

Facebook Project

Cynthia Kineza

October 16, 2017

A while ago, I worked on a data science project that required me to fetch data from Facebook pages. 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 and much of these factors increase risk. Out of all factors, having a BRCA1 gene mutation increase breast cancer risk significantly. The Breast Cancer Risk Assessment Tool (the Gail model) is often used by health care providers to estimate risk. Even though the tool can estimate a person's risk, it cannot provide you the information of whether or not you will get breast cancer. The tool determines a woman's risk of developing breast cancer within the next 5 years and within her lifetime (up to age 90). There are 7 key risk factors to evaluate breast cancer.

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, which provided me variables such as full name, message, likes count, comments counts and shares_count. In order to have access to the data from Facebook, I registered to have developer privileges:

Go to [link] (<https://developers.facebook.com/docs/graph-api>) if want to collect data on anything that is available publicly.

See [link] (<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 facebook pages

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 while some of 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), so I needed to find alternatives to estimate age of each individual. I used the census bureau information that provides the overall frequency of surnames as well as some of the basic demographic characteristics that are associated with surnames (gender). The US Census Bureau tabulates a list of surnames occurring > 100 times in its database (with frequency): all 152,000 of them.

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

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

All of those ranges in age and race are done by taking into account uncertainty in the absolute. 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. 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 already have done so, please share your experience :)

Main Packages

Gender Package: 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 Package] (<https://cran.r-project.org/web/packages/wru/>)
3. [Gender package] (<https://cran.r-project.org/web/packages/gender/>)
4. [Rfacebook package] (<https://cran.r-project.org/web/packages/Rfacebook/>)
5. [BCRA] (<https://cran.r-project.org/web/packages/BCRA/>)