Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/47/20241357

Prediction of patient breast cancer probability

Wenyang Qiu

Concordia University, 1455 Maisonneuve Ouest, Montreal, Canada

wenyang.qiu@mail.concordia.ca

Abstract. Breast cancer has a significant global impact; in 2015, it caused 570,000 deaths and 1.5 million yearly diagnoses. A challenge is that it has a poor prognosis for cure and is metastatic. The 21st century's health-conscious atmosphere emphasizes the need to lower the death toll from cancer, which will account for approximately one in six fatalities in 2020. According to malignancy, an estimated 7.8 million women are projected to be diagnosed with breast cancer throughout the upcoming five-year period. For the identification and prevention of cancer, proactive measures are required. Python algorithms, particularly linear regression, are extraordinarily useful for analyzing complex datasets. Using linear regression in Python to analyze data yields illuminating models that reveal morbidity trends. With the insights gained from these models, healthcare providers can provide patients with more individualized care. This proactive approach and implementing Python's linear regression algorithms enhance the understanding of cancer risk and allow for effective preventative measures. Greater public awareness of health issues has resulted in an emphasis on preventative measures against breast cancer and other cancers. With the help of Python's data-driven algorithms, society may acquire a more accurate understanding of cancer risks and make decisions that will enhance patient welfare.

Keywords: Breast Cancer, Early Detection, Risk factors, Linear regression.

1. Introduction

Breast cancer ranks among the most prevalent forms of cancer affecting women on a global basis. Based on statistical information, the phenomenon resulted in an estimated 570,000 fatalities in the entire year 2015. Cancer of the breast affects an important percentage of women globally, with an annual incidence of over 1.5 million cases, accounting for around 25% of all cancer diagnoses among females. Breast cancer accounted for approximately thirty percent (252,710 cases) of cases that were newly diagnosed in women in the United States in 2017, according to estimates [1]. In the 21st century, as society continues to advance, an increasing number of individuals are recognizing the paramount importance of health in their daily lives. The mortality rate, when considering the entire population, is recorded as 16.9 cases per 100 person-years. This discovery exhibits a bigger magnitude compared to nations with high economic levels. Significant prognostic indicators for death include advanced clinical stage, multiple medical conditions, menopausal state, and treatment with hormones [2]. When comparing with developed countries, it is evident that lifestyle has a significant role in cancer, particularly in relation to breast cancer death rates. In the context of the developing country, individuals have traditionally engaged in frequent bodily examinations. The timely detection and prompt intervention are essential in reducing the fatality rate associated with breast cancer. Cancer is a prominent health concern that

DOI: 10.54254/2755-2721/47/20241357

distinguishes itself from other medical conditions due to its significant worldwide impact. In the year 2020, cancer was responsible for about 10 million deaths, accounting for approximately one-sixth of all recorded fatalities. Breast cancer is distinguished among the other types of cancer as a large and distressing contributor to mortality. In the year 2020, breast cancer is projected to emerge as the most prevalent form of cancer on a global scale. Over the preceding five years, a total of 7.8 million women throughout the world have received a diagnosis of breast cancer. Cancer, an intricate assemblage of illnesses, comprises a multitude of varieties. Hence, prioritizing prevention, timely identification, and efficient management has substantial importance across various cancer classifications. Breast cancer has a significant prevalence, nevertheless, it is important to acknowledge the existence of other forms of cancer that also warrant considerable attention. This necessitates proactive endeavors to anticipate likely patterns in morbidity. By utilizing data, specifically incorporating variables such as age, gender, lifestyle behaviors, and dietary choices, it becomes feasible to predict the probability of acquiring different malignancies. Python programming, known for its versatility and robustness, has the potential to assist with the interpretation of this dataset. Using the Python programming language, sophisticated datasets may be analyzed and visualized, resulting in the generation of complete graphs and models. The graphical representations offer significant insights into the preventative actions that may be implemented to mitigate the risk of breast cancer and other diseases. This method enables both consumers and healthcare providers to make well-informed decisions and successfully execute preventative actions. The dynamic nature of health consciousness in the 21st century highlights the crucial significance of taking proactive actions to combat cancer, specifically emphasizing breast cancer. Using Python's powerful algorithms, such as linear regression, applying data-driven insights may greatly enhance society's comprehension of the risk factors linked to cancer. Linear regression is a very effective analytical method available in Python that enables the detection and measurement of correlations between variables. This capability is of utmost importance in the investigation of cancer risks. Using linear regression skills facilitates an improved understanding, which may subsequently guide the execution of tangible preventive actions. These metrics are crucial in protecting patients' health and well-being, based on dependable evidence obtained using algorithms.

2. Previous research

Breast cancer is prevalent in many online and offline sources, offering a substantial amount of information that may be utilized for analytical and predictive purposes. To undertake a comprehensive breast cancer prediction project, it is imperative to adopt a systematic methodology that encompasses the gathering, processing, and prediction of data. There are a total of six phases involved in the research process. The first imperative stage involves the collection of pertinent data. The criticality of guaranteeing the trustworthiness and relevance of data cannot be overstated when constructing a resilient dataset that faithfully represents the contemporary social milieu. The act of choosing reliable sources guarantees both coherence and genuineness. The aforementioned data serves as the fundamental basis for further analysis and prediction. The World Health Organization (WHO) is the main publisher of breast cancer mortality data. The complexity of the data provided by the WHO poses challenges for researchers. Consequently, only a limited set of pertinent data variables are utilized for predictive purposes. These variables include nation, year, sex, age group, number of deaths, and mortality rate per million people. The data collection process is depicted in Figure 1, where all data points are consolidated inside a single Excel sheet. After the data collection, the researcher will do the data processing to laundry the data readable. The data also gives proof that, More recently, women have an increased risk of breast cancer if they have mammographically dense breasts [3].

Table 1. Breast cancer death rate in Albania in Year 1987.

country name	year	sex	age group	number	Death rate per 100000 population
Albania	1987	All	[Unknown]	0	
Albania	1987	All	[85+]	18	180

DOI: 10.54254/2755-2721/47/20241357

Table 1. (continued).

Albania	1987	All	[80-84]	19	114.4578313
Albania	1987	All	[75-79]	34	110.3896104
Albania	1987	All	[70-74]	67	164.2156863
Albania	1987	All	[65-69]	76	117.4652241
Albania	1987	All	[60-64]	76	98.19121447
Albania	1987	All	[55-59]	70	70.56451613
Albania	1987	All	[50-54]	31	25.76891106
Albania	1987	All	[45-49]	22	15.49295775

3. Data Processing and Data Prediction

After the collection and organization of data in the Excel sheet, it frequently needs preprocessing. As seen in the diagram, the data undergoes many iterations of processing. The initial column comprised the statistics about downloads obtained from the official website of the World Health Organization (WHO). The table has been purged of extraneous data, specifically the area, country, and age group codes. Following the data cleaning process, the resultant table demonstrates data quality, which will be utilized in the section dedicated to data visualization. Table 1 only covered data about numerical values and mortality rates per 100,000 individuals within the population. During this stage of data processing, the reintegration and sorting of data is conducted to prevent data inaccuracies resulting from the removal of variables. The data analysis also revealed a congruent perspective concerning the overall level of physical activity shown by patients, encompassing both moderate and strenuous activities. Participation in regular physical activity, including aerobic activities, anaerobic exercises, and HIIT (High-Intensity Interval Training), has the potential to mitigate the risk of developing postmenopausal breast cancer. Furthermore, it has been shown that keeping an elevated level of body fat throughout early adulthood (between the ages of 18 and 30) as determined by the Body Mass Index (BMI) may serve as a preventative strategy against the development of breast cancer after menopause [4]. Based on the Python data analysis and the WHO report, being overweight may constitute another significant factor contributing to the occurrence of breast cancer in women. The significance of plant-based nutrition is particularly salient in the context of cancer prevention. Consistent adherence to a nutritious plant-based diet, along with the right inclusion of non-plant foods, proteins, lipids, and other essential ingredients, can potentially mitigate the likelihood of developing breast cancer. The range that offers the most favorable outcomes regarding risk reduction is situated within a moderate level of consumption. Nevertheless, the regular consumption of an unhealthy vegan diet may potentially elevate the likelihood of developing breast cancer [5]. The role of dietary factors is crucial in reducing the risk of breast cancer. It is recommended to prioritize the intake of a dietary pattern that is rich in nutrients and focuses on plant-based food. Based on the data gathered, it is imperative to incorporate nutritional information as a crucial element inside the table.

4. Code Implementation and Future Prediction

The prediction procedure is mainly dependent on the utilization of advanced algorithms. Python, a very adaptable programming language, is a valuable asset in developing these prediction models. The algorithms, guided by data-driven insights, yield forecasts that possess substantial utility in several fields. In addition to its implications for healthcare, the knowledge derived from these models can guide medical practitioners, researchers, and policymakers in making well-informed choices, developing efficacious preventative measures, and optimizing budget allocation. By utilizing the available data and resources, it is possible to enhance the accuracy and reliability of forecasts for breast cancer and other related types of malignancies. The prediction capabilities of the models are further enhanced by iterative improvement over time, which is based on real-world data. Figures 1 and 2 provide a solid basis for making future predictions by utilizing breast cancer data. The visualization charts derived from these

DOI: 10.54254/2755-2721/47/20241357

statistics provide significant insights into prospective patterns. The graphical representations in question are created using Python programming language. These visualizations are generated from data through a thorough and painstaking process of gathering and refinement, as described in the beginning stages of data handling and analysis. Moreover, the next prognostications must encompass a more comprehensive array of variables and supplementary data points for graphical depiction. Metastatic breast cancer is the major cause of cancer death in women. Metastatic breast cancer is well recognized as the leading factor contributing to the decline in health and mortality among women diagnosed with cancer. The lack of timely diagnosis may result in patients failing to get therapy within the most effective timeframe. Our comprehensive review of data has brought us to the conclusion that blood marker data may be utilized as a non-invasive machine learning approach in the early detection of breast cancer [6]. With a more comprehensive dataset, cancer scientists could potentially work towards reducing the mortality rate associated with breast cancer. Figure 1 displays the projected death rate for breast cancer after 2020. The data utilized in the Python programming does not account for the COVID-19 pandemic conditions. This paragraph gives a comprehensive analysis of the situation for breast cancer screening and diagnosis before and during the COVID-19 pandemic, encompassing 74 papers. None of these studies were evaluated to have the lowest risk of bias. The research indicated that screening volumes frequently declined during the pandemic, with over half of the studies reporting declines of more than 49%. Moreover, the majority of studies (66%) found a reduction of 25% or more in the number of breast cancer diagnoses, and the proportion of symptomatic patients was greater than that of screen-detected cases. During the pandemic, the distribution of breast cancer diagnoses showed a reduced proportion of early-stage cases (stage 0-1/I-II, or Tis and T1) and a larger proportion of advanced cases. However, total population rates were typically not published [7]. The COVID-19 pandemic may have introduced potential biases that might impact the veracity of the data findings about objective reality. To enhance the accuracy of future forecasts, it is imperative for the programming to incorporate a comprehensive consideration of diverse situations during the prediction process. Indeed, there is a notable emphasis on COVID-19 therapy within medical resources, leading to notable disparities between projected data and observed results.

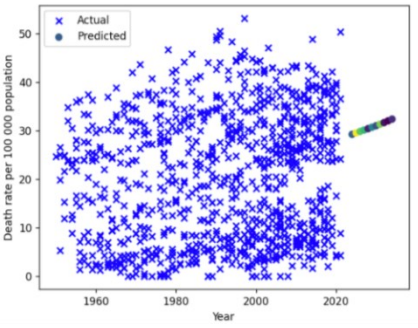

Figure 1. Predicted death rate per 100,000 population for breast cancer.

DOI: 10.54254/2755-2721/47/20241357

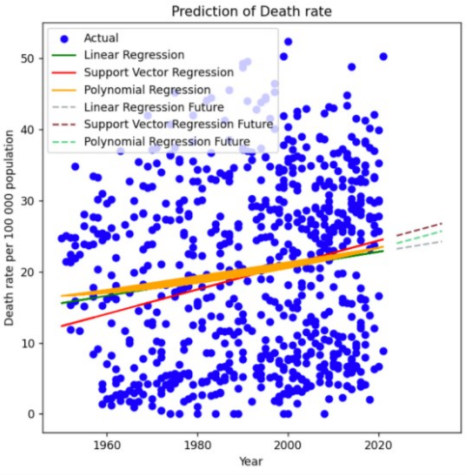

Figure 2. Predicted death rates per 100,000 population for breast cancer with different analysis

5. Discussion

To forecast future breast cancer deaths and mortality rates, the investigation employed historical data extracted from Figures 1 and 2. The results revealed a consistent trend that was originally identified while utilizing regression analysis with the SVR (Support Vector Regression) algorithm, polynomial model modeling, plus neural network techniques to forecast fatalities. The horizontal trajectory of the dashed line exemplifies the observed phenomenon. This implies that there is an anticipation of a consistent number of fatalities attributable to breast cancer in the future. The statement is supported by the progress made in medical technology, which includes improved methods of prevention and treatment that exceed previous standards, leading to significantly improved patient outcomes. It is worth noting that early-stage breast cancer demonstrates a notable cure percentage ranging from 80% to 90%, but mid to late-stage instances reveal a cure rate of around 40%. The study comprises several aspects that impact the prediction of the research outcomes. It may not be 100% accurate in representing all those factors in conjunction with other available resources results. Throughout the research, it was discovered that Argentina, Uruguay, and Venezuela had the greatest death rates within the Latin American and Caribbean regions. The countries of Guatemala, El Salvador, and Nicaragua had the most significant growth, with a rise of 0.4%. On the other hand, Argentina, Chile, and Uruguay observed a decline of 0.6%. Within the cohort of women aged below 50, it was observed that six nations exhibited a declining pattern, whereas five nations had an ascending pattern. Within the demographic of women aged 50 and over, it was observed that three nations exhibited a declining pattern, while ten nations had an ascending pattern. According to projections, it is anticipated that by the year 2030, the LAC region will experience a rise in mortality rates, with a particular focus on nations such as Guatemala, Nicaragua, and El Salvador [8]. In the LAC region, which encompasses various countries, most of the population is of Latin ethnicity. Despite this commonality, there exists a diversity of age ranges. Age constitutes a significant factor in breast cancer mortality, and its impact is notable across different regions and countries. As individuals age, breast cancer may be detected during various stages of its progression. The median age of those diagnosed with interval breast cancer was found to be 59 years, with a range of interquartile ranges spanning from 53 to 65 years. In contrast, individuals identified by the screening process had a median age of 60 years, accompanied by an interquartile range spanning from 55 to 65 years. In cases of breast cancer that were incidentally discovered, a greater percentage of the tumors were invasive and malignant (91% vs. 75%), exceeding the proportion of tumors detected by screening [9].

Additionally, these research findings are depicted in Figure 1. The following analysis applies to the prediction of mortality. The graph's dotted line clearly demonstrates an upward trajectory in the mortality rate, which contrasts the reported trend in death data. The transition is intricately linked to the demographic composition of the population. The pandemic's start precipitated a decline in fertility rates,

DOI: 10.54254/2755-2721/47/20241357

resulting in constrained population expansion. Simultaneously, the presence of an aging population leads to an increase in death rates. As a result, forthcoming predictions suggest an impending increase in death rates. Breast cancer is a heterogeneous disease that comes in several clinical and histological forms. Its clinical progression is difficult to predict using the current prognostic factors and its treatment is therefore not as effective as it should be [10]. Breast cancer has the potential to present itself in a range of clinical and histological variations, indicating that individual instances may display unique clinical manifestations and pathological features. The accurate prediction of the clinical progression of breast cancer is a significant challenge due to the limitations of known prognostic variables in providing exact projections. The probable consequence includes a potential restriction on the range of available treatment options and a potential decrease in the efficacy of treatment regimens. In prospective data forecasting, it is plausible to incorporate supplementary variables such as dietary patterns, levels of physical activity, unanticipated public health occurrences, and emotional welfare to augment the precision of the prognostications.

6. Conclusion

In conclusion, this research underscores the importance of considering both fatality and mortality rates when assessing the impact of breast cancer. The study also indicates that many additional factors, such as patient gender, emotional state, residential circumstances, and dietary habits, can potentially influence the onset of breast cancer. It is important to note that the research's prediction results are based on a constrained collection of variables and that more variables must be included if the forecasts are to be more accurate.

These results make it clear that additional studies and healthcare treatments are urgently needed to address the changing problems caused by breast cancer in the future.

References

[1] Sun, Y. S., Zhao, Z., Yang, Z. 2017 Risk factors and preventions of breast cancer Inter. J. of Bio. Sci. 13(11) 1387

[2] Misganaw, M., Zeleke, H., Mulugeta, H., and Assefa, B 2023 Mortality rate and predictors among patients with breast cancer at a referral hospital in northwest Ethiopia: A retrospective follow-up study Plos. One. 18(1) e0279656

[3] Cappello, N. M., Richetelli, D., and Lee, C. I. 2019 The impact of breast density reporting laws on women's awareness of density-associated risks and conversations regarding supplemental screening with providers J. Am. Coll. Radiol. 16(2) 139-146

[4] Wiseman, M. J. 2019 Nutrition and cancer: prevention and survival Brit. J. Nutr. 122(5) 481-487

[5] Shah, S., Mahamat-Saleh, Y., Ait-Hadad, W., Koemel, N. A., Varraso, R., Boutron-Ruault, M. C., and Laouali, N 2023 Long-term adherence to healthful and unhealthful plant-based diets and breast cancer risk overall and by hormone receptor and histologic subtypes among postmenopausal females The Am. J. of Clin. Nutr. 17(3) 467-476

[6] Botlagunta, M., Botlagunta, M. D., Myneni, M. B., Lakshmi, D., Nayyar, A., Gullapalli, J. S., and Shah, M. A. 2023 Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms Sci. Rep. 13(1) 485

[7] Li, T., Nickel, B., Ngo, P., McFadden, K., Brennan, M., Marinovich, M. L., and Houssami, N. 2023 A systematic review of the impact of the COVID-19 pandemic on breast cancer screening and diagnosis The Breast

[8] Torres-Román, J. S., Ybaseta-Medina, J., Loli-Guevara, S. 2023 Public Health Disparities in breast cancer mortality among Latin American women: trends and predictions for 2030 Y BMC 23(1) 1-9

[9] Celeste Damiani, Grigios Kalliatakis, Muthyala Sreenivas, 13 June 2023 Evaluation of an AI model to assess future breast cancer risk Evaluation of an AI Model to Assess Future Breast Cancer Risk

[10] Bertucci, F. and Birnbaum, D 2008 Reasons for breast cancer heterogeneity J. of bio. 7 1-4