RESEARCH ARTICLE

Epidemiological breast cancer prediction by country: A novel machine learning approach

Hasna EL HAJI[1]*, Nada SBIHI[1], Bassma GUERMAH[1], Amine SOUADKA[2], Mounir GHOGHO[1]

1 TICLab, International University of Rabat, Rabat, Morocco, **2** Surgical Oncology Department, National Institute of Oncology, Mohammed V University, Rabat, Morocco

* hasna.elhaji@uir.ac.ma

OPEN ACCESS

Citation: EL HAJI H, SBIHI N, GUERMAH B, SOUADKA A, GHOGHO M (2024) Epidemiological breast cancer prediction by country: A novel machine learning approach. PLoS ONE 19(8): e0308905. https://doi.org/10.1371/journal.pone.0308905

Editor: Zahra Cheraghi, Hamadan University of Medical Sciences, School of Public Health, IRAN, ISLAMIC REPUBLIC OF

Received: May 6, 2024

Accepted: July 27, 2024

Published: August 14, 2024

Data Availability Statement: The data supporting this study are publicly accessible from various sources. Incidence rates of breast cancer can be found at the Global Cancer Observatory website (https://gco.iarc.fr/). Information on the average age of women by country is available at World Data (https://www.worlddata.info/average-age.php). Data on the prevalence of insufficient physical activity, depression, overweight, obesity, BMI, alcohol consumption, and smoking rates are provided by the World Health Organization Global

Abstract

Breast cancer remains a significant contributor to cancer-related deaths among women globally. We seek for this study to examine the correlation between the incidence rates of breast cancer and newly identified risk factors. Additionally, we aim to utilize machine learning models to predict breast cancer incidence at a country level. Following an extensive review of the available literature, we have identified a range of recently studied risk factors associated with breast cancer. Subsequently, we gathered data on these factors and breast cancer incidence rates from numerous online sources encompassing 151 countries. To evaluate the relationship between these factors and breast cancer incidence, we assessed the normality of the data and conducted Spearman's correlation test. Furthermore, we refined six regression models to forecast future breast cancer incidence rates. Our findings indicate that the incidence of breast cancer is most positively correlated with the average age of women in a country, as well as factors such as meat consumption, CO2 emissions, depression, sugar consumption, tobacco use, milk intake, mobile cells, alcohol consumption, pesticides, and oral contraceptive use. As for prediction, the CatBoost Regressor successfully predicted future breast cancer incidence with an R squared value of 0.84 ± 0.03. An increased incidence of breast cancer is mainly associated with dietary habits and lifestyle. Our findings and recommendations can serve as a baseline for developing educational programs intended to heighten awareness amongst women in countries with heightened risk.

1 Introduction

Breast cancer (BC) is the most prevalent cancer diagnosed in women (over a third of all female cancers). Even if cardiovascular diseases are the leading cause of death, it was estimated that deaths caused by cancer will exceed those caused by cardiovascular diseases in a few decades [1]. According to the World Health Organization, in 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally [2]. Besides physical and psychological suffering induced by breast cancer, all relatives may be affected, especially the children. Therefore, socio-economic consequences may be severe since two out of every three employees are forced to interrupt their careers [3].

Health Observatory (https://www.who.int/data/gho/data/indicators). CO2 emissions data can be accessed through the International Energy Agency (https://www.iea.org/data-and-statistics/data-products?filter=emissions). Data on breastfeeding, contraception use, and world fertility rates are available from the United Nations Department of Economic and Social Affairs (https://www.un.org/development/desa/pd/data/world-contraceptive-use). Information on food supply quantities of meat, sugar, and milk is provided by the Food and Agriculture Organization (https://www.fao.org/faostat/en/#home). Additionally, mobile cell data can be accessed at OpenCellID (https://www.opencellid.org/#zoom=16&lat=37.77889&lon=-122.41942).

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors declare that they have no competing interests that could be perceived to bias this work. No financial support was received for this study.

In spite of high incidence rates, access to care and social support promotes health in developed countries [4]. Unlike in developing countries, where delays in diagnosis, high treatment costs, and a lack of support all contribute to increasing mortality rates [5, 6] and the situation worsened even further in the context of COVID-19, Morocco as an example [7, 8]. Thus, to take the first step toward BC prevention, it is crucial to understand the risk factors involved.

Despite the large number of breast cancer risk factors recently presented in the literature, the discussion and decision on the most relevant ones are still ambiguous. Expressly, the subject lacks a systematic literature review that identifies emerging risk factors and examines their association with the total incidence of breast cancer at a country level. It is particularly important to consider modifiable risk factors in order to provide some prevention recommendations. Education may be a fundamental part of this process, yet, only few programs focused on breast cancer education have emerged recently in this area [9]. Nonetheless, educational interventions for women at all phases of life are highly required. They may provide a good impact by raising awareness and encouraging self-care.

In addition to examining the emerging risk factors, it may also be important to find out their breakdown by country. It may even be more important to anticipate breast cancer future incidence rates per country. Some research studies have predicted the future incidence and mortality of breast cancer in some countries, such as Iran [10] Pakistan [11] and Japan [12]. Generally, they use time series models or they compute the number of new breast cancer incidence and deaths by multiplying the age-specific incidence (or mortality) rates estimated for a given year, by the corresponding expected population for a future interval of years. However, to the best of our knowledge, no study has been carried out in order to predict the incidence rate of breast cancer by country using machine learning and common risk factors shared between individuals in the same country.

Additionally, it should be highlighted that besides the investigated risk factors, it may also be relevant to incorporate some preventive reproductive factors, particularly breastfeeding and total fertility rates. We assumed that these factors could reduce the incidence rate of breast cancer since there is a growing body of research relating breast cancer risk reduction to breastfeeding [13]. Furthermore, we want to assess the association between fertility and country-specific breast cancer incidence.

The primary objective of this study is to review the most recent risk factors established through meta-analyses and systematic reviews. The secondary objective involves collecting and analyzing available data on risk and preventive factors from multiple countries, examining their correlation with reported breast cancer incidence rates. The third objective is to create a specific profile of modifiable risk factors that can be addressed within each country, such as external factors, population eating habits, and lifestyles. This analysis aims to propose potential solutions for mitigating the impact of these factors. Lastly, the study aims to develop a model that serves as a starting point for predicting future breast cancer incidence based on risk and preventive factors across countries.

The ultimate goal of this research is to conduct a larger study with the main objective of determining the current rates of each factor in each country and subsequently predicting future breast cancer incidence. This information can help identify high-risk countries and facilitate early detection strategies, particularly in developing nations where early detection programs are lacking, and cancer is often diagnosed at advanced stages. For example, if our model predicts a high future incidence rate in a particular country, this knowledge can assist the government in implementing appropriate actions, such as designing educational programs to raise awareness among the at-risk population.

2 Methods

Unlike traditional cancer studies that typically recruit cohorts, this study adopted an ecologic study design by gathering data at the country level. To predict breast cancer incidence, we examined the rates of risk factors across the entire population within each country. It is important to note that this study adheres to the guidelines outlined in Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) [14].

Through an extensive bibliographic study, we examined the non-genetic risk factors associated with breast cancer. Subsequently, we collected data on these factors from various countries. Additionally, we included information on two preventive factors, namely the total fertility rate and breastfeeding rate. It is important to note that we deliberately excluded genetic factors from our analysis. Genetic factors pertain to individual-level data, which is not feasible when considering the overall population. Our focus was on factors that encompass the entire population rather than specific individuals.

In order to identify the most influential factors, we conducted correlation analyses between the risk factors and breast cancer (BC) incidence. It is worth noting that some of these factors lack consensus, and our study aims to establish their significance. Additionally, to forecast the future incidence of breast cancer on a country-by-country basis, we rigorously tested and refined six regression models.

2.1 Data collection

2.1.1 Risk factors identification. In order to identify recent BC risk factors, we have conducted a systematic literature review on the related Meta-Analysis and Systematic Reviews published in 2021. Following the PRISMA protocol [15], we started the search process on February 2, 2022 by considering four digital libraries: Scopus, PubMed, Web of Science and Cochrane. The MeSH keyword used for the automatic search in the mentioned digital sources is “Breast Neoplasms”. The search strings used are: “Breast Neoplasms” AND “risk factors”. Furthermore, we have removed duplicates then we applied the following inclusion/exclusion criteria: *Inclusion criteria (IC)*

- IC: publications that include breast cancer non-genetic risk factors

Exclusion criteria (EC)

- EC1: studies not related to breast cancer and risk factors
- EC2: studies dealing with male breast cancer
- EC3: research articles that concern risk factors of breast cancer recurrence
- EC4: publications that treat genetic factors or family antecedent

2.1.2 Construction of the dataset. In order to create our dataset, we gathered reported BC incidence rates dating back to 2018 in several countries, as well as available data reflecting the risk and preventive factors we were able to access. We extracted reported incidence rates of breast cancer (rates per 100 000 females in 2018, age between 0 and 84) from the Global Cancer Observatory [16], the average age of women per country from WorldData [17], the prevalence of insufficient physical activity, depression, overweight, obesity, BMI, alcohol consumption and smoking rates from the World Health Organization [18], CO2 emissions from the International Energy Agency [19], breastfeeding, contraception use and world fertility data from United Nations [20]. We extracted food supply quantities of meat, sugar and milk from the Food and Agriculture Organization [21], mobile cells from the OpenCellID [22] and surface area from the World Bank Open Data [23] for data normalization. Notably, we could not find data on the consumption of sugar-sweetened beverages, so we considered data on white sugar consumption by country since sugar is also known to be a risk factor for breast cancer [24].

2.2 Breast cancer incidence

Prior to conducting the analysis, we addressed missing values in our dataset by imputation. Specifically, we utilized the k-Nearest Neighbor algorithm (KNN) [25]. The KNN algorithm was applied individually to each feature with missing values, considering all other available features as input. During the imputation process, the KNN algorithm iterates through the dataset to identify "k" similar or closely related examples, also known as neighbors, based on spatial proximity. For each example with missing values, the algorithm imputes those missing values with the mean value derived from its k-neighbors. This method has been shown to effectively handle missing data and maintain the integrity of the dataset [25].

To gain a more comprehensive understanding of breast cancer incidence across countries, we categorized them into quartiles, which are subgroups that divide the countries based on their breast cancer incidence rates into four equal parts. Each quartile represents 25% of the total number of countries and is determined by the combination of three values that serve as thresholds for this division. This approach allows for a more nuanced assessment of how countries rank in terms of breast cancer incidence.

2.3 Correlation between breast cancer incidence rates and the studied factors

2.3.1 Data normality assessment. To evaluate the normality of the distribution of each risk and preventive factor against the incidence of breast cancer, we employed both statistical and graphical methods. For each factor, we conducted the Shapiro-Wilk test [26] which is a statistical test that determines whether a sample comes from a normally distributed population. A p-value greater than 0.05 in the Shapiro-Wilk test indicates that the data do not significantly deviate from normality, suggesting a normal distribution. To support our findings, we also generated histograms and Q-Q (Quantile-Quantile) plots.

2.3.2 Correlation analysis. Given the results of the normality assessments, we proceeded with the appropriate correlation analysis. The presence of non-normally distributed data necessitated the use of a non-parametric correlation test. Therefore, we opted for Spearman's rank correlation coefficient [27], which measures the strength and direction of the association between two variables. For each risk and preventive factor, we calculated Spearman's rank correlation coefficient with the incidence of breast cancer. This approach allowed us to identify significant associations while accounting for the non-normal distribution of our data.

2.4 Breast cancer incidence rate prediction

One of the purposes of this study is to anticipate breast cancer future incidence rate in each country using machine learning, specifically a regression model. In this section, we go through the prediction pipeline in great detail.

2.4.1 Regression models. We have split the data into two sets, train with 67% and test with 33%. Then, we tested the following regression models:

- Linear Regression [28]
- Support Vector Regression (SVR) [29]
- Random Forest [30]

- 3 Gradient Boosting models based on decision trees: Catboost [31], XGBoost [32] and LightGBM [33]

Linear Regression [28] and Random Forest [30] are two fundamental regression models and boosting algorithms [34] are a class of ensemble learning where models are built sequentially so that each improves the error of the previous model. Catboost (Category Boosting) [31] applies gradient boosting on decision trees and achieve satisfactory results with no required parameter tuning. XGBoost (eXtreme Gradient Boosting) [32] is an ensemble model also based on parallel decision trees. By combining results from a set of simple and weak models, it generates a more accurate prediction. LightGBM (Light Gradient Boosted Machine) [33] is a gradient boosting framework that supports distributed learning, has a fast training phase and low memory usage and it provides efficient results. Finally, for SVR (Support Vector Regression) [29], typically Support Vector Machine (SVM) that is based on statistical learning theory can also perform as a regression method. Unlike classical regression models, where error is minimized by selecting the best hyperplane of fit, SVR sets a threshold error allowance around the regression hyperplane, ensuring that all data points within the threshold are not penalized.

2.4.2 Evaluation metrics. In order to evaluate the models, we used:

- R squared (R^2)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)

R-squared (R^2) measures how closely the data points match the fitted regression hyperplane. Mean squared Error (MSE) is an absolute measure of the quality of the fit. It considers the sum of squares of error. Consequently, the square root of MSE is the Root Mean Squared Error (RMSE). Lastly, Mean Absolute Error (MAE) is similar to Mean Squared Error (MSE), however, it is calculated by adding the absolute values of error. MAE is a more straightforward depiction of the sum of error terms than MSE or RMSE. MSE penalizes high prediction errors by squaring them, whereas MAE handles all errors similarly.

2.4.3 Features selection. We applied a Forward Feature Selection with cross-validation to each of the six regression models. Forward Feature Selection is an iterative approach for selecting features that resulted in good model performance. It starts with 0 features in the model and adds at each iteration the feature that best improves the model until adding a new variable does not enhance the model's performance. Cross-validation (CV) is used to test the performance of a machine learning model and it is an efficient method when the data is limited since it generates different test sets. Afterward, we provided a comparison between the 6 models in terms of the above mentioned evaluation metrics. Finally, we selected the best regression model and retrained it on the selected features 50 times with different random splits of the training/test sets. The final result is the average performance over the 50 runs with their standard deviation.

3 Results

3.1 Data collection

We have identified 1603 publications and then removed 342 duplicates. We conducted an analysis by inspecting each article's title, abstract, and keywords. We identified a total of 56 studies after application of inclusion and exclusion criteria as presented in Fig 1. Accordingly, based on the full text analysis of the selected articles, we have categorized the non-genetic risk factors as presented in Table 1.

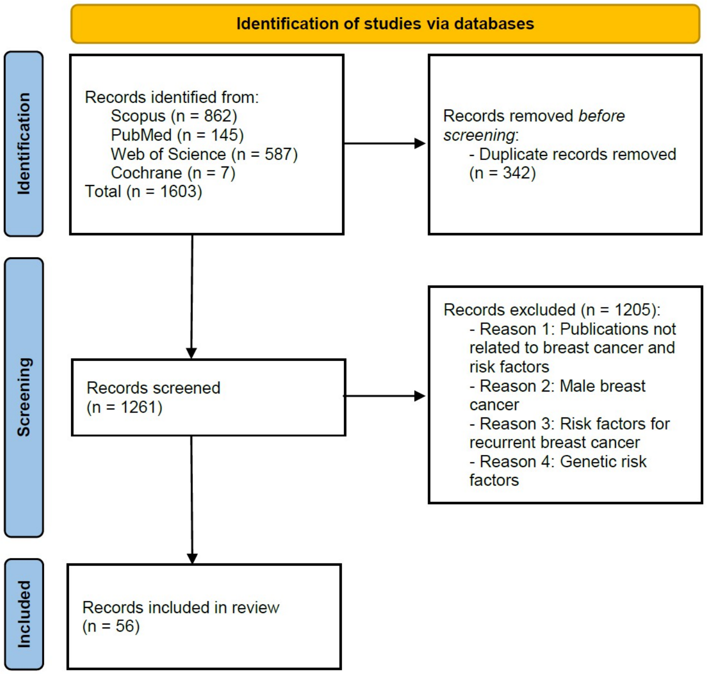
Identification of studies via databases
Identification
Screening
Included
Records identified from:
Scopus (n = 862)
PubMed (n = 145)
Web of Science (n = 587)
Cochrane (n = 7)
Total (n = 1603)
Records removed *before screening*:
- Duplicate records removed (n = 342)
Records screened (n = 1261)
Records excluded (n = 1205):
- Reason 1: Publications not related to breast cancer and risk factors
- Reason 2: Male breast cancer
- Reason 3: Risk factors for recurrent breast cancer
- Reason 4: Genetic risk factors
Records included in review (n = 56)

Fig 1. PRISMA flow diagram for articles' screening and identification.

https://doi.org/10.1371/journal.pone.0308905.g001

In order to create the dataset, we collected reported BC incidence rates dating back to 2018 across 151 countries, along with data on 13 risk factors (available data on factors from Table 1) in addition to 2 preventive factors (breastfeeding and total fertility rates). We refer the reader to Appendix 1 for details about the extracted risk and preventive factors.

3.2 Breast cancer incidence

Based on breast cancer incidence rates reported in 2018 by the Global Cancer Observatory [16], our interpretation revealed that countries with very high levels of breast cancer incidence rates (ranging from 109.5 to 200.7) predominantly include the United States, Canada, Japan, Australia, and most European countries.

The second category with high level (48.5–109.5) includes Russian Federation, China, some North African countries (Morocco, Algeria and Tunisia), Indonesia, South Africa, two Eastern European countries (Belarus and Ukraine), Turkey in addition to Latin America except Mexico, Peru and Ecuador which belong to the third category.

The latter which is of medium incidence rate (24.60–48.5) encompasses mainly some Central Asian countries, India, the vast majority of Gulf countries, some Sub-Saharan African countries, namely Ethiopia, Kenya, Gabon, Cameroon, Nigeria and Namibia.

The last category with low incidence rate (4.4–24.60) aggregates the rest of Sub-Saharan African countries in addition to Pakistan, Afghanistan, Mongolia and Maldives.

3.3 Correlation between breast cancer incidence rates and the studied factors

3.3.1 Data normality assessment. We conducted the Shapiro-Wilk test for each factor. The results of the test are displayed in Table 2. We also generated histograms and Q-Q plots (see Appendix 2) to visually inspect the data distribution for any deviations from normality.

Table 1. Categorization of breast cancer risk factors identified from recent biomedical literature (2021).

Risk factor category	Risk factor designation	Reference
Hormonal factors	Oral contraceptives	[35–38]
	Menopausal hormone therapy	[39]
	Repercussions of melatonin	[40]
Reproductive factors	Early menarche	[41, 42]
Health and physical conditions	Insufficient physical activity	[43, 44]
	Long-term weight gain	[45]
	Obesity	[46–48]
	Chronic inflammation	[49]
	Type 2 diabetes	[48, 50]
	Vitamin D deficiency	[51]
	Breast skin microbiota	[52]
	Bariatric surgery	[53]
	Preeclampsia	[54]
	Thyroid disease	[55]
	Sleep-disordered breathing	[56]
Psychological Factors	Bipolar disorder	[57]
	Stress	[58, 59]
	Trauma, grief and depression	[60]
Lifestyle	Smoking	[61]
	Alcohol consumption	[62]
	Sedentary work	[63, 64]
	Night-shift work	[65, 66]
	Excessive Smartphone use	[67]
Eating habits	Pickled foods	[68]
	Consumption of red and processed meat	[69–71]
	Intake of Isoflavones	[72]
	Ultraprocessed food intake	[73]
	Consumption of sugar-sweetened beverages	[74, 75]
	Pro-inflammatory diet	[76, 77]
	Milk consumption	[69]
Medications	Aspirine use	[78, 79]
	Antihypertensive medication use	[80]
External exposures	Light at night exposure	[81, 82]
	Exposure to Polychlorinated Biphenyls	[83]
	Exposure to Endocrine disruptors	[84]
	Ambient air pollution exposure	[85]
	Pesticide exposure	[86]
	Hair chemicals	[87]
	Exposure to Polycyclic Aromatic Hydrocarbons	[88]
Demographic factors	Age (advancing age)	[89]
	Low socioeconomic status	[90]

https://doi.org/10.1371/journal.pone.0308905.t001

The histograms and Q-Q plots corroborate the results of the Shapiro-Wilk test, providing a consistent evaluation of data normality. Only depression and insufficient physical activity were normally distributed (P-values of 0.883 and 0.234 respectively according to Table 2). Therefore, we opted to use the Spearman's correlation test to assess the correlation between the factors and breast cancer incidence rates.

Table 2. Normality test (Shapiro-Wilk) for each factor.

Factor	Statistic	P-value	Normal Distribution
CO2 emissions	0.73	< 0.05	False
Pesticides	0.42	< 0.05	False
Average age	0.94	< 0.05	False
Depression	0.99	0.883	True
Mobile cells	0.61	< 0.05	False
Alcohol consumption	0.96	< 0.05	False
Tobacco consumption	0.79	< 0.05	False
Sugar consumption	0.88	< 0.05	False
Meat consumption	0.94	< 0.05	False
Milk consumption	0.88	< 0.05	False
Obesity	0.97	< 0.05	False
Insufficient physical activity	0.98	0.234	True
Total Fertility Rate	0.89	< 0.05	False
Breastfeeding rate	0.96	< 0.05	False
Oral contraceptives	0.86	< 0.05	False

https://doi.org/10.1371/journal.pone.0308905.t002

3.3.2 Correlation analysis. Table 3 displays the results of the correlation test. The last column indicates the interpretation of Spearman's correlation coefficient ρ. Six interpretations are possible [91]:

- Perfect association: $|\rho| = 1$
- Very strong: $0.80 \leq |\rho| < 1$
- Moderate: $0.60 \leq |\rho| < 0.80$
- Fair: $0.30 \leq |\rho| < 0.60$
- Poor: $0 < |\rho| < 0.30$
- No association: $\rho = 0$

Table 3. Results of the Spearman's correlation test (correlation between BC incidence rate and each factor).

Factor	Correlation	P-value	Interpretation
Average age	0.88	< 0.05	Very strong
Meat consumption	0.79	< 0.05	Moderate
CO2 emissions	0.71	< 0.05	Moderate
Depression	0.71	< 0.05	Moderate
Sugar consumption	0.71	< 0.05	Moderate
Tobacco consumption	0.69	< 0.05	Moderate
Milk consumption	0.64	< 0.05	Moderate
Mobile cells	0.63	< 0.05	Moderate
Alcohol consumption	0.57	< 0.05	Fair
Pesticides	0.56	< 0.05	Fair
Oral contraceptives	0.52	< 0.05	Fair
Insufficient physical activity	0.40	< 0.05	Fair
Obesity	0.36	< 0.05	Fair
Total Fertility Rate	-0.86	< 0.05	Very strong
Breastfeeding rate	-0.60	< 0.05	Moderate

https://doi.org/10.1371/journal.pone.0308905.t003

All the factors are significantly associated with the incidence of breast cancer ($P - value < 0.05$) (Table 3). Concerning risk factors, average age is very strongly associated with BC incidence (correlations of 0.88). Meat consumption, CO2 emissions, depression, sugar intake, tobacco use, milk consumption, and mobile cells show moderate correlations with the occurrence of breast cancer (correlation of 0.79, 0.71, 0.71, 0.71, 0.69, 0.64 and 0.63 respectively). Alcohol consumption, pesticides, oral contraceptives, lack of physical activity, and obesity are all fairly linked to the incidence of the disease (0.57, 0.56, 0.52, 0.40 and 0.36 respectively). In terms of preventive factors, breast cancer incidence shows a very strong negative correlation with fertility (correlation of -0.86) and a moderate negative correlation with breastfeeding (correlation of -0.60). It makes sense that higher levels of risk factors are linked to an increased incidence of breast cancer, while higher levels of preventive factors are associated with a reduced incidence.

3.4 Breast cancer incidence rate prediction

Table 4 presents the results of prediction: comparing the performance of the 6 machine learning models in terms of R squared, RMSE and MAE.

CatBoost Regressor achieves an R squared score of **0.84 ± 0.03** in predicting BC incidence rate, making it the best performing model against the other 5 regressors. Such value of R squared is considered high and the prediction is accurate as it shows that the model can predict well the incidence rate value for a given country. The result is confirmed by the RMSE and MAE metrics which recorded smaller values (20.39 ± 2.26 and 14.99 ± 1.62 respectively) for CatBoost Regressor compared to the 5 remaining models.

The Forward Feature Selection enabled the selection of 13 variables from a total of 15. The selected features are: CO2 emissions, pesticides, average age, mobile cells, tobacco consumption, sugar, milk and meat consumption, insufficient physical activity, obesity, breastfeeding, total fertility rate and oral contraceptives.

The feature importance is computed using the impurity-based feature importance. Fig 2 represents the importance of each risk or preventive factor associated with breast cancer. The greater the score, the more important the feature. The importance of a feature is known as the Gini importance.

4 Discussion

Our study showed that age is the most correlated factor with breast cancer incidence (correlation of 0.88 according to Table 3). Other studies have confirmed that age is the most predominant risk factor for breast cancer [92], and the incidence rate increases significantly with age in different countries [93, 94]. Yet, younger women are not spared, for the latter, breast tumors are likely to appear in more complicated states, with positive lymph nodes, larger size and weaker survival [95].

Table 4. Evaluation results of regression models.

Regressor	Number of selected features	R squared	RMSE	MAE
CatBoost Regressor	13	**0.84 ± 0.03**	**20.39 ± 2.26**	**14.99 ± 1.62**
Light GBM Regressor	10	0.83 ± 0.04	21.24 ± 2.32	15.76 ± 1.66
Random Forest	10	0.83 ± 0.04	21.05 ± 2.55	15.23 ± 1.88
SVR	8	0.82 ± 0.03	22.11 ± 2.50	15.82 ± 1.70
XGBoost Regressor	15	0.81 ± 0.05	22.28 ± 3.06	16.06 ± 2.14
Linear Regression	2	0.78 ± 0.03	24.49 ± 2.31	19.27 ± 1.70

https://doi.org/10.1371/journal.pone.0308905.t004

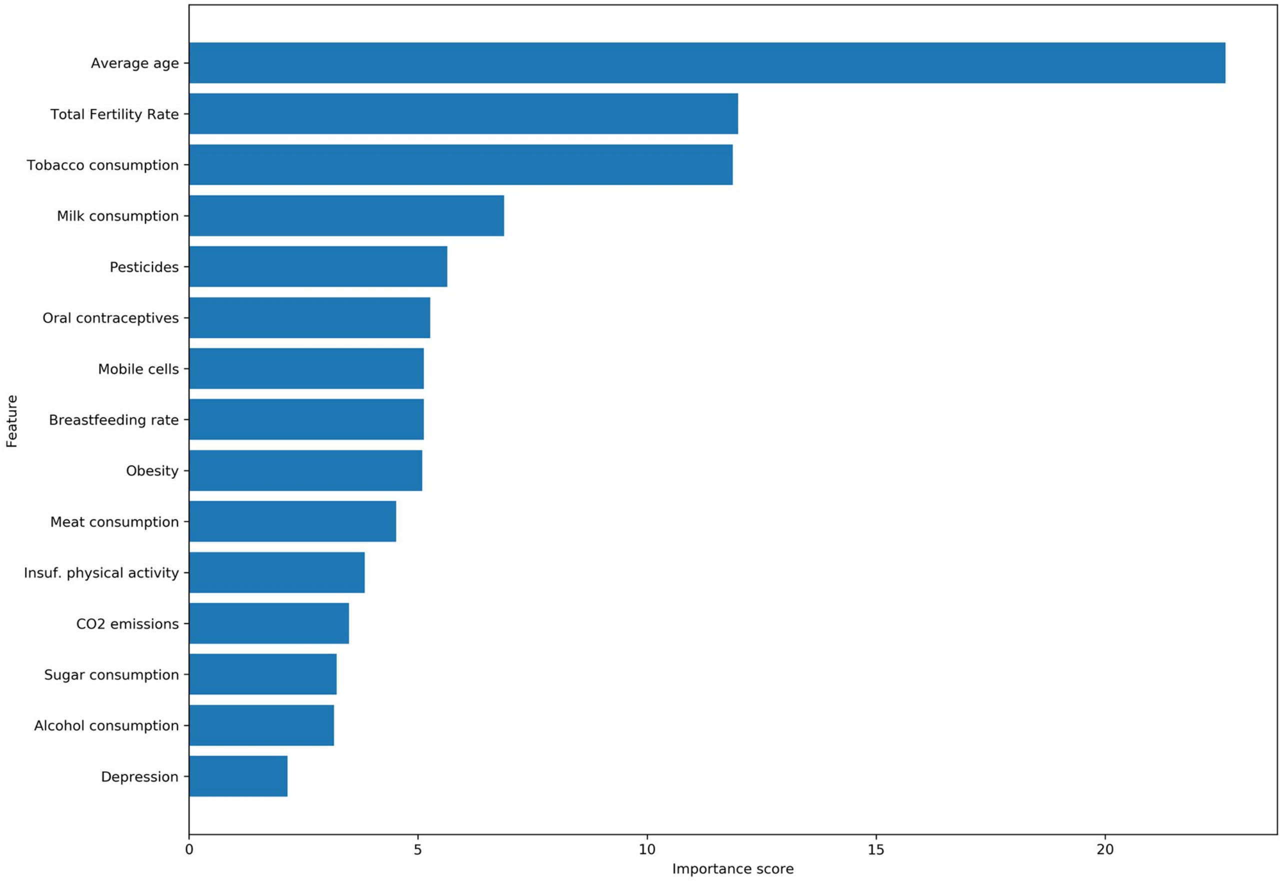

Fig 2. Feature importance using CatBoost.

https://doi.org/10.1371/journal.pone.0308905.g002

According to our study, meat consumption is the most influencing eating habit positively correlated with the incidence of the disease (correlation of 0.79 according to Table 3). According to the International Agency for Research on Cancer (IARC), meat consumption can range from a few percent to 100% depending on the country [96]. In a recent research conducted using the Sister Study cohort [97], authors examined different types of meat and breast cancer incidence. They found that red meat consumption may increase the risk of invasive breast cancer [98]. The same conclusion has been drawn by several meta-analyses and systematic reviews insistently on processed meat [68–70].

Important correlation was observed regarding the association of breast cancer with CO2 emissions, depression and sugar consumption (0.71 according to Table 3).

A meta-analysis of 18 epidemiological studies indicates that exposure to ambient air pollution may have a significant impact on breast cancer development. Pooled analysis found that nitrogen dioxide exposure increases breast cancer risk [85]. As far as air particles are concerned, particulate matter PM2.5 and PM10 did not have significant associations with BC risk. Authors argue that further studies, particularly in developing countries, are needed to draw a firm conclusion of causality [85].

A recent retrospective cohort study [99] highlights a significant connection between depression and an increased risk of cancer, particularly breast cancer. The study demonstrated that individuals with depression showed an 18% overall increase in cancer diagnosis risk. This risk was most pronounced for breast cancer, with a hazard ratio (HR) of 1.23 (95% CI: 1.12–1.35, $p < 0.0001$) [99]. Additionally, we found that depression prevalence in a country as a psychiatric morbidity is associated with BC incidence. A recent systematic review investigating the impact of psychological factors on breast cancer risk has found that from twenty studies published between 1988 and 2020, only five reported a significant association between depression and BC incidence [60]. The authors emphasized that psychological factors deserve further investigation.

As for sugar consumption, the results of a study involving participants from a large prospective cohort confirmed our finding. Higher sugar intake with its different types has been associated with an increased risk of breast cancer [24]. Notwithstanding, the authors claimed that experimental data are needed to clarify the mechanisms behind these findings. Equally important, consumption of sugar-sweetened beverages may also increase the risk of breast cancer [74, 75].

Conforming to our study, tobacco consumption is the first lifestyle habit positively correlated with BC incidence (correlation of 0.69 according to Table 3). In this regard, a Mendelian Randomization was applied to investigate whether there is a causality between smoking and breast cancer risk [100]. The findings showed a possible causal relationship between lifelong smoking exposure and BC risk.

As for milk consumption, the correlation was 0.64 (Table 3). A study evaluating the association between dairy and soy milk consumption and BC risk joined us in concluding that higher consumption of dairy milk was linked to an increased risk of breast cancer [101]. In spite of this finding, a meta-analysis of 8 studies does not provide significant evidence that milk consumption is associated with breast cancer risk [102]. Accordingly, milk consumption is not consistently linked to breast cancer. To make a decision about this risk factor, further research is needed.

As for mobile cells, we considered the number of mobile cells per country divided by country area. We intended to evaluate the association between radio-frequency radiation exposure and BC prevalence. We found a moderate association with BC prevalence (correlation of 0.63 according to Table 3). This association was not examined in recent literature. Though, a Taiwanese case-control study has assessed the link between the behavior of Smartphone users and breast cancer risk (addiction, use before sleep, closeness to breast, etc) [103]. According to the study, excessive smartphone use increases the risk of breast cancer significantly.

Our results indicate that alcohol consumption is also positively correlated with BC incidence (correlation of 0.57 according to Table 3). The dose-response relationship between different types of alcohol and breast cancer risk was assessed in a meta-analysis [104]. The authors found a significant association between total drinking and breast cancer risk. The latter gradually increased with alcohol consumption, especially among postmenopausal women regardless of the kind of alcohol consumed.

Concerning pesticides, these are considered as endocrine disruptors likely to alter hormonal activity or cause epigenetic damage [105]. Our results indicate a correlation of 0.56 between BC incidence and pesticides (Table 3). A systematic review of 63 studies published between 1960 and 2019 showed that 62% indicated an association between pesticide exposure and BC, while 38% indicated the opposite. According to the authors, exposure to some types of pesticides may increase the risk of breast cancer [86].

Our study has found a fair link between oral contraceptives and breast cancer (0.52 according to Table 3), which is consistent with the findings of many recent studies [35–38]. The latter are all meta-analyses of case-control studies investigating the link between oral contraceptives and breast cancer. It was established that taking an oral contraceptive pill was linked to a considerably higher risk of breast cancer in totality. It's important to note that the type of oral contraceptive, durability, dosage and age of starting use all have a part in this association.

Although obesity and insufficient physical activity are established risk factors for breast cancer [43, 44, 46–48], we could not uncover a strong link between these factors and the incidence of the disease since we found correlation of 0.36 and 0.40 respectively according to Table 3. Our result was also confirmed in a study rating BC risk factors [106]. Authors have demonstrated that obesity presents only a relatively modest risk for breast cancer, however, the risk of breast cancer is significantly increased by factors such as a genetic predisposition to the disease, a history of atypical hyperplasia and a history of neoplastic disease [106].

Regarding preventive factors, we observed a strong negative correlation between the total fertility rate and increased breast cancer risk (correlation of -0.86 according to Table 3). A prospective study conducted in Burkina Faso has provided supporting results [107]. The authors showed that, when comparing multiparous women to their non-multiparous counterparts, it was reported that multiparity decreases the risk of breast cancer [107]. This association is interesting and may be the subject of several observational studies to confirm or invalidate this point in other countries.

We also found that breastfeeding rate is moderately correlated with BC incidence (correlation of -0.60 according to Table 3). A systematic review and meta-analysis of studies published in the period of 1998–2021 uncovered a strong correlation between breastfeeding and risk of breast cancer [108]. The duration of breastfeeding was especially found to reduce BC risk [108].

A comparison between our findings and those of recent literature is given in Table 5. The table also provides some recommendations to reduce modifiable risk factors while emphasizing factors that need more research to determine their association with BC incidence.

In our view, there is a need for counseling about lifestyle habits (smoking and alcohol intake), as well as education about eating habits especially for those prone to other breast cancer risk factors, such as a family history of the disease. In fact, the most frequent risk factors responsible for BC onset are hereditary and genetic, such as breast cancer or ovarian cancer history and inherited mutations, in particular BRCA1 and BRCA2 [109].

As for the prediction, building an accurate model would highlight countries likely to register high incidence rates in the upcoming years. In terms of predictive ability, the CatBoost model yielded the best performance. It provided the most precise results when predicting breast cancer rates. In fact, a more precise regression is the one with a relatively high R squared (close to 1). For CatBoost Regressor, the average R squared is 0.84 ± 0.03.

CatBoost model was also used to identify factors that had a greater impact on breast cancer incidence prediction. Fig 2 shows that average age is the most significant predictive variable of breast cancer incidence. This is evident since the prevalence of BC increases after the age of 40 [93, 94], and a population with a median age higher than 40 is more likely to register a high incidence rate. Age is followed by total fertility rate and tobacco consumption (Fig 2).

4.1 Limitations and future directions

Our research has few limitations. First, given the lack of statistics regarding breast cancer, data were used from the majority of the developed countries but there was a lack of data concerning the United Kingdom. Also data from some Latin, Asian and African countries were not available. Nevertheless, this bias is frequently found in exploratory studies and can only be overcome if countries are more committed to communicating their data. Second, we did not consider genetic factors shared among individuals within the same population since related studies are heterogeneous and the only way would have been to do a meta-analysis on real aggregated data, which was not the scope of our study.

Table 5. Comparison between our findings and those of recent literature and some recommendations to reduce modifiable risk factors.

Factor	Association assessed by other studies (at individual level)	Association assessed by our study (at country level)	Recommendations
Average age	Positive association [92–95]	Very strong association	Unfortunately, age is not a modifiable factor. However, we can suggest strengthening the screening in the countries with high levels of BC incidence
Tobacco consumption	Positive association [100]	Moderate positive association	To quit smoking, we propose to strengthen psychological therapies for assisting women by encouraging them to participate in free smoking cessation programs
Meat consumption	Positive association [68–70, 98]	Moderate positive association	These findings bolster public health recommendations to reduce meat consumption
Milk consumption	• Positive association [101] • No significant association [102]	Moderate positive association	More studies are needed on milk consumption to come to a definite conclusion
Depression	Positive association [60, 99]	Moderate positive association	Screening for depression needs to become more sensitive
Alcohol consumption	Positive association [104]	Fair positive association	We join the 2020–2025 Dietary Guidelines for Americans. It is recommended that adults of legal drinking age may choose to abstain from alcohol consumption in order to reduce the risk of alcohol-related harms
Sugar consumption	Positive association [24]	Fair positive association	We insist on the American Heart Association recommendations. It suggests that, for women, added sugar should not exceed 100 calories a day (approximately 6 teaspoons) and also the consumption of sugar-sweetened beverages should be limited
Mobile cells	Positive association [103]	Fair positive association	The association between radiofrequency radiation exposure and BC incidence should receive more attention in future studies so that we can make recommendations accordingly
Oral contraceptives	Positive association [35–38]	Fair positive association	Generally, it may be beneficial to organize campaigns demonstrating the direct link between oral contraceptive use and breast cancer risk. Individually, every woman needs to discuss her contraception options with her physician.
CO2 emissions	Positive association [85]	Fair positive association	A recent literature review on ways to reduce carbon emissions from supply chains demonstrated the importance of coordinating with various means to reduce gas emissions. For instance how energy consumption is structured, production processes, and the optimal level of carbon emissions [110]
Pesticides	Positive association [86]	Poor positive association	Local fruits and vegetables are encouraged and they must be washed thoroughly before consumption.
Obesity	• Positive association [46–48] • No significant association [106]	Poor positive association	Even if we had not figure out a strong link between obesity and BC incidence, we recommend avoiding foods with high fat content and added sugar
Insufficient physical activity	Positive association [43, 44]	Poor positive association	Our study found a weak link between insufficient physical activity and the prevalence of BC, yet we emphasize the current physical activity guidelines for Americans. In fact, women need 150 minutes of moderately intense exercise per week

https://doi.org/10.1371/journal.pone.0308905.t005

As a future direction, breast cancer incidence rates will be retrieved for various years and we will examine trends in breast cancer prevalence and related risk factors resulting from the present study. Our intention is to explore whether the changes in risk factors have influenced BC incidence rates.

5 Conclusion

Breast cancer is a complex disease influenced by multiple factors. Through our statistical analysis, we have identified several noteworthy associations. Specifically, we observed a significant positive correlation between the incidence of breast cancer in a country and the average age of women within that country, CO2 emissions, pesticides, depression rates, lifestyle, eating habits, mobile cells, and the use of oral contraceptives. Conversely, preventive factors such as breastfeeding rates and total fertility rates in a given country displayed a negative association with breast cancer prevalence. These findings informed the inclusion of these features in our breast cancer prediction model at the country level.

The prediction task was formulated as a regression problem, aiming to estimate the incidence rate of breast cancer. Our model demonstrated strong performance, achieving a mean R-squared score of 0.84 ± 0.03, underscoring the predictive power of our approach and the robustness of the regression model.

While this work provides valuable insights into predicting future breast cancer incidence rates and understanding major risk factors, it is important to interpret the results with caution due to the ecologic study design. Our findings are based on group-level data, and as such, cannot be directly translated to individual risk factors or causality. Nonetheless, this study highlights key areas for public health interventions and offers targeted recommendations for modifying certain modifiable factors, particularly lifestyle choices and eating habits. By addressing these factors at the population level, there is potential to make a substantial impact on reducing the burden of breast cancer.

Supporting information

S1 Appendix.
(PDF)

Acknowledgments

We thank Othmane Cherqi for creating links between the National Institute of Oncology and the International University of Rabat in Morocco to carry out this work.

Author Contributions

Conceptualization: Hasna EL HAJI, Nada SBIHI, Bassma GUERMAH, Mounir GHOGHO.

Data curation: Hasna EL HAJI, Nada SBIHI.

Formal analysis: Hasna EL HAJI, Nada SBIHI, Amine SOUADKA.

Funding acquisition: Hasna EL HAJI.

Investigation: Hasna EL HAJI, Amine SOUADKA.

Methodology: Hasna EL HAJI, Nada SBIHI, Amine SOUADKA, Mounir GHOGHO.

Project administration: Hasna EL HAJI, Bassma GUERMAH, Amine SOUADKA, Mounir GHOGHO.

Software: Hasna EL HAJI.

Supervision: Nada SBIHI, Amine SOUADKA.

Validation: Bassma GUERMAH, Amine SOUADKA, Mounir GHOGHO.

Writing – original draft: Hasna EL HAJI, Nada SBIHI, Bassma GUERMAH.

Writing – review & editing: Hasna EL HAJI, Amine SOUADKA, Mounir GHOGHO.

References

1. Weir H. K., Anderson R. N., King S. M. C., Soman A., Thompson T. D., Hong Y., et al. (2016). Peer reviewed: heart disease and cancer deaths—trends and projections in the United States, 1969–2020. Preventing chronic disease, 13. https://doi.org/10.5888/pcd13.160211

2. "Breast cancer." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer. [Accessed: 15-Mar-2022].
3. Report of the Societal Cancer Observatory, National Cancer League, 2013.
4. Robertson, R., Gregory, S., & Jabbal, J. (2014). The social care and health systems of nine countries. Commission on the future of health and social care in England.: The King's Fund.
5. Souadka A., Houmada A., & Souadka A. (2021). Multidisciplinary team meeting as a highly recommended EUSOMA criteria evaluating the quality of breast cancer management between centers. Breast (Edinburgh, Scotland), 60, 310. https://doi.org/10.1016/j.breast.2021.11.001 PMID: 34764006
6. Knight S. R., Shaw C. A., Pius R., Drake T. M., Norman L., Ademuyiwa A. O., et al. (2021). Global variation in postoperative mortality and complications after cancer surgery: a multicentre, prospective cohort study in 82 countries. The Lancet, 397(10272), 387–397. https://doi.org/10.1016/S0140-6736(21)00001-5
7. Souadka A., Benkabbou A., Al Ahmadi B., Boutayeb S., & Majbar M. A. (2020). Preparing African anticancer centres in the COVID-19 outbreak. The lancet oncology, 21(5), e237. https://doi.org/10.1016/S1470-2045(20)30216-3 PMID: 32251622
8. Souadka A., Essangri H., Benkabbou A., Amrani L., & Majbar M. A. (2020). COVID-19 and Healthcare worker's families: behind the scenes of frontline response. EClinicalMedicine, 23. https://doi.org/10.1016/j.eclinm.2020.100373 PMID: 32368726
9. Del Carmen O. J. M., Emilia G. R. D., Mares B. H., & Marcela O. J. (2021). Educational interventions on breast cancer in men and women: a necessity in primary healthcare. ecancermedicalscience, 15. https://doi.org/10.3332/ecancer.2021.1255 PMID: 34267811
10. Valipour A. A., Mohammadian M., Ghafari M., & Mohammadian-Hafshejani A. (2017). Predict the future incidence and mortality of breast cancer in Iran from 2012-2035. Iranian journal of public health, 46(4), 579–580. PMID: 28540280
11. Zaheer S., Shah N., Maqbool S. A., & Soomro N. M. (2019). Estimates of past and future time trends in age-specific breast cancer incidence among women in Karachi, Pakistan: 2004-2025. BMC public health, 19(1), 1001. https://doi.org/10.1186/s12889-019-7330-z PMID: 31345204
12. Katayama K., & Narimatsu H. (2016). Prediction of female breast cancer incidence among the aging society in Kanagawa, Japan. PloS one, 11(8), e0159913. https://doi.org/10.1371/journal.pone.0159913 PMID: 27532126
13. Chowdhury R., Sinha B., Sankar M. J., Taneja S., Bhandari N., Rollins N., et al. (2015). Breastfeeding and maternal health outcomes: a systematic review and meta-analysis. Acta paediatrica (Oslo, Norway: 1992), 104(467), 96–113. https://doi.org/10.1111/apa.13102 PMID: 26172878
14. Knottnerus A., & Tugwell P. (2008). STROBE–a checklist to Strengthen the Reporting of Observational Studies in Epidemiology. Journal of clinical epidemiology, 61(4), 323. https://doi.org/10.1016/j.jclinepi.2007.11.006 PMID: 18313555
15. Page M. J., McKenzie J. E., Bossuyt P. M., Boutron I., Hoffmann T. C., Mulrow C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Bmj, 372.
16. "Global Cancer Observatory." [Online]. Available: https://gco.iarc.fr/. [Accessed: 15-Mar-2022].
17. "Average age by country." [Online]. Available: https://www.worlddata.info/average-age.php. [Accessed: 15-Mar-2022].
18. "Indicators." [Online]. Available: https://www.who.int/data/gho/data/indicators. [Accessed: 15-Mar-2022].
19. "IEA—International Energy Agency—IEA." [Online]. Available: https://www.iea.org/data-and-statistics/data-products?filter=emissions. [Accessed: 15-Mar-2022].
20. "World Contraceptive Use _ Population Division." [Online]. Available: https://www.un.org/development/desa/pd/data/world-contraceptive-use. [Accessed: 15-Mar-2022].
21. "FAOSTAT." [Online]. Available: https://www.fao.org/faostat/en/#home. [Accessed: 15-Mar-2022].
22. "OpenCelliD—Largest Open Database of Cell Towers & Geolocation—by Unwired Labs." [Online]. Available: https://www.opencellid.org/#zoom=16&lat=37.77889&lon=-122.41942. [Accessed: 15-Mar-2022].
23. "Surface area (sq. km) | Data." [Online]. Available: https://data.worldbank.org/indicator/AG.SRF.TOTL.K2/. [Accessed: 15-Mar-2022].
24. Debras C., Chazelas E., Srour B., Kesse-Guyot E., Julia C., Zelek L., et al. (2020). Total and added sugar intakes, sugar types, and cancer risk: results from the prospective NutriNet-Santé cohort. The American journal of clinical nutrition, 112(5), 1267–1279. https://doi.org/10.1093/ajcn/nqaa246 PMID: 32936868
25. Batista, G. E., & Monard, M. C. (2002). A study of K-nearest neighbour as an imputation method. International Conference on Health Information Science.

26. Shapiro S. S., & Wilk M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3/4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

27. Zar J. H. (1972). Significance testing of the Spearman rank correlation coefficient. Journal of the American Statistical Association, 67(339), 578–580. https://doi.org/10.1080/01621459.1972.10481251

28. Seber G. A., & Lee A. J. (2012). Linear regression analysis. John Wiley & Sons.

29. Smola A. J., & Schölkopf B. (2004). A tutorial on support vector regression. Statistics and computing, 14, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

30. Breiman L. (2001). Random forests. Machine learning, 45, 5–32. https://doi.org/10.1023/A:1017934522171

31. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

32. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

33. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

34. Freund Y., & Schapire R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

35. Ammembal A. M. K., & Udupa K. (2021). Combined Oral Contraceptives and Breast Cancer: an Unsolved Conundrum. Indian Journal of Gynecologic Oncology, 19(4), 1–5. https://doi.org/10.1007/s40944-021-00561-5

36. Barańska A., Błaszczuk A., Kanadys W., Malm M., Drop K., & Polz-Dacewicz M. (2021). Oral Contraceptive Use and Breast Cancer Risk Assessment: A Systematic Review and Meta-Analysis of Case-Control Studies, 2009–2020. Cancers, 13(22), 5654. https://doi.org/10.3390/cancers13225654 PMID: 34830807

37. Bonfiglio R., & Di Pietro M. L. (2021). The impact of oral contraceptive use on breast cancer risk: State of the art and future perspectives in the era of 4P medicine. In Seminars in Cancer Biology (Vol. 72, pp. 11–18). Academic Press.

38. Kanadys W., Baranska A., Malm M., Blaszczuk A., Polz-Dacewicz M., Janiszewska M., et al. (2021). Use of oral contraceptives as a potential risk factor for breast cancer: A systematic review and meta-analysis of case-control studies up to 2010. International journal of environmental research and public health, 18(9), 4638. https://doi.org/10.3390/ijerph18094638 PMID: 33925599

39. Rozenberg S., Di Pietrantonio V., Vandromme J., & Gilles C. (2021). Menopausal hormone therapy and breast cancer risk. Best Practice & Research Clinical Endocrinology & Metabolism, 35(6), 101577. https://doi.org/10.1016/j.beem.2021.101577 PMID: 34535397

40. Laborda-Illanes A., Sánchez-Alcoholado L., Boutriq S., Plaza-Andrades I., Peralta-Linero J., Alba E., et al. (2021). A New Paradigm in the Relationship between Melatonin and Breast Cancer: Gut Microbiota Identified as a Potential Regulatory Agent. Cancers, 13(13), 3141. https://doi.org/10.3390/cancers13133141 PMID: 34201776

41. Kim J. H., & Lim J. S. (2021). Early menarche and its consequence in Korean female: reducing fructose intake could be one solution. Clinical and experimental pediatrics, 64(1), 12. https://doi.org/10.3345/cep.2019.00353 PMID: 32403898

42. Fuhrman B. J., Moore S. C., Byrne C., Makhoul I., Kitahara C. M., De González A. B., et al. (2021). Association of the age at menarche with site-specific cancer risks in pooled data from nine cohorts. Cancer research, 81(8), 2246–2255. https://doi.org/10.1158/0008-5472.CAN-19-3093 PMID: 33820799

43. Słojewska K. (2021). The effect of physical activity on sex hormone levels in women. Implications for breast cancer risk. Nowotwory. Journal of Oncology, 71(6), 383–390. https://doi.org/10.5603/NJO.a2021.0067

44. Jurdana M. (2021). Physical activity and cancer risk. Actual knowledge and possible biological mechanisms. Radiology and oncology, 55(1), 7–17. https://doi.org/10.2478/raon-2020-0063 PMID: 33885236

45. Hao Y., Jiang M., Miao Y., Li X., Hou C., Zhang X., et al. (2021). Effect of long-term weight gain on the risk of breast cancer across women's whole adulthood as well as hormone-changed menopause stages: A systematic review and dose–response meta-analysis. Obesity Research & Clinical Practice, 15(5), 439–448. https://doi.org/10.1016/j.orcp.2021.08.004 PMID: 34456166

46. García-Estévez L., Cortés J., Pérez S., Calvo I., Gallegos I., & Moreno-Bueno G. (2021). Obesity and breast cancer: a paradoxical and controversial relationship influenced by menopausal status. Frontiers in Oncology, 3114. https://doi.org/10.3389/fonc.2021.705911 PMID: 34485137

47. Mohanty S. S., & Mohanty P. K. (2021). Obesity as potential breast cancer risk factor for postmenopausal women. Genes & Diseases, 8(2), 117–123. https://doi.org/10.1016/j.gendis.2019.09.006 PMID: 33997158

48. Scully T., Ettela A., LeRoith D., & Gallagher E. J. (2021). Obesity, type 2 diabetes, and cancer risk. Frontiers in Oncology, 3196. https://doi.org/10.3389/fonc.2020.615375 PMID: 33604295

49. Danforth D. N. (2021). The role of chronic inflammation in the development of breast cancer. Cancers, 13(15), 3918. https://doi.org/10.3390/cancers13153918 PMID: 34359821

50. Pearson-Stuttard J., Papadimitriou N., Markozannes G., Cividini S., Kakourou A., Gill D., et al. (2021). Type 2 diabetes and cancer: an umbrella review of observational and Mendelian randomization studies. Cancer Epidemiology and Prevention Biomarkers, 30(6), 1218–1228. https://doi.org/10.1158/1055-9965.EPI-20-1245 PMID: 33737302

51. Voutsadakis I. A. (2021). Vitamin D baseline levels at diagnosis of breast cancer: A systematic review and meta-analysis. Hematology/oncology and stem cell therapy, 14(1), 16–26. https://doi.org/10.1016/j.hemonc.2020.08.005 PMID: 33002425

52. Wang K., Nakano K., Naderi N., Bajaj-Elliott M., & Mosahebi A. (2021). Is the skin microbiota a modifiable risk factor for breast disease?: A systematic review. The Breast, 59, 279–285. https://doi.org/10.1016/j.breast.2021.07.014 PMID: 34329949

53. Lovrics O., Butt J., Lee Y., Lovrics P., Boudreau V., Anvari M., et al. (2021). The effect of bariatric surgery on breast cancer incidence and characteristics: A meta-analysis and systematic review. The American Journal of Surgery, 222(4), 715–722. https://doi.org/10.1016/j.amjsurg.2021.03.016 PMID: 33771341

54. Wang F., Zhang W., Cheng W., Huo N., & Zhang S. (2021). Preeclampsia and cancer risk in women in later life: a systematic review and meta-analysis of cohort studies. Menopause, 28(9), 1070–1078. https://doi.org/10.1097/GME.0000000000001806 PMID: 34374685

55. Chen S., Wu F., Hai R., You Q., Xie L., Shu L., et al. (2021). Thyroid disease is associated with an increased risk of breast cancer: a systematic review and meta-analysis. Gland Surgery, 10(1), 336. https://doi.org/10.21037/gs-20-878 PMID: 33633990

56. Wei L., Han N., Sun S., Ma X., & Zhang Y. (2021). Sleep-disordered breathing and risk of the breast cancer: A meta-analysis of cohort studies. International Journal of Clinical Practice, 75(11), e14793. https://doi.org/10.1111/ijcp.14793 PMID: 34482589

57. Anmella G., Fico G., Lotfaliany M., Hidalgo-Mazzei D., Soto-Angona Ó., Gimenez-Palomo A., et al. (2021). Risk of cancer in bipolar disorder and the potential role of lithium: International collaborative systematic review and meta-analyses. Neuroscience & Biobehavioral Reviews, 126, 529–541. https://doi.org/10.1016/j.neubiorev.2021.03.034 PMID: 33831461

58. Bowen D. J., Fernandez Poole S., White M., Lyn R., Flores D. A., Haile H. G., et al. (2021). The Role of Stress in Breast Cancer Incidence: Risk Factors, Interventions, and Directions for the Future. International Journal of Environmental Research and Public Health, 18(4), 1871. https://doi.org/10.3390/ijerph18041871 PMID: 33671879

59. Falcinelli M., Thaker P. H., Lutgendorf S. K., Conzen S. D., Flaherty R. L., & Flint M. S. (2021). The Role of Psychologic Stress in Cancer Initiation: Clinical Relevance and Potential Molecular Mechanisms. Cancer research, 81(20), 5131–5140. https://doi.org/10.1158/0008-5472.CAN-21-0684 PMID: 34266894

60. Pereira M. A., Araújo A., Simões M., & Costa C. (2021). Influence of Psychological Factors in Breast and Lung Cancer Risk-A Systematic Review. Frontiers in psychology, 12, 769394–769394. https://doi.org/10.3389/fpsyg.2021.769394 PMID: 35046872

61. Baron J. A., Nichols H. B., Anderson C., & Safe S. (2021). Cigarette smoking and estrogen-related cancer. Cancer Epidemiology and Prevention Biomarkers, 30(8), 1462–1471. https://doi.org/10.1158/1055-9965.EPI-20-1803

62. Papadimitriou N., Markozannes G., Kanellopoulou A., Critselis E., Alhardan S., Karafousia V., et al. (2021). An umbrella review of the evidence associating diet and cancer risk at 11 anatomical sites. Nature communications, 12(1), 4579. https://doi.org/10.1038/s41467-021-24861-8 PMID: 34321471

63. Lee J., Lee J., Lee D. W., Kim H. R., & Kang M. Y. (2021). Sedentary work and breast cancer risk: A systematic review and meta-analysis. Journal of Occupational Health, 63(1), e12239. https://doi.org/10.1002/1348-9585.12239 PMID: 34161650

64. Chong F., Wang Y., Song M., Sun Q., Xie W., & Song C. (2021). Sedentary behavior and risk of breast cancer: a dose–response meta-analysis from prospective studies. Breast Cancer, 28(1), 48–59. https://doi.org/10.1007/s12282-020-01126-8 PMID: 32607943

65. Manouchehri E., Taghipour A., Ghavami V., Ebadi A., Homaei F., & Latifnejad Roudsari R. (2021). Night-shift work duration and breast cancer risk: an updated systematic review and meta-analysis. BMC women's health, 21(1), 1–16. https://doi.org/10.1186/s12905-021-01233-4 PMID: 33653334

66. Van N. T. H., Hoang T., & Myung S. K. (2021). Night shift work and breast cancer risk: a meta-analysis of observational epidemiological studies. Carcinogenesis, 42(10), 1260–1269. https://doi.org/10.1093/carcin/bgab074 PMID: 34409980

67. Shih Y. W., & Tsai H. T. (2021). The association between smartphone use and breast cancer risk among Taiwanese women: A case–control study [response to letter]. Cancer Management and Research, 13, 89–90. https://doi.org/10.2147/CMAR.S296556 PMID: 33447081

68. Cao S., Lu S., Zhou J., Zhu Z., Li W., Su J., et al. (2021). Association between dietary patterns and risk of breast cancer in Chinese female population: a latent class analysis. Public Health Nutrition, 24(15), 4918–4928. https://doi.org/10.1017/S1368980020004826 PMID: 33256868

69. Kazemi A., Barati-Boldaji R., Soltani S., Mohammadipoor N., Esmaeilinezhad Z., Clark C. C., et al. (2021). Intake of various food groups and risk of breast cancer: A systematic review and dose-response meta-analysis of prospective studies. Advances in Nutrition, 12(3), 809–849. https://doi.org/10.1093/advances/nmaa147 PMID: 33271590

70. Huang Y., Cao D., Chen Z., Chen B., Li J., Guo J., et al. (2021). Red and processed meat consumption and cancer outcomes: Umbrella review. Food Chemistry, 356, 129697. https://doi.org/10.1016/j.foodchem.2021.129697 PMID: 33838606

71. Farvid M. S., Sidahmed E., Spence N. D., Mante Angua K., Rosner B. A., & Barnett J. B. (2021). Consumption of red meat and processed meat and cancer incidence: A systematic review and meta-analysis of prospective studies. European journal of epidemiology, 36(9), 937–951. https://doi.org/10.1007/s10654-021-00741-9 PMID: 34455534

72. Finkeldey L., Schmitz E., & Ellinger S. (2021). Effect of the Intake of Isoflavones on Risk Factors of Breast Cancer—A Systematic Review of Randomized Controlled Intervention Studies. Nutrients, 13 (7), 2309. https://doi.org/10.3390/nu13072309 PMID: 34371819

73. Lane M. M., Davis J. A., Beattie S., Gómez-Donoso C., Loughman A., O'Neil A., et al. (2021). Ultraprocessed food and chronic noncommunicable diseases: a systematic review and meta-analysis of 43 observational studies. Obesity reviews, 22(3), e13146. https://doi.org/10.1111/obr.13146 PMID: 33167080

74. Llaha F., Gil-Lespinard M., Unal P., de Villasante I., Castañeda J., & Zamora-Ros R. (2021). Consumption of sweet beverages and cancer risk. A systematic review and meta-analysis of observational studies. Nutrients, 13(2), 516. https://doi.org/10.3390/nu13020516 PMID: 33557387

75. Li Y., Guo L., He K., Huang C., & Tang S. (2021). Consumption of sugar-sweetened beverages and fruit juice and human cancer: a systematic review and dose-response meta-analysis of observational studies. Journal of Cancer, 12(10), 3077. https://doi.org/10.7150/jca.51322 PMID: 33854607

76. Hayati Z., Jafarabadi M. A., & Pirouzpanah S. (2021). Dietary inflammatory index and breast cancer risk: an updated meta-analysis of observational studies. European journal of clinical nutrition, 1–15. PMID: 34728816

77. Chen H., Gao Y., Wei N., Du K., & Jia Q. (2021). Strong association between the dietary inflammatory index (DII) and breast cancer: a systematic review and meta-analysis. Aging (Albany NY), 13(9), 13039. https://doi.org/10.18632/aging.202985 PMID: 33962395

78. Wang L., Zhang R., Yu L., Xiao J., Zhou X., Li X., et al. (2021). Aspirin Use and Common Cancer Risk: A Meta-Analysis of Cohort Studies and Randomized Controlled Trials. Frontiers in oncology, 11.

79. Ma S., Guo C., Sun C., Han T., Zhang H., Qu G., et al. (2021). Aspirin use and risk of breast cancer: a meta-analysis of observational studies from 1989 to 2019. Clinical Breast Cancer, 21(6), 552–565. https://doi.org/10.1016/j.clbc.2021.02.005 PMID: 33741292

80. Xie Y., Wang M., Xu P., Deng Y., Zheng Y., Yang S., et al. (2021). Association Between Antihypertensive Medication Use and Breast Cancer: A Systematic Review and Meta-Analysis. Frontiers in pharmacology, 12, 1169. https://doi.org/10.3389/fphar.2021.609901 PMID: 34054514

81. Urbano T., Vinceti M., Wise L. A., & Filippini T. (2021). Light at night and risk of breast cancer: a systematic review and dose–response meta-analysis. International journal of health geographics, 20(1), 1–26. https://doi.org/10.1186/s12942-021-00297-7 PMID: 34656111

82. Wu Y., Gui S. Y., Fang Y., Zhang M., & Hu C. Y. (2021). Exposure to outdoor light at night and risk of breast cancer: a systematic review and meta-analysis of observational studies. Environmental Pollution, 269, 116114. https://doi.org/10.1016/j.envpol.2020.116114 PMID: 33280921

83. Fiolet T., Mahamat-Saleh Y., Frenoy P., Kvaskoff M., & Mancini F. R. (2021). Background exposure to polychlorinated biphenyls and all-cause, cancer-specific, and cardiovascular-specific mortality: A systematic review and meta-analysis. Environment International, 154, 106663. https://doi.org/10.1016/j.envint.2021.106663 PMID: 34082240

84. Rocha P. R. S., Oliveira V. D., Vasques C. I., Dos Reis P. E. D., & Amato A. A. (2021). Exposure to endocrine disruptors and risk of breast cancer: A systematic review. Critical Reviews in Oncology/Hematology, 161, 103330. https://doi.org/10.1016/j.critrevonc.2021.103330 PMID: 33862246

85. Wei W., Wu B. J., Wu Y., Tong Z. T., Zhong F., & Hu C. Y. (2021). Association between long-term ambient air pollution exposure and the risk of breast cancer: a systematic review and meta-analysis. Environmental Science and Pollution Research, 28(44), 63278–63296. https://doi.org/10.1007/s11356-021-14903-5 PMID: 34227005

86. Ledda C., Bracci M., Lovreglio P., Senia P., Larrosa M., Martinez-Jarreta B., et al. (2021). Pesticide exposure and gender discrepancy in breast cancer. Eur Rev Med Pharmacol Sci, 25(7), 2898–2915. PMID: 33877654

87. Xu S., Wang H., Liu Y., Zhang C., Xu Y., Tian F., et al. (2021). Hair chemicals may increase breast cancer risk: A meta-analysis of 210319 subjects from 14 studies. PloS one, 16(2), e0243792. https://doi.org/10.1371/journal.pone.0243792 PMID: 33539348

88. Gamboa-Loira B., López-Carrillo L., Mar-Sánchez Y., Stern D., & Cebrián M. E. (2021). Epidemiologic evidence of exposure to polycyclic aromatic hydrocarbons and breast cancer: A systematic review and meta-analysis. Chemosphere, 133237. PMID: 34929281

89. Johansson A., Christakou A. E., Iftimi A., Eriksson M., Tapia J., Skoog L., et al. (2021). Characterization of Benign Breast Diseases and Association With Age, Hormonal Factors, and Family History of Breast Cancer Among Women in Sweden. JAMA network open, 4(6), e2114716. https://doi.org/10.1001/jamanetworkopen.2021.14716 PMID: 34170304

90. Lago-Peñas S., Rivera B., Cantarero D., Casal B., Pascual M., Blázquez-Fernández C., et al. (2021). The impact of socioeconomic position on non-communicable diseases: what do we know about it?. Perspectives in Public Health, 141(3), 158–176. https://doi.org/10.1177/1757913920914952 PMID: 32449467

91. Akoglu H. (2018). User's guide to correlation coefficients. Turkish journal of emergency medicine, 18 (3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001 PMID: 30191186

92. Thakur P., Seam R. K., Gupta M. K., Gupta M., Sharma M., & Fotedar V. (2017). Breast cancer risk factor evaluation in a Western Himalayan state: A case-control study and comparison with the Western World. South Asian journal of cancer, 6(3), 106–109. https://doi.org/10.4103/sajc.sajc_157_16 PMID: 28975116

93. Kim Y., Yoo K. Y., & Goodman M. T. (2015). Differences in incidence, mortality and survival of breast cancer by regions and countries in Asia and contributing factors. Asian Pacific journal of cancer prevention: APJCP, 16(7), 2857–2870. https://doi.org/10.7314/APJCP.2015.16.7.2857 PMID: 25854374

94. Mahouri K., Dehghani Zahedani M., & Zare S. (2007). Breast cancer risk factors in south of Islamic Republic of Iran: a case-control study. EMHJ-Eastern Mediterranean Health Journal, 13 (6), 1265–1273. https://doi.org/10.26719/2007.13.6.1265 PMID: 18341177

95. Assi H. A., Khoury K. E., Dbouk H., Khalil L. E., Mouhieddine T. H., & El Saghir N. S. (2013). Epidemiology and prognosis of breast cancer in young women. Journal of thoracic disease, 5 Suppl 1(Suppl 1), S2–S8. https://doi.org/10.3978/j.issn.2072-1439.2013.05.24 PMID: 23819024

96. "IARC Monographs evaluate consumption of red meat and processed meat | UICC." [Online]. Available: https://www.uicc.org/news/iarc-monographs-evaluate-consumption-red-meat-and-processed-meat. [Accessed: 16-Jun-2022].

97. "The Sister Study: What is the Sister Study." [Online]. Available: https://sisterstudy.niehs.nih.gov/English/about.htm. [Accessed: 16-Jun-2022]

98. Lo J. J., Park Y. M. M., Sinha R., & Sandler D. P. (2020). Association between meat consumption and risk of breast cancer: Findings from the Sister Study. International journal of cancer, 146(8), 2156–2165. https://doi.org/10.1002/ijc.32547 PMID: 31389007

99. Mössinger H., & Kostev K. (2023). Depression Is Associated with an Increased Risk of Subsequent Cancer Diagnosis: A Retrospective Cohort Study with 235,404 Patients. Brain sciences, 13(2), 302. https://doi.org/10.3390/brainsci13020302 PMID: 36831845

100. Park H. A., Neumeyer S., Michailidou K., Bolla M. K., Wang Q., Dennis J., et al. (2021). Mendelian randomisation study of smoking exposure in relation to breast cancer risk. British journal of cancer, 125 (8), 1135–1145. https://doi.org/10.1038/s41416-021-01432-8 PMID: 34341517

101. Fraser G. E., Jaceldo-Siegl K., Orlich M., Mashchak A., Sirirat R., & Knutsen S. (2020). Dairy, soy, and risk of breast cancer: those confounded milks. International journal of epidemiology, 49(5), 1526–1537. https://doi.org/10.1093/ije/dyaa007 PMID: 32095830

102. Chen L., Li M., & Li H. (2019). Milk and yogurt intake and breast cancer risk: A meta-analysis. Medicine, 98(12), e14900. https://doi.org/10.1097/MD.0000000000014900 PMID: 30896640

103. Shih Y. W., Hung C. S., Huang C. C., Chou K. R., Niu S. F., Chan S., et al. (2020). The Association Between Smartphone Use and Breast Cancer Risk Among Taiwanese Women: A Case-Control Study. Cancer management and research, 12, 10799–10807. https://doi.org/10.2147/CMAR.S267415 PMID: 33149685

104. Sun Q., Xie W., Wang Y., Chong F., Song M., Li T., et al. (2020). Alcohol Consumption by Beverage Type and Risk of Breast Cancer: A Dose-Response Meta-Analysis of Prospective Cohort Studies. Alcohol and alcoholism (Oxford, Oxfordshire), 55(3), 246–253. https://doi.org/10.1093/alcalc/agaa012 PMID: 32090238

105. Natarajan R., Aljaber D., Au D., Thai C., Sanchez A., Nunez A., et al. (2020). Environmental Exposures during Puberty: Window of Breast Cancer Risk and Epigenetic Damage. International journal of environmental research and public health, 17(2), 493. https://doi.org/10.3390/ijerph17020493 PMID: 31941024

106. Singletary S. E. (2003). Rating the risk factors for breast cancer. Annals of surgery, 237(4), 474–482. https://doi.org/10.1097/01.SLA.0000059969.64262.87 PMID: 12677142

107. Zouré A. A., Bambara A. H., Sawadogo A. Y., Ouattara A. K., Ouédraogo M., Traoré S. S., et al. (2016). Multiparity and breast cancer risk factor among women in Burkina Faso. Asian Pacific journal of cancer prevention: APJCP, 17(12), 5095. https://doi.org/10.22034/APJCP.2016.17.12.5095 PMID: 28122440

108. Qiu R., Zhong Y., Hu M., & Wu B. (2022). Breastfeeding and Reduced Risk of Breast Cancer: A Systematic Review and Meta-Analysis. Computational and mathematical methods in medicine, 2022, 8500910. https://doi.org/10.1155/2022/8500910 PMID: 35126640

109. Baretta Z., Mocellin S., Goldin E., Olopade O. I., & Huo D. (2016). Effect of BRCA germline mutations on breast cancer prognosis: A systematic review and meta-analysis. Medicine, 95(40), e4975. https://doi.org/10.1097/MD.0000000000004975 PMID: 27749552

110. Ibrahim, M. F., Putri, M. M., & Utama, D. M. (2020). A literature review on reducing carbon emission from supply chain system: drivers, barriers, performance indicators, and practices. In IOP Conference Series: Materials Science and Engineering (Vol. 722, No. 1, p. 012034). IOP Publishing.