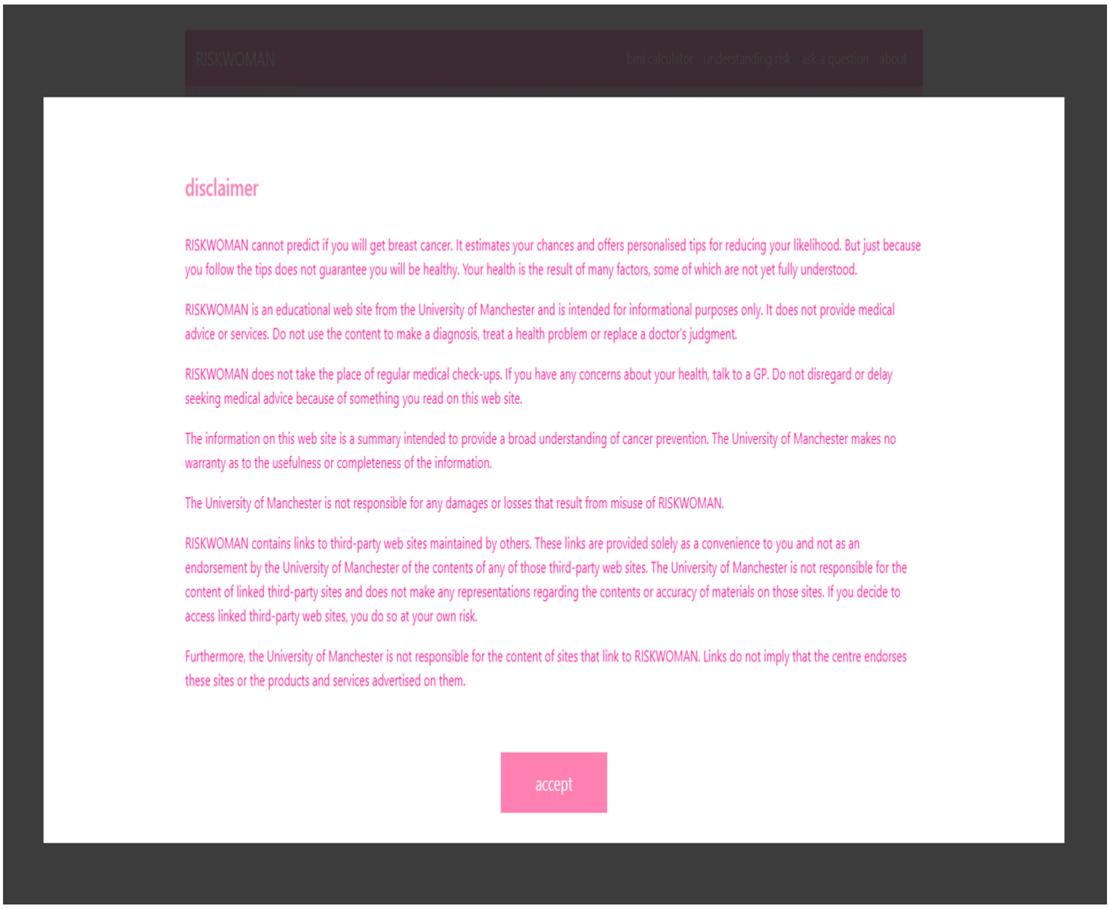
Figure 7.6: Results of the external validation, calibration curve and the AUC curve of the Canadian ATP cohort using the post- menopausal BC subjects and controls.

### 7.4.4 Developing breast cancer risk prediction web-based models (RiskWomen)

The next step was to develop a web-based page (RiskWomen) for the two epidemiological models as a start. The website was developed by two information technology specialists based on the information and data provided to them by our team. The made sure the website was simple, has all the required information, user-friendly, and eye catching to encourage females of all ages to use it. The website could be accessed by the desktop or smart phones. Even though, the website is finished but still not available for the public. Two focus groups were conducted in Limelight community centre. General females were asked to test the website and tell us their feedback on the model, webpage designs, wording, colors, their feeling before and after using the model and knowing their results At the beginning, the project was asked and the models and asked females who agreed to participate to sign a consent and handed them a leaflet containing all the required information.. Then they tried the website and the team was near if they needed our help and once, they finish their feedback was taken. Based on their feedback, the website was changed and improved accordingly. At this point, the website is still not ready for the public as more work and evaluation is needed. Snapshots of the preliminary model webpage is shown below.

Figure 7.7: Front page of the RiskWomen website





disclaimer

RISKWOMAN cannot predict if you will get breast cancer. It estimates your chances and offers personalised tips for reducing your likelihood. But just because you follow the tips does not guarantee you will be healthy. Your health is the result of many factors, some of which are not yet fully understood.

RISKWOMAN is an educational web site from the University of Manchester and is intended for informational purposes only. It does not provide medical advice or services. Do not use the content to make a diagnosis, treat a health problem or replace a doctor's judgment.

RISKWOMAN does not take the place of regular medical check-ups. If you have any concerns about your health, talk to a GP. Do not disregard or delay seeking medical advice because of something you read on this web site.

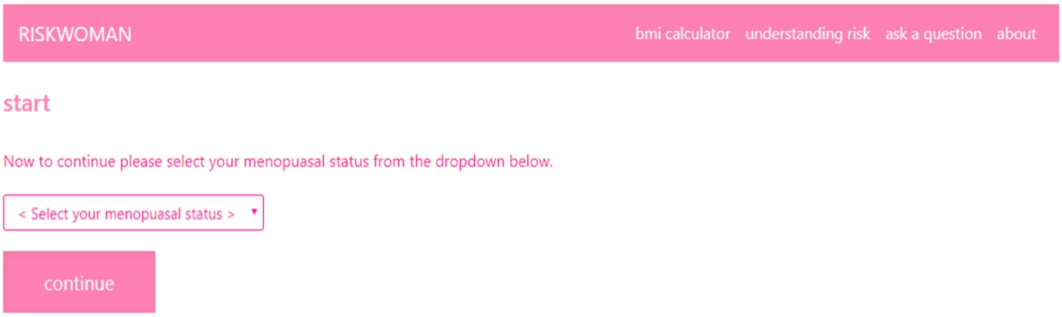
The information on this web site is a summary intended to provide a broad understanding of cancer prevention. The University of Manchester makes no warranty as to the usefulness or completeness of the information.

The University of Manchester is not responsible for any damages or losses that result from misuse of RISKWOMAN.

RISKWOMAN contains links to third-party web sites maintained by others. These links are provided solely as a convenience to you and not as an endorsement by the University of Manchester of the contents of any of those third-party web sites. The University of Manchester is not responsible for the content of linked third-party sites and does not make any representations regarding the contents or accuracy of materials on those sites. If you decide to access linked third-party web sites, you do so at your own risk.

Furthermore, the University of Manchester is not responsible for the content of sites that link to RISKWOMAN. Links do not imply that the centre endorses these sites or the products and services advertised on them.

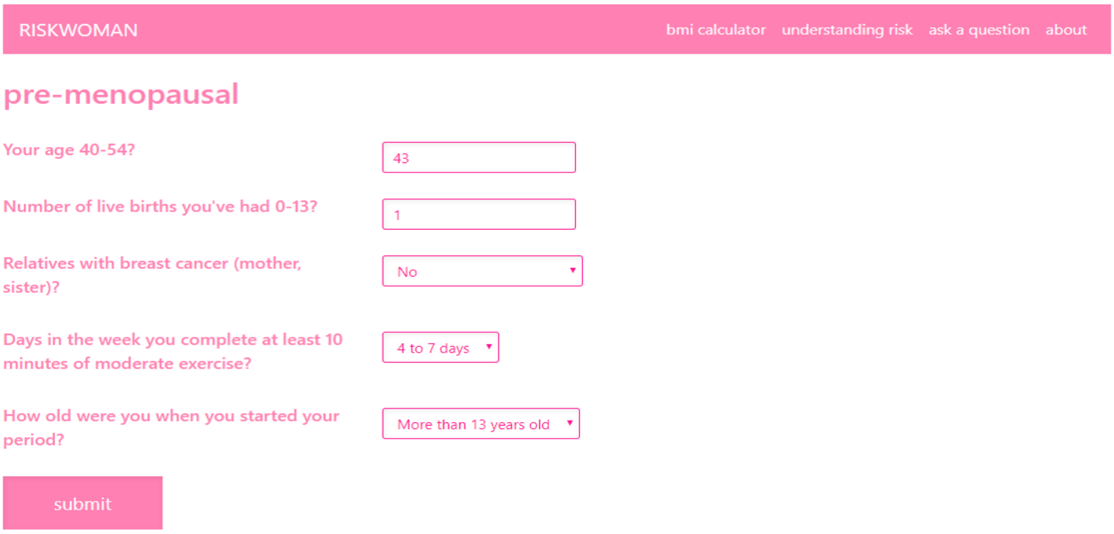
Figure 7.8: Disclaimer message provided by the website



start

Now to continue please select your menopuasal status from the dropdown below.

Figure 7.9: Selection of menopausal status to select which model is to be used (Pre- or post-menopausal model)



pre-menopausal

| | |
|---|---|
| Your age 40-54? | 43 |
| Number of live births you've had 0-13? | 1 |
| Relatives with breast cancer (mother, sister)? | No |
| Days in the week you complete at least 10 minutes of moderate exercise? | 4 to 7 days |
| How old were you when you started your period? | More than 13 years old |

Figure 7.10: An example of the two model (pre-menopausal model)

## 7.5 Discussion

Models to predict BC risk among the UK females were developed. The two models were based on menopausal status. Our models included factors which were considered modifiable such as physical activity, BMI, HRT use, number of live births (possibly modifiable) and alcohol intake. Many models in the literature focused mainly on family history, breast biopsies, and atypical hyperplasia [333]. The discrimination power (AUC) from both models was about 0.58 which is similar to the original Gail model (AUC= 0.56) [349]; however, it is still a modest AUC. Various possibilities could lead to the modest discriminatory power such as inclusion of weak evidence-based factors, lack of information of some of the significant factors such as breastfeeding, no inclusion of any clinical risk factors [333]. Moreover, BC heterogeneity might also contribute as different cancer types have different risk factors. In our study, it did not consider different types/subtypes of BC for model stratification. Incorporating BC pathology into prediction models enhances the probabilities of predicting BRCA1/2 carriers and BRCA1 in particular which might have implication on clinical decision-making [350]. Another possibility could be the effect of healthy volunteer selection bias in the UK Biobank cohort [351]. As a future plan, MR approach could be utilised to identify the causal risk factors (epidemiological, and clinical) for BC and incorporate it the newly developed risk prediction models.

The limited discrimination power of our two models may restrict its clinical utility, especially for targeting screening. Nevertheless, it can be used as an educational tool for the general public to increase their awareness about BC risk factors. Using our models can potentially benefit the general population over time to help educate and raise awareness for individuals to consider modifying their lifestyle habits. Potentially if the performance of the model can be further improved it can help in strategies to identify higher risk groups for more frequent mammogram screening.

The model's risk factors were applied to the ATP datasets as the external validation cohort to validate the epidemiological models. Our findings suggested good validation to the post-epidemiological model but not for the pre-epidemiological model. A possible explanation for this could be due to large number of missing values in the required risk factors which led to unreliable results, and small sample size of the incident pre-menopaused BC cases.

Additionally, the extended version of the epidemiological models by adding the PRS score improved the model's performance both pre- and post-menopausal models resulting in AUC of 0.664 from 0.586 in the pre-menopausal model and AUC of 0.646 from AUC of 0.580 in the post-menopausal model. Internal validation of the extended models using 10k cross folds also suggest similar values. As ATP have no genetic data, no external validation was done for the extended models. Further datasets from other cohorts with genotyping are being sought and will form the base of future work.

Moreover, a BC genetic risk prediction models were fitted using the PRS alone without any epidemiological risk factors for both pre- and post-menopausal cohorts. The discrimination power of these models AUC was 0.652 and 0.626 respectively. Even though these AUCs were very close to the AUC values of the extended model (combined genetic and epidemiolocal model), the epidemiological risk factors were incorporated to serve original aim of developing a risk prediction model with modifiable BC risk factors. Females can change their lifestyle based on their individualised risk of BC using these extended combined models. Furthermore, the AUCs were improved slightly by including both epi and genetic risk factors and in risk prediction research any improvement in the discrimination power is counted.

A review published in 2018 [333] revised the BC models, which included only modifiable risk factors as our two original models. None of these models [109, 221, 236, 346, 349, 352-358] built a separate model based on the menopausal status, but they used menopausal status as a risk factor fitted in the model. The smallest AUC of internal discrimination was 0.56 by the original Gail model [349], and the largest AUC was 0.65 [357] (among a subset in a Korean model). Many risk factors included in the previous models weren't significant in our UKB cohort such as: smoking, alcohol intake, age at first birth, age at subsequent birth, oral contraceptives, breastfeeding (data was not available in UK Biobank), pregnancy, menopause type, menstrual regularity (data was not available in UK Biobank), menstrual duration, gestation period, benign breast disease, history of breast biopsies and mammogram. Despite moderate model performance, calibration of these models was very good with E/O index very close to 1 (perfect agreement between expected and observed values).

The strengths of our project are the large sample size used in the development dataset (UKB), information availability on the most important BC risk factors, ability to add a genetic PRS, and an opportunity to develop two BC models by menopausal status. Depending on the objective of the user, they can choose between the epidemiological and genetic model or can use them both. On the other hand, there are some limitations which potentially affected our model performance. For example, breastfeeding which is previously documented as BC risk factor was not available in the UK Biobank. Another possible limitation is that UK Biobank participants tend to be healthier than the UK population. There is evidence that the participants are well educated, healthy and live in a less deprived area in the UK [351]. These advantages of the UK participants make them well aware of their health and usual risk factors of breast cancer are less prevalent among them. Subsequently, they might be not the real representation of the UK population. Finally, the models targeted Caucasian females only; hence they inherited a problem of generalizability to other ethnic groups.

**Acknowledgements**

We would like to express our appreciation to UK Biobank participants and staffs for this valuable source of data (application: 5791). We are also greatly thankful to the Ministry of higher education in Oman for supporting the first author to pursue her PhD degree. KM and AL are supported by the NIHR Manchester Biomedical Research Centre, the Alan Turing Institute under the EPSRC grant EP/N510129/1 and by the CRUK Integrated Cancer Epidemiology Programme ICEP [grant number C18281/A19169]). The ATP cohort is funded by Alberta Health, Alberta Cancer foundation, and Alberta Health Services.

## Supplementary materials

Table S1: Example of calculation for 5 years absolute risk using pre- and post-menopausal females (healthy and BC cases)

| Age | Breast cancer cases | Relative risk EE for all factors (r) = multiplication of all factors | (1/r) | Pop attributable risk (PAR) | (1-PAR) | age/sex incidence rate | baseline hazard ratio (h1) = (age/sex incidence rate) * (1-PAR) | h1*r = hazard ratio* RR | h2= mortality rate | h1*r + h2= m | (h1*r /h1*r + h2) = k<br>AD/AF | e (-5 *m) = c | Prob = (k *(1-c)) | 5-year AR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 42 | Healthy | 1.262 | 0.792 | 0.511 | 0.489 | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 | 0.475 | 0.981 | 0.009 | 0.009 |
| 55 | Healthy | 2.459 | 0.407 | 0.844 | 0.156 | 0.002 | 0.000 | 0.001 | 0.003 | 0.004 | 0.238 | 0.980 | 0.005 | 0.005 |
| 62 | Incident | 9.622 | 0.104 | 0.844 | 0.156 | 0.003 | 0.001 | 0.005 | 0.006 | 0.011 | 0.462 | 0.945 | 0.025 | 0.025 |

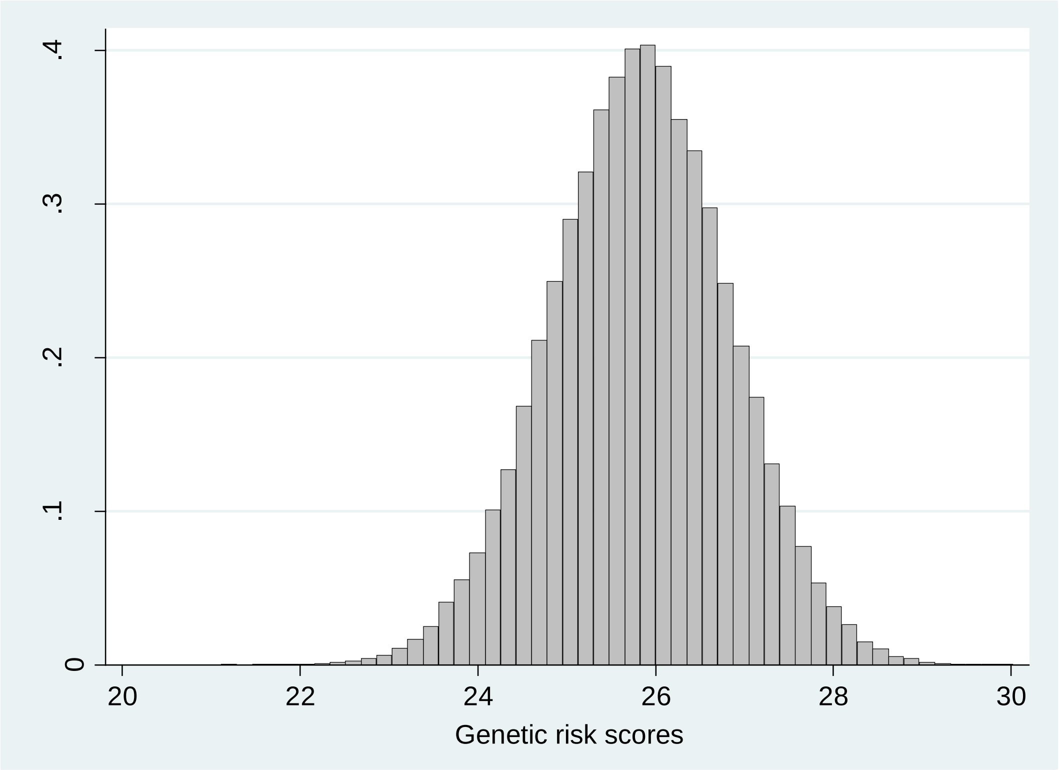


Figure S1: Distribution of standardised polygenic risk scores of the 305 selected SNPs among UK Biobank Caucasian females

Table S2: Multicollinearity test for pre- and postmenopausal variables

| **Variable** | **VIF** | **SQRT VIF** | **Tolerance** | **r-squared** |
|---|---|---|---|---|
| **Pre-menopausal variables** | | | | |
| **Age** | 1.01 | 1.01 | 0.9897 | 0.0103 |
| **Family history of BC** | 1.00 | 1.00 | 0.9992 | 0.0008 |
| **Menarche age** | 1.02 | 1.01 | 0.9791 | 0.0209 |
| **Standing height** | 1.02 | 1.01 | 0.9760 | 0.0240 |
| **Physical activity** | 1.02 | 1.01 | 0.9815 | 0.0185 |
| **BMI** | 1.05 | 1.02 | 0.9545 | 0.0455 |
| **Mean VIF** | 1.02 | | | |
| **Post-menopausal variables** | | | | |
| **Age** | 1.10 | 1.05 | 0.9111 | 0.0889 |
| **Family history of BC** | 1.00 | 1.00 | 0.9979 | 0.0021 |
| **Number of live** | 1.04 | 1.02 | 0.9588 | 0.0412 |

| **Variable** | **VIF** | **SQRT VIF** | **Tolerance** | **r-squared** |
|---|---|---|---|---|
| **births** | | | | |
| **Standing height** | 1.02 | 1.01 | 0.9789 | 0.0211 |
| **HRT use** | 1.02 | 1.01 | 0.9796 | 0.0204 |
| **BMI** | 1.06 | 1.03 | 0.9469 | 0.0531 |
| **Mean VIF** | 1.04 | | | |

# Chapter 8 : Conclusions and future work

## Introduction

This final chapter summarises work presented in this thesis. Six aims were listed at the beginning of the project and summaries of the key findings for each aim are presented in this chapter. These aims were:

(1) generating a systematic review of BC risk prediction models

(2) summarising the established risk factors for BC

(3) assessing the reproductive, lifestyle, anthropometric and dietary risk factors associated with the development of BC in UK females

(4) assessing the effect of lifestyle on different levels of genetic predisposition,

(5) developing an epidemiological risk prediction model for BC

(6) developing an extended epidemiological genetic risk prediction model by incorporating the PRS into BC prediction models

Implications of the presented work are discussed and evaluated and additionally the strengths and limitations of the work are considered. Based on the limitations of this study and those in the literature a number of recommendations are suggested for future work.

## 8.1 Summary of the key findings

### Aim 1: Systematic review of BC risk prediction models

Breast cancer risk prediction models are risk assessment tools allowing women to assess their personal risk of developing BC based on the combination of BC associated factors. Several BC models have been developed in the last 30 years, however, there have been limited reports on their performance and the majority have not been evaluated externally by using an independent cohort. Moreover, most models were not user-friendly and needed the input of clinical results which were not available to the participants themselves. Therefore, it was decided to develop a more user-friendly model based on the input of modifiable epidemiological risk factors so females could assess their risk and subsequently act accordingly. These models can help females to be aware of possible changes they may choose to adopt for a healthier lifestyle in order to decrease their BC risk. In total, 14 BC risk models were identified that were based around: age, ethnicity, height, weight, BMI, alcohol intake, smoking, physical activity, diet, age at menarche, age at menopause, age at first live birth, age at subsequent birth, HRT use, OC use, breastfeeding, pregnancy, parity, children number, menopause type, menstrual regularity, menstrual duration, menopausal status, gestation period, BC family history, age of onset of BC in relatives, benign breast disease, history of breast biopsies, and mammogram results. Not all of these factors were present in all reviewed models, rather a combination of these factors was present in each one of them and no two models had the same combination of risk factors. The above-mentioned risk factors were also a combination of  significant, probable and possible risk factors for BC [345]. Six out of 14 reviewed models validated their models externally, only one reported the utility performance and one reported the accuracy measures (sensitivity, specificity, positive predictive value, and negative predictive value). The best level of discrimination power achieved was 0.65 (divided as before and after 50 years of age) in a Korean population and the worst one for internal validation was 0.61 in a study of Asian and Pacific Islander American women.

The reviewed models' performance appeared to be only moderately good in differentiating between BC cases and non-cases although they may still serve as a good educational tool as part of cancer prevention.  They can be used to assess the public's knowledge of BC risk factors and to promote cancer risk reduction actions. Some of the models reviewed cannot be applied to other populations as the prevalence of risk factors vary considerably between different populations. It is recommended that researchers develop a more reliable and valid BC risk model which has good calibration, accuracy, discrimination, and wide utility where both internal and external validation indicates that it can be reliable for general use. It was confirmed that a literature review was critically required for the valid and most significant BC risk factors before developing a new BC risk prediction model.

**Aim 2: Summarising the established risk factors of BC**

An extensive literature review was performed to identify BC risk factors for potential inclusion into the risk prediction models. The results of this review are presented in chapter 2. All evidence was categorised into three levels (significant, probable and possible) based on the Harvard report evaluation criteria [42]. The included studies were MR studies, systematic review meta-analysis, prospective studies, and case-control studies. Nevertheless, bias, confounders, and reverse causation could affect the results of all mentioned studies, but it is less likely to be present in MR studies.

After evaluating the included studies, risk factors that showed significant evidence were genetic factors, family history of BC, radiation, height, physical activity, and BMI. Probable risk factors were age, smoking, OC use, HRT use, diet, early menarche age, late menopause age, benign breast diseases, breast density, null-parity, age at first birth, alcohol, and breastfeeding. Lastly, spontaneous miscarriage or induced abortion was the possible risk factor. Moreover, the magnitude (value) and the direction (increase or decrease) of the risk factor on BC varies depending on menopausal status. The most obvious evidence is BMI were being obese can protect premenopausal females against BC  [45] while being obese can increase the risk of BC among post-menopausal females [74].

These identified factors were set as priority factors.  Further work was carried out involved the explanatory analysis of these factors using the UKB female cohort.

**Aim 3: Assessing reproductive, lifestyle, anthropometric and diet risk factors associated with developing BC amongst UK females**

The results relating to aim 3 are presented fully in chapter 5.  In the pre-menopausal group, older age, height (> 169 cm) , low BMI (between 18.5-24.9) , low waist to hip ratio (<=0.80)

, positive first-degree family history of BC, early menarche age, nulliparity, late age at first live birth (> 25 years old) , high reproductive interval index (index > 16) and long duration of contraceptive use were all significantly associated with an increased BC risk. In the post-menopausal group, older age, being taller, having high BMI (BMI>=30), first degree BC family history, nulliparous, and high reproductive interval index (when no children reported) were all significantly associated with an increased risk of BC. The population attributable fraction (PAF) suggested that an early first live birth, lower reproductive interval index and increased number of children can all contribute to BC risk reduction by up to 50%. Many of these risk factors are modifiable and females can act upon changing them. The significant risk factors were used to develop a new risk prediction model for BC.

Further evidence from the literature review (diet, alcohol intake, smoking, physical activity, and family history of BC) were assessed against BC risk using data from the UKBiobank cohort. The BC model was developed using all available risk factors related to BC within UK Biobank. The models were divided on the basis of menopausal status as different risk factors were significant in pre-and post-menopausal women. Some risk factors were significant (e.g. smoking, alcohol intake and beef consumption among post-menopausal females) although they did not fit into the final regression model by using either stepwise or bootstrap regression approaches. Accordingly, they were removed from the final model developed.

**Aim 4: Assess the effect of lifestyle within groups based on different genetic predisposition**

In chapter 6 females were stratified into three groups based on their lifestyle: favourable, intermediate, and un-favourable lifestyles. A “healthy lifestyle” was defined based on five factors based on the Cancer Research UK guidelines (healthy weight, regular physical activity, limited alcohol intake, no contraceptive use and no or limited HRT use). Females practising an unhealthy lifestyle (intermediate and un-favourable) were at higher risk of developing BC compared healthy (favourable lifestyle) females.

The availability of genome-wide genotyping data for all participants in the UK Biobank cohort made an analysis of genetic predisposition possible. Three hundred and five BC-associated SNPs reported in the literature were used to calculate a polygenic risk score (PRS) and their aggregated score was tested against the lifestyle categories listed above. Furthermore, the PRS were grouped into three tertile groups of women based on their genetic risk scores (low, intermediate, and high). Females in the high and intermediate groups were at higher risk of BC compared to females with the lowest PRS. The influence of the three lifestyles in each genetic group was compared to test whether adhering to a healthy lifestyle could reduce the BC risk in each genetic risk strata. The findings suggested that females with intermediate and un-favourable lifestyle were at higher risk of BC compared to females with favourable lifestyle across all genetic groups.

This work highlights the potential benefit of healthy lifestyle regardless of a woman's genetic predisposition.  This screening programme provides a clear and positive message for cancer prevention; therefore, messages to promote healthy lifestyle can be offered to women who attend BC screening.

**Aim 5: Develop an epidemiological risk prediction model for BC**

A further goal of this study was to build a breast cancer risk prediction model which could ultimately be converted into a website for the public to assess their personalized BC risk based on their lifestyle. The results of chapters 2 and 5 were used to develop, two models based on pre-and post-menopausal status. In the pre-model, the factors included were age, family history, menarche age, height, physical activity and BMI. In the post-model, the factors included were age, family history, height, number of live births, HRT use, alcohol intake and BMI. Their overall performance was tested, and both showed good levels of calibration with modest discrimination. The premenopausal-AUC was 0.584 and the postmenopausal-AUC was 0.580. Implication of these models are discussed later in this chapter.

**Aim 6: Develop an extended epidemiological genetic risk prediction model by incorporating the PRS into the BC prediction models**

The generated PRS (from aim 4) was incorporated in these two models and their performance tested. The PRS supplemented models improved the level of discrimination power on both models resulting in a pre-AUC of 0.664 and a post-AUC of 0.647.

## 8.2 Implication of the developed models

The initial models were developed using modifiable risk factors without any clinical or genetic risk factors to facilitate wider usage among females. Two epidemiological models were developed based on female menopausal status (pre- and post-menopausal). The performance of the newly developed models (discrimination power of no more than 58% for both pre- and post-menopausal models) is close to the performance of the original Gail model of 56% [219]. Nevertheless, other epidemiological BC models reviewed in chapter 4 had better internal discrimination ranging from 56% to 65%. The Korean KoBCRAT model had the best discrimination power of 65% [216] which is better than other models developed. The Koran model included seven significant, three probable and three possible BC risk factors. The high discriminatory power in the KoBCRAT model as suggested by the study researchers, was due to the population characteristics used to build the original model.

One of the primary reasons for developing the BC risk prediction models was either to replace or to improve the effectiveness of the national mammogram screening programmes. However, until now there is no BC prediction model achieved that goal. Our model as well is not suitable for such mission due to high uncertainty and low discrimination power. Another reason for low discrimination power is no clinical risk factors (breast density, hyperplasia, benign breast diseases) were included in the model. As shown in chapter 2 (literature review) these factors have an impact on the BC development. Additionally, especially with large sample size is it likely to include statistically significant variables with no clinical impotence [27]. Moreover, no BC subtypes were considered at model development stage. It is well known that different BC subtypes have different genetic markers [359] and different risk factors [27] and these differences might affect the discrimination power of the developed models. Also, no consideration for oestrogen and progesterone receptor status could affected the performance of our model and its applicability.

Even though, our model discrimination powers were improved by including the polygenic risk scores of 305 pre-identified SNPs, yet it did not improve the applicability of the models. The AUC and C-statistics are hardly increased even with very significant BC predictors (Use and misuse of the receiver operating characteristic curve in risk prediction). According to systematic review published in 2019 [38] , the highest AUC was 0.71 reported by Eriksson *et al* [37], even with this improved power still it is not suitable to replace screening.

For the sake of clinical use of the model, it is a prerequisite to validate the model externally and internally and have at least AUC of 0.7 [360] and our models does not fulfil these criteria as the AUC < 0.7.

However, the model is intended to be used to increase awareness of the possible BC risk factors and encourage women to adopt healthier lifestyle. Although low in discriminatory power, these models can still be used as educational tool for the most significant risk factors associated with developing BC among UK females (using UKBiobank dataset). It can estimate the 5-year, 10-year, 20-year, and lifetime BC absolute risk of BC up to 90 years. Another benefit of the developed models could be to improve future BC models targeting UK population. An extensive risk factors analysis based on menopausal status has been conducted using the national wide UK biobank which could be beneficial for other research groups.

A future goal of this work is to make the model available for UK females through web-based access so it can be widely used by anyone. The intention was to follow the success of the BrightPink (https://brightpink.org/) web-page assessment tool for breast and ovarian cancer which provides an individual risk, thus helping females to manage their own health. By the end of the PhD project, initial development of the web-page was initiated by working with a web designer and the work will be continued beyond the duration of this Ph.D. The algorithm of risk assessment was provided to an IT technical team to build a website (RiskWomen). The prototype website was constructed however prior to implementation, more work is now required including language/comprehension testing using a message frame to interpret the risk and then further pilot studies, and focus groups for final development and implementation.

It is also worth noting that following access to UKBiobank genotyping data, a further model was developed by incorporating the genetic information into the BC prediction models. The PRS of 305 SNPs was added into the two epidemiological models. The discrimination power was subsequently improved to 66% among pre-model and 65% among post-menopausal model. The two extended models were internally validated, however, no external validation was performed due to the lack of an independent cohort with both genotype and epidemiological data. If such dataset becomes available, more work can be done beyond this PhD project to validate these models. Without external validation of these models, they cannot be recommended for the screening purposes. The developed four models are, however, the first population-based BC risk prediction models specific for UK Caucasian females.

### 8.3 Strengths and limitations

The strengths of the project will be presented for each paper separately.

For the first paper (risk models review), the strength is that this is the first published review summarising the non-clinical and non-genetic BC risk prediction.

For the second paper (risk factors of BC), the strengths are: (1) the use of large nation-wide prospective population-based cohort (UK Biobank). The UKBB provides high quality phenotypic exposures and genotyping data. Information on the most important environmental, reproductive, genetic, dietary, and behavioural risk factors were available for use in this study. (2) A prospective study design allowed the assessment of these exposures prior to BC development. The exposures were less subject to recall bias. (3) The investigation of the effect of the anthropometric and reproductive risk factors on BC development in the UK Biobank female cohort was the first study to report on such factors in the UK females. (4) The reproductive interval index is a new measure and is reported for the first time in this published article. (5) Estimation of the general PAF and the PAF of the subgroups for BC in the UK Biobank female cohort is presented as the attributable risk; this can inform public health/health professionals and promote lifestyle improvements in order to reduce BC incidence.

For the third paper (association of non-genetic factors with breast cancer risk) the strengths of the study are: (1) assessing together the lifestyle using modifiable risk factors in different genetically predisposed females, the study confirmed that adopting healthier lifestyle can reduce the risk of BC even among females with high genetic profile. (2) This is the first published study to investigate the combination of genetic and modifiable risk effects in UK females.

For the fourth paper (model development) its strengths are: the epidemiological models developed can potentially be used as an educational tool to raise awareness for BC prevention. The study also raises the possibility of incorporating the genetic PRS into models to enhance performance.

Limitations are reported according to the order of the published papers.

For the first paper (risk models review) the limitations are: (1) Absence of unified standards in defining model's performance (for example, what are acceptable measures of calibration and discrimination in preventive/ diagnostic / prognostic models? What is the utility cut-off in each type of model?). (2) Some reviewed model's results cannot be generalised to other populations as risk factor prevalence differ between different populations.

For the second paper (risk factors of breast cancer), limitations are: (1) The UK Biobank cohort itself is not fully representative of UK female population. There is evidence of UK Biobank participants being "healthier, better-educated, living in less deprived areas". This effect could underestimate or overestimate the real effect estimate of the exposure. (2) Volunteer selection bias is known to be present among UK Biobank [361] where participants are more health-conscious non-participants. This might lead to less precise generalizable prevalence or incidence rates of the outcome of interest [362]. To overcome this volunteer effect and to produce generalizable associations of exposures, sufficiently large sample size of participants with different level of exposures are needed in cohort studies [363, 364]. In conclusion, even though UKBiobank is not fully representative for the UK population and is not the best cohort to estimate the prevalence and incidence rates, it is a large enough cohort to estimate a reliable association between exposures and the interested outcome [351]. (3) Another possible limitation is the lack of some information relating to BC such as breastfeeding history, family history of ovarian cancer, BC onset of the family members and BC subtype (PR+, ER+, HER2+, and triple negative). As some of these missing variables are considered as significant risk factors for BC (family history of ovarian cancer) and some of them are needed for stratification (BC subtypes). These missing variables limits further data analysis.

For the third paper (association of non-genetic factors with breast cancer risk), limitations are: (1) The PRS used is restricted to Caucasian females only and cannot be generalised to other ethnic groups. (2) Another limitation is that the analysis was restricted to postmenopausal women; therefore, these results cannot be applied to premenopausal women. (3) Reverse causation could be an issue as lag measurement of two years was not taken into consideration in defining the incident cases of BC (4) deceased females were not included in the analysis which could lead to biased results of the interested estimate (5) the scoring method might be improved by assigned a weighted score for healthy lifestyle factors by using β coefficients of each lifestyle factor derived from the Cox proportional hazards regression model (6) the study did not assess the formal interaction between lifestyle and genetic factors due to insufficient study power, (7) while the study was satisfactorily powered to detect large effects for rarer exposures (<5%), more modest effects (< 30%) could be un-detected.

In the fourth paper (model development), the limitations are that the healthy volunteer effect of UKBiobank participants, the lack of some important information of BC risk factors such as breastfeeding history, and the developed models target Caucasian females only. Thus, it cannot be generalised to other ethnic groups or populations.

### 8.4 Recommendations and future work

- Pilot studies around risk communication are needed before making these models available to the public.
- External validation for the genetic models is needed to assess the model’s performance.
- Another cohort from the UK is ideally needed as external validation cohort to assess both epidemiological and genetic models.
- Another approach would be testing the causality of the BC risk factors using mendelian randomisation and building a more refined model(s).
- Assessment of the utility and clinical usefulness of the developed models is required.
- Clinical risk factors could be added to improve the model’s performance and widen its target audience.

# References

1. Parks, R.M., et al., *Breast Cancer Epidemiology*, in *Breast Cancer Management for Surgeons: A European Multidisciplinary Textbook*, L. Wyld, et al., Editors. 2018, Springer International Publishing: Cham. p. 19-29.
2. Ferlay, J., et al., *Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods.* International Journal of Cancer, 2019. **144**(8): p. 1941-1953.
3. Winters, S., et al., *Chapter One - Breast Cancer Epidemiology, Prevention, and Screening*, in *Progress in Molecular Biology and Translational Science*, R. Lakshmanaswamy, Editor. 2017, Academic Press. p. 1-32.
4. Ghoncheh, M., Z. Pournamdar, and H. Salehiniya, *Incidence and Mortality and Epidemiology of Breast Cancer in the World.* Asian Pacific Journal of Cancer Prevention, 2016. **17**(S3): p. 43-46.
5. Lee, A., et al., *BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors.* Genet Med, 2019. **21**(8): p. 1708-1718.
6. Lundqvist, A., et al., *Socioeconomic inequalities in breast cancer incidence and mortality in Europe—a systematic review and meta-analysis.* European Journal of Public Health, 2016. **26**(5): p. 804-813.
7. UK, C.R. *Breast cancer incidence (invasive) statistics*. 2018 20 June 2019; Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/incidence-invasive#heading-One.
8. UK, C.R., *Breast cancer statistics.* 2018.
9. McPherson, K., C.M. Steel, and J.M. Dixon, *Breast cancer—epidemiology, risk factors, and genetics.* BMJ : British Medical Journal, 2000. **321**(7261): p. 624-628.
10. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2019.* CA: A Cancer Journal for Clinicians, 2019. **69**(1): p. 7-34.
11. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA: A Cancer Journal for Clinicians, 2018. **68**(6): p. 394-424.
12. *Screening for breast cancer in England: past and future.* Journal of Medical Screening, 2006. **13**(2): p. 59-61.
13. Li, C.I., *Breast Cancer Epidemiology*. 2010: Springer. 1-403.
14. Marmot, M.G., et al., *The benefits and harms of breast cancer screening: an independent review.* British Journal of Cancer, 2013. **108**(11): p. 2205-2240.
15. Løberg, M., et al., *Benefits and harms of mammography screening.* Breast Cancer Research, 2015. **17**(1): p. 63.
16. Darby, S.C., et al., *Risk of Ischemic Heart Disease in Women after Radiotherapy for Breast Cancer.* 2013. **368**(11): p. 987-998.
17. Yeh, E.T.H. and C.L. Bickford, *Cardiovascular Complications of Cancer Therapy.* 2009. **53**(24): p. 2231-2247.
18. Brodersen, J. and V.D. Siersma, *Long-Term Psychosocial Consequences of False-Positive Screening Mammography.* 2013. **11**(2): p. 106-115.
19. Kalager, M., et al., *Prognosis in women with interval breast cancer: population based observational cohort study.* 2012. **345**: p. e7536.
20. Fred A. Mettler, J., et al., *Radiologic and Nuclear Medicine Studies in the United States and Worldwide: Frequency, Radiation Dose, and Comparison with Other Radiation Sources—1950–2007.* 2009. **253**(2): p. 520-531.
21. Yaffe, M.J. and J.G. Mainprize, *Risk of Radiation-induced Breast Cancer from Mammographic Screening.* 2011. **258**(1): p. 98-105.
22. Meads, C., I. Ahmed, and R. Riley, *A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance.* Breast Cancer Res Treat, 2012. **132**(2): p. 365-377.

23. Kannel, W.B., D. McGee, and T. Gordon, *A general cardiovascular risk profile: the Framingham Study.* Am J Cardiol, 1976. **38**(1): p. 46-51.
24. Freedman, A.N., et al., *Cancer Risk Prediction Models: A Workshop on Development, Evaluation, and Application.* Journal of the National Cancer Institute, 2005. **97**(10): p. 715-723.
25. Cintolo-Gonzalez, J.A., et al., *Breast cancer risk models: a comprehensive overview of existing models, validation, and clinical applications.* Breast Cancer Research and Treatment, 2017. **164**(2): p. 263-284.
26. Engel, C. and C. Fischer, *Breast Cancer Risks and Risk Prediction Models.* Breast Care, 2015. **10**(1): p. 7-12.
27. Anothaisintawee, T., et al., *Risk prediction models of breast cancer: a systematic review of model performances.* Breast Cancer Res Treat, 2012. **133**(1): p. 1-10.
28. Antoniou, A.C., et al., *The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions.* 2008(1532-1827 (Electronic)).
29. Steyerberg, E.W., et al., *Assessing the performance of prediction models: a framework for some traditional and novel measures.* Epidemiology, 2010. **21**(1): p. 128-138.
30. Win, A.K., et al., *Risk prediction models for colorectal cancer: a review.* Cancer Epidemiology Biomarkers & Prevention, 2012. **21**(3): p. 398-410.
31. Moons, K.G.M., et al., *Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker.* 2012. **98**(9): p. 683-690.
32. Gail, M.H., et al., *Projecting Individualized Probabilities of Developing Breast-Cancer for White Females Who Are Being Examined Annually.* Journal of the National Cancer Institute, 1989. **81**(24): p. 1879-1886.
33. Costantino, J.P., et al., *Validation studies for models projecting the risk of invasive and total breast cancer incidence.* J Natl Cancer Inst, 1999. **91**(18): p. 1541-8.
34. Gail, M.H., et al., *Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer.* J Natl Cancer Inst, 1999. **91**(21): p. 1829-46.
35. Freedman, A.N., et al., *Estimates of the number of US women who could benefit from tamoxifen for breast cancer chemoprevention.* J Natl Cancer Inst, 2003. **95**(7): p. 526-32.
36. Fund, W.C.R., *Diet, nutrition, physical activity and breast cancer* 2017.
37. Eriksson, M., et al., *A clinical model for identifying the short-term risk of breast cancer.* Breast Cancer Res, 2017. **19**(1): p. 29.
38. Louro, J., et al., *A systematic review and quality assessment of individualised breast cancer risk prediction models.* British Journal of Cancer, 2019. **121**(1): p. 76-85.
39. Sudlow, C., 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**(3): p. e1001779.
40. *About UK Biobank*. 2017; Available from: http://www.ukbiobank.ac.uk.
41. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.
42. Colditz, G.A., 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**(6): p. 477-88.
43. Samet, J.M., et al., *The IARC Monographs: Updated Procedures for Modern and Transparent Evidence Synthesis in Cancer Hazard Identification.* Journal of the National Cancer Institute, 2020. **112**(1): p. 30-37.
44. WCRF, *Judging the evidence* 2018.
45. Chan, D.S.M., et al., *World Cancer Research Fund International: Continuous Update Project—systematic literature review and meta-analysis of observational cohort studies on physical activity, sedentary behavior, adiposity, and weight change and breast cancer risk.* Cancer Causes & Control, 2019. **30**(11): p. 1183-1200.
46. Guo, W., et al., *Physical activity and breast cancer risk: results from the UK Biobank prospective cohort.* Br J Cancer, 2020. **122**(5): p. 726-732.
47. Monninkhof, E.M., et al., *Physical Activity and Breast Cancer: A Systematic Review.* Epidemiology, 2007. **18**(1): p. 137-157.

48. Papadimitriou, N., et al., *Physical activity and risks of breast and colorectal cancer: a Mendelian randomisation analysis.* Nature communications, 2020. **11**(1): p. 597-597.

49. McTiernan, A., et al., *Recreational physical activity and the risk of breast cancer in postmenopausal women: The women&#39;s health initiative cohort study.* JAMA, 2003. **290**(10): p. 1331-1336.

50. Wiseman, M., *The second World Cancer Research Fund/American Institute for Cancer Research expert report. Food, nutrition, physical activity, and the prevention of cancer: a global perspective.* Proc Nutr Soc, 2008. **67**(3): p. 253-6.

51. Monninkhof, E.M., et al., *Physical activity and breast cancer: a systematic review.* Epidemiology, 2007. **18**(1): p. 137-57.

52. Humans, I.W.G.o.t.E.o.C.R.t., *Alcohol consumption and ethyl carbamate.* IARC monographs on the evaluation of carcinogenic risks to humans, 2010. **96**: p. 3-1383.

53. Chen, W.Y., et al., *Moderate Alcohol Consumption During Adult Life, Drinking Patterns, and Breast Cancer Risk.* JAMA, 2011. **306**(17): p. 1884-1890.

54. Alimujiang, A. and G.A. Colditz, *What can we learn from the association between adolescent alcohol consumption and breast cancer risk?* Expert Review of Anticancer Therapy, 2019. **19**(4): p. 287-289.

55. Collaborative Group on Hormonal Factors in Breast Cancer, *Alcohol, tobacco and breast cancer – collaborative reanalysis of individual data from 53 epidemiological studies, including 58 515 women with breast cancer and 95 067 women without the disease.* British Journal of Cancer, 2002. **87**(11): p. 1234-1245.

56. Petri, A.L., et al., *Alcohol intake, type of beverage, and risk of breast cancer in pre- and postmenopausal women.* Alcohol Clin Exp Res, 2004. **28**(7): p. 1084-90.

57. Allen, N.E., et al., *Moderate Alcohol Intake and Cancer Incidence in Women.* Journal of the National Cancer Institute, 2009. **101**(5): p. 296-305.

58. White, A.J., et al., *Lifetime Alcohol Intake, Binge Drinking Behaviors, and Breast Cancer Risk.* American Journal of Epidemiology, 2017. **186**(5): p. 541-549.

59. Seitz, H.K., et al., *Epidemiology and Pathophysiology of Alcohol and Breast Cancer: Update 2012.* Alcohol and Alcoholism, 2012. **47**(3): p. 204-212.

60. Liu, Y., N. Nguyen, and G.A. Colditz, *Links between Alcohol Consumption and Breast Cancer: A Look at the Evidence.* Women's Health, 2015. **11**(1): p. 65-77.

61. Ingold, N., H.A. Amin, and F. Drenos, *Alcohol causes an increased risk of head and neck but not breast cancer in individuals from the UK Biobank study: A Mendelian randomisation analysis.* medRxiv, 2019: p. 19002832.

62. Goldvaser, H., et al., *The association between smoking and breast cancer characteristics and outcome.* BMC cancer, 2017. **17**(1): p. 624-624.

63. Secretan, B., et al., *A review of human carcinogens--Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish.* Lancet Oncol, 2009. **10**(11): p. 1033-4.

64. Johnson, K.C., et al., *Active smoking and secondhand smoke increase breast cancer risk: the report of the Canadian Expert Panel on Tobacco Smoke and Breast Cancer Risk (2009).* Tob Control, 2011. **20**(1): p. e2.

65. Alberg, A.J., D.R. Shopland, and K.M. Cummings, *The 2014 Surgeon General's report: commemorating the 50th Anniversary of the 1964 Report of the Advisory Committee to the US Surgeon General and updating the evidence on the health consequences of cigarette smoking.* Am J Epidemiol, 2014. **179**(4): p. 403-12.

66. Johnson, K.C., et al., *Active smoking and secondhand smoke increase breast cancer risk: the report of the Canadian Expert Panel on Tobacco Smoke and Breast Cancer Risk (2009).* Tobacco Control, 2011. **20**(1): p. e2.

67. Kabat, G.C., et al., *Smoking and alcohol consumption in relation to risk of triple-negative breast cancer in a cohort of postmenopausal women.* Cancer Causes Control, 2011. **22**(5): p. 775-83.

68. Kawai, M., et al., *Active smoking and the risk of estrogen receptor-positive and triple-negative breast cancer among women ages 20 to 44 years.* Cancer, 2014. **120**(7): p. 1026-34.

69. Wang, K., et al., *Smoking increases risks of all-cause and breast cancer specific mortality in breast cancer individuals: a dose-response meta-analysis of prospective cohort studies involving 39725 breast cancer cases.* Oncotarget, 2016. **7**(50): p. 83134-83147.

70. Kropp, S. and J. Chang-Claude, *Active and Passive Smoking and Risk of Breast Cancer by Age 50 Years among German Women.* American Journal of Epidemiology, 2002. **156**(7): p. 616-626.

71. Lawlor, D.A., G.D. Ebrahim S Fau - Smith, and G.D. Smith, *Smoking before the birth of a first child is not associated with increased risk of breast cancer: findings from the British Women's Heart and Health Cohort Study and a meta-analysis.* 2004(0007-0920 (Print)).

72. Andersen, Z.J., et al., *Active smoking and risk of breast cancer in a Danish nurse cohort study.* BMC Cancer, 2017. **17**(1): p. 556.

73. Elwood, P.C., et al., *Healthy living and cancer: evidence from UK Biobank.* Ecancermedicalscience, 2018. **12**: p. 792-792.

74. Al-Ajmi, K., et al., *Risk of breast cancer in the UK biobank female cohort and its relationship to anthropometric and reproductive factors.* Plos One, 2018. **13**(7).

75. Weiderpass, E., et al., *A Prospective Study of Body Size in Different Periods of Life and Risk of Premenopausal Breast Cancer.* Cancer Epidemiology Biomarkers &amp; Prevention, 2004. **13**(7): p. 1121-1127.

76. Cleary, M.P. and M.E. Grossmann, *Obesity and Breast Cancer: The Estrogen Connection.* Endocrinology, 2009. **150**(6): p. 2537-2542.

77. Ooi, B.N.S., et al., *The genetic interplay between body mass index, breast size and breast cancer risk: a Mendelian randomization analysis.* International Journal of Epidemiology, 2019. **48**(3): p. 781-794.

78. Johnson, K.E., et al., *Assessing a causal relationship between circulating lipids and breast cancer risk via Mendelian randomization.* bioRxiv, 2019: p. 794594.

79. Mørch, L.S., et al., *Contemporary Hormonal Contraception and the Risk of Breast Cancer.* New England Journal of Medicine, 2017. **377**(23): p. 2228-2239.

80. Hunter, D.J., et al., *Oral Contraceptive Use and Breast Cancer: A Prospective Study of Young Women.* Cancer Epidemiology Biomarkers &amp;amp; Prevention, 2010. **19**(10): p. 2496.

81. Collaborative Group on Hormonal Factors in Breast Cancer, *Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies.* 1996(0140-6736 (Print)).

82. Collaborative Group on Hormonal Factors in Breast Cancer, *Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52 705 women with breast cancer and 108 411 women without breast cancer.* The Lancet, 1997. **350**(9084): p. 1047-1059.

83. Beral, V. and C. Million Women Study, *Breast cancer and hormone-replacement therapy in the Million Women Study.* Lancet, 2003. **362**(9382): p. 419-27.

84. Beral, V., *Breast cancer and hormone-replacement therapy in the Million Women Study.* 2003(1474-547X (Electronic)).

85. Fahlen, M., et al., *Hormone replacement therapy after breast cancer: 10 year follow up of the Stockholm randomised trial.* Eur J Cancer, 2013. **49**(1): p. 52-9.

86. Jung, S., et al., *Fruit and vegetable intake and risk of breast cancer by hormone receptor status.* J Natl Cancer Inst, 2013. **105**(3): p. 219-36.

87. Papadimitriou, N., et al., *Circulating concentrations of micro-nutrients and risk of breast cancer: A Mendelian randomization study.* bioRxiv, 2019: p. 668186.

88. Dong, J.Y., et al., *Dairy consumption and risk of breast cancer: a meta-analysis of prospective cohort studies.* Breast Cancer Res Treat, 2011. **127**(1): p. 23-31.

89. zur Hausen, H., *Red meat consumption and cancer: Reasons to suspect involvement of bovine infectious factors in colorectal cancer.* 2012. **130**(11): p. 2475-2483.

90. Farvid, M.S., et al., *Consumption of red and processed meat and breast cancer incidence: A systematic review and meta-analysis of prospective studies.* 2018. **143**(11): p. 2787-2799.

91. Anderson, J.J., et al., *Red and processed meat consumption and breast cancer: UK Biobank cohort study and meta-analysis.* European Journal of Cancer, 2018. **90**: p. 73-82.

92. Taylor, V.H., M. Misra, and S.D. Mukherjee, *Is red meat intake a risk factor for breast cancer among premenopausal women?* Breast Cancer Res Treat, 2009. **117**(1): p. 1-8.

93. Pala, V., et al., *Meat, eggs, dairy products, and risk of breast cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort.* Am J Clin Nutr, 2009. **90**(3): p. 602-12.

94. Fu, Z., et al., *Well-done meat intake and meat-derived mutagen exposures in relation to breast cancer risk: the Nashville Breast Health Study.* Breast cancer research and treatment, 2011. **129**(3): p. 919-928.

95. Colditz, G.A. and B. Rosner, *Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study.* American journal of epidemiology, 2000. **152**(10): p. 950-964.

96. Kroman, N., et al., *Parity, age at first childbirth and the prognosis of primary breast cancer.* British Journal of Cancer, 1998. **78**(11): p. 1529-1533.

97. Lambertini, M., et al., *Reproductive behaviors and risk of developing breast cancer according to tumor subtype: A systematic review and meta-analysis of epidemiological studies.* Cancer Treatment Reviews, 2016. **49**: p. 65-76.

98. Layde, P., et al., *The independent associations of parity, age at first full term pregnancy, and.* J Clin Epidemiol. , 1989. **42**(0895-4356 (Print)): p. 63-73.

99. Ewertz, M., et al., *Age at first birth, parity and risk of breast cancer: a meta-analysis of 8 studies from the Nordic countries.* 1990(0020-7136 (Print)).

100. Peila, R., R. Arthur, and T.E. Rohan, *Risk factors for ductal carcinoma in situ of the breast in the UK Biobank cohort study.* Cancer Epidemiology, 2020. **64**: p. 101648.

101. Fortner, R.T., et al., *Parity, breastfeeding, and breast cancer risk by hormone receptor status and molecular phenotype: results from the Nurses' Health Studies.* Breast Cancer Research, 2019. **21**(1): p. 40.

102. Research, W.C.R.F.A.I.f.C., *Second Expert Report – Food, Nutrition, Physical Activity, and the Prevention of Cancer: A Global Perspective.* 2007.

103. Solans, M., et al., *A systematic review and meta-analysis of the 2007 WCRF/AICR score in relation to cancer-related health outcomes.* Annals of Oncology, 2020. **31**(3): p. 352-368.

104. Kohler, L.N., et al., *Adherence to Diet and Physical Activity Cancer Prevention Guidelines and Cancer Outcomes: A Systematic Review.* Cancer Epidemiology Biomarkers &amp;amp; Prevention, 2016. **25**(7): p. 1018.

105. Zhou, Y., et al., *Association Between Breastfeeding and Breast Cancer Risk: Evidence from a Meta-analysis.* Breastfeeding Medicine, 2015. **10**(3): p. 175-182.

106. Collaborative Group on Hormonal Factors in Breast Cancer, *Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease.* The Lancet, 2002. **360**(9328): p. 187-195.

107. Yang, L. and K.H. Jacobsen, *A Systematic Review of the Association between Breastfeeding and Breast Cancer.* Journal of Women's Health, 2008. **17**(10): p. 1635-1645.

108. Momenimovahed, Z. and H. Salehiniya, *Epidemiological characteristics of and risk factors for breast cancer in the world.* Breast cancer (Dove Medical Press), 2019. **11**: p. 151-164.

109. Lee, C., et al., *Computational Discrimination of Breast Cancer for Korean Women Based on Epidemiologic Data Only.* J Korean Med Sci, 2015. **30**(8): p. 1025-34.

110. Cancer Research UK, *Cancer Statistics :Breast Cancer* 2014, Cancer Research UK. p. 1-2.

111. Assi, H.A., et al., *Epidemiology and prognosis of breast cancer in young women.* J Thorac Dis, 2013. **5 Suppl 1**: p. S2-8.

112. Wendt, C. and S. Margolin, *Identifying breast cancer susceptibility genes – a review of the genetic background in familial breast cancer.* Acta Oncologica, 2019. **58**(2): p. 135-146.

113. Aloraifi, F., et al., *Gene analysis techniques and susceptibility gene discovery in non-BRCA1/BRCA2 familial breast cancer.* Surg Oncol, 2015. **24**(2): p. 100-9.

114. Godet, I. and D.M. Gilkes, *BRCA1 and BRCA2 mutations and treatment strategies for breast cancer.* Integr Cancer Sci Ther, 2017. **4**(1).

115. Kuchenbaecker, K.B., et al., *Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers.* Jama, 2017. **317**(23): p. 2402-2416.

116. Lalloo, F. and D.G. Evans, *Familial Breast Cancer.* 2012. **82**(2): p. 114.

117. Yari, K., et al., *The MMP-2 -735 C allele is a risk factor for susceptibility to breast cancer.* Asian Pac J Cancer Prev, 2014. **15**(15): p. 6199-203.

118. Skol, A.D., M.M. Sasaki, and K. Onel, *The genetics of breast cancer risk in the post-genome era: thoughts on study design to move past BRCA and towards clinical relevance.* Breast Cancer Research, 2016. **18**(1): p. 99.

119. Michailidou, K., et al., *Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer.* Nat Genet, 2015. **47**(4): p. 373-80.

120. Mavaddat, N., et al., *Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes.* American Journal of Human Genetics, 2019. **104**(1): p. 21-34.

121. Stratton, M.R. and N. Rahman, *The emerging landscape of breast cancer susceptibility.* Nature Genetics, 2008. **40**(1): p. 17-22.

122. Ghoussaini, M. and P.D. Pharoah, *Polygenic susceptibility to breast cancer: current state-of-the-art.* Future Oncol, 2009. **5**(5): p. 689-701.

123. Michailidou, K., et al., *Association analysis identifies 65 new breast cancer risk loci.* Nature, 2017. **551**(7678): p. 92-94.

124. Bravi, F., A. Decarli, and A.G. Russo, *Risk factors for breast cancer in a cohort of mammographic screening program: a nested case-control study within the FRiCaM study.* Cancer Med, 2018. **7**(5): p. 2145-2152.

125. Ahern, T.P., et al., *Family History of Breast Cancer, Breast Density, and Breast Cancer Risk in a U.S. Breast Cancer Screening Population.* Cancer Epidemiol Biomarkers Prev, 2017. **26**(6): p. 938-944.

126. Metcalfe, K.A., et al., *Breast cancer risks in women with a family history of breast or ovarian cancer who have tested negative for a BRCA1 or BRCA2 mutation.* Br J Cancer, 2009. **100**(2): p. 421-5.

127. Collaborative Group on Hormonal Factors in Breast Cancer, *Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies.* The Lancet Oncology, 2012. **13**(11): p. 1141-1151.

128. Clavel-Chapelon, F. and E.N.E. Group, *Differential effects of reproductive factors on the risk of pre- and postmenopausal breast cancer. Results from a large cohort of French women.* Br J Cancer, 2002. **86**(5): p. 723-7.

129. Laamiri, F.Z., et al., *Risk Factors for Breast Cancer of Different Age Groups: Moroccan Data?* Open Journal of Obstetrics and Gynecology, 2015. **Vol.05No.02**: p. 9.

130. Dai, Q., B. Liu, and Y. Du, *Meta-analysis of the risk factors of breast cancer concerning reproductive factors and oral contraceptive use.* Frontiers of Medicine in China, 2009. **3**(4): p. 452-458.

131. Kabat, G.C., et al., *A multi-center prospective cohort study of benign breast disease and risk of subsequent breast cancer.* Cancer Causes & Control, 2010. **21**(6): p. 821-828.

132. Hartmann, L.C., et al., *Benign breast disease and the risk of breast cancer.* N Engl J Med, 2005. **353**(3): p. 229-37.

133. Worsham, M.J., et al., *Risk factors for breast cancer from benign breast disease in a diverse population.* Breast Cancer Res Treat, 2009. **118**(1): p. 1-7.

134. Tice, J.A., et al., *Benign Breast Disease, Mammographic Breast Density, and the Risk of Breast Cancer.* JNCI: Journal of the National Cancer Institute, 2013. **105**(14): p. 1043-1049.

135. Hartmann, L.C., et al., *Benign Breast Disease and the Risk of Breast Cancer.* New England Journal of Medicine, 2005. **353**(3): p. 229-237.

136. Boyd, N.F., et al., *Mammographic breast density as an intermediate phenotype for breast cancer.* Lancet Oncol, 2005. **6**(10): p. 798-808.

137. Ginsburg, O.M., L.J. Martin, and N.F. Boyd, *Mammographic density, lobular involution, and risk of breast cancer.* Br J Cancer, 2008. **99**(9): p. 1369-74.

138. Tamimi, R.M., et al., *Endogenous hormone levels, mammographic density, and subsequent risk of breast cancer in postmenopausal women.* J Natl Cancer Inst, 2007. **99**(15): p. 1178-87.

139. Harvey, J.A. and V.E. Bovbjerg, *Quantitative assessment of mammographic breast density: relationship with breast cancer risk.* Radiology, 2004. **230**(1): p. 29-41.

140. Pinsky, R.W. and M.A. Helvie, *Mammographic breast density: effect on imaging and breast cancer risk.* J Natl Compr Canc Netw, 2010. **8**(10): p. 1157-64; quiz 1165.

141. McCormack, V.A. and I. dos Santos Silva, *Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis.* 2006(1055-9965 (Print)).

142. Vachon, C.M., et al., *Mammographic density, breast cancer risk and risk prediction.* Breast Cancer Res, 2007. **9**(6): p. 217-217.

143. Fund, W.C.R., *Height and birthweight and the risk of cancer* 2018.

144. Qian, F., et al., *Height and Body Mass Index as Modifiers of Breast Cancer Risk in BRCA1/2 Mutation Carriers: A Mendelian Randomization Study.* JNCI: Journal of the National Cancer Institute, 2018. **111**(4): p. 350-364.

145. Zhang, B., et al., *Height and Breast Cancer Risk: Evidence From Prospective Studies and Mendelian Randomization.* Journal of the National Cancer Institute, 2015. **107**(11): p. djv219.

146. Lahmann, P.H., et al., *Body size and breast cancer risk: Findings from the European prospective investigation into cancer and nutrition (EPIC).* 2004. **111**(5): p. 771.

147. Zhang, B., et al., *Height and Breast Cancer Risk: Evidence From Prospective Studies and Mendelian Randomization.* Journal of the National Cancer Institute, 2015. **107**(11 ).

148. Green, J., et al., *Height and cancer incidence in the Million Women Study: prospective cohort, and meta-analysis of prospective studies of height and total cancer risk.* Lancet Oncol, 2011. **12**(8): p. 785-94.

149. Van den Brandt, P.A., et al., *Pooled Analysis of Prospective Cohort Studies on Height, Weight, and Breast Cancer Risk.* American Journal of Epidemiology, 2000. **152**(6): p. 514-527.

150. Andersen Zj Fau - Baker, J.L., et al., *Birth weight, childhood body mass index, and height in relation to mammographic density and breast cancer: a register-based cohort study.* 2014(1465-542X (Electronic)).

151. Key, T.J., P.K. Verkasalo, and E. Banks, *Epidemiology of breast cancer.* The Lancet Oncology, 2001. **2**(3): p. 133-140.

152. *Breast cancer and abortion: collaborative reanalysis of data from 53 epidemiological studies, including 83 000 women with breast cancer from 16 countries.* The Lancet, 2004. **363**(9414): p. 1007-1016.

153. Guo, J., et al., *Association between abortion and breast cancer: an updated systematic review and meta-analysis based on prospective studies.* Cancer Causes & Control, 2015. **26**(6): p. 811-819.

154. Tong, H., et al., *No association between abortion and risk of breast cancer among nulliparous women: Evidence from a meta-analysis.* 2020. **99**(19): p. e20251.

155. Helm, J.S. and R.A. Rudel, *Adverse outcome pathways for ionizing radiation and breast cancer involve direct and indirect DNA damage, oxidative stress, inflammation, genomic instability, and interaction with hormonal regulation of the breast.* Archives of toxicology, 2020. **94**(5): p. 1511-1549.

156. Ozasa, K., et al., *Studies of the mortality of atomic bomb survivors, Report 14, 1950-2003: an overview of cancer and noncancer diseases.* Radiat Res, 2012. **177**(3): p. 229-43.

157. Eidemüller, M., et al., *Breast cancer risk and possible mechanisms of radiation-induced genomic instability in the Swedish hemangioma cohort after reanalyzed dosimetry.* Mutat Res, 2015. **775**: p. 1-9.

158. Little, M. and J. Boice Jr, *Comparison of breast cancer incidence in the Massachusetts tuberculosis fluoroscopy cohort and in the Japanese atomic bomb survivors.* Radiation research, 1999. **151**(2): p. 218-224.

159. Schulz, K.F. and D.A. Grimes, *Case-control studies: research in reverse.* The Lancet, 2002. **359**(9304): p. 431-434.

160. Kelsey, J.L., *Methods in observational epidemiology*. 2nd ed. Monographs in epidemiology and biostatistics. 1996, New York: Oxford University Press. viii, 432 p.

161. Mann, C.J., *Observational research methods. Research design II: cohort, cross sectional, and case-control studies.* Emergency Medicine Journal, 2003. **20**(1): p. 54.

162. Euser, A.M., et al., *Cohort Studies: Prospective versus Retrospective.* Nephron Clinical Practice, 2009. **113**(3): p. c214-c217.

163. Haidich, A.B., *Meta-analysis in medical research.* Hippokratia, 2010. **14**(Suppl 1): p. 29-37.

164. Oxman, A.D. and G.H. Guyatt, *The science of reviewing research.* Ann N Y Acad Sci, 1993. **703**: p. 125-33; discussion 133-4.

165. Antman, E.M., et al., *A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction.* Jama, 1992. **268**(2): p. 240-8.

166. Davies, N.M., M.V. Holmes, and G. Davey Smith, *Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians.* BMJ, 2018. **362**: p. k601.

167. Sheehan, N.A., et al., *Mendelian Randomisation and Causal Inference in Observational Epidemiology.* PLOS Medicine, 2008. **5**(8): p. e177.

168. Burgess, S., et al., *A robust and efficient method for Mendelian randomization with hundreds of genetic variants.* Nature Communications, 2020. **11**(1): p. 376.

169. Sheehan, N.A. and V. Didelez, *Epidemiology, genetic epidemiology and Mendelian randomisation: more need than ever to attend to detail.* Human Genetics, 2020. **139**(1): p. 121-136.

170. Liberati, A., et al., *The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration.* PLOS Medicine, 2009. **6**(7): p. e1000100.

171. Liberati, A., et al., *The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration.* BMJ, 2009. **339**.

172. Hewitt, J., et al., *Cohort profile of the UK Biobank: diagnosis and characteristics of cerebrovascular disease.* BMJ Open, 2016. **6**(3).

173. UKBiobank, *UK Biobank Malignant Cancer Summary Report*. 2020.

174. NHS. *Menopause* 2015; Available from: http://www.nhs.uk/conditions/menopause/Pages/Introduction.aspx.

175. Morris, D.H., et al., *Determinants of age at menarche in the UK: analyses from the Breakthrough Generations Study.* Br J Cancer, 2010. **103**(11): p. 1760-1764.

176. Newson, R., *PUNAF: Stata module to compute population attributable fractions for cohort studies*. 2012: Boston College Department of Economics.

177. StataCorp. [cited 2017 20 Sep]; Available from: https://www.stata.com/statamp/.

178. Dupont, W.D. and W.D. Plummer, Jr., *Power and sample size calculations. A review and computer program.* Control Clin Trials, 1990. **11**(2): p. 116-28.

179. CRUK, *Risk factors* 2017.

180. Khera, A.V., et al., *Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease.* New England Journal of Medicine, 2016. **375**(24): p. 2349-2358.

181. Lourida, I., et al., *Association of Lifestyle and Genetic Risk With Incidence of DementiaAssociation of Lifestyle and Genetic Risk With Incidence of DementiaAssociation of Lifestyle and Genetic Risk With Incidence of Dementia.* JAMA, 2019.

182. Corp., S. *ltable — Life tables for survival data*. Available from: https://www.stata.com/manuals13/stltable.pdf.

183. Microsoft. *Microsoft Excel*. Available from: https://products.office.com/en-gb/excel.

184. StataCorp. *StataMP* Available from: https://www.stata.com/statamp/.

185. Duncan, L., et al., *Analysis of polygenic risk score usage and performance in diverse human populations.* Nature Communications, 2019. **10**(1): p. 3328.

186. Marees, A.T., et al., *A tutorial on conducting genome-wide association studies: Quality control and statistical analysis.* 2018. **27**(2): p. e1608.

187. Coleman, J.R.I., et al., *Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray.* Briefings in functional genomics, 2016. **15**(4): p. 298-304.

188. Chang, C. 2019; Available from: https://www.cog-genomics.org/plink2.

189. Ali, A.M.G., et al., *Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer.* British Journal of Cancer, 2011. **104**(4): p. 693-699.

190. Bennett, D.A., *How can I deal with missing data in my study?* 2001. **25**(5): p. 464-469.

191. Miles, J., *Tolerance and Variance Inflation Factor*, in *Encyclopedia of Statistics in Behavioral Science*. 2005.

192. Horn, J., et al., *Reproductive factors and the risk of breast cancer in old age: a Norwegian cohort study.* 2013(1573-7217 (Electronic)).

193. WHO, *Obesity and overweight.* 2016.

194. WHO, *Waist circumference and waist–hip ratio*. 2008. p. 39.

195. UKBiobank. *Townsend deprivation index at recruitment*. [cited 2017 22 Sep]; Available from: http://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=189.

196. Aveyard, P., S. Manaseki, and J. Chambers, *The relationship between mean birth weight and poverty using the Townsend deprivation score and the Super Profile classification system.* Public Health, 2002. **116**(6): p. 308-314.

197. Riley, R.D., et al., *Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes.* 2019. **38**(7): p. 1276-1296.

198. Luque, M., Maringe, C and Nelson, P, *CVAUROC: Stata module to compute Cross-validated Area Under the Curve for ROC Analysis after Predictive Modelling for Binary Outcomes.* EconPapers, 2017.

199. Thrift, A.P., et al., *A model to determine absolute risk for esophageal adenocarcinoma.* Clin Gastroenterol Hepatol, 2013. **11**(2): p. 138-44 e2.

200. Jemal, A., et al., *Global cancer statistics.* CA: A Cancer Journal for Clinicians, 2011. **61**(2): p. 69-90.

201. Parkin, D.M. and L.M.G. Fernández, *Use of Statistics to Assess the Global Burden of Breast Cancer.* The Breast Journal, 2006. **12**: p. S70-S80.

202. Schreer, I. and J. Lüttges, *Breast cancer: early detection*, in *Radiologic-Pathologic Correlations from Head to Toe*. 2005, Springer. p. 767-784.

203. Tabar, L., et al., *Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening.* Lancet, 2003. **361**(9367): p. 1405-1410.

204. Gotzsche, P.C. and K.J. Jorgensen, *Screening for breast cancer with mammography.* Cochrane Database of Systematic Reviews, 2013(6).

205. Anderson, B.O., et al., *Early detection of breast cancer in countries with limited resources.* Breast J, 2003. **9 Suppl 2**: p. S51-9.

206. Yip, C.H., et al., *Guideline implementation for breast healthcare in low- and middle-income countries: early detection resource allocation.* Cancer, 2008. **113**(8 Suppl): p. 2244-56.

207. Li, J. and Z. Shao, *Mammography screening in less developed countries.* SpringerPlus, 2015. **4**: p. 615.

208. Institute, N.C. *The nation's investment in cancer research. A plan and budget proposal for the fiscal year 2006*. 2005 [cited 2018; Available from: https://www.cancer.gov/about-nci/budget/plan/.

209. Gerds, T.A., M. Cai T Fau - Schumacher, and M. Schumacher, *The performance of risk prediction models.* 2008(1521-4036 (Electronic)).

210. Anothaisintawee, T., et al., *Risk prediction models of breast cancer: a systematic review of model performances.* Breast Cancer Research and Treatment, 2012. **133**(1): p. 1-10.

211. Steyerberg, E.W., et al., *Assessing the performance of prediction models: a framework for traditional and novel measures.* Epidemiology, 2010. **21**(1): p. 128-38.

212. Moons, K.G., et al., *Prognosis and prognostic research: application and impact of prognostic models in clinical practice.* BMJ, 2009. **338**: p. b606.

213. Folkerd, E. and M. Dowsett, *Sex hormones and breast cancer risk and prognosis.* Breast, 2013. **22 Suppl 2**: p. S38-43.

214. Wright, C.E., et al., *Beliefs about weight and breast cancer: an interview study with high risk women following a 12 month weight loss intervention.* Hereditary Cancer in Clinical Practice, 2015. **13**(1): p. 1.

215. Eliassen, A.H., et al., *Adult weight change and risk of postmenopausal breast cancer.* JAMA, 2006. **296**.

216. Park, B., et al., *Korean risk assessment model for breast cancer risk prediction.* PLoS One, 2013. **8**(10): p. e76736.

217. Novotny, J., et al., *Breast cancer risk assessment in the Czech female population–an adjustment of the original Gail model.* Breast cancer research and treatment, 2006. **95**(1): p. 29-35.

218. Rosner, B. and G.A. Colditz, *Nurses' health study: log-incidence mathematical model of breast cancer incidence.* Journal of the National Cancer Institute, 1996. **88**(6): p. 359-364.

219. Gail, M.H., et al., *Projecting individualized probabilities of developing breast cancer for white females who are being examined annually.* J Natl Cancer Inst, 1989. **81**(24): p. 1879-86.

220. Rosner, B., G.A. Colditz, and W.C. Willett, *Reproductive risk factors in a prospective study of breast cancer: the Nurses' Health Study.* American journal of epidemiology, 1994. **139**(8): p. 819-835.

221. Ueda, K., et al., *Estimation of individualized probabilities of developing breast cancer for Japanese women.* Breast Cancer, 2003. **10**(1): p. 54-62.

222. Boyle, P., et al., *Contribution of three components to individual cancer risk predicting breast cancer risk in Italy.* European journal of cancer prevention, 2004. **13**(3): p. 183-191.

223. Lee, E.O., 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**(5): p. 400-406.

224. Gail, M.H., et al., *Projecting individualized absolute invasive breast cancer risk in African American women.* Journal of the National Cancer Institute, 2007. **99**(23): p. 1782-1792.

225. Matsuno, R.K., et al., *Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women.* Journal of the National Cancer Institute, 2011.

226. Banegas, M.P., et al., *Evaluating breast cancer risk projections for Hispanic women.* Breast cancer research and treatment, 2012. **132**(1): p. 347-353.

227. Pfeiffer, R.M., 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): p. e1001492.

228. Lee, C., et al., *Computational Discrimination of Breast Cancer for Korean Women Based on Epidemiologic Data Only.* Journal of Korean Medical Science, 2015. **30**(8): p. 1025-1034.

229. Meads, C., I. Ahmed, and R. Riley, *A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance.* Breast Cancer Research and Treatment, 2012. **132**(2): p. 365-377.

230. Parikh, R., et al., *Understanding and using sensitivity, specificity and predictive values.* Indian J Ophthalmol, 2008. **56**(1): p. 45-50.

231. Emmons, K.M., et al., *Tailored Computer-Based Cancer Risk Communication: Correcting Colorectal Cancer Risk Perception.* Journal of Health Communication, 2004. **9**(2): p. 127-141.

232. Karen M, E., et al., *A Qualitative Evaluation of the Harvard Cancer Risk Index.* Journal of Health Communication, 1999. **4**(3): p. 181-193.

233. Bandura, A., *Exercise of personal agency through the self-efficacy mechanism*, in *Self-efficacy: Thought control of action*. 1992, Hemisphere Publishing Corp: Washington, DC, US. p. 3-38.

234. Kim, D.J., B. Rockhill, and G.A. Colditz, *Validation of the Harvard Cancer Risk Index: a prediction tool for individual cancer risk.* Journal of Clinical Epidemiology, 2004. **57**(4): p. 332-340.

235. Karen M. Emmons, S.K.-W.K.A.L.C.R.R.G.C., *A Qualitative Evaluation of the Harvard Cancer Risk Index.* Journal of Health Communication, 1999. **4**(3): p. 181-193.

236. Matsuno, R.K., et al., *Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women.* J Natl Cancer Inst, 2011. **103**(12): p. 951-61.

237. Collins, G.S., et al., *External validation of multivariable prediction models: a systematic review of methodological conduct and reporting.* BMC Med Res Methodol, 2014. **14**: p. 40.

238. Pepe, M.S., et al., *Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker.* American Journal of Epidemiology, 2004. **159**(9): p. 882-890.

239. Rosner, B., et al., *Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers.* Am J Epidemiol, 2013. **178**(2): p. 296-308.

240. Colditz, G.A., et al., *Risk factors for breast cancer according to estrogen and progesterone receptor status.* J Natl Cancer Inst, 2004. **96**(3): p. 218-28.

241. Cook, N.R., *Use and Misuse of the Receiver Operating Characteristic Curve in Risk Prediction.* Circulation, 2007. **115**(7): p. 928-935.

242. Pfeiffer, R.M., 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 Medicine, 2013. **10**(7).

243. Timmers, J.M., et al., *Breast cancer risk prediction model: a nomogram based on common mammographic screening findings.* Eur Radiol, 2013. **23**(9): p. 2413-9.

244. McCowan, C., et al., *Identifying suspected breast cancer: development and validation of a clinical prediction rule.* British Journal of General Practice, 2011. **61**(586).

245. Cook, N.R., et al., *Mammographic Screening and Risk Factors for Breast Cancer.* American Journal of Epidemiology, 2009. **170**(11): p. 1422-1432.

246. Tice, J.A., et al., *Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model.* Annals of Internal Medicine, 2008. **148**(5): p. 337-W75.

247. Rosner, B., et al., *Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the Nurses' Health Study.* Breast Cancer Research, 2008. **10**(4).

248. Lee, S.M., J.H. Park, and H.J. Park, *Implications of systematic review for breast cancer prediction.* Cancer Nurs, 2008. **31**(5): p. E40-6.

249. Chlebowski, R.T., et al., *Predicting risk of breast cancer in postmenopausal women by hormone receptor status.* J Natl Cancer Inst, 2007. **99**(22): p. 1695-705.

250. Decarli, A., et al., *Gail model for prediction of absolute risk of invasive breast cancer: Independent evaluation in the Florence-European Prospective Investigation Into Cancer and Nutrition Cohort.* Journal of the National Cancer Institute, 2006. **98**(23): p. 1686-1693.

251. Chen, J.B., et al., *Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density.* Journal of the National Cancer Institute, 2006. **98**(17): p. 1215-1226.

252. Barlow, W.E., et al., *Prospective breast cancer risk prediction model for women undergoing screening mammography.* J Natl Cancer Inst, 2006. **98**(17): p. 1204-14.

253. Tice, J.A., et al., *Nipple aspirate fluid cytology and the Gail model for breast cancer risk assessment in a screening population.* Cancer Epidemiol Biomarkers Prev, 2005. **14**(2): p. 324-8.

254. Taplin, S.H., et al., *Revisions in the Risk-Based Breast-Cancer Screening-Program at Group Health Cooperative.* Cancer, 1990. **66**(4): p. 812-818.

255. Anderson, D.E. and M. Badzioch, *Risk of Familial Breast-Cancer.* Lancet, 1984. **1**(8373): p. 392-392.

256. Ottman, R., et al., *Practical Guide for Estimating Risk for Familial Breast-Cancer.* Lancet, 1983. **2**(8349): p. 556-558.

257. Lee, A.J., et al., *BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface.* British Journal of Cancer, 2014. **110**(2): p. 535-545.

258. McCarthy, A.M., et al., *Incremental impact of breast cancer SNP panel on risk classification in a screening population of white and African American women.* Breast Cancer Research and Treatment, 2013. **138**(3): p. 889-898.

259. Dite, G.S., et al., *Using SNP genotypes to improve the discrimination of a simple breast cancer risk prediction model.* Breast Cancer Research and Treatment, 2013. **139**(3): p. 887-896.

260. Biswas, S., et al., *Simplifying clinical use of the genetic risk prediction model BRCAPRO.* Breast Cancer Research and Treatment, 2013. **139**(2): p. 571-579.

261. Sueta, A., et al., *A genetic risk predictor for breast cancer using a combination of low-penetrance polymorphisms in a Japanese population.* Breast Cancer Research and Treatment, 2012. **132**(2): p. 711-721.

262. Huesing, A., et al., *Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status.* Journal of Medical Genetics, 2012. **49**(9): p. 601-608.

263. Darabi, H., et al., *Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement.* Breast Cancer Research, 2012. **14**(1).

264. Dai, J.C., et al., *Breast cancer risk assessment with five independent genetic variants and two risk factors in Chinese women.* Breast Cancer Research, 2012. **14**(1).

265. Biswas, S., et al., *Assessing the added value of breast tumor markers in genetic risk prediction model BRCAPRO.* Breast Cancer Res Treat, 2012. **133**(1): p. 347-55.

266. van Zitteren, M., et al., *Genome-Based Prediction of Breast Cancer Risk in the General Population: A Modeling Study Based on Meta-Analyses of Genetic Associations.* Cancer Epidemiology Biomarkers & Prevention, 2011. **20**(1): p. 9-22.

267. Crooke, P.S., et al., *Estrogen Metabolism and Exposure in a Genotypic-Phenotypic Model for Breast Cancer Risk Prediction.* Cancer Epidemiology Biomarkers & Prevention, 2011. **20**(7): p. 1502-1515.

268. Wacholder, S., et al., *Performance of Common Genetic Variants in Breast-Cancer Risk Models.* New England Journal of Medicine, 2010. **362**(11): p. 986-993.

269. Antoniou, A.C., et al., *The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions.* British Journal of Cancer, 2008. **98**(8): p. 1457-1466.

270. Tyrer, J., S.W. Duffy, and J. Cuzick, *A breast cancer prediction model incorporating familial and personal risk factors.* Statistics in Medicine, 2004. **23**(7): p. 1111-1130.

271. Evans, D.G.R., et al., *A new scoring system for the chances of identifying a BRCA1/2 mutation outperforms existing models including BRCAPRO.* Journal of Medical Genetics, 2004. **41**(6): p. 474-480.

272. Antoniou, A.C., et al., *The BOADICEA model of genetic susceptibility to breast and ovarian cancer.* British Journal of Cancer, 2004. **91**(8): p. 1580-1590.

273. Jonker, M.A., et al., *Modeling familial clustered breast cancer using published data.* Cancer Epidemiology Biomarkers & Prevention, 2003. **12**(12): p. 1479-1485.

274. Fisher, T.J., et al., *A simple tool for identifying unaffected women at a moderately increased or potentially high risk of breast cancer based on their family history.* Breast, 2003. **12**(2): p. 120-127.

275. Apicella, C., et al., *Log odds of carrying an Ancestral Mutation in BRCA1 or BRCA2 for a defined personal and family history in an Ashkenazi Jewish woman (LAMBDA).* Breast Cancer Research, 2003. **5**(6): p. R206-R216.

276. Frank, T.S., et al., *Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: Analysis of 10,000 individuals.* Journal of Clinical Oncology, 2002. **20**(6): p. 1480-1490.

277. de la Hoya, M., et al., *Association between BRCA1 and BRCA2 mutations and cancer phenotype in Spanish breast/ovarian cancer families: Implications for genetic testing.* International Journal of Cancer, 2002. **97**(4): p. 466-471.
278. Berry, D.A., et al., *BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes.* Journal of Clinical Oncology, 2002. **20**(11): p. 2701-2712.
279. Antoniou, A.C., et al., *A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes.* British Journal of Cancer, 2002. **86**(1): p. 76-83.
280. Vahteristo, P., et al., *A probability model for predicting BRCA1 and BRCA2 mutations in breast and breast-ovarian cancer families.* British Journal of Cancer, 2001. **84**(5): p. 704-708.
281. Gilpin, C.A., N. Carson, and A.G.W. Hunter, *A preliminary validation of a family history assessment form to select women at risk for breast or ovarian cancer for referral to a genetics center.* Clinical Genetics, 2000. **58**(4): p. 299-308.
282. Hartge, P., et al., *The prevalence of common BRCA1 and BRCA2 mutations among Ashkenazi Jews.* American Journal of Human Genetics, 1999. **64**(4): p. 963-970.
283. Parmigiani, G., D.A. Berry, and O. Aguilar, *Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2.* American Journal of Human Genetics, 1998. **62**(1): p. 145-158.
284. Frank, T.S., et al., *Sequence analysis of BRCA1 and BRCA2: Correlation of mutations with family history and ovarian cancer risk.* Journal of Clinical Oncology, 1998. **16**(7): p. 2417-2425.
285. Shattuck-Eidens, D., et al., *BRCA1 sequence analysis in women at high risk for susceptibility mutations. Risk factor analysis and implications for genetic testing.* Journal of the American Medical Association, 1997(0098-7484 (Print)).
286. Couch, F.J., et al., *BRCA1 mutations in women attending clinics that evaluate the risk of breast cancer.* New England Journal of Medicine, 1997. **336**(20): p. 1409-1415.
287. Claus, E.B., N. Risch, and W.D. Thompson, *Autosomal-Dominant Inheritance of Early-Onset Breast-Cancer - Implications for Risk Prediction.* Cancer, 1994. **73**(3): p. 643-651.
288. Claus, E.B., N. Risch, and W.D. Thompson, *The Calculation of Breast-Cancer Risk for Women with a First Degree Family History of Ovarian-Cancer.* Breast Cancer Research and Treatment, 1993. **28**(2): p. 115-120.
289. Wang, S., 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 Research, 2016. **76**(14 Supplement): p. 2590.
290. Hartge, P., et al., *The prevalence of common BRCA1 and BRCA2 mutations among Ashkenazi Jews.* Am J Hum Genet, 1999. **64**(4): p. 963-70.
291. de la Hoya, M., et al., *Association between BRCA1 and BRCA2 mutations and cancer phenotype in Spanish breast/ovarian cancer families: implications for genetic testing.* Int J Cancer, 2002. **97**(4): p. 466-71.
292. Decarli, A., et al., *Gail model for prediction of absolute risk of invasive breast cancer: independent evaluation in the Florence-European Prospective Investigation Into Cancer and Nutrition cohort.* J Natl Cancer Inst, 2006. **98**(23): p. 1686-93.
293. Rosner, B., et al., *Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the Nurses' Health Study.* Breast Cancer Res, 2008. **10**(4): p. R55.
294. Cook, N.R., et al., *Mammographic screening and risk factors for breast cancer.* Am J Epidemiol, 2009. **170**(11): p. 1422-32.
295. Hortobagyi, G.N., et al., *The global breast cancer burden: variations in epidemiology and survival.* Clin Breast Cancer, 2005. **6**(5): p. 391-401.
296. Kelsey, J.L., E.M. Gammon Md Fau - John, and E.M. John, *Reproductive factors and breast cancer.* 1993(0193-936X (Print)).
297. Friedenreich, C.M., *Review of anthropometric factors and breast cancer risk.* European Journal of Cancer Prevention, 2001. **10**(1): p. 15-32.

298. Hulka, B.S. and P.G. Moorman, *Breast cancer: hormones and other risk factors.* Maturitas, 2001. **38**(1): p. 103-113.

299. Balducci, L. and W.B. Ershler, *Cancer and ageing: a nexus at several levels.* Nat Rev Cancer, 2005. **5**(8): p. 655-662.

300. Petrisek, A., S. Campbell, and L. Laliberte, *Family history of breast cancer. Impact on the disease experience.* Cancer Pract, 2000. **8**(3): p. 135-42.

301. Kelsey, J.L. and M.D. Gammon, *The epidemiology of breast cancer.* CA Cancer J Clin, 1991. **41**(3): p. 146-65.

302. Pharoah, P.D., et al., *Family history and the risk of breast cancer: a systematic review and meta-analysis.* Int J Cancer, 1997. **71**(5): p. 800-9.

303. Miki, Y., et al., *A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.* Science, 1994. **266**(5182): p. 66-71.

304. Wooster, R., et al., *Identification of the breast cancer susceptibility gene BRCA2.* Nature, 1995. **378**(6559): p. 789-92.

305. Dartois, L., et al., *Proportion of premenopausal and postmenopausal breast cancers attributable to known risk factors: Estimates from the E3N-EPIC cohort.* International Journal of Cancer, 2016. **138**(10): p. 2415-2427.

306. Harvie, M., L. Hooper, and A.H. Howell, *Central obesity and breast cancer risk: a systematic review.* Obes Rev, 2003. **4**(3): p. 157-73.

307. Friedenreich, C.M., *Review of anthropometric factors and breast cancer risk.* Eur J Cancer Prev, 2001. **10**(1): p. 15-32.

308. Papa, V., et al., *Insulin-like growth factor-I receptors are overexpressed and predict a low risk in human breast cancer.* 1993(0008-5472 (Print)).

309. Bates, P., et al., *Mammary cancer in transgenic mice expressing insulin-like growth factor II (IGF-II).* 1995(0007-0920 (Print)).

310. Key, T.J., et al., *Body mass index, serum sex hormones, and breast cancer risk in postmenopausal women.* J Natl Cancer Inst, 2003. **95**(16): p. 1218-26.

311. Endogenous, H., et al., *Circulating sex hormones and breast cancer risk factors in postmenopausal women: reanalysis of 13 studies.* Br J Cancer, 2011. **105**(5): p. 709-22.

312. Sprague, B.L., et al., *Proportion of invasive breast cancer attributable to risk factors modifiable after menopause.* Am J Epidemiol, 2008. **168**(4): p. 404-11.

313. Nystrom, L., et al., *Breast cancer screening with mammography: overview of Swedish randomised trials.* Lancet, 1993. **341**(8851): p. 973-8.

314. Thorbjarnardottir, T., et al., *Oral contraceptives, hormone replacement therapy and breast cancer risk: A cohort study of 16 928 women 48 years and older.* Acta Oncologica, 2014. **53**(6): p. 752-758.

315. Beral, V., et al., *Breast cancer risk in relation to the interval between menopause and starting hormone therapy.* J Natl Cancer Inst, 2011. **103**(4): p. 296-305.

316. Prentice, R.L., et al., *Benefits and Risks of Postmenopausal Hormone Therapy When It Is Initiated Soon After Menopause.* American Journal of Epidemiology, 2009. **170**(1): p. 12-23.

317. Cancer, I.A.f.R.o., *Monograph on the Evaluation of Carcinogenic Risks to Humans. Combined Estrogen/Progestogen Contraceptives and Combined Estrogen/Progestogen Menopausal Therapy.* France IARC Press, 2008. **91**.

318. CRUK. *Breast cancer statistics*. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Four.

319. Nathanson, K.N., R. Wooster, and B.L. Weber, *Breast cancer genetics: What we know and what we need.* Nature Medicine, 2001. **7**(5): p. 552-556.

320. Pashayan, N., et al., *Polygenic susceptibility to prostate and breast cancer: implications for personalised screening.* British Journal of Cancer, 2011. **104**(10): p. 1656-1663.

321. Pharoah, P.D.P., et al., *Polygenic susceptibility to breast cancer and implications for prevention.* Nature Genetics, 2002. **31**(1): p. 33-36.

322. Mavaddat, N., et al., *Prediction of breast cancer risk based on profiling with common genetic variants.* J Natl Cancer Inst, 2015. **107**(5).

323. Al-Ajmi, K., et al., *Risk of breast cancer in the UK biobank female cohort and its relationship to anthropometric and reproductive factors.* PLoS One, 2018. **13**(7): p. e0201097.

324. Colditz, G.A., et al., *Harvard Report on Cancer Prevention Volume 4: Harvard Cancer Risk Index.* Cancer Causes & Control, 2000. **11**(6): p. 477-488.

325. Elwood, P.C., et al., *Healthy living and cancer: evidence from UK Biobank.* Ecancermedicalscience, 2018. **12**: p. 792.

326. Hamajima, N., et al., *Alcohol, tobacco and breast cancer--collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease.* Br J Cancer, 2002. **87**(11): p. 1234-45.

327. Dartois, L., et al., *Proportion of premenopausal and postmenopausal breast cancers attributable to known risk factors: Estimates from the E3N-EPIC cohort.* International Journal of Cancer, 2016. **138**(10): p. 2415-27.

328. Haffty, B.G., et al., *Outcome of conservatively managed early-onset breast cancer by BRCA1/2 status.* The Lancet, 2002. **359**(9316): p. 1471-1477.

329. de la Rochefordière, A., et al., *Age as prognostic factor in premenopausal breast carcinoma.* The Lancet, 1993. **341**(8852): p. 1039-1043.

330. Walker, R.A., et al., *Breast carcinomas occurring in young women (< 35 years) are different.* British Journal of Cancer, 1996. **74**(11): p. 1796-1800.

331. Mavaddat, N., et al., *Genetic susceptibility to breast cancer.* Molecular Oncology, 2010. **4**(3): p. 174-191.

332. Lewis, D.R., et al., *Adolescent and young adult cancer survival.* J Natl Cancer Inst Monogr, 2014. **2014**(49): p. 228-35.

333. Al-Ajmi, K., et al., *Review of non-clinical risk models to aid prevention of breast cancer.* Cancer Causes & Control, 2018. **29**(10): p. 967-986.

334. La Vecchia, C. and G. Carioli, *The epidemiology of breast cancer, a summary overview.* Epidemiology Biostatistics and Public Health, 2018. **15**(1).

335. Dumitrescu, R.G. and I. Cotarla, *Understanding breast cancer risk - where do we stand in 2005?* Journal of Cellular and Molecular Medicine, 2005. **9**(1): p. 208-221.

336. Hulka, B.S. and P.G. Moorman, *Breast cancer: hormones and other risk factors.* Maturitas, 2008. **61**(1-2): p. 203-13; discussion 213.

337. Guo, W.J., T.J. Key, and G.K. Reeves, *Adiposity and breast cancer risk in postmenopausal women: Results from the UK Biobank prospective cohort.* International Journal of Cancer, 2018. **143**(5): p. 1037-1046.

338. Ross, R.K., et al., *Effect of hormone replacement therapy on breast cancer risk: estrogen versus estrogen plus progestin.* J Natl Cancer Inst, 2000. **92**(4): p. 328-32.

339. Dumeaux, V., E. Alsaker, and E. Lund, *Breast cancer and specific types of oral contraceptives: a large Norwegian cohort study.* International Journal of Cancer, 2003. **105**(6): p. 844-50.

340. Kumle, M., et al., *Use of Oral Contraceptives and Breast Cancer Risk.* Cancer Epidemiology Biomarkers &amp;amp; Prevention, 2002. **11**(11): p. 1375.

341. Calle, E.E., et al., *Breast cancer and hormonal contraceptives: Collaborative reanalysis of individual data on 53297 women with breast cancer and 100239 women without breast cancer from 54 epidemiological studies.* Lancet, 1996. **347**(9017): p. 1713-1727.

342. Bernstein, L., et al., *Treatment with human chorionic gonadotropin and risk of breast cancer.* Cancer Epidemiol Biomarkers Prev, 1995. **4**(5): p. 437-40.

343. Murphy, B., et al., *The Use of Deceased Controls in Epidemiologic Research: A Systematic Review.* American Journal of Epidemiology, 2017. **186**(3): p. 367-384.

344. Barry, V., et al., *Disease fatality and bias in survival cohorts.* Environ Res, 2015. **140**: p. 275-81.

345. Al-Ajmi, K., et al., *Review of non-clinical risk models to aid prevention of breast cancer.* Cancer Causes Control, 2018. **29**(10): p. 967-986.

346. Rosner, B., G.A. Colditz, and W.C. Willett, *Reproductive risk factors in a prospective study of breast cancer: the Nurses' Health Study.* Am J Epidemiol, 1994. **139**(8): p. 819-35.

347. Surakasula, A., G.C. Nagarjunapu, and K.V. Raghavaiah, *A comparative study of pre- and post-menopausal breast cancer: Risk factors, presentation, characteristics and management.* J Res Pharm Pract, 2014. **3**(1): p. 12-8.

348. Chollet-Hinton, L., et al., *Breast cancer biologic and etiologic heterogeneity by young age and menopausal status in the Carolina Breast Cancer Study: a case-control study.* Breast Cancer Research, 2016. **18**(1): p. 79.

349. Gail, M.H., et al., *Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually.* JNCI: Journal of the National Cancer Institute, 1989. **81**(24): p. 1879-1886.

350. Mavaddat, N., et al., *Incorporating tumour pathology information into breast cancer risk prediction algorithms.* Breast Cancer Research, 2010. **12**(3): p. R28.

351. Fry, A., et al., *Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population.* Am J Epidemiol, 2017. **186**(9): p. 1026-1034.

352. Banegas, M.P., et al., *Evaluating breast cancer risk projections for Hispanic women.* Breast Cancer Res Treat, 2012. **132**(1): p. 347-53.

353. Boyle, P., et al., *Contribution of three components to individual cancer risk predicting breast cancer risk in Italy.* Eur J Cancer Prev, 2004. **13**(3): p. 183-91.

354. Colditz, G.A. and B. Rosner, *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**(10): p. 950-64.

355. Lee, E.O., 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 Nurs, 2004. **27**(5): p. 400-6.

356. Novotny, 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**(1): p. 29-35.

357. Pfeiffer, R.M., 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): p. e1001492.

358. Rosner, B. and G.A. Colditz, *Nurses' health study: log-incidence mathematical model of breast cancer incidence.* J Natl Cancer Inst, 1996. **88**(6): p. 359-64.

359. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.

360. Amir, E., et al., *Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme.* J Med Genet, 2003. **40**(11): p. 807-14.

361. Fry, A., et al., *Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population.* American Journal of Epidemiology, 2017. **186**(9): p. 1026-1034.

362. Batty, G.D., et al., *Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis.* BMJ, 2020. **368**: p. m131.

363. Rothman, K.J., J.E. Gallacher, and E.E. Hatch, *Why representativeness should be avoided.* Int J Epidemiol, 2013. **42**(4): p. 1012-4.

364. Ebrahim, S. and G. Davey Smith, *Commentary: Should we always deliberately be non-representative?* Int J Epidemiol, 2013. **42**(4): p. 1022-6.

# Appendices

Appendix 1: Assessment of the included risk factors

| **Variable** | **Obtaining method** | **Missingness # (%) *** |
|---|---|---|
| Menopausal status | Questionnaire | 0 (0%) |
| Menarche age | Questionnaire | 2,958 (2.3%) |
| Parity | Questionnaire | 56 (0.04%) |
| Family history | Questionnaire | 365 (0.3%) |
| Moderate physical activity | Questionnaire | 2,027 (1.6%) |
| HRT users | Questionnaire | 221 (0.2%) |
| OC users | Questionnaire | 156 (0.1%) |
| Smoking status | Questionnaire | 332 (0.3%) |
| Alcohol drinking status | Questionnaire | 82 (0.1%) |
| Age | Questionnaire | 0 (0%) |
| Age at first birth | Questionnaire | 42,783 (32.8%) |
| BMI | Calculated from measured variables | 281 (0.2%) |
| Waist to hip ratio | Calculated from measured variables | 184 (0.1%) |
| Reproductive interval index | Calculated | 44,436 (34.1%) |
| Deprivation score | Provided by the UK biobank team | 156 (0.1%) |
| Raw vegetables intake | Questionnaire | 0 (0%) |
| Processed meat intake | Questionnaire | 129 (0.1%) |
| Beef intake | Questionnaire | 325 (0.3%) |
| Number of Pregnancy termination | Questionnaire | 89,893 (68.9%) |
| Number of live births | Questionnaire | 56 (0.04%) |
| Hip circumference | Measured by UK biobank team | 185 (0.1%) |
| Waist circumference | Measured by UK biobank team | 182 (0.1%) |
| Standing height | Measured by UK biobank team | 195 (0.1%) |
| Sitting height | Measured by UK biobank team | 312 (0.2%) |

*Total females included in the analysis is 130,382

Appendix 2: Coding Stata do file for the included risk factors in the prediction model.

```
*** 1- Case and control coding
** ---------   Do file for UK biobank Project    --------- **
** ---------   Modified by: Kawthar Alajmi      --------- **
*** Available coding in this file: (Each section and subsection were bookmarked to make easier to find it)
   /// 1) Cases and controls identified using ICD10 variables
   /// 2) Cases and controls identified using ICD9 variables
   /// 3) Cases and controls identified using self-reported variables
*** 2- Family history of breast cancer coding
   /// Family history of a mother or sibling or both
*** 3- Anthropometric variables coding
   /// Available variables: BMI (continues and categorical), waist to hip ratio (continues and categorical), standing height (continues and categorical), and sitting height
*** 4- Reproductive variables coding
   /// Available variables: Menarche age (continues and categorical), menopause age, age at first birth (continues and categorical), interval of reproductive index (continues and categorical), premenopause_duration,  ///
   /// postmenopause_duration, HRT (status , start age , end age , and duration) , contraceptive use (status , start age , end age , and duration (numeric and categorical)), number of live births (continues and categorical) ///
   /// stillbirth group, termination number, miscarriages number, hysterectomy
*** 5- Menopausal status coding
   /// Information about the menopausal status classification (pre and post) based on reported status, hysterectomy and oophorectomy stats, current age, menarche age, and menopause age.

** 1- Case and control coding

** Coding of cases and controls based on ICD10 variables
** 0 = control , 100==ICD10/0 101==ICD10/1 102==ICD10/2 103==ICD10/3 104==ICD10/4 105==ICD10/5 106==ICD10/6 107==ICD10/7 108==ICD10/8 109==ICD10/9
** If participant had any of these 1cd10 then it will be considered as a case of Breast cancer
* ICD10=C500, C501, C502, C503, C504, C505, C506, C507, C508, C509
* ICD9=1740, 1741, 1742, 1743, 1744, 1745, 1746, 1747, 1748, 1749
* Self-reported = 1002

***     First by ICD10 for all the follow ups    ***
// I started by the from the last follow up to the first follow up in order
// so, the case appears for example in follow up 3 and 9
// The follow up 9 will be replace by 3 and we will know at what follow the participant go the cancer
// More accurate to consider as incident or prevalent case

** Cases and controls identified using ICD10 variables

gen bc_case10_new=.
 ** ICD10 **
replace bc_case10_new=100 if bc_case10_new==. & (strmatch( s_40006_0_0 , "C50?"))
replace bc_case10_new=100 if bc_case10_new==. & (strmatch( s_40006_0_0 , "C50"))
replace bc_case10_new=101 if bc_case10_new==. & (strmatch( s_40006_1_0 , "C50?"))
replace bc_case10_new=101 if bc_case10_new==. & (strmatch( s_40006_1_0 , "C50"))
replace bc_case10_new=102 if bc_case10_new==. & (strmatch( s_40006_2_0 , "C50?"))
replace bc_case10_new=102 if bc_case10_new==. & (strmatch( s_40006_2_0 , "C50"))
replace bc_case10_new=103 if bc_case10_new==. & (strmatch( s_40006_3_0 , "C50?"))
replace bc_case10_new=103 if bc_case10_new==. & (strmatch( s_40006_3_0 , "C50"))
replace bc_case10_new=104 if bc_case10_new==. & (strmatch( s_40006_4_0 , "C50?"))
replace bc_case10_new=104 if bc_case10_new==. & (strmatch( s_40006_4_0 , "C50"))
replace bc_case10_new=105 if bc_case10_new==. & (strmatch( s_40006_5_0 , "C50?"))
replace bc_case10_new=105 if bc_case10_new==. & (strmatch( s_40006_5_0 , "C50"))
replace bc_case10_new=106 if bc_case10_new==. & (strmatch( s_40006_6_0 , "C50?"))
replace bc_case10_new=106 if bc_case10_new==. & (strmatch( s_40006_6_0 , "C50"))
replace bc_case10_new=107 if bc_case10_new==. & (strmatch( s_40006_7_0 , "C50?"))
replace bc_case10_new=107 if bc_case10_new==. & (strmatch( s_40006_7_0 , "C50"))
replace bc_case10_new=108 if bc_case10_new==. & (strmatch( s_40006_8_0 , "C50?"))
replace bc_case10_new=108 if bc_case10_new==. & (strmatch( s_40006_8_0 , "C50"))
replace bc_case10_new=109 if bc_case10_new==. & (strmatch( s_40006_9_0 , "C50?"))
replace bc_case10_new=109 if bc_case10_new==. & (strmatch( s_40006_9_0 , "C50"))
replace bc_case10_new=110 if bc_case10_new==. & (strmatch( s_40006_10_0 , "C50?"))
replace bc_case10_new=110 if bc_case10_new==. & (strmatch( s_40006_10_0 , "C50"))
replace bc_case10_new=111 if bc_case10_new==. & (strmatch( s_40006_11_0 , "C50?"))
replace bc_case10_new=111 if bc_case10_new==. & (strmatch( s_40006_11_0 , "C50"))
replace bc_case10_new=113 if bc_case10_new==. & (strmatch( s_40006_13_0 , "C50?"))
replace bc_case10_new=113 if bc_case10_new==. & (strmatch( s_40006_13_0 , "C50"))
replace bc_case10_new=115 if bc_case10_new==. & (strmatch( s_40006_15_0 , "C50?"))
replace bc_case10_new=115 if bc_case10_new==. & (strmatch( s_40006_15_0 , "C50"))
replace bc_case10_new=116 if bc_case10_new==. & (strmatch( s_40006_16_0 , "C50?"))
replace bc_case10_new=116 if bc_case10_new==. & (strmatch( s_40006_16_0 , "C50"))

**    2- define the Controls into four groups   **
**  First: Carcinoma in situ ICD10: D050, D051, D057, D059 and ICD9: 2330
**  Second: Any other cancerous control (All except breast cancer)
**  Third: Other diseases but not cancers
**  Forth: apparently healthy controls - No diseases

** Defining the In situ cases of breast
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_0_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_1_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_2_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_3_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_4_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_5_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_6_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_7_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_8_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_9_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_10_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_11_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_13_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_15_0 , "D05?"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_16_0 , "D05?"))
** 3 digits
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_0_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_1_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_2_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_3_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_4_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_5_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_6_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_7_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_8_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_9_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_10_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_11_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_13_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_15_0 , "D05"))
replace bc_case10_new=3 if bc_case10_new==. & (strmatch( s_40006_16_0 , "D05"))
** Defining the Cancerous controls using ICD10  **
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_0_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_1_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_2_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_3_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_4_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_5_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_6_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_7_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_8_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_9_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_10_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_11_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_13_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_15_0 , "C???"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_16_0 , "C???"))
** 3 digits
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_0_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_1_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_2_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_3_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_4_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_5_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_6_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_7_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_8_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_9_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_10_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_11_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_13_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_15_0 , "C??"))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40006_16_0 , "C??"))
** Defining the Cancerous controls using ICD9  **
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_0_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_1_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_2_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_3_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_4_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_5_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_6_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_7_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_8_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_10_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_11_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_12_0 , "1???" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_14_0 , "1???" ))
** When only 3 digits in the coding **
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_0_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_1_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_2_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_3_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_4_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_5_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_6_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_7_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_8_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_10_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_11_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_12_0 , "1??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_14_0 , "1??" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_0_0 , "20??" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_1_0 , "20??" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_2_0 , "20??" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_3_0 , "20??" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_4_0 , "20??" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_5_0 , "20??" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_6_0 , "20??" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_7_0 , "20??" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_8_0 , "20??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_10_0 , "20??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_11_0 , "20??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_12_0 , "20??" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_14_0 , "20??" ))
** When only 3 digits in the coding **
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_0_0 , "20?" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_1_0 , "20?" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_2_0 , "20?" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_3_0 , "20?" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_4_0 , "20?" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_5_0 , "20?" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_6_0 , "20?" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_7_0 , "20?" ))
replace bc_case10_new =4 if bc_case10_new ==. & (strmatch( s_40013_8_0 , "20?" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_10_0 , "20?" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_11_0 , "20?" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_12_0 , "20?" ))
replace bc_case10_new=4 if bc_case10_new==. & (strmatch( s_40013_14_0 , "20?" ))
** Defining the Cancerous controls using Self   **
replace bc_case10_new=4 if bc_case10_new==. & n_20001_0_0!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_0_1!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_0_2!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_0_3!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_0_4!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_0_5!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_1_0!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_1_1!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_1_2!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_2_0!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_2_1!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_2_2!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_2_3!=.
replace bc_case10_new=4 if bc_case10_new==. & n_20001_2_4!=.
** Defining the insitu carcinoma in ICD10
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_0_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_1_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_2_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_3_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_4_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_5_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_6_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_7_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_8_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_9_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_10_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_11_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_13_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_15_0 , "D0??"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_16_0 , "D0??"))
** 3 digits
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_0_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_1_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_2_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_3_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_4_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_5_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_6_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_7_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_8_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_9_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_10_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_11_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_13_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_15_0 , "D0?"))
replace bc_case10_new=6 if bc_case10_new==. & (strmatch( s_40006_16_0 , "D0?"))
** ICD9
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_0_0 , "230?" "231?"
"232?" "233?""234?"))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_1_0 , "230?" "231?"
"232?" "233?""234?"))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_2_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_3_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_4_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_5_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_6_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_7_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_8_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_10_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_11_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_12_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_14_0 , "230?" "231?"
"232?" "233?""234?" ))
** When only 3 digits in the coding **
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_0_0 , "230" "231" "232"
"233" "234"))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_1_0 , "230" "231" "232"
"233" "234" ))
```

```
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_2_0 , "230" "231" "232"
"233" "234" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_3_0 , "230" "231" "232"
"233" "234" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_4_0 , "230" "231" "232"
"233" "234" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_5_0 , "230" "231" "232"
"233" "234" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_6_0 , "230" "231" "232"
"233" "234" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_7_0 , "230" "231" "232"
"233" "234"))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_8_0 , "230" "231" "232"
"233" "234" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_10_0 , "230" "231" "232"
"233" "234" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_11_0 , "230" "231" "232"
"233" "234" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_12_0 , "230" "231" "232"
"233" "234" ))
replace bc_case10_new =6 if bc_case10_new ==. & (strmatch( s_40013_14_0 , "230" "231" "232"
"233" "234" ))
** Defining the Neoplasms of unknown behaviour / nature
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_0_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_1_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_2_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_3_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_4_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_5_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_6_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_7_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_8_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_9_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_10_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_11_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_13_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_15_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_16_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
** 3 digits
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_0_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
```

```
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_1_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_2_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_3_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_4_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_5_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_6_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_7_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_8_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_9_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_10_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_11_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_13_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_15_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case10_new=7 if bc_case10_new==. & (strmatch( s_40006_16_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))

** ICD9
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_0_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_1_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_2_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_3_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_4_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_5_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_6_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_7_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_8_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_10_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_11_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_12_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_14_0 , "235?" "236?"
"237?" "238?""239?"))
** When only 3 digits in the coding **
```

```
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_0_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_1_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_2_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_3_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_4_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_5_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_6_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_7_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_8_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_10_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_11_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_12_0 , "235" "236" "237"
"238""239"))
replace bc_case10_new=7 if bc_case10_new ==. & (strmatch( s_40013_14_0 , "235" "236" "237"
"238""239"))
** Defining the other controls using non-cancer-self reported **

** first i had to code icd10 , icd9 and self-reported without any cancer code
** ppl with no code in the three categories means they don’t have cancer, but they have other
diseases
gen no_icd10=.
replace no_icd10=0 if  s_40006_0_0=="" & s_40006_1_0=="" & s_40006_2_0=="" &
s_40006_3_0=="" & s_40006_4_0=="" & ///
s_40006_5_0=="" & s_40006_6_0=="" & s_40006_7_0=="" & s_40006_8_0=="" &
s_40006_9_0=="" & s_40006_10_0=="" & s_40006_11_0=="" & s_40006_13_0=="" &
s_40006_15_0=="" & s_40006_16_0==""
replace no_icd10=1 if no_icd10==.
label define no10lb 0"No cancer Code" 1"Cancer code", modify
label values no_icd10 no10lb
gen no_icd9=.
replace no_icd9=0 if  s_40013_0_0=="" & s_40013_1_0=="" & s_40013_2_0=="" &
s_40013_3_0=="" & s_40013_4_0=="" & ///
s_40013_5_0=="" & s_40013_6_0=="" & s_40013_7_0=="" & s_40013_8_0=="" &
s_40013_10_0=="" & s_40013_11_0=="" & s_40013_12_0=="" & s_40013_14_0==""
replace no_icd9=1 if no_icd9==.
label values no_icd9 no10lb
gen no_self=.
replace no_self=0 if  n_20001_0_0==. & n_20001_0_1==. & n_20001_0_2==. & n_20001_0_3==.
& n_20001_0_4==. & ///
n_20001_0_5==. & n_20001_1_0==. & n_20001_1_1==. & n_20001_1_2==.  & n_20001_2_0==.
///
& n_20001_2_1==. & n_20001_2_2==. & n_20001_2_3==. & n_20001_2_4==.
replace no_self=1 if no_self==.
label values no_self no10lb
replace bc_case10_new=5 if bc_case10_new==. & no_icd10==0 & no_icd9==0 & no_self==0 //
this is defining the controls without cancer
```

```
** Defining the controls   **
replace bc_case10_new=0 if bc_case10_new==.
**  Define the incident and prevalent cases                          **
**  If age of diagnosis >= age when attended the centre = Incident case   **
**  If age of diagnosis < age when attended the centre = Prevalent case    **
** follow up = 0
replace bc_case10_new=1 if bc_case10_new==100 & n_40008_0_0 >= n_21003_0_0 &
n_40008_0_0!=.
replace bc_case10_new=2 if bc_case10_new==100 & n_40008_0_0 < n_21003_0_0
** follow up = 1
replace bc_case10_new=1 if bc_case10_new==101 & n_40008_1_0 >= n_21003_0_0 &
n_40008_1_0!=.
replace bc_case10_new=2 if bc_case10_new==101 & n_40008_1_0 < n_21003_0_0
** follow up = 2
replace bc_case10_new=1 if bc_case10_new==102 & n_40008_2_0 >= n_21003_0_0 &
n_40008_2_0!=.
replace bc_case10_new=2 if bc_case10_new==102 & n_40008_2_0 < n_21003_0_0
** follow up = 3
replace bc_case10_new=1 if bc_case10_new==103 & n_40008_3_0 >= n_21003_0_0 &
n_40008_3_0!=.
replace bc_case10_new=2 if bc_case10_new==103 & n_40008_3_0 < n_21003_0_0
** follow up = 4
replace bc_case10_new=1 if bc_case10_new==104 & n_40008_4_0 >= n_21003_0_0 &
n_40008_4_0!=.
replace bc_case10_new=2 if bc_case10_new==104 & n_40008_4_0 < n_21003_0_0
** follow up =5
replace bc_case10_new=1 if bc_case10_new==105 & n_40008_5_0 >= n_21003_0_0 &
n_40008_5_0!=.
replace bc_case10_new=2 if bc_case10_new==105 & n_40008_5_0 < n_21003_0_0
** follow up =7
replace bc_case10_new=1 if bc_case10_new==107 & n_40008_7_0 >= n_21003_0_0 &
n_40008_7_0!=.
replace bc_case10_new=2 if bc_case10_new==107 & n_40008_7_0 < n_21003_0_0
** follow up =9
replace bc_case10_new=1 if bc_case10_new==109 & n_40008_9_0 >= n_21003_0_0 &
n_40008_9_0!=.
replace bc_case10_new=2 if bc_case10_new==109 & n_40008_9_0 < n_21003_0_0
** follow up =13
replace bc_case10_new=1 if bc_case10_new==113 & n_40008_13_0 >= n_21003_0_0 &
n_40008_13_0!=.
replace bc_case10_new=2 if bc_case10_new==113 & n_40008_13_0 < n_21003_0_0
label define caselb 0"Healthy controls" 1"Incident" 2"Prevalent" 3"Breast insitu" 4"Cancer
Controls" 5"Other controls - No cancer" 6"Other In situ" 7"Unknown neoplasm" , modify
label values bc_case10_new caselb
label var bc_case10_new "Cases and Controls bc ICD10"
** Cases and controls identified using ICD9 variables
gen bc_case9_new=.
** ICD9 **
replace bc_case9_new=900 if bc_case9_new==. & (strmatch( s_40013_0_0 , "174?" ))
replace bc_case9_new=900 if bc_case9_new==. & (strmatch( s_40013_0_0 , "174" ))
replace bc_case9_new=901 if bc_case9_new==. & (strmatch( s_40013_1_0 , "174?" ))
replace bc_case9_new=901 if bc_case9_new==. & (strmatch( s_40013_1_0 , "174" ))
replace bc_case9_new=902 if bc_case9_new==. & (strmatch( s_40013_2_0 , "174?" ))
replace bc_case9_new=902 if bc_case9_new==. & (strmatch( s_40013_2_0 , "174" ))
replace bc_case9_new=903 if bc_case9_new==. & (strmatch( s_40013_3_0 , "174?" ))
replace bc_case9_new=903 if bc_case9_new==. & (strmatch( s_40013_3_0 , "174" ))
replace bc_case9_new=904 if bc_case9_new==. & (strmatch( s_40013_4_0 , "174?" ))
replace bc_case9_new=904 if bc_case9_new==. & (strmatch( s_40013_4_0 , "174" ))

replace bc_case9_new=905 if bc_case9_new==. & (strmatch( s_40013_5_0 , "174?" ))
replace bc_case9_new=905 if bc_case9_new==. & (strmatch( s_40013_5_0 , "174" ))
replace bc_case9_new=906 if bc_case9_new==. & (strmatch( s_40013_6_0 , "174?" ))
replace bc_case9_new=906 if bc_case9_new==. & (strmatch( s_40013_6_0 , "174" ))
replace bc_case9_new=907 if bc_case9_new==. & (strmatch( s_40013_7_0 , "174?" ))
replace bc_case9_new=907 if bc_case9_new==. & (strmatch( s_40013_7_0 , "174" ))
replace bc_case9_new=908 if bc_case9_new==. & (strmatch( s_40013_8_0 , "174?" ))
replace bc_case9_new=908 if bc_case9_new==. & (strmatch( s_40013_8_0 , "174" ))
replace bc_case9_new=910 if bc_case9_new==. & (strmatch( s_40013_10_0 , "174?" ))
replace bc_case9_new=910 if bc_case9_new==. & (strmatch( s_40013_10_0 , "174" ))
replace bc_case9_new=911 if bc_case9_new==. & (strmatch( s_40013_11_0 , "174?" ))
replace bc_case9_new=911 if bc_case9_new==. & (strmatch( s_40013_11_0 , "174" ))
replace bc_case9_new=912 if bc_case9_new==. & (strmatch( s_40013_12_0 , "174?" ))
replace bc_case9_new=912 if bc_case9_new==. & (strmatch( s_40013_12_0 , "174" ))
replace bc_case9_new=914 if bc_case9_new==. & (strmatch( s_40013_14_0 , "174?" ))
replace bc_case9_new=914 if bc_case9_new==. & (strmatch( s_40013_14_0 , "174" ))
** Defining the in-situ cases of breast
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_0_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_1_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_2_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_3_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_4_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_5_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_6_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_7_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_8_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_10_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_11_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_12_0 , "2330" ))
replace bc_case9_new=3 if bc_case9_new==. & (strmatch( s_40013_14_0 , "2330" ))
** Defining the Cancerous controls using ICD10  **
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_0_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_1_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_2_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_3_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_4_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_5_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_6_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_7_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_8_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_9_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_10_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_11_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_13_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_15_0 , "C???"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_16_0 , "C???"))
** 3 digits
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_0_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_1_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_2_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_3_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_4_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_5_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_6_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_7_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_8_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_9_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_10_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_11_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_13_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_15_0 , "C??"))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40006_16_0 , "C??"))
** Defining the Cancerous controls using ICD9  **
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_0_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_1_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_2_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_3_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_4_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_5_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_6_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_7_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_8_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_10_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_11_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_12_0 , "1???" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_14_0 , "1???" ))
** When only 3 digits in the coding **
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_0_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_1_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_2_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_3_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_4_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_5_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_6_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_7_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_8_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_10_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_11_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_12_0 , "1??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_14_0 , "1??" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_0_0 , "20??" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_1_0 , "20??" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_2_0 , "20??" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_3_0 , "20??" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_4_0 , "20??" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_5_0 , "20??" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_6_0 , "20??" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_7_0 , "20??" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_8_0 , "20??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_10_0 , "20??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_11_0 , "20??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_12_0 , "20??" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_14_0 , "20??" ))
** When only 3 digits in the coding **
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_0_0 , "20?" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_1_0 , "20?" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_2_0 , "20?" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_3_0 , "20?" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_4_0 , "20?" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_5_0 , "20?" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_6_0 , "20?" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_7_0 , "20?" ))
replace bc_case9_new =4 if bc_case9_new ==. & (strmatch( s_40013_8_0 , "20?" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_10_0 , "20?" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_11_0 , "20?" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_12_0 , "20?" ))
replace bc_case9_new=4 if bc_case9_new==. & (strmatch( s_40013_14_0 , "20?" ))
** Defining the Cancerous controls using Self   **
replace bc_case9_new=4 if bc_case9_new==. & n_20001_0_0!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_0_1!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_0_2!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_0_3!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_0_4!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_0_5!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_1_0!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_1_1!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_1_2!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_2_0!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_2_1!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_2_2!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_2_3!=.
replace bc_case9_new=4 if bc_case9_new==. & n_20001_2_4!=.
** Defining the insitu carcinoma in ICD10
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_0_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_1_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_2_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_3_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_4_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_5_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_6_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_7_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_8_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_9_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_10_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_11_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_13_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_15_0 , "D0??"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_16_0 , "D0??"))
** 3 digits
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_0_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_1_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_2_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_3_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_4_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_5_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_6_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_7_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_8_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_9_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_10_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_11_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_13_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_15_0 , "D0?"))
replace bc_case9_new=6 if bc_case9_new==. & (strmatch( s_40006_16_0 , "D0?"))
** ICD9
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_0_0 , "230?" "231?" "232?"
"233?""234?"))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_1_0 , "230?" "231?" "232?"
"233?""234?"))
```

```
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_2_0 , "230?" "231?" "232?"
"233?""234?" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_3_0 , "230?" "231?" "232?"
"233?""234?" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_4_0 , "230?" "231?" "232?"
"233?""234?" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_5_0 , "230?" "231?" "232?"
"233?""234?" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_6_0 , "230?" "231?" "232?"
"233?""234?" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_7_0 , "230?" "231?" "232?"
"233?""234?" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_8_0 , "230?" "231?" "232?"
"233?""234?" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_10_0 , "230?" "231?" "232?"
"233?""234?" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_11_0 , "230?" "231?" "232?"
"233?""234?" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_12_0 , "230?" "231?" "232?"
"233?""234?" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_14_0 , "230?" "231?" "232?"
"233?""234?" ))
** When only 3 digits in the coding **
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_0_0 , "230" "231" "232"
"233" "234"))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_1_0 , "230" "231" "232"
"233" "234" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_2_0 , "230" "231" "232"
"233" "234" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_3_0 , "230" "231" "232"
"233" "234" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_4_0 , "230" "231" "232"
"233" "234" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_5_0 , "230" "231" "232"
"233" "234" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_6_0 , "230" "231" "232"
"233" "234" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_7_0 , "230" "231" "232"
"233" "234"))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_8_0 , "230" "231" "232"
"233" "234" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_10_0 , "230" "231" "232"
"233" "234" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_11_0 , "230" "231" "232"
"233" "234" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_12_0 , "230" "231" "232"
"233" "234" ))
replace bc_case9_new =6 if bc_case9_new ==. & (strmatch( s_40013_14_0 , "230" "231" "232"
"233" "234" ))
** Defining the Neoplasms of unknown behaviour / nature
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_0_0 , "D37?" "D38?" "D39?"
"D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_1_0 , "D37?" "D38?" "D39?"
"D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_2_0 , "D37?" "D38?" "D39?"
"D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
```

```
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_3_0 , "D37?" "D38?" "D39?"
"D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_4_0 , "D37?" "D38?" "D39?"
"D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_5_0 , "D37?" "D38?" "D39?"
"D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_6_0 , "D37?" "D38?" "D39?"
"D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_7_0 , "D37?" "D38?" "D39?"
"D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_8_0 , "D37?" "D38?" "D39?"
"D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_9_0 , "D37?" "D38?" "D39?"
"D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_10_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_11_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_13_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_15_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_16_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
** 3 digits
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_0_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_1_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_2_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_3_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_4_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_5_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_6_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_7_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_8_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_9_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_10_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_11_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_13_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_15_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case9_new=7 if bc_case9_new==. & (strmatch( s_40006_16_0 , "D37" "D38" "D39"
"D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
** ICD9
```

```
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_0_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_1_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_2_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_3_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_4_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_5_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_6_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_7_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_8_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_10_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_11_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_12_0 , "235?" "236?" "237?"
"238?""239?"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_14_0 , "235?" "236?" "237?"
"238?""239?"))
** When only 3 digits in the coding **
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_0_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_1_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_2_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_3_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_4_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_5_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_6_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_7_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_8_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_10_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_11_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_12_0 , "235" "236" "237"
"238""239"))
replace bc_case9_new=7 if bc_case9_new ==. & (strmatch( s_40013_14_0 , "235" "236" "237"
"238""239"))
** Defining the other controls using non-cancer-self reported **
replace bc_case9_new=5 if no_icd10==0 & no_icd9==0 & no_self==0
** Defining the controls   **
replace bc_case9_new=0 if bc_case9_new==.
```

```
**  Define the incident and prevalent cases                            **
**  If age of diagnosis >= age when attended the centre = Incident case   **
**  If age of diagnosis < age when attended the centre = Prevalent case    **
**                  Based on ICD9
replace bc_case9_new=1 if bc_case9_new==900 & n_40008_0_0 >= n_21003_0_0 &
n_40008_0_0!=.
replace bc_case9_new=2 if bc_case9_new==900 & n_40008_0_0 < n_21003_0_0
** follow up = 1
replace bc_case9_new=1 if bc_case9_new==901 & n_40008_1_0 >= n_21003_0_0 &
n_40008_1_0!=.
replace bc_case9_new=2 if bc_case9_new==901 & n_40008_1_0 < n_21003_0_0
** follow up = 2
replace bc_case9_new=1 if bc_case9_new==902 & n_40008_2_0 >= n_21003_0_0 &
n_40008_2_0!=.
replace bc_case9_new=2 if bc_case9_new==902 & n_40008_2_0 < n_21003_0_0
** follow up = 3
replace bc_case9_new=1 if bc_case9_new==903 & n_40008_3_0 >= n_21003_0_0 &
n_40008_3_0!=.
replace bc_case9_new=2 if bc_case9_new==903 & n_40008_3_0 < n_21003_0_0
** follow up = 4
replace bc_case9_new=1 if bc_case9_new==904 & n_40008_4_0 >= n_21003_0_0 &
n_40008_4_0!=.
replace bc_case9_new=2 if bc_case9_new==904 & n_40008_4_0 < n_21003_0_0
** follow up = 5
replace bc_case9_new=1 if bc_case9_new==905 & n_40008_5_0 >= n_21003_0_0 &
n_40008_5_0!=.
replace bc_case9_new=2 if bc_case9_new==905 & n_40008_5_0 < n_21003_0_0
** follow up = 7
replace bc_case9_new=1 if bc_case9_new==907 & n_40008_7_0 >= n_21003_0_0 &
n_40008_7_0!=.
replace bc_case9_new=2 if bc_case9_new==907 & n_40008_7_0 < n_21003_0_0
label values bc_case9_new caselb
label var bc_case9_new "Cases and Controls bc ICD9"
** Cases and controls identified using self-reported variables
gen bc_case_self_new=.
***     Third by Self-reported for all the follow ups    ***
replace bc_case_self_new=400 if n_20001_0_0==1002 & bc_case_self_new==.
replace bc_case_self_new=401 if n_20001_0_1==1002 & bc_case_self_new==.
replace bc_case_self_new=402 if n_20001_0_2==1002 & bc_case_self_new==.
replace bc_case_self_new=403 if n_20001_0_3==1002 & bc_case_self_new==.
replace bc_case_self_new=404 if n_20001_0_4==1002 & bc_case_self_new==.
replace bc_case_self_new=405 if n_20001_0_5==1002 & bc_case_self_new==.
replace bc_case_self_new=410 if n_20001_1_0==1002 & bc_case_self_new==.
replace bc_case_self_new=411 if n_20001_1_1==1002 & bc_case_self_new==.
replace bc_case_self_new=412 if n_20001_1_2==1002 & bc_case_self_new==.
replace bc_case_self_new=420 if n_20001_2_0==1002 & bc_case_self_new==.
replace bc_case_self_new=421 if n_20001_2_1==1002 & bc_case_self_new==.
replace bc_case_self_new=422 if n_20001_2_2==1002 & bc_case_self_new==.
replace bc_case_self_new=423 if n_20001_2_3==1002 & bc_case_self_new==.
replace bc_case_self_new=424 if n_20001_2_4==1002 & bc_case_self_new==.
** Defining the Cancerous controls using ICD10  **
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_0_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_1_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_2_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_3_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_4_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_5_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_6_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_7_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_8_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_9_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_10_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_11_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_13_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_15_0 , "C???"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_16_0 , "C???"))
** 3 digits
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_0_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_1_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_2_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_3_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_4_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_5_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_6_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_7_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_8_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_9_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_10_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_11_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_13_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_15_0 , "C??"))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40006_16_0 , "C??"))
** Defining the Cancerous controls using ICD9  **
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_0_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_1_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_2_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_3_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_4_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_5_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_6_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_7_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_8_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_10_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_11_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_12_0 , "1???" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_14_0 , "1???" ))
** When only 3 digits in the coding **
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_0_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_1_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_2_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_3_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_4_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_5_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_6_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_7_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_8_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_10_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_11_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_12_0 , "1??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_14_0 , "1??" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_0_0 , "20??" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_1_0 , "20??" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_2_0 , "20??" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_3_0 , "20??" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_4_0 , "20??" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_5_0 , "20??" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_6_0 , "20??" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_7_0 , "20??" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_8_0 , "20??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_10_0 , "20??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_11_0 , "20??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_12_0 , "20??" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_14_0 , "20??" ))
** When only 3 digits in the coding **
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_0_0 , "20?" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_1_0 , "20?" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_2_0 , "20?" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_3_0 , "20?" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_4_0 , "20?" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_5_0 , "20?" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_6_0 , "20?" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_7_0 , "20?" ))
replace bc_case_self_new =4 if bc_case_self_new ==. & (strmatch( s_40013_8_0 , "20?" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_10_0 , "20?" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_11_0 , "20?" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_12_0 , "20?" ))
replace bc_case_self_new=4 if bc_case_self_new==. & (strmatch( s_40013_14_0 , "20?" ))
** Defining the Cancerous controls using Self   **
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_0_0!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_0_1!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_0_2!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_0_3!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_0_4!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_0_5!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_1_0!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_1_1!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_1_2!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_2_0!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_2_1!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_2_2!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_2_3!=.
replace bc_case_self_new=4 if bc_case_self_new==. & n_20001_2_4!=.
** Defining the insitu carcinoma in ICD10
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_0_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_1_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_2_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_3_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_4_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_5_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_6_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_7_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_8_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_9_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_10_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_11_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_13_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_15_0 , "D0??"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_16_0 , "D0??"))
** 3 digits
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_0_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_1_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_2_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_3_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_4_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_5_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_6_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_7_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_8_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_9_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_10_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_11_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_13_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_15_0 , "D0?"))
replace bc_case_self_new=6 if bc_case_self_new==. & (strmatch( s_40006_16_0 , "D0?"))
** ICD9
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_0_0 , "230?" "231?"
"232?" "233?""234?"))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_1_0 , "230?" "231?"
"232?" "233?""234?"))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_2_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_3_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_4_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_5_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_6_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_7_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_8_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_10_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_11_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_12_0 , "230?" "231?"
"232?" "233?""234?" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_14_0 , "230?" "231?"
"232?" "233?""234?" ))
** When only 3 digits in the coding **
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_0_0 , "230" "231"
"232" "233" "234"))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_1_0 , "230" "231"
"232" "233" "234" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_2_0 , "230" "231"
"232" "233" "234" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_3_0 , "230" "231"
"232" "233" "234" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_4_0 , "230" "231"
"232" "233" "234" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_5_0 , "230" "231"
"232" "233" "234" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_6_0 , "230" "231"
"232" "233" "234" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_7_0 , "230" "231"
"232" "233" "234"))
```

```
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_8_0 , "230" "231"
"232" "233" "234" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_10_0 , "230" "231"
"232" "233" "234" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_11_0 , "230" "231"
"232" "233" "234" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_12_0 , "230" "231"
"232" "233" "234" ))
replace bc_case_self_new =6 if bc_case_self_new ==. & (strmatch( s_40013_14_0 , "230" "231"
"232" "233" "234" ))
** Defining the Neoplasms of unknown behaviour / nature
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_0_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_1_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_2_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_3_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_4_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_5_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_6_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_7_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_8_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_9_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_10_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_11_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_13_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_15_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_16_0 , "D37?" "D38?"
"D39?" "D40?" "D41?" "D42?" "D43?" "D44?" "D45?" "D46?" "D47?" "D48?"))
** 3 digits
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_0_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_1_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_2_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_3_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_4_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_5_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_6_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
```

```
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_7_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_8_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_9_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_10_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_11_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_13_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_15_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
replace bc_case_self_new=7 if bc_case_self_new==. & (strmatch( s_40006_16_0 , "D37" "D38"
"D39" "D40" "D41" "D42" "D43" "D44" "D45" "D46" "D47" "D48"))
** ICD9
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_0_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_1_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_2_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_3_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_4_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_5_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_6_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_7_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_8_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_10_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_11_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_12_0 , "235?" "236?"
"237?" "238?""239?"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_14_0 , "235?" "236?"
"237?" "238?""239?"))
** When only 3 digits in the coding **
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_0_0 , "235" "236"
"237" "238""239"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_1_0 , "235" "236"
"237" "238""239"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_2_0 , "235" "236"
"237" "238""239"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_3_0 , "235" "236"
"237" "238""239"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_4_0 , "235" "236"
"237" "238""239"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_5_0 , "235" "236"
"237" "238""239"))
```

```
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_6_0 , "235" "236"
"237" "238""239"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_7_0 , "235" "236"
"237" "238""239"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_8_0 , "235" "236"
"237" "238""239"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_10_0 , "235" "236"
"237" "238""239"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_11_0 , "235" "236"
"237" "238""239"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_12_0 , "235" "236"
"237" "238""239"))
replace bc_case_self_new=7 if bc_case_self_new ==. & (strmatch( s_40013_14_0 , "235" "236"
"237" "238""239"))
** Defining the other controls using non-cancer-self reported **
replace bc_case_self_new=5 if no_icd10==0 & no_icd9==0 & no_self==0
** Defining the controls   **
replace bc_case_self_new=0 if bc_case_self_new==.
**  Define the incident and prevalent cases                              **
**  If age of diagnosis >= age when attended the centre = Incident case   **
**  If age of diagnosis < age when attended the centre = Prevalent case    **
**                 Based on self-reported
replace bc_case_self_new=1 if bc_case_self_new==412 & n_20007_1_2 >= n_21003_0_0 &
n_20007_1_2!=.
replace bc_case_self_new=2 if bc_case_self_new==412 & n_20007_1_2 < n_21003_0_0 &
n_20007_1_2>0
replace bc_case_self_new=1 if bc_case_self_new==411 & n_20007_1_1 >= n_21003_0_0 &
n_20007_1_1!=.
replace bc_case_self_new=2 if bc_case_self_new==411 & n_20007_1_1 < n_21003_0_0 &
n_20007_1_1>0
replace bc_case_self_new=1 if bc_case_self_new==410 & n_20007_1_0 >= n_21003_0_0 &
n_20007_1_0!=.
replace bc_case_self_new=2 if bc_case_self_new==410 & n_20007_1_0 < n_21003_0_0 &
n_20007_1_0>0
** follow up = 20
replace bc_case_self_new=1 if bc_case_self_new==420 & n_20007_2_0 >= n_21003_0_0 &
n_20007_2_0!=.
replace bc_case_self_new=2 if bc_case_self_new==420 & n_20007_2_0 < n_21003_0_0 &
n_20007_2_0>0
** follow up = 21
replace bc_case_self_new=1 if bc_case_self_new==421 & n_20007_2_1 >= n_21003_0_0 &
n_20007_2_1!=.
replace bc_case_self_new=2 if bc_case_self_new==421 & n_20007_2_1 < n_21003_0_0 &
n_20007_2_1>0
** follow up = 3
replace bc_case_self_new=1 if bc_case_self_new==403 & n_20007_0_3 >= n_21003_0_0 &
n_20007_0_3!=.
replace bc_case_self_new=2 if bc_case_self_new==403 & n_20007_0_3 < n_21003_0_0 &
n_20007_0_3>0
** follow up = 2
replace bc_case_self_new=1 if bc_case_self_new==402 & n_20007_0_2 >= n_21003_0_0 &
n_20007_0_2!=.
replace bc_case_self_new=2 if bc_case_self_new==402 & n_20007_0_2 < n_21003_0_0 &
n_20007_0_2>0
** follow up = 1
replace bc_case_self_new=1 if bc_case_self_new==401 & n_20007_0_1 >= n_21003_0_0 &
n_20007_0_1!=.
```

```
replace bc_case_self_new=2 if bc_case_self_new==401 & n_20007_0_1 < n_21003_0_0 &
n_20007_0_1>0
** follow up = 0
replace bc_case_self_new=1 if bc_case_self_new==400 & n_20007_0_0 >= n_21003_0_0 &
n_20007_0_0!=.
replace bc_case_self_new=2 if bc_case_self_new==400 & n_20007_0_0 < n_21003_0_0 &
n_20007_0_0>0
label values bc_case_self_new caselb
label var bc_case_self_new "Cases and Controls bc self"
** Combining the three sources of cases and controls identification
 ** defining cases / controls using ICD10,9, self
** defining new cases and controls without any death in the dataset
g bc_new=.
replace bc_new=1 if (bc_case10_new==1 | bc_case9_new==1 | bc_case_self_new==1) & dead==0
replace bc_new=2 if bc_new==. & (bc_case10_new==2 | bc_case9_new==2 |
bc_case_self_new==2) & dead==0
replace bc_new=3 if bc_new==. & (bc_case10_new==5 | bc_case9_new==5 |
bc_case_self_new==5) & dead==0
replace bc_new=4 if bc_new==. & dead==0
replace bc_new=5 if dead==1
label define casenlb  1"Incident" 2"Prevalent" 3"controls without cancer" 4"Cancer Controls"
5"Dead" 6"breast case but couldn’t be identified inc or prev" , modify
label values bc_new casenlb
label var bc_new "Cases and Controls using app 5791"
gen bc_analysis=.
replace bc_analysis=0 if bc_new==3
replace bc_analysis=1 if bc_new==1
label define clb 0"Controls" 1"Incident cases", modify
label values bc_analysis clb
gen dead=0
replace dead=1 if s_40001_0_0!=""
replace dead=1 if s_40001_1_0!=""
replace dead=1 if s_40002_0_0!=""
replace dead=1 if s_40002_0_1!=""
replace dead=1 if s_40002_0_2!=""
replace dead=1 if s_40002_0_3!=""
replace dead=1 if s_40002_0_4!=""
replace dead=1 if s_40002_0_5!=""
replace dead=1 if s_40002_0_6!=""
replace dead=1 if s_40002_0_7!=""
replace dead=1 if s_40002_0_8!=""
replace dead=1 if s_40002_0_9!=""
replace dead=1 if s_40002_0_10!=""
replace dead=1 if s_40002_0_11!=""
replace dead=1 if s_40002_0_13!=""
replace dead=1 if s_40002_1_0!=""
replace dead=1 if s_40002_1_1!=""
replace dead=1 if s_40002_1_2!=""
replace dead=1 if s_40002_1_3!=""
replace dead=1 if s_40002_1_4!=""
replace dead=1 if s_40002_1_5!=""
replace dead=1 if s_40002_1_6!=""
label define deadlb 0"alive" 1"dead"
label values dead deadlb
gen bc_alone_nodead=bc_alone
replace bc_alone_nodead=. if dead==1
```

```
***2- Family history coding

** 2- Family history of breast cancer
** Family history of breast cancer _ mother (only don't know and prefer not to say coded as
missing while . coded as not having FH)

gen fh_mot_BC=0
replace fh_mot_BC=. if  n_20110_0_0==-11 | n_20110_0_0==-13 | n_20110_0_0==-17
replace fh_mot_BC=1 if  n_20110_0_0==5 | n_20110_0_1==5 | n_20110_0_2==5|
n_20110_0_3==5| n_20110_0_4==5| n_20110_0_5==5| n_20110_0_6==5| n_20110_0_7==5|
n_20110_0_8==5| n_20110_0_9==5| n_20110_0_10==5 // 5 indicates breast cancer among mother
replace fh_mot_BC=1 if  n_20110_1_0==5 | n_20110_1_1==5 | n_20110_1_2==5|
n_20110_1_3==5| n_20110_1_4==5| n_20110_1_5==5| n_20110_1_6==5 // 5 indicates breast
cancer among mother
replace fh_mot_BC=1 if  n_20110_2_0==5 | n_20110_2_1==5 | n_20110_2_2==5|
n_20110_2_3==5| n_20110_2_4==5| n_20110_2_5==5| n_20110_2_6==5 // 5 indicates breast
cancer among mother
label define fh_mlb 0"No Mother history of BC " 1"Mother history of BC"
label values fh_mot_BC fh_mlb
label var fh_mot_BC "Mother history of Breast cancer"
** Family history of breast cancer _ Sibling (only don't know and prefer not to say coded as
missing while . coded as not having FH)
gen fh_sib_BC=0
replace fh_sib_BC=. if  n_20111_0_0==-11 |  n_20111_0_0==-13 |  n_20111_0_0==-17
replace fh_sib_BC=1 if  n_20111_0_0==5 | n_20111_0_1==5 | n_20111_0_2==5|
n_20111_0_3==5| n_20111_0_4==5| n_20111_0_5==5| n_20111_0_6==5| n_20111_0_7==5|
n_20111_0_8==5 | n_20111_0_9==5| n_20111_0_10==5| n_20111_0_11==5 // 5 indicates breast
cancer among mother
replace fh_sib_BC=1 if  n_20111_1_0==5 | n_20111_1_1==5 | n_20111_1_2==5|
n_20111_1_3==5| n_20111_1_4==5| n_20111_1_5==5| n_20111_1_6==5| n_20111_1_7==5    // 5
indicates breast cancer among mother
replace fh_sib_BC=1 if  n_20111_2_0==5 | n_20111_2_1==5 | n_20111_2_2==5|
n_20111_2_3==5| n_20111_2_4==5| n_20111_2_5==5| n_20111_2_6==5| n_20111_2_7==5    // 5
indicates breast cancer among mother
label define fh_slb 0"No Sibling history of BC " 1"Sibling history of BC"
label values fh_sib_BC fh_slb
label var fh_sib_BC "Sibling history of Breast cancer"
** Combined Family history of BC mother and siblings (None , mother or sib , mother and sib)
gen com_fh=0
replace com_fh=1 if fh_mot_BC==1 | fh_sib_BC==1
replace com_fh=2 if fh_mot_BC==1 & fh_sib_BC==1
replace com_fh=. if fh_mot_BC==. & fh_sib_BC==.
label define comfhlb 0"No FH of BC " 1"Mother or Sibling history of BC " 2"Mother and Sibling
history of BC"
label values com_fh comfhlb
label var com_fh "Combined FH of Breast cancer none/M or S/ M and S)"
** Combined Family history of BC mother and siblings (Yes, No)
gen com_fh1= com_fh
replace com_fh1=1 if com_fh==1 | com_fh==2
label define com_fhlb1 0"No" 1"Yes"
label values com_fh1 com_fhlb1
label var com_fh1 "Combined FH of Breast cancer Yes/No"
** Combined Family history of BC mother and siblings (None , mother alone, sib alone , mother
and sib)
g fh=.
replace fh=0 if com_fh1==0
replace fh=1 if fh_mot_BC==1 | fh_sib_BC==1
replace fh=2 if fh_mot_BC==1 & fh_sib_BC==0
replace fh=3 if fh_sib_BC==1 & fh_mot_BC==0
replace fh=4 if fh_mot_BC==1 & fh_sib_BC==1
label var fh "FH mot/ sib alone then together"
label define fh_lb 0"No FH" 1"Mother or Sibling" 2"Mother only" 3"Sibling only" 4"Mother and Sibling history"
label values fh fh_lb


*** 3- Anthropometric variables coding

** 3- Anthropometric variables coding
** BMI group
gen bmi_cat=.
replace bmi_cat=0 if n_21001_0_0 >=18.500 & n_21001_0_0<=24.999 & n_21001_0_0!=.
replace bmi_cat=1 if n_21001_0_0 >=25.000 & n_21001_0_0<=29.999 & n_21001_0_0!=.
replace bmi_cat=2 if n_21001_0_0 >=30.000 & n_21001_0_0!=.
replace bmi_cat=3 if n_21001_0_0 < 18.500  & n_21001_0_0!=.
replace bmi_cat=n_21001_1_0 if bmi_cat==.
replace bmi_cat=n_21001_2_0 if bmi_cat==.
replace bmi_cat=0 if bmi_cat >=18.500 & bmi_cat<=24.999 & bmi_cat!=.
replace bmi_cat=1 if bmi_cat >=25.000 & bmi_cat<=29.999 & bmi_cat!=.
replace bmi_cat=2 if bmi_cat >=30.000 & bmi_cat!=.
replace bmi_cat=3 if bmi_cat < 18.500  & bmi_cat>=10
label define bmilb 0"Healthy"  1"Overweight" 2"Obese" 3"Underweight", modify
label var bmi_cat "BMI cat var"
label values bmi_cat bmilb
** BMI continous variable
gen bmi_num=n_21001_0_0
replace bmi_num=n_21001_1_0 if bmi_num==.
replace bmi_num=n_21001_2_0 if bmi_num==.
label var bmi_num "BMI as numerical value"
** Sitting height **
g sit_ht= n_20015_0_0
replace sit_ht= n_20015_1_0 if sit_ht==.
replace sit_ht= n_20015_2_0 if sit_ht==.
label var sit_ht "Sitting ht of the three follow ups"
** Standing height **
g stand_ht= n_50_0_0
replace stand_ht= n_50_1_0 if stand_ht==.
replace stand_ht= n_50_2_0 if stand_ht==.
label var stand_ht "Standing ht of the three follow ups"
** Standing height categories
g ht_cat_females=.
sum stand_ht if bc_new==3 & n_31_0_0==0  // mean of the control group among females only
replace ht_cat_females=0 if stand_ht< 149.78355
replace ht_cat_females=1 if stand_ht>= 149.78355 & stand_ht<= 175.08125
replace ht_cat_females=2 if stand_ht> 175.08125 & stand_ht!=.
label variable ht_cat "Categorical standing ht (mean +- SD of controls)"
label define htlb 0"Below the mean group" 1"Within the mean group"  2"Above the mean group" , replace
label values ht_cat_females htlb
** Waist to hip ratio as numeric value
gen waist_to_hip= n_48_0_0/n_49_0_0 if n_48_0_0!=. & n_49_0_0!=.
replace waist_to_hip= n_48_1_0/n_49_1_0 if waist_to_hip==.
replace waist_to_hip= n_48_2_0/n_49_2_0 if waist_to_hip==.
label var waist_to_hip "waist to hip as numeric variable"
```

```
** Waist to hip ratio as groups
gen wth_gp=.
replace wth_gp=0 if waist_to_hip <= 0.80000
replace wth_gp=1 if waist_to_hip>0.80000 & waist_to_hip<=0.85000
replace wth_gp=2 if waist_to_hip>0.85000 & waist_to_hip!=.
label define wthlb 0"Low" 1"Moderate" 2"High" , modify
label values wth_gp wthlb
label var wth_gp "Waist to Hip group"

*** 4- Reproductive variables coding

** 4- Reproductive variables coding
** 1- Menarche age
** generate menarche age var without -3 and -1
gen menarche_age1= n_2714_0_0
replace menarche_age1=. if n_2714_0_0==-3 | n_2714_0_0==-1
label var menarche_age1 "menarche age without -3 & -1"
** Menarche age group 1: generate menarche age var with 3 groups and without -3 and -1
gen menarche_gp= n_2714_0_0
replace menarche_gp=0 if n_2714_0_0>0 & n_2714_0_0<=10
replace menarche_gp=1 if n_2714_0_0>10 & n_2714_0_0<15
replace menarche_gp=2 if n_2714_0_0>=15 & n_2714_0_0!=.
replace menarche_gp=. if n_2714_0_0==-3 | n_2714_0_0==-1   ///** missing data
label define menarlb 0"<= 10yrs" 1"11-15yrs" 2">=15yrs"
label values menarche_gp menarlb
label var menarche_gp "menarche groups"
** Menarche age group 2:generate menarche age var with 3 groups and without -3 and -1
gen menar_age_11y = .
replace menar_age_1=0 if menarche_age1>11 & menarche_age1!=.
replace menar_age_1=1 if menarche_age1<=11 & menarche_age1!=.
label define menar11 0">11 years old" 1"<=11 years old"
label values menar_age_1 menar11
label var menar_age_1 "Menarche age two gps > o < 11 years"
** Menarche age group 3:generate menarche age var with 3 groups and without -3 and -1
gen menar_age_13y = .
replace menar_age_13y=0 if menarche_age1>13 & menarche_age1!=.
replace menar_age_13y=1 if menarche_age1<=13 & menarche_age1!=.
label define menar11 0">13 years old" 1"<=13 years old", replace
label values menar_age_13y menar11
label var menar_age_13y "Menarche age two gps > o < 13 years"
** 2- First birth age
** recode the age at first birth without -3 and -4
gen livebth_st_age1= n_2754_0_0
replace livebth_st_age1=. if n_2754_0_0==-3 | n_2754_0_0==-4
label var livebth_st_age1 "first birth age counting no children as 0"
** 3- Bump variable = Age at first birth - Age at menarche **
gen bump= livebth_st_age - menarche_age1 if livebth_st_age!=. // minus values will be created
among women with no children
replace bump=99 if bump<0 // all 99 means thy didn’t had children to cal the bump
replace bump=99 if parity_gp==0 & bump==.  // ages were missing but they didn’t have children
label var bump "first birth age - menarche age"
sum bump if bump!=99 & bc_both==0 , d  // to get the percentiles for the groups from controls
only (bc_alone)
gen bump_gp= .
replace bump_gp= 0 if bump>= 0 & bump<=12
replace bump_gp= 1 if bump> 12 & bump<=16
replace bump_gp= 2 if bump> 16 & bump!=99
replace bump_gp= 4 if bump==99
label var bump_gp "Bump group divided by percentile to  groups"
label define bumlb 0"Low" 1"Moderate" 2"High" 3"Very High" 4"No children", replace
label values bump_gp bumlb

** 4- Menopause age
** generate menopause age var without -3 and -1
gen menopause_age1= n_3581_0_0
replace  menopause_age1=. if n_3581_0_0==-3 | n_3581_0_0==-1
label var menopause_age1 "menopause age without -3 & -1"
** 5- Duration of pre_menopause = menopause age - menarche age
gen pre_duration= menopause_age1 - menarche_age1 if  menopause_age1!=.
** 6- Duration of post_menopause = ag when attended the centre - menopause age
gen post_duration= n_21003_0_0 - menopause_age1  if  menopause_age1!=.
** 7- HRT use : recode had HRT var without -3 and -1
gen HRT_gp= n_2814_0_0
replace HRT_gp=. if n_2814_0_0==-3 | n_2814_0_0==-1
label var HRT_gp "Had HRT without -3 & -1"
label define HRTlb 0"No" 1"Yes"
label values HRT_gp HRTlb
** 8- HRT start age: generate HRT starting age var without -3 and -1
gen HRTst_age1= n_3536_0_0
replace  HRTst_age1=. if n_3536_0_0==-3 | n_3536_0_0==-1
replace  HRTst_age1=n_3536_1_0 if HRTst_age1==. & n_3536_1_0!=-3 & n_3536_1_0!=-1
label var HRTst_age1 "HRT starting age without -3 & -1"
** 9- HRT start age:generate HRT last age var without -3 and -1
gen HRTend_age1= n_3546_0_0
replace  HRTend_age1=. if n_3546_0_0==-3 | n_3546_0_0==-1
replace  HRTend_age1=n_3546_1_0 if HRTend_age1==. & n_3546_1_0!=-3 & n_3546_1_0!=-1
label var HRTend_age1 "HRT last used age without -3 & -1"
** 10- HRT duration: duration of HRT usage = last age used HRT - age start using HRT
gen HRT_duration=  HRTend_age1 - HRTst_age1 if  HRTend_age1!=. & HRTst_age1!=. &
HRTend_age1!=-11
replace HRT_duration=0 if HRT_gp==0 & HRT_duration==.
replace HRT_duration=. if HRT_duration<0
 /// for any result appeared as minus (means end age was younger than start age)
label var HRT_duration "HRT use duration from two follow ups"
** 11-OC use: recode had Contraceptive pills var without -3 and -1
gen contra_gp= n_2784_0_0
replace contra_gp=. if n_2784_0_0==-3
replace contra_gp=. if n_2784_0_0==-1
label var contra_gp "Had contraceptive pills without -3 & -1"
label values contra_gp HRTlb
** 12- OC start age: generate Contraceptive pills start var without -3 and -1
gen contra_st_age1= n_2794_0_0
replace  contra_st_age1=. if n_2794_0_0==-3
replace  contra_st_age1=. if n_2794_0_0==-1
replace  contra_st_age1=n_2794_1_0 if contra_st_age1==. & n_2794_1_0!=-3 & n_2794_1_0!=-1
/// add values from second follow-up
label var contra_st_age1 "Contraceptive pills starting age without -3 & -1"
** 13- OC end age: generate Contraceptive pills last age var without -3 and -1 (-11 is still on pills)
gen contra_end_age1= n_2804_0_0
replace  contra_end_age1=. if n_2804_0_0==-3 | n_2804_0_0==-1
replace  contra_end_age1=n_2804_1_0 if  contra_end_age1==. & n_2804_1_0!=-3 &
n_2804_1_0!=-1
/// add values from second follow-up
label var contra_end_age1 "Contraceptive pills last used age without -3 & -1"
```

```
** 14- OC duration: duration of Contraceptive pills usage = eng age - start age if they use OC
gen Conta_duration=  contra_end_age1 - contra_st_age1 if  contra_end_age1!=. & contra_end_age1!=-11 & contra_st_age1!=.
replace Conta_duration=0 if  n_2784_0_0==0 & Conta_duration==.
replace Conta_duration=. if Conta_duration<0 /// for any result appeared as minus (means end age was younger than start age)
label var Conta_duration "Contraceptive use duration from two follow ups"
** 15- OC duration group: duration of Contraceptive pills usage
g oc_dur_gp=.
replace oc_dur_gp=1 if  Conta_duration<5 & Conta_duration>=0
replace oc_dur_gp=0 if contra_gp==0
replace oc_dur_gp=2 if  Conta_duration>=5 & Conta_duration!=.
label define durlb 0"No OC" 1"< 5 yrs" 2">= 5yrs"
label values oc_dur_gp durlb
label var oc_dur_gp "OC duration groups >or < 5 yrs"
** 16- number of live births: recode number of live births or parity
gen live_birth_num= n_2734_0_0
replace  live_birth_num=.  if n_2734_0_0 ==-3 | n_2804_0_0==-4
label var live_birth_num "Number of live births without -3 & -4"
** 17- Age at first live birth: recode age at fist live birth var without -3 & -4
gen livebth_st_age1= n_2754_0_0
replace  livebth_st_age1=. if n_2754_0_0==-3 | n_2754_0_0==-4
label var livebth_st_age1 "First live birth age without -3 & -4"
** 18- Age at first live birth group: recode age at fist live birth group var
gen livebth_st_gp= livebth_st_age1
replace  livebth_st_gp=0 if livebth_st_age1 <20
replace  livebth_st_gp=1 if livebth_st_age1 >=20 & livebth_st_age1 <=24
replace  livebth_st_gp=2 if livebth_st_age1 >=25 & livebth_st_age1 <=29
replace  livebth_st_gp=3 if livebth_st_age1 >=30 & livebth_st_age1!=.
label var livebth_st_gp "First live birth age group"
label define livelb 0"<20 yrs" 1"20-24 yrs" 2"25-29yrs" 3">=30yrs"
label values livebth_st_gp livelb
** 19- Age at last live birth:recode age at last live birth var without -3 & -4
gen livebth_end_age1= n_2764_0_0
replace  livebth_end_age1=. if n_2764_0_0==-3 | n_2764_0_0==-4
label var livebth_end_age1 "last live birth age without -3 & -4"
** 20- Ever had stillbirth or termination group
gen stillbrth_gp= n_2774_0_0
replace  stillbrth_gp=. if n_2774_0_0==-3 | n_2774_0_0==-1
label var stillbrth_gp "Ever had stillbirth or termination without -3 & -1"
** 21- Termination number: recode number of pregnancy termination
gen termination_num= n_3849_0_0
replace  termination_num=. if n_3849_0_0==-3 | n_3849_0_0==-1
label var termination_num "Number of  pregnancy termination -3 & -1"
** 22- Miscarriages number: recode number of spontaneous miscarriages
gen spon_misc_num= n_3839_0_0
replace  spon_misc_num=. if n_3839_0_0==-3 | n_3839_0_0==-1
label var spon_misc_num "Number of spontaneous miscarriage -3 & -1"
** 23- Hysterectomy history: recode Ever had a hysterectomy
gen hyster_gp= n_3591_0_0
replace  hyster_gp=. if n_3591_0_0==-3 | n_3591_0_0==-5
replace  hyster_gp=n_3591_1_0 if hyster_gp==.
replace  hyster_gp=. if hyster_gp==-3 | hyster_gp==-5 // to remove any -3/-1 from second cohort
label var hyster_gp "Ever had a hysterectomy without -3 & -5"
 label values hyster_gp HRTlb

** 24- Hysterectomy age: recode age at hysterectomy
```

```
gen hyst_age= n_2824_0_0
replace  hyst_age=. if n_2824_0_0==-3 | n_2824_0_0==-1
label var hyst_age "Age at hysterectomy without  -3 & -1"
** 25- Hysterectomy age group:Cat age of hyst
g hys_age_gp=.
replace hys_age_gp=0 if hyster_gp==0
replace hys_age_gp=1 if hyst_age<30 & hyst_age>0
replace hys_age_gp=2 if hyst_age>=30 & hyst_age<34
replace hys_age_gp=3 if hyst_age>=34 & hyst_age<39
replace hys_age_gp=4 if hyst_age>=39 & hyst_age<=44
replace hys_age_gp=5 if hyst_age>=45 & hyst_age!=.
label define hysagelb 0"No Hys" 1"<30 yrs" 2"30-34" 3"35-39" 4"40-44" 5">=45"
label values hys_age_gp hysagelb
label var hys_age_gp "Hysterectomy age categories"
** 26- Oophorectomy history: recode Ever had an oophorectomy
gen bi_ooph_gp= n_2834_0_0
replace bi_ooph_gp=. if n_2834_0_0==-3 | n_2834_0_0==-5
label var bi_ooph_gp "Ever had a Bilateral oophorectomy  without -3 & -5"
label values bi_ooph_gp HRTlb
** 27- Oophorectomy age: recode age at bilateral ooph
gen bi_ooph_age= n_3882_0_0
replace  bi_ooph_age=. if n_3882_0_0==-3 | n_3882_0_0==-1
label var hyst_age "Age at Bilateral oophorectomy without  -3 & -1"
** 28- Oophorectomy age group: Cat age of ooph
g bi_age_gp=.
replace bi_age_gp=0 if bi_ooph_gp==0
replace bi_age_gp=1 if bi_ooph_age<30
replace bi_age_gp=2 if bi_ooph_age>=30 & bi_ooph_age<34
replace bi_age_gp=3 if bi_ooph_age>=34 & bi_ooph_age<39
replace bi_age_gp=4 if bi_ooph_age>=39 & bi_ooph_age<=44
replace bi_age_gp=5 if bi_ooph_age>=45 & bi_ooph_age!=.
label define biagelb 0"No ooph" 1"<30 yrs" 2"30-34" 3"35-39" 4"40-44" 5">=45"
label values bi_age_gp biagelb
label var bi_age_gp "Bilateral oophorectomy age categories"
** 29- Combined history of Hysterectomy and Oophorectomy: combined variable of hys+ooph
g hys_bi_com=0
replace hys_bi_com=. if hyster_gp==. & bi_ooph_gp==.
replace hys_bi_com=1 if hyster_gp==1 | bi_ooph_gp==1
replace hys_bi_com=2 if hyster_gp==1 & bi_ooph_gp==1
label define comlb 0 "No hys No bi" 1"One of them" 2"Both of them"
label values hys_bi_com comlb
label var hys_bi_com "Combined cat var for hys + ooph"
** 30- Parity status group 1 (yes/no)
gen parity_gp= n_2734_0_0
replace parity_gp=. if n_2734_0_0==-3
replace parity_gp=1 if n_2734_0_0>0 & n_2734_0_0!=.
label values parity_gp HRTlb
** 31- Parity group 1 (None, 1-2, 3-4, >4 children)
g parity_1=.
replace parity_1=0 if  live_birth_num==0
replace parity_1=1 if live_birth_num==1 | live_birth_num==2
replace parity_1=2 if live_birth_num>2 & live_birth_num<=4
replace parity_1=3 if  live_birth_num>4 & live_birth_num!=.
label define parlb 0 "No children" 1"1-2 children" 2"3-4 children" 3"> 4 children"
label values parity_1 parlb
label var parity_1 "New parity group"
```

```
*** 5- Menopausal status coding

** 5- Menopausal status coding
** The original Menopausal Status
gen meno_status= n_2724_0_0
replace meno_status=. if  n_2724_0_0==-3  // -3:Prefer not to answer consider as missing
replace meno_status=1 if  n_2724_0_0==2   // 2:Not sure - had a hysterectomy consider as had menopause
replace meno_status=1 if  n_2724_0_0==3 &  n_2724_1_0==1   // recode 3: not sure other reasons to Yes if in the next follow up they answered as yes
replace meno_status=1 if  n_2724_0_0==3 &  n_2724_1_0==2   // recode 3: not sure other reasons to Yes if in the next follow up they answered as Not sure had hysterectomy
replace meno_status=0 if  n_2724_0_0==3 &  n_2724_1_0==0   // recode 3: not sure other reasons to No if in the next follow up they answered as No
replace meno_status=. if  n_2724_0_0==3 &  n_2724_1_0==-3   // recode 3: not sure other reasons to Missing if in the next follow up they answered as prefer not to say
replace meno_status=. if  n_2724_0_0==3 &  n_2724_1_0==3   // recode 3: not sure other reasons to Missing if in the next follow up they answered as not sure other reasons
replace meno_status=. if  n_2724_0_0==3 &  n_2724_1_0==.   // recode 3: not sure other reasons to Missing if in the next follow up they answered as not sure other reasons
label var meno_status "Menopause status from two follow ups"
label define menolb 0"Pre_menopause" 1"Post_menopause"
label values meno_status menolb
** Updated menopausal status based on NHS recommendations
g meno_status_new = .
replace meno_status_new= 0 if meno_status==0 & menarche_age1>=7 & hys_bi_com==0 & n_21003_0_0 <55
replace meno_status_new= 1 if meno_status==1 & hys_bi_com==0 & menopause_age1 >=40 & menopause_age1 !=.
label define meno_lb 0"Pre_menopause" 1"Post_menopause", replace
label values meno_status_new meno_lb
label var meno_status_new "New menopausal status"
***************************************************************************************
```

# UK Biobank

# Accessing UK Biobank Data

Version 2.1

http://www.ukbiobank.ac.uk/

February 2020

This document details the means by which data supplied by UK Biobank can be obtained and manipulated once access has been approved.

## Contents

# 0 What’s new in this version

This section summarises the changes in this document between version 2.0 and this version 2.1.

- This new “What’s new in this version” has been added as Section 0.

- The name given to the keyfile sent in the notification email has been changed to include the run id as well as the application id. So where before it might have had a name like k56789.key, it will now be called something like k56789r23456.key, where 56789 is the relevant application id and 23456 the specific run id of this dataset.

  This is to make it clear that to decrypt a main dataset it is necessary to use the keyfile specific to that run. The sections of this document that refer to the keyfile have been altered to reflect this change.

- In section 8.2.5 regarding ukbgene and how to run Linux if you only have a Windows computer available, the advice has been altered to provide a more straightforward way of doing this than previously.

# 1 Introduction

## 1.1 The notification email

This guide is intended for researchers who have had an application for access to UK Biobank data, or a request for additional data on an existing project, approved and have received a notification email that their data is now available for download.

Note that only the Principal Investigator (PI) of a project and those collaborators with “delegate” status on the Application Management System (AMS) will receive the notification email and be able to download the main dataset and access the Data Portal on Showcase. The project PI can assign delegate rights to other collaborators in the Collaborators tab on the AMS.

The notification email will contain a 32-character MD5 Checksum within the main body of the email. This is needed to download the main dataset.

It will also have an attachment, with a name of the form: k56789r23456.key where 56789 will be replaced by the relevant application id and 23456 the run id, called an (authentication) keyfile containing a 64-character password. This password is needed to decrypt the main dataset, and the keyfile itself is needed to use other utilities to download bulk and genetics data as well as returned datasets.

## 1.2 The formats of data available

The data available to download from UK Biobank comes in a variety of formats which need to be accessed in different ways:

- The main dataset – this is downloaded, decrypted and converted according to the instructions given in section 2.

- Bulk images/files (e.g. MRI Images, ECG data) – these are downloaded using the ukbfetch utility as explained in section 3.

- Genetics data – this is downloaded using either the ukbgene utility or ukbfetch depending on the type of data. Also, some genetics data can be downloaded from the European Genome-Phenome Archive (EGA), and some genetics fields are included as part of the main dataset rather than needing a separate download. See section 4 for further details.

- Record-level hospital and primary care (GP) data – this is accessed via the Data Portal in the Downloads page of Showcase. See section 5 for details.

- Returned datasets – these are datasets returned from researchers who have used UK Biobank data in their research, but which have not been incorporated directly into the main resource. See section 6 for details.

## 1.3 Multiple downloads and refreshes

The main dataset can be downloaded multiple times without limit, but will become inaccessible after a year. This is in order to prevent the data of participants who have subsequently withdrawn from the study being released again.

Periodically, the UK Biobank Showcase resource will be updated with new data. Currently, this typically happens 2-3 times a year. Researchers will be notified by email when a Showcase update has been made.

In order to gain access to updated data for fields in a previous data basket a researcher can request a "refresh" of that basket through AMS. A refresh of a dataset is a new extraction of the fields in the basket, and will include any additional data added to Showcase when it was updated. It will also remove the data for participants who have withdrawn since the basket was last released.

In order to request a refresh of a basket, a researcher will need to login to the Access Management System (AMS), navigate to their project (click Projects then View/Update), then click on the Data tab, and then on the "Go to Showcase to refresh or download data" button which will lead to the Downloads page. Next click 'Application' (at the top of the page) and then select the basket to be refreshed.

It will only be possible to refresh a basket that contains new data subsequent to a Showcase Update. If the selected basket can be refreshed a button 'Request Refresh' will be visible. Clicking on this button will then show the refresh requested as 'Queued'. A new notification email will be sent when the refreshed basket is available to download.

If a Showcase update includes new fields that were not previously included in a basket for the project, then a "Change Request" can be submitted for access to the new fields.

Periodically UK Biobank will send out an email to researchers containing a list of all participants who have withdrawn consent for their data to be used. These participants should be removed from any unpublished analyses.

## 1.4 Getting help

If you are having difficulties with any aspect of the data download process we have collected some previously encountered issues in the Appendix of this document.

If you are unable to find a solution then please contact the Access Management Team (AMT) at access@ukbiobank.ac.uk quoting your Application ID and the Run ID to which the problem relates.

It will help us to solve your issue more quickly if you provide screenshots of your problem, the steps you have followed up until the point the issue occurred, including any error messages received, as well as (where appropriate) listings of the contents of the folder you are working from.

If you find any errors in this document, or any parts that are unclear or incomplete, we would be grateful if you would pass them on to the AMT at the address above.

# 2 Downloading a main dataset

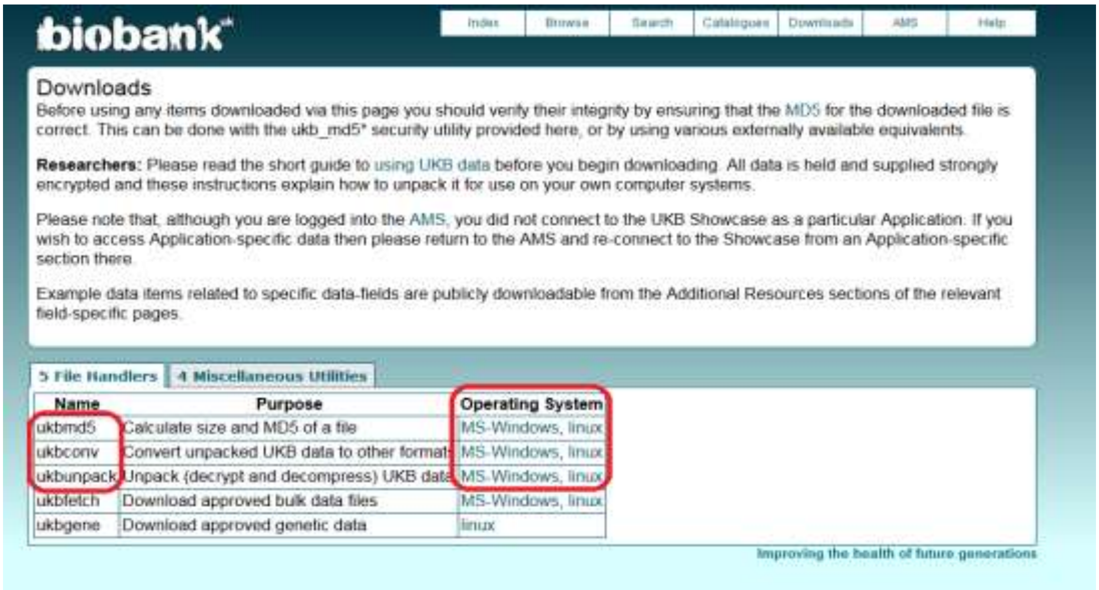
Downloading a main dataset requires several steps. The encrypted dataset must be downloaded through AMS. It must then be decrypted (“unpacked”), and then converted to a suitable format for use. A number of “helper programs” need to be downloaded to accomplish these steps.

## 2.1 Helper programs & encoding file

There are three helper programs required for decrypting and converting the main dataset:

- **ukbmd5** – for ensuring the encrypted main dataset has downloaded correctly;
- **ukbunpack** – for decrypting the downloaded main dataset;
- **ukbconv** – for converting the decrypted dataset into a suitable format.

These are provided in the File Handlers tab in the Download section of the Showcase website, as shown in figure 2.1.1.


biobank
Index
Browse
Search
Catalogues
Downloads
AMS
Help
5 File Handlers
4 Miscellaneous Utilities
Improving the health of future generations

Downloads

Before using any items downloaded via this page you should verify their integrity by ensuring that the MD5 for the downloaded file is correct. This can be done with the ukb_md5* security utility provided here, or by using various externally available equivalents.

**Researchers:** Please read the short guide to using UKB data before you begin downloading. All data is held and supplied strongly encrypted and these instructions explain how to unpack it for use on your own computer systems.

Please note that, although you are logged into the AMS, you did not connect to the UKB Showcase as a particular Application. If you wish to access Application-specific data then please return to the AMS and re-connect to the Showcase from an Application-specific section there.

Example data items related to specific data-fields are publicly downloadable from the Additional Resources sections of the relevant field-specific pages.

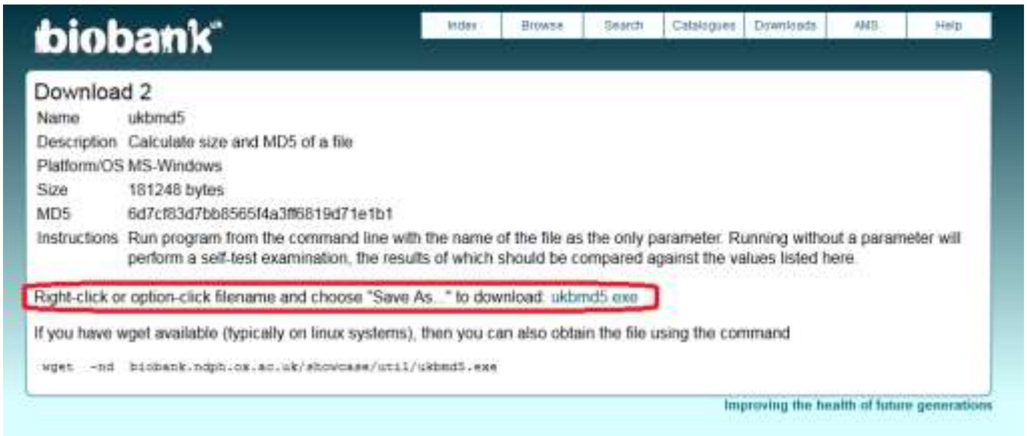
| Name | Purpose | Operating System |
|---|---|---|
| ukbmd5 | Calculate size and MD5 of a file | MS-Windows, linux |
| ukbconv | Convert unpacked UKB data to other format | MS-Windows, linux |
| ukbunpack | Unpack (decrypt and decompress) UKB data | MS-Windows, linux |
| ukbfetch | Download approved bulk data files | MS-Windows, linux |
| ukbgene | Download approved genetic data | linux |

**Figure 2.1.1: Helper programs**

The helper programs are supplied in two separate formats for compatibility with Windows or Linux operating systems. The Windows format is distinguished by the suffix ".exe".

The helper programs can be downloaded one at a time by selecting the required operating system version. This will open a new page, where the download can be found (figure 2.1.2).

Each program can be downloaded by clicking on it. We recommend that the helper programs are saved in a single file folder. A Linux command is also provided to perform the download.



Download 2

| | |
|---|---|
| Name | ukbmd5 |
| Description | Calculate size and MD5 of a file |
| Platform/OS | MS-Windows |
| Size | 181248 bytes |
| MD5 | 6d7cf83d7bb8565f4a3ff6819d71e1b1 |
| Instructions | Run program from the command line with the name of the file as the only parameter. Running without a parameter will perform a self-test examination, the results of which should be compared against the values listed here. |

Right-click or option-click filename and choose "Save As..." to download: ukbmd5.exe

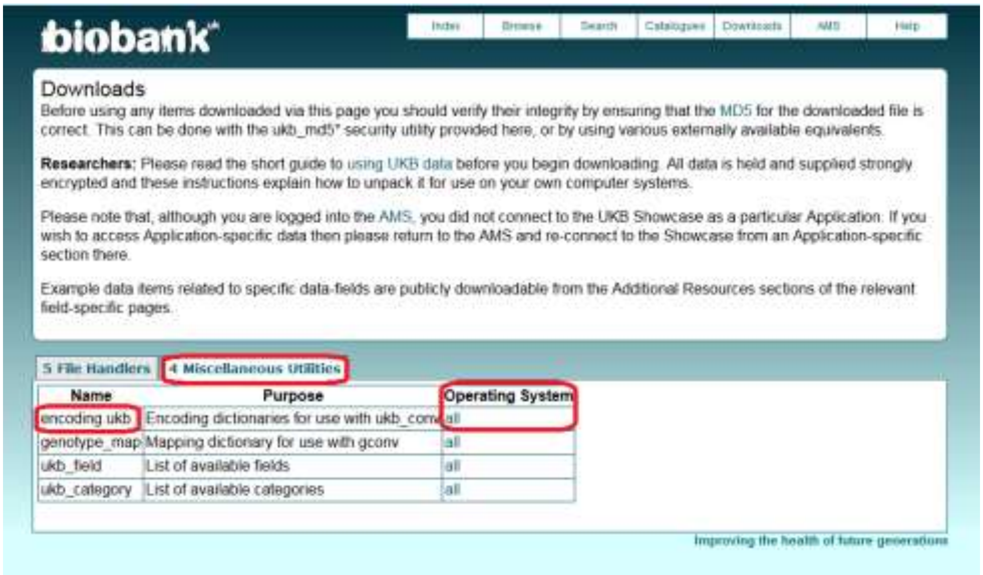
If you have wget available (typically on linux systems), then you can also obtain the file using the command

```
wget  -nd  biobank.ndph.ox.ac.uk/showcase/util/ukbmd5.exe
```

**Figure 2.1.2: Download page**

As part of the conversion process into certain formats (section 2.6), the converter program “ukbconv” will look for a file called “encoding.ukb”, which is used to assign coded definitions to variables in the dataset, and is compatible for use with both Windows and Linux systems.

The file encoding.ukb is provided in the Miscellaneous Utility tab in the Download section of the Showcase website, as shown in figure 2.1.3.  We recommend that you download “encoding.ukb” and save it along with the helper programs in the same folder.



Downloads

Before using any items downloaded via this page you should verify their integrity by ensuring that the MD5 for the downloaded file is correct. This can be done with the ukb_md5* security utility provided here, or by using various externally available equivalents.

**Researchers:** Please read the short guide to using UKB data before you begin downloading. All data is held and supplied strongly encrypted and these instructions explain how to unpack it for use on your own computer systems.

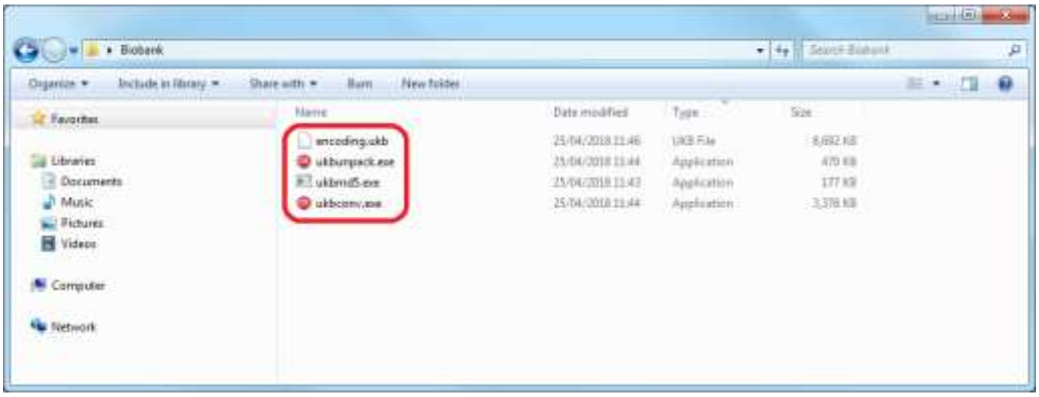
Please note that, although you are logged into the AMS, you did not connect to the UKB Showcase as a particular Application. If you wish to access Application-specific data then please return to the AMS and re-connect to the Showcase from an Application-specific section there.

Example data items related to specific data-fields are publicly downloadable from the Additional Resources sections of the relevant field-specific pages.

| Name | Purpose | Operating System |
|---|---|---|
| encoding.ukb | Encoding dictionaries for use with ukb_conv | all |
| genotype_map | Mapping dictionary for use with gconv | all |
| ukb_field | List of available fields | all |
| ukb_category | List of available categories | all |

**Figure 2.1.3: Encoding file**

At this point in the process you should have a folder containing four files similar to that shown in Figure 2.1.4.



| Name | Date modified | Type | Size |
|---|---|---|---|
| encoding.ukb | 25/04/2018 11:46 | UKB File | 8,692 KB |
| ukbunpack.exe | 25/04/2018 11:44 | Application | 470 KB |
| ukbmd5.exe | 25/04/2018 11:43 | Application | 177 KB |
| ukbconv.exe | 25/04/2018 11:44 | Application | 3,378 KB |

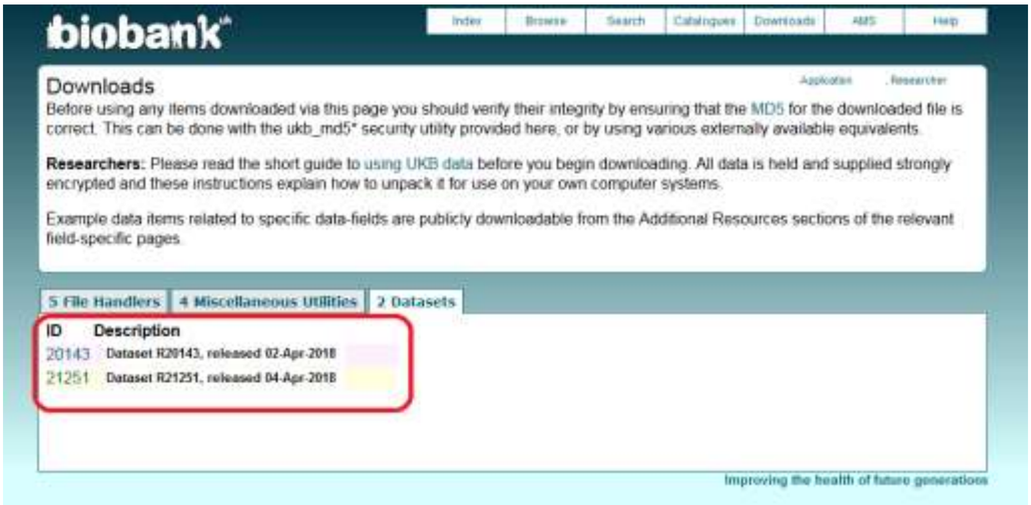
**Figure 2.1.4: Helper programs & encoding file**

## 2.2 Downloading a main dataset from Showcase

To download a dataset, you must first login to the Access Management System (AMS), navigate to the Projects tab and select the relevant application ID. Then click the blue button View/Update, then click on the Data tab at the top right, and then on the "Go to Showcase to refresh or download data" button which will lead to the Downloads page.

Only the project Principal Investigator (PI) and collaborators with delegate access are able to access the Data Download page. The project PI can assign delegate rights to other collaborators by using the Collaborators tab on the AMS.

Your dataset will be shown in the Dataset tab, as shown in figure 2.2.1:



Downloads

Before using any items downloaded via this page you should verify their integrity by ensuring that the MD5 for the downloaded file is correct. This can be done with the ukb_md5* security utility provided here, or by using various externally available equivalents.

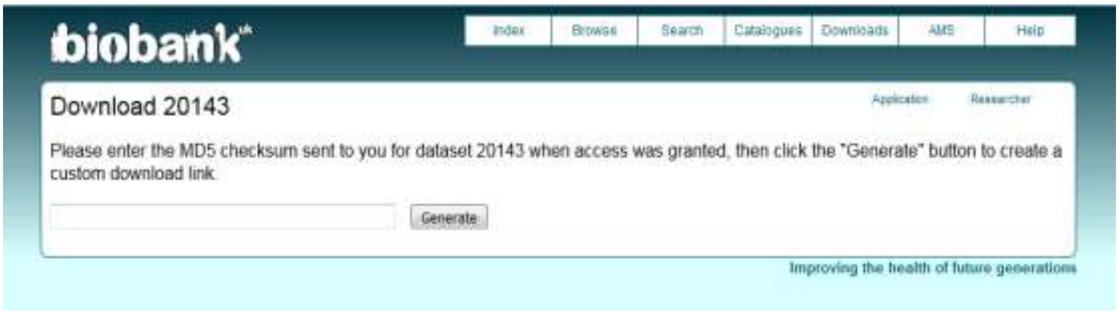
**Researchers**: Please read the short guide to using UKB data before you begin downloading. All data is held and supplied strongly encrypted and these instructions explain how to unpack it for use on your own computer systems.

Example data items related to specific data-fields are publicly downloadable from the Additional Resources sections of the relevant field-specific pages.

| ID | Description |
|---|---|
| 20143 | Dataset R20143, released 02-Apr-2018 |
| 21251 | Dataset R21251, released 04-Apr-2018 |

**Figure 2.2.1: Location of datasets**

Click on the ID for the dataset you wish to download, which will take you to the authentication screen:



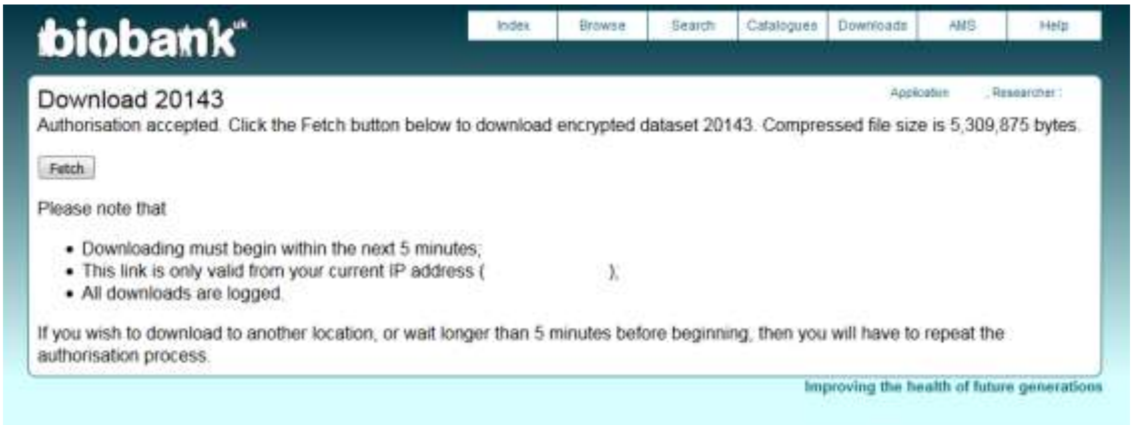
Download 20143

Please enter the MD5 checksum sent to you for dataset 20143 when access was granted, then click the "Generate" button to create a custom download link.

**Figure 2.2.2: Authentication screen**

Enter the 32-character MD5 checksum (included in the main body of the notification email for the dataset). Then click Generate.

This will open a new page with a link to your dataset as shown in Figure 2.2.3:



Download 20143

Authorisation accepted. Click the Fetch button below to download encrypted dataset 20143. Compressed file size is 5,309,875 bytes.

Please note that

- Downloading must begin within the next 5 minutes;
- This link is only valid from your current IP address ( );
- All downloads are logged.

If you wish to download to another location, or wait longer than 5 minutes before beginning, then you will have to repeat the authorisation process.

**Figure 2.2.3: Link to download dataset**

Click the Fetch button to download the encrypted dataset. Then save your dataset in the same file directory as the helper programs.

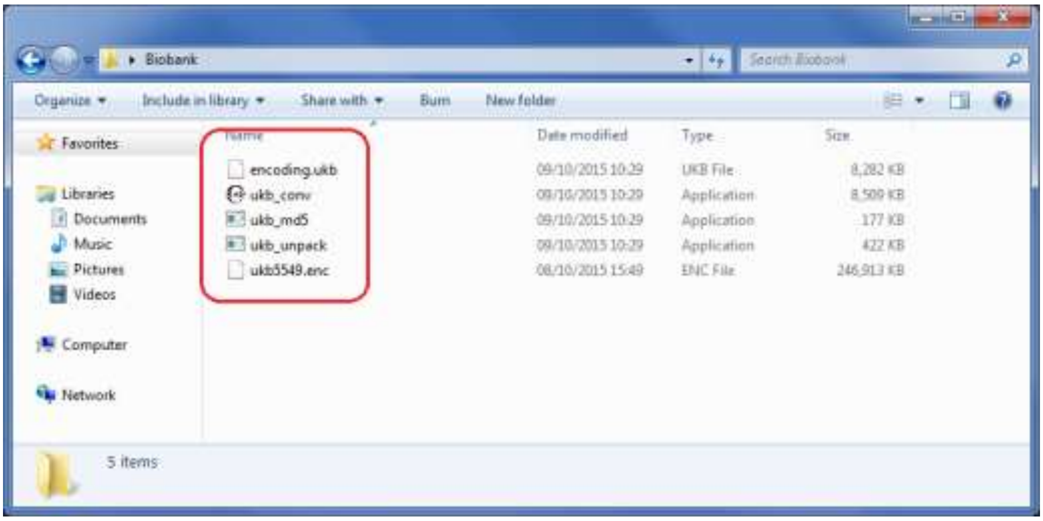
## 2.3 Open a command prompt / terminal

In order to proceed with the download process: validating, decrypting & converting the downloaded file, it is necessary to be able to run the helper programs (see section 2.1) using command line instructions from a command prompt in Windows or a terminal window in Linux.

For guidance on how to work with the command prompt in Window please see Section 8.1. The next few sections assume basic familiarity with command-line interfaces.

## 2.4 Validating the download (ukbmd5)

At this point you should now have five files in your folder, similar to as shown in figure 2.4.1. Note that the number used in the .enc file will be specific to your dataset; it should be the basket's run ID.



| Name | Date modified | Type | Size |
|---|---|---|---|
| encoding.ukb | 09/10/2015 10:29 | UKB File | 8,282 KB |
| ukb_conv | 09/10/2015 10:29 | Application | 8,509 KB |
| ukb_md5 | 09/10/2015 10:29 | Application | 177 KB |
| ukb_unpack | 09/10/2015 10:29 | Application | 422 KB |
| ukb5549.enc | 08/10/2015 15:49 | ENC File | 246,913 KB |

**Figure 2.4.1: The helper programs & dataset for run ID 5549**

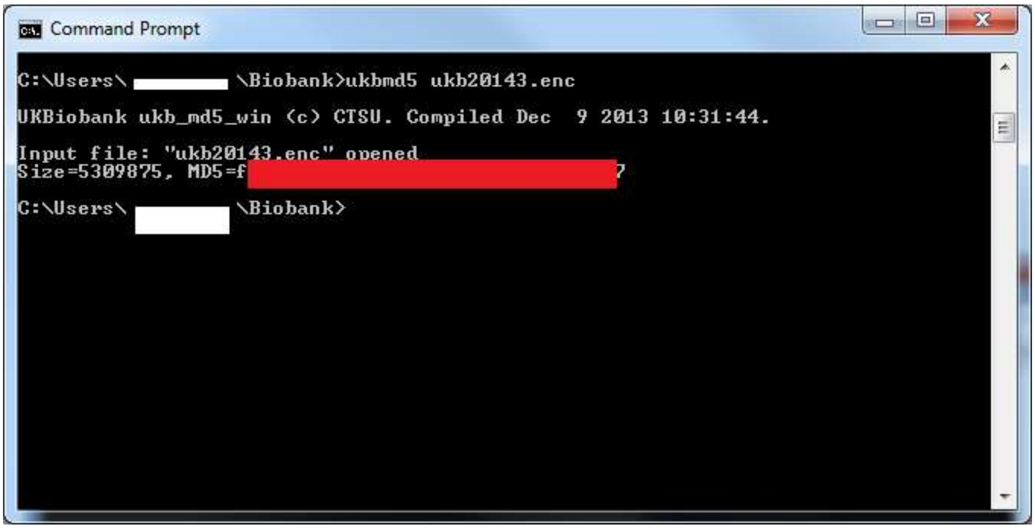
Open a command prompt / terminal and set the folder containing the above five files as your current working directory (by entering the command “cd” followed by the location; e.g. “cd username”).

You can verify the integrity of the files that you have downloaded by typing the command:

```
ukbmd5 ukb23456.enc
```

replacing **`ukb23456.enc`** with the name of your dataset file.

You should get output similar to Figure 2.4.2 below:



```
C:\Users\[redacted]\Biobank>ukbmd5 ukb20143.enc

UKBiobank ukb_md5_win (c) CTSU. Compiled Dec  9 2013 10:31:44.

Input file: "ukb20143.enc" opened
Size=5309875, MD5=f[redacted]

C:\Users\[redacted]\Biobank>
```

**Figure 2.4.2: Validation using ukbmd5**

where the red bar will be replaced by the 32-character MD5 checksum that you used to download the data. If the MD5 checksum is different to the one in your notification email it indicates that something has gone wrong in the download. In this case, you should delete the dataset and download it again.

## 2.5 Decrypting the dataset (ukbunpack)

Datasets are supplied in a compressed encrypted format. The ukbunpack program decrypts and uncompresses the downloaded file into a custom UK Biobank format.

To use the program, type the command:

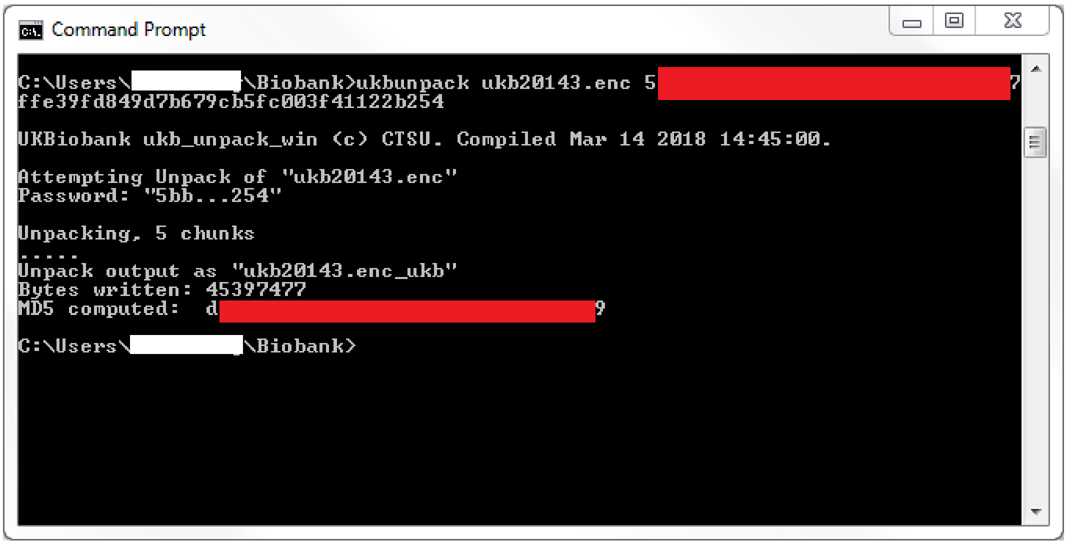
```
ukbunpack ukb23456.enc keyvalue
```

replacing:

- `ukb23456.enc` with the name of your dataset file;
- `keyvalue` with the 64-character password from the second line of the keyfile attachment in your notification email (this will have a name of the form **k56789r23456.key** where 56789 is replaced by your application ID and 23456 by the run ID of the dataset). Note that the keyvalue is not the same as the MD5 checksum. The keyfile may not open directly from an email server (due to the .key extension) but as it is simply a text file it can be downloaded and then opened in a text editor such as Notepad.

**Each .enc file has a different keyvalue** that can be found as an attachment to the notification email for the release of that particular dataset. Although each keyfile is given the same name, the keyfiles are not interchangeable, and the passwords of datasets released for the same project will each be different.

After the command above has been entered, the file will be decrypted ("unpacked") as shown in figure 2.5.1. This could take a few minutes.



```
C:\Users\[redacted]\Biobank>ukbunpack ukb20143.enc 5[redacted]7
ffe39fd849d7b679cb5fc003f41122b254

UKBiobank ukb_unpack_win (c) CTSU. Compiled Mar 14 2018 14:45:00.

Attempting Unpack of "ukb20143.enc"
Password: "5bb...254"

Unpacking, 5 chunks
.....
Unpack output as "ukb20143.enc_ukb"
Bytes written: 45397477
MD5 computed:  d[redacted]9

C:\Users\[redacted]\Biobank>
```

**Figure 2.5.1: Decrypting / unpacking a dataset**

This process will create a new file in your directory, named: **ukb23456.enc_ukb**, where 23456 is replaced by the run ID of your dataset.

Note that an alternative way of using ukbunpack is using the command:

```
ukbunpack ukb23456.enc k56789r23456.key
```

where **`ukb23456.enc`** is as before and **`k56789r23456.key`** is the keyfile from the notification email, which must have been placed in the same folder as ukbunpack and your .enc file.

## 2.6 Conversion of the dataset (ukbconv)

The result of the unpacking program is a dataset in a custom UK Biobank format (the .enc_ukb file shown above). The ukbconv program can be used to convert this into various other formats.

The ukbconv program is run via the command:

```
ukbconv ukb23456.enc_ukb option
```

where **ukb23456.enc_ukb** is the file generated from the previous unpacking step (with 23456 replaced by the run ID of your dataset) and **option** is replaced by one of: docs, csv, txt, r, sas, stata or bulk depending on the output desired.

The various options do the following:

- **docs**: generates a data dictionary for your dataset (see section 2.6.1);

- **csv, txt, r, sas** or **stata**: converts the dataset into a csv file, tab-separated txt file, or a file suitable for one of the statistics packages R, SAS and Stata (see section 2.6.2);

- **bulk**: creates a "bulk" file which is used in conjunction with ukbfetch to download bulk data (see section 3.2.2 for further details). This option is only relevant for the downloading of bulk data items such as MRI images etc.

In all cases the original .enc_ukb file remains intact so the converter may be used multiple times to generate different outputs.

### 2.6.1 Creating a data dictionary

The option 'docs' creates an HTML document that lists information about the structure of the dataset. The first nine rows of such a file are shown below for illustration:

| Column | UDI* | Count[†] | Type | Description |
|---:|---|---:|---|---|
| 0 | eid | 502619 | Sequence | Encoded anonymised participant ID |
| 1 | 31-0.0 | 502619 | Categorical (single) | Sex<br>Uses data-coding 9 |
| 2 | 34-0.0 | 502619 | Integer | Year of birth |
| 3 | 46-0.0 | 499206 | Integer | Hand grip strength (left) |
| 4 | 46-1.0 | 20202 | | |
| 5 | 46-2.0 | 23075 | | |
| 6 | 47-0.0 | 499273 | Integer | Hand grip strength (right) |
| 7 | 47-1.0 | 20217 | | |
| 8 | 47-2.0 | 23070 | | |

** UDI - the Unique Data Identifier for an item of data within the UK Biobank repository. The format for standard data fields is field_id-instance_index.array_index with genomic SNPs begin prefixed by "affy".*

† *Count - the number of non-empty rows present in this dataset.*

See section 2.8 on the structure of a main dataset for an explanation of what is meant by “instance index” and “array index” for a main dataset.

Running the ‘docs’ option also creates a text file called fields.ukb giving a list of all the fields contained in the dataset; this file can be useful when using some of the other options for ukbconv as detailed in the next section. A log file is also created to summarise the results of the conversion process.

### 2.6.2 Converting to a csv or statistics package format

Using the option: csv, txt, r, sas or stata with ukbconv transforms this dataset into various standard formats.

To convert the dataset into a standard format we use the command:

```
ukbconv ukb23456.enc_ukb option
```

where `ukb23456.enc_ukb` is the file generated from the previous unpacking step (with 23456 replaced by the run ID of your dataset), and `option` is replaced by one of: csv, txt, r, sas or stata depending on the output desired.

Assuming the file encodings.ukb (see section 2.1) is contained in the folder where the conversion is taking place the options r, sas and stata will not only convert the data, but replace all categorical Data-Codings with their meanings.

All four options generate the following two files:

- fields.ukb – a simple text file, giving a list of all the Showcase field numbers appearing in the dataset

- ukb23456.log – a log file used to summarise the result of the conversion process, giving the date & time, name of the output file, application identifier, basket identifier, number of variables, and the time required to convert.

In addition, depending on the conversion type, a number of other files will be generated as shown in the table below:

| Format | File generated | Description |
|---|---|---|
| **csv** | ukb23456.csv | Comma-separated file output with all fields double-quoted (to account for possible text fields containing commas).<br>The Data-Codings will be retained rather than replaced by their meanings. |
| **txt** | ukb23456.txt | A basic tab-separated text file output. As for csv above, the Data-Codings will be retained rather than replaced by their meanings. |
| **r** | ukb23456.tab | This is the actual file containing the data, in a tab-separated format. This file could potentially be imported directly into R, but none of the values will be coded. |
| | ukb23456.R | This file should be opened and executed in R (or any other R environment, such as RStudio). It contains a list of commands that will import the dataset (as a data.frame named bd) and recode all categorical variables. |
| **sas** | ukb23456.sd2 | This is the actual file containing the data as a SAS Data Set. This file could potentially be imported directly into SAS, but none of the values will be coded. |
| | ukb23456.sas | This file is the SAS program that should be opened and executed. It contains a list of commands that will import the dataset (as a dataset named WORK.LABELLED_LFVPWW) and recode all categorical variables. |
| **stata** | ukb23456.raw | This is the actual file containing the data. This file could potentially be imported directly into Stata, but none of the values will be coded. |
| | ukb23456.do | This file should be opened and executed in Stata. It contains a list of commands that will import the dataset and recode all categorical variables. |
| | ukb23456.dct | A dictionary of values used by ukb23456.do to format and label variables in the imported dataset. |

**Table 2.6.2: Conversion formats**

For example, with regards to the encodings.ukb file, if the csv option is used to convert the dataset then the field corresponding to “Sex” (field 31) will contain 0 and 1 (meaning Female and Male respectively). If the file is converted using the r option, then whilst the .tab file will still contain 0s and 1s in this column, if the .R file is opened and run, the dataset will be displayed with the column for field 31 showing “Female” and “Male”.

Note that if the file encodings.ukb is not present when the conversion into R, SAS or Stata format is run, then the conversion will still proceed but without the categorical variables being recoded.

Please note that large datasets may take a considerable amount of time (possibly hours) to convert, depending on the speed of the local system.

### 2.6.3 Optional parameters for ukbconv

Various optional parameters can be applied to the conversion, in particular to restrict which columns are included in the output. The full list is shown in the table below:

| Flag | Meaning |
|---|---|
| -s | Specify a single field (only) to include in the output |
| -i | Specify a subset of fields to include in the output |
| -x | Specify a subset of fields to exclude from the output |
| -o | Specify an alternative name for the output file |
| -e | Specify an alternative file from which to extract encoding information |

**Table 2.6.3: Optional parameters for ukbconv**

Options are included by adding them to the end of the ukbconv command. So for example the command:

```
ukbconv ukb23456.enc_ukb r -s20002
```

would convert the dataset into an R format, keeping only the eid column and all columns relating to field 20002. Note that since field 20002 (Non-cancer illness code, self-reported) has numerous different instance and array indices this will produce multiple additional columns (see section 2.8 for an explanation of instance and array indices).

When selecting subsets of fields using the options -i or -x, the file defining the parameters should be in text format (.txt), with one field-ID per row. To assist with preparing this file, the converter outputs the file named “field.ukb” each time, and this lists all the available fields associated with the dataset. This can be edited to identify the particular fields which are to be included in the subset.

Note that running the converter twice, using the same subset file but with -i and -x on alternate runs, will split the dataset into two complimentary parts.

By default ukbconv will look for the encoding file “encoding.ukb”, as described in the previous section, but by using the -e option a different filename can be used as the source for the Data-Coding definitions.

## 2.7 Decryption and conversion example

A researcher has been notified by email that data for their application 56789 is available for download. The email provides the run ID 23456 for the dataset. The 32-character MD5 Checksum is:

“abcdef0123456789abcdef0123456789”

and the 64-character Password (contained in the second line of the attached text file k56789r23456.key) is:

“a1b2c3d4a1b2c3d4a1b2c3d4a1b2c3d4a1b2c3d4a1b2c3d4a1b2c3d4a1b2c3d4”.

We assume that the three helper programs and encoding file have already been downloaded in accordance with section 2.1 and that the dataset is being downloaded into the same folder.

1. The researcher (either the PI or a collaborator with delegate access) logs on to the Application Management System (AMS), clicks Projects and then clicks on the blue button View/Update for project 56789. They select the Data tab at the top right of the page, and select the option to go to the Showcase download page. From this page they select the Dataset tab. An entry with run ID 23456 should be listed.

2. They click on the (run) ID 23456 for the entry and on the following screen enter the MD5 checksum given above (from the main body of the notification email):

   abcdef0123456789abcdef0123456789

   into the box and click Generate. This will open the download page; they click Fetch to initiate the download of file ukb23456.enc and save it in the same folder as the helper programs and encoding file.

3. To verify that the file has arrived intact they open a command prompt, navigate to the appropriate folder and enter:

   ```
   ukbmd5 ukb23456.enc
   ```

   This displays an MD5 value which matches the MD5 Checksum from the notification email (the one used to download the dataset). If the MD5 checksum had not matched, the researcher would need to repeat the download operation. If there was still no match they would need to contact the Access Management Team (AMT) for further assistance.

4. They next unpack (decrypt) the data by entering into the command prompt:

   ```
   ukbunpack ukb23456.enc a1b2c3d4a1b2c3d4...a1b2c3d4
   ```

   where we have truncated the 64-character keyfile so it fits onto the line above, but the researcher would have needed to include all 64 of the characters.

   This will produce a file ukb23456.enc_ukb.

5. To create a comma separated variable (csv) version of the data, they enter into the command prompt:

   ```
   ukbconv ukb23456.enc_ukb csv
   ```

   This will produce a file `ukb23456.csv` which can be processed by standard programs.

## 2.8 The structure of a main dataset

Having followed the above steps a researcher will now have a main UK Biobank dataset. We here give some indication of what this would look like, focusing in particular on the meanings of the column headers.

A main dataset will be rectangular with one participant per row, and columns headers giving the Showcase field number that the data in the that column relates to together with the “instance index” and “array index” of that item. Broadly speaking, the instance index is used to distinguish data for a field which was gathered at different times, and the array index is used to distinguish multiple pieces of data for that field which were gathered at the same time.

These will display differently depending on the format that the dataset has been converted to (see table 2.8.2 at the end of this section). The example we give in table 2.8.1 below shows a small portion of a sample dataset as it would appear in .csv format opened in Excel:

| **eid** | **53-0.0** | **53-1.0** | **53-2.0** | **20002-0.0** | **20002-0.1** | **20002-1.0** | **20002-1.1** | **20002-2.0** | **20002-2.1** | **…** |
|---|---|---|---|---|---|---|---|---|---|---|
| 1256847 | 11/04/2007 | | 03/01/2017 | 1077 | | | | 1077 | 1075 | |
| 8645816 | 29/10/2009 | | | | | | | | | |
| 4652658 | 15/08/2009 | | | | | | | | | |
| 2328974 | 12/07/2008 | 09/03/2013 | | | | 1002 | | | | |
| 3315794 | 22/02/2010 | 01/12/2012 | 19/11/2018 | 1111 | | 1111 | | 1111 | 1065 | |
| 9497726 | 25/02/2006 | | | | | | | | | |
| 4582852 | 06/06/2008 | | | 1222 | 1265 | | | | | |
| … | | | | | | | | | | |

**Table 2.8.1: A portion of a sample main dataset**

The eid is the encoded participant identifier for the project in question. The remaining column headers are in the format `F-I.A` where `F` is the field number, `I` is the instance index and `A` is the array index.

Two fields are shown in the sample dataset: Field 53 (Date of attending assessment centre) and Field 20002 (Non-cancer illness code, self-reported). In each case there are three “instances” of the variable (the first number after the -). Using the “Instance” tab on the fields pages on Showcase we see that these correspond the visit type: 0 for the initial (baseline) visit, 1 for the repeat assessment and 2 for the first imaging assessment.

The columns 53-0.0, 53-1.0 and 53-2.0 then hold the dates each participant attended that particular type of assessment centre. In the above, all participants attended a baseline assessment centre (this would always be the case), but only two attended the repeat assessment, one of whom also attended an imaging centre. The first participant attended an imaging centre, but did not attended the repeat assessment.

At each assessment centre visit a participant can self-report illnesses, which are coded in Field 20002. The illnesses are coded using Coding 6, as indicated on the Field 20002 page on Showcase. Clicking on the “6” of “Coding 6” on that page allows us to see a list giving the meanings of the codes given above.

For example: looking at the participant with eid 3315794 we see that at each of their three assessment centre visits they self-reported having asthma (code 1111). As the “first” condition reported this is assigned to have array index 0 (the final number in 20002-0.0 etc). At their imaging assessment visit (instance 2) they also report hypertension (code 1065), and this being the second reported condition at that visit it is assigned to array index 1, i.e. in field 20002-2.1.

Note that in reality field 20002 has array indices running from 0 to 33 (indicating at least one participant self-reported 34 illness codes), and so the real dataset would be considerably wider than that shown above even with only these two fields in it.

Note also that due to the nature of field 20002 being a self-report field (i.e. reported at an assessment centre), we can only have data for a particular instance index for field 20002 if that same instance index in field 53 has a value. For example, since the participant with eid 4582852 only attended baseline assessment they can only have values for field 20002 with instance index equal to 0.

The instance index is not exclusively used to refer to the assessment centre visit. For example, the “Diet by 24-hour recall” fields (see Category 100090) use instance 0 to refer to the baseline assessment centre (as above), but then instances 1 to 4 refer to the four on-line cycles of this questionnaire. As another example, reports from the cancer register (see Category 100092) are given a new instance index for each additional type of cancer reported.

As indicated above, the column headers appear slightly differently depending on which package you are using. The various output formats display the headers as follows:

| File type | Column header | Notes | Example |
|---|---|---|---|
| csv | `F-I.A` | | 31-0.0 |
| R | `f.F.I.A` | with f. preceding all fields | f.31.0.0 |
| SAS | `a_F_I_A` | a indicates the type of variable, e.g. a will be n for numerical fields and s for string fields. | n_31_0_0 |
| Stata | `a_F_I_A` | As for SAS. | n_31_0_0 |

**Table 2.8.2: Column header formats**

where, as previously, `F` represents the field number, `I` the instance index and `A` the array index

# 3 Bulk data

This section deals with accessing bulk data, such as imaging data (e.g. brain MRIs), accelerometer and ECG data, i.e. fields for which each item is a complex/compound dataset in itself.

These are accessed using a command line utility ukbfetch. This can be downloaded from the File Handlers tab on the Download section of Showcase. Both a Windows and Linux version of ukbfetch are available.

The ukbconv program will also usually be needed to generate a “bulk file” allowing the download of multiple bulk items at once (see section 3.2.2).

If you have a bulk data field in your project basket, there will be a column for it in your main dataset, however only the field ID will be present rather than the actual contents of the bulk data. The purpose is to indicate which participants have that bulk field available.

Note that ukbfetch creates a temporary file during the download, and then checks the MD5 checksum of the resulting file against its expected value. If the checksums do not agree then the download will fail. There is hence no separate validation step needed.

The sizes of some of the bulk fields are given in Section 8.4 in the Appendix for reference.

## 3.1 Connectivity and authentication

The bulk repository consists of a pair of mirrored systems each connected to the UK JANET network by independent links. The system names are:

- biota.ndph.ox.ac.uk
- chest.ndph.ox.ac.uk

To access bulk data your computer must be able to make http (Port 80) connections to at least one, and preferably both, of these systems. Please note that navigating to the above websites is not part of the download process; you simply need to ensure that your computer is able to connect to them. For most researchers this will not be a problem; however, please see section 8.2.6 for a way of checking this if you believe this may be an issue on your system. It is not possible to use a proxy server when using the ukbfetch utility.

In order to use ukbfetch it is necessary for you to authenticate yourself to the system. To do this you will need the authentication “keyfile” containing the 64-character password which was attached to the email notifying you that your data was ready to download (called k56789r23456.key where 56789 is replaced by your application ID and 23456 the run ID of the data extract). This is a simple text file containing your Application ID on the first line and the 64-character decryption password for that dataset on the second line.

The authentication keyfile should be saved in the folder where you will be running ukbfetch. The utility expects by default that the authentication keyfile has been renamed as ".ukbkey" (i.e. this is its full name with no other file extension). However, it is still possible to run the utility with the keyfile named differently by using the -a option (see section 3.2 for further details).

## 3.2 Using ukbfetch

The following two sections give general instructions for accessing Bulk data using the ukbfetch utility. Further details are given in UKB Resource 644.

### 3.2.1 Downloading a single bulk item

We assume for illustration that a participant with eid 2143432 has data for the bulk Field 20252 (T1 structural brain images - NIFTI). In a main dataset this will be indicated by the cell corresponding to the row with eid 2143432 and the column 20252-2.0 having the value 20252 in it (we are assuming the particular Field-Instance-Array format for a .csv file here; see section 2.8 for more information about this).

Note here that the instance index is 2 because the field was collected at a first imaging centre (instance 2) and the array index is 0 because only a single item of data was collected for this field at that centre.

To download the brain image for this participant we would use the command:

```
ukbfetch -e2143432 -d20252_2_0
```

assuming that the authentication keyfile has been renamed as .ukbkey (and placed in the same folder as ukbfetch). If the keyfile is instead called k56789r23456.key (for example) then the command would be:

```
ukbfetch -e2143432 -d20252_2_0 -ak56789r23456.key
```

Note that there must be no spaces between the flags (-e, -a etc) and the following arguments.

### 3.2.2 Creating and using a bulk file

To download many bulk fields at once, ukbconv can be used to generate a “bulk file” which lists participant eids and field numbers (including instance & array indices) for which that bulk field exists.

For example, let us assume we want to download all the T1 structural brain images for all participants at once.

Firstly, to generate the bulk file we run the command:

```
ukbconv ukb23456.enc_ukb bulk -s20252
```

where ukb23456.enc_ukb is our unpacked (but not converted) main dataset (see sections 2.5 & 2.6), and 23456 would be replaced by the run ID corresponding to your dataset.

The above command would output a file called **`ukb23456.bulk`** the first few lines of which would look something like:

```
3422567 20252_2_0
5321753 20252_2_0
2457842 20252_2_0
```

i.e. a simple list with each row the eid of a participant and the Field_Instance_Array of the relevant data.

Note that we cannot specify particular instance and array indices in the ukbconv call as the -s flag does not have this functionality. If this were a problem the bulk file could be edited using an appropriate software package to keep only the particular instances/arrays required.

Note that the -i flag for ukbconv can replace the -s flag to select a group of fields rather that a single one as in the example (see section 2.6.3).

Next, using our bulk file we can now use the command:

```
ukbfetch -bukb23456.bulk
```

to download every file for Field 20252. Once again there should be no space between the -b flag and the filename.

We can limit the number of files we download at once using ukbfetch by using the -s and -m flags. There is a limit of 50,000 files per ukbfetch call and so this will sometimes be an essential element of the process.

The flag -s gives the starting row of our bulk file to work from, and the -m flag sets how many rows from the bulk file we process.

For example, we could download 5000 files at a time for the above field by running the following commands one by one:

```
ukbfetch -bukb23456.bulk -s1 -m5000
ukbfetch -bukb23456.bulk -s5001 -m5000
ukbfetch -bukb23456.bulk -s10001 -m5000
ukbfetch -bukb23456.bulk -s15001 -m5000
ukbfetch -bukb23456.bulk -s20001 -m5000
```

Assuming that there are less than 25000 participants with this field, which is true at the time of writing, this would download all files for field 20252.

These commands could also be added to a batch file / shell script and run in one go. In this case there is -o flag which can be used to specify a different name for the logfile for each call of the ukbfetch utility.

Further details for using ukbfetch are given in UKB Resource 644.

# 4 Genetics data

This section deals with accessing the genetics data. There are a variety of different types of genetics data available through UK Biobank, and different methods are used for downloading the different types.

Note that genetics fields will have a corresponding column in your main dataset, but in the same way as for bulk fields only the field ID will be present rather than the actual contents of the bulk data. The purpose is to indicate which participants have that genetics field available.

## 4.1 Genotype fields

Some Genotype data fields appear in the main dataset, some can be downloaded using the ukbgene utility, some need to be downloaded using the ukbfetch utility, and some can be downloaded from the European Genome-phenome Archive (EGA). Some types of genetics data can be downloaded using more than one of these methods.

Most of the relevant information is shown on the page Category 263 (Genotypes) and in UKB Resource 664. Access via the EGA requires an account to be set up through the Access Management Team (AMT)

Some information about the method of download is also given on the Notes tab of the relevant field. For example, the CEL files, Field 22002, need to be downloaded using ukbfetch using the same methods as described in section 3.

Further information about the Genotyping data, including the size of the files, and about using the EGA is available in UKB Genotyping and Imputation Data Release March 2018 – FAQ which can be found in the “Useful resources” section (towards the bottom) of the page: UKB Genetic data.

The data in the Genotype BED and BGEN files appear in a common order for all researchers. In order to match your participant eids to the data (which is done by position) it is necessary to use ukbgene to download appropriate FAM and sample files.

## 4.2 Using ukbgene

The ukbgene utility can be used to download some parts of the Genotype data, as described in Category 263 (Genotypes) and in UKB Resource 664, in particular it is used to create the FAM and Sample files for a project to match the project eids, by position, to the BED and BGEN files.

Only a Linux version of ukbgene is available (see section 8.2.5 for a way to proceed if you do not have Linux available).

Note that ukbgene creates a temporary file during the download, and then checks the MD5 checksum of the resulting file against its expected value. If the checksums do not agree then the download will fail. There is hence no separate validation step needed.

MD5 checksums for Genotyped files are also available at Resource 998 and Resource 997.

### 4.2.1 Connectivity & authentication

The bulk repository consists of a pair of mirrored systems each connected to the UK JANET network by independent links. The system names are:

- biota.ndph.ox.ac.uk
- chest.ndph.ox.ac.uk

To access bulk data your computer must be able to make http (Port 80) connections to at least one, and preferably both, of these systems. Please note that navigating to the above websites is not part of the download process; you simply need to ensure that your computer is able to connect to them. For most researchers this will not be a problem; however, please see section 8.2.6 for a way of checking this if you believe this may be an issue on your system. It is not possible to use a proxy server when using the ukbgene utility.

In order to use ukbgene it is necessary for you to authenticate yourself to the system. To do this you will need the “keyfile” containing the 64-character password which was attached to the email notifying you that your data was ready to download (called k56789r23456.key where 56789 is replaced by your application ID and 23456 by the run ID of the data extract). This is a simple text file containing your Application ID on the first line and the 64-character decryption password for that dataset as the second line.

The authentication keyfile should be saved in the folder where you will be running ukbgene. The utility expects the authentication keyfile to be renamed as ".ukbkey" (i.e. this is its full name with no other file extension). However, it is still possible to run the utility with the keyfile named differently by using the -a option (see section 4.1.3).

### 4.2.2 A ukbgene example

A researcher has gained access to the Genotype calls for chromosome 5 by including Field 22105 (Chromosome 5 genotype results) in their project basket, which has subsequently been approved.

They have downloaded ukbgene from Download 665 by running the wget command given on a Linux terminal. To make ukbgene an executable file they have then run:

```
chmod 755 ukbgene
```

They have also saved their authentication keyfile k56789r23456.key from their notification email (where 56789 is their application ID) into the same folder as ukbgene, and renamed it as .ukbkey (this being the full filename).

To download the Genotype call .bed file for Chromosome 5, they enter the command:

```
ukbgene cal -c5
```

To download the associated FAM file (i.e. the link file giving the order that their project eids appear in the .bed file) they use the command:

```
ukbgene cal -c5 -m
```

Note that sometimes ./ukbgene needs to be used in place of ukbgene because of the way a Linux system is set up (see section 8.2.1). If the researcher had not renamed their keyfile, and left it with the filename k56789r23456.key, then they would have had to replace the above commands with:

```
ukbgene cal -c5 -ak56789r23456.key
```

and

```
ukbgene cal -c5 -m -ak56789r23456.key
```

Further information about the various options available with ukbgene are given in UKB Resource 664.

## 4.3 Exome sequences

A description of the Exome sequence fields are contained in Category 170 (Exome sequences) on Showcase.

The PLINK format Exome fields (23170 and 23160) are downloaded by using ukbgene as described in the Notes tabs for those fields. Note that in each case the data for all chromosomes is contained within a single file (the -c1 flag used in the download does not indicate that only chromosome 1 is included).

The remaining Exome fields are downloaded using ukbfetch as described in section 3.2.

Further information about the Exome sequence data is contained in UKB 50k Exome Sequencing Data Release July 2019 – FAQs which can be found in the “Useful resources” section (towards the bottom) of the following page: UKB Genetic data.

# 5 Record-level Hospital inpatient & GP data

The record-level hospital inpatient data and primary care (GP) data is available from the record repository accessed via the Data Portal on Showcase.

The record repository is divided into a number of database tables and access to each table is granted to a research project on a table-by-table basis by including a specific data-field in a project basket. For example, including Field 41259 in a basket will give access to the main HESIN table.

The main dataset will include a column for each such field but the values shown in that column will be a count of the number of rows that each participant has in the corresponding table.

To access the Data Portal, a researcher will need to login to the Access Management System, navigate to their project (click Projects then View/Update), then click on the Data tab, and then on the Data Download option at the bottom of the page.

This will lead to the Downloads page where, if approved for record data, there will be a Data Portal tab. Clicking on the Connect button will open up the portal into which SQL statements may be entered and the results viewed and downloaded.

Only the project Principal Investigator (PI) and collaborators with delegate access are able to access the Data Portal. The project PI can assign delegate rights to other collaborators by using the Collaborators tab on the AMS.

## 5.1 Record-level Hospital inpatient data

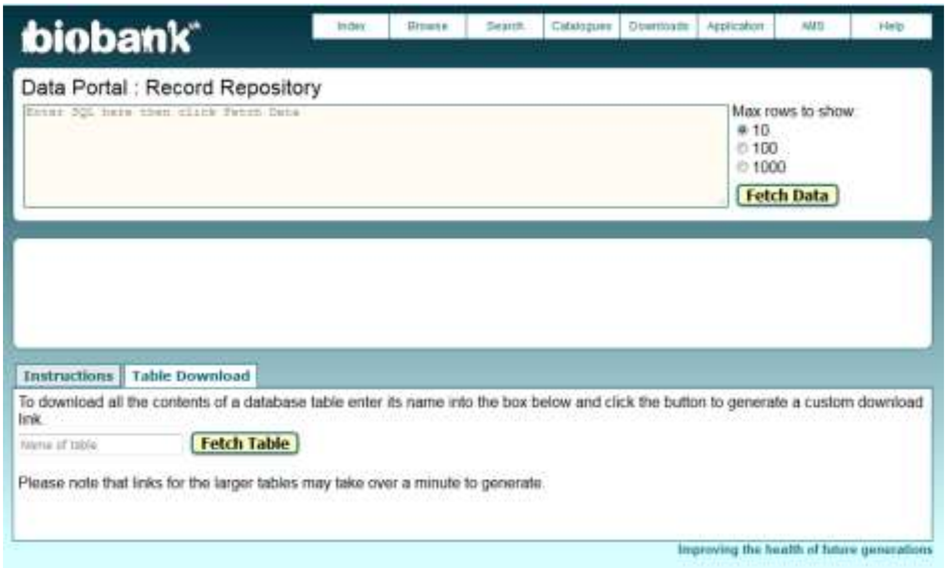
### 5.1.1 General structure of the hospital inpatient data

The Hospital inpatient data has been divided into six interrelated tables as described in the Hospital inpatient overview document and the Inpatient data dictionary. These are available on the Resources tab of Category 2000 (Hospital inpatient) on Showcase.

### 5.1.2 Downloading tables from the data portal

Once a researcher has accessed the data portal they can download each complete table as shown below, or query the data prior to downloading it (see section 5.1.3).

To download a complete table click on the ‘Table Download’ tab in the bottom panel, enter the name of the table you wish to download (e.g. hesin_diag) and click on the ‘Fetch Table’ button as shown in Figure 9.3.1.



Data Portal : Record Repository

Enter SQL here then click Fetch Data

Max rows to show:
● 10
○ 100
○ 1000

To download all the contents of a database table enter its name into the box below and click the button to generate a custom download link.

Name of table

Please note that links for the larger tables may take over a minute to generate.

**Figure 9.3.1: Table download tab**

This will generate a custom download link that you can paste into a web browser and a wget command for those using a linux system. The resulting dataset will be provided as a tab separated text file (.txt). Please note it can take some time to download the complete tables.

### 5.1.3 Using SQL to query the tables

An alternative to downloading whole tables is to use SQL statements to select data of interest prior to download.

SQL (Structured Query Language) is the control language used to manage and manipulate information within most modern relational databases. If you do not know SQL already then there are a number of free tutorials available on the web.

Each major database uses a slightly different dialect to that of other vendors, however most common commands are identical across them. The UK Biobank system uses the Ingres platform to host its relational databases. A reference manual is available online and can be located by an internet search for “Ingres 10.2 SQL Reference Guide”.

Some examples of SQL statements that can be used to do simple explorations of the inpatient data without downloading it are given in section 8.3.

## 5.2 Record-level GP data

The record-level primary care (GP) data is divided into 3 tables as described in the resources in Category 3000 (Primary care) on Showcase.

The tables are accessed via the same Data Portal as the Hospital inpatient data as described above in section 5.1, and SQL statements can be used to manipulate the data without downloading it, if desired.

# 6 Returned datasets

“Returns” are datasets returned by researchers who have used UKB data in their research. Some returned datasets are incorporated into the main resource, but those that have not been need to be downloaded using the ukblink utility.

The ukblink utility can be downloaded from the File Handlers tab on the Download section of Showcase. Both a Windows and Linux version of ukblink are available.

## 6.1 Authentication

In order to provide authentication for the download you will need to have your authentication keyfile in the same folder as ukblink (this is the attachment to your notification email with a name like k56789r23456.key). This is the same requirement as for ukbfetch and ukbgene (see, for example, section 3.1 for more details).

## 6.2 Using ukblink

We use Return 1362 as an example. We assume that we have been granted access to this dataset, that we have downloaded the ukblink utility (and if using Linux, made it executable; see section 8.2.1) and moved our keyfile into the same folder.

To download it we use the command:

```
ukblink -r1362
```

assuming our keyfile has been renamed as .ukbkey. Otherwise we use:

```
ukblink -r1362 -ak56789r23456.key
```

assuming the keyfile still has its original filename. (In Linux we may need to replace ukblink by ./ukblink, see section 8.2.1.)

Note that files will download as generic .dat files. More recent Returns are in fact all .zip files and renaming them as such should allow standard unzipping programs to be run. Older files may either be .zip or .7z files. A list of which of the older Returns has which type of zipped file format is included in the Appendix (section 8.5).

Some returned datasets provide participant-level data, and for these the ukblink utilitiy also allows the creation of a bridge to connect your project eids with those used in the Return.

Return 1362 is an example of a Return that includes participant-level data (this can be seen from its Showcase page in the “Personal” row). In order to download the bridge we need to know the Application that this Return was generated as part of. This can be determined from the first line of its Showcase page where we can see that it was part of Application 2964.

Hence, to generate the appropriate bridge file we use the command:

```
ukblink -b2964
```

(adding -`ak56789r23456.key` if appropriate).

Further details for accessing Returns using the ukblink utility are given in UKB Resource 655.

# 7 Bridges

## 7.1 Linking to Genetic data

Given the size of the genetics data, some projects will be given approval to link to a institution-held copy of the data rather than each project being required to have a separate copy. Any project accessing genetics data, even through a dataset downloaded by a different project, must have the relevant genetic fields included in an approved basket for their own project.

The genetics data appears in a common order for different projects, and the appropriate link file (FAM or sample file) for a project then provides the order in which the participants appear in the data.

All that is necessary for a new project to link to a genetics dataset downloaded by another project is for them to generate the appropriate link file using ukbgene, so as to determine the order that their eids appear in the genetics data.

Note that if the ‘owner’ of a genetics dataset, i.e. the project who originally downloaded it, is approached to share a genetics dataset they should confirm with UK Biobank (at access@ukbiobank.ac.uk) that the appropriate approvals are in place before allowing access to the data. They should also ensure that they have seen the fully executed MTA for the other project, with genetic data included.

Note that approval to reuse a genetics dataset in this way does not permit projects to share processed data with each other directly, or to construct a bridge to share other elements of UK Biobank data. This would constitute a breach of the project’s Material Transfer Agreement (MTA).

## 7.2 Bridge files for bulk fields

In some instances UK Biobank will release bridging files to link two separate UK Biobank applications together, in order for bulk images and other bulk fields to be shared between projects.

UK Biobank is currently reviewing its procedures with regards to bridging files and will be providing updated information in due course.

# 8 Appendix

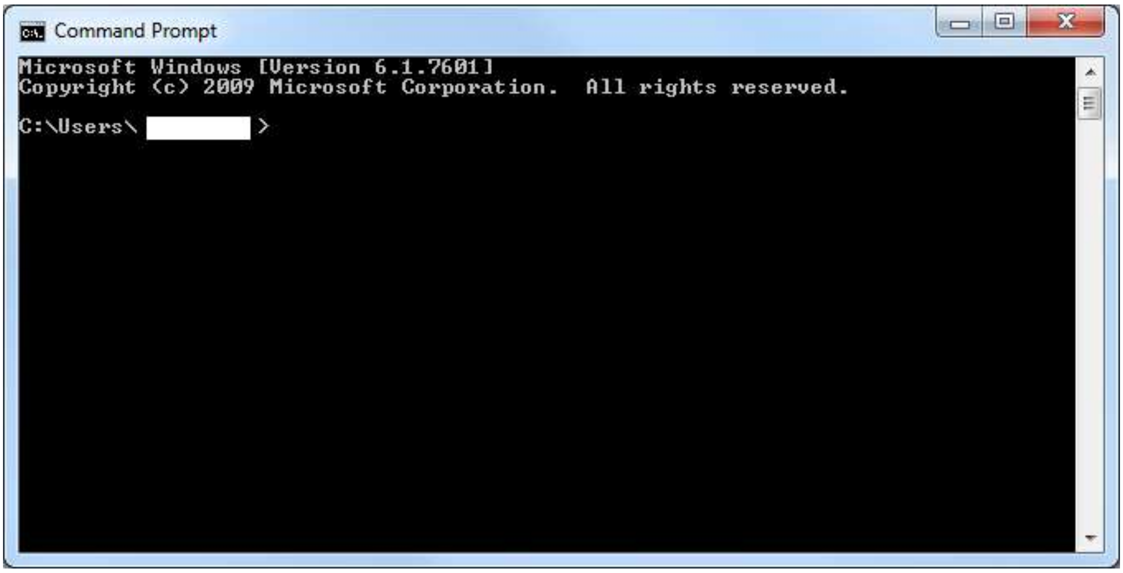
## 8.1 Using a command prompt in Windows

If you are using Windows:

- **Windows XP:** go to Start > All Programs > Accessories > Command Prompt
- **Windows Vista:** go to Start > type cmd in the Search bar, and click on Command Prompt once it has appeared
- **Windows 7:** go to Start > All Programs > Accessories > Command Prompt
- **Windows 8:** go to Start > type cmd in the Search bar, and click on Command Prompt once it has appeared
- **Windows 10:** go to the Search icon > type cmd in the Search bar, and click on Command Prompt once it has appeared

For any version of Windows, if the Command Prompt does not appear by following the steps above, please press the following combination of keys: Windows+R. (The Windows key is located between the Ctrl and the Alt keys on your keyboard). This will open a small window named "Run". Type cmd in the "Open:" space, then click OK. This will open a Command Prompt window.

Once it is opened, the Command Prompt window should display only a bit of text at the top, and then a blinking cursor preceded by a directory address on your computer (by default, this should be **C:\Users\YourUserName**), as shown in Figure 8.1.1.

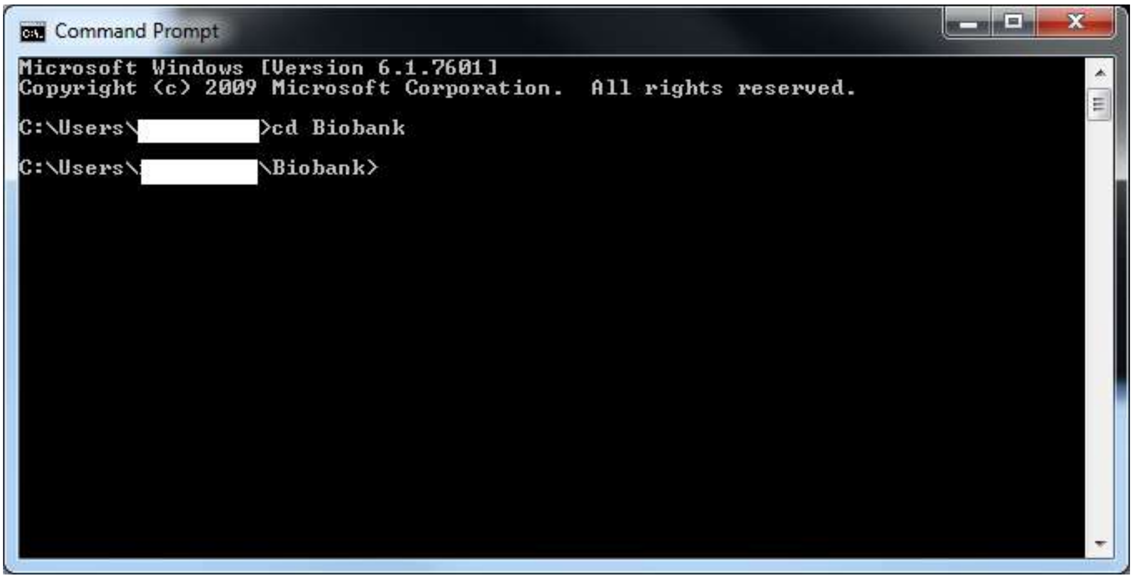


```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\[redacted]>
```

**Figure 8.1.1: The command prompt window**

The next step is to navigate to the directory in which you previously downloaded all the helper files, the encoding file and your dataset. To do this, type **cd** followed by the path that you wish to navigate to, from the current folder.

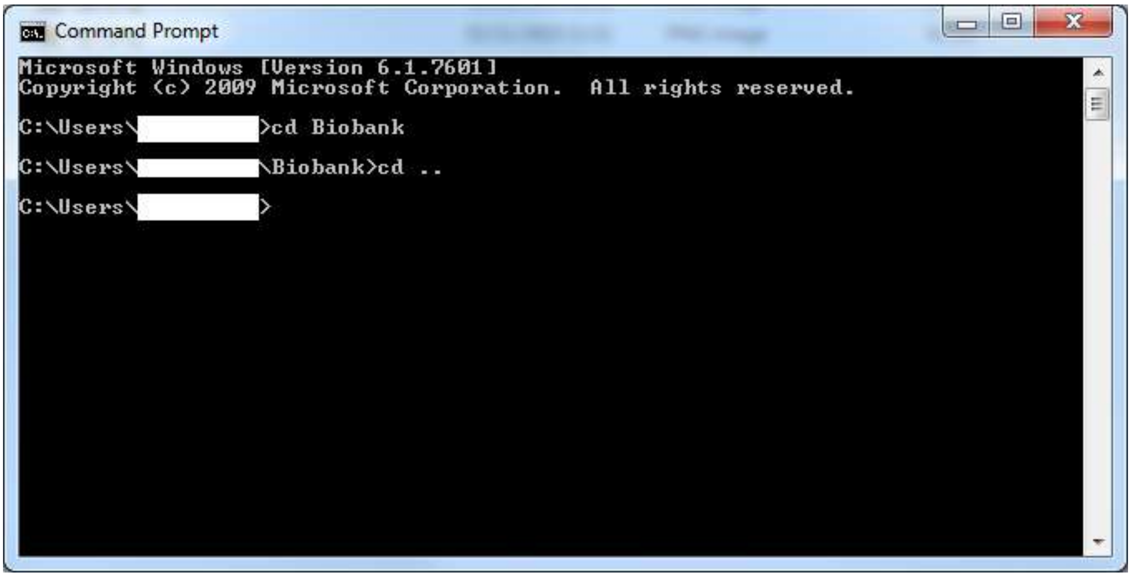
In our example, we downloaded the files in a directory named Biobank, which is located in the home directory for the user. All we need to do is type **cd Biobank** and press Enter to navigate to the Biobank directory, as shown in Figure 8.1.2.



```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\[redacted]>cd Biobank

C:\Users\[redacted]\Biobank>
```

**Figure 8.1.2: Changing the directory**

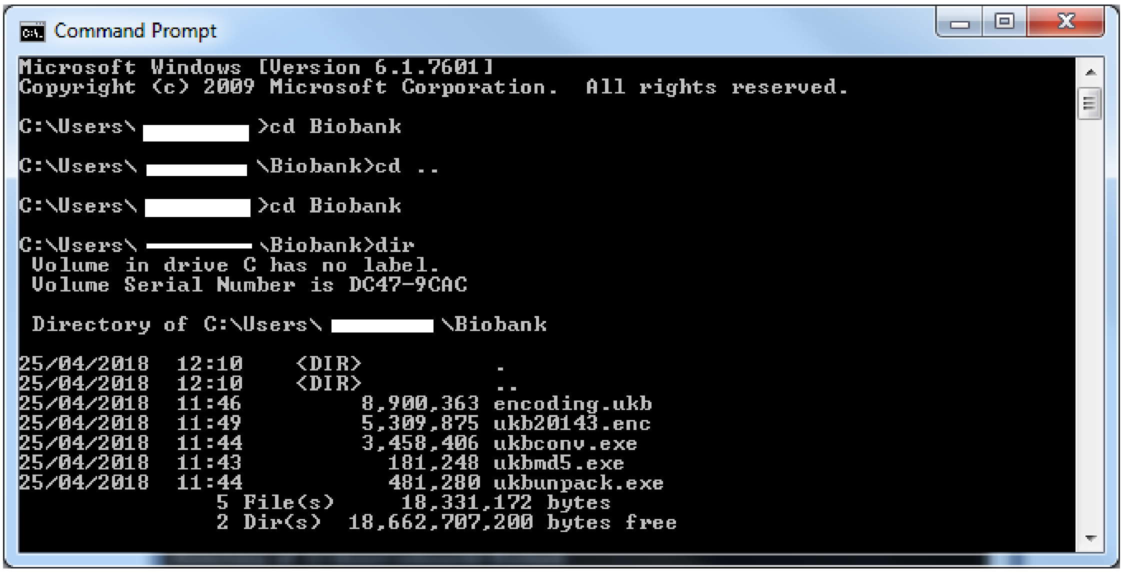
Note that you can also use cd followed by two dots (**cd ..**) to go back to the parent directory, as shown in Figure 8.1.3.



```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\[redacted]>cd Biobank

C:\Users\[redacted]\Biobank>cd ..

C:\Users\[redacted]>
```

**Figure 8.1.3: Moving up a directory**

Use the **cd** command to navigate to the chosen directory. Once you are in the right directory, you can use the **dir** command to list all the files in the current directory (Figure 8.1.4). This allows you to check that you are indeed in the right place: the **dir** command should display the name of the 5 files that you previously downloaded.



```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\[redacted]>cd Biobank

C:\Users\[redacted]\Biobank>cd ..

C:\Users\[redacted]>cd Biobank

C:\Users\[redacted]\Biobank>dir
 Volume in drive C has no label.
 Volume Serial Number is DC47-9CAC

 Directory of C:\Users\[redacted]\Biobank

25/04/2018  12:10    <DIR>          .
25/04/2018  12:10    <DIR>          ..
25/04/2018  11:46         8,900,363 encoding.ukb
25/04/2018  11:49         5,309,875 ukb20143.enc
25/04/2018  11:44         3,458,406 ukbconv.exe
25/04/2018  11:43           181,248 ukbmd5.exe
25/04/2018  11:44           481,280 ukbunpack.exe
               5 File(s)     18,331,172 bytes
               2 Dir(s)  18,662,707,200 bytes free
```

**Figure 8.1.4: Displaying the contents of a directory**

## 8.2 Issues with helper files & utilities

### 8.2.1 General

- If you are trying to run ukbunpack, ukbconv etc in a Windows environment and receive an “Access is denied” error, then it is likely you do not have permissions set to run executable files which are unknown to the system. You may need to log on as an Administrator or contact your local IT support for assistance.

- If you are working in a Linux environment having downloaded a utility such as ukbgene, it will not by default be recognised as being executable. To fix this use the command:

  ```
  chmod 755 ukbgene
  ```

  You may also find that your system cannot locate “ukbgene” because it does not search the current working directory when looking for executable files. The easiest way around this is to prefix the command as follows:

```
./ukbgene <other parameters>
```

to indicate to your system that ukbgene is located in the current directory (designated by a dot . ).

- A “malloc” error, meaning a “memory allocation error”, is encountered when your computer runs out of working memory during the download. This is particularly prone to happening when the Windows versions of the helper files / utilities are being used, and so if this happens we recommend you use the Linux versions instead.

### 8.2.2 ukbunpack

- When attempting to unpack the dataset, if you receive the error:

  *FAIL: Unpack : failed to get uncompressed data - uncompression failed*

  you are probably using the wrong 64-character Password. Please note that the main dataset can only be unpacked using the Password from the keyfile k56789r23456.key contained in the notification email for that particular dataset, i.e. for the dataset released as run 23456. You cannot reuse the keyfile Password from a different data release on the same project (even though they are sent out with the same filenames).

### 8.2.3 ukbconv

- While using the ukbconv utility, some researchers, depending on the variables in their dataset, may see the following error message appear in the command-line terminal:

  *Rosetta error: ROSETTA Error: member "eXXX" not found*
  *Validity error: ROSETTA Error: member "eXXX" not found*
  *(XXX can be any integer)*

  This bug is being investigated at the moment, but this message does not affect the conversion process in any way, and has no consequence on the data being extracted. Researchers can directly open the files generated by ukbconv without worrying about these errors.

### 8.2.4 ukbfetch

- If you are running ukbfetch with a bulk file and are receiving an error indicating that it cannot find data for a particular eid/field combination, then this might be because you created a bulk file (using ukbconv) containing fields which are not accessed using ukbfetch. For instance, if ukbconv is run with the bulk option without specifying a particular field (or set of fields), it will include genomics fields that are downloaded using ukbgene in amongst the bulk fields. Hence, most fields appearing in the bulk file will fail to download because it is not possible to access their data in this way.

### 8.2.5 ukbgene

- When running ukbgene you may find that the data appears to be “fetched” properly, but then cannot be “written”, causing the process to abort. This is most likely due to the large size of some of the genetics data (particularly the imputed data) overwhelming the local storage available during the download. We recommend contacting your local IT support to deal with this issue.

- If you only have a Windows computer available, it is possible to set up a Linux shell to run within it from which you can run ukbgene. Googling “running linux on windows” or similar will provide links describing how to do this.

### 8.2.6 ukbfetch / ukbgene / ukblink

- If you are uncertain whether your IT system will allow you to access the websites:

  - biota.ndph.ox.ac.uk
  - chest.ndph.ox.ac.uk

  needed for bulk and some genetics data, then the command line utility ping can be used to check the connection. From a Windows command line the command:

  ```
  ping biota.ndph.ox.ac.uk
  ```

  will send four signals to the website and report if a reply is received. In Linux the command:

  ```
  ping biota.ndph.ox.ac.uk –w4
  ```

  has the same effect (the –w flag is to limit the number of signals sent, which otherwise will continue until Ctrl-C is entered).

- The keyfile (received as the attachment to your notification email) needs to be in the same directory as the utility. Both utilities by default expect it to have been renamed as .ukbkey (note that this as its full name, with no other file extension). This can cause problems in Windows, and hides the file in Linux (ls -a will show such “hidden files”). If you prefer to give a different name to the keyfile, then ukbfetch, ukbgene or ukblink can still be run, but you will need to add -ak56789r23456.key to the end of your command where k56789r23456.key is replaced by the name of your keyfile.

- If you get the error:

  *Invalid authentication file*
  *File names must be 1-64 characters long*

  it is because you have put a space between the -a and the keyfile name.

- When using a Linux system, if you receive an error along the lines of:

  *`GLIBC_2.14' not found (required by ukbfetch)*

  it means that your local Linux libraries are not compatible with our standard versions of the utility ukbfetch (in this example), ukbgene or ukblink. In each case it is possible to create a version of the utility that will run on your system. See Resource 645 (for ukbfetch), Resource 665 (for ukbgene) or Resource 656 (for ukblink) for further details.

## 8.3 SQL Examples

The following gives some simple examples of how data can be investigated using SQL directly in the Data Portal. These examples should be read in conjunction with the inpatient data dictionary (see the resources tab in [Category 2000]() on Showcase).

Note that SQL generally ignores whitespace (including linebreaks) so the spacing in the examples can be altered without having any effect. Also, it is not necessary for SQL statements such as SELECT to be written in uppercase. This is done simply for clarity, and a lower-case “select” will work just the same.

**Example 1:** To fetch all the fields from table **hesin**, enter:

```
SELECT * FROM hesin
```

To select just the first 100 records we would use:

```
SELECT FIRST 100 * FROM hesin
```

To select the next 100 rows (i.e. rows 101 to 200) we can use:

```
SELECT FIRST 100 * FROM hesin
OFFSET 100
```

**Example 2:** To select just a subset of fields from **hesin** and join these to the primary ICD-10 diagnosis from the **hesin_diag** table:

```
SELECT hesin.eid,
       hesin.ins_index,
       dsource,
       epistart,
       epiend,
       admidate,
       disdate,
       diag_icd10
FROM hesin JOIN hesin_diag USING(eid, ins_index)
WHERE level=1
```

The `level=1` picks only the primary/main diagnosis, and means that only one row in **hesin_diag** will be joined to each **hesin** row. If this condition were omitted the query would return multiple rows for each hesin row, one for each associated diagnosis code.

The eid and ins_index fields need to be prefixed by the table name because they appear in both tables in the join.

Note that the above query partially recreates the structure of the old hesin table where the primary/main diagnosis was included in the main table and secondary diagnoses were split into a separate child table.

**Example 3:** To return all the OPCS4 operation codes, primary and secondary, and episode start dates for records starting from 1st July 2010, and link them to the appropriate participants (via the **hesin** table):

```
SELECT hesin.eid,
       epistart,
       level,
       oper4
FROM hesin JOIN hesin_oper USING(eid, ins_index)
WHERE epistart >= '2010-07-01'
```

**Example 4:** Return a subset of fields from records where the ICD-10 code I21.1 (Acute transmural myocardial infarction of inferior wall) appears as the primary diagnosis:

```
SELECT  hesin.eid,
        hesin.ins_index,
        dsource,
        epistart,
        epiend,
        admidate,
        disdate,
        diag_icd10
FROM hesin JOIN hesin_diag USING(eid, ins_index)
WHERE level=1 and diag_icd10='I211'
```

Note that the decimal point in the code I21.1 is not present in the data, i.e. it appears as I211.

**Example 5:** Continuing from Example 4, we can search more generally for all codes starting I21 (i.e I21.0 to I21.4, and I21.9) by amending the above slightly to:

```
SELECT  hesin.eid,
        hesin.ins_index,
        dsource,
        epistart,
        epiend,
        admidate,
        disdate,
        diag_icd10
FROM hesin JOIN hesin_diag USING(eid, ins_index)
WHERE level=1 and diag_icd10 LIKE 'I21%'
```

The “like” searches for a pattern in the field, and the % functions as a “wildcard” matching any sequence of characters.

If we were interested in secondary diagnoses as well, but not in “external causes”, we could replace this with:

```
SELECT  hesin.eid,
        hesin.ins_index,
        dsource,
        epistart,
        epiend,
        admidate,
        disdate,
        diag_icd10
FROM hesin JOIN hesin_diag USING(eid, ins_index)
WHERE level in (1,2) and diag_icd10 LIKE 'I21%'
```

i.e allow level to be either 1 or 2 (primary or secondary).

**Example 6:** Following on from example 5, we could instead count the number of distinct participants having each of these codes in the inpatient data as follows:

```
SELECT diag_icd10, COUNT(DISTINCT hesin.eid) as number_of_pts
FROM hesin JOIN hesin_diag USING(eid, ins_index)
WHERE level=1 and diag_icd10 LIKE 'I21%'
GROUP BY diag_icd10
```

Which will provide a short table giving the number of distinct participants for which each of the codes I210 – I219 appears as a primary diagnosis in their data. Note that if a participant had both (for example) I214 and I219 in their data they would be counted for both.

## 8.4 Sizes of bulk fields

The following table gives the approximate size, per participant, of a number of the bulk fields available:

| Field | Name | Estimated size per participant (MB) |
|---|---|---|
| 20158 | DXA images | 2 |
| 20201 | Dixon technique for internal fat - DICOM | 71 |
| 20202 | Pancreatic fat - DICOM | 9 |
| 20203 | Liver images | 1 |
| 20204 | OCRM experimental sequence - DICOM | 3 |
| 20206 | Measurements of pancreas volume - DICOM | 2 |
| 20207 | Scout images for heart MRI - DICOM | 7 |
| 20208 | Long axis heart images - DICOM | 9 |
| 20209 | Short axis heart images - DICOM | 81 |
| 20210 | Aortic distensibilty images - DICOM | 6 |
| 20211 | Cine tagging images - DICOM | 5 |
| 20212 | Left ventricular outflow tract images - DICOM | 4 |
| 20213 | Blood flow images - DICOM | 5 |
| 20214 | Experimental shMOLLI sequence images - DICOM | 4 |
| 20215 | Scout images for brain scans - DICOM | 5 |
| 20217 | Functional brain images - task - DICOM | 244 |
| 20218 | Multiband diffusion brain images - DICOM | 128 |
| 20224 | Phoenix - DICOM | <1 |
| 20225 | Functional brain images - resting - DICOM | 360 |
| 20249 | Functional brain images - task - NIFTI | 453 |
| 20250 | Multiband diffusion brain images - NIFTI | 1047 |
| 20251 | Susceptibility weighted brain images - NIFTI | 33 |
| 20252 | T1 structural brain images - NIFTI | 51 |
| 20253 | T2 FLAIR structural brain images - NIFTI | 34 |
| 25747 | Eprime advisor file | <1 |
| 25748 | Eprime txt file | <1 |
| 25749 | Eprime ed2 file | <1 |
| 25750 | rfMRI full correlation matrix, dimension 25 | <1 |
| 25751 | rfMRI full correlation matrix, dimension 100 | <1 |
| 25752 | rfMRI partial correlation matrix, dimension 25 | <1 |
| 25753 | rfMRI partial correlation matrix, dimension 100 | <1 |

## 8.5 File types of returned datasets

The following table gives the zipped format used for older returned datasets. If a return is not on this table then it will be a newer file in .zip format. The file downloaded by ukblink should be renamed to the correct file type, and standard utilities used to unzip the file.

| Return id | Title | Extension |
|---|---|---|
| 124 | Derived variables from application 735/ 15716 - myopia variables | 7z |
| 146 | 5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study | zip |
| 147 | Built Environment Data for Bristol | zip |
| 164 | Suitability of UK BIOBANK Retinal Images for Automatic Analysis of morphometric properties of the vasculature | zip |
| 210 | Built Environment Data - Newcastle and Middlesbrough | 7z |
| 263 | Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan. | zip |
| 265 | The effect of functional hearing and hearing aid usage on verbal reasoning in a large community-dwelling population | zip |
| 362 | Built Environment Data for Birmingham and Nottingham | zip |
| 363 | Built Environment Data for Oxford | zip |
| 403 | New reference values for body composition by bioelectrical impedance analysis in the general population: results from the UK Biobank | 7z |
| 408 | Parental diabetes and birthweight in 236 030 individuals in the UK Biobank Study | 7z |
| 421 | Chronic widespread bodily pain is increased among individuals with history of fracture: findings from UK Biobank | 7z |
| 423 | Do smoking habits differ between women and men in contemporary Western populations? Evidence from half a million people in the UK Biobank study. | zip |
| 424 | Characteristics of rheumatoid arthritis and its association with major comorbid conditions: cross-sectional study of 502 649 UK Biobank participants. | 7z |

| | | |
|---|---|---|
| 463 | Heaviness, health and happiness: a cross-sectional study of 163066 UK Biobank participants | 7z |
| 464 | Psychiatry Gender differences in the association between adiposity and probable major depression: a cross-sectional study of 140,564 UK Biobank participants | 7z |
| 473 | The effect of functional hearing loss and age on long- and short-term visuospatial memory: evidence from the UK Biobank resource | 7z |
| 474 | Better visuospatial working memoery in adults who report profound deafness comapred to those with normal or poor hearing: Data from the UK Biobank resource | 7z |
| 501 | Cognitive function and lifetime features of depression and bipolar disorder in a large population sample: Cross-sectional study of 143,828 UK Biobank participants | 7z |
| 504 | Low birth weight and features of neuroticism and mood disorder in 83545 participants of the UK Biobank cohort | 7z |
| 508 | Prevalence and Characteristics of Probable Major Depression and Bipolar Disorder within UK Biobank: Cross-Sectional Study of 172,751 Participants. | 7z |
| 509 | Associations between single and multiple cardiometabolic diseases and cognitive abilities in 474 129 UK Biobank participants. | 7z |
| 511 | Adiposity among 132,479 UK Biobank participants; contribution of sugar intake vs other macronutrients | 7z |
| 513 | Cognitive Test Scores in UK Biobank: Data Reduction in 480,416 Participants and Longitudinal Stability in 20,346 Participants. | 7z |
| 526 | Change in commute mode and body mass index: prospective longitudinal evidence from UK Biobank | 7z |
| 527 | Active commuting and obesity in mid-life: cross-sectional, observational evidence from UK Biobank | 7z |
| 529 | Lifestyle factors and prostate-specific antigen (PSA) testing in UK Biobank: Implications for epidemiological research | 7z |
| 534 | Ethnic differences in sleep duration and moring-evening type in a population | 7z |

| | | |
|---|---|---|
| 535 | Smoking, screen-based sedentary behaviour, and diet associated with habitual sleep duration and chronotype: data from the UK Biobank | 7z |
| 536 | Interactive effects of sleep duration and morning/ evening preference on cardiovascular risk factors | 7z |
| 542 | Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index | 7z |
| 547 | The influence of social interaction and physical health on the association between hearing and depression with age and gender | 7z |
| 584 | Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk | zip |
| 702 | Case-control association mapping by proxy using family history of disease | zip |
| 717 | Gemome-wide association study identifies 74 loci associated with educational attainment | zip |
| 718 | Genetic variants associated with subjective well-being, depressive symptoms,and neuroticism identified through genome-wide analysis | zip |
| 723 | Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence | zip |
| 726 | Linkage disequilibrium - dependent architecture of human complex traits shows action of negative selection | zip |
| 735 | Red blood cell distribution width: Genetic evidence for aging pathways in 116,666 volunteers | zip |
| 736 | Mixed model association for biobank-scale data sets. | zip |
| 739 | Genome-wide association analyses for lung funtion and chronic obstructive pulmonary disease identify new loci and potnential druggable targets | zip |
| 744 | Genome-wide association study reveals ten loci associated with chronotype in the UKBiobank. | zip |
| 745 | Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits | zip |

| | | |
|---|---|---|
| 749 | Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=122117) | zip |
| 752 | Genome-wide associations for birth weight and correlations with adult disease | zip |
| 760 | Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112151) | zip |
| 762 | Molecular genetic aetiology of general cognitive function is enriched in evolutionarily conserved regions | zip |
| 776 | Rare coding variants pinpoint genes that control human hematological traits | zip |
| 777 | An erythroid-specific ATP2B4 enhancer mediates red blood cell hydration and malaria susceptibility | zip |
| 783 | Cognitive performance among carriers of pathogenic copy number variants: Analysis of 152,000 UK Biobank subjects | 7z |
| 792 | The 'Cognitive footprint' of psychiatric and neurological conditions: cross-sectional study in the UK Biobank Cohort | 7z |
| 793 | Visualization of cancer and cardiovascular disease co-occurrence with network methods | 7z |
| 796 | Psychological distress, neuroticism, and cause-specific mortality: early prospective idence from UK Biobank | 7z |
| 981 | Volumetric measurements of body composition derived from abdominal MRI - application 23889 | zip |
| 1362 | Derived variable from cardiac MRI | zip |
| 1363 | Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort | zip |
| 1364 | Built Environment data for Edinburgh and Glasgow | zip |
| 1365 | Built Environment data for Greater London Authority | zip |
| 1366 | Built Environment data for Liverpool, Manchester and Bury | zip |
| 1367 | Built Environment data for Leeds and Sheffield | zip |
| 1368 | Built Environment data for Stoke | zip |
| 1369 | Built Environment data for Wales | zip |

| | | |
|---|---|---|
| 1455 | Genetic evidence that lower circulating FSH levels lengthen menstrual cycle, increase age at menopause and impact female reproductive health | zip |
| 1456 | Events in Early Life are Associated with Female Reproductive Ageing: A UK Biobank Study | zip |
| 1458 | Vitreoretinal interface abnormalities in middle-aged adults with visual impairment in the UK Biobank study: prevalence, impact on visual acuity and associations | zip |
| 1461 | Monocular and binocular visual impairment in the UK Biobank study: prevalence, associations and diagnoses | zip |
| 1465 | Cost-effectiveness of the polypill versus risk assessment for prevention of cardiovascular disease | zip |
| 1468 | Sex differences in body anthropometry and composition in individuals with and without diabetes in UK Biobank | zip |
| 1469 | Women's reproductive health factors and body adiposity: findings from the UK Biobank | zip |
| 1470 | Differences in morning-evening type and sleep duration between black and white adults: Results from a propensity-matched UK Biobank sample | zip |
| 1472 | Calcium and Vitamin D supplementation are not associated with risk of incident ischemic cardiac events or death: Findings from the UK Biobank Cohort | zip |
| 1475 | Number of offspring and cardiovascular disease risk in men and women | zip |
| 1476 | Chronic multisite pain in major depression and bipolar disorder: cross-sectional study of 149, 611 participants in UK Biobank | zip |
| 1480 | Associations between active commuting and incident cardiovascular disease, cancer and mortality: prospective cohort study | zip |
| 1491 | Human CCL3L1 copy number variation, gene expression, and the role of the CCL3L1-CCR5 axis in lung function | zip |
| 1502 | Long-term intra-individual reproducibility of heart rate dynamics during exercise and recovery in the UK Biobank cohort | zip |
| 1504 | Bone mineral density and risk of type 2 diabetes and coronary heart disease: A Mendelian randomization study | zip |

| | | |
|---|---|---|
| 1522 | Genetic prediction of male pattern baldness | zip |
| 1541 | Self-Reported Facial Pain in UK Biobank Study: Prevalence and Associated Factors | zip |

Cancer Causes & Control (2018) 29:967–986
https://doi.org/10.1007/s10552-018-1072-6

REVIEW ARTICLE

# Review of non-clinical risk models to aid prevention of breast cancer

Kawthar Al-Ajmi[1] · Artitaya Lophatananon[1] · Martin Yuille[1] · William Ollier[1] · Kenneth R. Muir[1]

Received: 28 September 2017 / Accepted: 10 August 2018 / Published online: 3 September 2018

## Abstract

A disease risk model is a statistical method which assesses the probability that an individual will develop one or more diseases within a stated period of time. Such models take into account the presence or absence of specific epidemiological risk factors associated with the disease and thereby potentially identify individuals at higher risk. Such models are currently used clinically to identify people at higher risk, including identifying women who are at increased risk of developing breast cancer. Many genetic and non-genetic breast cancer risk models have been developed previously. We have evaluated existing non-genetic/non-clinical models for breast cancer that incorporate modifiable risk factors. This review focuses on risk models that can be used by women themselves in the community in the absence of clinical risk factors characterization. The inclusion of modifiable factors in these models means that they can be used to improve primary prevention and health education pertinent for breast cancer. Literature searches were conducted using PubMed, ScienceDirect and the Cochrane Database of Systematic Reviews. Fourteen studies were eligible for review with sample sizes ranging from 654 to 248,407 participants. All models reviewed had acceptable calibration measures, with expected/observed (E/O) ratios ranging from 0.79 to 1.17. However, discrimination measures were variable across studies with concordance statistics (C-statistics) ranging from 0.56 to 0.89. We conclude that breast cancer risk models that include modifiable risk factors have been well calibrated but have less ability to discriminate. The latter may be a consequence of the omission of some significant risk factors in the models or from applying models to studies with limited sample sizes. More importantly, external validation is missing for most of the models. Generalization across models is also problematic as some variables may not be considered applicable to some populations and each model performance is conditioned by particular population characteristics. In conclusion, it is clear that there is still a need to develop a more reliable model for estimating breast cancer risk which has a good calibration, ability to accurately discriminate high risk and with better generalizability across populations.

**Keywords** Assessment risk tool · Calibration · Discrimination · Risk factors · Risk prediction · Concordance and E/O statistics

✉ Kenneth R. Muir
kenneth.muir@manchester.ac.uk

Kawthar Al-Ajmi
kawthar.alajmi@postgrad.manchester.ac.uk

Artitaya Lophatananon
artitaya.lophatananon@manchester.ac.uk

Martin Yuille
Martin.Yuille@manchester.ac.uk

William Ollier
Bill.Ollier@manchester.ac.uk

[1] Division of Population Health, Health Services Research and Primary Care, Faculty of Biology, Medicine and Health, Centre for Epidemiology, The University of Manchester, Manchester M139 PL, UK

## Abbreviations

| | |
|---|---|
| CDSR | Cochrane Database of Systematic Reviews |
| ROC | Area under a receiver operating characteristic curve |
| AUC | Area under a receiver operating characteristic curve |
| NPV | Negative predictive value |
| PPV | Positive predictive value |
| TP | True positive |
| TN | True negative |
| FP | False positive |
| FN | False negative |
| NCC | National Cancer Centre cohort |
| KMCC | Korean Multi-center Cancer Cohort |
| NHW | Non-Hispanic white |

| | |
|---|---|
| NCI | National Cancer Institute |
| E/O | Expected/observed |

## Introduction

Breast cancer is the most common cancer among females in high-, middle- and low-income countries and it accounts for 23% of all new female cancers globally [1, 2]. While there has been a significant reduction in mortality, incidence rates have continued to rise [3]. Breast cancer incidence rates are high in North America, Australia, New Zealand, and Western and Northern Europe. It has intermediate levels of incidence in South America, Northern Africa, and the Caribbean but is lower in Asia and sub-Saharan Africa [1].

Early detection of breast cancer improves prognosis and increases survival. Mammographic imaging is the best method available for early detection [4] contributing substantially in reducing the deaths caused by breast cancer [5]. Unfortunately mammography mass screening still leads to some levels of over-diagnosis and over-treatment [6]. As yet routine mammography screening is not readily available globally, particularly in some developing countries [7, 8]. This is supported by the observations that for every million adult women there are only four mammogram screening machines in Sudan has four mammogram machines, whereas Mexico has 37 and Canada has 72 [9]. Under these circumstances, it is clearly more appropriate to prioritize access to mammographic screening or other targeted interventions (such as tamoxifen chemoprevention) for higher-risk individuals who could be identified using a sensitive and specific risk prediction model [10]. Such risk prediction models are individualized statistical methods to estimate the probability of developing certain medical diseases. This is based on specific risk factors in currently healthy individuals within a defined period of time [11]. Such prediction models have a number of potential uses such as planning intervention trials, designing population prevention policies, improving clinical decision-making, assisting in creating benefit/risk indices and estimating the burden cost of disease in population [10].

A general case can also be made for using risk models for certain diseases. For example, their use can allow the application of risk-reducing interventions that may actually prevent the disease in question. If their application can be based on use of existing health records this will avoid increasing levels of anxiety in at least low to moderate risk individuals. The National Cancer Institute of the USA (NCI) has confirmed that the application of "risk prediction" approaches has an extraordinary chance of enhancing "The Nation's Investment in Cancer Research" [12]. This provides an explanation for the rapid increase in the number of models now being reported in the literature [11, 13]. It is clear that not all developed models are valid or can be widely used across populations. The minimum performance measures required for a useful and robust risk prediction model in clinical decision making are discrimination and calibration [14].

We recognize that risk models are increasingly now being used as part of a "triage" assessment for mammography and/or for receipt of other more personalized medical care. There is a growing interest in applying risk prediction models as educational tools.

The models developed can differ significantly with regard to; the specific risk factors that are included; the statistical methodology used to estimate, validate and calibrate risk; in the study design used; and in the populations investigated to assess the models. These differences make it essential that any assessment of model usefulness takes into account both their internal and external validity. Here, we focus on the reliability, discriminatory accuracy and generalizability of breast cancer risk models that exclude clinical (any variable which needs physician input e.g., presence of atypical hyperplasia) and any genetic risk factors. Accurate assessment of risk using easily acquired data is essential as a first stage of tackling the rising burden of breast disease globally. Well-validated models with high predictive power are preferable although this is not the case for all models. The usability of any model is dependent on the purpose the model will be used for and its target populations [15]. Furthermore, it has been suggested that adapting existing predictive models to the local circumstances of a new population rather than developing a new model for each time is a better approach [16].

This review focuses on breast cancer risk predicting models that incorporated modifiable risk factors and/or factors that can be self-reported. Such models could be applied as an educational tool and potentially used to advice at risk individuals on appropriate behavioural changes.

## Methods

### Databases

The following databases were searched for all related publications (up to July 2016): PubMed (https://www.ncbi.nlm.nih.gov/pubmed/); ScienceDirect (http://www.sciencedirect.com/); the Cochrane Database of Systematic Reviews (CDSR) (http://www.cochranelibrary.com/). Terms used for the search were "assessment tool, assessment model, risk prediction model, predictive model, prediction score, risk index, breast cancer, breast neoplasm, breast index, Harvard model, Rosner and Colditz model, and Gail model". Risk models were retrieved based on any study design, study population or types of risk factors.

A Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach was applied for selecting reviewed articles [17]. A total of 61 genetic and non-genetic breast cancer risk models were identified and then filtered to include only risk models with non-clinical factors (Fig. 1). These models contain variables which are considered to be modifiable and/or self-reported by the respondents. For this review, 14 studies were eventually considered to be eligible. No literature reviews were found on breast cancer risk models solely focusing on epidemiological risk factors although all the selected reviews summarized generic composite risk models. The literature search was extended to include publications relating to systematic reviews and meta-analyses; this did not reveal any appropriate publications.

### Confidence in risk factors

Details relating to the degree of confidence in variables used as risk factors in the risk models were taken from the Harvard report [18]. The degree of confidence was categorized as either:

- definite (an established association between outcome and exposure where chance, bias [systematic error], confounders [misrepresentation of an association by unmeasured factor/s] are eliminated with significant confidence)
- probable (an association exists between the outcome and the exposure where chance, bias, confounders cannot be eliminated with sufficient confidence—inconsistent results found with different studies)
- possible (inconclusive or insufficient evidence of an association between the outcome and the exposure)

## Results

### Potential risk factors included in breast cancer non-clinical predictive models

The variables used in the 14 models under review and specifies the degree of confidence (definite, probable or possible) in those variables as risk factors for breast cancer based on the current literature are summarized in Table 1.

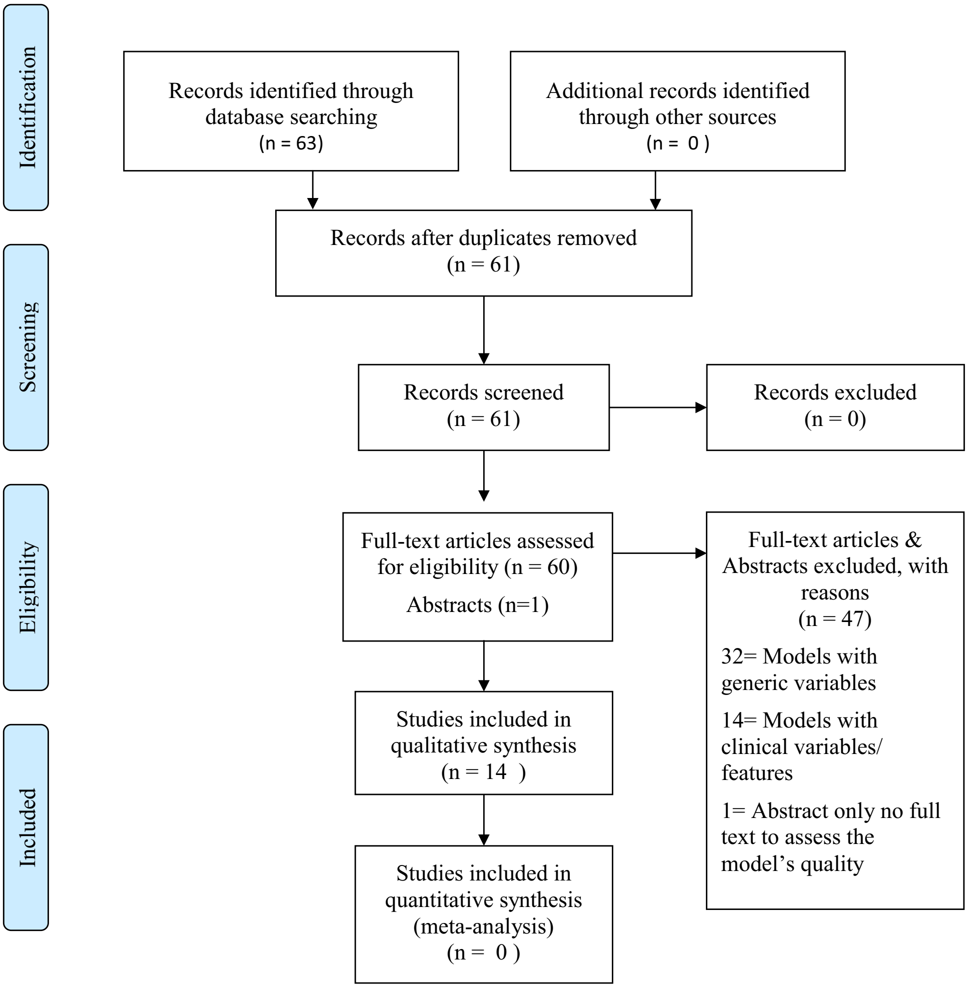

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 63)
Additional records identified through other sources (n = 0 )
Records after duplicates removed (n = 61)
Records screened (n = 61)
Records excluded (n = 0)
Full-text articles assessed for eligibility (n = 60)
Abstracts (n=1)
Full-text articles & Abstracts excluded, with reasons (n = 47)
32= Models with generic variables
14= Models with clinical variables/ features
1= Abstract only no full text to assess the model's quality
Studies included in qualitative synthesis (n = 14 )
Studies included in quantitative synthesis (meta-analysis) (n = 0 )

Fig. 1 Identification of eligible risk models using PRISMA flowchart

**Table 1** Breast cancer risk factors included in the 14 models

| Name of model | Gail [37] | Rosner [42] | Rosner [25] | Colditz [50] | Ueda [38] | Boyle [39] | Lee [36] | Novotny [24] | Gail [32] | Matsuno [51] | Banegas [40] | Pffeifer [31] | Park [23] | Lee [33] | Effect | Level of evidence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Basic characteristics | | | | | | | | | | | | | | | | |
| Age | Yes | Yes | Yes | Yes | | Yes | Yes | Yes | Yes | | Yes | | | Yes | Increased risk | Definite |
| Ethnicity | | | | | | | | | | Yes | | | | | Jewish increased risk | Definite |
| Height | | | | Yes | | | | | | | | | | | Increased risk | Definite |
| Weight | | | | Yes | | | | | | | | | | | Increased risk in post-menopausal | Probable |
| BMI | | | | Yes | Yes | Yes | | | | | | Yes | Yes | Yes | | Probable |
| Alcohol intake | | | | Yes | | Yes | Yes | | | | | Yes | Yes | | Increased risk | Probable |
| Smoking | | | | | | | Yes | | | | | | Yes | | Increased risk | Possible |
| Physical activity | | | | | | Yes | | | | | | | Yes | Yes | Decreased risk | Possible |
| Diet | | | | | | Yes | | | | | | | | | Decreased risk | Probable |
| Hormonal and reproductive factors | | | | | | | | | | | | | | | | |
| Age at menarche | Yes | Yes | Yes | Yes | Yes | Yes | | Yes | Yes | Yes | Yes | | Yes | Yes | Increased risk | Definite |
| Age at first live birth | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Increases risk | Definite |
| Age at subsequent birth | | Yes | Yes | | | | | | | | | | | | Increases risk | Definite |
| Age at menopause | | Yes | Yes | Yes | | | | | | | | Yes | Yes | Yes | Increased risk | Definite |
| Hormone replacement therapy use | | | | Yes | | Yes | | | | | | Yes | Yes | | Increases risk | Definite |
| Oral contraceptive use | | | | Yes | | | | Yes | | | | | Yes | | Increases risk | Definite |
| Breast feeding | | | | | | | Yes | | | | | | Yes | | Decreases risk | Probable |

**Table 1** (continued)

| Name of model | Gail [37] | Rosner [42] | Rosner [25] | Colditz [50] | Ueda [38] | Boyle [39] | Lee [36] | Novotny [24] | Gail [32] | Matsuno [51] | Banegas [40] | Pfeifer [31] | Park [23] | Lee [33] | Effect | Level of evidence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pregnancy | | | | | | | | | | | | | Yes | | Decreases risk | Possible |
| Parity | | | | Yes | | | | | | | | Yes | | | Decreases risk | Definite |
| Children number | | | | Yes | | | | | | | | | | Yes | Decreases risk | Possible |
| Menopause type | | | | Yes | | | | | | | | | | | Surgical menopause reduces risk | Possible |
| Menstrual regularity | | | | | | | Yes | | | | | | | | Menstrual regularity and duration—inconsistent results | Possible |
| Menstrual duration | | | | | | | Yes | | | | | | | Yes | | Possible |
| Menopausal status | | | | | | | | | | | | Yes | | Yes | Post-menopause increases risk | Possible |
| Gestation period | | | | | | | | | | | | | | Yes | Increases risk | Possible |
| Family history of breast and/or ovarian cancer or diseases | | | | | | | | | | | | | | | | |
| Family history of breast cancer | Yes | | | Yes | Yes | Yes | Yes | Yes | | Yes | | Yes | Yes | Yes | Increases risk | Definite |
| First-degree relatives with breast cancer | Yes | | | Yes | | | | | Yes | | Yes | | | | Increases risk | Definite |
| Age of onset of breast cancer in a relative | | | | | | Yes | | | | | | | | | Increases risk | Probable |
| Benign breast disease | | | | Yes | | | | Yes | | | | Yes | | | Increases risk | Probable |
| History of breast biopsies | Yes | | Yes | | | | | Yes | Yes | Yes | Yes | | Yes | | Increases risk | Definite |
| Mammogram | | | | | | | | | | | | | | Yes | Increases risk | Probable |

**Table 1** (continued)

| Name of model | Gail [37] | Rosner [42] | Rosner [25] | Colditz [50] | Ueda [38] | Boyle [39] | Lee [36] | Novotny [24] | Gail [32] | Matsuno [51] | Banegas [40] | Pfeifer [31] | Park [23] | Lee [33] | Effect | Level of evidence |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Summary of risk factors included in each model | | | | | | | | | | | | | | | | |
| Definite factors | 5 | 5 | 6 | 10 | 3 | 5 | 3 | 6 | 3 | 5 | 5 | 5 | 7 | 5 | Max of 10 and min of 3 factors | |
| Probable factors | 0 | 0 | 0 | 4 | 1 | 3 | 2 | 1 | 0 | 1 | 0 | 3 | 3 | 2 | Max of 4 and min of 0 factors | |
| Possible factors | 0 | 0 | 0 | 2 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 1 | 3 | 5 | Max of 5 and min of 0 factors | |
| Total factors | 5 | 5 | 6 | 16 | 4 | 8 | 8 | 7 | 3 | 5 | 5 | 9 | 13 | 12 | Max of 16 and min of 3 factors | |

Age, age at first birth, age at menarche, family history of breast cancer, and self-reported history of biopsies were the most common variables used amongst the 14 models selected. These variables are considered as definite risk factors for developing breast cancer [18]. Other additional variables were observed in fewer models. These included ethnicity (Jewish—definite), definite hormonal replacement therapy, diet (some probable and others possible), physical activity (possible), height (definite), weight (probable- for pre-menopausal women and definite for post-menopausal women). Among pre-menopausal females, weight is considered to be a protective factor [19]. In contrast amongst post-menopausal women, weight is considered to be a risk factor [20–22] as is parity, oral contraceptive pill use (definite), pregnancy history, timing and type of menopause (definite), menstrual regularity (possible), menstrual duration and gestation period (probable), smoking (possible), mammogram screening (probable) and age of onset of breast cancer in a relative (definite).

The largest number of definite factors included in a model ($n = 10$ variables) was seen in the study reported by Colditz and Rosner [18]. This was followed by studies by reported by Park [23], Novotny [24] and Rosner [25]. We evaluated the number of the definite, probable and possible variables in the models to compare their performance based on the type and number of the variable included.

## Evaluation measures of the risk models

The most important measures used to assess the performance of the models were considered to be as follows:

- Calibration (reliability): the E/O statistic measures the calibration performance of the predictive model. Calibration involves comparing the expected versus observed numbers of the event using goodness-of-fit or chi square statistics. A well-calibrated model will have a number close to 1 indicating little difference between the E and O events. If the E/O statistic is below 1.0 then the event incidence is underestimated, while if the E/O ratio is above 1.0 then incidence is overestimated [14, 26].
- Discrimination (precision): the C statistic (Concordance statistic) measures the discrimination performance of the predictive model and corresponds to the area under a receiver operating characteristic curve. This statistic measures how efficiently the model is able to discriminate affected individuals from un-affected individuals. A C-statistic of 0.5 indicates no discrimination between individuals who go on to develop the condition and those who do not. In contrast, a C-statistic of 1 implies perfect discrimination [27, 28]. Good discrimination is important for screening individuals and for effective clinical decision making [10].

**Table 2** Formulas used to calculate the accuracy of the model

| Term | Definition | Equation |
|---|---|---|
| Sensitivity | Probability of a test will indicate 'positive' among those with the disease | (TP)/(TP + FN) |
| Specificity | Probability of a test will indicate 'negative' among those without the disease | (TN)/(TN + FP) |
| Positive predictive value | Probability of a patient having disease when test is positive | (TP)/(TP + FP) |
| Negative predictive value | Probability of a patient not having disease when test is negative | (TN)/(FN + TN) |

*TP* True positive, *TN* true negative, *FP* false positive, *FN* false negative

- Accuracy: is tested by measuring of 'sensitivity', 'specificity', 'positive predictive value' (PPV) and negative predictive value (NPV). All of these terms are defined in Table 2. These measures indicate how well the model is able to categorize specific individuals into their real group (i.e., 100% certain to be affected or unaffected). Accuracy is equally important for both individual categorisation and for clinical decision making. Nevertheless, even with good specificity or sensitivity, low positive predictive values may be found in rare diseases [10] as the predictive values also depend on disease prevalence. With high prevalence, PPV will increase while NPV will decrease [29].
- Utility: this evaluates the ease with which the target groups (public, clinicians, patients, policy makers) can submit the data required by the model. Utility evaluation assesses lay understanding of risk, risk perception, results interpretation, level of satisfaction and worry [30]. This evaluation usually uses surveys or interviews [26].Calibration and discrimination were the most common measures used to assess the breast cancer risk models under review and these measures are summarized in Fig. 2. Internal calibration was performed in just three of the 14 models with values ranging from 0.92 to 1.08. These calibration values represented a good estimate of the affected cases using these models. For external calibration, six of the 14 models used an independent cohort. Rosner [25] and Pfeiffer [31] reported the highest with E/O values of 1.00 and followed by Colditz [18] with an E/O of 1.01.

The C-Statistic values measuring internal discrimination ranged across studies from 0.61 to 0.65. The Park [23] model achieved the best outcome (C-Statistic = 0.64). Additionally, Park [23] showed the highest value with a C-Statistic of 0.89 when applied to subjects recruited from the NCC (National Cancer Centre) screening program. The lowest C-Statistic (0.56) was observed in the Gail model [32]. Overall, this demonstrates that the models have better calibration than discrimination. Accuracy was only evaluated in the Lee model [33]. Sensitivity, specificity and overall accuracy were calculated. The values indicate low accuracy with values ranging from 0.55 to 0.66 (Table 3).

In qualitative research relating to the impact and utility [34] of the Harvard Cancer Risk Index (HCRI) [18], nine focus groups (six female, three male) showed good overall satisfaction with HCRI. Participants appreciated both the detailed explanation and the updated inclusion of risk factors. On the other hand, some participants criticized the absence of what they considered to be important factors (e.g., environmental factors and poverty). Some participants

**Table 3** Summary of the evaluation measures of the risk models

| Model | Calibration | | | Discrimination | | | Accuracy | Utility |
|---|---|---|---|---|---|---|---|---|
| | Derived model | Internal | External | Derived model | Internal | External | Sensitivity, specificity, PPV, NPV | |
| Gail [37] | | | 0.79–1.12 | | | 0.58–0.67 | | |
| Rosner [42] | – | – | – | – | – | | – | – |
| Rosner [25] | | – | 1.00 (0.93–1.07)[d] | – | | 0.57 (0.55–0.59)[d] | | |
| Colditz [50] | – | – | 1.01 (0.94–1.09)[d] | – | | 0.64 (0.62–0.66)[d] | – | Good[e] |
| Ueda [38] | – | – | – | – | – | | – | – |
| Boyle [39][a] | | (a) 0.96 (0.75–1.16) cohort1<br>(b) 0.92 (0.68–1.16) cohort2 | – | 0.59 | | | – | – |
| Lee [36] | – | – | – | – | – | | – | – |
| Novotny [24] | – | – | – | – | – | | – | – |
| Gail [32] | – | 1.08 (0.97–1.20) | 0.93 (0.97–1.20)[f] | – | | 0.56 (0.54–0.58)[f] | – | – |
| Matsuno [51] | 1.17 (0.99–1.38) | | | | 0.614 (0.59–0.64) | | – | – |
| Banegas [40][b] | – | (a) 1.08 (0.91–1.28); Hispanic<br>(b) 0.98 (0.96–1.01); NHW | – | – | – | – | – | – |
| Pfeiffer [31] | | | 1.00 (0.96–1.04) | | | 0.58 (0.57–0.59) | | |
| Park [23][c] | – | – | (a) 0.97(0.67–1.40); KMCC<br>(b) 0.96 (0.70–1.37); NCC | – | (a) 0.63 (0.61–0.65) < 50 years (KMCC)<br>(b) 0.65 (, 0.61–0.68) ≥ 50 years (KMCC) | (a) 0.61(0.49–0.72); KMCC<br>(b) 0.89(0.85–0.93); NCC | – | – |

**Table 3** (continued)

| Model | Calibration | | | Discrimination | | | Accuracy | Utility |
|---|---|---|---|---|---|---|---|---|
| | Derived model | Internal | External | Derived model | Internal | External | Sensitivity, specificity, PPV, NPV | |
| Lee [33] | | | | | Overall: 0.62 (0.620–0.623)<br>Under 50: 0.61 (0.60–0.61)<br>Above 50: 0.64 (0.63–0.64) | | (a) Sensitivity<br>Overall: 0.55 (0.54–0.56)<br>$<$ 50: 0.61 (0.60–0.62)<br>$>$ 50:0.59 (0.59–0.60)<br>(b) Specificity<br>Overall: 0.66 (0.65–0.67)<br>$>$ 50: 0.58 (0.57–0.59)<br>$<$ 50:0.64 (0.63–0.65)<br>(c) Accuracy<br>Overall: 0.60 (0.60–0.61)<br>$>$ 50:0.59 (0.59–0.60)<br>$<$ 50:0.61 (0.61–0.62) | – |

[a] Boyle [39] used two cohorts for calibration (1-cohort with complete follow-up and 2-cohort with 5 years of follow-up at most)

[b] Banegas [40] used two cohorts for calibration (1-Hispanic and 2-non-Hispanic white (NHW))

[c] Park [23] used two cohorts for calibration and discrimination, using two Korean cohorts: 1-the Korean Multi-center Cancer Cohort (KMCC) and 2-National Cancer Centre (NCC) cohort

[d] [49]

[e] [52]

[f] [11]

believed that some of the factors on which subjects had been assessed might cause anxiety. It is also noted, however, that the case has been made that such anxiety provides motivation for action to mitigate risk [35].

### Overview of current models

All the models described (except for Lee et al. 2004) [36] are extended versions of either the Gail model or the Rosner and Colditz model (Tables 4, 5). The Gail model developed in 1989 [37] was the first risk model for breast cancer and included the following variables: age, menarche age, age at first birth, breast cancer history in first-degree relatives, history of breast biopsies and history of atypical hyperplasia. The range of calibration of the Gail modified models was E/O = (0.93–1.17) and the discrimination range was C-Statistics = (0.56–0.65). This indicates that these models are well calibrated, although discrimination could be improved.

Ueda et al. [38] modified the Gail model by including age at menarche, age at first delivery, family history of breast cancer and BMI in post-menopausal women, as risk factors in his model for Japanese women. However, as with the original Gail model, no validation was performed. In the Boyle model [39], more factors were included such as alcohol intake, onset age of diagnosis in relatives, one of the two diet scores and BMI and HRT. This results in calibration with E/O close to unity and less acceptable discrimination of C-stat = 0.59. The Novotny model [24] added the number of previous breast biopsies performed on a woman and her history of benign breast disease. However, no validation assessment was performed for this model. Newer models [32, 40, 41] included the number of benign biopsies. This resulted in acceptable calibration but less acceptable discrimination (Gail [32]: E/O = 0.93; C-stat = 0.56; Matsuno: E/O = 1.17, C-statistic = 0.614; and Banegas E/O = 1.08). Park et al. [23] included menopausal status, number of pregnancies, duration of breastfeeding, oral contraceptive usage, exercise, smoking, drinking, and number of breast examinations as risk factors. This model has an E/O = 0.965; C-stat = 0.64. However, the C-statistic reported from the external validation cohort was high compared to the original C-statistic. They reported a C-statistic of 0.89 using the NCC cohort. This discrepancy was claimed to be caused by the population characteristics (participants were 30 years and above, recruited from cancer screening program, from a teaching hospital in an urban area) [23]. In the same year, Pfeiffer et al. [23] developed a model where parity was considered as a factor and had E/O of 1.00 and a C-statistic of 0.58. The later Gail model published in 2007 used logistic regression to derive relative risks. These estimates are then combined with attributable risks and cancer registry incidence data to obtain estimates of the baseline hazards [32].

The Rosner and Colditz model of 1994 [42] was based on a cohort study of more than 91,000 women. The model used Poisson regression (rather than logistic regression as in the Gail model). The variables were as follows: age, age at all births, menopause age, and menarche age. This model was not validated. A new version in 1996 [25] included one modification (current age was excluded) and gave an E/O = 1.00 and a C-statistic = 0.57. In 2000, Colditz et al. [18] modified the model with risk factors for: benign breast disease, use of post-menopausal hormones, type of menopause, weight, height, and alcohol intake. This model gave an E/O = 1.01; C-statistic = 0.64.

Lee et al. [36] used two control groups: a “hospitalised” group and a nurses and teachers group. The risk factors in the hospitalized controls were as follows: family history, menstrual regularity, total menstrual duration, age at first full-term pregnancy, and duration of breastfeeding. The risk factors in the nurses/teachers control group were as follows: age, menstrual regularity, alcohol drinking status and smoking status. This model was not based on Gail or Rosner and Colditz. Hosmer–Lemeshow goodness of fit was used to assess model fit which had a $p$ value = 0.301 in (hospital controls) and $p$ value = 0.871 in (nurse/teacher controls). No calibration or discrimination measures were reported.

Lee [33] used three evaluation techniques to assess the discrimination and the accuracy of their model: support vector machine, artificial neural network and Bayesian network. Of the three, support vector machine showed the best values among the Korean cohort. However, accuracy and discrimination were less acceptable in this model.

In summary, calibration performance is similar between models (Modified Gail and modified Rosner, Colditz), yet modified Gail models showed better discrimination performance with the C-statistic of the Park model being 0.89.

## Discussion

There is increasing interest among clinicians, researchers and the public in the use of risk models. This makes it important that we fully evaluate model development and application. Each risk model should be assessed before it can be recommended for any clinical application. Performance assessment should involve the use of an independent population [43] separate from the population used to build the model. We have reviewed breast cancer risk models that include non-genetic and non-clinical risk factors but exclude clinical risk factors. By using PubMed, ScienceDirect, Cochrane library and other research engines, 14 models met these criteria. The most recent model examined was developed in 2015 [33]. Most models were based on two earlier risk models developed over 20 years ago—the Gail model [37] and the Rosner and Colditz model [42]. The

**Table 4** Characteristic summary of the reviewed breast cancer risk models

| Author/model | Study design | Participants | Ethnicity | Outcome | Statistical method | Effect estimates | Sample size | Risk factors considered in the models | Age target | Stratification |
|---|---|---|---|---|---|---|---|---|---|---|
| Gail [37] | Case–control | White American females from the Breast Cancer Detection Demonstration Project (BCDDP) | American–Caucasian | Invasive breast cancer + in situ carcinoma | unconditional logistic regression | Relative risk | 2,852 cases<br>3,146 controls | Age at menarche, age at first live birth, number of previous biopsies, and number of first-degree relatives with breast cancer | Any age | None |
| Rosner [42] | Cohort | Registered nurses | American–Caucasian | Invasive breast cancer | Poisson regression | Cumulative incidence | 2,341 cases, 91,523 controls | Age, age at all births, menopause age, menarche age | 30–55 years | Number of births |
| Rosner [25] | Cohort | Registered nurses | American–Caucasian | Invasive breast cancer | Poisson regression | Relative risk | 2,249 cases, 89,132 controls | Menarche age, first live birth age, subsequent births age, menopause age | Any age | None |
| Colditz [50] | Cohort | General women | American–Caucasian | Invasive breast cancer | Poisson regression | Cumulative incidence | 1,761cases<br>56,759 controls | Benign breast disease, use of HRT, weight, height, menopausal type, and alcohol intake | Women aged 30–55 years | None |
| Ueda [38] | Case–control | General women | Japanese–Asian | Invasive breast cancer | Conditional logistic regression | Relative risk | 376 cases<br>430 controls | Menarche, first birth age, family history, and BMI in post-menopausal women | Any age | Menopausal status |
| Boyle [39] | Case–control | General women | Italian–Caucasian | Invasive breast cancer | Conditional logistic regression | Absolute + relative risk | 2,569 cases<br>2,588 controls | Menarche age, first birth age, alcohol intake, family history, age of diagnosis in relatives, and one of the two diet scores. BMI and HRT were included only for women older > 50 | 23–74 years (cases)<br>20–74 years (controls) | Age (< 50 and > 50) |
| Lee [36] | Case–control | 1-General women<br>2-Well educated (nurse/teacher) | Korean–Asian | Invasive breast cancer | Hosmer–Lemeshow goodness of fit | Probability | 384 cases<br>270 controls | With hospitalized controls: family history, menstrual regularity, total menstrual duration, first full-term pregnancy age, breastfeeding duration while with nurse/teacher controls: age, menstrual regularity, drinking status, smoking status | Age at least 20 years | None |

**Table 4** (continued)

| Author/model | Study design | Participants | Ethnicity | Outcome | Statistical method | Effect estimates | Sample size | Risk factors considered in the models | Age target | Stratification |
|---|---|---|---|---|---|---|---|---|---|---|
| Novotny [24] | Case–control | General women | Czeck females–Caucasian | Invasive breast cancer | Unconditional Logistic regression | Relative risk | 4,598 matched pairs | Age at birth of first child, family history of breast cancer, No. of previous breast biopsy, menarche age, parity, history of benign breast disease | Age matched | None |
| Gail [32] | Case–control | General women | African American | Invasive breast cancer | Conditional logistic regression | Absolute + relative risk | 1,607cases<br>1,647 controls | Menarche age, No. of affected mother or sisters, No. of benign biopsy | 35–64 years | Age (< 50 and > 50) |
| Matsuno [51] | Case–control | General women | Asian and Pacific Islander American | Invasive breast cancer | Conditional logistic regression | Absolute + relative + attributable risks | 589 cases<br>952 controls | Menarche age, age at first live birth, No. of biopsies, family history, ethnicity | Any age | Ethnicity |
| Banegas [40] | Longitudinal study | General women | Hispanic | Invasive breast cancer | Cox proportional hazards regression | Relative risk | 6,353 cases<br>128,976 controls | Age, age at first live birth, menarche age, No. of first-degree relatives with breast cancer, No. of breast biopsies | Post-menopausal participants aged ≥ 50 | None |
| Pfeiffer [31] | Prospective study | White over 50 years old | White and non-Hispanic Caucasian | Invasive breast cancer | Cox proportional hazards regression | Relative and attributable risks | 7,695 cases<br>240,712 controls | BMI, oestrogen and progestin MHT use, other MHT use, parity, age at first birth, premenopausal, age at menopause, benign breast diseases, family history of breast or ovarian cancer, and alcohol consumption | 50 and above | None |

**Table 4** (continued)

| Author/model | Study design | Participants | Ethnicity | Outcome | Statistical method | Effect estimates | Sample size | Risk factors considered in the models | Age target | Stratification |
|---|---|---|---|---|---|---|---|---|---|---|
| Park [23] | Case–control | General women | Korean–Asian | Invasive breast cancer | Unconditional Logistic regression | Absolute risk | 3,789 cases<br>3,789 controls | Family history, menarche age, menopausal status, menopause age, pregnancy, first full-term pregnancy age, No. of pregnancies, breastfeeding duration, OC usage, HRT, exercise, BMI, smoking, drinking, No. of breast examinations | Any age | Age (< 50 and > 50) |
| Lee [33] | Case–control | General women | Asian | Invasive breast cancer | Conditional logistic regression | | 2,291 cases and 2,283 controls | First full-term pregnancy age, children No., menarche age, BMI, family history, menopausal status, regular mammography, exercises, oestrogen exposure duration, gestation period, menopause age | Any age | Age (< 50 and > 50) |

**Table 5** Models reviewed in this article

| | Title | Size of study | Population | First author | References |
|---|---|---|---|---|---|
| Included in this review | Projecting individualized probabilities of developing breast cancer for white females who are being examined annually | 2,852 cases<br>3,146 controls | Caucasian | Gail 1989 | [37] |
| | Reproductive risk factors in a prospective study of breast cancer: the Nurses' Health Study | 2,341 cases, 91,523 controls | Caucasian | Rosner 1994 | [42] |
| | Nurses' health study: log-incidence mathematical model of breast cancer incidence | 2,249 cases, 89,132 controls | Caucasian | Rosner 1996 | [25] |
| | Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study | 1,761cases<br>56,759 controls | Caucasian | Colditz | [50] |
| | Estimation of individualized probabilities of developing breast cancer for Japanese women | 376 cases<br>430 controls | Asian | Ueda | [38] |
| | Contribution of three components to individual cancer risk predicting breast cancer risk in Italy | 2,569 cases<br>2,588 controls | Caucasian | Boyle | [39] |
| | Determining the Main Risk Factors and High-risk Groups of Breast Cancer Using a Predictive Model for Breast Cancer Risk Assessment in South Korea | 384 cases<br>270 controls | Asian | Lee | [36] |
| | Breast cancer risk assessment in the Czech female population–an adjustment of the original Gail model | 4,598 matched pairs | Caucasian | Novotny | [24] |
| | Projecting individualized absolute invasive breast cancer risk in African American women | 1,607cases<br>1,647 controls | African | Gail | [32] |
| | Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women | 589 cases<br>952 controls | Asian | Matsuno | [51] |

**Table 5** (continued)

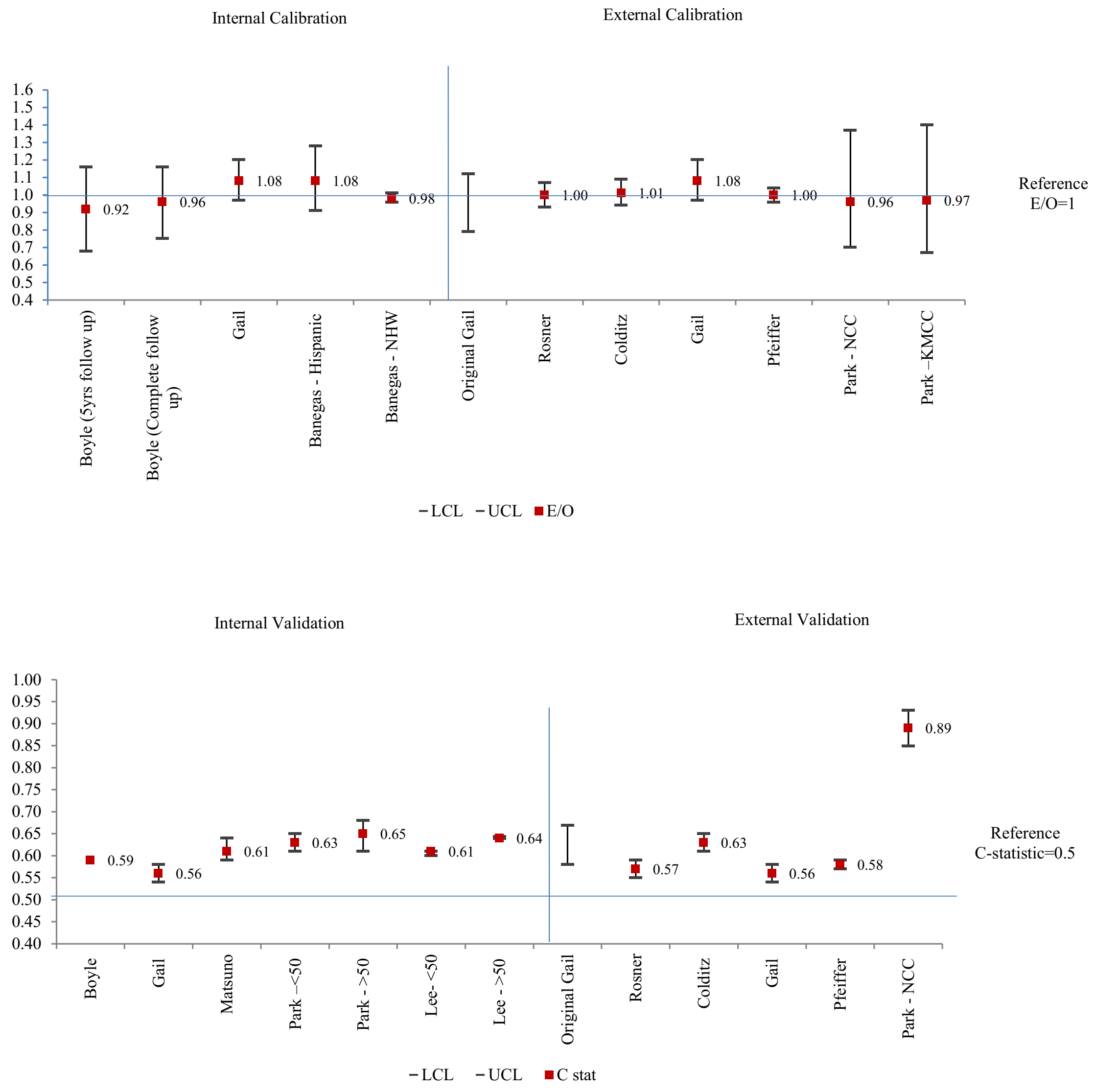
| | Title | Size of study | Population | First author | References |
|---|---|---|---|---|---|
| | Evaluating breast cancer risk projections for Hispanic women | 6,353 cases<br>128,976 controls | Hispanic | Banegas | [40] |
| | Risk Prediction for Breast, Endometrial, and Ovarian Cancer in White Women Aged 50 y or Older: Derivation and Validation from Population-Based Cohort Studies | 42,821 cases<br>114,931 controls | White, non-Hispanic women aged 50+ | Pfeiffer | [53] |
| | Korean risk assessment model for breast cancer risk prediction | 3,789 cases<br>3,789 controls | Asian | Park | [23] |
| | Computational Discrimination of Breast Cancer for Korean Women Based on Epidemiologic Data Only | 2,291 cases and 2,283 controls | Asian | Lee | [33] |
| Excluded from this review | [54–101] | | | | |

modified versions of these two original models varied in the risk factors included and the estimation methods used. In 2012, there were two literature reviews published which analysed breast cancer risk prediction models [11, 28]; however, our review focuses particular on modifiable risk factors and/or self-reported factors and we have updated the models published after 2012 [23, 31, 33].

Most models with modifiable risk factors included report acceptable calibration, with E/O close to 1 but less acceptable discrimination with C-statistic close to 0.5. Calibration and validation were improved when more definite factors were included. A possible explanation for less acceptable discrimination performance could be the inclusion of weaker evidence-based factors (probable and possible risk factors). All the models had combinations of probable and possible factors with no single model restricted to the inclusion of the definite factors.

Various factors affect model performance. Inclusion of less significant factors is likely to occur in studies with small sample sizes [11, 28]. Some important clinical risk factors were not included and this may affect the model's final performance [44]. Breast cancer heterogeneity may also contribute to poor performance as different cancer types may have different risk factors [11]. Most of the models included in this review did not stratify breast cancer into its subtypes during model development. Rosner and Colditz however evaluated the model's performance based on breast cancer subtypes (ER±, PR ± or HR2±) and concluded that risk factors vary according to the subtypes [45, 46]. Finally, even when strong risk factors are included in a model, significant increases in C-statistic have not been seen [47].

Model performance statistics were affected by the criteria used to stratify the analysis. Four models were stratified by age (below 50 and above 50). One model was further stratified by menopausal status [38], one by ethnicity [41] and one by number of births [42]. Breast cancer risk models could be improved if appropriate factors were used to stratify the population. For example, pre-menopausal and post-menopausal females have different risk factors in breast cancer development. The models that applied menopausal status have some limitation in that this may not be applicable to women who have had hysterectomy. For example, in the US, hysterectomy is the second most common procedure performed and the likelihood of oophorectomy varies by age at hysterectomy [48]. Hence, completion of risk assessment outside of a clinical setting is problematic as women may be challenged to define their menopausal status. Even though the overall performance of these models appears to be moderate in differentiating between cases and non-cases, they may still serve as a good educational tool as part of cancer prevention. Utility evaluation assesses the public's knowledge of breast cancer risk factors rather well and could be used to promote cancer risk reduction actions.



[i] Boyle model & Park model were adjusted for age when C-statistic were calculated

**Fig. 2** Calibration and discrimination performances of the 13 breast cancer risk models

A significant limitation in the development of risk models is the absence of consensus standards for defining and classifying a model's performance. For example what is the level of good or acceptable calibration or measures of discrimination? what are acceptable measures of specificity and sensitivity in diagnostic/prognostic/preventive models? how close to unity should calibration and discrimination be for a model to be considered valid? what is the utility cut-off in each type of model? All of these questions are hard to answer without global agreement. However, this lack of consensus is understandable as these values vary depending on the type of the model type (diagnostic, prognostic, preventive), goal (clinical tool, educational tool, screening tool), targeted audience (public, high-risk patients, patients visiting the clinic) and the disease itself and its types or subtypes (such as breast cancer, familial breast cancer, lobular/ductal/invasive/in situ carcinoma breast cancer). This suggests that the closer value of E/O and C-statistics to 1, the better model performance. Such a pragmatic attitude permits us to begin to focus on improving the availability of effective risk reduction actions.

Furthermore, some of the models reviewed cannot be applied to some of the populations as the risk factors may vary between different populations. For example, alcohol consumption would not be applicable to Muslim women. We recommend that researchers develop a more reliable and valid breast cancer risk model which has good calibration, accuracy, discrimination and utility where both internal and external validations indicate that it can be reliable for general use. In order to improve our models, the following should be considered: (1) the model type (diagnostic, prognostic, preventive), goal (clinical tool, educational tool, screening tool), targeted audience (public, high-risk patient), (2) inclusion of definite risk factors while incorporating the clinical and/or genetic risk factors where possible, (3) dividing the model into disease subtypes, age and menopausal status, (4) ensuring that a model is developed that can be validated externally.

**Acknowledgments** We are very grateful to Ministry of Higher Education in Oman for supporting the first author with her PhD studies without their financial support we would not be able to publish this paper.

**Author contributions** Conception and design: KM, AL and KA. Development of methodology: KM, AL and KA. Acquisition of data: KM, KA and AL. Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): KA, AL. Writing, review, and/or revision of the manuscript: KA, KM, AL, MY, WO.

**Funding** Kenneth R. Muir and Artitaya Lophatananon were funded by ICEP ("This work was also supported by CRUK [Grant Number C18281/A19169]").

## Compliance with Ethical Standards

**Conflict of interest** None of the authors have any competing interests.

## References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011) Global cancer statistics. CA: a cancer. J Clin 61(2):69–90. https://doi.org/10.3322/caac.20107
2. Li CI (2010) Breast cancer epidemiology. Springer, Berlin. https://doi.org/10.1007/978-1-4419-0685-4
3. Parkin DM, Fernández LMG (2006) Use of statistics to assess the global burden of breast cancer. Breast J 12:S70–S80. https://doi.org/10.1111/j.1075-122X.2006.00205.x
4. Schreer I, Lüttges J (2005) Breast cancer: early detection. In: Gourtsoyiannis NC, Ros PR (eds) Radiologic-pathologic correlations from head to toe: understanding the manifestations of disease. Springer, Berlin, pp 767–784. https://doi.org/10.1007/3-540-26664-x_35
5. Tabar L, Yen MF, Vitak B, Chen HHT, Smith RA, Duffy SW (2003) Mammography service screening and mortality in breast cancer patients: 20-year follow-up before and after introduction of screening. Lancet 361(9367):1405–1410. https://doi.org/10.1016/S0140-6736(03)13143-1
6. Gotzsche PC, Jorgensen KJ (2013) Screening for breast cancer with mammography. Cochrane Database Syst Rev. https://doi.org/10.1002/14651858.CD001877.pub5
7. Anderson BO, Braun S, Lim S, Smith RA, Taplin S, Thomas DB (2003) Early detection of breast cancer in countries with limited resources. Breast J 9(Suppl 2):S51–S59
8. Yip CH, Smith RA, Anderson BO, Miller AB, Thomas DB, Ang ES, Caffarella RS, Corbex M, Kreps GL, McTiernan A (2008) Guideline implementation for breast healthcare in low- and middle-income countries: early detection resource allocation. Cancer 113(8 Suppl):2244–2256. https://doi.org/10.1002/cncr.23842
9. Li J, Shao Z (2015) Mammography screening in less developed countries. SpringerPlus 4:615. https://doi.org/10.1186/s40064-015-1394-8
10. Freedman AN, Seminara D, Gail MH, Hartge P, Colditz GA, Ballard-Barbash R, Pfeiffer RM (2005) Cancer risk prediction models: a workshop on development, evaluation, and application. J Natl Cancer Inst 97(10):715–723. https://doi.org/10.1093/jnci/dji128
11. Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A (2012) Risk prediction models of breast cancer: a systematic review of model performances. Breast Cancer Res Treat 133(1):1–10. https://doi.org/10.1007/s10549-011-1853-z
12. National Cancer Institute (2005) The nation's investment in cancer research. A plan and budget proposal for the fiscal year 2006. https://www.cancer.gov/about-nci/budget/plan/. 2018
13. Gerds TA, Cai T, Schumacher M (2008) The performance of risk prediction models. Biom J 50(4):457–479. https://doi.org/10.1002/bimj.200810443
14. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW (2010) Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology 21(1):128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2
15. Moons KG, Altman DG, Vergouwe Y, Royston P (2009) Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 338:b606. https://doi.org/10.1136/bmj.b606
16. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M (2012) Risk prediction models: II. External validation, model updating, and impact assessment. Heart 98(9):691–698. https://doi.org/10.1136/heartjnl-2011-301247
17. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, Clarke M, Devereaux PJ, Kleijnen J, Moher D (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 339. https://doi.org/10.1136/bmj.b2700
18. Colditz GA, Atwood KA, Emmons K, Monson RR, Willett WC, Trichopoulos D, Hunter DJ (2000) Harvard report on cancer prevention volume 4: Harvard Cancer Risk Index. Risk Index Working Group, Harvard Center for Cancer Prevention. Cancer Causes Control 11(6):477–488
19. Weiderpass E, Braaten T, Magnusson C, Kumle M, Vainio H, Lund E, Adami H-O (2004) A prospective study of body size in different periods of life and risk of premenopausal breast cancer. Cancer Epidemiol Biomark Prev 13(7):1121
20. Eliassen A, Colditz GA, Rosner B, Willett WC, Hankinson SE (2006) Adult weight change and risk of postmenopausal breast cancer. JAMA 296(2):193–201. https://doi.org/10.1001/jama.296.2.193
21. Folkerd E, Dowsett M (2013) Sex hormones and breast cancer risk and prognosis. Breast 22(Suppl 2):S38–43. https://doi.org/10.1016/j.breast.2013.07.007
22. Wright CE, Harvie M, Howell A, Evans DG, Hulbert-Williams N, Donnelly LS (2015) Beliefs about weight and breast cancer: an interview study with high risk women following a 12 month

weight loss intervention. Hereditary Cancer Clin Pract 13(1):1. https://doi.org/10.1186/s13053-014-0023-9

23. Park B, Ma SH, Shin A, Chang M-C, Choi J-Y, Kim S, Han W, Noh D-Y, Ahn S-H, Kang D (2013) Korean risk assessment model for breast cancer risk prediction. PLoS ONE 8(10):e76736
24. Novotny J, Pecen L, Petruzelka L, Svobodnik A, Dusek L, Danes J, Skovajsova M (2006) Breast cancer risk assessment in the Czech female population—an adjustment of the original Gail model. Breast Cancer Res Treat 95(1):29–35
25. Rosner B, Colditz GA (1996) Nurses' health study: log-incidence mathematical model of breast cancer incidence. J Natl Cancer Inst 88(6):359–364
26. Win AK, MacInnis RJ, Hopper JL, Jenkins MA (2012) Risk prediction models for colorectal cancer: a review. Cancer Epidemiol Biomark Prev 21(3):398–410
27. Engel C, Fischer C (2015) Breast cancer risks and risk prediction models. Breast Care 10(1):7–12
28. Meads C, Ahmed I, Riley R (2012) A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. Breast Cancer Res Treat 132(2):365–377. https://doi.org/10.1007/s10549-011-1818-2
29. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R (2008) Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol 56(1):45–50
30. Emmons KM, Wong MEI, Puleo E, Weinstein N, Fletcher R, Colditz G (2004) Tailored computer-based cancer risk communication: correcting colorectal cancer risk perception. J Health Commun 9(2):127–141. https://doi.org/10.1080/10810730490425295
31. Pfeiffer RM, Park Y, Kreimer AR, Lacey JV Jr, Pee D, Greenlee RT, Buys SS, Hollenbeck A, Rosner B, Gail MH (2013) 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 10(7):e1001492
32. Gail MH, Costantino JP, Pee D, Bondy M, Newman L, Selvan M, Anderson GL, Malone KE, Marchbanks PA, McCaskill-Stevens W (2007) Projecting individualized absolute invasive breast cancer risk in African American women. J Natl Cancer Inst 99(23):1782–1792
33. Lee C, Lee JC, Park B, Bae J, Lim MH, Kang D, Yoo K-Y, Park SK, Kim Y, Kim S (2015) Computational discrimination of breast cancer for Korean women based on epidemiologic data only. J Korean Med Sci 30(8):1025–1034. https://doi.org/10.3346/jkms.2015.30.8.1025
34. Karen ME, Susan K-W, Kathy A, Lisa C, Rima R, Graham C (1999) A qualitative evaluation of the harvard cancer risk index. J Health Commun 4(3):181–193. https://doi.org/10.1080/108107399126904
35. Bandura A (1992) Exercise of personal agency through the self-efficacy mechanism. In: Self-efficacy: thought control of action. Hemisphere Publishing Corp, Washington, DC, pp 3–38
36. Lee EO, Ahn SH, You C, Lee DS, Han W, Choe KJ, Noh D-Y (2004) 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 Nurs 27(5):400–406
37. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81(24):1879–1886
38. Ueda K, Tsukuma H, Tanaka H, Ajiki W, Oshima A (2003) Estimation of individualized probabilities of developing breast cancer for Japanese women. Breast Cancer 10(1):54–62
39. Boyle P, Mezzetti M, La Vecchia C, Franceschi S, Decarli A, Robertson C (2004) Contribution of three components to individual cancer risk predicting breast cancer risk in Italy. Eur J Cancer Prev 13(3):183–191
40. Banegas MP, Gail MH, LaCroix A, Thompson B, Martinez ME, Wactawski-Wende J, John EM, Hubbell FA, Yasmeen S, Katki HA (2012) Evaluating breast cancer risk projections for Hispanic women. Breast Cancer Res Treat 132(1):347–353
41. Matsuno RK, Costantino JP, Ziegler RG, Anderson GL, Li H, Pee D, Gail MH (2011) Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. J Natl Cancer Inst 103(12):951–961. https://doi.org/10.1093/jnci/djr154
42. Rosner B, Colditz GA, Willett WC (1994) Reproductive risk factors in a prospective study of breast cancer: the Nurses' Health Study. Am J Epidemiol 139(8):819–835
43. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu LM, Moons KG, Altman DG (2014) External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 14:40. https://doi.org/10.1186/1471-2288-14-40
44. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P (2004) Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 159(9):882–890. https://doi.org/10.1093/aje/kwh101
45. Rosner B, Glynn RJ, Tamimi RM, Chen WY, Colditz GA, Willett WC, Hankinson SE (2013) Breast cancer risk prediction with heterogeneous risk profiles according to breast cancer tumor markers. Am J Epidemiol 178(2):296–308. https://doi.org/10.1093/aje/kws457
46. Colditz GA, Rosner BA, Chen WY, Holmes MD, Hankinson SE (2004) Risk factors for breast cancer according to estrogen and progesterone receptor status. J Natl Cancer Inst 96(3):218–228
47. Cook NR (2007) Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 115(7):928–935. https://doi.org/10.1161/circulationaha.106.672402
48. Laughlin GA, Barrett-Connor E, Kritz-Silverstein D, von Mühlen D (2000) Hysterectomy, oophorectomy, and endogenous sex hormone levels in older women: the rancho Bernardo study. J Clin Endocrinol Metab 85(2):645–651. https://doi.org/10.1210/jcem.85.2.6405
49. Kim DJ, Rockhill B, Colditz GA (2004) Validation of the Harvard Cancer Risk Index: a prediction tool for individual cancer risk. J Clin Epidemiol 57(4):332–340. https://doi.org/10.1016/j.jclinepi.2003.08.013
50. Colditz GA, Rosner B (2000) Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study. Am J Epidemiol 152(10):950–964
51. Matsuno RK, Costantino JP, Ziegler RG, Anderson GL, Li H, Pee D, Gail MH (2011) Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. J Natl Cancer Inst 103:951–961
52. Emmons KM, Koch-Weser S, Atwood K, Conboy L, Rudd R, Colditz G (1999) A qualitative evaluation of the Harvard Cancer Risk Index. J Health Commun 4(3):181–193. https://doi.org/10.1080/108107399126904
53. Pfeiffer RM, Park Y, Kreimer AR, Lacey JV, Pee D, Greenlee RT, Buys SS, Hollenbeck A, Rosner B, Gail MH, Hartge P (2013) 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 10(7):e1001492
54. Timmers JM, Verbeek AL, IntHout J, Pijnappel RM, Broeders MJ, den Heeten GJ (2013) Breast cancer risk prediction model: a nomogram based on common mammographic screening findings. Eur Radiol 23(9):2413–2419. https://doi.org/10.1007/s00330-013-2836-8
55. McCowan C, Donnan PT, Dewar J, Thompson A, Fahey T (2011) Identifying suspected breast cancer: development and validation

of a clinical prediction rule. Brit J Gen Pract 61 (586). https://doi.org/10.3399/bjgp11X572391

56. Cook NR, Rosner BA, Hankinson SE, Colditz GA (2009) Mammographic screening and risk factors for breast cancer. Am J Epidemiol 170(11):1422–1432. https://doi.org/10.1093/aje/kwp304
57. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K (2008) Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. Ann Intern Med 148(5):337–375. https://doi.org/10.7326/0003-4819-148-5-200803040-00004
58. Rosner B, Colditz GA, Iglehart JD, Hankinson SE (2008) Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the Nurses' Health Study. Breast Cancer Res 10(4):R55. https://doi.org/10.1186/bcr2110
59. Lee SM, Park JH, Park HJ (2008) Implications of systematic review for breast cancer prediction. Cancer Nurs 31(5):E40–E46. https://doi.org/10.1097/01.NCC.0000305765.34851.e9
60. Chlebowski RT, Anderson GL, Lane DS, Aragaki AK, Rohan T, Yasmeen S, Sarto G, Rosenberg CA, Hubbell FA, Women's Health Initiative I (2007) Predicting risk of breast cancer in postmenopausal women by hormone receptor status. J Natl Cancer Inst 99(22):1695–1705. https://doi.org/10.1093/jnci/djm224
61. Decarli A, Calza S, Masala G, Specchia C, Palli D, Gail MH (2006) Gail model for prediction of absolute risk of invasive breast cancer: Independent evaluation in the Florence-European Prospective Investigation Into Cancer and Nutrition Cohort. J Natl Cancer I 98(23):1686–1693. https://doi.org/10.1093/jnci/djj463
62. Chen JB, Pee D, Ayyagari R, Graubard B, Schairer C, Byrne C, Benichou J, Gail MH (2006) Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. J Natl Cancer I 98(17):1215–1226. https://doi.org/10.1093/jnci/djj332
63. Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA, Tice JA, Buist DS, Geller BM, Rosenberg R, Yankaskas BC, Kerlikowske K (2006) Prospective breast cancer risk prediction model for women undergoing screening mammography. J Natl Cancer Inst 98(17):1204–1214. https://doi.org/10.1093/jnci/djj331
64. Tice JA, Miike R, Adduci K, Petrakis NL, King E, Wrensch MR (2005) Nipple aspirate fluid cytology and the Gail model for breast cancer risk assessment in a screening population. Cancer Epidemiol Biomark Prev 14(2):324–328. https://doi.org/10.1158/1055-9965.EPI-04-0289
65. Taplin SH, Thompson RS, Schnitzer F, Anderman C, Immanuel V (1990) Revisions in the Risk-Based Breast-Cancer Screening-Program at Group Health Cooperative. Cancer 66 (4):812–818. https://doi.org/10.1002/1097-0142(19900815)66:4%3C812::Aid-Cncr2820660436%3E3.0.Co;2-1
66. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ (1989) Projecting individualized probabilities of developing breast-cancer for white females who are being examined annually. J Natl Cancer Inst 81(24):1879–1886. https://doi.org/10.1093/jnci/81.24.1879
67. Anderson DE, Badzioch M (1984) Risk of familial breast-cancer. Lancet 1(8373):392–392
68. Ottman R, Pike MC, King MC, Henderson BE (1983) Practical guide for estimating risk for familial breast-cancer. Lancet 2(8349):556–558
69. Lee AJ, Cunningham AP, Kuchenbaecker KB, Mavaddat N, Easton DF, Antoniou AC, Modifiers CI, Consortium BCA (2014) BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. Brit J Cancer 110(2):535–545. https://doi.org/10.1038/bjc.2013.730
70. McCarthy AM, Armstrong K, Handorf E, Boghossian L, Jones M, Chen JB, Demeter MB, McGuire E, Conant EF, Domchek SM (2013) Incremental impact of breast cancer SNP panel on risk classification in a screening population of white and African American women. Breast Cancer Res Treat 138(3):889–898. https://doi.org/10.1007/s10549-013-2471-8
71. Dite GS, Mahmoodi M, Bickerstaffe A, Hammet F, Macinnis RJ, Tsimiklis H, Dowty JG, Apicella C, Phillips KA, Giles GG, Southey MC, Hopper JL (2013) Using SNP genotypes to improve the discrimination of a simple breast cancer risk prediction model. Breast Cancer Res Treat 139(3):887–896. https://doi.org/10.1007/s10549-013-2610-2
72. Biswas S, Atienza P, Chipman J, Hughes K, Barrera AMG, Amos CI, Arun B, Parmigiani G (2013) Simplifying clinical use of the genetic risk prediction model BRCAPRO. Breast Cancer Res Treat 139(2):571–579. https://doi.org/10.1007/s10549-013-2564-4
73. Sueta A, Ito H, Kawase T, Hirose K, Hosono S, Yatabe Y, Tajima K, Tanaka H, Iwata H, Iwase H, Matsuo K (2012) A genetic risk predictor for breast cancer using a combination of low-penetrance polymorphisms in a Japanese population. Breast Cancer Res Treat 132(2):711–721. https://doi.org/10.1007/s10549-011-1904-5
74. Huesing A, Canzian F, Beckmann L, Garcia-Closas M, Diver WR, Thun MJ, Berg CD, Hoover RN, Ziegler RG, Figueroa JD, Isaacs C, Olsen A, Viallon V, Boeing H, Masala G, Trichopoulos D, Peeters PHM, Lund E, Ardanaz E, Khaw KT, Lenner P, Kolonel LN, Stram DO, Le Marchand L, McCarty CA, Buring JE, Lee IM, Zhang SM, Lindstrom S, Hankinson SE, Riboli E, Hunter DJ, Henderson BE, Chanock SJ, Haiman CA, Kraft P, Kaaks R, Bpc3 (2012) Prediction of breast cancer risk by genetic risk factors, overall and by hormone receptor status. J Med Genet 49(9):601–608. https://doi.org/10.1136/jmedgenet-2011-100716
75. Darabi H, Czene K, Zhao WT, Liu JJ, Hall P, Humphreys K (2012) Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement. Breast Cancer Res 14(1):R25. https://doi.org/10.1186/bcr3110
76. Dai JC, Hu ZB, Jiang Y, Shen H, Dong J, Ma HX, Shen HB (2012) Breast cancer risk assessment with five independent genetic variants and two risk factors in Chinese women. Breast Cancer Res 14(1):R17. https://doi.org/10.1186/bcr3101
77. Biswas S, Tankhiwale N, Blackford A, Barrera AM, Ready K, Lu K, Amos CI, Parmigiani G, Arun B (2012) Assessing the added value of breast tumor markers in genetic risk prediction model BRCAPRO. Breast Cancer Res Treat 133(1):347–355. https://doi.org/10.1007/s10549-012-1958-z
78. van Zitteren M, van der Net JB, Kundu S, Freedman AN, van Duijn CM, Janssens ACJW (2011) Genome-based prediction of breast cancer risk in the general population: a modeling study based on meta-analyses of genetic associations. Cancer Epidem Biomark 20(1):9–22. https://doi.org/10.1158/1055-9965.Epi-10-0329
79. Crooke PS, Justenhoven C, Brauch H, Dawling S, Roodi N, Higginbotham KSP, Plummer WD, Schuyler PA, Sanders ME, Page DL, Smith JR, Dupont WD, Parl FF, Consortium G (2011) Estrogen metabolism and exposure in a genotypic-phenotypic model for breast cancer risk prediction. Cancer Epidemiol Biomark 20(7):1502–1515. https://doi.org/10.1158/1055-9965.Epi-11-0060
80. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, Thun MJ, Cox DG, Hankinson SE, Kraft P, Rosner B, Berg CD, Brinton LA, Lissowska J, Sherman ME, Chlebowski R, Kooperberg C, Jackson RD, Buckman DW, Hui P, Pfeiffer R, Jacobs KB, Thomas GD, Hoover RN, Gail MH, Chanock SJ, Hunter DJ (2010) Performance of common

genetic variants in breast-cancer risk models. New Engl J Med 362(11):986–993. https://doi.org/10.1056/NEJMoa0907727

81. Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, Risch HA, Eyfjord JE, Hopper JL, Southey MC, Olsson H, Johannsson O, Borg A, Passini B, Radice P, Manoukian S, Eccles DM, Tang N, Olah E, Anton-Culver H, Warner E, Lubinski J, Gronwald J, Gorski B, Tryggvadottir L, Syrjakoski K, Kallioniemi OP, Eerola H, Nevanlinna H, Pharoah PDP, Easton DF (2008) The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. Brit J Cancer 98(8):1457–1466. https://doi.org/10.1038/sj.bjc.6604305
82. Tyrer J, Duffy SW, Cuzick J (2004) A breast cancer prediction model incorporating familial and personal risk factors. Stat Med 23(7):1111–1130. https://doi.org/10.1002/sim.1668
83. Evans DGR, Eccles DM, Rahman N, Young K, Bulman M, Amir E, Shenton A, Howell A, Lalloo F (2004) A new scoring system for the chances of identifying a BRCA1/2 mutation outperforms existing models including BRCAPRO. J Med Genet 41(6):474–480. https://doi.org/10.1136/jmg.2003.017996
84. Antoniou AC, Pharoah PPD, Smith P, Easton DF (2004) The BOADICEA model of genetic susceptibility to breast and ovarian cancer. Brit J Cancer 91(8):1580–1590. https://doi.org/10.1038/sj.bjc.6602175
85. Jonker MA, Jacobi CE, Hoogendoorn WE, Nagelkerke NJD, de Bock GH, van Houwelingen JC (2003) Modeling familial clustered breast cancer using published data. Cancer Epidemiol Biomark 12(12):1479–1485
86. Fisher TJ, Kirk J, Hopper JL, Godding R, Burgemeister FC (2003) A simple tool for identifying unaffected women at a moderately increased or potentially high risk of breast cancer based on their family history. Breast 12(2):120–127. https://doi.org/10.1016/S0960-9776(02)00285-0
87. Apicella C, Andrews L, Hodgson SV, Fisher SA, Lewis CM, Solomon E, Tucker K, Friedlander M, Bankier A, Southey MC, Venter DJ, Hopper JL (2003) Log odds of carrying an Ancestral Mutation in BRCA1 or BRCA2 for a defined personal and family history in an Ashkenazi Jewish woman (LAMBDA). Breast Cancer Res 5(6):R206–R216. https://doi.org/10.1186/bcr644
88. Frank TS, Deffenbaugh AM, Reid JE, Hulick M, Ward BE, Lingenfelter B, Gumpper KL, Scholl T, Tavtigian SV, Pruss DR, Critchfield GC (2002) Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: Analysis of 10,000 individuals. J Clin Oncol 20(6):1480–1490. https://doi.org/10.1200/Jco.20.6.1480
89. de la Hoya M, Osorio A, Godino J, Sulleiro S, Tosar A, Perez-Segura P, Fernandez C, Rodriguez R, Diaz-Rubio E, Benitez J, Devilee P, Caldes T (2002) Association between BRCA1 and BRCA2 mutations and cancer phenotype in Spanish breast/ovarian cancer families: Implications for genetic testing. Int J Cancer 97(4):466–471. https://doi.org/10.1002/ijc.1627
90. Berry DA, Iversen ES, Gudbjartsson DF, Hiller EH, Garber JE, Peshkin BN, Lerman C, Watson P, Lynch HT, Hilsenbeck SG, Rubinstein WS, Hughes KS, Parmigiani G (2002) BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. J Clin Oncol 20(11):2701–2712. https://doi.org/10.1200/Jco.2002.05.121
91. Antoniou AC, Pharoah PDP, McMullan G, Day NE, Stratton MR, Peto J, Ponder BJ, Easton DF (2002) A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. Brit J Cancer 86(1):76–83. https://doi.org/10.1038/sj/bjc/6600008
92. Vahteristo P, Eerola H, Tamminen A, Blomqvist C, Nevanlinna H (2001) A probability model for predicting BRCA1 and BRCA2 mutations in breast and breast-ovarian cancer families. Brit J Cancer 84(5):704–708. https://doi.org/10.1054/bjoc.2000.1626
93. Gilpin CA, Carson N, Hunter AGW (2000) A preliminary validation of a family history assessment form to select women at risk for breast or ovarian cancer for referral to a genetics center. Clin Genet 58(4):299–308. https://doi.org/10.1034/j.1399-0004.2000.580408.x
94. Hartge P, Struewing JP, Wacholder S, Brody LC, Tucker MA (1999) The prevalence of common BRCA1 and BRCA2 mutations among Ashkenazi Jews. Am J Hum Genet 64(4):963–970. https://doi.org/10.1086/302320
95. Parmigiani G, Berry DA, Aguilar O (1998) Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. Am J Hum Genet 62(1):145–158. https://doi.org/10.1086/301670
96. Frank TS, Manley SA, Olopade OI, Cummings S, Garber JE, Bernhardt B, Antman K, Russo D, Wood ME, Mullineau L, Isaacs C, Peshkin B, Buys S, Venne V, Rowley PT, Loader S, Offit K, Robson M, Hampel H, Brener D, Winer EP, Clark S, Weber B, Strong LC, Rieger P, McClure M, Ward BE, Shattuck-Eidens D, Oliphant A, Skolnick MH, Thomas A (1998) Sequence analysis of BRCA1 and BRCA2: correlation of mutations with family history and ovarian cancer risk. J Clin Oncol 16(7):2417–2425. https://doi.org/10.1200/Jco.1998.16.7.2417
97. Shattuck-Eidens D, Oliphant A, Fau-McClure M, McClure M, Fau-McBride C, McBride C, Fau-Gupte J, Gupte J, Fau-Rubano T, Rubano T, Fau-Pruss D, Pruss D, Fau-Tavtigian SV, Tavtigian SV, Fau-Teng DH, Teng DH, Fau-Adey N, Adey N, Fau-Staebell M, Staebell M, Fau-Gumpper K, Gumpper K, Fau-Lundstrom R, Lundstrom R, Fau-Hulick M, Hulick M, Fau-Kelly M, Kelly M, Fau-Holmen J, Holmen J, Fau-Lingenfelter B, Lingenfelter B, Fau-Manley S, Manley S, Fau-Fujimura F, Fujimura F, Fau-Luce M, Luce M, Fau-Ward B, Ward B, Fau-Cannon-Albright L, Cannon-Albright L, Fau-Steele L, Steele L, Fau-Offit K, Offit K, Fau-Thomas A, Thomas A et al (1997) BRCA1 sequence analysis in women at high risk for susceptibility mutations. Risk factor analysis and implications for genetic testing. J Am M Assoc (0098-7484 (Print))
98. Couch FJ, DeShano ML, Blackwood MA, Calzone K, Stopfer J, Campeau L, Ganguly A, Rebbeck T, Weber BL, Jablon L, Cobleigh MA, Hoskins K, Garber JE (1997) BRCA1 mutations in women attending clinics that evaluate the risk of breast cancer. New Engl J Med 336(20):1409–1415. https://doi.org/10.1056/Nejm199705153362002
99. Claus EB, Risch N, Thompson WD (1994) Autosomal-dominant inheritance of early-onset breast-cancer—implications for risk prediction. Cancer 73 (3):643–651. https://doi.org/10.1002/1097-0142(19940201)73:3%3C643::Aid-Cncr2820730323%3E3.0.Co;2-5
100. Claus EB, Risch N, Thompson WD (1993) The calculation of breast-cancer risk for women with a first degree family history of ovarian-cancer. Breast Cancer Res Treat 28(2):115–120. https://doi.org/10.1007/Bf00666424
101. Wang S, Ogundiran T, Ademola A, Olayiwola OA, Adeoye A, Adeniji-Sofoluwe A, Morhason-Bello I, Odedina S, Agwai I, Adebamowo C, Obajimi M, Ojengbede O, Olopade OI, Huo D (2016) Abstract 2590: development and validation of a breast cancer risk prediction model for black women: findings from the Nigerian breast cancer study. Cancer Res. https://doi.org/10.1158/1538-7445.AM2016-2590

# Terms and Conditions

RESEARCH ARTICLE

# Risk of breast cancer in the UK biobank female cohort and its relationship to anthropometric and reproductive factors

**Kawthar Al-Ajmi, Artitaya Lophatananon, William Ollier, Kenneth R. Muir***

Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, United Kingdom

* kenneth.muir@manchester.ac.uk

OPEN ACCESS

**Citation:** Al-Ajmi K, Lophatananon A, Ollier W, Muir KR (2018) Risk of breast cancer in the UK biobank female cohort and its relationship to anthropometric and reproductive factors. PLoS ONE 13(7): e0201097. https://doi.org/10.1371/journal.pone.0201097

**Editor:** David Meyre, McMaster University, CANADA

**Received:** March 29, 2018

**Accepted:** July 9, 2018

**Published:** July 26, 2018

**Data Availability Statement:** Upon publishing this article, we have fulfilled our proposed work agreement with the UK Biobank and have returned our data that we used for these analyses to the UKBiobank as part of the agreement. However, the data from the UK Biobank (www.ukbiobank.ac.uk) are third party and their legal agreement means that we do not have permission to share the data. The UK Biobank data used in this study can however be accessed by applying through the UK Biobank Access Management System (www.ukbiobank.ac.uk/register-apply).

## Abstract

### Background

Anthropometric and reproductive factors have been reported as being established risk factors for breast cancer (BC). This study explores the contribution of anthropometric and reproductive factors in UK females developing BC in a large longitudinal cohort.

### Methods

Data from the UK Biobank prospective study of 273,467 UK females were analyzed. Relative risks (RRs) and 95% confidence intervals (CIs) for each factor were adjusted for age, family history of BC and deprivation score. The analyses were stratified by the menopausal status.

### Results

Over the 9 years of follow up the total number of BC cases were 14,231 with 3,378 (23.7%) incident cases with an incidence rate of 2.09 per 1000 person-years. In pre-menopausal, increase in age, height, having low BMI, low waist to hip ratio, first degree family history of BC, early menarche age, nulliparous, late age at first live birth, high reproductive interval index, and long contraceptive use duration were all significantly associated with an increased BC risk. In post-menopausal, getting older, being taller, having high BMI, first degree BC family history, nulliparous, late age at first live birth, and high reproductive interval index were all significantly associated with an increased risk of BC. The population attributable fraction (PAF) suggested that an early first live birth, lower reproductive interval index and increased number of children can contribute to BC risk reduction up to 50%.

### Conclusions

This study utilizes the UK Biobank study to confirm associations between anthropometric and reproductive factors and the risk of breast cancer development. Result of attributable fraction of risk contributed by each risk factor suggested that lifetime risk of BC can be reduced by controlling weight, reassessing individual approaches to the timing of childbirth

**Funding:** Kenneth R. Muir and Artitaya Lophatananon received funding through the ICEP supported by CRUK [grant number C18281/A19169]). We also would like to thank Ministry of Higher Education in Oman for their financial support to the first author (Kawthar Al-Ajmi). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

**Abbreviations:** RR, Relative risk; BMI, Body mass index; BC, Breast cancer; CI, Confidence interval; EPIC, The European Prospective Investigation into Cancer and Nutrition; HRT, Hormonal replacement therapy; IGF-1, Insulin-like growth factor 1; ICD, The International Classification of Diseases; IQR, Interquartile range; LCL, Lower confidence limit; NHS, National Health Service; OC, Oral contraceptive; PAF, Population attributable fractions; SD, Standard deviation; UCL, Upper confidence limit; WHR, Waist to hip ratio.

and options for contraception and considering early screening for women with family history in the first degree relative.

## Introduction

Breast cancer is the most common cancer in females, globally accounting for 23% of all new female cancers [1–4]. In the UK, BC accounts for 15% of all newly diagnosed cancer cases in the population regardless of gender [5]. Global variations in BC incidences arise mainly from the availability of early detection and treatment facilities; however other factors may also affect this variation. Factors such as population structure (age, ethnicity, and race), life expectancy, environment, lifestyle, prevalence of risk factors, health insurance status, availability of new treatments, and pathology can enhance this variation [6]. Several risk factors have been reported in the literature. Reproductive risk factors including, early age at menarche, late menopause age, late age at first birth, low parity, hormonal replacement therapy usage, contraceptive use, hysterectomy and bilateral oophorectomy have all been identified as conferring risk for developing BC [7, 8]. Another major factor for increasing BC incidences is the accumulated effect of anthropometric factors. Increased height, weight, hip circumference, waist circumference, body mass index (BMI), and waist to hip ratio (WHR) have been reported as increasing BC risk depending on the menopausal status of women [9]. Given the unique opportunity the UK biobank [2] project offers for assessing a wide range of disease risk factors in a large longitudinal cohort, we have measured the effect of anthropometric and reproductive factors on BC risk. This study is the first study to explore the relationships of risk factors and breast cancer in the UK Biobank initiative. This landmark national cohort provides an important dataset based on half a million UK residents. The recruitment was undertaken and 22 regional centers to seek distributed population coverage across the UK. The cohort also has broad-scale genotyping performed which will allow further investigations of the possible combined effects of the genetic and the epidemiological risk factors reported in this paper.

## Materials and methods

### Study population and study design

UK Biobank is a national-based health project that aims to improve the diagnosis, treatment, and prevention of diseases such as cancers, diabetes, stroke, heart disease, osteoporosis, arthritis, eye diseases, dementia and depression [2]. A total of 502,650 participants aged between 39 to 71 years were enrolled in the study between 2006 and 2010 and they continue to be clinically followed up. Details can be found at http://www.ukbiobank.ac.uk/. In addition to the collection of biological samples (blood, saliva and urine), health, demographic and anthropometric data were collected in 22 UK assessment facilities across England, Wales and Scotland. Detailed physical / physiological measurements were further supported by the administration of questionnaires and eye examination. Many participants completed additional detailed questionnaires on work history, diet, and cognitive function. Anonymized data are now available to researchers across the world [2, 3]. Our study acquired data on the female cohort (273,467 female participants) from UK Biobank. The UK Biobank female cohort had a mean follow up time of 6.9 years (at 2016). Data on exposures were defined prior to the development of BC in cases or prior to the first assessment date in controls.

**Defining breast cancer cases and controls.** BC was defined as a malignant neoplasm of the breast. The UK Biobank database contained record of all cancers including their subtype occurring either before or after participant enrollment using the International Classification of Diseases (ICD10, ICD9) and their self-reported data. Details of codes used to identify BC cases are summarized in S1 Table.

**Breast cancer cases.** In the database, each participant had 9 follow-up time point records for ICD10, 11 follow-up time point records for ICD9 and 9 follow-up time point records for self-reported status of cancer. The case-control groups were identified by utilizing all these three data sources. The codes for BC are presented in S1 Table. Cases were characterized as incident or prevalent using 'age or date when they attended the center' and 'age when first reported BC cancer'. With cases defined by ICD10 and ICD9, if their 'attending age' was greater than 'cancer diagnosis age' then this was considered as a prevalent case. Subjects were considered to be incident cases if their 'attending age' was less than their 'cancer diagnosis age'. For self-reported cases, the same criteria were applied. Age when first attended the assessment center was compared with the interpolated age of the participant when cancer was first diagnosed. To combine and classify the type of cases from 3 different sources, we applied the following criteria:

1. If the BC cases appeared as being incident using any of these three identification methods then the cases were deemed to be incident cases.

2. Prevalent cases were defined using combination of rules a) only if the participant has been identified as a prevalent case by any of the three methods and b) none of these methods define the same participant as being an incident case.

In total, there were 14,231 BC cases with 3,378 being incident cases and 10,853 prevalent cases.

**Controls.** Female participants were defined as controls if they had no record of cancer, *in-situ* carcinoma or an undefined neoplasm (232,476 controls).

**Exclusion criteria.** In the case group, we excluded 10,853 (3.97%) prevalent BC cases. In the control group, participants were excluded due to following reasons; other type of cancers (23,540), breast *in situ* carcinoma (636), other *in situ* carcinomas (2,463) and unknown neoplasm (121).

**Exposures.** Reproductive variables included menarche age, menopause age, menopausal status, parity (yes/no), number of children, age at first live birth, pregnancy history, pregnancy termination and number of terminations, reproductive interval index (difference between menarche age and age at first birth), history of oral contraceptive (OC) use and its duration, and history of hormonal replacement therapy (HRT) use and its duration. Anthropometric variables included BMI, waist to hip ratio (WHR) and height (sitting and standing).

## Statistical analysis

To assess associations between exposures and BC risk in the cohort, we computed relative risk (RR) and 95% confident intervals (95% C.I.) using a binomial generalised linear regression model. Regression analyses were performed for each independent variable and were adjusted for age, family history of BC in first degree relatives, and deprivation score. The independent variables list and description are presented in S2 Table.

All analyses were stratified by menopausal status: pre- and post-menopausal. The criteria for pre-menopausal were females aged $\leq$ 55 years old (according to the NHS the menopause age in the UK is between 40 to 55 years [10]) who reported that they still had periods and did not report a history of hysterectomy or bilateral oophorectomy, and menarche age $\geq$ 7 years old (the menarche age in the UK ranges from 7 to 20 years [11]). Post-menopausal females were defined as those who reported no longer having periods and did not report a history of hysterectomy or bilateral oophorectomy and their menopause age $\geq$ 40 years old. These criteria were employed to minimise inclusion of both pre-mature and the medically induced pre- or post-menopausal women. After further application of criteria, 61,903 participants were in pre-menopause group and 133,704 participants were in post-menopause group.

To compute BC incidence within the cohort, we used the STATA *stptime* command to obtain the overall person-time of observation and disease incidence rate. To calculate time for each participant, we subtracted the endpoint (either the date of cancer diagnosis or the end of the follow-up—January 1st, 2016) with the date of study enrolment. Incidence rates were estimated for the whole cohort and pre- and post-menopausal separately. Moreover, population attributable fractions (PAF) were calculated using the *punaf* command [12] where the fraction was estimated compared to whole cohort and compared to the most significant subgroup associated with the BC. This was done to estimate how much risk could be eliminated by controlling that risk factor in both groups.

All statistical analysis was performed using STATA MP 14.1 software for Windows [13]. Results with 95% confident intervals not including 1 were considered as being statistically significant.

## Results

The UK biobank female cohort consisted of 273,476 female participants with a mean age of 56.3 years (SD ±8.00). The follow up time was 9.8 years up to January 2016 where the database was frozen for this analysis. The total number of BC cases was 14,231 with 3,378 (1.24%) incident cases and 10,853 (3.97%) prevalent cases. The total number of controls was 232,476 (85.01%). The remaining participants were either females with other cancer 23,540 (8.61%) or with breast *in situ* carcinoma 636 (0.23%), or other *in situ* carcinoma 2,463 (0.90%) or unknown neoplasm 121 (0.04%). A total of 3,162 (93.60%) of incident cases were identified by ICD10 and the rest 216 (6.40%) were identified by self-reporting. All the BC cases identified by ICD9 were solely prevalent cases. When further applying criteria for menopause status, the total number of pre-menopausal females was 61,903 (31.65%) and post-menopausal was 133,704 (68.35%). Out of the total pre-menopausal females, 618 (1.07%) were incident cases and 57,089 (98.93%) were controls. For post-menopausal females, 1,757 (1.53%) were incident cases and 112,757 (98.47%) were controls (Fig 1). The BC incidence rate of the whole cohort was 2.09 per 1000 person-years. The pre-menopause BC incidence rate was 1.55 per 1000 person-years and the post-menopause BC incidence rate was 2.24 per 1000 person-years. The incidence rate ratio between the pre- and post-menopausal females is 1.45 with 95% CI 1.32–1.59.

Comparisons of mean values of age, deprivation score, anthropometric and reproductive variables (all continuous variables) of the participants conditioned on the menopausal status are summarised in Table 1. In both pre- and post-menopause groups, cases were older than controls and the mean age differences were statistically significant (*Student's t-test p-values*$<$*0.05*). Results using the Townsend deprivation score showed that case's mean score were significantly lower than control mean score in both pre- and post-menopause females (*Student's t-test p-values* $< 0.05$).

For anthropometric variables, in the pre-menopausal group, the mean values of standing and sitting height in cases were higher as compared to controls (*Student's t-test p-values*$<$*0.05*). On the other hand, mean values of BMI, waist circumference and waist to hip ratio were significantly lower in cases as compared with controls (*Student's t-test p-values*$<$*0.05*). In the post-menopause case group, the mean values of standing and sitting height, BMI, waist circumference, and hip circumferences were higher when compared with controls (*Student's t-test p-values*$<$*0.05*).

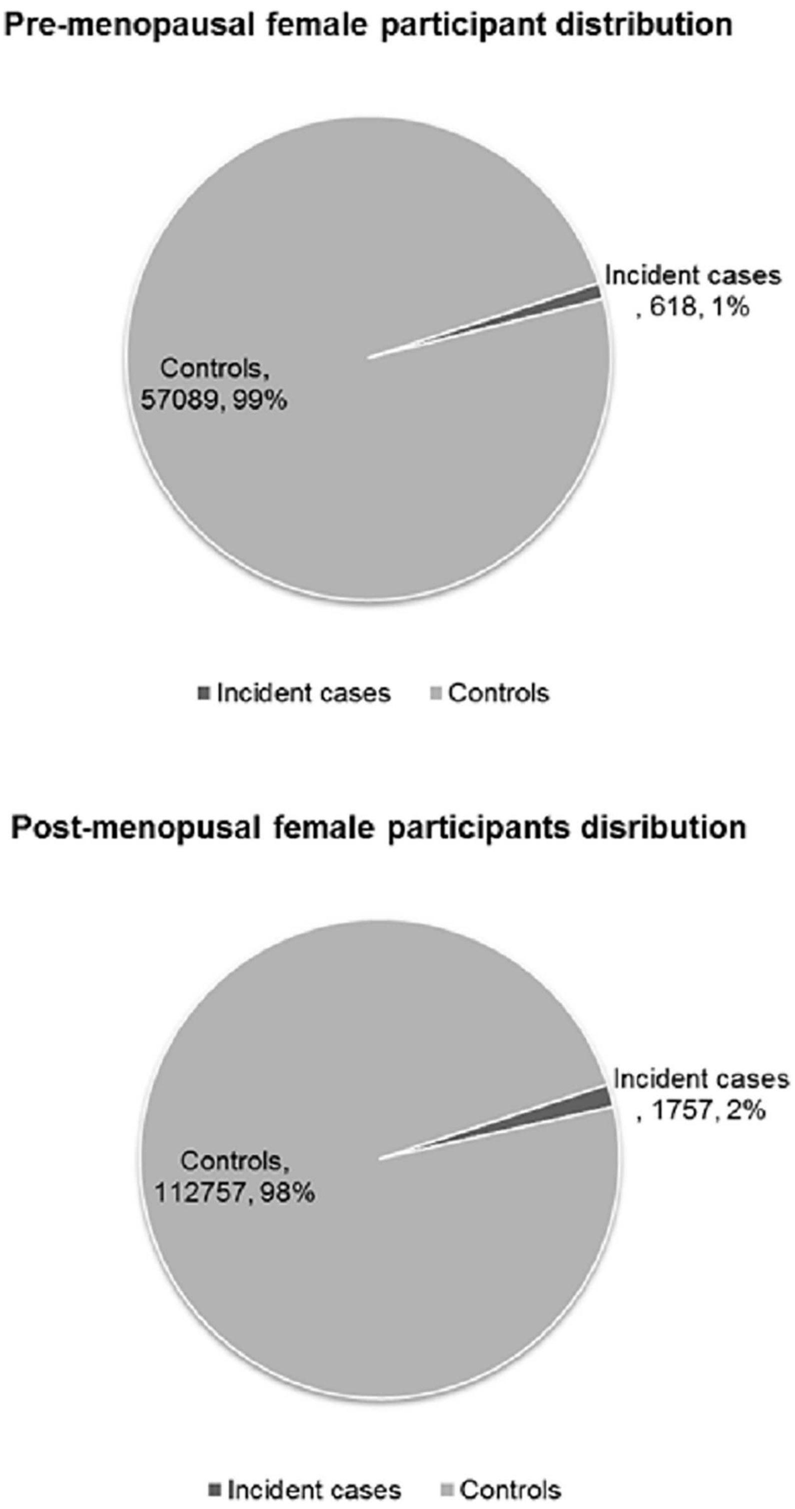


**Fig 1. UK biobank data distribution based on menopausal status.**

https://doi.org/10.1371/journal.pone.0201097.g001

Analysis of reproductive factors in pre-menopause case group, showed higher mean values of age at first birth, reproductive interval index, and contraceptive use duration as compared

**Table 1. Mean comparisons between cases and controls in pre- and post-menopause status.**

| | **Pre-menopausal** | | | | **Post-menopausal** | | | |
|---|---|---|---|---|---|---|---|---|
| **Variables** | **No. (cases/controls)** | **Case's mean** | **Control's mean** | **P-value*** | **No. (cases/controls)** | **Case's mean** | **Control's mean** | **P-value*** |
| Age (year) | (618/ 57,089) | 46.43 | 45.83 | <0.001 | (1,757 /112,757) | 60.67 | 59.76 | <0.001 |
| Deprivation score | (618/56,999) | -1.49 | -1.09 | 0.001 | (1,755 /112,639) | -1.72 | -1.48 | 0.006 |
| Body shape measures | | | | | | | | |
| BMI (kg/m2) | (612/ 56,847) | 25.95 | 26.43 | 0.026 | (1,750/112,270) | 27.45 | 27.01 | <0.001 |
| Waist Circumference (cm) | (613 /56,890) | 80.97 | 82.23 | 0.012 | (1,752/112,426) | 86.03 | 84.78 | <0.001 |
| Hip Circumference (cm) | (613 /56,889) | 102.16 | 102.51 | 0.408 | (1,752/112,423) | 104.32 | 103.12 | <0.001 |
| Waist to Hip ratio | (613 /56,883) | 0.79 | 0.80 | <0.001 | (1,752/112,416) | 0.82 | 0.82 | 0.114 |
| Standing Height (cm) | (612 /56,896) | 164.70 | 164.04 | 0.011 | (1,751/112,391) | 162.61 | 161.91 | <0.001 |
| Sitting height (cm) | (603 /56,406) | 87.86 | 87.54 | 0.031 | (1,724/111,654) | 86.36 | 86.03 | <0.001 |
| Reproductive factors measures | | | | | | | | |
| Menarche age (year) | (605 /55,286) | 12.95 | 13.05 | 0.105 | (1,727/110,214) | 12.93 | 12.98 | 0. 178 |
| Menopause age (year) | N/A | | | | (1,757/112,757) | 50.85 | 50.58 | 0.007 |
| Number of live births | (618 /57,053) | 1.49 | 1.57 | 0.095 | (1,754/112,685) | 1.77 | 1.88 | <0.001 |
| Age at first birth (year) | (336 /33,071) | 27.70 | 27.03 | 0.015 | (1,171/79,421) | 25.46 | 25.30 | 0.231 |
| Number of Pregnancy termination | (221 / 20,149) | 0.61 | 0.69 | 0.127 | (529/34,166) | 0.47 | 0.52 | 0.140 |
| Reproductive interval index (year) | (521/47,237) | 14.66 | 13.93 | 0.011 | (1,483 /96,718) | 12.50 | 12.29 | 0.131 |
| Contraceptive use duration (year) | (519/ 50,012 | 11.62 | 9.99 | <0.001 | (1,610/ 102,760) | 7.51 | 7.68 | 0.386 |
| HRT duration (year) | (609/56,210) | 0.05 | 0.03 | 0.200 | (1,553/ 102,786) | 2.25 | 1.92 | <0.001 |
| Total | 618 / 57,089 | | | | 1,757 / 112,757 | | | |

*Student's t-test*

https://doi.org/10.1371/journal.pone.0201097.t001

with controls (*Student's t-test p values <0.05*). In addition, among the post-menopausal group, mean values of menopause age and duration of HRT use were significantly higher in cases compared with controls. In contrast, mean values of number of live births were lower in cases as compared to controls in post-menopausal females.

Relative risks (RRs) of the key characteristics and anthropometric measures of pre- and post-menopausal females are illustrated in Table 2. For both pre-and post-menopausal females, age as a continuous variable showed a slight increased risk of developing BC (RR = 1.05, 95%CI; 1.02–1.07) and RR = 1.03, 95%CI; 1.02–1.04, respectively). Results of Townsend deprivation score showed a decreased risk of BC associated with increased deprivation score (more deprived) among both pre- (RR = 0.96, 95%CI; 0.94–0.99) and post-menopausal (RR = 0.97, 95%CI; 0.96–0.99) females.

Family history of BC is a well-defined risk factor for BC. The strength of this risk factor varies according to the number and relationship of the affected family members. Females who reported having had a family history of BC were at increased risk for developing BC in both pre- and post-menopausal females with (RR = 1.77, 95%CI; 1.43–2.19) and (RR = 1.58, 95%CI; 1.40–1.79), respectively. Both pre- and post-menopause subjects with their siblings affected with BC were at increased risk of 82% (pre-menopause) and 61% (post-menopause) respectively. Similar results were also seen in subjects who reported only their mother affected with BC with increased risk of 72% in pre- and 57% in post-menopausal women. All of these significant associations were stronger among pre-menopausal compared to post-menopausal women. In the post-menopause group, subjects with both mother and sibling affected with BC

**Table 2. Relative risk of key characteristics and anthropometric factors in pre- and post- menopausal females.**

| Menopausal status | Pre-menopausal | | | | | Post-menopausal | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variables | Number of cases/controls | RR | P-value | LCL | UCL | Number of cases/controls | RR | P-value | LCL | UCL |
| Age in years (Continuous) * | 618/ 57,089 | 1.046 | <0.001 | 1.024 | 1.069 | 1,757 /112,757 | 1.033 | <0.001 | 1.024 | 1.042 |
| Deprivation score (Continuous) ** | 618/56,999 | 0.962 | 0.004 | 0.937 | 0.988 | 1,755 /112,639 | 0.973 | 0.001 | 0.957 | 0.990 |
| Family history *** | | | | | | | | | | |
| No | 520/ 51,547 | Ref | | | | 1,458/ 99,998 | Ref | | | |
| Yes | 97/ 5,326 | 1.770 | <0.001 | 1.427 | 2.194 | 290/ 12,367 | 1.582 | <0.001 | 1.397 | 1.792 |
| Mother BC history *** | | | | | | | | | | |
| No mother BC history | 532/ 51,750 | Ref | | | | 1,529/ 102,184 | Ref | | | |
| Mother BC history | 78/ 4,360 | 1.724 | <0.001 | 1.362 | 2.181 | 192/ 8,145 | 1.569 | <0.001 | 1.353 | 1.820 |
| Sibling BC history *** | | | | | | | | | | |
| No sibling BC history | 579/ 54,125 | Ref | | | | 1,553 / 103,570 | Ref | | | |
| Sibling BC history | 23/ 1,108 | 1.823 | 0.004 | 1.206 | 2.756 | 120/ 4,782 | 1.613 | <0.001 | 1.343 | 1.938 |
| Family history- Combined*** | | | | | | | | | | |
| No family history at all | 520/ 51,547 | Ref | | | | 1,458/ 99,998 | Ref | | | |
| Mother or Sister BC history | 93/ 5,184 | 1.756 | <0.001 | 1.408 | 2.190 | 268/11,807 | 1.540 | <0.001 | 1.351 | 1.754 |
| Mother and Sister BC history | 4/142 | 2.592 | 0.054 | 0.982 | 6.837 | 22/560 | 2.594 | <0.001 | 1.717 | 3.920 |
| BMI in kg/m$^2$ (Continuous) | 612/ 56,847 | 0.983 | 0.041 | 0.968 | 0.999 | 1,750/112,270 | 1.018 | <0.001 | 1.009 | 1.027 |
| BMI–categorical | | | | | | | | | | |
| BMI—Healthy (18.5–24.9) | 326/26,983 | Ref | | | | 626/44,215 | Ref | | | |
| BMI—Overweight (25–29.9) | 186/18,319 | 0.839 | 0.055 | 0.701 | 1.004 | 681/42,624 | 1.102 | 0.078 | 0.989 | 1.228 |
| BMI—Obese (> = 30) | 100/11,545 | 0.733 | 0.007 | 0.586 | 0.918 | 443/25,431 | 1.241 | 0.001 | 1.098 | 1.401 |
| Waist to Hip (Continuous) | 613 /56,883 | 0.131 | 0.001 | 0.038 | 0.446 | 1,752/112,416 | 1.520 | 0.226 | 0.772 | 2.994 |
| Waist to Hip–categorical | | | | | | | | | | |
| Waist to Hip—Low (< = 0.80) | 362/30,170 | Ref | | | | 678/45,184 | Ref | | | |
| Waist to Hip—Moderate (0.81–0.85) | 139/13,993 | 0.829 | 0.060 | 0.682 | 1.008 | 475/30,741 | 1.010 | 0.869 | 0.898 | 1.135 |
| Waist to Hip—High (>0.85) | 112/12,720 | 0.744 | 0.006 | 0.602 | 0.920 | 599/36,491 | 1.073 | 0.213 | 0.961 | 1.198 |
| Sitting Height in cm (Continuous) | 603 /56,406 | 1.023 | 0.041 | 1.001 | 1.046 | 1,724/111,654 | 1.032 | <0.001 | 1.019 | 1.046 |
| Standing Height in cm (Continuous) | 612 /56,896 | 1.017 | 0.010 | 1.004 | 1.030 | 1,751/112,391 | 1.021 | <0.001 | 1.013 | 1.029 |
| Standing Height in cm–categorical | | | | | | | | | | |
| Below mean ± SD (150.20–156.06 cm) | 57/6,447 | Ref | | | | 285/21,259 | Ref | | | |
| Within mean ± SD (159.21–165.71 cm) | 388/37,137 | 1.181 | 0.243 | 0.893 | 1.562 | 1,153/ 75,173 | 1.168 | 0.019 | 1.025 | 1.330 |
| Above mean ± SD (169.02–175.00 cm) | 167/13,314 | 1.429 | 0.021 | 1.057 | 1.933 | 313/ 15,964 | 1.533 | <0.001 | 1.305 | 1.802 |

All adjusted for age + Family history of BC + deprivation score

*adjusted for deprivation score only

** no adjustment

***Adjusted for age + deprivation score

https://doi.org/10.1371/journal.pone.0201097.t002

were almost at three-fold increase BC risk (RR = 2.59, 95%CI; 1.72–3.92). Despite a similar relative risk estimate, no association was reported in pre-menopause group.

For anthropometric exposures treated as being continuous variables, increasing BMI (RR = 0.98, 95%CI; 0.97–1.00), and waist to hip ratio (RR = 0.13, 95%CI; 0.04–0.45) were associated with reduced BC risk among the pre-menopause group. The WHR as a categorical variable (low as reference group, moderate and high) showed significant risk reduction only in the high WHR group (RR = 0.74 with 95%CI; 0.60–0.92). BMI as a categorical variable showed that obese women with a BMI ≥30 had 26.7% decreased BC risk compared to women with normal range BMI. For height, per 1 cm of increased height (cm), BC risk was increased by

2%. Height as a categorical variable showed that women in the tallest group (height ranges from 168.8 to 199 cm) had their BC risk increased by 43% compared to shorter females with height ranges from 152.20 to 156.06 cm.

In post-menopausal women, increasing BMI, standing height and sitting height were associated with a slight increased risk of BC of 2%, 2% and 3%, respectively. BMI as a categorical variable showed that obese subjects had 24.1% increased risk for BC (RR = 1.24, 95%CI; 1.10–1.40) when compared to the normal BMI group. For height treated as a categorical variable, results suggested that the tallest group (height ranges from 168.8 to 199 cm, mean = 172.0) were at 53% increased risk of BC (RR = 1.53, 95%CI; 1.31–1.80) when compared to the reference group (height ranges from 100 to 156 cm, mean = 153.1).

## Reproductive factors and breast cancer

RRs for the reproductive factors and BC risk are presented in Table 3. For the pre-menopause group, menarche age as continuous variable showed a slight risk reduction (RR = 0.95, 95%CI; 0.90–1.00). When menarche age was grouped into >13 years old (as a reference group) versus ≤13 years old, a moderate increased risk was observed (RR = 1.23, 95% CI; 1.04–1.45). For the post-menopause group, age at menarche did not show any significant association with BC risk (confidence interval value included 1).

Parous women were at reduced BC risk in both pre- (RR = 0.76, 95% CI; 0.64–0.91) and post-menopausal women (RR = 0.82, 95% CI; 0.73–0.93) when compared to nulliparous women. The 'number of children' when treated as a continuous variable showed moderate decreased BC risk (pre-menopause group RR = 0.93, 95% CI; 0.86–0.99 and post-menopause group (RR = 0.90, 95% CI; 0.86–0.94). In contrast, increasing maternal age at live birth showed very slight increased BC risk in both pre- (2%) and post-menopausal women (1%). Further analysis was carried out in parous women to explore the association of age at live birth and BC risk. Age at first live birth as categorical variable (< 20 years old as the reference group, 20–24, 25–29, and ≥30 years old) showed that among pre-menopausal females, BC risk was almost double when they reported having had their first child at age ≥30 years old and at age 25–29 years as compared to women who reported having their first baby at age <20 years old (RR 1.94; 95% CI, 1.06–3.54 and RR = 1.88 with 95% CI; 1.04–3.42, respectively). This effect was not seen in post-menopausal females (all 95% CI values included 1). Both pregnancy termination history (ever versus none) and number of terminations were not significantly associated with BC development in both pre- and post-menopausal females (all 95% CI values included 1).

The Reproductive Interval Index (the difference between age at first child and the age of menarche) based on the interquartile range of the control group (low as reference group, moderate, high, and no children) only showed statistically significant increased risk in 'high' (RR = 1.42, 95% CI; 1.10–1.84) and 'no children' groups (RR = 1.53, 95% CI; 1.21–1.94) in pre-menopausal females. In post-menopausal group, only females reporting no children showed an increased risk of BC (RR = 1.33, 95% CI; 1.16–1.53) when compared to the low index group.

History of oral contraceptive (OC) pills used showed no association with BC risk in both pre- and post-menopause groups. Within the OC use group, however, OC duration showed a slight increased BC risk in pre-menopause women of 2% but not in post-menopausal women. Hormone replacement therapy (HRT) was not associated with risk of BC in pre-menopause UK females. In the post-menopause group, women who reported using HRT were at moderate significant increased risk (RR = 1.14, 95%CI; 1.04–1.26).

Women in both pre- and post-menopausal groups who reported having had mammograms were at increased risk of BC of 19% and 26%, respectively.

**Table 3. Relative risks of the reproductive factors based on the menopausal status.**

| Menopausal status | | Pre-menopausal | | | | | Post-menopausal | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variables | No. cases/controls | RR | P-value | LCL | UCL | No. cases/controls | RR | P-value | LCL | UCL |
| Menarche age in years (Continuous)* | 605 /55,286 | 0.948 | 0.042 | 0.900 | 0.998 | 1,727/110,214) | 0.987 | 0.388 | 0.958 | 1.017 |
| Menarche age–categorical | | | | | | | | | | |
| Menarche age (>13) | 198/ 20,785 | Ref | | | | 625 /40,534 | Ref | | | |
| Menarche age (≤13) | 407/ 34,501 | 1.228 | 0.017 | 1.037 | 1.454 | 1,102/69,680 | 1.029 | 0.569 | 0.933 | 1.134 |
| Menopause age in years (Continuous)* | Not applicable | | | | | 1,757/112,757 | 1.006 | 0.284 | 0.995 | 1.018 |
| Parity | | | | | | | | | | |
| No | 188/15,024 | Ref | | | | 326/18,855 | Ref | | | |
| Yes | 430/42,029 | 0.764 | 0.002 | 0.643 | 0.908 | 1,428/93,830 | 0.821 | 0.001 | 0.728 | 0.926 |
| Number of births (Continuous) | 618 /57,053 | 0.925 | 0.024 | 0.864 | 0.990 | 1,754/112,685 | 0.899 | <0.001 | 0.863 | 0.937 |
| First live birth age in years (Continuous) | 336 /33,071 | 1.022 | 0.055 | 1.000 | 1.045 | 1,171/79,421 | 1.010 | 0.142 | 0.997 | 1.023 |
| First live birth age–categorical | | | | | | | | | | |
| First live birth age (<20) | 12/2,422 | Ref | | | | 97/7,330 | Ref | | | |
| First live birth age (20–24) | 74/7,873 | 1.719 | 0.082 | 0.933 | 3.168 | 369/27,992 | 0.966 | 0.763 | 0.773 | 1.207 |
| First live birth age (25–29) | 138/12,625 | 1.882 | 0.038 | 1.036 | 3.417 | 492/31,181 | 1.091 | 0.435 | 0.876 | 1.360 |
| First live birth age (≥30) | 112/10,151 | 1.938 | 0.031 | 1.062 | 3.539 | 186/12,918 | 1.055 | 0.669 | 0.825 | 1.350 |
| pregnancy termination | | | | | | | | | | |
| No | 117/9,544 | Ref | | | | 321/ 19,771 | Ref | | | |
| Yes | 104/10,605 | 0.835 | 0.181 | 0.641 | 1.088 | 208/14,395 | 0.981 | 0.834 | 0.823 | 1.171 |
| Pregnancy termination number (Continuous) | 221 / 20,149 | 0.898 | 0.232 | 0.753 | 1.071 | 529/34,166) | 0.973 | 0.673 | 0.858 | 1.104 |
| Reproductive interval index in years (Continuous) | 521/47,237 | 1.003 | 0.002 | 1.001 | 1.005 | 1,483 /96,718 | 1.003 | <0.001 | 1.001 | 1.004 |
| Reproductive interval index–categorical | | | | | | | | | | |
| Low index (≤12) | 109/12,673 | Ref | | | | 585/41,334 | Ref | | | |
| Moderate index (12.01–16) | 98/9,499 | 1.146 | 0.329 | 0.872 | 1.506 | 359/22,601 | 1.128 | 0.073 | 0.989 | 1.287 |
| High index (>16.01) | 126/10,041 | 1.421 | 0.008 | 1.098 | 1.838 | 213/13,928 | 1.130 | 0.128 | 0.965 | 1.323 |
| No children | 188/15,024 | 1.530 | <0.001 | 1.208 | 1.937 | 326/18,855 | 1.333 | <0.001 | 1.163 | 1.528 |
| Contraceptive use | | | | | | | | | | |
| No | 53/ 6,297 | Ref | | | | 366/23,896 | Ref | | | |
| Yes | 565/50,646 | 1.261 | 0.106 | 0.952 | 1.670 | 1,389/88,638 | 1.124 | 0.053 | 0.998 | 1.265 |
| Contraceptive duration in years (Continuous) | 519/ 50,012 | 1.024 | <0.001 | 1.013 | 1.034 | 1,610/ 102,760 | 1.003 | 0.319 | 0.997 | 1.010 |
| HRT use | | | | | | | | | | |
| No | 599/ 55,336 | Ref | | | | 943 /65,669 | Ref | | | |
| Yes | 18/1,565 | 0.945 | 0.813 | 0.590 | 1.513 | 811/46,830 | 1.141 | 0.006 | 1.038 | 1.255 |
| HRT duration in years (Continuous) | 609/56,210 | 1.063 | 0.298 | 0.947 | 1.193 | 1,553/ 102,786 | 1.013 | 0.054 | 1.000 | 1.025 |
| Mammogram history | | | | | | | | | | |
| No | 359 /37,546 | Ref | | | | 50/5,408 | Ref | | | |
| Yes | 285/19,341 | 1.190 | 0.054 | 0.997 | 1.420 | 1,706/107,289 | 1.260 | 0.120 | 0.942 | 1.686 |

All adjusted for age, family history of BC and deprivation score

* adjusted more for BMI.

https://doi.org/10.1371/journal.pone.0201097.t003

PAF were calculated for the modifiable risk factors only based on the menopause status (Table 4). Two fractions were estimated; the PAF among the studied population and the PAF among the sub-population (the exposed significant group) to evaluate how many cases could be avoided if a particular factor was eliminated. Among pre-menopausal females these modifiable factors were the strongest in reducing the BC risk. Giving birth at age <30 can eliminate about 44.6% of the BC cases in general population, and about 48.4% among females who had

**Table 4. Population attributable fraction (PAF) among modifiable breast cancer risk factors according to the menopausal status.**

| Variables | Pre-menopausal | | Post-menopausal | |
|---|---|---|---|---|
| | PAF in population | PAF in subpopulation group | PAF in population | PAF in subpopulation group |
| BMI | | | | |
| BMI—Healthy (18.5–24.9) | Ref | | | |
| BMI—Obese (> = 30) | -0.091 | -0.707 | 0.083 | 0.194 |
| Waist to Hip ratio | | | | |
| Waist to Hip—Low (< = 0.80) | Ref | | | |
| Waist to Hip—High (>0.85) | -0.080 | -0.662 | *NS* | *NS* |
| Parity (Yes/No) | | | | |
| Yes | Ref | | | |
| No | 0.072 | 0.092 | 0.033 | 0.179 |
| Number of births | | | | |
| None | Ref | | | |
| More than one child | 0.088 | 0.247 | 0.046 | 0.211 |
| First live birth age | | | | |
| First live birth age (<20) | Ref | | | |
| First live birth age (25–29) | *NS* | 0.469 | *NS* | *NS* |
| First live birth age (≥30) | 0.446 | 0.484 | *NS* | *NS* |
| Reproductive interval index | | | | |
| Low index (≤12) | Ref | | | |
| High index (>16.01) | 0.149 | 0.296 | *NS* | *NS* |
| No children | 0.223 | 0.346 | 0.089 | 0.250 |
| HRT use (No /Yes) | | | | |
| No | Ref | | | |
| Yes | *NS* | *NS* | 0.058 | 0.125 |

https://doi.org/10.1371/journal.pone.0201097.t004

first children at age ≥30years old and about 46.9% of cases among females who had first children at age 25–29. Followed by low reproductive interval index with about 34.6% of BC cases can be eliminated among null-parous females and about 29.6% of BC cases can be eliminated among females with high index (>16.01). Being parous can eliminate only 9.2% of the cases without taking into consideration the number of children they gave birth to. Finally, having BMI ≥30 and WTH >0.85 can eliminate 70% and 66.2% of the cases among pre-menopausal women, respectively.

Among post-menopausal women; reducing BMI <30 can eliminate 8.3% among general population and 19.4% among obese females; being parous can eliminate 17.9% among null parous females; having more than one child can eliminate 21.1% % among females with <1 child; not using HRT can eliminate 12.5% of cancer cases among users.

The most effective preventative factors identified were giving birth at earlier age, having more than one child, reducing the reproductive interval index, and reducing weight.

A summary for the significant factors associated with development of BC among UK females is presented in S3 Table.

## Discussion

This study explores the effect of anthropometric and reproductive factors on risk of developing BC in the UK Biobank female cohort. The BC incidence rate in the pre-menopause group was 1.55 per 1000 person-years and 2.24 per 1000 person-years in the post-menopause group. McPherson *et al* reported a similar finding that in every 1000 UK women over 50 years old, two females will be diagnosed with BC [14] which suggests that UK biobank is a representative cohort of the UK female population.

Findings from previous studies suggested that differences in risk factors and incidences of BC were based on the menopausal status [4, 5, 15]. Some of the risk factors were common across pre- and post-menopause groups while other factors showed different effects. We therefore stratified all the analyses by menopausal status.

## Age

For both pre- and post-menopausal groups, age is associated with increasing risk of developing BC. Age is a well-established risk factor for BC [16]. BC incidence increases with age during the reproductive years by the double in every 10 years up until the menopause [5, 15]. A potential explanation could be cells becoming more susceptible to environmental carcinogens and modification in the biological ageing which stimulates or allows tumour growth and metastasis [17].

## Family history

Family history of BC is also a well-established risk factor. Our findings suggested that females with a first degree relative (sibling or mother) affected with BC were at high risk of developing BC. Regardless of menopause status, the estimated risks were higher in females who reported only their sibling(s) affected with BC as compared to females who reported only their mother affected with BC. The estimated risks were even higher when both mother and sister were affected with BC. Evidence of family history of BC in the first degree relatives and BC risk has been well documented by many studies with different study designs [14, 18]. The variation of reported estimated risks was due to family history nature such as affected age, number and type of the affected family members [19, 20]. It is known that BRCA1 and BRCA2 gene mutations are responsible for this strong association for cases diagnosed at young age [21, 22]. The stronger effect of family history among pre-menopausal females in this study suggested a component of familial BC [20]. Possible explanations to higher estimated risks observed in subjects with sibling affected include recall bias. With self-reported data, maternal history is more likely to be incomplete as compared to the sibling history. Another possibility is the confounder effect such as parity; mothers of subjects were obviously parous while sisters could be either parous or nulliparous. It is known that parity is a protective factor against BC hence if subject's sisters were null-parous; one would expect to observe higher risk. Sisters are more likely to share the same or similar environmental factors than mother and a daughter. Finally, multiple family relatives having an early onset or bilateral cancer increases the risk even more [15].

## Deprivation score

Deprivation score data was available for the dataset. Our result suggested that the most deprived females appeared to have lower BC risk compared to least deprived females in the UK Biobank cohort. Our cohort appeared to be mainly from least deprived districts like Bristol (8.8%), Leeds (8.9%), Newcastle (7.4%), and Nottingham (6.8%). Most deprived districts included Stockport (0.76%), Manchester (2.7%), and Birmingham (4.9) contributed less in this cohort. This sampling distribution could have an effect on the association direction between deprivation and BC.

## Variables related to body size

Inverse associations were observed with BMI and waist to hip ratio in the pre-menopausal group. While among post-menopausal females, increased risks were reported. A Norwegian prospective study suggested a decreased risk of BC among overweight and obese females who had no family history of BC. Nevertheless once a female has a family history, that protection effect disappeared in both overweight and obese pre-menopausal females [23]. A meta-analysis conducted in 2012 showed no significant effect of BMI on the incidence of pre-menopausal BC [24]. Our results however suggested that risk was reduced even when family history of BC was present among pre-menopausal females. One study reported an estimation of 3% risk increase in BC for every 1 kg/m$^2$ in post-menopausal females [25], while another study reported that weight gains of 5–12 kg increases the post-menopausal BC risk by 50% and modest weight loss (5–10%) can decrease BC risk by 25–40% [26]. Furthermore, overweight and obesity are associated with poor prognosis and increased BC mortality [27]. BMI is a modifiable factor and can contribute to reduce the BC risk by 10.0% in pre- and 5.1% in post menopause women [28]. Our study confirmed a BC risk reduction of 8.3% if females reduced their BMI lower than 30 among general population but if obese females (BMI≥30) reduced their BMI to normal BMI range, a 19.4% of BC risk will be eliminated among post-menopaused females. Another way to assess central adiposity among individuals is by measuring WHR (waist to hip ratio). A systematic review on the relationship of WHR and BC concluded that 24% risk reduction was associated with small WHR in post-menopausal females. In contrast among pre-menopausal the effect was very little [29]. Another review suggested the same conclusion; pre-menopausal BC is not associated with WHR however, 1.4 to 5.4 times of BC risk was proven among post-menopausal females [30]. Our study showed BC risk reduction was associated with increased WHR up to 25.6% in pre-menopausal females but failed to prove any association with post-menopausal females. The findings on height and BC risk supported adult height being associated with BC risk in both pre- and post-menopausal groups. The EPIC cohort study [31] reported a positive association between height and post-menopausal BC (RR 1.10 with 95% CI 1.05–1.16). Furthermore, a meta-analysis of 159 prospective studies showed a pooled BC RR of 1.17 (95% CI = 1.15–1.19) per 10cm increase in height [32, 33]. Another pooled analysis also suggested positive association among post-menopausal females (RR = 1.07 with 95% CI: 1.03, 1.12) [34]. No association was reported in pre-menopausal females (RR 1.02 with 95% CI: 0.96, 1.10). Not all prospective studies confirmed the positive association. A register-based cohort study with 13,572 participants concluded no statistical evidence of association between height and BC risk [35]. Evidence from case-control studies was inconsistent. Our study showed an increased risk of 18% per 10cm increase in height among pre-menopausal and 23% per 10cm increase in height among post-menopausal. All the results mentioned previously were for standing height; we examined sitting height and found a BC risk association with sitting height. Taller sitting height is associated with 25.5% BC risk increase per each 10 cm increase in pre- and 37.0% in post-menopausal per 10 cm increase.

The relationship between height and BC suggests a protective effect among females with short stature rather than a continuous increased risk with the increasing of female's height. One possible explanation is that short females would be exposed to lower levels of insulin like growth factor 1 (IGF 1) throughout childhood and adolescence. IGF-1 is considered to be a strong mitogen for BC cells and IGF-1 receptors are expressed in breast tumour tissues 10 folds higher than normal breast tissues [36, 37].

## Reproductive factors

Our findings suggested protective effect of factors related to childbearing and having more children among pre- and post-menopausal females. Risk factors in pre-menopausal females were early menarche age (<13 years old), late age at first live birth (>25 years of age), high reproductive interval index, and increased duration of OC used were considered as risk factors for BC in pre-menopausal females. Factors such as nulliparous, high reproductive interval index and increased duration of OC used were risk factors in post-menopausal females.

Increased production of steroid hormone starts around the time of menarche and decreases significantly near the menopause [4]. Hormones produced by the ovary directly affect the breast function and development. Studies showed long period of hormonal exposure increases the risk to develop BC. Late menarche and early menopause are known to be protective factors as the period of hormonal exposure is reduced. Lengthening the reproductive years by an early menarche of one year has a stronger effect than delaying the menopause by one year [4]. The strength of menarche age and menopause age on BC development can be affected by BMI [38, 39]. The association between the BC and menopause age can be weaker among post-menopausal females with high BMI as seen in the meta-analysis [4]. Our results showed an evidence of BC risk reduction by late age of menarche but not by early age the menopause age as the previous studies even when BMI was adjusted for in the analysis. A meta-analysis of 120,000 BC cases and 300,000 controls done by a collaborative research group confirmed the existing association between early menarche and developing risk of BC. Extra risk is associated with lengthening female's reproductive years by one year during menarche rather than lengthening one year at menopause [4]. The RR associated with early menarche was 1.05 (95% CI 1.04–1.06) and the RR associated with late menopause was 1.03(95% CI 1.03–1.03) [4].

Childbearing in a known protective factor against BC although other factors might help confound this protection, such as breast feeding [40]. Combination of both factors can help protect females even more. Unfortunately there were no data available on breastfeeding in our cohort and unable to assess this effect. In the case of parity, our results showed a significant evidence of risk reduction among both pre- and post- menopausal females with a stronger effect among pre-menopausal. Likewise, as the number of children increases, the protective effect increases. Our results suggested an elimination of 9.2% among pre- and 17.9% among post- BC risk associated with being a parous female while other study reported a lower yet an affective risk reduction of 13.3% for the same factor [41]. As the number of children increases, the attributed risk reduction increases accordingly with reduction of 5.2% among pre- and 5.4% among post-menopausal females [28]. Nevertheless, our results suggested a higher reduction among pre- (8.8%) and a lower reduction percentage among post- menopausal women (4.6%).

Termination of pregnancy, whether induced or natural did not appear to affect the BC risk. Thus, younger age at childbirth is a protective factor against BC and this was observed among pre-menopausal females with p values $<0.05$. Studies showed early pregnancy causes permanent morphological changes to the breast and makes it more resistant to carcinogenic changes [7]. Our study supported the elimination of 44.6% of BC risk if females in general had their first child in their 20s rather than ≥30 years old among pre-menopaused females. This reduction can reach up to 48.4% among females who had their first child at age of ≥30 if they had their first child in their twenties. Furthermore we explored the reproductive interval variable (duration between the menarche and first child) and the results supported evidence reported in the literature that as the duration increases the risk also increases. Long term hormonal exposure has been confirmed to be a risk for BC [15]. Our study showed a BC reduction of 14.9% in pre-menopausal women if they have reproductive interval of $<$ 16 and this reduction can reach up to 29.6% if those females with reproductive interval of ≥16 had interval of 12 or less among pre-menopausal females.

Mammogram history suggested borderline significant increased risk in pre-menopausal women and no association in post-menopausal women. The mammogram itself *per se* is not a risk factor for BC but women who reported having had a mammogram were more likely to be diagnosed. Mammogram screening is proved to reduce the BC mortality by 29% among females aged between 50–69 years [42].

## Hormone use

Oral contraceptive use is known to be a risk factor of BC and this risk rises with longer duration of use [43]. It has been proposed that using OC can activate breast tumours which are already present. Oestrogen is recognized as enhancing tumour growth, and with OC and later HRT use these hormones promotes the tumour growth even more [43]. Our findings suggested a positive association between BC and OC duration amongst pre- menopausal females only. Moreover, HRT users showed 14.1% more risk for developing BC among our cohort. Extensive evidence showed an increase in BC incidence in current HRT users and that risk returns to normal soon after use terminates. Combined oestrogen-progesterone therapy revealed higher risk compared to oestrogen only preparations including results from the Women Health Initiative study (WHI). Recent results from WHI found both oestrogen only and combined formulations convey greater risk for BC if the females started the HRT in less than 5 years after the menopause compared to longer gap [38, 44–47]. The study also carried out further analysis of HRT. Their results showed attenuated BC risk among obese females which is driven by hormonal adiposity of the breast. Endogenous oestrogen rises with the increase of the BMI among HRT non-users which increases the breast adiposity [38]. Another major study carried out in the UK (Million Women Study) identified that BC risk is associated with current use of HRT and the risk is considerably greater among combined oestrogen- progesterone users than other types of HRT [48]. According to our analysis stopping HRT can reduce the risk by 5.8% and by 12.5% risk among HRT users. The Million Woman Study estimated this figure to be 4.6% [41] and a more recent study has put this figure higher at 14.5% [28].

In conclusion, we carried out an analysis to confirm risk and protective factors and BC risk in the UK Biobank female cohort. The findings suggest that protective factors in women included reducing BMI, waist to hip ratio, increasing the numbers of births, having birth at an early age, minimising the use of oral contraceptive and HRT and their durations. Most of our findings are in keeping with evidence reported from the other UK large cohort studies such as the One Million Women and EPIC studies. Evidence from this large study can be further used in translational research such as prevention programmes. Our study has some strengths and limitations. The strengths of this study are: large nation-wide prospective population-based cohort with a follow up time of 9 years and a sizable number of incident cases (UK Biobank). Furthermore, to our knowledge, this is the first study investigating the effect of the anthropometric and reproductive factors with BC risk among the UK Biobank female cohort. The results of this study can be used to inform BC prevention strategies and be used to educate the public and form a basis for building risk prediction models for BC for the UK population. Additionally reproductive interval index is a new measure and only reported by our study using UK data. Estimation of the general PAF and the PAF of the subgroups for BC in the UK Biobank female cohort is novel. The attributable risks calculated for the modifiable factors can be translated into action to reduce BC incidence.

One of the study limitations is that the UK Biobank cohort is not the best representation of UK female population. A recent study investigated the sociodemographic characteristics of the UK biobank participants compared to normal UK population [49] found an evidence of "healthy volunteer" selection bias among the participants. UK biobank participants tend to be healthier, more educated and living in less deprived areas. This effect is common with other volunteer cohorts. Nonetheless, to overcome this limitation and to produce more generalizable associations it is very essential to use large sample size with high internal validity [50, 51]. Our study used a decent sample size and confirmed the expected associations which are similar to the published literature.

Another possible limitation is the lack of information such as breastfeeding history, ovarian cancer family history, BC onset of the family members and BC subtype (PR+, ER+, HER2+, triple negative). Some of the risk factors may affect BC subtype differently [52]. Finally, small sample size in some of the associations such as family history of breast cancer. There were only4 observations among pre-menopaused with both mother and sister family history which can affect the strength of the findings.

## Supporting information

**S1 Table. Codes used to identify breast cancer cases and controls.**
(DOCX)

**S2 Table. Classification of the variables included in the analysis.**
(DOCX)

**S3 Table. Summary of the significant factors associated with breast cancer among both pre- and post-menopausal females in the UK.**
(DOCX)

## Acknowledgments

We would like to express our appreciation to UK biobank participants and staffs for providing the research community with this valuable source of data. This work used data from the application number 5791.

## Author Contributions

**Conceptualization:** Kawthar Al-Ajmi, Kenneth R. Muir.

**Formal analysis:** Kawthar Al-Ajmi.

**Investigation:** Kawthar Al-Ajmi, Artitaya Lophatananon.

**Methodology:** Kawthar Al-Ajmi, Artitaya Lophatananon.

**Supervision:** Artitaya Lophatananon.

**Validation:** Kawthar Al-Ajmi, Artitaya Lophatananon.

**Writing – original draft:** Kawthar Al-Ajmi, Artitaya Lophatananon.

**Writing – review & editing:** Kawthar Al-Ajmi, Artitaya Lophatananon, William Ollier.

## References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA: A Cancer Journal for Clinicians. 2011; 61(2):69–90. https://doi.org/10.3322/caac.20107 PMID: 21296855
2. About UK Biobank 2017. Available from: http://www.ukbiobank.ac.uk.
3. Hewitt J, Walters M, Padmanabhan S, Dawson J. Cohort profile of the UK Biobank: diagnosis and characteristics of cerebrovascular disease. BMJ Open. 2016; 6(3).
4. Collaborative Group on Hormonal Factors in Breast Cancer. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. The Lancet Oncology. 2012; 13(11):1141–51. https://doi.org/10.1016/S1470-2045(12)70425-4 PMID: 23084519

5. Cancer Research UK. Cancer Statistics: Breast Cancer Cancer Research UK, 2014 November. Report No.

6. Hortobagyi GN, de la Garza Salazar J, Pritchard K, Amadori D, Haidinger R, Hudis CA, et al. The global breast cancer burden: variations in epidemiology and survival. Clin Breast Cancer. 2005; 6(5):391–401. PMID: 16381622.

7. Horn J, Asvold Bo Fau—Opdahl S, Opdahl S Fau—Tretli S, Tretli S Fau—Vatten LJ, Vatten LJ. Reproductive factors and the risk of breast cancer in old age: a Norwegian cohort study. 2013;(1573–7217 (Electronic)). https://doi.org/10.1007/s10549-013-2531-0 PMID: 23605085

8. Kelsey JL, Gammon Md Fau—John EM, John EM. Reproductive factors and breast cancer. 1993; (0193-936X (Print)). doi: D—PIP: 084709.

9. Friedenreich CM. Review of anthropometric factors and breast cancer risk. European journal of cancer prevention. 2001; 10(1):15–32. PubMed PMID: 00008469-200102000-00003. PMID: 11263588

10. NHS. Menopause 2015. Available from: http://www.nhs.uk/conditions/menopause/Pages/Introduction.aspx.

11. Morris DH, Jones ME, Schoemaker MJ, Ashworth A, Swerdlow AJ. Determinants of age at menarche in the UK: analyses from the Breakthrough Generations Study. Br J Cancer. 2010; 103(11):1760–4. https://doi.org/10.1038/sj.bjc.6605978 PMID: 21045834

12. Newson R. PUNAF: Stata module to compute population attributable fractions for cohort studies: Boston College Department of Economics; 2012.

13. StataCorp. [cited 2017 20 Sep]. Available from: https://www.stata.com/statamp/.

14. McPherson K, Steel CM, Dixon JM. Breast cancer—epidemiology, risk factors, and genetics. BMJ: British Medical Journal. 2000; 321(7261):624–8. PubMed PMID: PMC1118507. PMID: 10977847

15. Hulka BS, Moorman PG. Breast cancer: hormones and other risk factors. Maturitas. 2001; 38(1):103–13. http://dx.doi.org/10.1016/S0378-5122(00)00196-1. PMID: 11311599

16. Winters S, Martin C, Murphy D, Shokar NK. Chapter One—Breast Cancer Epidemiology, Prevention, and Screening. In: Lakshmanaswamy R, editor. Progress in Molecular Biology and Translational Science. 151: Academic Press; 2017. p. 1–32. https://doi.org/10.1016/bs.pmbts.2017.07.002

17. Balducci L, Ershler WB. Cancer and ageing: a nexus at several levels. Nat Rev Cancer. 2005; 5 (8):655–62. https://doi.org/10.1038/nrc1675 PMID: 16056261

18. Petrisek A, Campbell S, Laliberte L. Family history of breast cancer. Impact on the disease experience. Cancer Pract. 2000; 8(3):135–42. PMID: 11898138.

19. Kelsey JL, Gammon MD. The epidemiology of breast cancer. CA Cancer J Clin. 1991; 41(3):146–65. Epub 1991/05/01. PMID: 1902137.

20. Pharoah PD, Day NE, Duffy S, Easton DF, Ponder BA. Family history and the risk of breast cancer: a systematic review and meta-analysis. International journal of cancer. 1997; 71(5):800–9. Epub 1997/05/29. PMID: 9180149.

21. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science (New York, NY). 1994; 266 (5182):66–71. Epub 1994/10/07. PMID: 7545954.

22. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature. 1995; 378(6559):789–92. Epub 1995/12/21. https://doi.org/10.1038/378789a0 PMID: 8524414.

23. Weiderpass E, Braaten T, Magnusson C, Kumle M, Vainio H, Lund E, et al. A Prospective Study of Body Size in Different Periods of Life and Risk of Premenopausal Breast Cancer. Cancer Epidemiology Biomarkers & Prevention. 2004; 13(7):1121–7.

24. Cheraghi Z, Poorolajal J, Hashem T, Esmailnasab N, Doosti Irani A. Effect of Body Mass Index on Breast Cancer during Premenopausal and Postmenopausal Periods: A Meta-Analysis. PLoS One. 2012; 7(12):e51446. https://doi.org/10.1371/journal.pone.0051446 PMID: 23236502

25. Folkerd E, Dowsett M. Sex hormones and breast cancer risk and prognosis. Breast (Edinburgh, Scotland). 2013; 22 Suppl 2:S38–43. Epub 2013/10/01. https://doi.org/10.1016/j.breast.2013.07.007 PMID: 24074790.

26. Wright CE, Harvie M, Howell A, Evans DG, Hulbert-Williams N, Donnelly LS. Beliefs about weight and breast cancer: an interview study with high risk women following a 12 month weight loss intervention. Hereditary Cancer in Clinical Practice. 2015; 13(1):1. https://doi.org/10.1186/s13053-014-0023-9 PMID: 25648828

27. Cleary MP, Grossmann ME. Obesity and Breast Cancer: The Estrogen Connection. Endocrinology. 2009; 150(6):2537–42. https://doi.org/10.1210/en.2009-0070 PMID: 19372199

28. Dartois L, Fagherazzi G, Baglietto L, Boutron-Ruault M-C, Delaloge S, Mesrine S, et al. Proportion of premenopausal and postmenopausal breast cancers attributable to known risk factors: Estimates from the E3N-EPIC cohort. International journal of cancer. 2016; 138(10):2415–27. https://doi.org/10.1002/ijc.29987 PMID: 26756677

29. Harvie M, Hooper L, Howell AH. Central obesity and breast cancer risk: a systematic review. Obes Rev. 2003; 4(3):157–73. PMID: 12916817.

30. Friedenreich CM. Review of anthropometric factors and breast cancer risk. European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP). 2001; 10(1):15–32. PMID: 11263588.

31. Lahmann PH, Hoffmann K, Allen N, van Gils CH, Khaw K- T, Tehard B, et al. Body size and breast cancer risk: Findings from the European prospective investigation into cancer and nutrition (EPIC). 2004; 111(5):771.

32. Zhang B, Shu X- O, Delahanty RJ, Zeng C, Michailidou K, Bolla MK, et al. Height and Breast Cancer Risk: Evidence From Prospective Studies and Mendelian Randomization. Journal of the National Cancer Institute. 2015; 107(11%U http://jnci.oxfordjournals.org/content/107/11/djv219.abstract).

33. Green J, Cairns BJ, Casabonne D, Wright FL, Reeves G, Beral V, et al. Height and cancer incidence in the Million Women Study: prospective cohort, and meta-analysis of prospective studies of height and total cancer risk. Lancet Oncol. 2011; 12(8):785–94. https://doi.org/10.1016/S1470-2045(11)70154-1 PMID: 21782509; PubMed Central PMCID: PMCPMC3148429.

34. van den Brandt PA, Spiegelman D, Yaun S-S, Adami H-O, Beeson L, Folsom AR, et al. Pooled Analysis of Prospective Cohort Studies on Height, Weight, and Breast Cancer Risk. American Journal of Epidemiology. 2000; 152(6%U http://aje.oxfordjournals.org/content/152/6/514.abstract):514-27.

35. Andersen Zj Fau—Baker JL, Baker Jl Fau—Bihrmann K, Bihrmann K Fau—Vejborg I, Vejborg I Fau—Sorensen TIA, Sorensen Ti Fau—Lynge E, Lynge E. Birth weight, childhood body mass index, and height in relation to mammographic density and breast cancer: a register-based cohort study. 2014; (1465-542X (Electronic)). doi: D—NLM: PMC3978910 EDAT- 2014/01/22 06:00 MHDA- 2015/04/17 06:00 CRDT- 2014/01/22 06:00 PHST- 2013/07/04 [received] PHST- 2014/01/06 [accepted] PHST- 2014/01/20 [aheadofprint] AID—bcr3596 [pii] AID— PST—epublish. https://doi.org/10.1186/bcr3596 PMID: 24443815

36. Papa V, Gliozzo B Fau—Clark GM, Clark Gm Fau—McGuire WL, McGuire Wl Fau—Moore D, Moore D Fau—Fujita-Yamaguchi Y, Fujita-Yamaguchi Y Fau—Vigneri R, et al. Insulin-like growth factor-I receptors are overexpressed and predict a low risk in human breast cancer. 1993;(0008–5472 (Print)). PMID: 8339284

37. Bates P, Fisher R Fau—Ward A, Ward A Fau—Richardson L, Richardson L Fau—Hill DJ, Hill Dj Fau—Graham CF, Graham CF. Mammary cancer in transgenic mice expressing insulin-like growth factor II (IGF-II). 1995;(0007–0920 (Print)). doi: D—NLM: PMC2033962 EDAT- 1995/11/01 MHDA- 1995/11/01 00:01 CRDT- 1995/11/01 00:00 PST—ppublish. PMID: 7577466

38. Key TJ, Appleby PN, Reeves GK, Roddam A, Dorgan JF, Longcope C, et al. Body mass index, serum sex hormones, and breast cancer risk in postmenopausal women. J Natl Cancer Inst. 2003; 95 (16):1218–26. PMID: 12928347.

39. Endogenous H, Breast Cancer Collaborative G, Key TJ, Appleby PN, Reeves GK, Roddam AW, et al. Circulating sex hormones and breast cancer risk factors in postmenopausal women: reanalysis of 13 studies. Br J Cancer. 2011; 105(5):709–22. https://doi.org/10.1038/bjc.2011.254 PMID: 21772329; PubMed Central PMCID: PMCPMC3188939.

40. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. The Lancet. 2002; 360(9328):187–95. http://dx.doi.org/10.1016/S0140-6736(02)09454-0.

41. Sprague BL, Trentham-Dietz A, Egan KM, Titus-Ernstoff L, Hampton JM, Newcomb PA. Proportion of invasive breast cancer attributable to risk factors modifiable after menopause. Am J Epidemiol. 2008; 168(4):404–11. https://doi.org/10.1093/aje/kwn143 PMID: 18552361; PubMed Central PMCID: PMCPMC2727276.

42. Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials. Lancet (London, England). 1993; 341 (8851):973–8. Epub 1993/04/17. PMID: 8096941.

43. Thorbjarnardottir T, Olafsdottir EJ, Valdimarsdottir UA, Olafsson O, Tryggvadottir L. Oral contraceptives, hormone replacement therapy and breast cancer risk: A cohort study of 16 928 women 48 years and older. Acta Oncologica. 2014; 53(6):752–8. https://doi.org/10.3109/0284186X.2013.878471 PMID: 24460068

**44.** Beral V, Reeves G, Bull D, Green J. Breast cancer risk in relation to the interval between menopause and starting hormone therapy. J Natl Cancer Inst. 2011; 103(4):296–305. Epub 2011/02/01. https://doi.org/10.1093/jnci/djq527 PMID: 21278356; PubMed Central PMCID: PMCPMC3039726.

**45.** Prentice RL, Manson JE, Langer RD, Anderson GL, Pettinger M, Jackson RD, et al. Benefits and Risks of Postmenopausal Hormone Therapy When It Is Initiated Soon After Menopause. American Journal of Epidemiology. 2009; 170(1):12–23. https://doi.org/10.1093/aje/kwp115 PMID: 19468079

**46.** Cancer IAfRo. Monograph on the Evaluation of Carcinogenic Risks to Humans. Combined Estrogen/Progestogen Contraceptives and Combined Estrogen/Progestogen Menopausal Therapy. France IARC Press. 2008; 91.

**47.** Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52 705 women with breast cancer and 108 411 women without breast cancer. The Lancet. 1997; 350(9084):1047–59.

**48.** Beral V, Million Women Study C. Breast cancer and hormone-replacement therapy in the Million Women Study. Lancet (London, England). 2003; 362(9382):419–27. PMID: 12927427.

**49.** Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. American Journal of Epidemiology. 2017; 186(9):1026–34. https://doi.org/10.1093/aje/kwx246 PMID: 28641372

**50.** Rothman KJ, Gallacher JEJ, Hatch EE. Why representativeness should be avoided. International Journal of Epidemiology. 2013; 42(4):1012–4. https://doi.org/10.1093/ije/dys223 PMID: 24062287

**51.** Ebrahim S, Davey Smith G. Commentary: Should we always deliberately be non-representative? International Journal of Epidemiology. 2013; 42(4):1022–6. https://doi.org/10.1093/ije/dyt105 PMID: 24062291

**52.** Yang XR, Sherman ME, Rimm DL, Lissowska J, Brinton LA, Peplonska B, et al. Differences in Risk Factors for Breast Cancer Molecular Subtypes in a Population-Based Study. Cancer Epidemiology Biomarkers &amp; Prevention. 2007; 16(3):439. https://doi.org/10.1371/journal.pone.0158887

JAMA Network Open

Original Investigation | Oncology

# Association of Nongenetic Factors With Breast Cancer Risk in Genetically Predisposed Groups of Women in the UK Biobank Cohort

Kawthar Al Ajmi, MSc; Artitaya Lophatananon, PhD; Krisztina Mekli, PhD; William Ollier, PhD; Kenneth R. Muir, PhD

## Abstract

**IMPORTANCE** The association between noninherited factors, including lifestyle factors, and the risk of breast cancer (BC) in women and the association between BC and genetic makeup are only partly characterized. A study using data on current genetic stratification may help in the characterization.

**OBJECTIVE** To examine the association between healthier lifestyle habits and BC risk in genetically predisposed groups.

**DESIGN, SETTING, AND PARTICIPANTS** Data from UK Biobank, a prospective cohort comprising 2728 patients with BC and 88 489 women without BC, were analyzed. The data set used for the analysis was closed on March 31, 2019. The analysis was restricted to postmenopausal white women. Classification of healthy lifestyle was based on Cancer Research UK guidance (healthy weight, regular exercise, no use of hormone replacement therapy for more than 5 years, no oral contraceptive use, and alcohol intake <3 times/wk). Three groups were established: favorable (≥4 healthy factors), intermediate (2-3 healthy factors), and unfavorable (≤1 healthy factor). The genetic contribution was estimated using the polygenic risk scores of 305 preselected single-nucleotide variations. Polygenic risk scores were categorized into 3 tertiles (low, intermediate, and high).

**MAIN OUTCOMES AND MEASURES** Cox proportional hazards regression was used to assess the hazard ratios (HRs) of the lifestyles and polygenic risk scores associated with a malignant neoplasm of the breast.

**RESULTS** Mean (SD) age of the 2728 women with BC was 60.1 (5.5) years, and mean age of the 88 489 women serving as controls was 59.4 (4.9) years. The median follow-up time for the cohort was 10 years (maximum 13 years) (interquartile range, 9.44-10.82 years). Women with BC had a higher body mass index (relative risk [RR], 1.14; 95% CI, 1.05-1.23), performed less exercise (RR, 1.12; 95% CI, 1.01-1.25), used hormonal replacement therapy for longer than 5 years (RR, 1.23; 95% CI, 1.13-1.34), used more oral contraceptives (RR, 1.02; 95% CI, 0.93-1.12), and had greater alcohol intake (RR, 1.11; 95% CI, 1.03-1.19) compared with the controls. Overall, 20 657 women (23.3%) followed a favorable lifestyle, 60 195 women (68.0%) followed an intermediate lifestyle, and 7637 women (8.6%) followed an unfavorable lifestyle. The RR of the highest genetic risk group was 2.55 (95% CI, 2.28-2.84), and the RR of the most unfavorable lifestyle category was 1.44 (95% CI, 1.25-1.65). The association of lifestyle and BC within genetic subgroups showed lower HRs among women following a favorable lifestyle compared with intermediate and unfavorable lifestyles among all of the genetic groups: women with an unfavorable lifestyle had a higher risk of BC in the low genetic group (HR, 1.63; 95% CI, 1.13-2.34), intermediate genetic group (HR, 1.94; 95% CI, 1.46-2.58), and high genetic group (HR, 1.39; 95% CI, 1.11-1.74) compared with the reference group of favorable lifestyle. Intermediate lifestyle was also associated with a higher risk of BC among the low genetic group (HR, 1.40; 95% CI, 1.09-1.80) and the intermediate genetic group (HR, 1.37; 95% CI, 1.12-1.68).

(continued)

## Key Points

**Question** Is adhering to healthier lifestyle habits associated with a reduced breast cancer risk even among genetically predisposed groups?

**Findings** This cohort study evaluated 2728 women with breast cancer and 88 489 controls and noted lower risks of breast cancer among women who practice a healthy lifestyle. Factors included in this lifestyle were exercise, healthy weight, low alcohol intake, and no oral contraceptive use, as well as avoiding or limiting use of hormonal replacement therapy to less than 5 years, among low, intermediate, and high genetic risk groups.

**Meaning** Following a healthier lifestyle appears to be associated with a decreased level of risk of breast cancer across all strata of genetic risk.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

*Abstract (continued)*

**CONCLUSIONS AND RELEVANCE** In this cohort study of data on women in the UK Biobank, a healthier lifestyle with more exercise, healthy weight, low alcohol intake, no oral contraceptive use, and no or limited hormonal replacement therapy use appeared to be associated with a reduced level of risk for BC, even if the women were at higher genetic risk for BC.

*JAMA Network Open.* 2020;3(4):e203760. doi:10.1001/jamanetworkopen.2020.3760

## Introduction

Breast cancer (BC) is the most common cancer in women as well as the second most common cause of cancer-related death in women.[1,2] In the UK, it is estimated that more than 55 000 new cases of BC occur annually.[2] Both genetic and lifestyle factors play crucial roles in the complex mechanism of BC. Evidence supporting the genetic component of BC is seen with highly penetrant rare gene variants, such as in the *BRCA1* and *BRCA2* genes. These particular variants, however, account for just a small proportion (<5%) of BC cases[3] and for 1.5% to 2% in familial BC cases.[4] Genome-wide association studies have identified a number of single-nucleotide variations (SNVs) associated with risk for BC development, although these SNVs individually contribute only a small genetic proportion or are in genes exhibiting medium to low penetrance. The cumulative genetic contribution and effects of all such BC-associated SNVs is referred to as a polygenic risk score (PRS). This aggregated PRS is present in a substantial proportion of all patients with BC (88%).[5-7] The application of genetic risk stratification to individuals as a clinical tool for aiding BC screening is now on the horizon.[6] Mavaddat et al[8] showed that women at the top 5% of the PRS can develop BC at age 37 years, while those in the lowest 20% of the PRS will likely never develop BC.

Some lifestyle and behavioral factors can play an important role in and contribute to the risk of BC.[9-14] However, few studies have investigated the contribution and role of lifestyle risk exposures in BC in women exhibiting different PRSs. Whereas inherited genetic risk for disease is not modifiable, this factor is not the case for most known nongenetic risk factors. The central hypothesis examined in this study is that, regardless of a person's PRS, overall BC risk can be reduced by following a favorable lifestyle.

## Methods

Data from women within the UK Biobank longitudinal cohort study were used. The data set for the analysis was closed on March 31, 2019. The UK Biobank is a national cohort including 502 650 men and women aged between 39 and 71 years. Patients were enrolled between 2006 and 2010 and continue to be longitudinally followed up for capture of subsequent health events. Participants gave the UK Biobank written informed consent to use their data and samples for health-related research purposes. Ethics approval for use of UK Biobank data was obtained from the North West-Haydock research ethics committee. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline for cohort studies.

In this analysis, the inclusion criteria to select study participants were (1) British women who were white (age, 40-71 years), (2) postmenopausal women who did not report a history of hysterectomy or bilateral oophorectomy and reported no longer menstruating, and (3) women with a menopause age of 40 years or older. Deceased participants were excluded from our analysis. Of the UK Biobank cohort of 273 402 female participants, 114 723 women (42.0%) fulfilled our inclusion criteria.

The study outcome was defined as women with a malignant neoplasm of the breast. Cases and controls were identified according to the criteria summarized in eFigure 1 in the Supplement. We used 3 coding systems to identify patients with BC and those serving as controls: *International*

*Statistical Classification of Diseases and Related Health Problems, Tenth Revision*; *International Classification of Diseases, Ninth Revision*; and self-reported (eTable 1 in the Supplement). If patients with breast cancer appeared to have an incident case of BC according to any of these 3 coding systems, they were deemed incident cases (age at cancer diagnosis was older than age when they attended the assessment center of the UK Biobank study). Cases were considered prevalent only if they were defined as such according to any of the 3 coding systems, which was applicable only if none of the 3 approaches had described the BC case as being an incident case. A total of 2728 postmenopausal women with incident cases of BC were eligible for the analysis. Controls were defined as women without a history of any cancer, carcinoma in situ, or unknown neoplasm. The final number of controls selected by menopausal status and our set criteria was 88 489. eFigure 1 in the Supplement illustrates the number of study participants in the case and control selection process.

Cancer Research UK[15] has reported risk factors for BC development as being either modifiable or nonmodifiable. Based on their list, we identified the 5 modifiable factors: weight, alcohol intake, physical activity, oral contraceptive use, and hormonal replacement therapy (HRT) intake for more than 5 years. We developed a scoring system based on the presence or absence of these 5 factors to derive favorable lifestyle, intermediate lifestyle, and unfavorable lifestyle. This approach was adopted from similar studies on coronary heart disease[16] and dementia.[17] The details of the 5 factors and score definition are presented in **Table 1**. Eligible participants were stratified into 3 categories: favorable lifestyle (≥4 healthy factors present), intermediate lifestyle (2 or 3 healthy factors present), and unfavorable lifestyle (≤1 healthy factor present).

A PRS was derived based on the Mavaddat score[5] using the UK Biobank high-density genome-wide SNV data set available for 488 377 of their participants. The SNV data were used from individuals who were included on the basis of being female (matched genetic and self-reported sex) and their genetic ethnic grouping (white). During the quality control process, individuals with missingness (>2%), outliers for heterozygosity, and duplicates, as well as those who were biologically related, were excluded.

The PRS for BC was constructed using the 313 SNVs previously determined to contribute some risk by the hard threshold approach used by Mavaddat et al.[5] Of these 313 SNVs, 306 were present in the UK Biobank data set; however, SNV rs10764337 was triallelic and excluded. The final number of SNVs used for PRS construction was therefore 305, and their details are presented in eTable 2 in the Supplement. Forty of 305 SNVs had been directly genotyped and successfully passed the marker test applied by UK Biobank.[18] The remaining 265 SNVs had been imputed. The quality of the imputation was estimated using the information scores available, which is a number between 0 and

Table 1. Criteria for Healthy Lifestyle Classification

| Factor | UK Biobank cohort | Code |
|---|---|---|
| **Healthy lifestyle criteria** | | |
| Healthy weight | Healthy: BMI <25 | 1 |
| | Unhealthy: BMI ≥25 | 0 |
| Regular physical activity | Healthy: ≥1 time/wk | 1 |
| | Unhealthy: none | 0 |
| Alcohol intake | Healthy: none or <3 times/wk | 1 |
| | Unhealthy: ≥3 times/wk | 0 |
| Oral contraceptive use | Healthy: none | 1 |
| | Unhealthy: any use | 0 |
| Hormone replacement therapy | Healthy: none or <5 y | 1 |
| | Unhealthy: use for ≥5 y | 0 |
| **Lifestyle classification** | | |
| Favorable | Presence of 4-5 healthy lifestyle factors | Sum: ≥4 |
| Intermediate | Presence of 2-3 healthy lifestyle factors | Sum: 2 or 3 |
| Unfavorable | Presence of ≤1 healthy lifestyle factor | Sum: 0 or 1 |

Abbreviations: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared).

1 where 0 indicates complete uncertainty and 1 indicates complete certainty. The lowest information score was 0.86. Linkage disequilibrium was assessed, and no $r^2$ value between any 2 SNVs reached 0.9. Plinkopen source software version 1.90 was used to carry out the quality control processes.[19]

Individual participant PRS was created by adding the number of risk alleles at each SNV and then multiplying the sum by the effect size as the previously published estimated effect size.[5] The raw PRS was standardized by dividing each raw PRS by the SD of the PRS derived from the control group. No transformation to the PRS data was applied because the scores were normally distributed (eFigure 2 in the Supplement). A tertile genetic risk classification using standardized PRS values from controls was generated. Each participant was then assigned to a genetic risk group: low (1st tertile up to 33.33%), intermediate (2nd tertile between 33.34% and 66.67%), and high (3rd tertile from 66.68% to 100%).

### Statistical Analysis

Relative risks (RRs) and 95% CIs of the basic risk factors were computed with an adjustment for age and family history using a binomial generalized linear regression model. Cox proportional hazards regression was applied to assess the hazard ratios (HRs) of the lifestyles and BC risk. We first computed HRs for each genetic stratum with the low genetic risk group as a reference group and for each lifestyle (favorable, intermediate, and unfavorable) stratum with the favorable category as a reference group. The HRs in each lifestyle stratum were calculated within each genetic risk group. All analyses were adjusted for age and family history. The Cox proportional hazards regression model assumption for each analysis was tested. A 2-sided *P* value <.05 was considered significant. The Ltable[20] command was used to compute a 10-year cumulative BC incidence for each lifestyle category within each genetic risk stratum. Results presented in graphic bar charts were generated using Microsoft Excel 2016 (Microsoft Corp).[21] All analyses were performed using Stata/MP software version 14 (StataCorp LLC).[22]

## Results

The median follow-up time for the cohort was 10 years (maximum, 13 years) (interquartile range, 9.44-10.82 years). The total number of the incident cases was 2728 patients with BC, and the total number of controls was 88 489. The mean (SD) age of the patients was 60.1 (5.5) years and for controls was 59.4 (4.9) years. The mean (SD) body mass index (BMI) measures (calculated as weight in kilograms divided by height in meters squared) were 27.3 (5.0) for patients and 26.9 (4.9) for controls. In addition, patients used more HRT (30.4%) compared with controls (25.2%). Furthermore, women with BC more often reported no regular physical activity (13.3%) compared with controls (12.0%).

**Table 2** presents the distribution of the general characteristics and estimated RR results. A 1-year increase in age was associated with a 2.3% increase in BC development risk. Having 1 female first-degree family member (either mother or sister) with BC was associated with a 48.6% increase in BC risk, while having both mother and sister affected was associated with a doubling of the risk of BC compared with women without a family history of BC. An unhealthy weight (BMI ≥25) was associated with a 13.9% increased risk of BC (RR, 1.14; 95% CI, 1.05-1.23). Participants who reported that they did not have regular physical activity were had a 12.2% increased risk of BC (RR, 1.12; 95% CI, 1.01-1.25), and alcohol intake 3 or more times per week was associated with an increased BC risk of 10.7% (RR, 1.11; 95% CI, 1.03-1.19). Use of HRT for 5 or more years was associated with an increased BC risk of 22.9% (RR, 1.23; 95% CI, 1.13-1.34). History of oral contraceptive use did not show any association with BC risk among women in the UK Biobank (RR, 1.02; 95% CI, 0.93-1.12); however, this factor was retained as part of lifestyle classification. Overall, 20 657 women (23.3%) followed a favorable lifestyle, 60 195 women (68.0%) followed an intermediate lifestyle, and 7637 women (8.6%) followed an unfavorable lifestyle. Intermediate and unfavorable lifestyles were both

Table 2. Relative Risks for Basic Characteristics, Lifestyles, and Genetic Categories

| | Frequency, No. (%) | | |
|---|---|---|---|
| Risk factor | Cases | Controls | RR (95% CI) |
| Age[a] | 2728 (2.99) | 88 489 (97.01) | 1.02 (1.02-1.03) |
| Family history[b] | | | |
| No family history | 2276 (83.80) | 78 408 (88.84) | 1 [Reference] |
| Mother or sister BC history | 412 (15.17) | 9405 (10.66) | 1.49 (1.34-1.65) |
| Mother and sister BC history | 28 (1.03) | 440 (0.50) | 2.10 (1.46-3.01) |
| Weight | | | |
| Healthy | 995 (36.55) | 35 537 (40.25) | 1 [Reference] |
| Unhealthy | 1727 (63.45) | 52 749 (59.75) | 1.14 (1.05-1.23) |
| Regular physical activity | | | |
| ≥1 time/wk | 2329 (86.74) | 76 466 (88.00) | 1 [Reference] |
| No physical activity | 356 (13.26) | 10 423 (12.00) | 1.12 (1.01-1.25) |
| Alcohol intake | | | |
| No intake or <3 times/wk | 1566 (57.40) | 52 892 (59.80) | 1 [Reference] |
| Intake ≥3 times/wk | 1162 (42.60) | 35 557 (40.20) | 1.11 (1.03-1.19) |
| Oral contraceptive intake | | | |
| No | 561 (20.58) | 17 240 (19.50) | 1 [Reference] |
| Yes | 2165 (79.42) | 71 149 (80.50) | 1.02 (0.93-1.12) |
| HRT intake | | | |
| No | 1895 (69.64) | 66 093 (74.82) | 1 [Reference] |
| Yes | 826 (30.36) | 22 244 (25.18) | 1.23 (1.13-1.34) |
| Healthy lifestyle score | | | |
| Favorable | 530 (19.43) | 20 657 (23.34) | 1 [Reference] |
| Intermediate | 1909 (69.98) | 60 195 (68.03) | 1.25 (1.13-1.37) |
| Unfavorable | 289 (10.59) | 7637 (8.63) | 1.44 (1.25-1.65) |
| PRS category | | | |
| Low | 440 (19.67) | 24 297 (33.70) | 1 [Reference] |
| Intermediate | 655 (29.28) | 23 983 (33.27) | 1.49 (1.32-1.68) |
| High | 1142 (51.05) | 23 814 (33.03) | 2.55 (2.28-2.84) |

Abbreviations: BC, breast cancer; HRT, hormone replacement therapy; PRS, polygenic risk score; RR, relative risk.

[a] No adjustment.

[b] Adjusted for age only.

Table 3. Breast Cancer HRs Based on Lifestyles, Stratified by the Genetic Risk Group

| | | Frequency, No. (%) | | |
|---|---|---|---|---|
| Genetic risk group | Healthy lifestyle score[a] | Cases | Controls | HR (95% CI) |
| Low | Favorable lifestyle | 75 (17.05) | 5550 (22.84) | 1 [Reference] |
| | Intermediate lifestyle | 317 (72.05) | 16 540 (68.07) | 1.40 (1.09-1.80) |
| | Unfavorable lifestyle | 48 (10.91) | 2204 (9.08) | 1.63 (1.14-2.34) |
| | PH assumption *P* value | .99 | | |
| | *P* value | .004 | | |
| Intermediate | Favorable lifestyle | 117 (17.86) | 5582 (23.27) | 1 [Reference] |
| | Intermediate lifestyle | 458 (69.92) | 16 336 (68.11) | 1.37 (1.12-1.68) |
| | Unfavorable lifestyle | 80 (12.21) | 2065 (8.61) | 1.94 (1.46-2.58) |
| | PH assumption *P* value | .08 | | |
| | *P* value | <.001 | | |
| High | Favorable lifestyle | 236 (20.67) | 5571 (23.39) | 1 [Reference] |
| | Intermediate lifestyle | 792 (69.35) | 16 278 (68.35) | 1.13 (0.98-1.31) |
| | Unfavorable lifestyle | 114 (9.98) | 1965 (8.25) | 1.39 (1.11-1.74) |
| | PH assumption *P* value | .69 | | |
| | *P* value | .007 | | |

Abbreviations: BC, breast cancer; HR, hazard ratio; PH, proportional hazards.

[a] Adjusted for age and family history of BC.

associated with higher risk of BC compared with the favorable lifestyle (intermediate: RR, 1.25; 95% CI, 1.13-1.37; unfavorable: RR, 1.44; 95% CI, 1.25-1.65).

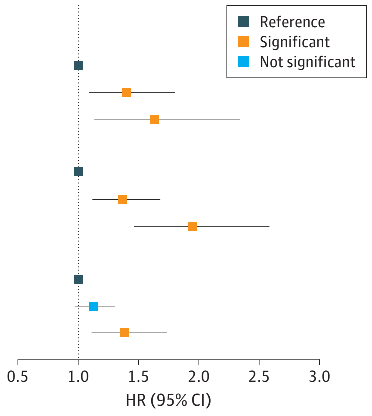
The mean standardized PRS of the cases was 26.26 (range, 21.63-29.40), which is higher than the mean standardized PRS of the control group (25.807; range, 21.119-29.941). This difference was examined using a *t* test, and a significant difference between the mean score was apparent between cases and controls (*P* < .001). Moreover, the estimated HR for overall BC among postmenopausal women per unit of increased PRS was 1.55 (95% CI, 1.48-1.61). Analysis of the PRS tertile groups indicated a gradient of increased BC risk across tertiles (for second tertile vs first tertile, *P* < .001; for third tertile vs first tertile, *P* < .001). Women in the higher genetic risk group (3rd tertile) were at significantly higher risk of BC (RR, 2.55; 95% CI, 2.28-2.84) compared with women in the low genetic risk group after adjusting for age and family history. Similarly, women in the intermediate risk group showed a moderate increased risk (RR, 1.49; 95% CI, 1.32-1.68) compared with those in the low genetic risk group.

Results of estimated HRs for lifestyle and BC risk in each genetic risk group are presented in **Table 3**. The results of Cox proportional hazards regression model assumption testing in the low, intermediate, and high genetic risk groups suggested no statistically significant violation of Cox proportional hazards regression model assumption. In the low genetic risk group, significantly increased HRs were observed in both the unfavorable lifestyle (HR, 1.63; 95% CI, 1.13-2.34) and intermediate lifestyle (HR, 1.40; 95% CI, 1.09-1.80) groups compared with the favorable lifestyle group. In the intermediate genetic risk group, significantly increased HRs were shown in the unfavorable (HR, 1.94; 95% CI, 1.46-2.58) and intermediate (HR, 1.37; 95% CI, 1.12-1.68) lifestyle groups. In the higher genetic risk strata, a significant HR was observed in the unfavorable lifestyle group (HR, 1.39; 95% CI, 1.11-1.74) compared with favorable lifestyle. All of the above results suggest that, within the same genetic risk group, adhering to a less healthy lifestyle (intermediate and unfavorable lifestyle) is associated with an increased risk of BC. **Figure 1** shows a forest plot of HRs according to genetic risk group and lifestyle categories.

The results of the 10-year cumulative incidence rate of BC in all genetic risk groups suggest incremental rates of increase from favorable to intermediate to unfavorable (**Figure 2**) lifestyle. A favorable lifestyle had the lowest 10-year cumulative BC incidence rate across all genetic risk groups (low, 3%; intermediate, 5%; and high, 9%). Similar findings in the 10-year cumulative BC incidence rate were observed for an unfavorable lifestyle across the genetic risk groups (low, 5%; intermediate, 9%; and high, 12%).

Figure 1. Association of Breast Cancer With Lifestyle and Genetic Factors

| Genetic risk group | HRs (95% CI) |
|---|---|
| Low | |
| Favorable lifestyle | 1.000 (1.000-1.000) |
| Intermediate lifestyle | 1.401 (1.090-1.802) |
| Unfavorable lifestyle | 1.630 (1.135-2.342) |
| Intermediate | |
| Favorable lifestyle | 1.000 (1.000-1.000) |
| Intermediate lifestyle | 1.372 (1.119-1.682) |
| Unfavorable lifestyle | 1.945 (1.463-2.587) |
| High | |
| Favorable lifestyle | 1.000 (1.000-1.000) |
| Intermediate lifestyle | 1.130 (0.977-1.307) |
| Unfavorable lifestyle | 1.391 (1.112-1.740) |



The hazard ratio (HR) of each genetic group was stratified based on the favorable, intermediate, and unfavorable lifestyles, with favorable lifestyle as the reference group in the 3 genetic groups.

## Discussion

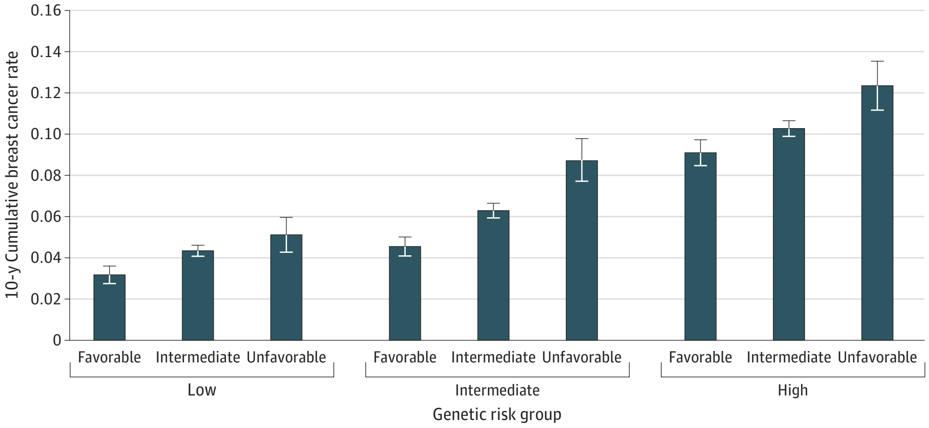
It has been estimated that BC could be prevented in 23% of patients in the UK.[2] Thus, it is important to understand the contribution of modifiable risk factors to BC and how they affect or add to the inherited genetic factors. This study therefore investigated the association between genetic and lifestyle factors with BC risk and tested the hypothesis that BC risk in postmenopausal women can be modified or reduced by improving lifestyle habits, even for the highest genetic risk group. We opted to investigate our hypothesis only in postmenopausal women because of the high proportion of BC incidence and prevalence in this group.[15,23] Furthermore, BC in premenopausal women is usually a more aggressive disease, likely caused by high penetrance genes,[24-27] resulting in a less-favorable prognosis.[28]

This study used genetic and lifestyle data generated by UK Biobank, a longitudinal study of the contribution of genetic, environmental, and lifestyle risk factors in disease. Participants were grouped by their level of polygenetic risk for BC using the SNV data available within the UK Biobank database. The 305 SNVs included in the PRS were mainly common variants with limited contribution to BC risk. Aggregated effect sizes of these SNVs were used to develop a standardized PRS.

Although many risk/protective factors contribute to BC development,[29] we selected 5 robust modifiable risk factors, recognized previously by Cancer Research UK as being associated with BC in white females.[9,30-32] The frequencies of these modifiable risk factors are high in women in the UK, and if they can be modified can potentially reduce BC incidence. The prevalence of these 5 modifiable risk factors in the UK Biobank female cohort were as follows: 63.4% exhibiting unhealthy weight in patients with BC vs 59.8% in controls, 13.3% of patients with BC having no regular exercise vs 12.0% of controls, 42.6% of patients with BC with regular alcohol intake vs 40.2% of controls, 79.4% of patients with BC who used oral contraceptives vs 80.5% of controls, and 30.4% of patients with BC who received HRT vs 25.2% of controls.

The findings from other large cohorts, including the Million Women Study and the Breast Cancer Association Consortium, have indicated that BC risk increase is associated with unhealthy weight,[9,12,33] no or limited exercise,[12,13] level of alcohol intake,[12-14] use of oral contraceptives,[9,12] and use of HRT.[9-12,34] The Cancer Research UK suggested that the relative contributions of these factors to BC development are 2% for HRT, 8% for obesity, 8% for alcohol intake, and less than 1% for use of oral contraceptives.[2] The results from our study are in keeping with the Cancer Research UK in that maintaining a healthy weight is associated with reduced BC risk by 13.9%, participating in regular exercise is associated with reduced BC risk by 12.2%, maintaining alcohol intake at less than 3 times a week is associated with reduced BC risk by 10.7%, and avoiding HRT use is associated with reduced BC risk by 22.9%. Our findings therefore support the selection of these modifiable lifestyle risk

Figure 2. Ten Year Cumulative Breast Cancer Incidence Rate of UK Biobank Postmenopausal Women, Classified According to Genetic and Lifestyle Factors



The error bars represent the mean rate with the maximum and minimum incidence rate.

factors for BC, with the exception of oral contraceptive use. Thus, further studies are needed to investigate whether there is a causal association between new risk factors and BC using, for example, a mendelian randomization approach.

Even though oral contraceptive use has been suggested previously to be associated with BC, this risk factor did not show any association in our study. Possible explanations for this observation could be that we did not take into account other related factors that could be associated with the results, including the type of oral contraceptive used,[35] the duration of use,[36] and age at the time when the drugs were stopped.[37] Furthermore, women who have had human chorionic gonadotropin injections as part of infertility or weight loss treatments showed a lower risk of BC.[38] All of these factors may have implications in BC risk. For example, if women stopped oral contraceptive use for more than 10 years before their enrollment in the UK Biobank study, their BC risk will be reduced or returned to the same risk of women who never used oral contraceptives.[37]

Exhibiting 2 or 3 of these healthy lifestyle factors (intermediate lifestyle) was associated with increased risk of BC by 24.5% compared with an increase of 43.6% in women who adhered to none or 1 of these factors (unfavorable lifestyle). Our findings suggest that women may be able to alter or reduce their risk of developing BC by following healthier lifestyles. While we did not set out to look for a formal interaction owing to limited study power, the results showed no significant interaction between lifestyles and genetic risk groups, and the 2 variables were considered as independent in the analysis. Further analysis demonstrated that a high PRS was associated with higher risk of BC. This level of increased risk is in line with other published findings.[5] The HRs derived from our analysis were generated by including only postmenopausal women. In contrast, the study by Mavaddat et al[5] reported HRs that were derived from both premenopausal and postmenopausal women.

The beneficial risk-reducing association of adhering to healthy lifestyles across all genetic risk stratification groups supports our hypothesis that BC risk reduction is seen regardless of the effect size of the PRS. We also found an association between 10-year cumulative BC incidence rate and both lifestyle and genetic factors when assessed together. This increase suggests that BC incidence may be reduced by following favorable lifestyles even in women with high genetic risk.

This study suggests that the lifestyle followed by women may contribute to reducing the incidence of BC in those who have an increased genetic predisposition for this condition. Similar approaches have been used to investigate complex risk factors associated with dementia[17] and coronary heart disease.[16] Both studies came to a conclusion similar to ours. In the dementia study, by adhering to favorable lifestyles (no current smoking, moderate alcohol intake, healthy diet, and regular exercise), the level of dementia was reduced. Similarly, in coronary artery disease, no smoking, no obesity, healthy diet, and regular exercise were associated with a reduction in the extent of coronary heart disease in participants, and this result was also observed in patients within the highest PRS group.

## Strengths and Limitations

A strength of our study is that a large sample size was analyzed and the selection of participants was spread across the UK. Furthermore, the quality and comprehensive nature of the phenotypic exposures assessed by UK Biobank were robust and of high standards. Our use of a prospective study design allowed exposure assessment before BC development in the cohort. However, the study has some limitations. The PRS used was restricted to white women and therefore presents a limitation on its generalizability to a wider range of racial/ethnic groups. Additional validation of these PRSs in other populations is needed to further understand its utility in genetic risk stratification. Our analysis was restricted to postmenopausal women; therefore, these results cannot be applied to premenopausal women.

However, the benefits reported herein for healthy lifestyle factors may also be seen in younger women. In addition, our analysis did not investigate the various known pathologic-based subtypes of BC, including *ER* positive and negative, *PR* positive and negative, and *ERBB2* (formerly *HER2* or *Her2/neu*) positive and negative.

## Conclusions

The results of this study suggest that promotion of healthy lifestyles through adequate levels of exercise, healthy weight, no or limited alcohol intake, and avoidance of hormonal replacement therapy should be encouraged to reduce the risk of BC. Following a healthy lifestyle appears to be associated with a reduced level of BC risk in all 3 genetic risk strata, further illustrating the importance of lifestyle factors in common diseases with a genetic predisposition, such as BC.

## ARTICLE INFORMATION

**Accepted for Publication:** February 25, 2020.

**Published:** April 24, 2020. doi:10.1001/jamanetworkopen.2020.3760

**Corresponding Author:** Kenneth R. Muir, PhD, Division of Population Health, Health Services Research and Primary Care, Faculty of Biology, Medicine and Health, Centre for Epidemiology, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom (kenneth.muir@manchester.ac.uk).

**Author Affiliations:** Division of Population Health, Health Services Research and Primary Care, Faculty of Biology, Medicine and Health, Centre for Epidemiology, University of Manchester, Manchester, United Kingdom (Al Ajmi, Lophatananon, Ollier, Muir); Cathie Marsh Institute for Social Research, School of Social Sciences, Faculty of Humanities, University of Manchester, Manchester, United Kingdom (Mekli); School of Healthcare Science, Faculty of Science and Engineering, Manchester Metropolitan University, Manchester, United Kingdom (Ollier).

**Author Contributions:** Drs Muir and Lophatananon had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

*Concept and design:* Al Ajmi, Ollier, Muir.

*Acquisition, analysis, or interpretation of data:* Al Ajmi, Lophatananon, Mekli.

*Drafting of the manuscript:* Al Ajmi.

*Critical revision of the manuscript for important intellectual content:* All authors.

*Statistical analysis:* Al Ajmi, Lophatananon, Mekli.

*Administrative, technical, or material support:* Lophatananon.

*Supervision:* Lophatananon, Ollier, Muir.

**Conflict of Interest Disclosures:** None reported.

**Funding/Support:** Drs Lophatananon and Muir are supported by the National Institute for Health Research Manchester Biomedical Research Centre, Alan Turing Institute under Engineering and Physical Sciences Research Council grant EP/N510129/1, and Cancer Research UK Integrated Cancer Epidemiology Programme grant C18281/A19169.

**Role of the Funder/Sponsor:** The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

**Additional Information:** Continued longitudinal follow-up for capture of subsequent health events in patients enrolled between 2006 and 2010 is available at https://www.ukbiobank.ac.uk/.

**Additional Contributions:** We thank the UK Biobank staff and participants for these data.

## REFERENCES

**1**. McPherson K, Steel CM, Dixon JM. ABC of breast diseases: breast cancer-epidemiology, risk factors, and genetics. *BMJ*. 2000;321(7261):624-628. doi:10.1136/bmj.321.7261.624

**2**. Cancer Research UK. Breast cancer statistics. Accessed April 5, 2019. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Four

**3**. Nathanson KL, Wooster R, Weber BL. Breast cancer genetics: what we know and what we need. *Nat Med*. 2001;7(5):552-556. doi:10.1038/87876

**4**. Venkitaraman AR. Cancer susceptibility and the functions of *BRCA1* and *BRCA2*. *Cell*. 2002;108(2):171-182. doi:10.1016/S0092-8674(02)00615-3

**5**. Mavaddat N, Michailidou K, Dennis J, et al; ABCTB Investigators; kConFab/AOCS Investigators; NBCS Collaborators. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet*. 2019;104(1):21-34. doi:10.1016/j.ajhg.2018.11.002

**6**. Pashayan N, Duffy SW, Chowdhury S, et al. Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. *Br J Cancer*. 2011;104(10):1656-1663. doi:10.1038/bjc.2011.118

**7**. Pharoah PDP, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet*. 2002;31(1):33-36. doi:10.1038/ng853

**8**. Mavaddat N, Pharoah PD, Michailidou K, et al. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst*. 2015;107(5):djv036. doi:10.1093/jnci/djv036

**9**. Al-Ajmi K, Lophatananon A, Ollier W, Muir KR. Risk of breast cancer in the UK Biobank female cohort and its relationship to anthropometric and reproductive factors. *PLoS One*. 2018;13(7):e0201097. doi:10.1371/journal.pone.0201097

**10**. Beral V; Million Women Study Collaborators. Breast cancer and hormone-replacement therapy in the Million Women Study. *Lancet*. 2003;362(9382):419-427. doi:10.1016/S0140-6736(03)14065-2

**11**. Beral V, Reeves G, Bull D, Green J; Million Women Study Collaborators. Breast cancer risk in relation to the interval between menopause and starting hormone therapy. *J Natl Cancer Inst*. 2011;103(4):296-305. doi:10.1093/jnci/djq527

**12**. Colditz GA, Atwood KA, Emmons K, 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(6):477-488. doi:10.1023/A:1008984432272

**13**. Elwood PC, Whitmarsh A, Gallacher J, et al. Healthy living and cancer: evidence from UK Biobank. *Ecancermedicalscience*. 2018;12:792. doi:10.3332/ecancer.2018.792

**14**. Hamajima N, Hirose K, Tajima K, et al; Collaborative Group on Hormonal Factors in Breast Cancer. Alcohol, tobacco and breast cancer—collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. *Br J Cancer*. 2002;87(11):1234-1245. doi:10.1038/sj.bjc.6600596

**15**. Cancer Research UK. Risk factors 2017. Accessed March 26, 2020. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/risk-factors

**16**. Khera AV, Emdin CA, Drake I, et al. Genetic risk, adherence to a healthy lifestyle, and coronary disease. *N Engl J Med*. 2016;375(24):2349-2358. doi:10.1056/NEJMoa1605086

**17**. Lourida I, Hannon E, Littlejohns TJ, et al. Association of lifestyle and genetic risk with incidence of dementia. *JAMA*. 2019;322(5):430-437. doi:10.1001/jama.2019.9879

**18**. Bycroft C, Freeman C, Petkova D, et al Genome-wide genetic data on ~500,000 UK Biobank participants. Preprint. Posted online July 20, 2017. bioRxiv 166298. doi:10.1101/166298

**19**. Chang C. PLINK 1.90 beta. Updated 2019. Accessed June 28, 2019. https://www.cog-genomics.org/plink2

**20**. Stata.com. Ltable—life tables for survival data. Accessed May 5, 2019. https://www.stata.com/manuals13/stltable.pdf

**21**. Microsoft. Microsoft Excel. Accessed July 24, 2019. https://products.office.com/en-gb/excel

**22**. StataCorp LLC. Stata/MP. Accessed November 1, 2018. https://www.stata.com/statamp/

**23**. Dartois L, Fagherazzi G, Baglietto L, et al. Proportion of premenopausal and postmenopausal breast cancers attributable to known risk factors: estimates from the E3N-EPIC cohort. *Int J Cancer*. 2016;138(10):2415-2427. doi:10.1002/ijc.29987

**24**. Haffty BG, Harrold E, Khan AJ, et al. Outcome of conservatively managed early-onset breast cancer by *BRCA1/2* status. *Lancet*. 2002;359(9316):1471-1477. doi:10.1016/S0140-6736(02)08434-9

**25**. de la Rochefordière A, Asselain B, Campana F, et al. Age as prognostic factor in premenopausal breast carcinoma. *Lancet*. 1993;341(8852):1039-1043. doi:10.1016/0140-6736(93)92407-K

**26**. Walker RA, Lees E, Webb MB, Dearing SJ. Breast carcinomas occurring in young women (< 35 years) are different. *Br J Cancer*. 1996;74(11):1796-1800. doi:10.1038/bjc.1996.632

**27**. Mavaddat N, Antoniou AC, Easton DF, Garcia-Closas M. Genetic susceptibility to breast cancer. *Mol Oncol*. 2010;4(3):174-191. doi:10.1016/j.molonc.2010.04.011

**28**. Lewis DR, Seibel NL, Smith AW, Stedman MR. Adolescent and young adult cancer survival. *J Natl Cancer Inst Monogr*. 2014;2014(49):228-235. doi:10.1093/jncimonographs/lgu019

**29**. Al-Ajmi K, Lophatananon A, Yuille M, Ollier W, Muir KR. Review of non-clinical risk models to aid prevention of breast cancer. *Cancer Causes Control*. 2018;29(10):967-986. doi:10.1007/s10552-018-1072-6

**30**. La Vecchia C, Carioli G. The epidemiology of breast cancer, a summary overview. *Epidemiol Biostat Public Health*. 2018;15(1):e12853. doi:10.2427/12853

**31**. Dumitrescu RG, Cotarla I. Understanding breast cancer risk—where do we stand in 2005? *J Cell Mol Med*. 2005;9(1):208-221. doi:10.1111/j.1582-4934.2005.tb00350.x

**32**. Hulka BS, Moorman PG. Breast cancer: hormones and other risk factors. *Maturitas*. 2008;61(1-2):203-213. doi:10.1016/j.maturitas.2008.11.016

**33**. Guo W, Key TJ, Reeves GK. Adiposity and breast cancer risk in postmenopausal women: results from the UK Biobank prospective cohort. *Int J Cancer*. 2018;143(5):1037-1046. doi:10.1002/ijc.31394

**34**. Ross RK, Paganini-Hill A, Wan PC, Pike MC. Effect of hormone replacement therapy on breast cancer risk: estrogen versus estrogen plus progestin. *J Natl Cancer Inst*. 2000;92(4):328-332. doi:10.1093/jnci/92.4.328

**35**. Dumeaux V, Alsaker E, Lund E. Breast cancer and specific types of oral contraceptives: a large Norwegian cohort study. *Int J Cancer*. 2003;105(6):844-850. doi:10.1002/ijc.11167

**36**. Kumle M, Weiderpass E, Braaten T, Persson I, Adami HO, Lund E. Use of oral contraceptives and breast cancer risk: The Norwegian-Swedish Women's Lifestyle and Health Cohort Study. *Cancer Epidemiol Biomarkers Prev*. 2002;11(11):1375-1381.

**37**. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. *Lancet*. 1996;347(9017):1713-1727. doi:10.1016/S0140-6736(96)90806-5

**38**. Bernstein L, Hanisch R, Sullivan-Halley J, Ross RK. Treatment with human chorionic gonadotropin and risk of breast cancer. *Cancer Epidemiol Biomarkers Prev*. 1995;4(5):437-440.

**SUPPLEMENT.**
**eFigure 1.** Number of Participants in Each Filter Step
**eFigure 2.** Distribution of Polygenic Risk Scores
**eTable 1.** Identification Codes for Patients With Breast Cancer and Controls in UK Biobank Cohort
**eTable 2.** 305 SNPs Used in Calculating the Polygenic Risk Scores of the UK Biobank Females