|  |
| --- |
| **Risk of Developing Breast Cancer in US Women [[1]](#footnote-2)**  Landon Leininger 1, Isabella Spinelli 1, Demetri Tsitsios 1  1 School of Biomedical Engineering, Science, & Health Systems, Drexel University, USA  Course : BMES 550  Instructor: Ahmet Sacan  Date : 2022-12-07 |

[[2]](#footnote-3)\*abstract

Breast cancer is a disease that affects a large number of women in the United States, and it is a disease that has numerous risk factors associated with it. In using a woman’s known risk factors to determine her risk of developing breast cancer, the woman can be more knowledgeable of this risk and work to control certain risk factors that she is capable of influencing. This graphical user interface (GUI) calls for the user to input her own risk factors, such as age, race/ethnicity, etc., and it outputs the percentage of known US women with these same risk factors who have a history of breast cancer. This application involves a MATLAB App Designer GUI with numerous dropdown menus corresponding to different risk factors. It interacts with an SQL database of mammogram data from US women. A separate MATLAB function searches women in this database who have risk factors associated with those inputted by the user, and it outputs the subsequent percentage who have a history of breast cancer. The results are shown in a pie chart with up to three sections: percentage of women with a history of breast cancer, percentage of women with no history of breast cancer, and percentage of women with an unknown history of breast cancer, all with the risk factors inputted by the user.

# introduction

Aside from heart disease, cancer is the leading cause of death in the United States [1]. While there are countless different types of cancer, breast cancer is the second most common cancer to affect US women. Around 264,000 US women are diagnosed with breast cancer each year in the United States, and about 42,000 die each year as a result of it [2]. Therefore, this disease not only affects a large number of women in the United States, but it affects their loved ones as well.

There are a variety of risk factors associated with the risk of developing breast cancer. These risk factors include age, breast density, and familial history of breast cancer. While women do not have control over these risk factors, there are also other risk factors that they can control. These include amount of physical activity, weight, hormone usage, reproductive history, and drinking alcohol. Women who are not physically active, obese, use hormone replacement therapies, have children after the age of 30, and drink alcohol regularly are more likely to develop breast cancer than others [3].

The target population for this application is US women. With this application, they will be able to input their own risk factors that are associated with developing breast cancer. These risk factors include age, race/ethnicity, familial history of breast cancer, age at first menstrual cycle, age at first birth, breast density, hormone replacement therapy status, menopausal status, and body mass index (BMI). This application will then search a database of over six million mammograms of US women, and it will output the percentage of women with these same risk factors who have a history of breast cancer. While some risk factors are out of the user’s control, the results outputted by this application can influence the user to control the risk factors that they are able to change. Furthermore, this application can be used over time as the user’s risk factors change. If this application is successful, it will be a useful tool for calculating risk of developing breast cancer based on real-world data.

Other breast cancer risk calculators are publicly available, including one provided by the National Cancer Institute. However, this tool lacks consideration of a handful of risk factors, including breast density, hormone replacement therapy status, menopausal status, and BMI [4]. Our application is more comprehensive as it accounts for the risk factors highlighted in the National Cancer Institute’s tool as well as these missing risk factors.

# Dataset

This dataset was provided by the Breast Cancer Surveillance Consortium (BCSC). Mammograms are performed on US women, and subsequent data on risk factors are collected during these mammograms. With patient consent, individual patient data is submitted to the BCSC and inputted into this database. There are no experimental groups in this dataset as the collected data is purely observational.

# methods and IMplementation

There are three programming languages used in this project: MATLAB, MATLAB App Designer, and SQLite. A GUI was produced using MATLAB R2022b App Designer that allows users to input their risk factors for breast cancer. Once the user selects any combination of risk factors, a pie chart of the percentage of known US women with breast cancer history, without breast cancer history, and unknown breast cancer history is displayed. The flowchart of the user GUI selection and risk calculation is shown below in **Figure 1**. The database was created by taking three (.csv) data files from the BCSC website and adding them to an SQLite database using a MATLAB function (dbcreation.m) [5]. A connection to this database was also created using a separate MATLAB function (dbconnection.m). Another MATLAB function (riskcalc.m) utilizes SQLite queries to return the proper columns of the database in order to determine the number of patients within the database that had a previous breast cancer diagnosis, did not have a previous breast cancer diagnosis, and those who do not know. This is done after the user inputs an array of risk factors. This same function then calculates the percentages of people from each of these three categories for this array of risk factors. Once the percentages of women are calculated, the GUI displays the results in a pie chart.

Database

The database consists of 6,788,436 entries of mammogram data, where each row represents a unique combination of risk factors, whether this particular combination resulted in a breast cancer diagnosis (or unknown), and the frequency of that unique combination. As such, the database resulted in 1,522,340 rows of unique combinations. The nine risk factors which the GUI inquires are as follows: age group, race/ethnicity, whether the patient has a first degree relative with breast cancer, the age of menarche (first menstrual cycle), the age of first child birth, BI-RADS breast density, whether the patient is currently undergoing hormone replacement therapy, the patient’s menopause status, and BMI group. In addition to these columns, two other columns consist of the year in which the mammogram was taken and whether or not the patient had undergone a previous breast biopsy or aspiration, neither of which are considered risk factors. All the entries of the database are numeric and have distinct meanings which are defined on the BCSC website. In general, the risk factors are allocated into groups which correspond to a number; for risk factors which are an answer of “yes, no, or unknown,” the database designates 1, 0, and 9, respectively, to represent those answers. The database has two additional columns for breast cancer history and count. The breast cancer history column represents whether the patient with those particular risk factors has a history of breast cancer. The count column represents the frequency of the unique combination [6].

SQL Query and Percentage Calculation

An SQL query is spliced together in the riskcalc.m MATLAB function depending on which risk factors the user selects in the GUI dropdown menus. The function will query SQLite first for the breast cancer history and count for all the rows where the risk factors match the user input. If the query finds rows which correspond to the user input, a table will be loaded into matlab. Then, the frequency of patients who have a history of breast cancer, no history of breast cancer, and an unknown history of breast cancer will be compiled. This will then be used to calculate the percentages of the patients selected for these three categories.

If the table initially loaded into MATLAB is empty, this indicates that the particular combination of user-inputted risk factors is not present in the database. So, an estimation is calculated based on the next-closest match. A new SQL query is spliced where the user age group is subtracted to all rows of the age group column of the database via a difference. For any other risk factor, a logical comparison of whether the user input does not match the database is found. The distinction is, a difference in age is a difference in years, whereas a difference in ethnicity is only an indication of whether or not the ethnicity is the same. As such, these comparisons are added up as a new column called distance (dist), and the first 10,000 rows which have the lowest value of distance are selected. These rows are then used to calculate a percentage of the patients who have a history of breast cancer, no history of breast cancer, and an unknown history of breast cancer.

Results

Once the risk calculation function completes, the percentages are outputted and used in the MATLAB App Designer function, pie(). The pie chart updates whenever the user changes risk factors via the dropdown menus and presses the “Calculate Risk” button.

Diagram

Description automatically generated

**Figure 1. GUI Flowchart.** This figure displays a flowchart of the breast cancer risk GUI application.

# Experiments and REsults

Two user cases of different breast cancer risk factors were analyzed using our MATLAB GUI. After selecting the appropriate risk factors from a dropdown menu, the risk calculation is performed and plotted as a pie chart back to the user on the GUI. The nine different parameters used for the breast cancer GUI are based on the nine different risk factors captured in the BCSC dataset. The first user case explores how the GUI works when certain risk factors are not selected. The results can be found in **Figure 2**.

Graphical user interface, application

Description automatically generated

Chart, pie chart

Description automatically generated

**Figure 2. GUI Display for Partial Risk Factors.** The top image displays the GUI with only four risk factors selected before submission. The bottom image displays the GUI after the submission button is clicked.

The second user case shows how the GUI analyzes the risk factors when all nine risk factors are inputted by the user. The resulting risk factor percentages can be found in **Figure 3**.

Graphical user interface, application, email

Description automatically generated

Chart, pie chart

Description automatically generated

**Figure 3. GUI Display for All Risk Factors.** The top image displays the GUI with all nine risk factors selected before submission. The bottom image displays the GUI after the submission button is clicked.

Based on the risk factors selected by the user, different risk estimations are produced. The GUI outputs the percentages of women in the database with a history of breast cancer, without a history of breast cancer, and an unknown history of breast cancer. The percentages belong to the population of patients in the BCSC database that possess the same risk factors inputted by the user. If an exact match cannot be determined, the risk calculation identifies the next closest match and outputs the corresponding percentages back to the user. If no dropdown is selected, the GUI will return an error message over the previous results.

# DISCUSSION

The results show that the GUI successfully calculates percentages of patients in the database with varying history (or unknown history) of breast cancer. This can signify to the user of the GUI that their particular combination of risk factors could mean they are more likely to develop breast cancer. This is solely based on the frequency of patients with the same or similar risk factors and how many patients within the database actually had a previous breast cancer diagnosis.

Currently, common risk factors used to screen women for breast cancer include age, race/ethnicity, and first degree relation to someone with a previous breast cancer diagnosis [3]. These factors are most commonly collected from women by their healthcare providers. However, in the example shown in **Figure 2**, the GUI displays a 0% for US women who had previously been diagnosed with breast cancer. The likelihood that there really is a 0% for any category in the real world is low. Therefore, this demonstrates the limitations of the GUI.

Major limitations of the study include the limited sampling frame of the database. The BCSC database consists of approximately seven million data entries, which equates to only 1.5 million unique combinations of risk factors. The stark similarities in the risk factors among all the participants could be because they are from specific cities in only a limited number of states in the United States (WA, NH, VT, NC, IL, and CA). This study could be improved if patient data was collected from more varied locations across the United States and possibly the world. With this improved database, more unique combinations of risk factors could be obtained, and less estimations of risk would need to be calculated via the riskcalc.m nearest neighbor distance calculation.

To improve the application of this GUI, an additional risk factor could be added to the patient data collection and, subsequently, the risk percentage calculation. The second to fourth digit ratio (2D:4D) has been found to have a direct association to breast cancer risk due to prenatal testosterone [7]. With this easy-to-obtain risk factor, more users of this GUI can input this information and obtain a more accurate measurement of risk.

# References

1. Centers for Disease Control and Prevention. (2022, September 6). *FastStats - Leading Causes of Death*. Centers for Disease Control and Prevention. Retrieved December 3, 2022, from https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

2. Centers for Disease Control and Prevention. (2022, September 26). *Basic Information About Breast Cancer | CDC*. Centers for Disease Control and Prevention. Retrieved December 3, 2022, from https://www.cdc.gov/cancer/breast/basic\_info/index.htm

3. Centers for Disease Control and Prevention. (2022, September 26). *What Are the Risk Factors for Breast Cancer? | CDC*. Centers for Disease Control and Prevention. Retrieved December 3, 2022, from https://www.cdc.gov/cancer/breast/basic\_info/risk\_factors.htm

4. National Institutes of Health. (n.d.). *Breast Cancer Risk Assessment Tool*. National Cancer Institute. Retrieved December 3, 2022, from https://bcrisktool.cancer.gov/calculator.html

5. Breast Cancer Surveillance Consortium. (n.d.). *Risk Factor Dataset Download*. BCSC. Retrieved December 6, 2022, from https://www.bcsc-research.org/data/rf/risk-factor-dataset-download

6. Breast Cancer Surveillance Consortium. (n.d.). *Risk Factor Dataset Documentation*. BCSC. Retrieved December 6, 2022, from https://www.bcsc-research.org/data/rf/documentation

7. Muller, D. C., Baglietto, L., Manning, J. T., McLean, C., Hopper, J. L., English, D. R., Giles, G. G., & Severi, G. (2012). Second to fourth digit ratio (2D : 4D), breast cancer risk factors, and breast cancer risk: A prospective cohort study. *British Journal of Cancer*, *107*(9), 1631–1636. https://doi.org/10.1038/bjc.2012.418

1. [↑](#footnote-ref-2)
2. [↑](#footnote-ref-3)