Correlation Analysis - A Statistical Perspective on Breast Cancer Epidemiology

Abstract

Summary: The epidemiology survey is crucial in determining the causes of medical diseases and offering future treatment possibilities. Many statistical methods, such as regression, weight calculation, and cluster analysis, have been utilized in this area to determine the relationship between factors and the disease. We conducted a cross-sectional epidemiological survey in South China, distributed questionnaires, and analyzed the data to better understand the primary factors influencing breast cancer and the treatment approaches associated with them. Our findings show that in the group we studied, behavioral and reproductive factors have the strongest link with breast cancer, which has been determined as causal and supported by biological evidence. Our model's predicted accuracy was 93.50 percent. The most common symptoms of breast cancer include a lump or thickening in/near the breast or underarm area, irregular menstruation, and feelings of anxiety or depression. The majority of our responders chose general breast inspection and mammogram for inspection, resection surgery, and medications for therapy.

Keywords: Breast Cancer, Epidemiology, Correlation Analysis, Phi Coefficient, Phi-K Coefficient, Pearson Correlation Coefficient, Principle Component Analysis

Contents

Abstract **2**

Contents **3**

1. Introduction **4**

1.1 Breast Cancer Overview 4

1.2 Pathogenesis 4

1.3 Inspection, Treatment, and Prognosis 4

2. Association Analysis **5**

2.1 Factors' Determination 5

2.2 Conduct of Research 5

2.3 Association Among Factors 6

2.3.1 Associations Among Binary Random Variables 6

2.3.2 Association Analysis Involving Non-Binary R.V.s 7

2.4 Association Between Factors and the Disease 10

2.4.1 Behavioral Factors 11

2.4.2 Reproductive Factors 12

2.5 Robustness Analysis 12

3. Future Prospects in Accordance **14**

3.1 Symptoms Emerging 14

3.2 Therapeutics Development 14

3.2.1 Inspection Method 14

3.2.2 Treatment Method 15

3.3 Guideline for Breast cancer 16

4. Resources **17**

4.1 Resources for Cancer Epidemiology 17

4.2 Resources for Patients 17

5. Conclusions **18**

Citations **19**

1. Introduction

1.1 Breast Cancer Overview

Breast cancer is a disease characterized by the growth of malignant cells in the mammary glands.[1] According to the data gathered by International Agency for Research on Cancer, breast cancer takes up 11.7% of the new cancer incidence, second to none, and 6.9% of the deaths in 2020.[2] Besides its occurrence rate, its geological distribution is also noteworthy. Breast cancer affects men and women worldwide despite diverse races, but the underlying socioeconomic status.[3] Some risk factors and protective factors have been established and many are still under investigation.[4–6]

Breast cancer often emerges in the form of a lump in the breast, which is regarded as the first visible sign.[7] The early symptoms of breast cancer include swelling of all or part of a breast, skin dimpling, breast or nipple pain, nipple retraction, nipple or breast skin that is red, dry, flaking, or thickened, nipple discharge, swollen lymph nodes under the arm or near the collar bone.[8–10] The trend of the report on the symptoms in the last year of life includes the increase of the prevalence of any pain for all decedents, prevalence of depression, congestive heart failure or chronic lung disease, and frailty.[11] Currently, a lot of effort is being made to gather the related information as accurately and effectively as possible due to the vastness of information. Many novel methods are applied to electronic health records to extract the symptoms described by the patients such as machine learning, natural language processing, and automatic recognition.[12–16] Those techniques are also applicable to other forms of illness.

1.2 Pathogenesis

According to the relevant research, breast cancer is not categorized as an infectious disease.[17] Currently there is no evidence supporting that breast cancer is caused by a virus or bacterial infection, which is distinct from cervical cancer.[18–20] The occurrence of breast cancer is due to a genetic mutation in mammary gland epithelial cells, most significantly, the mutation in BRCA1, BRCA2, TP53, CDH1, and PTEN.[5,21] The proteins coded by BRCA1 and BRCA2 make an important impact on suppressing the tumorigenesis of breast cancer.[22,23] The related genes are inherited, hence the clustering phenomenon of breast cancer in families.[24] Normally, those who have a family history of breast cancer are more likely to have a higher risk. However, the occurrence of breast cancer is the result of the interactions between genes and the environment, while certain environmental factors can also make an impact on tumorigenesis.[4,5,25]

1.3 Inspection, Treatment, and Prognosis

Due to the high mortality rate and severe prevalence, it's of vital importance to develop inspection methods as well as treatment methods. The major inspection methods can be categorized by whether intrusive or by inspection accuracy. Mammography diagnosis is performed the most frequently, followed by MRI and ultrasound.[26] Besides the old school inspection methods, small-scale, affordable instruments are coming into public sight. Based on the differences in physical properties between the normal tissue and tumor tissue, such as impedance, surface tension, local temperature, and absorption ability to microwave in different wavelengths, many inspection methods are developed.[27–34] In recent years, the photoacoustic effect is drawing more and more attention for its physical advantage in imaging.[35–37]

With the advancement in molecular biology, the stratification of breast cancer is now based on the expression status of certain receptors, for instance, estrogen receptor, progesterone receptor, and HER2.[38,39] Such a stratification method lays a foundation for more targeted treatment. The current stratification of breast cancer cannot satisfy the demand for targeting therapy.[40] Nowadays many researchers focus on a more detailed classification of the subtypes of cancer using clustering methods as well as machine learning techniques.[41] [42] Current treatment methods mainly consist of surgery, radiation therapy, chemotherapy, radiotherapy, endocrine therapy, and targeted therapy.[5], [9], [43] Some methods of treatment are used as adjuvant treatment. The treatment prognosis analysis is getting more attention, for it plays a central role in medical decision-making.[44,45] The prognosis is highly dependent on the stratification of breast cancer. For instance, the luminal-A type of tumors grow slowly and have the highest survival rate. For comparison, luminal-B type tumors and HER-2 overexpressing tumors have a relatively worse prognosis but a higher survival rate. The TNBC type of tumor has the worst prognosis, lowest survival rate, and highest recurrence rate.[46,47]

2. Association Analysis

2.1 Factors' Determination

Here we adopt the categorization method used in multiple studies.[4–6,25,48] To elucidate the causation of breast cancer, we firstly select factors for a primary epidemiological survey, which are: age, gender, educational level (>=undergraduate or college), family history of related illness, behavioral factors, and reproductive factors. Furthermore, to gather the symptoms emerging when suffering from breast cancer, which serves as indicators for later detection of the disease, we focus on the emergence of implications and symptoms in the early stage of the disease.

Although we are willing to extend the possible factors as many as possible, due to the limitations of time and resources, we can only cover a narrow range of all possible factors in favor. More still require future exploration. For instance, social connection level, exposure to artificial light, exposure to certain chemicals or radiations, ethnicity, and vitamin supplementation.[5,21,49,50]

2.2 Conduct of Research

To conduct the study, we firstly designed a questionnaire based on previous studies. The correlation between possible factors and questions on our questionnaire is reflected in *Table 1*. Besides, we also surveyed the inspection methods as well as treatment methods, which will be further discussed in *Section 4*. Moreover, to roughly understand crowd awareness of the disease, we asked about their familiarity with the condition, as well as methods of treatment. The entire content of the questionnaire can be found in **Appendix A**. Even though many potential factors are worth investigating, they are difficult to quantify most of the time. Instead of measuring the quantity by gradient, such as how many cigarettes one smokes in a week on average, we specifically designed the questions to be binary. Although the precision of the data is reduced to a certain level, we obtain more accuracy by making the respondents less dependent on their subjective memory, which is a major source of error.[51,52]

Table 1

Age	
Gender	
Education Background	
Family History	
Behavioral Factors	Alcohol Consumption
	Diet
	Cigarette
	Stay up
	Exercise
Reproductive Factors	Times of given birth
	Breastfeed
	Pregnancy termination occurrence
	Menarche Time
	Menstrual Cycle

The questionnaires are handed out through digital media in the form of online questionnaires. The initial date is 2022, February 18th, and the end date is 2022, July 10th. The period spanned 142 days. The total number of the collected validated questionnaire is 657. The validity of our questionnaire is based on the respondents' answers to an attention question.

Since our research identifies the cases before determining the factors, and the prevalence of both the factors and the disease is determined, thus, it can be regarded as a cross-sectional study.[25,53] Of note, although we sought possible factors which were selected as wide and comprehensive as possible, they only serve as possible sources of exposure rather than controlled sources of exposure. In the following sections, we use statistical methods to analyze the association among all potential factors in the first place. After identifying all independent factors, we will determine the relationship between these factors and the disease.

2.3 Association Among Factors

Before we dive into our analysis, a few assumptions must be made for both clarity and convenience.

Assumption 1. The threshold for the existence of potential strong association measured by the phi coefficient is set at 0.7 and -0.7;[54]

To determine the association between factors and the disease, we first examine the associations among factors, in which the underlying relationship may reside. Since the favored answers to the questions on our questionnaire are designed to be binary, e.g., whether the respondent suffers from breast cancer, then we adopt the logistic model to demonstrate the associations. Note that not all of the questions are binary, we discuss the binary questions and non-binary questions separately.[55]

We here assume the total number of the respondents to be N for generality. For the i-th respondent, we define the random variables as shown in *Table 2* and two generalized random vectors $\boldsymbol{X_i}$ and $\boldsymbol{U_i}$.

$$\begin{cases} \boldsymbol{X_i} = [X_{i1}\ X_{i2}\ \dots\ X_{i8}]^T \\ \boldsymbol{U_i} = [U_{i1}\ U_{i2}\ \dots\ U_{i5}]^T \end{cases} \tag{1}$$

Table 2

Symbol	Symbol Interpretation
Y_i	0 for not suffering from breast cancer, 1 for otherwise
X_{i1}	0 for male, 1 for otherwise
X_{i2}	0 for education background lower than college or undergraduate, 1 for otherwise
X_{i3}	0 for none of the family members have ever suffered from the disease, 1 for otherwise
X_{i4}	0 for never alcohol drinker, 1 for otherwise
X_{i5}	0 for never smoker, 1 for otherwise
X_{i6}	0 for none or seldom physical exercise, 1 for otherwise
X_{i7}	0 for having an unhealthy diet, 1 for otherwise
X_{i8}	0 for possessing the habit of staying up, 1 for otherwise
U_{i1}	0 for younger than 17, 1 for 18-29, 2 for 30-39, 3 for 40-49, 4 for older than 50
U_{i2}	0 to 3 is equal to the time the respondent had given birth, 4 for more
U_{i3}	0 for doesn't feed by breast, 1 for less than 12 months, 2 for 13 to 24 months, 3, for 25 to 36 months, 4 for more
U_{i4}	0 to 2 is equal to the time the respondent had terminated pregnancy, 3 for more
U_{i5}	0 for never had menarche, 1 for younger than 11, 2 for 11 to 14, 3 for 15-18, 4 for later than 18

Table 2: By defining the random variables, we can introduce the concept of probability and thus make use of the multivariate logistic regression in later discussion.

2.3.1 Associations Among Binary Random Variables

Since the case we are studying here only involves binary random variables, then we deal with the variables in an iterative sequence to study each random variable pair's association. For each pair consisting of 2 random variables, for instance, X_i and X_j, there are merely 4 possible cases, we label them in alphabetical order as shown below.

Table 3

$\boldsymbol{X_i \backslash X_j}$	**0**	**1**
0	Case A	Case B
1	Case C	Case D

The binary variables here are symmetrical in significance, in other words, each value of the variable is of equal weight when determining the association. To measure the association between two binary variables, we use the phi coefficient introduced by Udny Yule.[56] Phi correlation coefficient is a metric used to determine the correlation between two binary variables, which is a special case for the Pierson coefficient.[57] It's widely used in the field of bioinformatics. Currently, many means of information processing are based on the phi coefficient. For instance, the optimal classifier for imbalanced data.[58] The range of phi coefficient is $[-1, 1]$, and the closer the phi coefficient is to 1, the stronger positive correlation the random variable pair exhibits. Similarly, the closer the phi coefficient is to -1, the stronger the negative correlation it exhibits. When

the phi coefficient is closer to 0, it indicates weak or no correlation.

We here denote the number of respondents in each case as $n_j, j = A, B, C, D$. Naturally, the following equation stands. The contingency table is shown in *Table 4*.

$$\mathrm{N} = n_A + n_B + n_C + n_D \tag{2}$$

Table 4

i\j	0	1	Total
0	n_A	n_B	$n_{0\cdot}$
1	n_C	n_D	$n_{1\cdot}$
Total	$n_{\cdot 0}$	$n_{\cdot 1}$	N

The formula for the phi coefficient is

$$\varphi = \frac{N * n_{11} - n_{1\cdot} * n_{\cdot 1}}{\sqrt{n_{1\cdot} * n_{\cdot 1} * (N - n_{1\cdot}) * (N - n_{\cdot 1})}} \tag{3}$$

Through iterative computation, we obtain the result as shown in **Appendix B**. An intuitive illustration is shown in **Figure 1**.

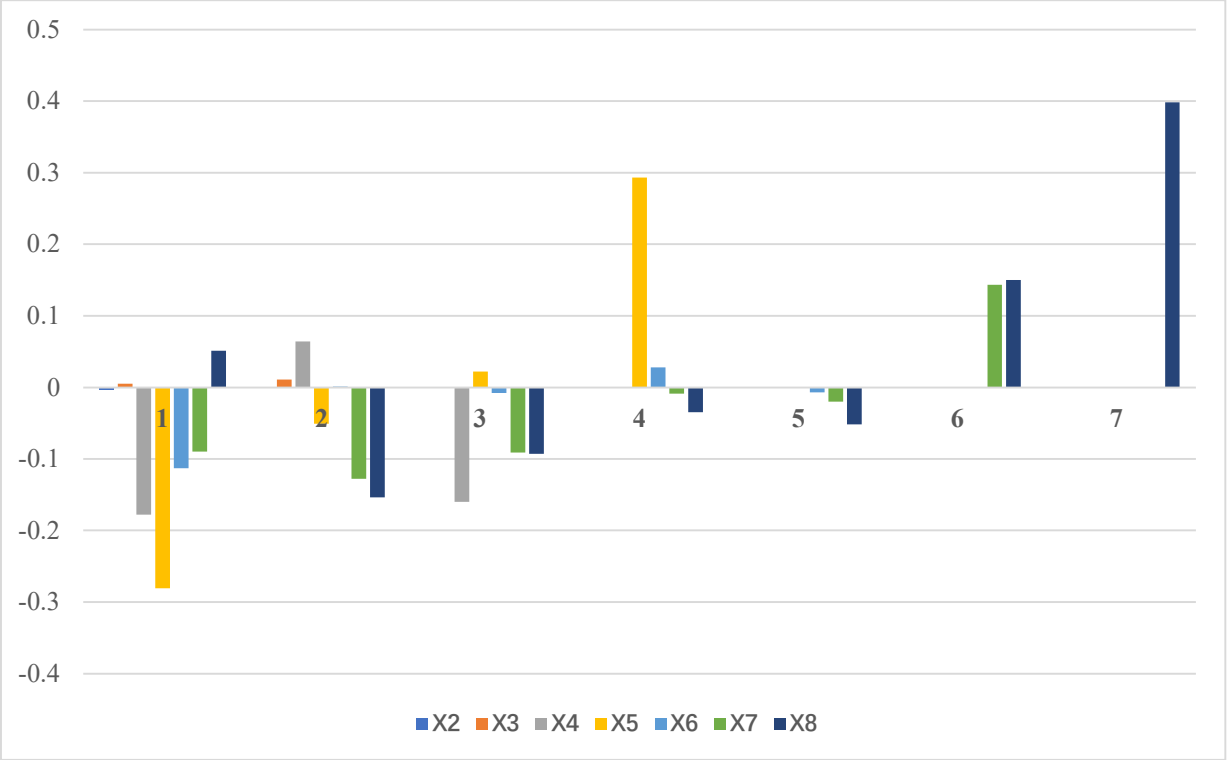

Figure 1

Figure 1: The vertical axis represents the phi coefficient and the horizontal label represents each random variable.

Recalling *Assumption 1* we made, since all of the phi coefficients are below 0.5, we can regard the binary variables as linearly irrelevant. Although some of the random variables show weak correlation. Interestingly, the habit of staying up and having an unhealthy diet produces the largest positive phi coefficient, which indicates that those who possess the habit of staying up tend to have an unhealthier diet. The second-largest phi coefficient corresponds to smoking and drinking, which means that those who possess the habit of drinking alcohol are also likely to smoke cigarettes. Moreover, the random variables X_1 and X_5 show a negative correlation with a phi coefficient equal to -0.281, which indicates that men are more likely to have the habit of smoking cigarettes than women. The phi coefficients of the other random variable pairs are less than 0.2 in absolute value.

2.3.2 Association Analysis Involving Non-Binary R.V.s

To elucidate the association between binary random variables and non-binary random variables, we first notice that some random variables

are already correlated biologically, for instance, gender (X_2) and reproductive factors. We seek to identify those correlated random variable pairs in the first place and label them as BC (biologically correlated). The BC pairs are shown in *Table 5*.

Table 5

BC Pair	Justification
(X_1, U_2)	Only female has the possibility to bear children
(X_1, U_3)	Only female has the possibility to feed by breast
(X_1, U_4)	Only female has the possibility to terminate a pregnancy
(X_1, U_5)	Only female has the possibility to have menstrual
(U_2, U_3)	Only a female given birth can she has the possibility to feed by breast

The mathematical tool we are going to adopt here is the ϕ_K coefficient introduced by M. Baak, R. Koopman, H. Snoek, and S. Klous, which serves as an extended and amended measure of Pearson's correlation coefficient. As constructed, Pearson's correlation coefficient is only applicable when studying interval variables.[59] Moreover, along with Pearson's correlation coefficient, many statistical measures can only capture linear correlation, which would produce enormous errors and false interpretations when the association is not linear.[54,60] The variables we are studying here are categorical. Some coefficients, for instance, Cramér's V, are specifically designed to measure the correlation between categorical variables.[59] However, their values are dependent on the number of rows and columns of the contingency table.[60]

We assume that there are r rows and k columns in the contingency table and (i,j) indicates a cell in the contingency table, we refer to the cell as C_{ij}. The total case number $N = r \cdot k$. For each random variable pair, here denoted by x and y, their probability distribution function characterized by ρ is

$$f(x, y \mid \bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_y^2} - \frac{2\rho(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}\right]\right) \tag{4}$$

$\bar{x}, \bar{y}$ denote the average value separately, and σ_x, σ_y stand for the standard deviation. ρ is the Pearson correlation coefficient. We denote the observed value and the expected value of cell (i,j) by O_{ij}, E_{ij} respectively. The formula of Pearson's χ^2 test is

$$\chi^2 = \sum_{i,j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{5}$$

Using the definition of the probability density function, the probability of cell (i,j) is

$$F_{ij}(\rho) = \int_{C_{ij}} f(x, y \mid \rho)\mathrm{d}x\mathrm{d}y \tag{6}$$

By assigning the value

$$\begin{cases} O_{ij} = N \cdot F_{ij}(\rho) \\ E_{ij} = N \cdot F_{ij}(0) \end{cases} \tag{7}$$

Substitute to *Formula (5)*, we obtain

$$\chi^2(\rho, N, r, k) = N \sum_{i,j}^{r,k} \frac{\left(F_{ij}(\rho) - F_{ij}(0)\right)^2}{F_{ij}(0)} \tag{8}$$

Notice the observed value O_{ij} is constructed independent of the true observed value, we then introduce the effective number of freedom n_f and the number of empty observe cells n_e to reflect the statistical noise.

$$n_\mathrm{f} = (r-1)(k-1) - n_\mathrm{e} \tag{9}$$

We define the pedestal statistics χ_p^2 and χ_{max}^2 as follows.[59]

$$\chi_\mathrm{p}^2 = n_\mathrm{f} + c \cdot \sqrt{2n_\mathrm{f}} \tag{10}$$

$$\chi_{max}^2(N, r, k) = N \cdot \min(r-1, k-1) \tag{11}$$

Where c is added to exclude the outliers. To amend the *Formula (8)* by extending for case $\rho = 0$ and case $\rho = 1$, we get

$$\chi^2(\rho, N, r, k) = \chi_\mathrm{p}^2 + \left\{\frac{\chi_{max}^2(N, r, k) - \chi_\mathrm{p}^2}{\chi^2(1, N, r, k)}\right\} \cdot \chi^2(\rho, N, r, k) \tag{12}$$

The correlation coefficient ϕ_K can be calculated as follow.

$$\phi_K = \begin{cases} 0, & \chi^2 < \chi_{\mathrm{p}}^2 \\ \rho', & \chi^2 \geq \chi_p^2 \end{cases} \tag{13}$$

Where ρ' is the solution to *Equation (12)*, which can be obtained using Brent's method.[14] Note that χ_{p}^2 and $\chi_{max}^2(N, r, k)$ are real values depending on $\mathrm{r}, \mathrm{k}, \mathrm{N}$. After calculation with SPSS and Python, we obtain the result shown in **Figure 2**. The code we used can be found in the citation here.[61]

Figure 2

Figure 2. This figure is drawn using Matplotlib. The number in each cell indicates the ϕ_K coefficient of the corresponding random variable pair. The closer ϕ_K is to 1, the tighter correlation the random variable pair exhibits. On the contrary, the closer ϕ_K is to 0, the more independence the random variable pair exhibits. Some data are adjusted to obtain a valid correlation coefficient since some factors are biologically associated. For instance, gender (X_1) and the times given birth (U_2).

From **Figure 2**, we can see that the random variable pairs that exhibit strong correlation are $(X_7, X_8), (U_1, U_3), (U_1, U_2)$, excluding the biologically correlated factors. Recalling the conclusions we made in *Section 2.3.1*, the first random variable pair indicates that those who possess the habit of staying up tend to have an unhealthy diet. The increased value may imply that the relationship between the two factors may not be linear so the phi coefficient captured less correlation. The 2nd pair indicates that the breastfeeding period of time is correlated with the age of the respondent with $\phi_K = 0.53$. The 3rd pair indicates that the time of giving birth is correlated with the age of the respondent with $\phi_K = 0.81$, which shows a strong correlation.

Here we note that all the correlated factors are categorized into the reproductive factor, we consider constructing a combined random variable representing their overall effect to satisfy the linear independence requirement of the logistic regression we are going to apply in the next section. The method we are going to apply to integrate the associated random variables is the principal component analysis. The principal component analysis is commonly used in exploratory data analysis and is able to conduct dimensionality reduction by vector projection.[62] The principal components are exactly the eigenvectors of the data's covariance matrix, which is obtained in former procedures. The covariance matrix of U_1, U_2, U_3 is shown below.

$$C_0 = \begin{bmatrix} 1 & 0.264 & 0.170 \\ 0.264 & 1 & 0.293 \\ 0.170 & 0.293 & 1 \end{bmatrix} \tag{14}$$

By calculating the eigenvalues and corresponding eigenvectors using the SVD decomposition of C_0, we obtain the adjusted coordinate system.[63] Now that the greatest variance by a scalar projection of the data lies on the first principal component, the second greatest variance on the

second principal component and so on. The scree diagram is depicted in **Figure 3**.

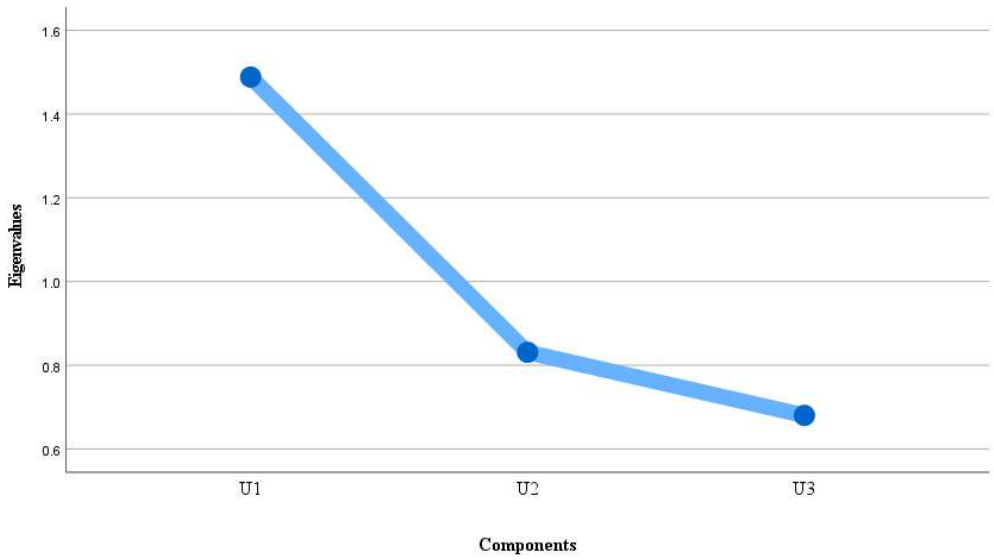

Figure 3

Figure 3. The scree diagram shows each random variable's eigenvalue at the same time. The eigenvalues in a decreasing sequence are 1.489, 0.831, 0.680.

Since the last two variables' eigenvalues are less than 1, we perform the dimensionality reduction by decorrelating U_2, U_3.[64] The feature vector is denoted by $\boldsymbol{v}$. The transformed dataset of the combined random variable can be found in **Appendix B**.

$$\boldsymbol{v} = \begin{bmatrix} 0.766 \\ 0.688 \\ 0.654 \end{bmatrix} \tag{15}$$

To summarize the analysis above, we have identified the association among all possible factors selected and decomposed the associated factors using PCA. We will discuss each factor's association with the disease in the next section.

2.4 Association Between Factors and the Disease

Through the analysis above, the independent factors are identified. Considering the interaction between the correlated factors, we combine their effect into an effect-equivalent combined random variable to represent the reproductive factors' effect. We define a new set of variables in *Table 6*.

Table 6

Symbol	Symbol Interpretation
V_{i1}	0 for male, 1 for otherwise
V_{i2}	0 for education background lower than college or undergraduate, 1 for otherwise
V_{i3}	0 for none of the family members have ever suffered from the disease, 1 for otherwise
V_{i4}	0 for never alcohol drinker, 1 for otherwise
V_{i5}	0 for never smoker, 1 for otherwise
V_{i6}	0 for none or seldom physical exercise, 1 for otherwise
V_{i7}	0 for having an unhealthy diet, 1 for otherwise
V_{i8}	0 for possessing the habit of staying up, 1 for otherwise
V_{i9}	The integrated random variable of U_{i1}, U_{i2}, U_{i3}
V_{i10}	0 to 2 is equal to the time the respondent had terminated pregnancy, 3 for more
V_{i11}	0 for never had menarche, 1 for younger than 11, 2 for 11 to 14, 3 for 15-18, 4 for later than 18

Logistic regression is a common approach to determining the probability of the correlation between the occurrence of two events, especially for problems touching upon discrete cases. In this study, we use the logistic regression formula to model the relationship between breast cancer

status and the linear combination of the potential factors we established.

We define $D = \{1, 2, \cdots, M\}$, $E = \{0, 1, \cdots, M\}$, then

$$p := p(V_k, \beta_i) = \frac{1}{1 + e^{-(\beta_0 + \sum_{k=1}^{m} \beta_k V_k)}}, k \in D, i \in E \tag{16}$$

Where p is the probability for an individual to have breast cancer and $\{V_k\}$ is the set of potential factors,and β_k is the weight for the k-th potential factor respectively. Using logit transformation, we convert the non-linear relationship between p and V_k into a linear relationship between $logit(p)$ and V_k, which is

$$logit(p) = ln\frac{p}{1-p} = \beta_0 + \sum_{k=1}^{M} \beta_k V_k \tag{17}$$

To obtain the most likely value of β_i, we use the maximum likelihood method by

$$L(\beta_i, y_k) = \sum_{k=1}^{M} y_k ln\big(p(V_k)\big) \sum_{k=1}^{M} (1 - y_k) ln\big(1 - p(V_k)\big) \tag{18}$$

Where y_k is the observed breast cancer status variable for the k-th individual. Taking the partial derivative for each β_i and assigning them to 0, we obtain the following set of equations. V_{nk} is the value of V_n of k-th individual.

$$\frac{\partial L}{\partial \beta_k} = \sum_{k=1}^{M} y_k V_{nk} - \sum_{k=1}^{m} p(V_k) V_{nk} = 0 \tag{19}$$

The result of $\{\beta_i\}$ is shown in *Table 7*. We then study each independent factor's association with breast cancer. Bradford Hill criteria are brought up in the study of association and causality between environmental factors and the disease.[65] It's widely accepted in determining the causal relationship between factors and the disease and is adopted by multiple studies.[66] The sequence is based on each factor's weight as in *Formula (16)*, when β increases, the probability of having breast cancer p increases.

Table 7

Coefficient	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
Value	0.000	-3.179	0.361	-10.016	2.340	19.945	-1.100	19.945	22.893	15.781	1.999	-2.233
Sequence	NA	10	7	11	5	2	8	3	1	4	6	9

We firstly analyze the factors with relatively small coefficients for they exhibit less correlation with the disease. Any factors with a coefficient value less than 1 are discussed here. Firstly, the education background (V_2) has a coefficient equal to 0.361, which means somehow the higher the education background is, the higher an individual is more likely to have breast cancer. However, the majority of our sample has an education background of undergraduate, increasing the probability that a respondent has a high education background. Thus, we cannot deduce that the educational background is associated with the disease. Secondly, the habit of suitable physical exercise.

The coefficient for gender (V_1) is -3.179, less than 0, which means that males are more likely to have breast cancer. This result is obviously biased according to multiple studies conducted.[25,49] The reason for our false result is that female takes up the majority of the respondent group, thus, the effect of gender has been diminished. As mentioned, the majority of our respondents are healthy undergraduate students, and the effect of menarche time (V_{11}) is reduced. Lastly, in our data, we discovered that the respondents who have breast cancer don't have any family members suffering from breast cancer. Although the significance of family history is recognized, we cannot draw a solid conclusion from our data.[24]

Through our analysis above, we have recognized the downsides of our data. However, due to the particularity of our sample, some effects are diminished yet some are amplified. We discuss the amplified effects further by categorizing them into behavioral factors and reproductive factors

2.4.1 Behavioral Factors

1. Staying up (V_8)

Staying up causes hormone imbalance in humans. By suppressing the proliferation of breast cancer cells, melatonin reduces the risk of tumorigenesis.[67] However, staying up late reduces the production of melatonin, which increases the risk of breast cancer. This effect is amplified especially in the group consisting of those who are deprived of sleep such as night shift workers.[68,69] Being exposed to light causes the downregulating of melatonin. Although the relationship between the secretion of melatonin and sleep time is yet to be explored, generally, as sleep time increases, so does the probability of not having breast cancer.[70] Thus, we should try to avoid staying up and arrange our sleep time rationally.

2. Unhealthy Diet (V_7)

Consuming food containing high content of sugar, fat, and salt chronically is considered as having an unhealthy diet.[71] An unhealthy diet is a major reason causing obesity. Obesity is tightly correlated to breast cancer, for women who pass their menopause, the obese group has a higher risk of having breast cancer.[72] However, obesity during childhood and adolescence somehow drops this risk, but the underlying principle remains unknown.[73] By having a healthy diet, especially by including more vegetables and fruits in the diet, the possibility of being diagnosed with breast cancer in middle age is lowered.[74] Besides this, having a healthy diet contributes to an overall benefit for health.

3. Smoking (V_5)

As is suggested by multiple breast cancer epidemiology surveys, the risk of having breast cancer is apparently higher in the smoking group.[75] Those who have a family history of smoking have a higher relative risk.[76] Of all the chemicals in the cigarette, nicotine is significantly correlated with lung metastases of breast cancer. The carcinogens contained in the cigarette are founded in breast tissues.[77]

4. Alcohol (V_4)

Drinking alcohol is an indirect cause of multiple malignant tumors, namely, it affects hormone levels prominently.[78] By reducing alcohol intake, a variety of diseases can be avoided, including oropharyngeal cancer, throat cancer, and esophageal cancer. According to relevant research, the intake of alcohol is monotonously positively correlated with the risk of having breast cancer.[79] By limiting alcohol intake to an extent, we can significantly reduce our risk of having breast cancer.

2.4.2 Reproductive Factors

1. Combined (V_9)

The combined factor is consisted of age, times of given birth, and breastfeeding time period. The underlying core of the reproductive factors is the change in hormone levels they arouse. The menarche time, times of giving birth, breastfeeding, and menopause are the main causes of hormone level shifts in females.[80] Additionally, the hormone level is affected by age. The lag of 2 years of menarche corresponds to a 10% lower risk of having breast cancer.[72] The difference in menopause time correspond to a 17% difference in breast cancer risk.[72,81] Moreover, lifelong infertility and late pregnancy increase the risk of having breast cancer.[81] Thus, taking personal factors into account is of vital importance when considering the risk of having breast cancer.

2. Pregnancy Termination (V_{10})

Full pregnancy reduces the risk of having breast cancer, in comparison, the effect of incomplete pregnancy on breast cancer is yet to be elucidated.[82] In some epidemiology research, neither the times of termination of pregnancy nor the pregnancy period of time shows a significant correlation with breast cancer, which is contrary to our result.[82,83] More relevant research is required to draw a conclusion.

2.5 Robustness Analysis

For the PCA process in *Section 2.3*, we use the Kaiser-Meyer-Olkin test to test the sample adequacy and Bartlett's test to test the null hypothesis.[84,85] Applying *Formula (20)* below, we obtain the sample's KMO index $KMO = 0.587$. In the formula, r_{jk} is the correlation between the desired variable pair and p_{jk} is the partial correlation. Since our KMO index falls in the interval $[0.5,0.6]$, thus our data is acceptable.[86] In *Formula (21*), $\mathrm{N} = \sum_{i=1}^{k} n_i$ and $S_p^2 = \frac{1}{N-k}\sum_i (n_i - 1)S_i^2$, which is the pooled estimate for the variance. The approximate χ^2 of our sample is 42.262, the significance level is 0.000, thus our conclusions hold.

$$KMO = \frac{\sum_{j \neq k} \sum_{jk} r_{jk}^2}{\sum_{j \neq k} \sum_{jk} r_{jk}^2 + \sum_{j \neq k} \sum_{jk} p_{jk}^2} \tag{20}$$

$$\chi^2 = \frac{(N-k)\ln\left(S_p^2\right) - \sum_{i=1}^{k} (n_i - 1)\ln\left(S_i^2\right)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right) - \frac{1}{N-k}\right)} \tag{21}$$

Moreover, we applied the log-likelihood test, and the degree of freedom test, to obtain the significance of our data. The detail of the test result can be found in **Appendix B**. Finally we obtain the estimation accuracy of our model, which is listed in *Table 8*. Although in some cases our model exhibits low estimation accuracy, the overall accuracy is 93.50%.

Table 8

Real Value	Evaluation		
	0	1	**Accuracy**
0	397	9	97.80%
1	21	32	60.40%
Accuracy	91.10%	8.90%	93.50%

3. Future Prospects in Accordance

3.1 Symptoms Emerging

Using the same methods applied in *Section 2.3*, *Section 2.4*, we analyze the association between the arise of symptoms and the occurrence of the disease. The symptoms we surveyed are listed in *Table 9*

Table 9

Number	Symptom
1	A Lump or Thickening In/Near the Breast or in the Underarm Area
2	Nipple Discharge
3	Irregular Menstruation, Anxious or Depressed
4	Local Skin Hyperthermia, Redness, Skin Abscess
5	Swollen Lymph Nodes
6	Dimpling of The Breast Tissue, Orange Peel Alike
7	Inverted Nipples

Those symptoms are selected because they are distinctive and common for breast cancer patients.[8–10] As **Figure 4** shows below, most of the patients have a lump or thickening in/near the breast or in the underarm area. The second largest group has irregular menstruation, anxious or depressed feelings. Other symptoms are nipple discharge, inverted nipples, and local skin hyperthermia, redness, skin abscess.

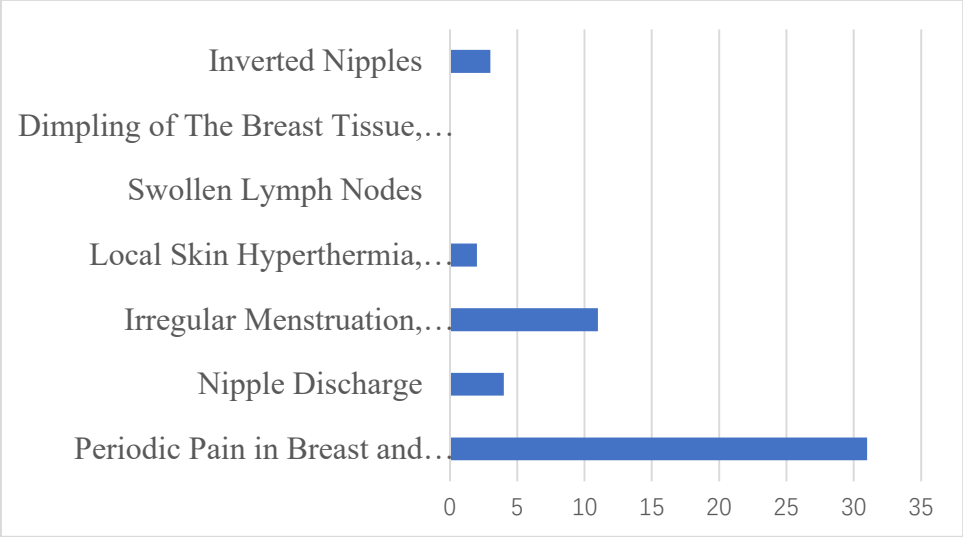

Figure 4

3.2 Therapeutics Development

3.2.1 Inspection Method

Question 18 on our questionnaire surveyed the inspection technique patients took during the diagnosis process while **Question 19** on our questionnaire surveyed the treatment methods patients adapted. According to the data obtained from the questionnaire, which is summarized here in a pie chart, as shown in **Figure 5**, the majority of our respondents selected general breast inspection (38%), followed by breast mammogram (22%), molybdenum target mammogram (14%), tissue biopsy inspection (10%), and blood test (6%). Each of the rest inspection methods is selected by less than 5% of the crowd. Due to our limited sample size, we only regard those higher than 5% as frequently chosen methods.

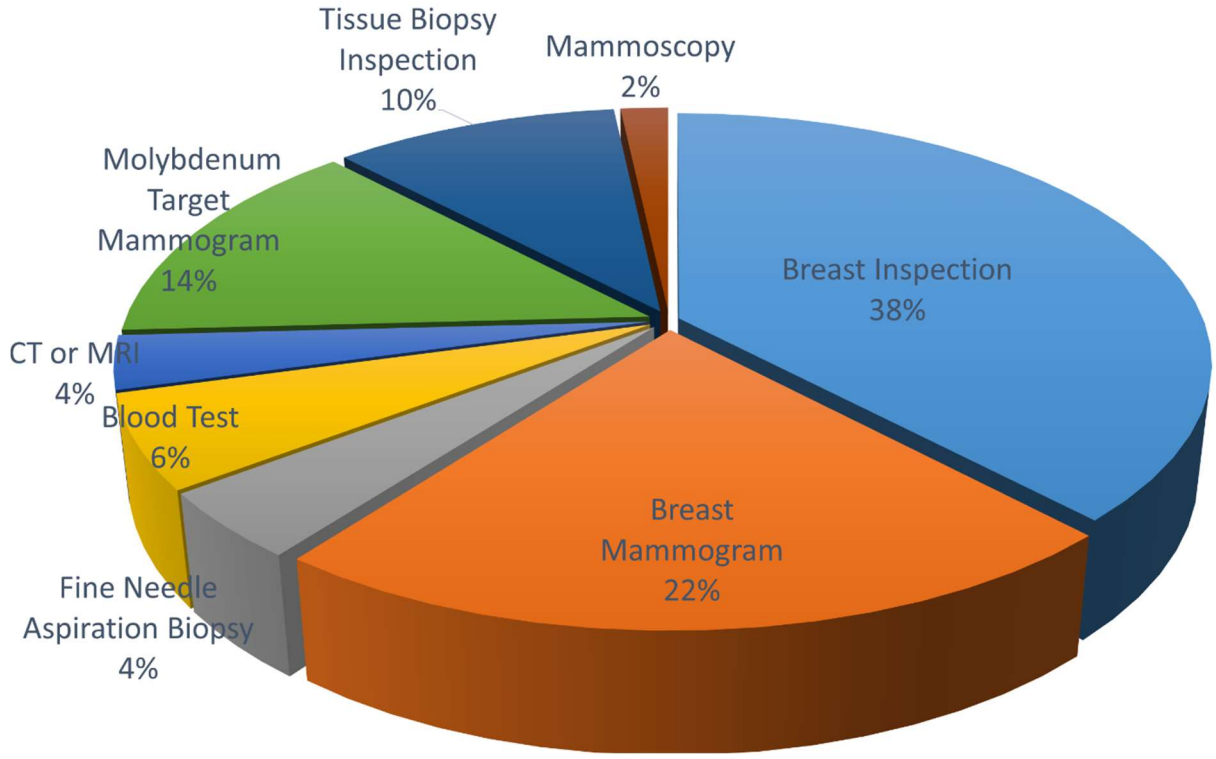

Figure 5 Pie Chart of Inspection Methods

Figure 5: Many respondents took more than one type of inspection, here we count the times each inspection method was taken. The crowd here only consists of those who are suffering from or have suffered from breast cancer. Other respondents are not taken into account for irrelevance.

3.2.2 Treatment Method

We describe the effectiveness of the treatment method by both absolute measurements and relative measurements to avoid unintended intuition from data, i.e., we take the selection count and phi coefficient both into account. For the i-th respondent, we label them as shown in *Table 10*.

Table 10

Number	Treatment	Notation
1	Drugs	T_{i1}
2	Resection Operation	T_{i2}
3	Laser Therapy	T_{i3}
4	Chemotherapy	T_{i4}
5	Radiotherapy	T_{i5}
6	Endocrine Therapy	T_{i6}
7	Targeted Therapy	T_{i7}
8	Self-healing	T_{i8}

Table 10: T_{i9} for other treatment methods, which are surveyed in our questionnaire in the form of blank-filling.

Since there are multiple choices of treatment methods, we simplify the cases by dividing each respondent's choice into multiple binary choices, which is whether a respondent selected a certain treatment method. Using the similar approach we proposed in *Section 2*, we regard the selection of each treatment method independently. We use C_i to represent the cure status of the i-th respondent, 1 for cured, 0 for still suffering from the disease. We measure the association between cure status and the treatment measures selected using the phi coefficient. The result is

depicted in *Table 11*. For the whole data content, please see **Appendix B**.

Table 11 Phi Coefficient Table for Treatment Methods and Cure Status

Treatment	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
Select Count	26	11	0	1	1	2	0	18
Phi Coefficient	0.138	0.506	None	-0.078	0.137	0.000	None	-0.236

We notice that the 3rd and 7th treatment methods are selected by none, their phi coefficients are labeled as none, thus we discard them in the later discussion. Also, we eliminate chemotherapy (4th), radiotherapy (5th), and endocrine therapy (6th) because their selection count is less than 5. If they are included, the result would be biased. Of all the other 3 treatment methods we surveyed, the drug is selected by the majority and the following are self-healing, then resection operation. Of note, in drug selection, Chinese traditional medicine is under debate because though it exhibits medical effect, its mechanism is still unclear and requires further investigation.[87] When we focus on the correlation between treatment method selection and cure status, we can see that even though the drug is selected by the most, it is less related to cure status than resection operation. Self-healing's phi coefficient is -0.236, which indicates a weak negative correlation. By further observation of data, the cured respondents usually use combined treatment methods, such as resection operation combined with medication.

3.3 Guideline for Breast cancer

Summarizing the factors verified in *Section 2*, we give some suggestions that if followed, one can definitely benefit in daily life in column A. The results collected from our questionnaire regarding the symptoms are listed in column B. Based on literature reviews and our data, we provide inspection methods and treatment methods in columns C and D. We hope this guideline can raise more awareness and provide an epidemiological perspective at the same time to those who are willing to know more about breast cancer. The whole guideline can be found in **Appendix C**.

4. Resources

4.1 Resources for Cancer Epidemiology

During our literature review and analysis, we find several databases that are of significant use. We here list them in *Table 12*.

Table 12

Institute	Website
Health Sciences Library System	https://www.hsls.pitt.edu/obrc/
International Agency for Research on Cancer	https://gco.iarc.fr/projects#datavisualization
American Cancer Society	https://www.cancer.org/
CDC	https://www.cdc.gov/cancer/breast/statistics/index.htm
China National Central Cancer Registry	https://ghdx.healthdata.org/series/china-national-central-cancer-registry
World Cancer Research Fund International	https://www.wcrf.org/cancer-trends/breast-cancer-statistics/
Cancer Incidence in Five Continents	CI5 - Home (iarc.fr)
Physician Data Query	https://www.cancer.gov/publications/pdq

The databases we mentioned here are mostly related to epidemiological research, while biomedical data are insufficient. To obtain more biomedical or biological data on breast cancer, please refer to the citations here.[88,89]

Moreover, respecting epidemiological statistics, we highly recommend the software named *Epi info* developed by CDC, which can be found on the website we cite here.[90] *Jointpoint Trend Analysis Software* for trend analysis is also worth exploring.[91]

4.2 Resources for Patients

For those who are unfortunately suffering from breast cancer, we summarized the sites and associations from our questionnaire and the internet where one can apply for aid and reach out for help. The detailed information is listed in *Table 13* below.

Table 13

Institute	Website
National Breast Cancer Foundation	https://www.nationalbreastcancer.org/
Breast Cancer Foundation	https://www.bcf.org.sg/
Cancer Support Community	https://www.cancersupportcommunity.org/breast-cancer
Cancer Care	https://www.cancercare.org/diagnosis/breast_cancer
TNBC Foundation	https://tnbcfoundation.org/
Breast Cancer Now	https://breastcancernow.org/
NBCCEDP	https://www.cdc.gov/cancer/nbccedp/
National Mammography Program	https://www.nationalbreastcancer.org/national-mammography-program

To reach out for local support, please consult local professionals.

5. Conclusions

In this study, we conducted an epidemiology survey of breast cancer using online questionnaires. From the data collected, we established two major risk factors in correlation with breast cancer: behavioral factors and reproductive factors. Behavioral factors include the habit of staying up, smoking, unhealthy diet, alcohol drinking, while reproductive factors include age, times of giving labor, the period of time for breastfeeding, and times of pregnancy termination. During our analysis of the association among factors with the ϕ coefficient and amended Pearson coefficient ϕ_K, we find that the habit of staying up is moderately correlated with having an unhealthy diet with $\phi = 0.398$, and the habit of smoking is weakly correlated with drinking alcohol with $\phi = 0.293$, and smoking is weakly correlated with the male gender with $\phi = -0.281$. Using principal component analysis, we combined several associated variables. Finally, using logistic regression, the estimated accuracy of our model reached 93.50%. By Kaiser-Meyer-Olkin test and Bartlett test, we justified our result.

The most frequently reported symptoms of breast cancer are a lump or thickening in/near the breast or in the underarm area and irregular menstruation, anxious or depressed feeling. The majority of our respondents selected general breast inspection and breast mammogram for inspection, respectively, resection operation, and drugs for treatment. According to the results obtained, we give a guideline including life habit recommendation, symptoms for self-examine, inspection method recommendation, and treatment method recommendation. Along the research process, we documented our data and resources, which can be found in **Appendix 2** and *Section 4*.

Author Contributions: Conceptualization, Hongying Zhong, Ruotao Yu, Shengqi Liu; draft preparation, Ruotao Yu, Shengqi Liu, Qianhui Li, Jiarui Chen; writing, review, and editing, Ruotao Yu, Shengqi Liu, Qianhui Li, Jiarui Chen; code and Figure: Shengqi Liu, Ruotao Yu, Peixin Liu; Guideline, Jiarui Chen; supervision, Hongying Zhong;

All authors have read and agreed to the published version of the manuscript.

Funding: No funding received.

Conflicts of Interest: The authors declare no conflict of interest.

Citations

1. Britannica, T. E. of E. https://www.britannica.com/science/breast-cancer. *Encyclopedia Britannica* (2022).
2. Hanahan, D. & Weinberg, R. A. The Hallmarks of Cancer. *Cell* **100**, 57–70 (2000).
3. Campbell, J. B. Breast cancer-race, ethnicity, and survival: A literature review. *Breast Cancer Res. Treat.* **74**, 187–192 (2002).
4. McPherson, K., Steel, C. M. & Dixon, J. M. ABC of breast diseases: Breast cancer - Epidemiology, risk factors, and genetics. *British Medical Journal* vol. 321 624–628 at https://doi.org/10.1136/bmj.321.7261.624 (2000).
5. Kukasiewicz, S. *et al.* Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—An updated review. *Cancers* vol. 13 1–30 at https://doi.org/10.3390/cancers13174287 (2021).
6. Verma, R., Bowen, R. L., Slater, S. E., Mihaimeed, F. & Jones, J. L. Pathological and epidemiological factors associated with advanced stage at diagnosis of breast cancer. *British Medical Bulletin* vol. 103 129–145 at https://doi.org/10.1093/bmb/lds018 (2012).
7. Mandell, J. B. Bathsheba's breast Women, cancer & history. *Journal of Clinical Investigation* vol. 115 1397 at https://doi.org/10.1172/JCI25456 (2005).
8. Teunissen, S. C. C. M. *et al.* Symptom Prevalence in Patients with Incurable Cancer: A Systematic Review. *J. Pain Symptom Manage.* **34**, 94–104 (2007).
9. Shah, R., Rosso, K. & David Nathanson, S. Pathogenesis, prevention, diagnosis and treatment of breast cancer. *World J. Clin. Oncol.* **5**, 283–298 (2014).
10. Society, A. C. https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-cancer-signs-and-symptoms.html#references. (2021).
11. Singer, A. E. *et al.* Symptom trends in the last year of life. *Ann. Intern. Med.* **162**, 175–183 (2015).
12. Forsyth, A. W. *et al.* Machine Learning Methods to Extract Documentation of Breast Cancer Symptoms From Electronic Health Records. *J. Pain Symptom Manage.* **55**, 1492–1499 (2018).
13. Szlosek, D. A. & Ferretti, J. M. Using Machine Learning and Natural Language Processing Algorithms to Automate the Evaluation of Clinical Decision Support in Electronic Medical Record Systems. *eGEMs (Generating Evid. Methods to Improv. patient outcomes)* **4**, 5 (2016).
14. Brent, R. P. *Powell's Algorithm*. *Algorithms for minimization without derivatives* (1973).
15. Carrell, D. S. *et al.* Using natural language processing to improve efficiency of manual chart abstraction in research: The case of breast cancer recurrence. *Am. J. Epidemiol.* **179**, 749–758 (2014).
16. Rajkomar, A., Dean, J. & Kohane, I. Machine Learning in Medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).
17. Greaves, M. & Hughes, W. Cancer cell transmission via the placenta. *Evol. Med. Public Heal.* **2018**, 106–115 (2018).
18. Schiffman, M., Castle, P. E., Jeronimo, J., Rodriguez, A. C. & Wacholder, S. Human papillomavirus and cervical cancer. *Lancet* **370**, 890–907 (2007).
19. DeVita , Lawrence, Theodore S., Rosenberg, Steven A., V. T. *DeVita, Hellman, and Rosenberg's cancer : principles & practice of oncology*. (Wolters Kluwer/Lippincott Williams & Wilkins, 2008).
20. Greig, J. M. & Ellis, C. J. Biological Agents. *Occup. Hyg. Third Ed.* **100**, 344–359 (2008).
21. Shiovitz, S. & Korde, L. A. Genetics of breast cancer: A topic in evolution. *Ann. Oncol.* **26**, 1291–1299 (2015).
22. Hoskins, L. M., Roy, K., Peters, J. A., Loud, J. T. & Greene, M. H. Disclosure of Positive BRCA1/2-Mutation Status in Young Couples: The Journey From Uncertainty to Bonding Through Partner Support. *Fam. Syst. Health* **26**, 296–316 (2008).
23. Thompson, D., Easton, D. F. & Consortium, the B. C. L. Cancer Incidence in BRCA1 Mutation Carriers. *JNCI J. Natl. Cancer Inst.* **94**, 1358–1365 (2002).
24. Jonker, M. A. *et al.* Modeling Familial Clustered Breast Cancer Using Published Data. *Cancer Epidemiol. Biomarkers Prev.* **12**, 1479–1485 (2003).
25. Mcpherson, K., Steel, C. M. & Dixon, J. M. Breast cancer — epidemiology , risk factors , and genetics Risk factors for breast cancer. *Mortality* **321**, (2000).
26. Smetherman, D. H. Screening, Imaging, and Image-Guided Biopsy Techniques for Breast Cancer. *Surg. Clin. North Am.* **93**, 309–327 (2013).
27. Anupama, P. *et al.* Design of a highly accurate data acquisition device for thermal imaging based early detection of breast cancer. *IEEE*

Reg. 10 Annu. Int. Conf. Proceedings/TENCON **2019-Octob**, 2439–2443 (2019).

28. Zhao, Q., Zhang, J., Wang, R. & Cong, W. Use of a thermocouple for malignant tumor detection. *IEEE Eng. Med. Biol. Mag.* **27**, 64–66 (2008).
29. Arcarisi, L. *et al.* Palpreast-A new wearable device for breast self-examination. *Appl. Sci.* **9**, (2019).
30. Pardo, A. *et al.* Modeling and Synthesis of Breast Cancer Optical Property Signatures with Generative Models. *IEEE Trans. Med. Imaging* **40**, 1687–1701 (2021).
31. Yun, X., Johnston, R. H. & Fear, E. C. Radar-based microwave Imaging for breast cancer detection: Tumor sensing with cross-polarized reflections. *IEEE Antennas Propag. Soc. AP-S Int. Symp.* **3**, 2432–2435 (2004).
32. Mashekova, A. *et al.* Early detection of the breast cancer using infrared technology – A comprehensive review. *Therm. Sci. Eng. Prog.* **27**, 101142 (2022).
33. Cherepenin, V. *et al.* A 3D electrical impedance tomography (EIT) system for breast cancer detection. *Physiol. Meas.* **22**, 9–18 (2001).
34. Ng, E. Y. K., Vinitha Sree, S., Ng, K. H. & Kaw, G. The use of tissue electrical characteristics for breast cancer detection: A perspective review. *Technol. Cancer Res. Treat.* **7**, 295–308 (2008).
35. Rao, A. P., Bokde, N. & Sinha, S. Photoacoustic imaging for management of breast cancer: A literature review and future perspectives. *Appl. Sci.* **10**, (2020).
36. Nyayapathi, N. & Xia, J. Photoacoustic imaging of breast cancer: a mini review of system design and image features. *J. Biomed. Opt.* **24**, 1 (2019).
37. Mallidi, S., Luke, G. P. & Emelianov, S. Photoacoustic imaging in cancer detection, diagnosis, and treatment guidance. *Trends Biotechnol.* **29**, 213–221 (2011).
38. Benichou, J., Gail, M. H. & Mulvihill, J. J. Graphs to estimate an individualized risk of breast cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **14**, 103–110 (1996).
39. Gail, M. H. *et al.* Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **81**, 1879–1886 (1989).
40. Yang, Q. *et al.* [Clinical features and prognosis analysis of different breast cancer molecular subtypes]. *Zhonghua Zhong Liu Za Zhi* **33**, 42–46 (2011).
41. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10869–10874 (2001).
42. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
43. Waks, A. G. & Winer, E. P. Breast Cancer Treatment: A Review. *JAMA - J. Am. Med. Assoc.* **321**, 288–300 (2019).
44. Dawood, S. *et al.* Defining breast cancer prognosis based on molecular phenotypes: results from a large cohort study. *Breast Cancer Res. Treat.* **126**, 185–192 (2011).
45. Weigel, M. T. & Dowsett, M. Current and emerging biomarkers in breast cancer: prognosis and prediction. *Endocr. Relat. Cancer* **17**, R245–R262 (2010).
46. Wang, D.-Y., Jiang, Z., Ben-David, Y., Woodgett, J. R. & Zacksenhaus, E. Molecular stratification within triple-negative breast cancer subtypes. *Sci. Rep.* **9**, 19107 (2019).
47. Russnes, H. G., Lingjærde, O. C., Børresen-Dale, A.-L. & Caldas, C. Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. *Am. J. Pathol.* **187**, 2152–2162 (2017).
48. Kelsey, J. L. Breast cancer epidemiology: Summary and future directions. *Epidemiol. Rev.* **15**, 256–263 (1993).
49. Block, W. D. & Muradali, D. Breast cancer in men. *Cmaj* **185**, 1247 (2013).
50. Lin, F. R., Niparko, J. K. & Ferrucci, and L. Social networks, social support and burden in relationships, and mortality after breast cancer diagnosis in the Life After Breast Cancer Epidemiology (LACE) Study. *Breast Cancer Res Treat* **23**, 1–7 (2013).
51. Scott, J. & Huskisson, E. C. Accuracy of subjective measurements made with or without previous scores: an important source of error in serial measurement of subjective states. *Ann. Rheum. Dis.* **38**, 558–559 (1979).
52. Bertrand, M. & Mullainathan, S. Do People Mean What They Say? Implications for Subjective Survey Data. *Am. Econ. Rev.* **91**, 67–72 (2001).
53. Services, H., Control, D. & Cdc, P. *Principles of Epidemiology in Public Health Practice*. (2012).

54. Schober, P. & Schwarte, L. A. Correlation coefficients: Appropriate use and interpretation. *Anesth. Analg.* **126**, 1763–1768 (2018).
55. Janse, R. J. *et al.* Conducting correlation analysis: important limitations and pitfalls. *Clin. Kidney J.* **14**, 2332–2337 (2021).
56. Yule, G. U. On the Methods of Measuring Association Between Two Attributes. *Society* **75**, 579–652 (1912).
57. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables. *Philosophical Magazine* vol. 50 157–175 at https://sci-hub.tw/10.1080/14786440009463897 (1900).
58. Boughorbel, S., Jarray, F. & El-anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. 1–17 (2017).
59. David, F. N. & Cramer, H. Mathematical Methods of Statistics. *Biometrika* **34**, 374 (1947).
60. Baak, M., Koopman, R., Snoek, H. & Klous, S. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Comput. Stat. Data Anal.* **152**, (2020).
61. https://github.com/KaveIO/PhiK/blob/master/phik/notebooks/phik_tutorial_spark.ipynb.
62. Pearson, K. LIII. *On lines and planes of closest fit to systems of points in space*. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
63. Jolliffe, I. Principal Component Analysis. *Encyclopedia of Statistics in Behavioral Science* at https://doi.org/https://doi.org/10.1002/0470013192.bsa501 (2005).
64. Ledesma, R. D., Valero-Mora, P. & Macbeth, G. The Scree Test and the Number of Factors: a Dynamic Graphics Approach. *Span. J. Psychol.* **18**, E11 (2015).
65. Hill, A. President' s Address The Environment and Disease: Association or causation? *Proc R Soc Med* 295–300 (1965).
66. Fredricks, D. N. & Relman, D. A. Sequence-based identification of microbial pathogens: A reconsideration of Koch's postulates. *Clin. Microbiol. Rev.* **9**, 18–33 (1996).
67. Cos, S. & Sánchez-Barceló, E. J. Melatonin and Mammary Pathological Growth. *Front. Neuroendocrinol.* **21**, 133–170 (2000).
68. Megdal, S. P., Kroenke, C. H., Laden, F., Pukkala, E. & Schernhammer, E. S. Night work and breast cancer risk: A systematic review and meta-analysis. *Eur. J. Cancer* **41**, 2023–2032 (2005).
69. Viswanathan, A. N. & Schernhammer, E. S. Circulating melatonin and the risk of breast and endometrial cancer in women. *Cancer Lett.* **281**, 1–7 (2009).
70. Wu, A. H. *et al.* Sleep duration, spot urinary 6-sulfatoxymelatonin levels and risk of breast cancer among Chinese women in Singapore. *Int. J. Cancer* **132**, 891–896 (2013).
71. Tremellen, K. Chapter 3.5 - Treatment of Sperm Oxidative Stress: A Collaborative Approach Between Clinician and Embryologist. in (eds. Henkel, R., Samanta, L. & Agarwal Antioxidants and Impact of the Oxidative Status in Male Reproduction, A. B. T.-O.) 225–235 (Academic Press, 2019). doi:https://doi.org/10.1016/B978-0-12-812501-4.00021-3.
72. Hsieh, C.-C., Trichopoulos, D., Katsouyanni, K. & Yuasa, S. Age at menarche, age at menopause, height and obesity as risk factors for breast cancer: Associations and interactions in an international case-control study. *Int. J. Cancer* **46**, 796–800 (1990).
73. Farvid, M. S. *et al.* Fruit and vegetable consumption in adolescence and early adulthood and risk of breast cancer: population based cohort study. *BMJ* **353**, i2343 (2016).
74. Baer, H. J., Tworoger, S. S., Hankinson, S. E. & Willett, W. C. Body Fatness at Young Ages and Risk of Breast Cancer Throughout Life. *Am. J. Epidemiol.* **171**, 1183–1194 (2010).
75. Park, H. A. *et al.* Mendelian randomisation study of smoking exposure in relation to breast cancer risk. *Br. J. Cancer* **125**, 1135–1145 (2021).
76. Jones, M. E., Schoemaker, M. J., Wright, L. B., Ashworth, A. & Swerdlow, A. J. Smoking and risk of breast cancer in the Generations Study cohort. *Breast Cancer Res.* **19**, 118 (2017).
77. Tyagi, A. *et al.* Nicotine promotes breast cancer metastasis by stimulating N2 neutrophils and generating pre-metastatic niche in lung. *Nat. Commun.* **12**, 474 (2021).
78. LoConte, N. K., Brewster, A. M., Kaur, J. S., Merrill, J. K. & Alberg, A. J. Alcohol and Cancer: A Statement of the American Society of Clinical Oncology. *J. Clin. Oncol.* **36**, 83–93 (2017).
79. Key, J. *et al.* Meta-analysis of Studies of Alcohol and Breast Cancer with Consideration of the Methodological Issues. *Cancer Causes Control* **17**, 759–770 (2006).
80. Łukasiewicz, S. *et al.* Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment

Strategies—An Updated Review. *Cancers* vol. 13 at https://doi.org/10.3390/cancers13174287 (2021).

81. Kelsey, J. L., Gammon, M. D. & John, E. M. Reproductive factors and breast cancer. *Epidemiol. Rev.* **15**, 36–47 (1993).
82. Lash, T. L. & Fink, A. K. Null association between pregnancy termination and breast cancer in a registry-based study of parous women. *Int. J. Cancer* **110**, 443–448 (2004).
83. Breast cancer and abortion: collaborative reanalysis of data from 53 epidemiological studies, including 83 000 women with breast cancer from 16 countries. *Lancet* **363**, 1007–1016 (2004).
84. Bartlett, M. S. Properties of Sufficiency and Statistical Tests. *Proc. R. Soc. Lond. A. Math. Phys. Sci.* **160**, 268–282 (1937).
85. Kaiser, H. F. & Rice, J. Little Jiffy, Mark Iv. *Educ. Psychol. Meas.* **34**, 111–117 (1974).
86. Dziuban, C. D. & Shirkey, E. C. When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychol. Bull.* **81**, 358–361 (1974).
87. Cui, Y. *et al.* Use of complementary and alternative medicine by Chinese women with breast cancer. *Breast Cancer Res. Treat.* **85**, 263–270 (2004).
88. Chowdhury, S. & Chakraborty, P. pratim. Universal health coverage - There is more to it than meets the eye. *J. Fam. Med. Prim. Care* **6**, 169–170 (2017).
89. Clare, S. E. & Shaw, P. L. “Big data” for breast cancer: Where to look and what you will find. *npj Breast Cancer* **2**, 1–5 (2016).
90. CDC. https://www.cdc.gov/epiinfo/index.html.
91. Regression, J. https://surveillance.cancer.gov/joinpoint/.