Gail Score Project

Cynthia Kineza

I. Introduction:

- Breast Cancer

Breast Cancer occurs when there is a collection of cells that grow abnormally or without control, called tumors. It is the second most common cancer affecting women. About 231,840 cases of invasive breast cancer and 60,290 of non-invasive breast cancer were diagnosed in the United States in 2015 alone[4]. Some of the factors that increase the risk of breast cancer are: gender, age, race, family history[4]. The latest statistics of breast cancer in the United Statesa are[3]:

- 1 in 8 women (about 12%) will develop breast cancer at some point in her lifetime.
- In 2016, an estimated 246,660 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., in addition to 61,000 new cases of non-invasive breast cancer.
- For women in the U.S., breast cancer death rates are higher than those for any other cancer(besides lung cancer)
- A woman's risk of breast cancer nearly doubles if she has a first-degree relative who has been diagnosed with breast cancer
- The most significant risk factors for breast cancer are gender (being a woman) and age (growing older). Given the above statistics,creating tools to assess risk level for breast cancer is primordial. The more you know, the better prepared you are to take actions that can help reduce your risk. Fortunately, the Gail Model was designed to educate women about their potential risk for breast cancer.
- Projet Overview and Goal

  The Gail Model is a statistical breast cancer risk assessment algorithm which was developed by Dr. Mitchell Gail and his colleagues with the Biostatistics Branch of NCI's Division of Cancer Epidemiology and Genetics. It was developed as part of a huge screening study of 280,000 women between 35 and 74 years of age. Initially, The Gail model was shown to be a reasonable tool for estimating breast cancer risk in white women, and later on other researchers subsequently supplemented the model to provide accurate risk assessments for African, Hispanic, and Asian women [2] using data from the Contraceptive and Reproductive Experiences (CARE) Study for African american women and data from the Asian American Breast Cancer Study (AABCS) for Asian-American and Pacific Islander women The Gail Model calculates a woman's risk of developing breast cancer within the next five years and within her lifetime (up to age 90 years old)[1]. The goal of this project is to develop a model for inferring the variables needed to calculate the Gail score for a woman based on her Facebook profile. In addition, include estimates of uncertainty in the predicted score. I will also observe any trends in absolute risks based on age and race.
- Description of Data

Data being used is obtained from Facebook page sites, "IhadCancer" and "TheBreastCanceSite". The focus was to extract comments and posts from those breast cancer support groups, and extract demographic information of women. The information extracted was used to estimate the Gail Score.

- Key Variables:

  Age ,Age at time of first menstrual period ,Age at time of first live birth of a child ,Ethnicity, Medical History,Mutation in either BRCA1 or BRCA2 gene,How many first degree relative have had breast cancer and *Breast Biopsy*

- Key Variables received from facebook:
  First and Last Names
- Variables created using Facebook Names:
  Age ,*Race* and *Ethniciy*
- Limitations in Data Extraction

The biggest hurdle was finding specific information about the Gail score. Variables such as mutation genes, medical history, breast biopsy age at time of first menstrual period were not easily available on a social media site like Facebook while some other variables were much easier to find (age, race/ethnicity, gender) using various R packages and the Census Bureau public data. Although all the variables were not utilized, it was sufficient to make inferences using estimated race,gender and age.

- Tools
  - Rfacebook package provides a series of functions that allow R users to access Facebook's API to get information about users and posts, and collect public status updates
  - Diverse R packages and functions to extract, clean and transform race, age and gender information ("Gender","WRU" packages)

II. Exploratory Data Analysis

- Data cleaning

Data Cleaning Process was consisted of recognizing and separating human names from organization names, and removing the later. Finding the gender and ethnicity of the participants using packages "wru" and "gender", using Social Security Administration to extract name frequency during a certain period of time so we can estimate age of participants based on their first names, using "BCRA"(Breast Cancer Risk Assessment tool) package to extract the absolute risk. The model used to predict the gail score of women was based of outputs generated by the "BCRA" package.

I ended up with 1,051 Women after going through data cleaning. I considered women that are older than 35(Breast Cancer Risk Assessment Tool only calculates risk for women 35 years of age or older), but younger than 90 (I used a limit of 90 years old since it is considered the limit of a lifetime by NIH (National Health Institute)).

III. Methodology/Analysis

- Coefficients of Interest

In order to calculate the absolute risk of the Gail Score Model, a dataset was constructed containing all the required input data needed to perform risk projections: ID, Initial Age,BrCA projection age, Number of Biopsies, Did biospy display atypical hyperplasia?, Age at menarchy, Age at first live birth , number of first degree relative with BrCa and Race. The followwing variables were missing in my final dataset but did not present an obstacle into calculating the absolute risk: number of biopsies, Age at first live birth, age at menarchy and the number of 1st degree relatives with BrCa; the missing columns were replaced with the number 99 to show missing value.

In order to account for uncertainty in age and race, a list of combinations were created to generate absolute risks:

- Using the range of ages (all plausible ages of a woman from first to third quartile using the Census bureau database of firstnames), and a fixed race, a range of potential absolute risks was derived
- Using a range of races (obtained through probabilities generated by package "wru" using surnames as indicators giving a list of probabilities that the individual is "white", "black","hispanic" or from "other" race) and the median estimated age, a range of potential risk was derived
- Using both the plausible ranges of age and race, a range of potential risk was calculated

(Attached below are the figures of two women selected to demonstrate the trend)

Moreover, by using the variability in both age and race ranges, we obtained variability in absolute risks with a projection in 5 years and a lifetime risk. Note that here in calculating the lifetime risk we fix the projection age of BrCA to 90 years, while a regular projection in 5 years is obtained by adding 5 years to the actual estimated age.

All of those ranges in age and race are done by taking into account uncertainty in the absolute.

Figure 1: Absolute risks density for two women

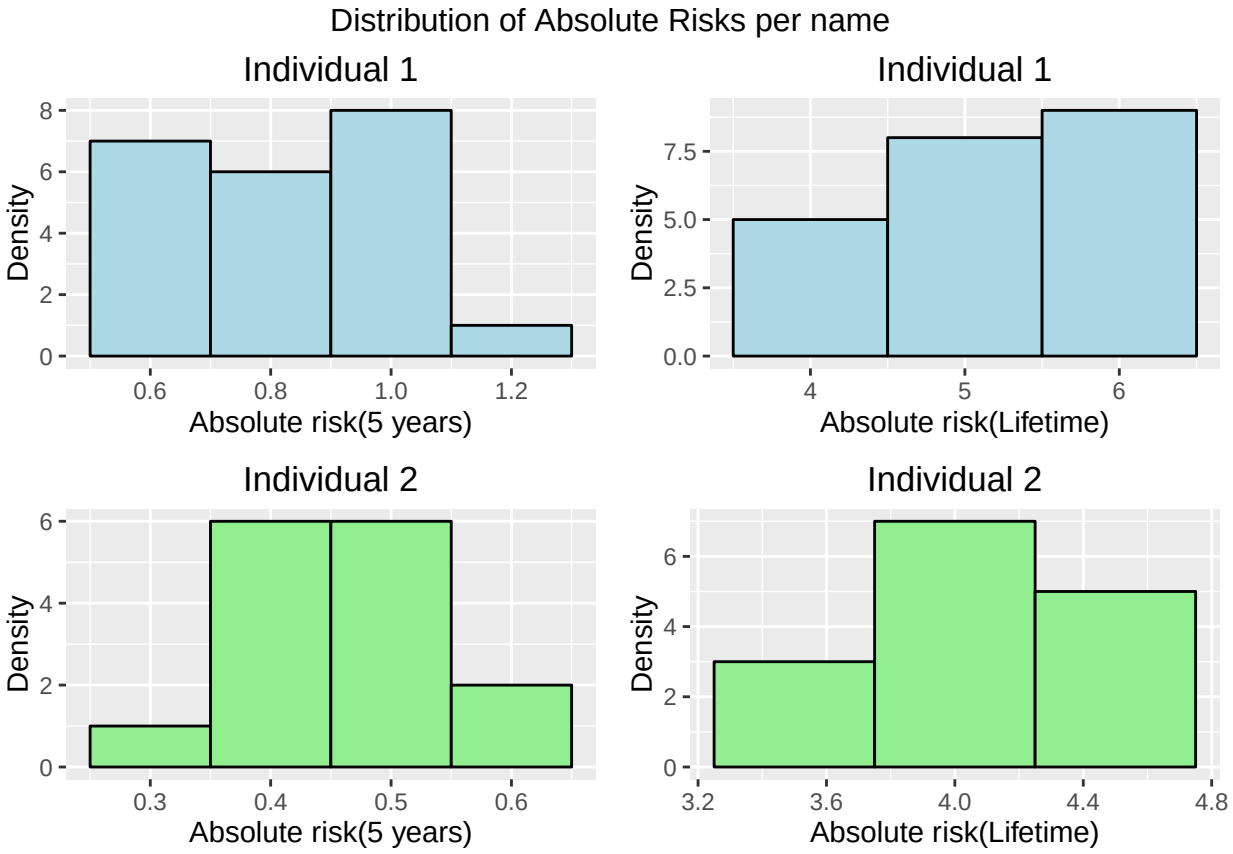
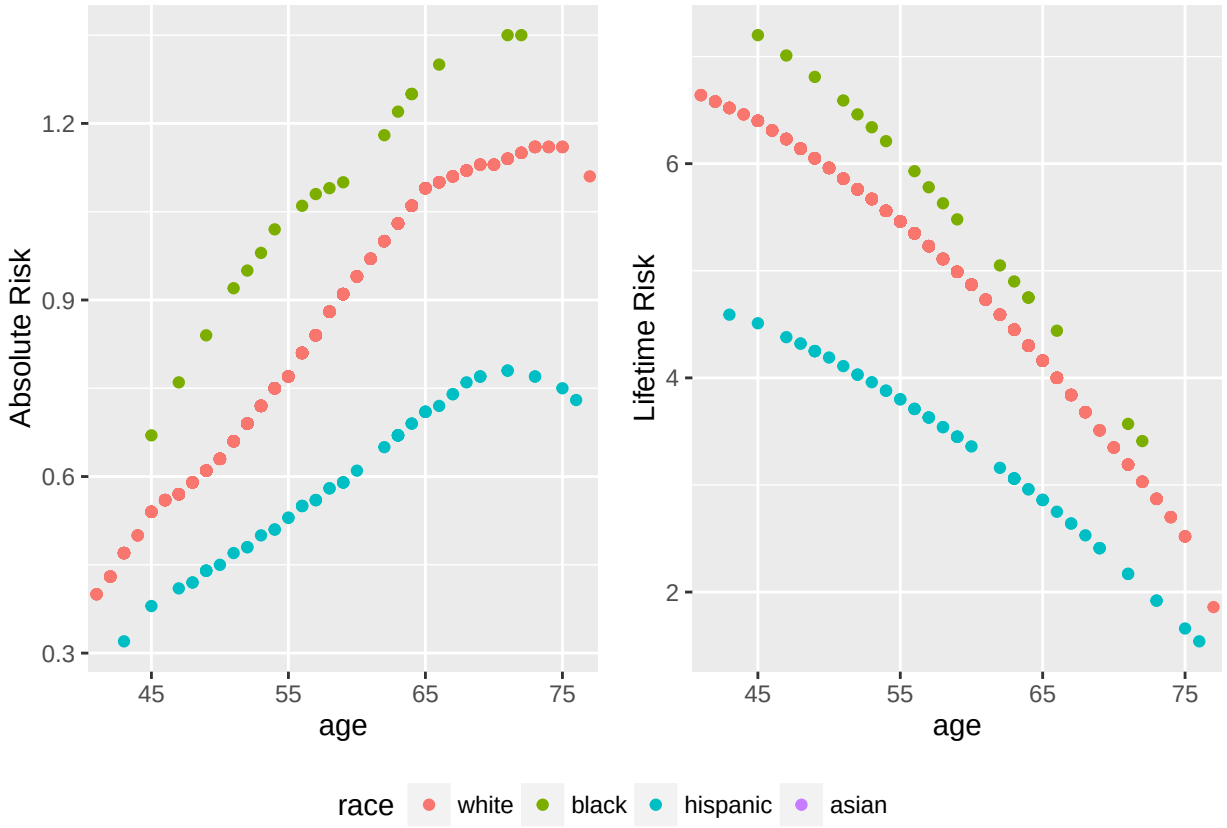


Figure 2



Observation: absolute risk increases as age increases, and it looks like black women have a higher absolute risk in this particular sample size. In fact, based on recent findings, African American women are more likely than white women to develop breast cancer before age 40 years, to be diagnosed with estrogen receptor (ER) –negative tumors, and to have higher 5-year breast cancer mortality rates[7].

IV. Measure of Uncertainty

The measure of uncertainty is measured for age and race. For race, using the package "gender", we have a range of probabilities for being white, hispanic and black. Those probabilities were used to estimate uncertainty in absolute risk. For age, we estimate the 1st quartile, median and third quartile and estimate the variability of the obsolute risk based on the variability of age. I came up with an estimated age and an estimated race. Then the risk was estimated as the range of risks corresponding to the range of ages that are plausible from the estimate of the age. By finding an estimate of age, I derived an estimate of confidence interval for the age. Then evaluated the range of potential risks for the range of potential ages. Note that the actual gail score is obtained by estimation, therefore the range of absolute risks of a woman has been based on her estimated age and race.

Figure 3:

Table of confidence intervals of absolute risk of these 4 subjects: Life.1qrt represent first quartile of absolute risk given combination of possible race and age values, while 5yearsRisk.1stQ is for 5 year projection first quartile of absolute risk given combination of possible ranges and ages.

I estimated the confidence interval of potential absolute risks that were generated from a range of age and race.

Firstname	Lastname	Life.1stQ	Life.Median	Life.3rdQ	5yearsRisk.1stQ	5yearsRisk.Median	5yearsRisk.3rdQ
Alice	B.	2.74	3.65	4.73	0.77	1.07	1.15
Alyce	J.	3.22	4.31	5.64	1.2	2.84	3.38
Anita	G.	4.09	4.88	5.69	1.52	2.47	3.2
Ann	H.	3.68	4.45	5.47	0.88	1.72	2.14

V. Interpretation of Results

- Summary

For a given woman, her absolute risk increased based on her age and race. According to the Susan Komen webpage,a Foundation dedicated to education and research about causes, treatment, and the search for breast cancer cure, women with a 5-year risk of 1.67 percent or higher are classified as "high-risk."[5]. In the above table, the 95% confidence interval of Alice B's absolute risk of 5 years projection falls between 0.77 and 1.15, while for Ann H. her confidence interval is [0.88,2.14], therefore we could make assumptions that Ann H. has a somewhat higher risk than Alice B considering where the ranges of possible absolute risks fall in. These kind of comparaison can be made to assess risk levels of different women in our generated dataset.

VI. Limitations

Finding medical history has been challenging, as filtering based on key words can mislead especially for large datasets. I therefore decided to ignore some variables and made the necessary adjustment when estimating the Gail Score. Another limitation has been based on age, I am calculating the age based on Social Security Administration based data, of names frequency. The biggest issue here is that we are estimating the age, therefore lacking precision. In addition, there is a limitation in estimating race, as I estimated race based solely on surnames, which leaves room for errors and uncertainty.

VII. Challenge Results

Although the tool can estimate your risk, it cannot tell whether or not you will get breast cancer.In fact, even if a woman's risk may be accurately estimated, these predictions do not allow one to say precisely which woman will develop breast cancer. The distribution of risk estimates for women who develop breast cancer overlaps the estimates of risk for women who do not.[8] Therefore it should be used with precaution.

VIII. Discussion/Future Work

It would be useful if women between 35 and 90 utilized the information provided by the Gail Score so that they can take the appropriate measure to prevent having the disease. Even tough the model doesn't tell a person if they will get breast cancer it calculates the risk. Knowing that a person has a high risk to have cancer due to family history or their age, it can motivate them to do screening more frequently. Furthermore,it would have been interesting to see the trend in absolute risk if information such as family history, Age at time of first menstrual period,Age at time of first live birth of a child, and number of biopsies were easily available on Facebook, I omitted them in my analysis as i found that using just the messages posted in facebook wouldn't provide enough information or could make false assumptions on medical status of an individual. Nevertheless, the Gail Score calculation had just the minimum required information (age, race and gender) to make appropriate inferences.

IX. Works Cited

(1) Breast Cancer Risk Calculations: The Gail Model. Steven Halls. October 31, 2016

(2) Breast Cancer Risk Assessment Tool. National Cancer Institute. May 16, 2011

(3) The Breast Cancer.org.September 30, 2016

(4) All About Breast Cancer.Christopher Dolinsky MD,Christine Hill-Kayser MD,Karen Arnold-Korzeniowski BSN RN.Oncolink. January 20, 2016

(5) Estimating Breast Cancer Risk.Susan G. Komen.October 28, 2016

(6) Stackoverflow. various tutorials

(7) Prospective Approach to Breast Cancer Risk Prediction in African American Women: The Black Women's Health Study Model. Deborah A. Boggs, Lynn Rosenberg, and Julie R. Palmer, Slone Epidemiology Center at Boston University, Boston, MA; and Lucile L. Adams-Campbell, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC.Journal of Clinical Oncology. March 20 1025

(8) Implementation of Breast Cancer Risk Assessment Tool using SAS®. Yuqin Li, Lihua Chen,Xiaohai Wan, Alan Chiang.PharmaSUG2013 – Paper PO06

X. Main Packages

(1) wru. https://cran.r-project.org/web/packages/wru/wru.pdf

(2) Gender.https://cran.r-project.org/web/packages/gender/gender.pdf

(3) BCRA. https://cran.r-project.org/web/packages/BCRA/BCRA.pdf

(4) RFacebook. https://cran.r-project.org/web/packages/Rfacebook/Rfacebook.pdf

(5) tm. https://cran.r-project.org/web/packages/tm/tm.pdf

(6) Parallel. //http://gforge.se/2015/02/how-to-go-parallel-in-r-basics-tips/

(7) Cluster. //https://www.r-bloggers.com/how-to-go-parallel-in-r-basics-tips/

Special Thanks Thanks to fellow classmate Prosenjit Kundu for his insights. Our discussions were very valuable.