**Using TCGA breast cancer MongoDB dataset combined with R shiny**

**Background knowledge:**

Breast cancer is a term used to describe a group of malignant tumors that develop from cells normally found in the breast. A malignant tumor is made up of abnormal cells that have the ability to invade surrounding tissues and spread to other parts of the body. Breast cancer is the second leading cause of cancer death in women. It occurs almost entirely in women. Overall, the average risk of a woman in the United States developing breast cancer in her life is about 13%. This means there is a 1 in 8 chance she will develop breast cancer. Last but not least, most breast cancers are found in women who are 50 years old or older.

**Goal:**

In this project, the goal is to study the highly expressed SNP in early stage breast cancer patients over 50 years old post-menopausal women.

**Database:**

For this project, we use MongoBD as the project dataset. MongoDB is a NoSQL database which can handle large volumes of data at high speed. Plus, it is simple for developers to learn and use, while still providing all the capabilities needed to meet the most complex requirements at any scale. MongoDB stores data in JSON-like documents, meaning fields can vary from document to document and data structure can be changed over time

Last but not least, MongoDB is free to use.

**Description:**

For this Final project, we are going to use TCGA breast cancer MongoBD dataset combine with R shiny. First, this project dataset is obtained from TCGA and it contains 50 breast cancer patients and 57 columns. Columns include: px (patient’s number), gender, race, vital status, days to last follow-up, age at diagnosis, pathologic T, pathologic N, pathologic M, pathologic stage, surgical procedure first, drug name 1 to 8, radiation type 1, anatomic treatment site 1, pathogenic Chr Pos 1 to 18, and Exon 1 RPKM 1 to 18. Then, In MongoBD, we create a database called Mongo\_clingen\_final and import the dataset which is called px\_final.json by using mongoimport from a new terminal. After importing the dataset, we are ready to start making queries.

**Queries:**

(1)

The first query is to see how many women in all races are older than 50 years old in my dataset. Due to the article mentions that most breast cancers are found in women who are 50 years old or older. In the query, it shows that there are 34 women are older than 50 years old [1].

> db.px\_final.count ( { "gender" : "FEMALE", "race" : {"$in" : ["WHITE", "BLACKORAFRICANAMERICAN", "ASIAN", "AMERICANINDIANORALASKANATIVE", "HISPANICORLATINO"]}, "age\_at\_diagnosis" : {"$gte" : 50} })

34

(2)

The second query is to see how many post-menopausal patients use Arimidex within the first query. Arimidex is used to treat post-menopausal women diagnosed with either early-stage or advanced-stage, hormone receptor-positive breast cancer [2]. The average age of menopause range from 45 and 55 [3], so we decide to select the median value 50. In the end, within the first query, there are 12 patients use Arimidex as their first drug.

> db.px\_final.count ( { "gender" : "FEMALE", "race" : {"$in" : ["WHITE", "BLACKORAFRICANAMERICAN", "ASIAN", "AMERICANINDIANORALASKANATIVE", "HISPANICORLATINO"]}, "age\_at\_diagnosis" : {"$gte" : 50}, "drug\_name\_1" : {"$in" : ["Arimidex"]}})

12

(3)

Then, continue with the second query, we would like to know how many people’s first surgical procedure are Lumpectomy. Lumpectomy is one of the common surgery to remove breast cancer. It is a surgery to remove the cancer as well as some surrounding normal tissue [4]. How much breast is removed depends on where and how big the tumor is. There are only 9 patients go through Lumpectomy after my second query.

> db.px\_final.count ( { "gender" : "FEMALE", "race" : {"$in" : ["WHITE", "BLACKORAFRICANAMERICAN", "ASIAN", "AMERICANINDIANORALASKANATIVE", "HISPANICORLATINO"]}, "age\_at\_diagnosis" : {"$gte" : 50}, "drug\_name\_1" : {"$in" : ["Arimidex"]}, "surgical\_procedure\_first" : "Lumpectomy"})

9

(4)

Within these 9 patients, I want to check how many patients are in the early stage of breast cancer. For example, StageIA and StageIB. When the tumor is small, invasive, and has not spread to the lymph nodes [5]. It is considered as Stage IA. When the cancer has spread to the lymph nodes and the cancer in the lymph node is larger than 0.2 mm but less than 2 mm in size. It is considered as Stage IB. As the result, there are 3 patients fit the query.

> db.px\_final.count ( { "gender" : "FEMALE", "race" : {"$in" : ["WHITE", "BLACKORAFRICANAMERICAN", "ASIAN", "AMERICANINDIANORALASKANATIVE", "HISPANICORLATINO"]}, "age\_at\_diagnosis" : {"$gte" : 50}, "drug\_name\_1" : {"$in" : ["Arimidex"]}, "surgical\_procedure\_first" : "Lumpectomy", "pathologic\_stage" : {"$in" : ["StageIA","StageIB"]}})

3

(5)

For the last query, we would like to know how many of these 3 patients have the specific Pathogenic\_Chr\_Pos\_1 which is “13\_32903685” or “13\_32890572” reported as highly expressed [6, 7] . It turns out that only one of them fit the query and it’s highly expressed Pathogenic\_Chr\_Pos\_1 is “13\_32890572”. The Pathogenic\_Chr\_Pos\_1 “13\_32890572” is located in Chormosome 13 position 32890572 in GRCh37. It’s a single nucleotide variant with a G to A mutation and it related with BRCA2. However, it is interpreted as Benign [8]. So, it is in line with the former queries about first drug usage, first surgery, and the early stage.

> db.px\_final.count ( { "gender" : "FEMALE", "race" : {"$in" : ["WHITE", "BLACKORAFRICANAMERICAN", "ASIAN", "AMERICANINDIANORALASKANATIVE", "HISPANICORLATINO"]}, "age\_at\_diagnosis" : {"$gte" : 50}, "drug\_name\_1" : {"$in" : ["Arimidex"]}, "surgical\_procedure\_first" : "Lumpectomy", "pathologic\_stage" : {"$in" : ["StageIA","StageIB"]}, "Pathogenic\_Chr\_Pos\_1" : {"$in" : ["13\_32903685", "13\_32890572"]}, "Exon\_1\_RPKM\_1" : {"$in" : ["13\_32903685", "13\_32890572"]}})

1

**Conclusion:**

From this study, we first select patients in all races who are older than 50 years old. Second, we make query of their first drug Arimidex. Third, we make query of their first surgical procedure which is Lumpectomy. Then, we narrow down our query with early stage. Last, we make query of their highly expressed pathogenic Chromosome Positions. From the result, we can conclude that the highly expressed SNP position 32890572 in Chromosome 13 is interpret as Benign. We can also speculate the result from their drug usage and surgical procedure.

**R shiny:**

After making queries, we use R shiny to display graphs. First, we use libraries of shiny, ggplot2, reconnect, and reshape2. Second, we keep patients who are greater than 50 years old. Last, we select three columns which are Pathogenic\_Chr\_Pos\_1, Drug\_name\_1, and Age\_at\_diagnosis. In user interface, we can change the title, text, and sidebar buttons with three different positions which are “13\_32890572”, “13\_32903685”, and “13\_32906729”. In the first output, we select position “13\_32890572”, it shows that most of patients who are older than 50 years old mainly use Arimidex as their first drug in the dataset and the second large portion of people use Tamoxifen. In the second output, we select position “13\_32903685”. Still, the majority of people use Arimidex as their first drug. Last, with the position “13\_32906729”, most people use Anastrozole as their first drug. Anastrozole is used to treat early hormone receptor-positive breast cancer and Arimidex is the brand name for anastrozole. So, it makes sense that people usually take Arimidex as their first drug because their stage might not be severe at this point and these chromosome position also interpret as Benign.

**Reference:**

1. Centers for Disease Control and Prevention. *CDC website*. 2021 [cited 2022 April 26]; Available from: <https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm#:~:text=Most%20breast%20cancers%20are%20found,factors%20have%20the%20same%20effect>.

2. Breastcancer.org. *Breastcancer.org website*. 2022 [cited 2022 April 26]; Available from: <https://www.breastcancer.org/treatment/hormonal-therapy/arimidex>.

3. University of Utah. *University of Utah website*. 2022 [cited 2022 April 26]; Available from: <https://healthcare.utah.edu/womenshealth/gynecology/menopause/postmenopause.php#:~:text=Most%20women%20reach%20this%20milestone%20somewhere%20between%20ages%2045%20and%2055.,-A%20New%20Pause>.

4. American Cancer Society, I. *merican Cancer Society, Inc. website*. 2022 [cited 2022 April 26]; Available from: <https://www.cancer.org/cancer/breast-cancer/treatment/surgery-for-breast-cancer.html#:~:text=This%20surgery%20is%20also%20called,several%20different%20types%20of%20mastectomies>.

5. American Society of Clinical Oncology. *ASCO website*. 2005-2022 [cited 2022 April 26]; Available from: <https://www.cancer.net/cancer-types/breast-cancer/stages>.

6. National Library of Medicine. *ClinVar*. 2022 [cited 2022 April 26]; Available from: <https://www.ncbi.nlm.nih.gov/clinvar/variation/126192/?new_evidence=true>.

7. National Library of Medicine. *ClinVar*. 2015 [cited 2022 April 26]; Available from: <https://www.ncbi.nlm.nih.gov/clinvar/variation/125965/?new_evidence=true>.

8. Capalbo, C., et al., *BRCA1 and BRCA2 genetic testing in Italian breast and/or ovarian cancer families: mutation spectrum and prevalence and analysis of mutation prediction models.* Ann Oncol, 2006. **17 Suppl 7**: p. vii34-40.