A facebook project: gail score model

Cynthia Kineza

2017-10-17

A while ago, I worked on a data science project that required me to collect data from Facebook. The goal was to estimate risk of breast cancer in women who had commented on various Facebook breast cancer support groups.

Background: Gail Score Model

In order to estimate the risk of breast cancer, a health care provider looks at how many risk factors a woman has, many of them increase risk at different degrees. For example, having a BRCA1 gene mutation increase breast cancer risk significantly compared to other risk factors.

The Breast Cancer Risk Assessment Tool (the Gail model) is used to estimate breast cancer risk. Even though the tool can estimate a person's risk, it cannot provide you the adequate information of whether or not you will have the disease. The tool determines a woman's risk of developing breast cancer within the next 5 years and within her lifetime (up to age 90). Below are the seven key risk factors to evaluate such risk:

Age

Age at first period

Age at the time of the birth of a first child (or has not given birth)

Family history of breast cancer (mother, sister or daughter)

Number of past breast biopsies

Number of breast biopsies showing atypical hyperplasia

Race/ethnicity

Algorithm

I built an algorithm that calculated Gail Score for women to estimate their lifetime risk of breast cancer based on the data collected from Facebook pages variables such as full name, message, likes count, comments count and shares count were extracted. In order to have access to the data, I registered to have developer privileges:

Refer to (https://developers.facebook.com/docs/graph-api) : For collecting data on anything that is available publicly.

Refer to: (https://developers.facebook.com/docs/graph-api/reference/v2.7/) : Choose any field from which you want to extract data such as "pages". In my case, I used few of them.

First, install package RFacebook, then follow these steps:

my_authorization <- fbOAuth(app_id = "enter your ID here" , app_secret = "secret code enter here")

save(my_authorization , file = "my_authorization")

load("my_authorization")

savepage <- getPage(page="Include here the name of the facebook page", token=my_authorization, n=5000, feed=TRUE)

Process

Few challenges presented themselves when I was building my algorithm. 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 impossible to obtain from a social media site/ or even through comments that I was going to parse. However, some other variables were much easier to estimate (age, race/ethnicity, gender) with the use of various R packages and the Census Bureau public data.

The names I had collected from facebook pages did not include birthdates (not surprising). Hence, I needed to find alternatives to estimate age of each individual who posted a comment.I used the census bureau information that provides the overall frequency of surnames as well as some of the basic demographic characteristics such as gender. The US Census Bureau tabulates a list of surnames occurring > 100 times in its database (with frequency): all 152,000 of them for numerous years.

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

Using a range of age (age of a woman given first to third quartile coming from the Census bureau database of firstnames) with a fixed race. Then, a range of potential absolute risks was derived

Using a range of race types (generated by package "wru" using surnames as indicators, then list of probabilities that the individual is "white", "black","hispanic" or from "other" race was inferred) with the median estimated age: a range of potential risks was derived

Using both the plausible age and race ranges: a range of potential risks was derived

All of those ranges of age and race are done by taking into account uncertainty in the absolute. Moreover, by using the variability in both age and race, we obtained variability in absolute risks with a projection in 5 years and a lifetime risk. In calculating the lifetime risk I fixed the projection age of BrCA to 90 years, while a regular projection in 5 years was obtained by adding 5 years to the actual estimated age.

This was a fun project, I hope you are inspired to use social media data in solving everyday problems. And if you have already done so, please share your experience :)

Main Packages

Gender: This package allows you to estimate the gender based on names from a chosen timeline

wru: Who are You? The method utilizes the Bayes'Rule to compute the posterior probability of each racial category for any given individual.

BCRA: Provides risk projections of invasive breast cancer based on Gail model according to National Cancer Institute's Breast Cancer Risk Assessment Tool algorithm for specified race/ethnic groups and age intervals.

Resources

1. The Breast Cancer Risk Assessment Tool (the Gail Model): (https://www.cancer.gov/bcrisktool/)
2. wru: (https://cran.r-project.org/web/packages/wru/)
3. Gender: (https://cran.r-project.org/web/packages/gender/)
4. Rfacebook: (https://cran.r-project.org/web/packages/Rfacebook/)
5. BCRA: (https://cran.r-project.org/web/packages/BCRA/)