Case Study: Text Mining Breast Cancer Studies

Allan Kimaina, Julia Mullokandova, and Wei Wang
Brown University Providence, RI

Abstract

As of 2018 in the United States, there are more than 3.1 million women with history of breast cancer, and death rates higher than those for any other cancer. However, these death rates have been decreasing due to treatment advances, earlier detection through screening and increased awareness. This paper reviews biomedical research literature on breast cancer with an aim to discover the progress and overlap made in research related to diagnosis and treatment of this disease. The study required conducting text mining on 1000 articles in each category related to breast cancer over the last ten years. The results highlight how Natural Language Processing (NLP) can be used to derive information from free text reports, which converts to useful clinical information for discovering therapies of a particular disease. This project provides opportunities for further advances in oncology by using published data for analysis.

Introduction

Breast cancer is the most common type of cancer diagnosed among women and is the second leading cause of cancer death.[5] Approximately 252,710 new cases of invasive breast cancer and 63,410 cases of in situ breast carcinoma were expected to be diagnosed among US women in 2017, and 40,610 women are expected to die from this disease alone.[5] Overall, breast cancer mortality rates have decreased by 39% through 2015 and this is a result of improvement in treatment and detection screening by mammography.[5] Cancer is a result of damage (mutation) to a cell's DNA, so that the cell loses normal functionality and instead gains the ability to indefinitely multiply until normal tissue functions are impaired.[7] Most breast cancers begin in the cells that line the ducts, few start in the cells lining the lobules, and in cells of the other tissues in the breast.[3] Cancerous DNA mutations may occur from a complex mixture of inherited and external factors, where these mutations are usually located in cell division genes.[7] However, some other risk factors include age, mutation in BRCA1 or BRCA2 gene, radiation exposure, high breast density on a mammogram, family history of the disease, and so on.[3] The aims of this study include: 1) the progress and overlap made in research related to diagnosis and treatment of this disease 2) to what extent can computational methods convert text data into useful clinical information 3) which knowledge resources can manage text mining of cancer related information 4) how Natural Language Processing (NLP) can be used to structure information from free text reports.[13] With the amount of articles published it is not uncommon for researchers to encounter new insights they were unaware of.

Text mining is used to discover these knowledge patterns or hypotheses in helping to solve biomedical questions.[4] Particularly in medicine, data mining helps to analyze patient disease history in order to identify and understand their future visits to clinicians, assists in discovering the successful treatments amongst the different diseases of the patients, and allows to analyze patient past disease history to find out the chances of future problems.[12] Text mining cannot occur on its own instead it is part of a process for finding the necessary information to analyze the problem. It initializes with integration of the data from all the sources, selecting the required data based on specific criteria, cleaning the data for missing value, errors, or particularly in our case "stop words", which provide no real meaning to the text. Next, we transform the data into a form feasible for analysis, and then concentrate on data mining, which is used to find out various similarities in the text. The final part is to evaluate the pattern observed and apply it to new discovery knowledge, which will assist in providing better decisions in a related field (Fig. 1).[12] The two major steps are: identifying biomedical entities and concepts of interests from free text using NLP (Natural Language Processing) techniques; and then further analyses are performed to see whether these entities have any relationships.[4]

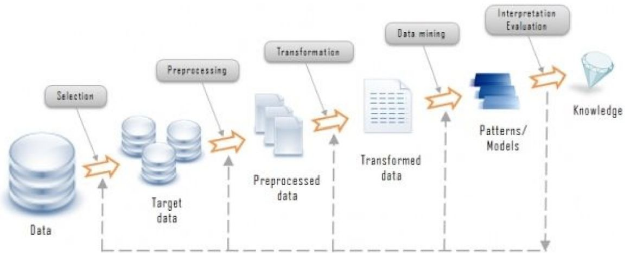

Figure 1. Overview of Text Mining Process

Methods

In this study, we utilized Julia programming language, PubMed Central, and text mining techniques to conduct the necessary research. Julia is a high-level, high-performance dynamic programming language for numerical computing. The language provides a sophisticated compiler, distributed parallel execution, numerical accuracy, and an extensive mathematical function library.[1] In using Julia, our program had a good performance, approaching that of statistically-compiled languages like C, the utilization of built-in package manager, and user defined types such as the functions we developed are as fast and compact as the built-ins; and those being just a few of the highly sophisticated features.[1] PubMed Central (PMC) is a free repository that archives publicly accessible full-text scholarly articles published within the biomedical and life sciences journal literature. In using PMC, we were able to access the entire article for analysis. The majority of our work was presented using Jupyter to provide a powerful browser-based graphical notebook interface to Julia.[1]

Eligibility Criteria

For this study, we focused on articles pertaining to the various areas of research in breast cancer. We defined eligible articles as those containing mesh term equivalence for breast cancer, which was \“breast neoplasms\”[mh] that specifically focuses on human only studies. We selected for that specific term in combination with the following mesh term qualifiers: diagnosis (DI), drug therapy (DT), and therapy (TH). The query terms were determined from the MeSH browser from: https://meshb.nlm.nih.gov/search. There was an exclusion criteria set to eliminate articles older than ten years (2008-2018). The sample size was set to 1000 articles for each of the categories (Diagnosis, Treatment) specifically 100 articles per year were selected.

Searches and Study Selection

To initialize the search, the package Requests was called for E-Utilities to be used in Julia to pull articles from PMC related to breast cancer. The defined mesh terms were utilized to conduct the search. The DataFrames package assisted with using the function searchBreastCancerArticles, which outputs a dataframe of articles stating the pmcid, pmid, the title, and the date created. For each of the three groups, a text file output was created with information pertaining to each individual article selected based on the search in a MEDLINE format.

Data Extraction

In order for the data to be extracted and processed for analysis, the use of the Taro package was required. Taro works with document files in Julia, particularly the feature we focused on is extracting raw text from PDF files. A function defined as fetchFullArticleFromPmc assisted in extracting text from the full article pdf. In order to avoid any issues with obtaining full articles, we used PMC. The article was pulled based on pmcid and stored into a temporary directory.

Data Cleaning

Prior to conducting textual analysis, it was necessary to utilize the TextAnalysis package for cleaning the data. First, we converted the raw text to a string document. Next, we removed the following from the texts: cases, punctuations, numbers, prepositions, pronouns, stop words, non-letters, and custom stop words (a list of words, which appeared frequently in the text and provided no meaning to our analysis). We performed a before and after data cleaning analysis to depict the removal of noise in the data (Fig. 2). Prior to the cleaning, the data contained numbers, articles, prepositions, article description terms, etc.

Data Analysis

In analyzing the articles, we created a corpus or a collection of documents to ease the process. This was accomplished from taking the file document, converting it to a string document, tokenizing it, and finally producing an NGram document. Next, we standardized the corpus to verify that all of the articles are of a single type followed by preprocessing the corpus. The package Query filters, sorts, joins, and wrangles data from any iterable data source. After these steps are complete, Corpus statistics were conducted using lexicon, which consists of all the terms that occur in any document and will output the frequency across all the articles. With this information, other packages such as Gadfly, Plotly, and Vega will be incorporated to depict the output of the textual analysis graphically. Plotly specifically assists in producing interactive and publishable graphs such as histograms, scatterplots, bar graphs, time series, 3D graphs, etc. Alongside Gadfly and Vega all three of these packages allow to efficiently plot the data with distinguished visualization.

Results

Using the specified criteria, we identified the most frequent words, word pairs, and three word phrases utilized in the breast cancer biomedical literature pertaining to diagnosis. According to Figure 3, the lexicon frequency analysis for an individual word depicted that cancer, breast, cells, patients, study, etc. were the most often used words. The output of frequent individual words had a sample size of 60 terms. Next, the word pair combinations were studied, which had a decreased output to 28 terms. The most significant ones appeared to be breast cancer, cancer cells, cancer patients, mda mb, etc (Fig. 4). Finally, we computed the frequency for the output of three word phrases, which yielded also 28 phrase terms. Figure 5 displays the most prevalent phrases to be breast cancer patients, breast cancer cells, breast cancer res, breast cancer risk, negative breast cancer, cancer res treat, etc.

A trend analysis of the diagnostic techniques was conducted. Initially, we had 13 various techniques, but we narrowed down to 5. The techniques selected for elimination were a result of them being tools for prognosis of the disease or for a follow up analysis after diagnosis. Figure 6 shows that the top five diagnosis procedures for breast cancer include MRI, tomography, mammography, ultrasound, and biopsy.

In 2008, biopsy was the most popular diagnostic technique and in a year mammography had soared to be the most prevalent. However, over the ten year period mammography began to decline after the 2009 spike. MRI, tomography, and ultrasound procedures increased overall except for a sudden dip in the plot in 2015 where they all decreased significantly. Finally, biopsy remained in a pretty consistent frequency range except for a significant drop in 2013, and a noticeable increase the following year in 2014. As of 2018, biopsy appears to be the most prevalent procedure for diagnosis breast cancer followed by mammography, ultrasound, MRI, and tomography.

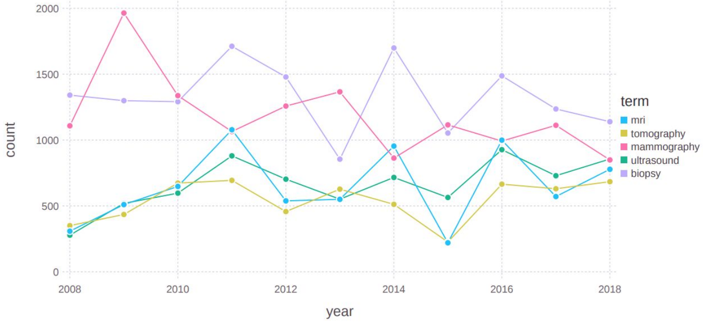

Figure 6. Trend analysis of the top 5 diagnostic techniques for breast cancer

We generated the different types of diagnosed breast cancer. These were composed into a trend plot to analyze the changes over time (Fig. 7). Ductal carcinoma in situ (DCIS) had a tendency to decrease since 2008, but was the most common diagnosis made in 2018. It had a significant increase in 2017 prior to decreasing. Next, we identified lobular carcinoma in situ (LCIS), which remained relatively constant with the diagnosis counts throughout the ten years. Invasive ductal carcinoma (IDC) increased from 2010 with the highest peak in 2014, while relatively decreasing after that. Invasive Lobular carcinoma (ILC) increases relatively with time with a sudden decrease for today's diagnosis count. Paget is a disease of the nipple, which remained relatively low and constant within the time interval, but a major spike was observed in 2015. Metaplastic diagnosis type remained relatively low with a decrease to zero by 2018. Inflammatory breast cancer (IBC) remained at a constant zero count for diagnosis type except in 2012, and with a sudden increase after 2016. Finally, mucinous carcinoma remained consistently low with a major diagnosis increase in 2012, and a significant decrease in 2014.

Additionally, we noticed patterns in text related to bodily chemicals and hormones, which fuel breast cancer growth. The most popular were estrogen and progesterone, which are hormones associated with developing and maintaining female characteristics. HER2 was the subsequent most common protein. Figure 8 depicts the trends in the research of these bodily chemicals when a subject is diagnosed with breast cancer. Progesterone and estrogen have been increasing significantly since 2008. Neu was the next most prevalent as it remained steady in the counts discovered in the research papers until 2016 with a major drop and then increasing from then on. Finally, sentinel lymph nodes remained relatively consistent with a count of around 500-800 over the ten year period.

We were also interested in observing the trends in the types of breast cancer tumors (Fig. 9). There are three different kinds: benign, malignant, and metastatic. With benign, doctors will usually just leave the tumor alone. Malignant is an aggressive form, which invades and damages surrounding tissue. Metastatic is when a cancer cell of a malignant tissue spreads to other parts of the body, usually through a lymph system and it forms a secondary tumor. Metastatic is the most common in the trend plot with an increase after 2010. Followed by malignant, which remains relatively steady after the increase from the 2009 decrease. As presumed, benign was the least common in the studies, and decreasing from 2011 to remain at a steady state.

Finally, we analyzed the trends in phenotypes of breast cancer cell lines (Fig. 10). Although we had an output of 10 common cell lines, the focus remained on the top 4, which include MDA, MCF, SUM, and EM. The remaining, HCC, UACC, EFM, KPL, CAL, and NME, appeared at or less than 500 counts in the studies conducted. With time, EM increased the most with a spike in 2016 to over a count of 1500, while MCF and MDA decreased then increased for some time, but overall decreased to a somewhat stable state. SUM remained consistently below a count of 1000 for the ten year interval.

Focusing on the treatment, we developed several trend plots after conducting textual analysis on the breast cancer research articles over a ten year time frame. Initially, we started off with a treatment trend analysis, identifying six of the most prevalent treatments mentioned in the studies. This plot demonstrates that surgery and chemotherapy are the most popular forms of treatment (Fig. 11). The remaining method of treatments: hormonal, targeted, systemic, and radiation appear relatively constant throughout the ten years. The unique changes in the plot relate to radiation increasing after 2016, while targeted and chemotherapy have spikes at 2011 and 2017, respectively.

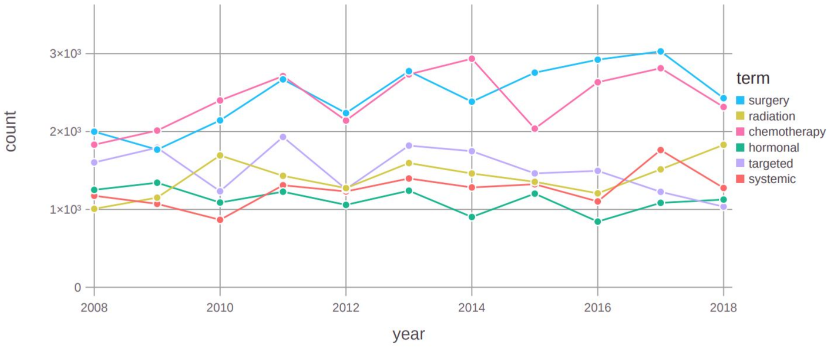

Figure 11. Trend analysis in the treatment methods for breast cancer

Within the surgery category, there are two types possible: lumpectomy and mastectomy (Fig. 12). Mastectomy is the most prominent type of surgery with comparatively increasing trends, while lumpectomy is much less prominent with an approximate 700 count or less. Delving deeper into this category, we learn the different techniques within mastectomy. Mastectomy contains a lot more possible procedural methods with the most significant being implants and latissimus dorsi flap (Fig. 13). The implants were at a maximum level of all techniques in 2011 and remained relatively high in comparison to the other procedures. Latissimus dorsi flap reaches its maximum point in 2016. The remaining techniques: Transverse Rectus Abdominis Muscle (TRAM) flap, Deep Inferior Epigastric Artery Perforator (DIEP) flap, and Gluteal free flap remained at a count below 100 with some even remaining at the zero level.

Under the treatment types, we selected for radiation therapy schedules to understand whether it is more likely that studies would specify whether the radiation administered was before the surgery to shrink the tumor (Neoadjuvant) or after the surgery (Adjuvant) (Fig. 14). It was determined that both techniques were increasing with time since 2008, however, adjuvant seems to be the more preferred method. There are also three types of radiation that were observed in our textual analysis, which include irradiation (PBI), Intensity-Modulated Radiation Therapy (IMRT), and Proton therapy (Fig. 15). PBI has been increasing since 2008 except for the decrease after 2013, but it soared up in usage from 2014. IMRT remains comparatively stable except for an increase peaked at 2013. For proton radiation, the largest count for usage appeared in 2011, but besides that it remained below a 300 usage count.

Using trends in chemotherapy, we selected for all the drugs mentioned in the articles, which outputs a list of 15 drugs, but we narrowed our interpretation to the top 5- paclitaxel, docetaxel, cyclophosphamide, doxorubicin, and fluorouracil. Paclitaxel is the only drug to reach a count of over 1000, which occurred in 2014. The other four followed similar trends with alternating increasing and decreasing, and only fluorouracil had major dips in the trend, particularly in 2017. A patient may also receive combinations of these drugs (Fig. 17), and the highest trend was over 2000 counts for AC (doxorubicin and cyclophosphamide), EC (epirubicin, cyclophosphamide) proceeded that with a little over 1000 counts, TC (docetaxel and cyclophosphamide) was closer to 1000 counts, and the rest remained pretty low and consistent with gradual decreases over time for CAF (cyclophosphamide, doxorubicin, and 5-FU), CEF (cyclophosphamide, epirubicin, and 5-FU), CMF (cyclophosphamide, methotrexate, and 5-FU), and TAC (docetaxel, doxorubicin, and cyclophosphamide).

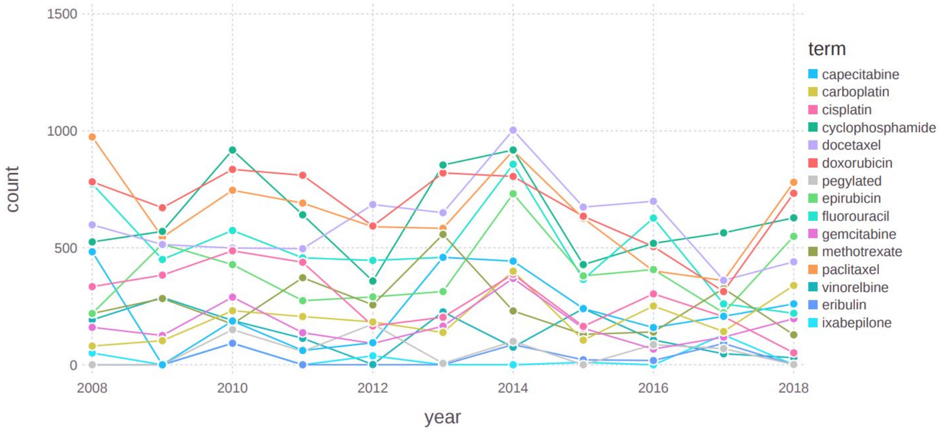

Figure 16. Trend analysis of chemotherapy drugs for breast cancer

Hormonal therapy trends takes into consideration the Tamoxifen drug and Aromatase inhibitors (AIs) (Fig. 18). Both of these procedures are decreasing with time, but Tamoxifen still remains at a higher usage count. Finally, the targeted therapy as a part of personalized medicine is relevant in displaying Trastuzumab as the most common drug used, while the remaining pertuzumab, ado, and neratinib were consistently low with counts computed from their mentionings in the studies (Fig. 19).

Discussion

In this project, we explored text mining of breast cancer research literature to assist us in understanding the trends in research. The results highlight the common diagnostic and treatment methods as well as, a deeper insight into the clinical and biological aspects of the studies conducted in order to help treat this disease. With text mining, we are able to interpret terms either individually or in a set. We noticed that for diagnosis, individual words provided less meaning as compared to those in pairs or a phrase. For example, some of the individual words were “breast”, “cancer”, “cell”, “women”, etc. Alone these words could be considered for any topic and does not have to pertain to breast cancer, which makes it more difficult to interpret and draw conclusions. However, when looking at word pairs or even a set of three we have a much better idea of what we are studying. An example includes: “breast cancer”, “mda mb”, “lymph node”, “estrogen receptor”, etc. and for three word phrases the output provided: “growth factor receptors”, “metastatic breast cancer”, “mda mb cells”, etc. As observed in the word pair combinations we had “mda mb”, which is not as clear as when the word cells is placed after it. With this we conclude that it is a type of cell line pertaining to breast cancer.

In interpreting the trends of diagnostic tools, we observed that biopsy was the most prevalent technique. A breast biopsy is a test that removes tissue or sometimes fluid from the suspicious area. The removed cells are examined under a microscope and further tested to check for the presence of breast cancer. A biopsy is the only diagnostic procedure that can definitely determine if the suspicious area is cancerous.[10] Mammography is a common screening tool, but it has slowly been on a decline for use with diagnosis. This is related to taking longer to perform than screening mammography and the total dose of radiation is higher because more x-ray images are needed to obtain views of the breast from several angles.[11] When assessing these studies we noticed that breast physical examinations and breast self-examinations were under researched among the diagnosis methods. This seems rather odd because it is a cost-effective alternative to all the other techniques.

The breast cancer types held their consistencies with detailed information regarding their histories. IDC is the most common type of breast cancer, and the trend in diagnosis increased from 2010 with the highest peak in 2014, while relatively decreasing after that, but still remaining one of the most prevalent.[10] ILC is the next most commonly diagnosed type of breast cancer, and increases relatively with time.[10] DCIS had the highest count of all other types of breast cancer especially in 2009, seeing as how most cancers initialize with the cells lining the ducts.[10] Finally, IBC and metaplastic are rare types of breast cancers and are currently under researched.[10] They are found to be less than 1% of all invasive breast cancer, but an aggressive form of locally advanced breast cancer.[10] With this idea, the analysis supported that as their appearance count in the research articles appeared to be relatively low over the ten year interval.

The final part of diagnosis was the phenotypes of the breast cell lines. The MDA-MB-231 cell line is an epithelial, human breast cancer cell line that was established from a pleural effusion of a 51-year-old caucasian female with a metastatic mammary adenocarcinoma1, and is one of the most commonly used breast cancer cell lines in medical research laboratories.[9] Hence, this term not only appears in the trend plot as the most prevalent, but it also does in the lexicon frequency plot. The MCF7 line retains several characteristics of differentiated mammary epithelium including ability to process estradiol via cytoplasmic estrogen receptors and the capability of forming domes. The cells also express the WNT7B oncogene.[8] Another common cell line is SUM, such as SUM102PT developed from a patient with minimally invasive, ER negative, PR negative and HER2 positive human breast carcinoma.[15] The cell line is representative of a class of human breast cancers characterized by a high level of EGFR expression in the absence of gene amplification.[15]

Treatment of such a complex disease creates an astounding amount of information available for analysis. With that, we were able to interpret the various treatment options for breast cancer patients, which include: surgery, radiation therapy, chemotherapy, hormonal therapy, and targeted therapy to target the cancer specific genes, proteins, or tissue environment that contributes to cancer growth and survival.[11] Surgery and chemotherapy appeared to be the most popular, since these ideas go hand in hand. The reasoning is drugs are administered either before or after surgery to make sure the tumor shrinks, or all the cancer cells have been eliminated, respectively.[11] In another trend plot we studied radiotherapy adjuvant versus neoadjuvant, which means either after or before surgery, respectively. This plot displays the removal of the tumor prior to administering radiation in order to guarantee it does not spread to other body tissues. PBI is the most common radiation form seems due to being administered to the tumor site directly, rather than the breast, which can kill healthy breast cells.

Lumpectomy is a type of surgical technique removing the tumor and a small, cancer-free margin of healthy tissue around the tumor, while most of the breast remains.[2] Mastectomy is the surgical removal of the entire breast.[2] Mastectomy had a higher turnout rate in the research studies, perhaps due to the guarantee of the complete removal of the cancerous tissue, but cosmetically is not the best option. With mastectomy, a breast implant using saline-filled or silicone gel-filled forms to reshape the breast is a common outcome.[2] The other techniques of conducting TRAM (uses muscle and tissue from the lower stomach wall), DIEP (takes tissue from the abdomen and the surgeon attaches the blood vessels to the chest wall) or even gluteal (uses tissue and muscle from the buttocks to create the breast, and the surgeon also attaches the blood vessels) is less common.[2] This results in an extended surgery and higher risk of complications so women prefer implants.

Finally, since chemotherapy is becoming more popular, we determine the types of drugs available for treating breast cancer. Paclitaxel, Docetaxel, and fluorouracil are one of the top most popular drugs administered to patients.[2] They function as antimicrotubule agents, inhibiting the microtubule structures within the cell.[2] Microtubules are part of the cell's apparatus for dividing and replicating itself.[2] Inhibition of these structures ultimately results in cell death.[2] The remaining two drugs in the top five category are useful for performing better in combination. Research has shown that combinations of certain drugs are sometimes more effective than single drugs for adjuvant treatment.[2] Doxorubicin and Cyclophosphamide function as an antitumor drug and alkylating agent, respectively.[2] By attaching alkyl groups to DNA it interferes with the cell's DNA and inhibits cancer cell growth.[2] With hormonal therapy, blocking the hormones can help prevent a cancer recurrence when used either by itself or after adjuvant or neoadjuvant chemotherapy.[11] Tamoxifen is a drug that blocks estrogen from binding to breast cancer cells.[11] It is effective for lowering the risk of recurrence in the breast that had cancer, the risk of developing cancer in the other breast, and the risk of distant recurrence.[11] Aromatase inhibitors (AIs) decrease the amount of estrogen made in tissues other than the ovaries in postmenopausal women by blocking the aromatase enzyme.[11] Tamoxifen appeared to have the highest count in the research papers probably as a result of focusing on blocking estrogen, which in the diagnosis section was one of the main hormones assessed for causing breast cancer.[11] Lastly, as a form of personalized medicine, Trastuzumab was the most used drug. It works by targeting the HER2/neu receptor on cancer cells.[11] HER2/neu is a growth hormone necessary for helping to manage how a breast cancer grows, divides, and repairs itself.[11] Some cancerous breast tissue have too much HER2 (HER2/neu overexpression), triggering the cells to divide and multiply very rapidly.[11] Trastuzumab attaches to the HER2 receptors to prevent cells from multiplying, preventing further cancer growth and slowing cancer progression.[11]

Limitations

In conducting this project, there were a few challenges that we encountered, which might have limited us in obtaining the best possible results. First, we only utilized PubMed Central as the database for retrieving the full text articles. Ideally, it would have been best to search all of PubMed, or other databases such as Public Library of Science (PLOS), Sciencedirect, etc. This could have provided us with more of a variety of documents related to breast cancer, creating a larger and more representative sample size, which would generate statistically better results. Although we also might of had to adjust for redundancy in article retrieval.

Second, completing the data cleaning process was a long and difficult procedure. With Julia language, the TextAnalysis package made for cleaning text was not adequate enough for accomplishing the necessary task completely. The first issue stemmed from the package failure to recognize some of the UTF-8 encodings. The second issue arose when creating the custom stop words. In generating this list ourselves, it can potentially be prone to human error where we could miss words that should be on the list such as author names or terms pertaining to the references in each paper, or placing words on the list that should not be there such as abbreviations representing potential genes, etc. The concern is how to standardize the text to keep it from generating inconsistencies and missing data that will hamper our analysis.

The last limitation we noticed is in querying the dataframe. Some of the lexicon may be confounded in context. For example, the term “surgery” in the articles can imply surgery for the treatment of breast cancer, but can also relate to another surgical procedure, which the subject underwent. This creates a minor challenge in analyzing the trend of the study. Hence, it would be best for the future to utilize the power of machine learning algorithms to assist with removing bias.

Conclusion

In conclusion, we determined the most prominently researched diagnosis and treatment options that have been published in breast cancer studies. We observed changes in trends of research over the ten year period utilizing NLP as a main tool. The transformation of the textual data is provided as useful clinical information for advancing further studies being conducted on complex diseases such as cancer, specifically with reporting under researched topics in the oncology field. We discovered valuable insight in conducting text mining, however, this method cannot occur on its own, but rather is part of a process for finding the necessary information to analyze the problem.

References

1. Bezanson J, Karpinski S, Shah V, Edelman A. A Summary of Features [Internet]. The Julia Language. [cited 2018Apr18]. Available from: https://julialang.org/
2. Breast Cancer Information and Awareness [Internet]. Breastcancer.org. Breastcancer.org; [cited 2018May9]. Available from: http://www.breastcancer.org/
3. Chaurasia V, Pal S. Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability [Internet]. SSRN. SSRN; 2017 [cited 2018Mar16]. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2994925

4. Chen H, Fuller SS, Friedman C, Hersh W. KNOWLEDGE MANAGEMENT, DATA MINING, AND TEXT MINING IN MEDICAL INFORMATICS . In: Medical Informatics Knowledge Management and Data Mining in Biomedicine. 1st ed. New York, NY: Springer US; 2005. p. 3–33.
5. DeSantis CE, Ma J, Goding Sauer A, Newman LA, Jemal A. Breast Cancer Statistics, 2017, Racial Disparity in Mortality by State [Internet]. Wiley Online Library. CA Cancer Journal; 2017 [cited 2018Mar16]. Available from: http://onlinelibrary.wiley.com/doi/10.3322/caac.21412/epdf
6. Gupta G, Bhathal GS. Introduction to Data Mining. In: SENTIMENT ANALYSIS OF ENGLISH TWEETS USING DATA MINING: Data Mining, Sentiment Analysis. BookRix; 2018.
7. Jurca G, Addam O, Aksac A, Gao S, Özyer T, Demetrick D, et al. Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends [Internet]. BMC Research Notes. BioMed Central; 2016 [cited 2018Mar16]. Available from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4845430/
8. MCF7 (ATCC® HTB-22™) [Internet]. MCF7 ATCC® HTB-22™. ATCC; [cited 2018May8]. Available from: https://www.atcc.org/Products/All/HTB-22.aspx#characteristics
9. MDA-MB-231 (ECACC catalogue no. 92020424) [Internet]. European Collection of Authenticated Cell Cultures. Public Health England; [cited 2018May7]. Available from: https://www.phe-culturecollections.org.uk/media/133182/mda-mb-231-cell-line-profile.pdf
10. Nbcf. Information, Awareness & Donations :: The National Breast Cancer Foundation [Internet]. www.nationalbreastcancer.org. 2016 [cited 2018May8]. Available from: http://www.nationalbreastcancer.org/
11. NCI. Comprehensive Cancer Information [Internet]. National Cancer Institute. NIH; [cited 2018May8]. Available from: https://www.cancer.gov/
12. Sarkar IN. Methods in biomedical informatics: a pragmatic approach. 1st ed. Cambridge, MA: Elsevier/AP, Academic Press is an imprint of Elsevier; 2013.
13. Spasic I, Livsey J, Keane JA, Nenadic G. Text mining of cancer-related information: Review of current status and future directions [Internet]. Egyptian Journal of Medical Human Genetics. Elsevier; 2014 [cited 2018May4]. Available from: https://www.sciencedirect.com/science/article/pii/S1386505614001105
14. Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, et al. Biomedical text mining and its applications in cancer research [Internet]. Journal of Biomedical Informatics. Academic Press; 2012 [cited 2018Mar16]. Available from: https://www.sciencedirect.com/science/article/pii/S153204641200171
15. May 11, 2018 | Hangzhou, China, May 16, 2018 | Orlando, Florida. SUM Breast Cancer Cell Lines [Internet]. BioIVT. [cited 2018May8]. Available from: https://www.bioivt.com/sum-breast-cancer-cell-lines/

Contributions

Allan Kimaina: Conducted text mining and analysis on the treatment of breast cancer. Wrote up abstract, results, discussion for his part. Julia Mullokandova: Searched, fetched and conducted data cleaning on full text articles. Wrote up methods and assisted with results and discussion. Wei Wang: Conducted text mining and analysis on diagnosis of breast cancer. Wrote up the conclusion, results and discussion for his part.

Appendix

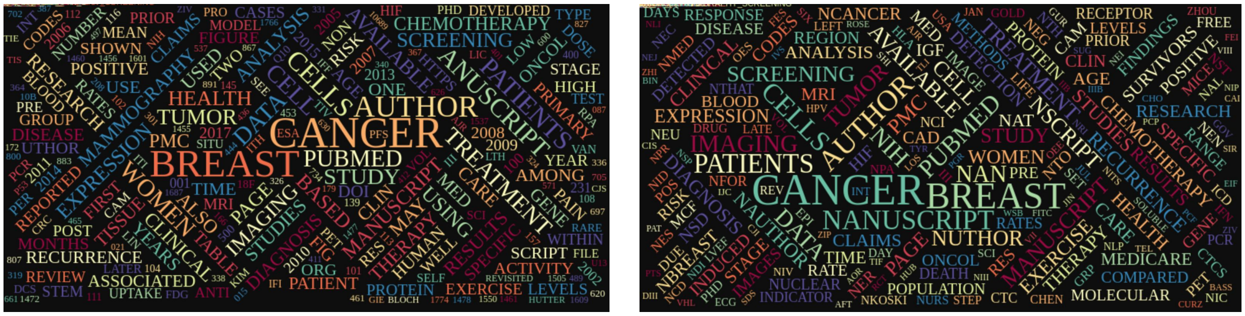

Figure 2. Word Cloud Before (Left) and After (Right) Data Cleaning

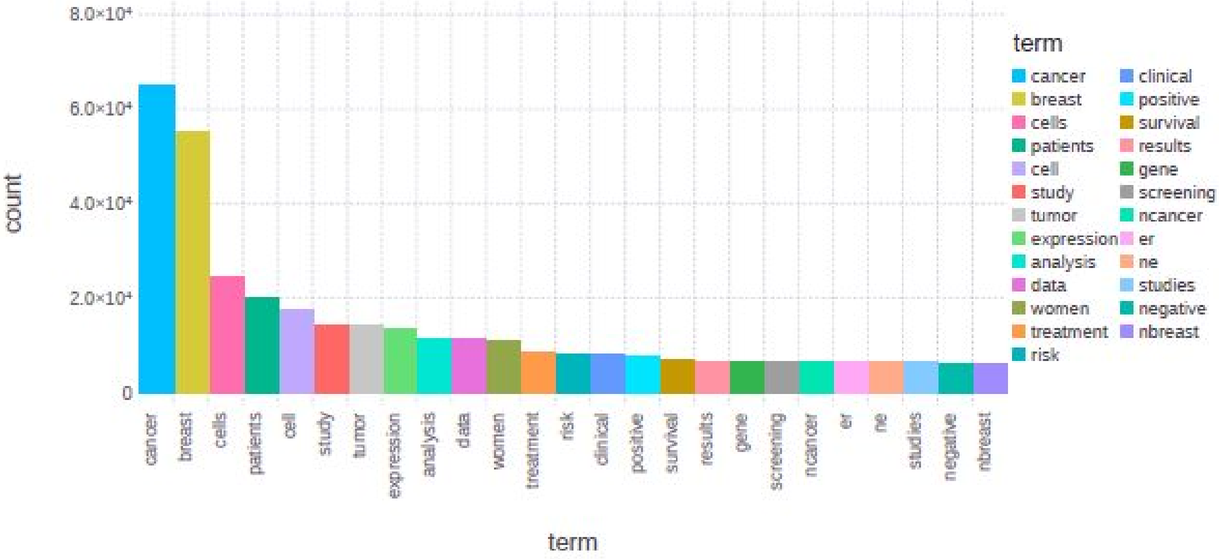

Figure 3. Lexicon frequency analysis for each individual word

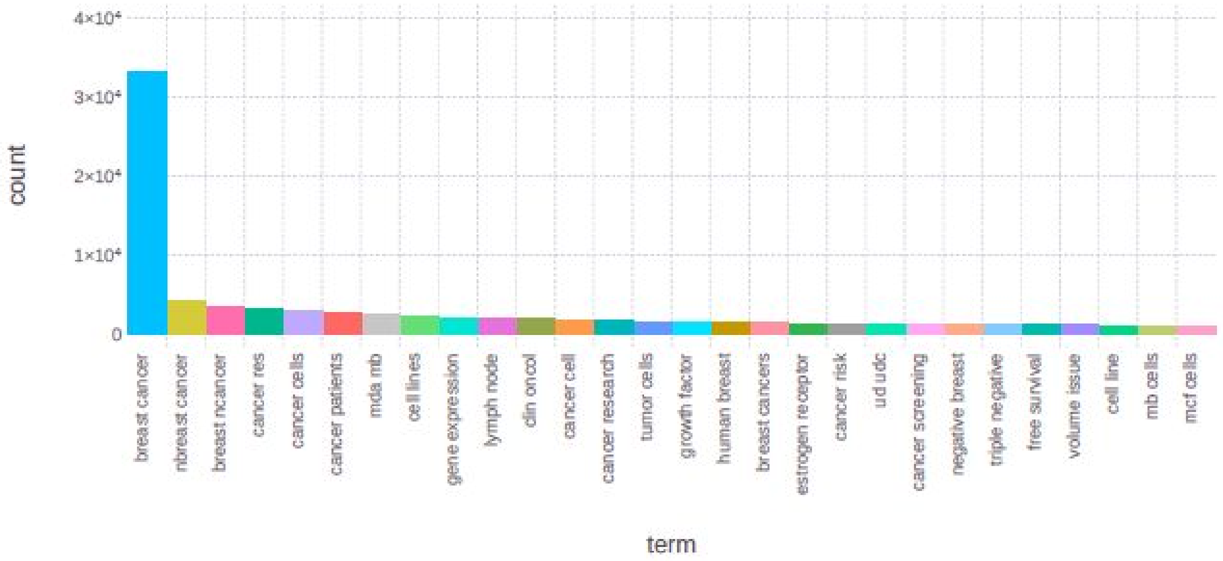

Figure 4. Lexicon frequency analysis for word pair combinations

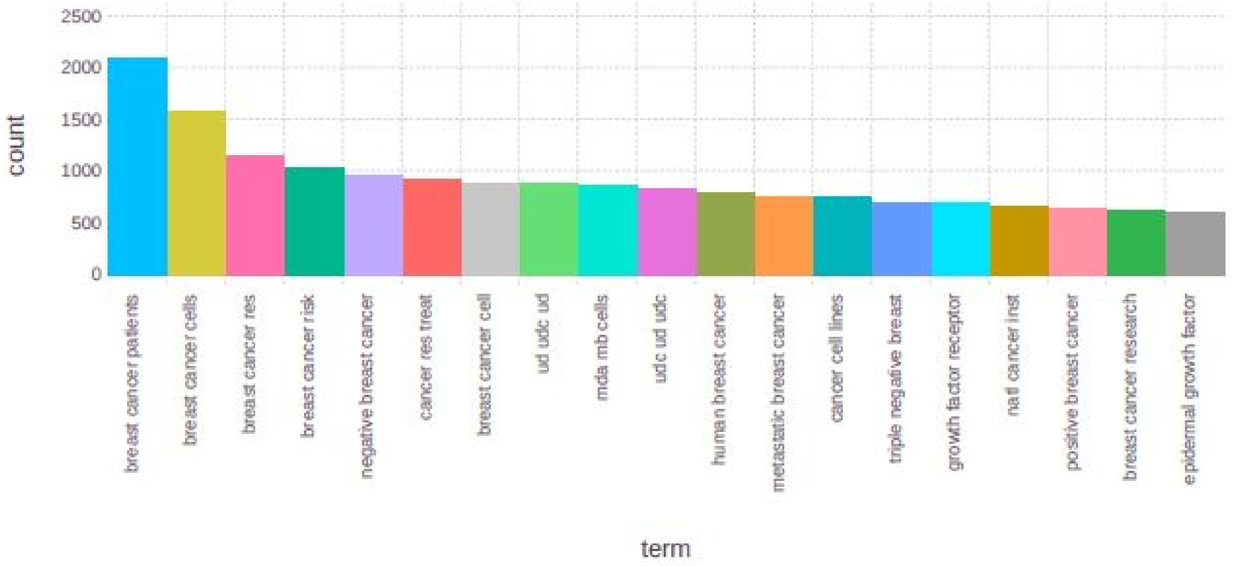

Figure 5. Lexicon frequency analysis for three word phrases

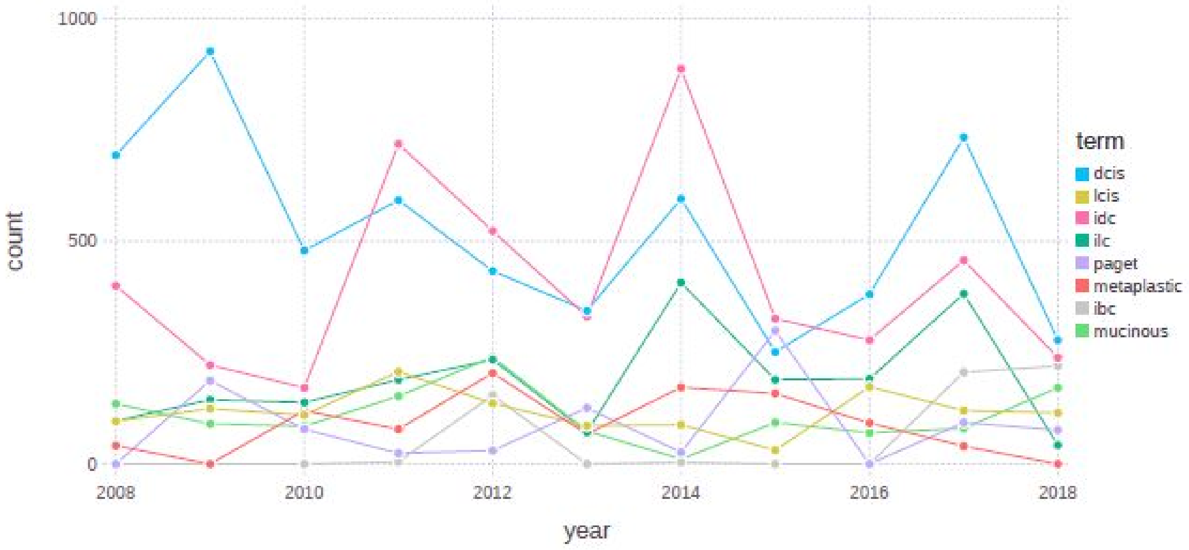

Figure 7. Trend analysis in types of breast cancer

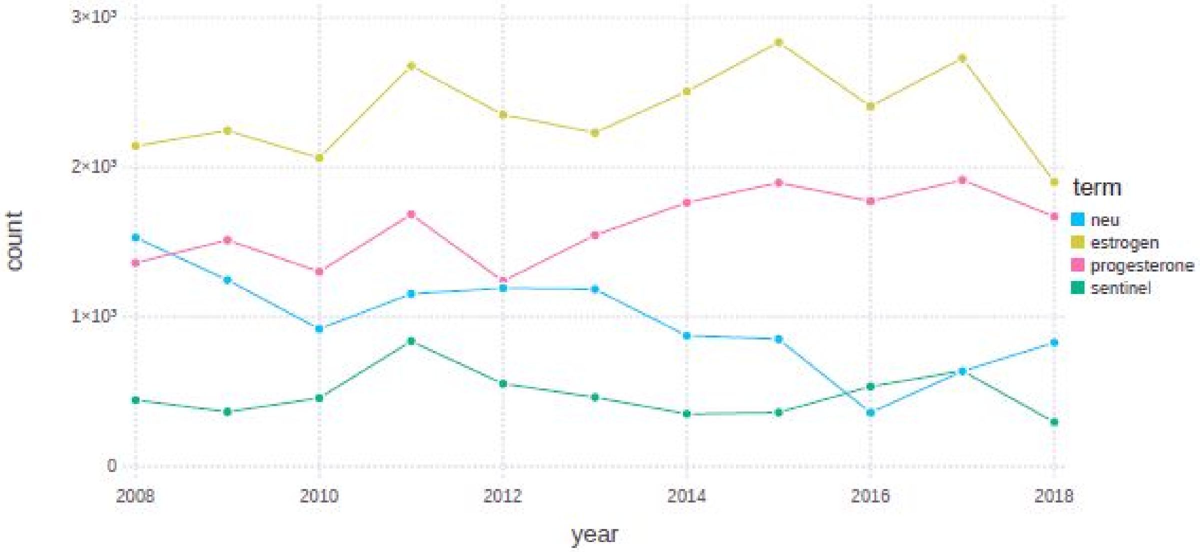

Figure 8. Trend analysis in body chemicals that fuel cancer growth

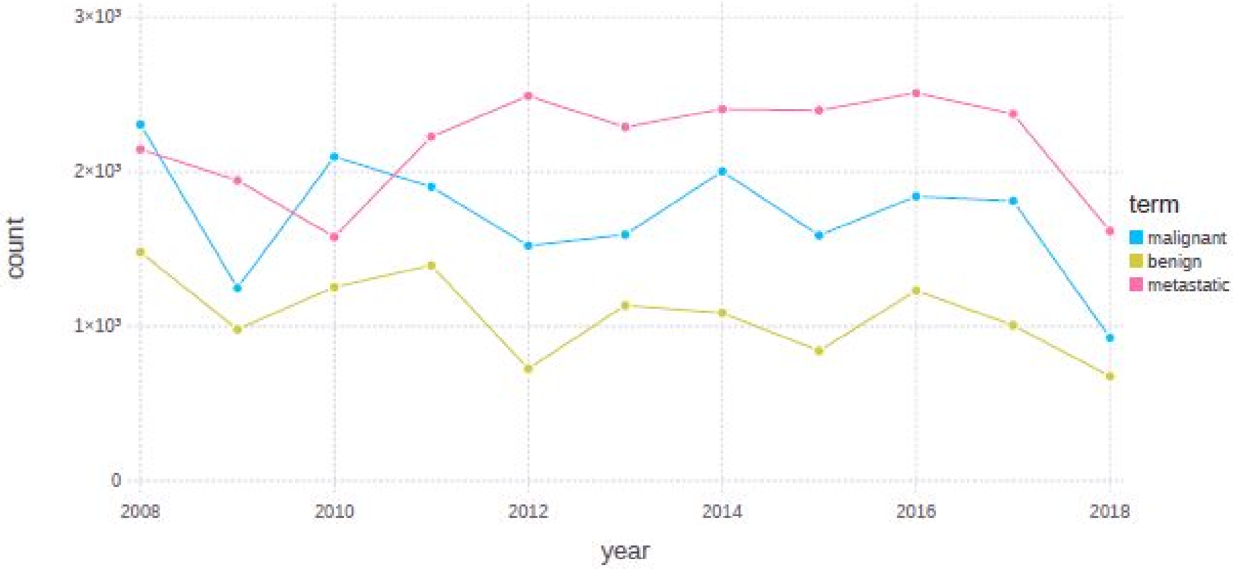

Figure 9. Trend analysis in breast cancer tumors

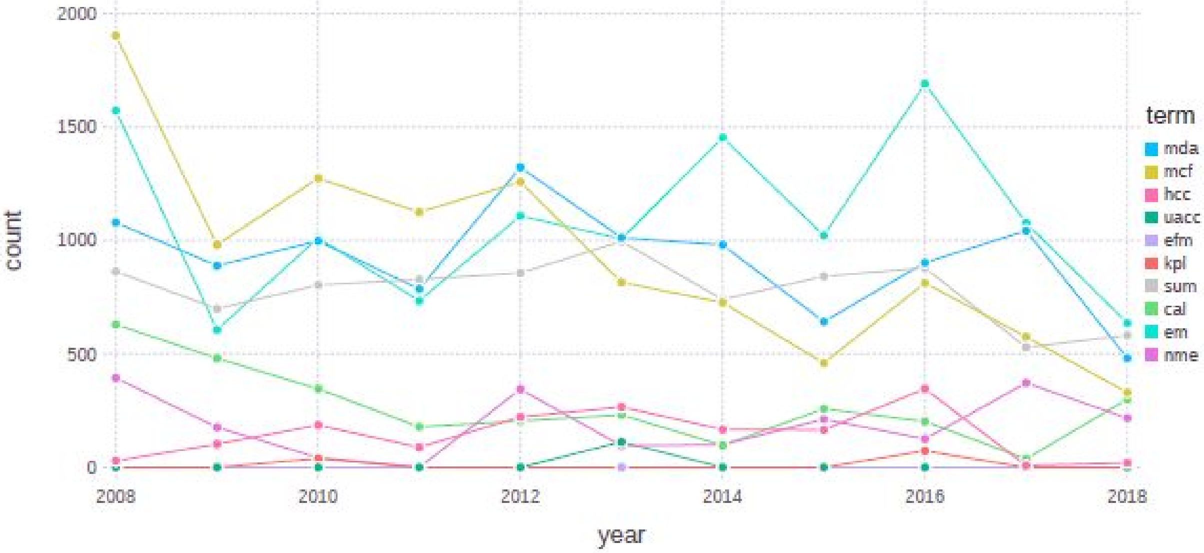

Figure 10. Trend analysis in phenotypes of breast cell lines

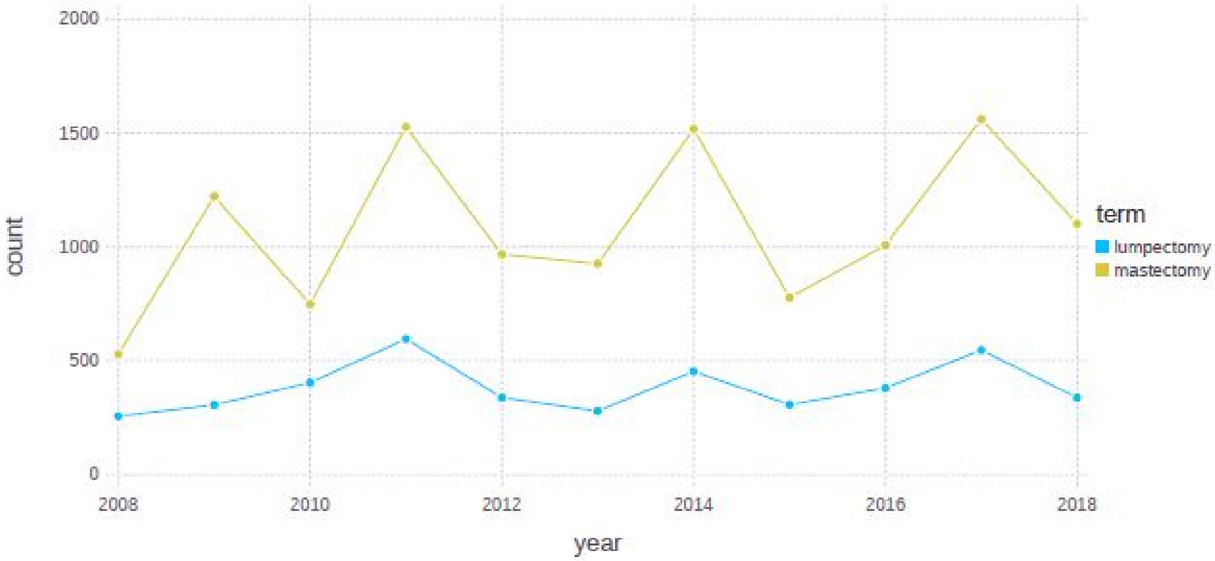

Figure 12. Breast cancer surgery trends

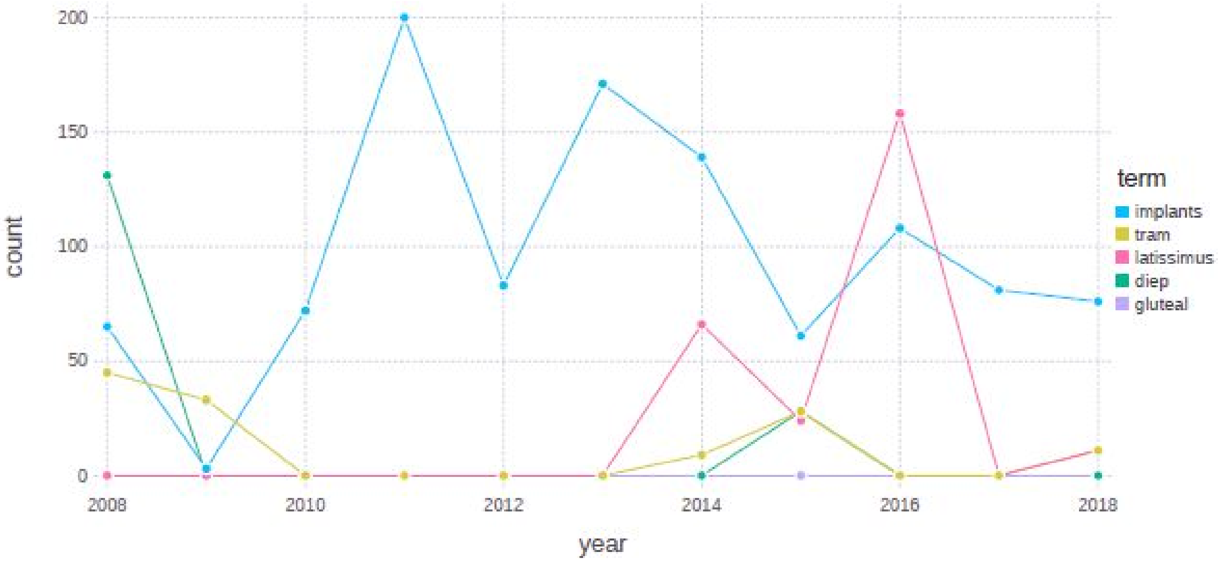

Figure 13. Breast cancer mastectomy surgery trends

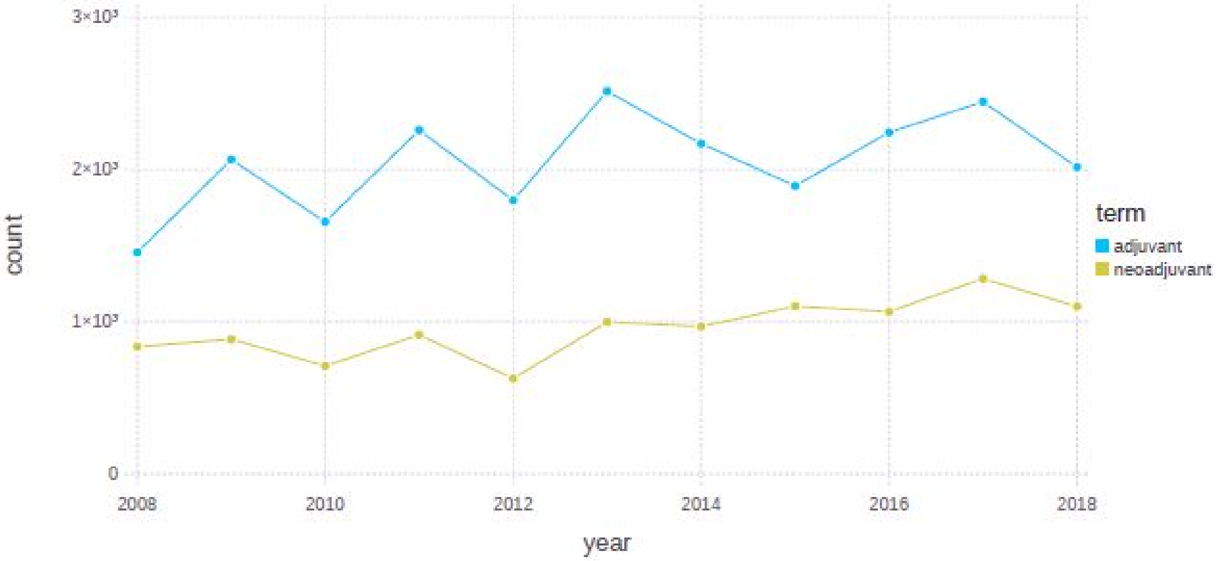

Figure 14. Trends in schedule of Radiation therapy

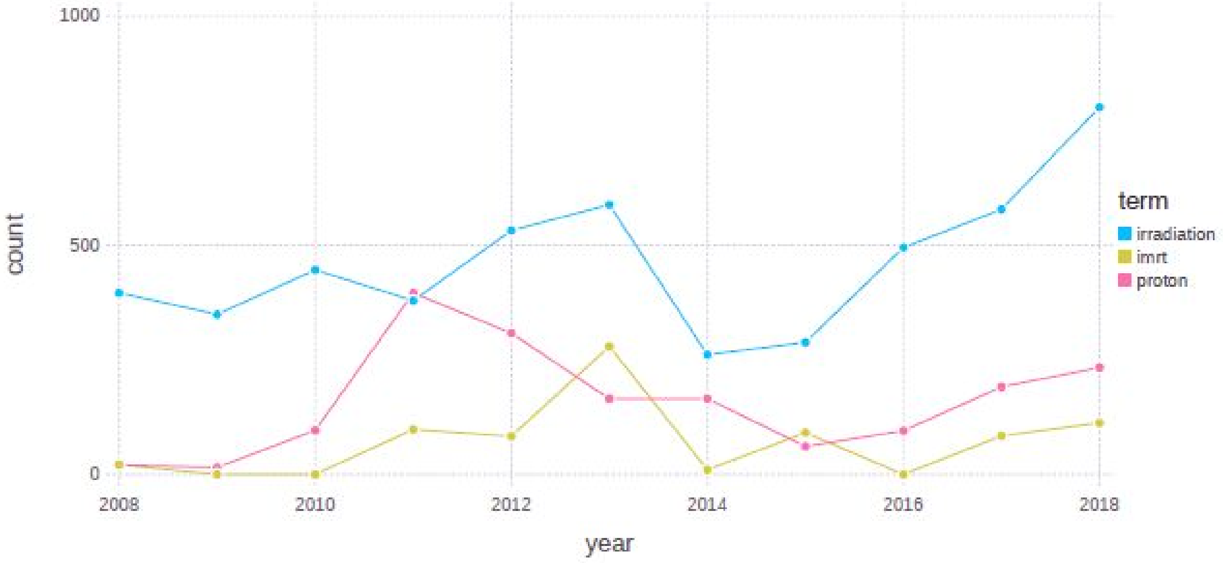

Figure 15. Trends in Types of Radiation therapy

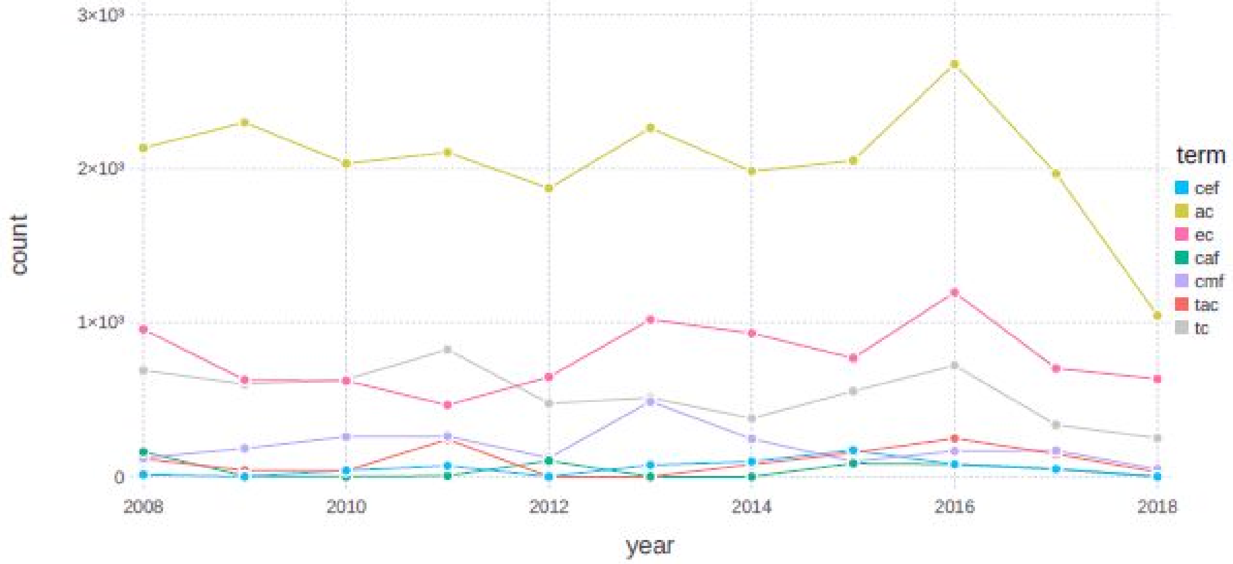

Figure 17. Drug Combination Analysis

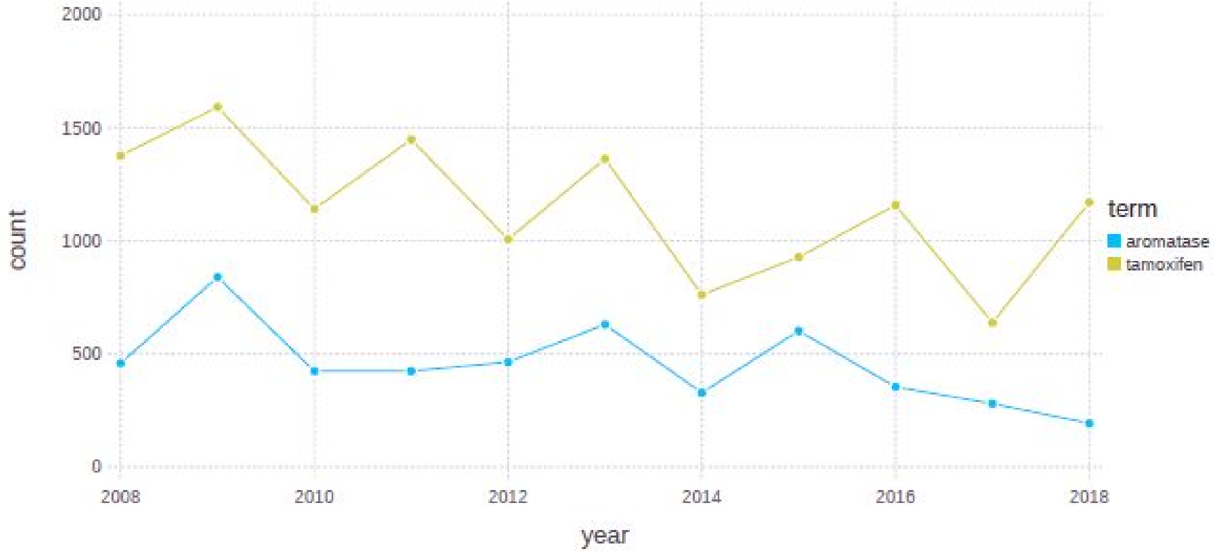

Figure 18. Trend analysis of hormonal therapy

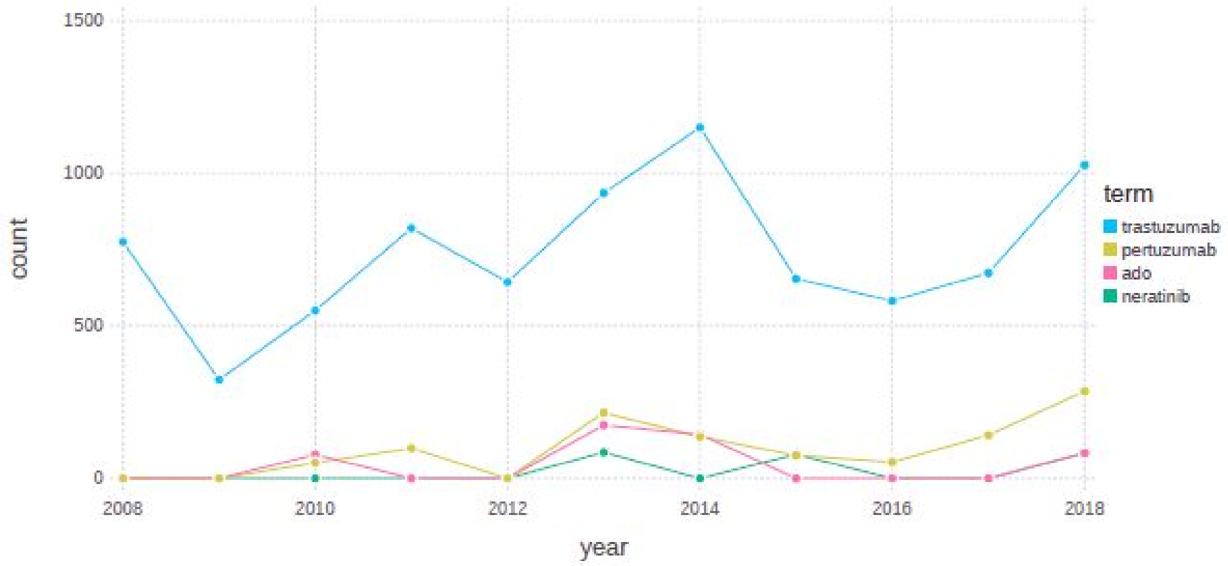

Figure 19. Trend analysis of targeted therapy