Johns Hopkins Bloomberg School of Public Health

Advanced Data Science - I

Gail Score Prediction based on Facebook Profiles

Author:
Prosenjit Kundu

Supervisor:
Dr. Jeff Leek
Dr. Elizabeth Colantuoni
Dr. John Muschelli

A project submitted in fulfillment of the requirements
for the Advanced Data Science - I

in the

Dept. of Biostatistics

October 28, 2016

Abstract

The main grail of this project is to estimate the score(the absolute risk of breast cancer) of an woman based on her facebook profile. Data is collected from a post in the The Breast Cancer Site page of facebook. The subjects considered are the first 2000 users who commented on the post. Based on some assumptions, covariates such as gender, race and age are predicted from names, surnames and first names of a subject using information from Social Security Administration. The Gail model is implemented to estimate the absolute risk for five years and lifetime risk for each of the subjects. The uncertainity in the gail score for each subject is calculated by considering the uncertainity in age and race.

Keywords : Absolute Risk, Breast Cancer, Facebook, Gail Model, Median, Name-Age Calculator, Quantiles.

Acknowledgements

I would like to express my deepest gratitude to Dr. Elizabeth Colantuoni, Dr. Jeff Leek and Dr. John Muschelli whose constructive suggestions helped me to being motivated according to the flavor of the given problem and also helped me to delve into many nooks and corners. I am also thankful to Kayode Sosina for his guidance in parallel computing.

Contents

List of Figures

List of Tables

Chapter 1

Gail Score Prediction

1.1 Introduction

In the U.S., breast cancer is the second most common cancer in women after skin cancer. It can occur in both men and women, but it is rare in men(National Cancer Institute). Breast cancer is caused due to genetic effects(mutation of BRCA1 and BRCA2) and environmental effects. Absolute risk of breast cancer or the Gail risk score is defined as the likelihood that a person who is free of breast cancer at a given age will develop that cancer over 5 years. An woman with a risk score of 1.4 over 5 years will be interpreted as out of a population of women with the same covariates(same age, race, etc), 1.4 % of the women will develop breast cancer in 5 years. It does not say anything about which of the in the population is going to develop breast cancer.

Women who have a Gail risk score of 1.66 or higher have a higher than average risk for developing breast cancer.

Significance: It is very important to know the risk of having a breast cancer so that the lifestyles or eating habits could be changed, appropriate medicines could be taken before the start of cancer and many other things causing cancer could be avoided accordingly.

Gail score is calculated based on the Breast Cancer Risk Assessment Tool from National Cancer Institute. The seven risk factors include age, age at first period, age at the time of the birth of her first child (or has not given birth at all), family history of breast cancer (mother, sister, or daughter), number of past breast biopsies, number of breast biopsies showing atypical hyperplasia, and race/ethnicity. In this project women with no medical history of breast cancer are considered .

Objective : Estimate the absolute risk of breast cancer for five years and lifetime risk(upto 90 years) for a woman. Estimate the uncertainity in the score.

1.2 Data

Data is collected from the Facebook post. Top 2000 users who commented on that post is taken as the sample for this project. The data is read into R by creating an account in the facebook developer API and then using the package Rfacebook. The following information for each of the users(of the 1999 users) or subjects are available from the Facebook:

- ID
- name
- first name
- last name
- picture url
- comments

Required variables for calculating the absolute risk :

- Medical history(yes or no)
- Age(35 - 90 years)
- Race

1.2.1 Processing the data

- Gender of the subject is determined from the names of the subjects using gender package in R which uses the information from the U.S. Social Security Administration baby name data. The proportion of females and males are calculated from the baby data from 1932 to 2012 for a particular name. Based on the proportions the person with that name is classified as male(if the proportion is > 0.5) or female. Male users are eliminated as the breast cancer for woman is considered. It worked fine with no misclassification based on a random subsample(size 25) of the given sample by just looking into the profile pics.

- Presence of medical history of breast cancer for an woman is determined by looking some of the appropriate words/sentences like "B/breast C/cancer S/survivor", "B/breast C/cancer F/free", "B/breast C/cancer D/diagnosed" in the comments assuming women commented about themselves. Women with medical history of breast cancer are removed as the gail score/absolute risk is calculated for woman free of breast cancer.

- Age is determined from the first names using the Name-Age Calculator which uses information from social security records(after 1985). The empirical distribution is seen for a particular first name and the median age is taken to be the estimated age of that woman with that first name. To get an uncertainity in the age, the 25th and 75th quantiles of the empirical distribution are taken as the bounds to form a set of probable ages. Women with first names that are not there in the records of the name age calculator are eliminated.

- Race/ethnicity is determined from the surnames assuming the surnames of the target women are same as of that before their marriage. Probabilities for five categories of ethnicity("White", "Black", "Hispanic", "Asian", "Others") are calculated using a Bayesian approach from the wru package in R by using information from the U.S. Census 2000 Surname List and Spanish Surname List. It estimates the prior probabilities $P(R_i = r|S_i = s)$ by calculating the proportions. For any surname S_i that appears on the Spanish Surname List, we set the prior probability to 1 for Latinos and 0 for every other racial group. The most probable race among the five is taken as the estimated race for the woman with that surname. Predicted races with probability $<$ 0.01 are ignored and the rest are taken as a set of probable ages to account for uncertainity in race.

- The final data set on which the gail score and its uncertainity is estimated thus contains the names, first names, last names, median age(with a range of probable ages) and most probable race(with a range of probable races) for each woman.

1.3 Method

The probability of an woman at age α with an age-dependent relative risk $r(t)$ who will develop breast cancer by age $\alpha + \tau$ is determined by

$$P(\alpha, \tau, r) = \int_{\alpha}^{\alpha+\tau} h_1(t)r(t)e^{-\int_{\alpha}^{t} h_1(u)r(u)du} \left\{ \frac{S_2(t)}{S_2(\alpha)} \right\} dt \qquad (1.1)$$

where h_1 is the baseline age-specific hazard of developing breast cancer, S_2 is the probability of surviving the death due to other causes, that is surviving the competing risks up to age t, S_1 is the probability of surviving the death due to breast cancer. The details are provided in the supplementary material.

- The absolute risk for five years and lifetime risk is calculated from the above formula where the coefficients (that is the β's) in the logistic regression are taken as the estimates estimated from a case-control data for different races .
- For an woman, a probable range of risk scores is calculated by taking different combinations of probable age and race.
- The 2.5th and 97.5th quantiles of the probable risk scores are taken to get a confidence interval for the risk score.

1.4 Results and Interpretations

1.4.1 Tables

1.4.2 Figures

1.4.3 Interpretation

- From Table 1.1, we see that the estimated(median) gail score for Marilyn Muller is 1.11 which says that 0.0111 % of the population of woman who are white, aged 66 years, with no medical history of breast cancer and with no information on age at first period, age at the time of the birth of her first child (or has not given birth at all), family history

TABLE 1.1: Estimates of absolute risk(5 years) with confidence interval for four women.

Name	Median Age (in years)	Most Probable Race	Absolute risk (5 years)	Confidence Interval
Marilyn Muller	66	White	1.11	(0.27, 1.35)
Denise Gallego Moreno	52	Hispanic	0.48	(0.24, 1.07)
Sue Wedin Jones	63	White	1.03	(0.32, 1.33)
Michele Byberg	48	White	0.60	(0.31, 0.73)

TABLE 1.2: Estimates of lifetime risk with confidence interval for four women.

Name	Median Age (in years)	Most Probable Race	Lifetime risk (upto 90 years)	Confidence Interval
Marilyn Muller	66	White	3.99	(0.72, 5.36)
Denise Gallego Moreno	52	Hispanic	4.03	(1.84, 7.22)
Sue Wedin Jones	63	White	4.44	(0.99, 5.85)
Michele Byberg	48	White	6.14	(3.93, 6.54)

of breast cancer (mother, sister, or daughter), number of past breast biopsies, number of breast biopsies showing atypical hyperplasia will have breast cancer in the next five years. The 95% confidence interval for the true gail score for Marilyn Muller is (0.27, 1.35) which implies that out of 1000 women with Marilyn Muller as their full name, approximately 950 of them will be having their absolute risk of breast cancer for five years between 0.27 and 1.35.

- From Table 1.2, we see that the estimated(median) lifetime risk for Marilyn Muller is 3.99 which says that 0.0399 % of the population of woman who are white, aged 66 years, with no medical history of breast cancer and with no information on age at first period, age at the time of the birth of her first child (or has not given birth at all), family history of breast cancer (mother, sister, or daughter), number of past breast biopsies, number of breast biopsies showing atypical hyperplasia will develop breast cancer in their lifetime(upto 90 years). The 95% confidence interval for the true gail score for Marilyn Muller is (0.72, 5.36) which implies that out of 1000 women with Marilyn Muller as their full name, approximately 950 of them will be having their lifetime risk of breast cancer between 0.72 and 5.36.

- From Figure 1.1, the absolute risk for five years increases with the age of an woman.

- From Figure 1.2, the lifetime absolute risk(risk upto 90 years) decreases as the age of an woman increases.

- There are no women in the study sample whose estimated gail score is greater than 1.66 implying that women who have the set of covariates same as that of predicted for the study sample are at a lower risk of having breast cancer.

1.5 Data Limitations

- Although the gender is determined from the Social Security Admnistration data based on full names, this method might not work for any name which is not recorded in the Social Security database.
- Similarly for determinig age from the name-age calculator, the method might not work for those women whose first names are not recorded on Social Security database.
- This method might not work for well-mixed population as there is a big assumption that surnames of women being same as that after their marriage which might occur if an woman marries a man with the same surname. This is not true for a general population.
- The presence of medical history of an woman is determined by searching some of the key words like "Breast Cancer Survivor", "Breast Cancer Free", "Diagnosed" and based on the assumption that these users commented about themselves. The results might vary based on the set of key words and might be wrong if someone commented about their daughter or other relatives.

1.6 Conclusion

It is very difficult to get information on age at menarchy, age at th time of first birth of child , presence of BRCA gene of a woman based on facebook profiles. It is also difficult to find some basic information such as gender, date of birth due to privacy issues. A post from the Breast Cancer Site in facebook is studied. The information such as full name, comments and picture url of the users are easily available and is taken as the raw data. Different characteristics such as age, gender, race, medical history of breast cancer are predicted/estimated based on the data. Rate of misclassification(comparing these algorithms and by manually) is negligible(shown in supplementary material). Absolute risk of breast cancer are estimated from the Gail model and their uncertainities are calculated based on empirical quantiles. It is observed that the absolute risk increases with age and the lifetime risk decreases with age. None of the women have their median absolute risks to be greater than 1.66 indicating they have low risk of having breast cancer. There are some limitations to this method which are discussed in the data limitations section. The whole procedure takes approximately 18 mins in R to compile and produce the results.

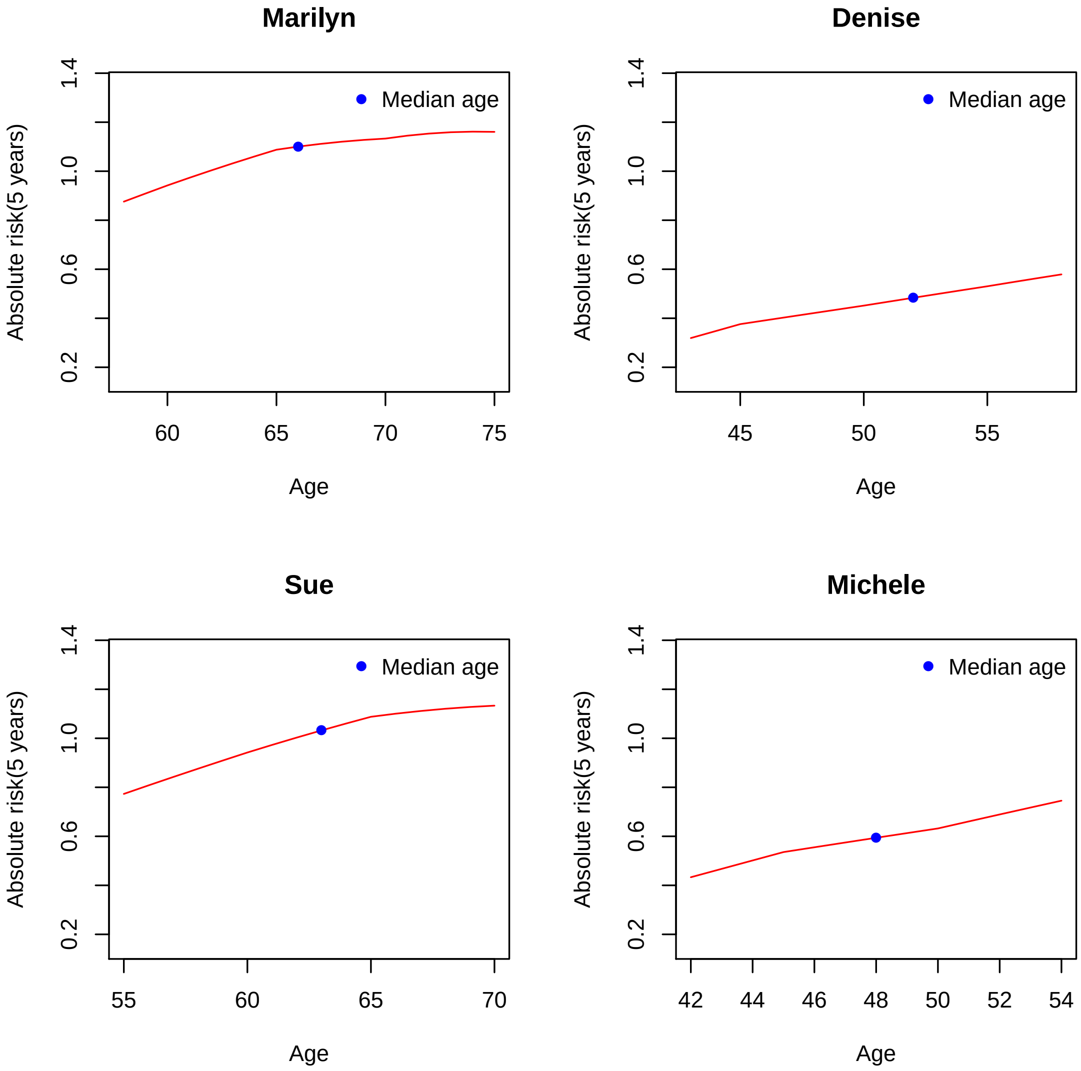

FIGURE 1.1: **Gail Score** Plot of absolute risk for five years versus age for four women from facebook profiles with their races fixed at their most probable races. Blue dot represents the gail score at the median age

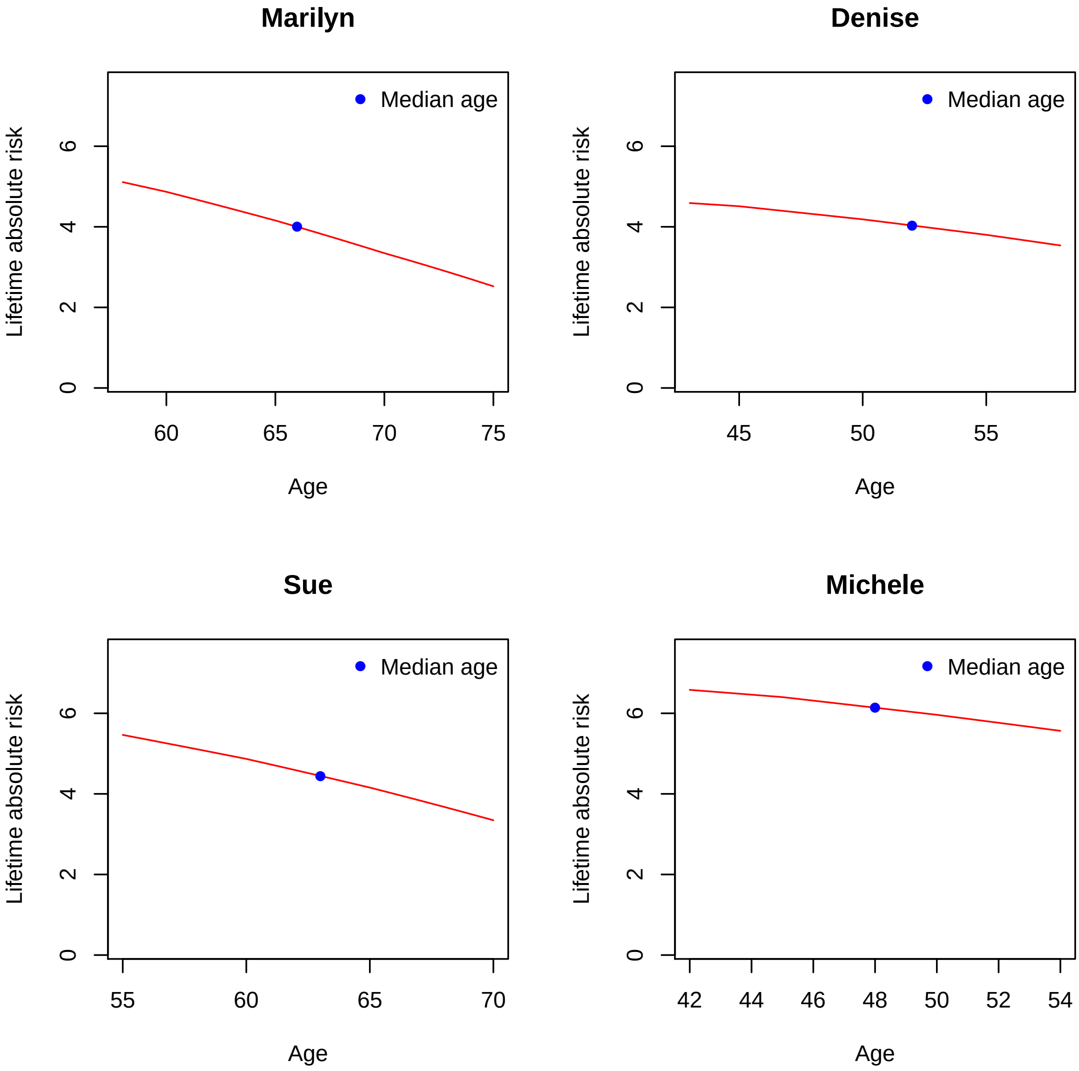

FIGURE 1.2: **Lifetime risk** Lifetime absolute risk(upto 90 years) versus age(in years) for four women with their races fixed at their most probable races. Blue dot represents the lifetime risk the median age.

Bibliography

[1] Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Shairer C, Mulvihill JJ: Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 81(24):1879-86, 1989.

[2] Gail MH, Costantino JP, Pee D, Bondy M, Newman L, Selvan M, Anderson GL, Malone KE, Marchbanks PA, McCaskill-Stevens W, Norman SA, Simon MS, Spirtas R, Ursin G, and Bernstein L. :Projecting Individualized Absolute Invasive Breast Cancer Risk in African American Women. J Natl Cancer Inst 99(23):1782-1792, 2007.

[3] Kosuke Imai, Kabir Khanna: Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records, Political Analysis (2016) 24:263–272.

[4] Yuqin Li, Lihua Chen, Xiaohai Wan, Alan Chiang: Implementation of Breast Cancer Risk Assessment Tool using SAS.