Comparison of Altimetric Scores and the Severity of Breast Cancer

Alexander Al-Taraireh, Brady Balk, and Ben Ford

Department of Computer Science, Northern Illinois University

CSCI 636 - MSTR Pattern Recognition & Data Mining I

Hamed Alhoori

December 6th, 2022

Abstract

Objective:

This study is intended to prove that you can use Altmetrics to determine the severity of a case of breast cancer. We believe that the reason a piece of research would be published and shared in greater volume would be in relation to the severity of the breast cancer.

Methods:

The first step to our method was to preprocess the data that we got from the dataset. This meant that we had to filter out nonconforming data to be able to find the specific cancer (breast cancer) that we wanted to work with. The next step was to clean that data by filtering out those with a score of 0 via Altmetrics. We also had to make sure that the data included a stage number, otherwise it would have been no good, so those without a cancer stage number were also filtered out. After this the data was sorted in a descending order and was ready to be loaded into our code. We loaded the data in and using a variety of libraries such as NumPy, Matplotlib, SciPy, and other Python libraries, we were able to produce statistics for the model and then plot out the model using a power regression function rather than linear as a graphical aid.

Results:

Our initial hypothesis was that we believed that the higher the Altmetric score, the higher the stage of breast cancer would be thus the severity would be higher. However, after cleaning and processing the data, our results showed a different hypothesis. It was that there looked to be an inverse relationship between the Altmetric score and the severity of breast cancer, meaning the higher the score, the lower the severity or stage of breast cancer. So, we revised the hypothesis and then ran the data into our code to generate results. The results showed that there was indeed an inverse relationship between the Altmetric score and the severity of breast cancer. The shape of the graphed power regression line was concave up, decreasing, further supporting that hypothesis.

Conclusion:

With our findings, we have reason to believe there are more papers covering breast cancer of a lower severity because of the treatability in the beginning stages of breast cancer. It has been proven that it is extremely valuable to spread awareness and encourage people to seek out a screening for cancer. Catching a tumor before it has metastasized has effectively improved the survival rates of breast cancer. This paired with the advancements of treatment options for cancer neutralizing the disease before it progresses to the later stages is more effective for preserving human life than focusing on advancing research and treatment options for the later stages of breast cancer.

Introduction & Related Work

Cancer is defined as a disease in which some part of the body's cells begins to grow uncontrollably and spread to other areas of the body. Normally the rate at which human cells divide and grow is regulated by replacing old or damaged cells that have died but, in the instance, where this methodical process breaks down, mutated or damaged cells rapidly divide

when they shouldn't. At this stage the cells form into tumors or lumps of tissue and are categorized as cancerous or benign. Cancerous tumors will metastasize or travel to different areas of the body wreaking havoc and destroying the body.

Breast cancer begins in the cells lining the ducts or lobules in the tissue of the breast. Initially the growth will be confined to the ducts or lobules with both minimal symptoms and potential to metastasize but after some time has passed it may progress, invading surrounding breast tissue and nearby lymph nodes with the potential of metastasizing to the rest of the body. The lymph nodes are a network of ducts or vessels that work as part of your immune system to carry aptly named lymph fluid through the body tissues to the blood. The lymph node's connection with your blood stream is what allows the developed cancer cells to travel to other areas of your body, however, cancer that has spread to your lymphatic system is not guaranteed to have metastasized.

According to the World Health Organization ("Breast Cancer"), in 2020 there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally. Additionally, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years making it the most prevalent cancer. Early detection programs along with new mods of treatment to eradicate the disease have improved the survival rates of Breast Cancer starting in 1980. Breast cancer's prevalence paired with the effectiveness of treatments if it is in its early stages magnifies the significance of spreading awareness to both the masses and the scientific community. Even if all potentially modifiable risk factors are completely avoided by a patient, at most it only reduced the chances of developing breast cancer by 30%. The greatest risk factor is the Female gender as only 0.5-1% of all breast cancers occur in men.

Fortunately, there has been a large and effective emphasis for spreading awareness of breast cancer. A study using Google Trends and Twitter to compare the effectiveness of prostate cancer awareness month and breast cancer awareness month (Johnson, Bradley S., et al.) found that while prostate awareness month did not generate substantial internet interest breast cancer awareness month succeeded in doing so despite both sharing common features. Breast cancer and prostate cancer are both the most common hormonal-dependent cancers with a similar prognosis and survival rates. Breast cancer awareness month is so effective at spreading awareness that the percentage of American women who received mammograms doubled only ten years after starting.

The awareness generated for breast cancer extends beyond spreading information to the populous as shown in a study looking for a correlation between cancer research trends and the importance of cancers based on mortality and diagnosis rates (Chou, Peter, er al.). News for Cancer research is generally infrequent at best, only covering large medical breakthroughs, danger to the public, medical research, and is very rarely ever covered in depth. Due to its infrequent coverage the public is not aware of the process in which fields of medical research receive their funding or how the funding is used. Using computational tools such as Python, R, and Microsoft Excel it was found that breast cancer was the most researched cancer overall by a substantial margin. Breast cancer was receiving more research attention than the mortality and diagnosis rate indicated. There were four times more research papers exclusively on breast cancer dwarfing prostate cancer, the second most researched cancer. However, despite the quantity of papers, majority of the papers in the field of cancer did not gain significant traction in online or media outlets.

With the volume of papers present in the world it's extremely critical to be able to identify what the best sources are for both doctors and researchers to develop or identify the best treatments for the patients. Identifying the highest quality articles is valuable for all areas of science, and because of this many different measurement tools have been developed over the years. As stated in a study on the comparison of cancer research on social media versus academia (Celik, Emir, et al.) citation of a research article is one of the primary and important indicators of quality. Despite how powerful, concise, and used this method is, citations alone are not a sufficient indicator of quality. Its greatest weakness is the amount of time it takes for the article to accumulate citations. In medicine especially, any time wasted in finding a cure might mean it is too late for a patient to be treated.

The issue of untimely data collection is best highlighted by a study using Google Trends data to examine interest in cancer screening (Shootman, M, et al.). As aforementioned, the best-case scenario for all types of cancer is early diagnosis. There are several surveillance systems that are in place to track self-reported cancer screening such as the Behavioral Risk Factor Surveillance System (BRFSS), the Health Information National Trends Survey (HINTS), and the National Health Interview Survey (NHIS). The traditional survey methods these databases make use of are drastically ill equipped for the digital age and are expensive to maintain due to their collection methods. The methods require a participation of many people, and in order to ensure accuracy there are strict limitations they rely on adding potential bias and time complexity. A new method of rapid data collection is desperately needed.

In the hopes of shortening the time it takes to identify and qualify scientific articles new tools have been used for both data collection and quantifying their usage. The two most stand out tools I've found are Google Trends and Altmetrics. Google Trends excels in the collection of data measuring public interest. It works by allowing the user to search for key words in other users of Google's searches within a given range and/or region. A study using Google Trends to measure public interest in breast cancer screening in Malaysia (Mohamad, Mazlyfarnia, and Hui Sin Kok) was able to determine that public interest in breast cancer screening strongly correlates with breast cancer awareness month, or pink October. The use of Google Trends allowed them to survey an extremely large pool of people comparing the relative search volume of breast cancer with both other months of the year and other years. Despite how cheap, powerful, and fast of a tool it is for data collection, Google Trends still has some limitations. First, Google is not the only search engine available. The usage of other search engines is not tracked by Google Trends and other web services don't off their own version of it for additive data. Secondly, regional and other availability disparities exist especially in lesser developed countries skewing data toward more urban environments.

A similar study using Google Trends data to study public interest in breast cancer screening in Brazil (Vasconcellos-Silva, Paulo Roberto, et al.) found stand out increases in interest in early breast cancer screening during breast cancer awareness month over a term of five years. The data also showed a sharp downward trend during December and January consistently over the five-year period of interest they studied. Without Google Trends, collecting public interest over a five-year period would be no simple task. The costs of surveying the same number of people for the same amount of time would make it next to impossible to pull off without government backing. That fact alone shows the value of new data collection tools in the digital age.

To answer our question of if Altmetric numbers correlate to severe breast cancers it was important for us to understand how Altmetrics work and what they are quantifying. Altmetrics are a qualitative data tool that is designed to be complementary to the traditional citation-based metrics we have covered in this paper ("What are Altmetrics?"). They are sourced entirely from the internet using a vast range of potential publication methods such as peer reviews, public documents, blogs, mainstream media coverage, and social media networks. Altmetrics show you how often a certain scholarly material or journal article is being referenced. More specifically, they are showcasing the attention and influence research is garnering on the internet.

Aside from quantifying the attention research is gathering Altmetrics also indicate three other main things. First, with a clear record of attention the metrics indicate how many people have been exposed to or engaged with specific research. Understanding the impact of your research is very valuable especially when trying to spread awareness like what is classically done during cancer screening campaigns. Second, you can determine where, why, and how research is being shared and discussed. In the face of data, it's important to distinguish between positive and negative attention especially in the medical field. Mistaking research that has received a lot of negative attention due to the high score as trustworthy can be catastrophic. Lastly, you can use Altmetric scores to detect a breakthrough or change in a field of study. Any of these three dimensions of indication can tell a more detailed story of a paper's value than citation counts are able to achieve on their own.

A previous study attempted to determine if Altmetrics correlate with the quality of papers (Bornmann, Lutz, and Robin Haunschild). The papers were selected based on the traditional method of peer assessments and citations to determine if the top papers based on peer review aligned with their Altmetric scores. Since Altmetrics use non-traditional metrics, they represent an alternative form of impact instead of using citations. The widespread coverage of social media makes altmetrics synonymous with a heterogeneous set of metrics gaining popularity among researchers, publishers, and research funders. For the purpose of determining the quality of research it is questionable to include the measurement of tweets in the altmetric score since the authors of tweets generally work outside the science space and are easily influenced by each other on the platform. It was determined that while altmetric scores don't indicate quality, they are a good measurement of other dimensions.

Another study wanted to determine if altmetrics could predict future citation counts in critical care medicine publications (Lehane, Daniel, and Colin S Black). Social media is increasingly being used for the distribution of medical research and the traditional methods of tracking citations do not account for this. In this instance, altmetrics are the needed measure of the spread of a publication on the internet. It's also important to note that scientific communities are increasing their social media presence to reach a wider audience. Social media is being used to educate the masses and distribute as much knowledge as possible. Because social media is being used by more scientists, papers that generated a lot of scientific impact were found to have spread on social media. However, in the case where the paper generated a large social media impact, it was found that they were much less likely to also have a substantial scientific impact. In short, articles that were highly sited were likely to generate a larger altmetric score but those with a high altmetric score were unlikely to generate a high citation count. For that reason, it was concluded that altmetrics couldn't be used to predict future citation counts. This is unfortunate,

since the process of spreading information on social media is a lot faster than the process of generating citations for an important paper.

Lastly, I'd like to refer to a study that wanted to test if highly cited cancer articles would correlate positively with altmetric scores. In the comparison of attention for cancer research on social media versus academia (Çelik, Emir, et al.) its focus was centered around determining the effectiveness of social media in the dissemination, promotion, and general display of medical literature. Gathering the top 50 most cited articles they compared the altmetric attention scores discovering that while there was a positive correlation between the average citations per year and the altmetric attention score, there was no correlation found between citation number and altmetric attention score. Determining once again that altmetric scores provide useful but different information regarding the impact of scientific materials.

What all three of these studies have in common is they have determined that despite altmetrics being a valuable tool for measuring the dissemination of scientific materials on social media it is not a replacement for traditional tools of determining the success or measuring the quality of a paper. At best it can server as a secondary or supporting data point of the paper's distribution on the internet or more specifically media platforms. For that reason, we were inspired to find if the qualitative metrics of altmetrics could be used to determine the severity of the breast cancer the paper is covering. Cancer and other serious illnesses do not discriminate between age, race, income, or religion. Being able to better predict the seriousness of an illness at an early stage will give the patients more time to live and explore treatment options. Using altmetrics, we want to be able to provide an accurate model that will provide a reasonable degree of certainty a patient's data aligns with the illness they have matched with. We've hypothesized that the more severe a breast cancer is the higher its altmetric score would be. This correlation between the severity of breast cancer and how often research is published would generate an organized model for a treatment option to be selected from.

Innovation

Innovation comes from making a unique and impactful project. Our project will offer innovation in that we are looking at finding out an answer to a question that has not been done yet. We are striving to answer a specific problem posed by a very problematic disease. Typically, things like this are looked at as a whole, that is cancer is examined as a disease as whole. This is because many variants have the same outcomes; sickness, tumors, and sometimes even death. However, we are looking at a very specific subset of cancer, breast cancer. This gives our project the uniqueness it needs. That targeting of a specific subset of a larger problem is what gives this project innovation. We further improved that innovation by applying a power regression model. This type of model isn't unique by itself, but when combined with a unique problem and dataset, it becomes innovative.

Materials & Method

The first step to our method was to find a dataset that would fit with what we wanted to research. We needed something that had a lot of data in it, as well as data that was good. Using the provided options, we found a dataset that fit our criteria with tons of data to use. It contained thousands of examples of research publications and their corresponding Altmetric score, which

was exactly what we needed. The next step was we had to narrow down the dataset to only include our specific type of cancer that we were researching. Using a filter, everything but breast cancer related items were filtered out and rendered useless for this project. After this, we had to make sure that a particular stage of cancer was included in the title of the work. Another filter was used that looked for the word stage and a number (severity), so anything without that was once again filtered out. The final step of cleaning our data was that we needed it in an orderly fashion to make loading it in easier. So, we sorted the data by descending order, to have the highest Altmetric scores first to make the visualization step easier. The next step was to load the finalized data into our Python code. This was achieved by using multiple libraries to load, plot, and calculate the statistics of the data, including slope, R-value, and R-squared. Using Pandas, we loaded the data into different paired data arrays via the NumPy array function. The next step was we needed some of the statistics for our data set. So, we used NumPy to calculate some of the more integral statistics that would help to prove the validity of our model to be printed out. The last step in the process was to get the data graphed visually, making it easier for people to see the trend. So, we used the SciPy Python package to implement a power regression model. We chose to use SciPy because it has the functionality we need built in, so generating a power regression was as simple as defining what we needed and then SciPy did the math intensive portions automatically. However, to display the data we went with the Matplotlib Python library because it is very intuitive and produces great visual aids. All we had to do was call the power regression that SciPy made and Matplotlib automatically graphed it and displayed it. The output of the statistics was simply handled by standard output of Python, as there was no special library needed to output those results.

Results

Obtaining results that matched our original hypothesis was impossible after we cleaned all of our data needed to generate a model. We realized that our hypothesis would have to change, and thus we revised it to fit what we believed the new data was going to show. After loading the data, performing the calculations of the statistics, and plotting the data out on graphs; our results shows that there is in fact an inverse relationship between the Altmetric score of a publication on breast cancer and the severity of the breast cancer. That is, the higher the score, the lower the severity of the breast cancer.

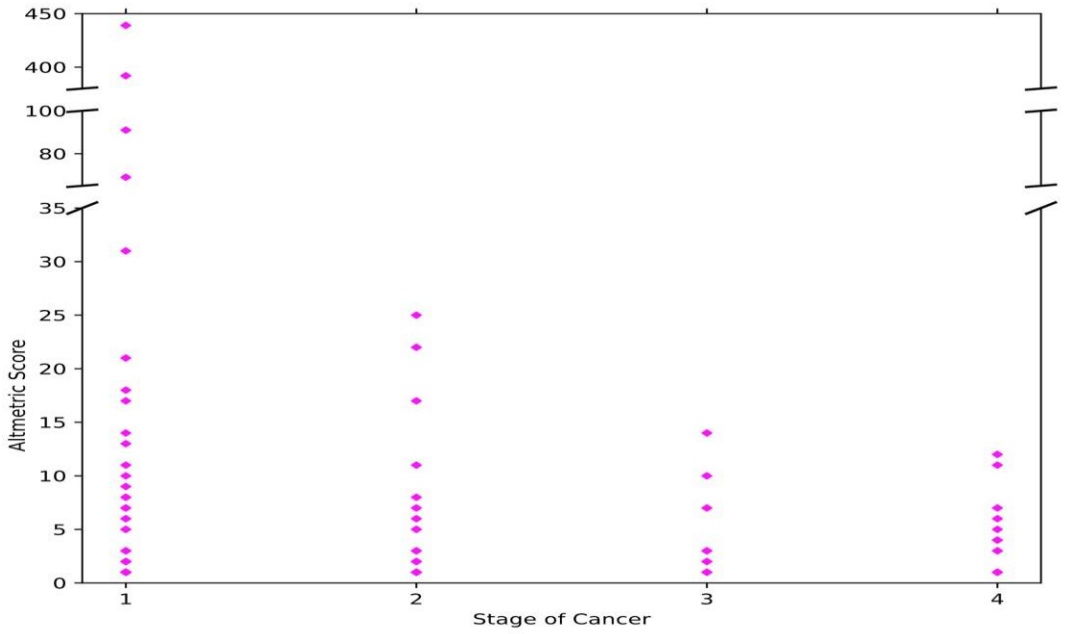
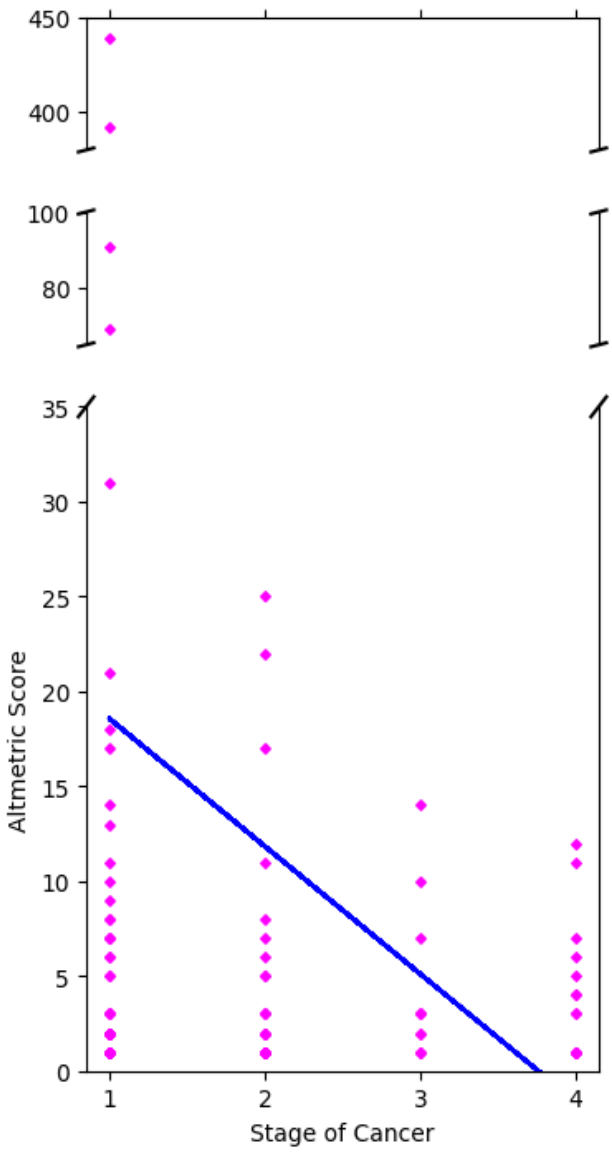


Figure 1.1

The picture above (fig 1.1) are our initial results of plotting the data, before any analysis and trying to find a best fit line for the model. It clearly shows that most of the data, especially the heavy weighted data, favors the left end of the graph (i.e., a lower stage of breast cancer). The right edge of the graph, the higher the stage of breast cancer, shows data points with much lower scores. This distribution will affect the line of the model, likely pulling it down quickly at first and then flattening out towards the right side of graph. This would indicate that the function for the regression line would be power rather than linear like we first thought. The shape would also likely be concave up and decreasing because of where the data points lie.

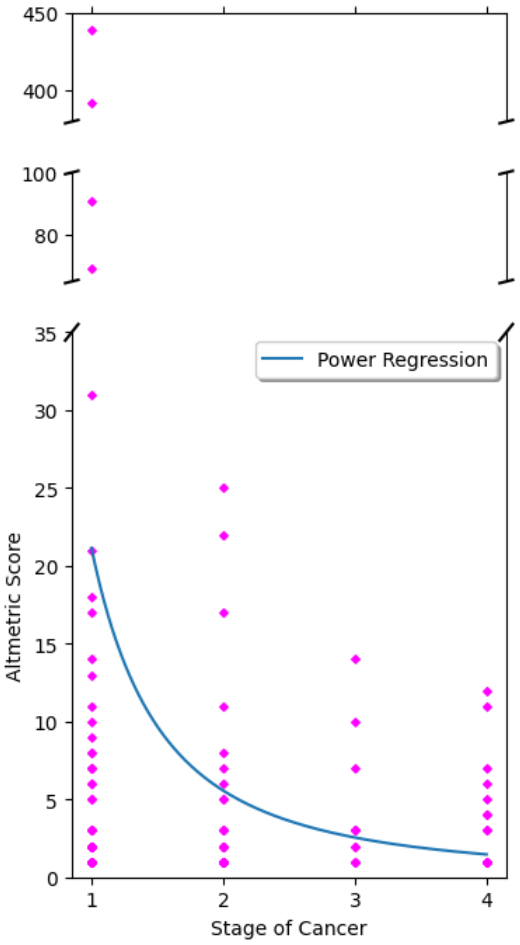


As shown to the left (fig 1.2) linear regression was not a good fit. It lead to terrible statistics including a very low R^2 value, thus proving it was way under fit for the data. The full results are as follows:

Linear regression statistics:

- Slope: -6.723917672107885
- Y - int: 25.28715400993614
- R value: -0.12657699565901473
- R^2: 0.016021735830062233
- P value: 0.17766251635443928
- Std error: 4.957025291233932

Figure 1.2



Optimal values (min. sum of squares)

- Altmetric Score: 21.134262589266086

- Stage of Cancer: 1.9247412788518965

Variance in Y - Values (Altmetric Scores): 50.584731855121774

Covariance between Almetric Score and Stage of Cancer: 3.106389970302665

Covariance score is postive, larger than R^2, thus indicating power regression is a good fit

Covariance Matrix Row 1

[50.58473186 3.10638997]

Covariance Matrix Row 2

[3.10638997 5.12180728]

Standard deviation (Sigma) of Altmetric Scores = 7.112294415666563

Figure 1.3

As predicted, the figure above (fig 1.3), shows our results ended up with a concave up, decreasing type of line in a power regression shape. This indicates that the majority of points are near the left (lower stage of cancer), as these points have higher scores. The points to the right (higher stage of cancer) have less of score, thus the line starts high and must come down to fit the lower scores. As the data points on the x-axis are the stages of cancer and there are only 4 distinct stages (without going into subclasses), the plot of the line had to be power rather than regression. This factor also contributed to the shape of the line. Nonetheless, it shows that our hypothesis of there being an inverse relationship between the Altmetric score and the severity of breast cancer is not rejected.

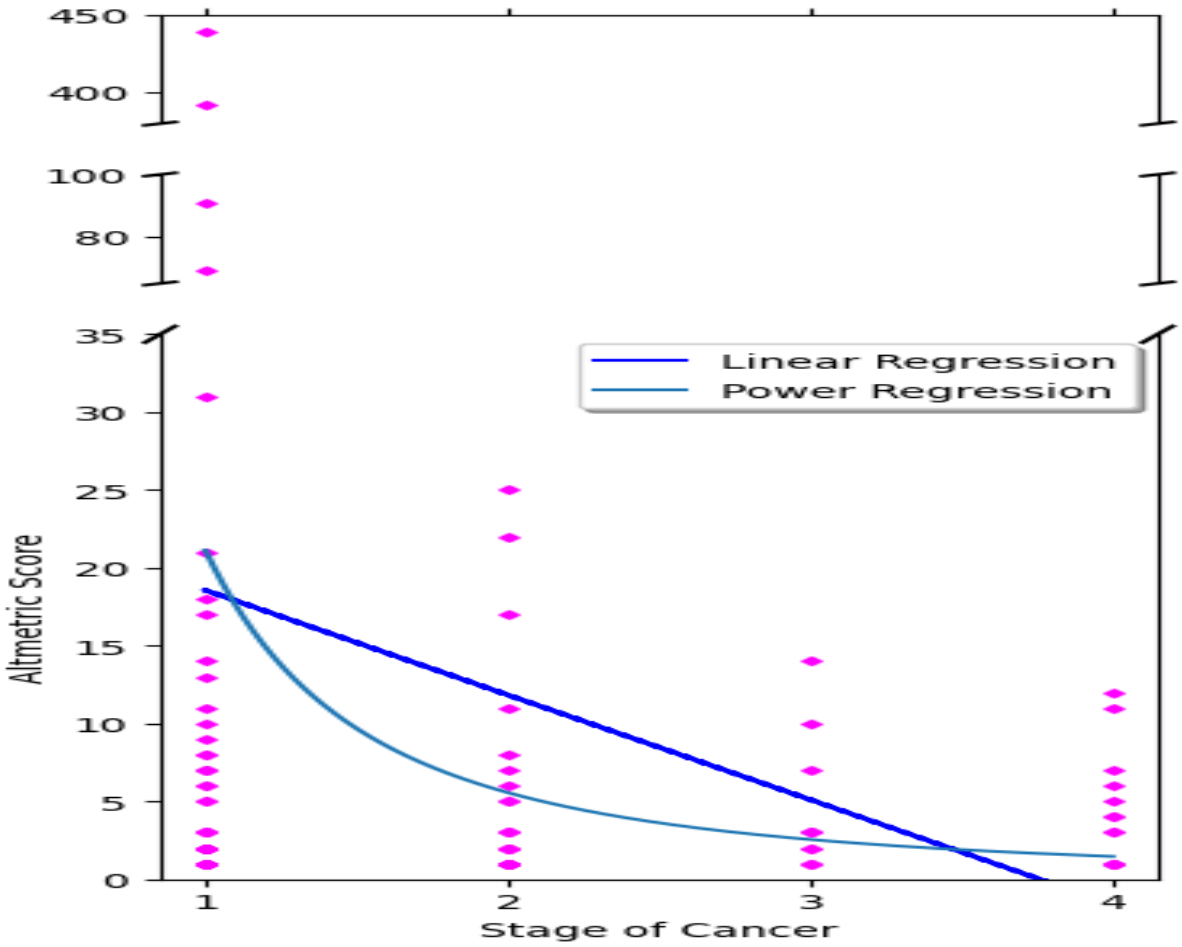


Figure 1.4

The above figure (fig 1.4) shows the two models going head-to-head. It is clear that the linear regression model does not fit the data that well when the power regression model is also displayed. The power regression model takes into account the weight of all of the scores of the data points per stage and that is why it fits so well.

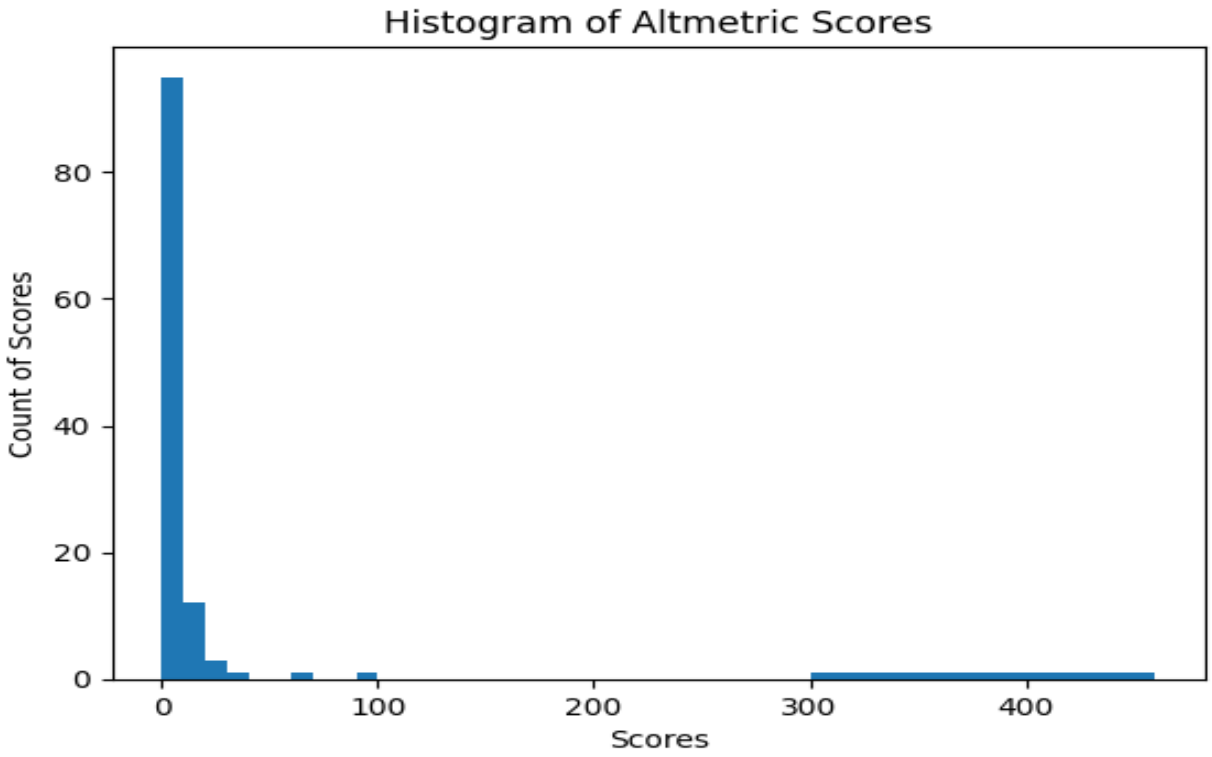


Figure 1.5

The above figure (fig 1.5) is a histogram of the Almetric scores. Bins were in intervals of 10 starting at 0 then going to 100, 300, 400, and 460. The reason is so many scores fall in the 0-100 range, but there are some out outliers between 300 and 460. It illustrates the sheer number of scores that are present in the 0-100 range, with most being paired with a stage 1 diagnosis. This is the reason the power regression model fits best, rather than the linear regression model.

Discussion

Our results indicate that there is an inverse correlation between altmetric attention scores and the severity of breast cancer. Doing further research on this relationship may prove beneficial and provide some insight for what kind of medical research generates public interest and spreads the easiest on internet platforms.

Our findings share an interesting similarity to a study done on online attention to oral cancer research (Hassona, Yazan, et al.). The objective of the study was to identify and analyze research articles about oral cancer that gathered the most online attention. It was found that of 7,940 articles that were identified the contents of the paper were mostly discussed on Twitter, Mendeley, Facebook, and news outlets however the ones that generated the greatest attention were mainly focused on treatment outcomes and quality of life. These being topics of general interest such as vaccination, risk factors, HPV transmission, and treatment. In align with these interests, our findings showed that breast cancer cases that demonstrated treatments, quality of life after those treatments and other positive recovery statistics governed higher altmetric attention rates.

Media's preference for positive treatments and information on quality of life doesn't seem to be the only reason for lower severities of breast cancer to be generating higher altmetric attention scores. Closely associated with breast cancer specifically are the massive campaigns every October to spread awareness and encourage self-administered checks and screening for early breast cancer detection. As we learned, the awareness garnered for breast cancer also greatly inflated the number of medical resources directed into research generating a larger quantity of papers on the topic. The campaigns target early detection because it has proven to be the most effective countermeasure for all cancerous diseases but is especially effective for breast cancer given how non-life threatening it is in the early stages. These factors combined will all contribute to larger media coverage and attention being directed to articles on the early stages of breast cancer.

Overall, media coverage of medical materials can be easily swayed by individuals with a lot of online credibility or movements started by organizations such as charities that spend a lot of money in order to spread awareness on an issue. A paper covering altmetric mentions and the identification of research impact (Loach, Tamar, and Jonathan Adams) aiming to find if media mentions could provide indicators of outcomes linked to social and technical innovation for stakeholders in networks beyond academia. Noting that a high altmetric score does not correlate to a high number of citations it becomes difficult to prove that papers have an impact on the general population and not just in academics. What was found during the statistical analysis is that papers that were associated with diseases covered by charities with large research funds had higher altmetric scores than lowers spending charities coverings diseases such as Musculo-skeletal or Immune. The standout exception of this trend was cardiovascular research which garnered more attention than its spending would indicate.

With this relationship in mind, the Breast Cancer Research Foundation, the largest private funder of metastatic breast cancer research, alone has committed $47.4 million to fund breast cancer research for 2021-22. They'll be funding 250 scientists and leading academic and medical institutions across 14 countries (BCRF). The research is centralized around developing innovative treatments, studying the basic make up of the cancer cells, and improving quality of life. Their organization alone boasts one of the largest international efforts exclusive to metastatic breast cancer research.

Breast Cancer Research Foundation's contributions are still only a fraction of the attention that breast cancer gets in the scope of funding. Awareness campaigns for breast cancer are so effective that breast cancer has become one of the most over-funded cancer's every year. A study on where public funding for cancer research goes (Trasta, Anthi) found that the distribution of funds for common complex diseases is extremely inequitable. In 2015, the mortality of lung cancer was triple that of breast cancer, however, the funding allocated for breast cancer was double that of lung cancer.

The funding and media attention of breast cancer is stand out when compared to other diseases. The high cash flow and mass attention given to building awareness every October has a very high potential of effecting the altmetric scores of breast cancer research papers making it difficult to draw any conclusions on correlations between altmetric scores and breast cancer research. Since other factors need to be considered when using the quantitative data, it has become clear that altmetric measurements are best used with other statistics as complimentary data to other

measures of social impact. It would be beneficial for altmetric measurements to be filtered to exclude mentions that were the result of a campaign or movement artificially inflating the attention given to specific research.

Conclusion

Our results showed that there was an inverse relationship between altmetric scores and the severity of breast cancer. With our findings, we have reason to believe there are more papers covering breast cancer of a lower severity because of the treatability in the beginning stages of breast cancer. It has been proven that it is extremely valuable to spread awareness and encourage people to seek out a screening for cancer. Catching a tumor before it has metastasized has effectively improved the survival rates of breast cancer. This paired with the advancements of treatment options for cancer neutralizing the disease before it progresses to the later stages is more effective for preserving human life than focusing on advancing research and treatment options for the later stages of breast cancer.

References

Lehane, Daniel J, and Colin S Black. “Qualitative Research - Journals.sagepub.com.” *Sage Journals*, 2 Feb. 2020, https://journals.sagepub.com/doi/full/10.1177/14687941221132956.

Bayar, Bahattin, and Rıfat Peksöz. “Analysis of Trending Topics in Breast Cancer Articles from an Altmetric Perspective.” *Cureus*, 5 July 2022, https://www.cureus.com/articles/101281-analysis-of-trending-topics-in-breast-cancer-articles-from-an-altmetric-perspective.

Bcrf. “BCRF Announces 2021-22 Investment.” *Breast Cancer Research Foundation*, BCRF, 11 Feb. 2022, https://www.bcrf.org/blog/bcrf-2021-2022-research-investment/.

Bornmann, Lutz, and Robin Haunschild. “Do Altmetrics Correlate with the Quality of Papers? A Large-Scale Empirical Study Based on F1000prime Data.” *PloS One*, U.S. National Library of Medicine, 23 May 2018, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5965816/.

“Breast Cancer.” *World Health Organization*, World Health Organization, 26 Mar. 2021, https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

Celik, Emir, et al. “Comparison of Attention for Cancer Research on Social Media versus ...” *Research Gate*, Jan. 2020, https://www.researchgate.net/profile/Betul-Borku-Uysal/publication/340264974_Comparison_of_Attention_for_Cancer_Research_on_Social_Media_Versus_Academia_An_Altmetric_Score_Analysis/links/61ebb8c88d338833e3877961/Comparison-of-Attention-for-Cancer-Research-on-Social-Media-Versus-Academia-An-Altmetric-Score-Analysis.pdf.

Chou, Peter, et al. “Correlation between Cancer Research Trends and the Importance of Cancers Based on Mortality and Diagnosis Rates: An Analysis of Altmetric Data.” *STEM Fellowship Journal*, 23 Nov. 2017, https://journal.stemfellowship.org/doi/10.17975/sfj-2017-014.

Goksoy, Beslen, and Hilmi Bozkurt. “Social Attention of the Top 50 Scientific Articles on Gastric Cancer ...” *Research Gate*, Dec. 2020, https://www.researchgate.net/profile/Beslen-Goksoy/publication/346653939_Social_attention_of_the_top_50_scientific_articles_on_gastric_cancer_Bibliometric_and_altmetric_analysis/links/600cb23a45851553a0678de9/Social-attention-of-the-top-50-scientific-articles-on-gastric-cancer-Bibliometric-and-altmetric-analysis.pdf.

Hassona, Yazan, et al. “The Online Attention to Oral Cancer Research: An ... - Wiley Online Library.” *Wiley Online Library*, 29 Apr. 2019, https://onlinelibrary.wiley.com/doi/abs/10.1111/odi.13111.

Johnson, Bradley S., et al. “Using Google Trends and Twitter for Prostate Cancer Awareness: A Comparative Analysis of Prostate Cancer Awareness Month and Breast Cancer Awareness Month.” *Cureus*, 13 Feb. 2021, https://www.cureus.com/articles/47045-using-google-

trends-and-twitter-for-prostate-cancer-awareness-a-comparative-analysis-of-prostate-cancer-awareness-month-and-breast-cancer-awareness-month.

Loach, Tamar, and Jonathan Adams. “Altmetric ‘Mentions’ and the Identification of Research Impact - NESTA.” *Nesta*, Feb. 2015, https://media.nesta.org.uk/documents/nesta_altmetrics_working_paper.pdf.

Mohamad, Mazlyfarina, and Hui Sin Kok. “Using Google Trends Data to Study Public Interest in Breast Cancer Screening in Malaysia.” *Asian Pacific Journal of Cancer Prevention : APJCP*, U.S. National Library of Medicine, 25 May 2019, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6857874/.

Munnolli, Satish S, and Shamprasad M Pujar. “Scientometric Study of Indian Cancer Research Based on Scopus Database.” *Taylor & Francis*, 20 Oct. 2017, https://www.tandfonline.com/doi/abs/10.1080/09737766.2017.1284729.

Pujar, Shamprasad. “The Impact of Indian Cancer Research in Social Media: A Study Using ...” *Research Gate*, Nov. 2015, https://www.researchgate.net/profile/Satish_Munnolli/publication/286921576_The_Impact_of_Indian_Cancer_Research_in_Social_Media_A_Study_using_Altmetric_explorer/links/566fb4ab08ae4d9a42572aa2/The-Impact-of-Indian-Cancer-Research-in-Social-Media-A-Study-using-Altmetric-explorer.pdf.

Rooney, Michael Kevin, et al. “Article Tools.” *Journal of Clinical Oncology*, 2 June 2022, https://ascopubs.org/doi/abs/10.1200/JCO.2022.40.16_suppl.11014?role=tab.

Schootman, M, et al. "The Utility of Google Trends Data to Examine Interest in Cancer Screening." *BMJ Open*, British Medical Journal Publishing Group, 1 June 2015, https://bmjopen.bmj.com/content/5/6/e006678.

Taylor, Mike. "City University of New York." *Research Trends*, 6 Jan. 2013, https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1229&context=le_pubs.

Trasta, Anthi. "Where Does Public Funding for Cancer Research Go: Allocation of Research Funding for Cancer and COPD Is Not Always Proportional to Disease Burden." *EMBO Reports*, U.S. National Library of Medicine, 19 Feb. 2018, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5836059/.

Vasconcellos-Silva, Paulo Roberto, et al. "Using Google Trends Data to Study Public Interest in Breast Cancer Screening in Brazil: Why Not A Pink February?" *JMIR Public Health and Surveillance*, U.S. National Library of Medicine, 6 Apr. 2017, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5399222/.

"What Are Altmetrics?" *Altmetric*, 17 Dec. 2018, https://www.altmetric.com/about-altmetrics/what-are-altmetrics/.

"What Is Breast Cancer?" *American Cancer Society*, 19 Nov. 2021, https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html.

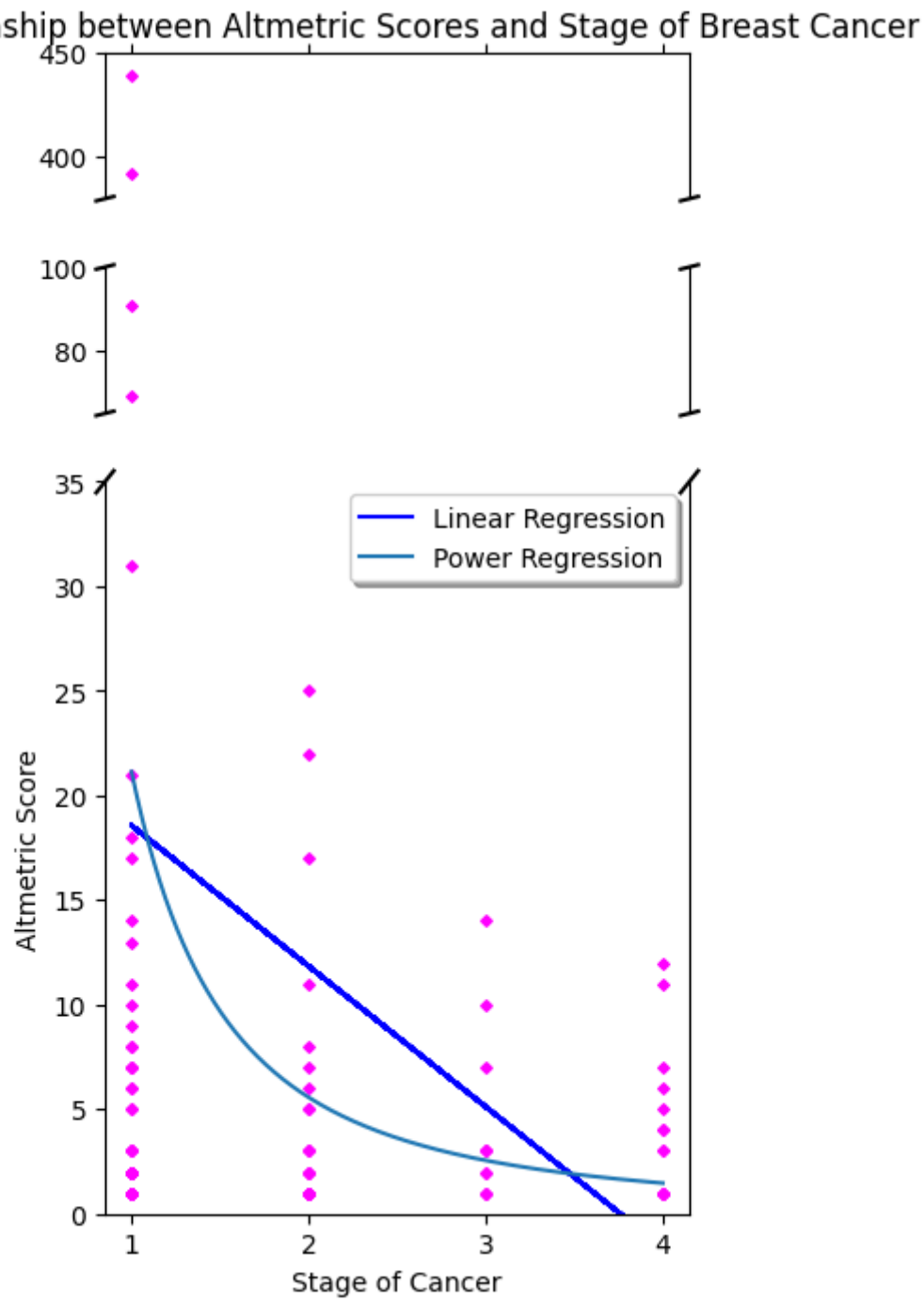
"What Is Cancer?" *National Cancer Institute*, 5 May 2021, https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

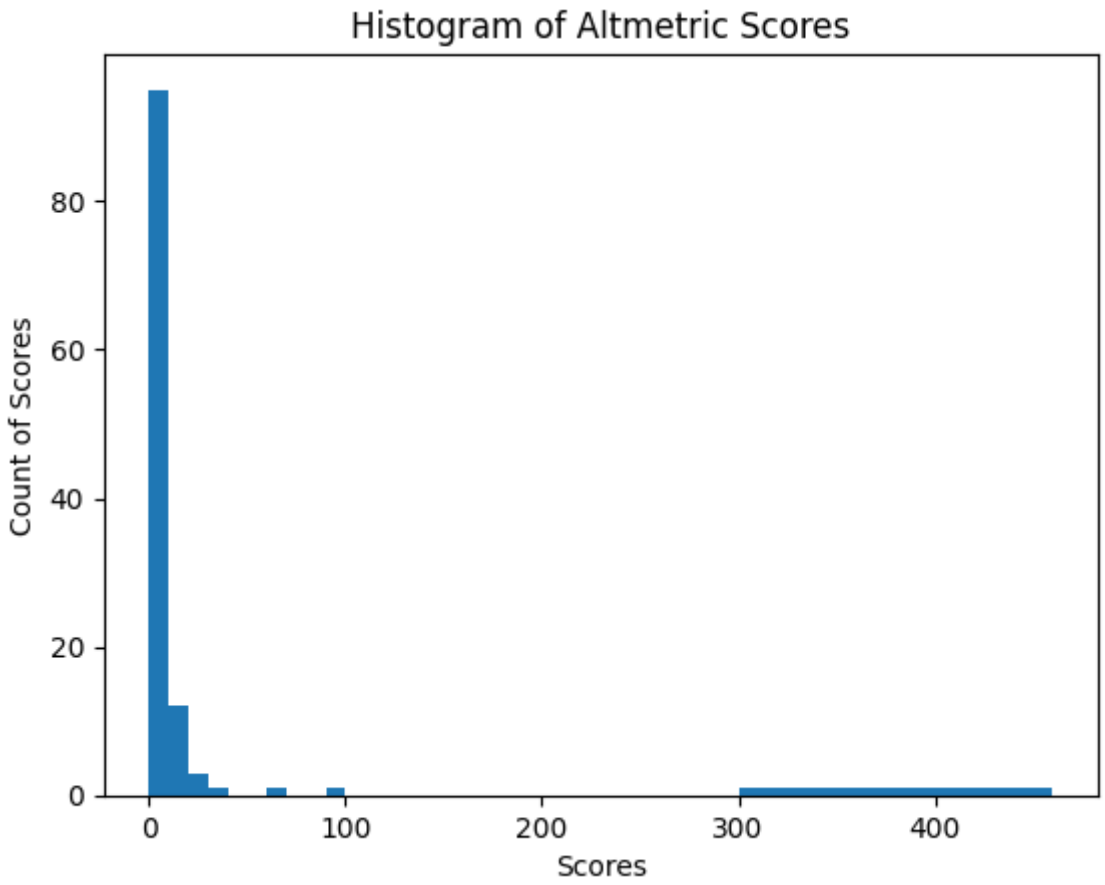
Çelik, Emir, et al. “Comparison of Attention for Cancer Research on Social Media versus Academia: An Altmetric Score Analysis.” *DSpace@Biruni*, Akademi Doktorlar Yayınevi, 1 Jan. 1970, http://openaccess.biruni.edu.tr/xmlui/handle/20.500.12445/1255.

In [104…

```
# Relation Between The Altmetric Score and Severity of Cancer
# Programmers: Ben Ford, Brady Balk, Alexander Al-Taraireh
# The purpose of this graph is to demonstrate and present the relationship betw
# DATA SOURCE: https://www.dropbox.com/s/8jjzdcbtgyx9lht/Dimensions-Clinical-Tr
# Libraries:
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import linregress
from scipy.optimize import curve_fit
# Creating the data:
# x represents the Stage of Cancer
# y represents the Altmetric Score
x = np.array([2,3,1,1,2,4,1,1,1,2,2,1,2,2,2,2,2,1,1,1,1,1,1,1,1,1,2,4,4,4,4,2,2
y = np.array([1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
# Get linear equation
(slope, intercept, rvalue, pvalue, stderr) = linregress(x,y)
# Power regression equation
popt, pcov = curve_fit(lambda fx,a,b: a*fx**-b,  x,  y)
x_linspace = np.linspace(min(x), max(x), 456)
power_y = popt[0]*x_linspace**-popt[1]
perr = np.sqrt(np.diag(pcov))
# Changing shape with marker
f, (ax,ax2,ax3) = plt.subplots(3,1,sharex=True, figsize=(4, 8), gridspec_kw={'h
ax.scatter(x, y, s=8, color=[1,0,1], marker="D")
ax.title.set_text('Relationship between Altmetric Scores and Stage of Breast Ca
ax2.scatter(x, y, s=8, color=[1,0,1], marker="D")
ax3.scatter(x, y, s=8, color=[1,0,1], marker="D")
ax.set_ylim(380, 450)
ax2.set_ylim(65,100)
ax3.set_ylim(0, 35)
# Hiding the spines between ax, ax2, and ax3
ax.spines['bottom'].set_visible(False)
ax2.spines['top'].set_visible(False)
ax2.spines['bottom'].set_visible(False)
ax2.tick_params(bottom=False)
ax.xaxis.tick_top()
ax.tick_params(labeltop=False)
ax3.spines['top'].set_visible(False)
# Will show how large to make the diagnal lines in the axes coordinates
d = .015
# Arguments to pass to plot, so that they wont be repeated
kwargs = dict(transform=ax.transAxes, color='k', clip_on=False)
# Top-left diagonal
ax.plot((-d, +d), (-d, +d), **kwargs)
# Top-right diagonal
ax.plot((1 -d, 1 + d), (-d, +d), **kwargs)
# Switch to the bottom axes
kwargs.update(transform=ax2.transAxes)
# Bottom-left diagonal
ax2.plot((-d, +d), (1 -d, 1 + d), **kwargs)
# Bottom-right diagonal
ax2.plot((1 -d, 1 + d), (1 -d, 1 + d), **kwargs)
# Top-left diagonal 2
ax2.plot((-d, +d), (-d, +d), **kwargs)
# Top-right diagonal 2
ax2.plot((1 -d, 1 + d), (-d, +d), **kwargs)
```

```
# Switch to the bottom axes
kwargs.update(transform=ax3.transAxes)
# Bottom-left diagonal
ax3.plot((-d, +d), (1 -d, 1 + d), **kwargs)
# Bottom-right diagonal
ax3.plot((1 -d, 1 + d), (1 -d, 1 + d), **kwargs)
# Tick marks for x axis
plt.xticks([1,2,3,4])
# Linear Regression calcuation
y_pred = intercept + slope*x
# Plot linear regression
plt.plot(x,y_pred, color="blue", label="Linear Regression")
# Power regression
plt.plot(x_linspace, power_y, label='Power Regression')
# Labels for x and y axes
plt.xlabel('Stage of Cancer')
plt.ylabel('Altmetric Score')
# Show the graph
plt.legend(loc='best', fancybox=True, shadow=True)
plt.show()
plt.hist(y, bins = [0,10,20,30,40,50,60,70,80,90,100,300,400,460])
plt.xlabel("Scores")
plt.ylabel("Count of Scores")
plt.title('Histogram of Altmetric Scores')
plt.show()
print ("Slope: ",slope, "Y - int: ",intercept, "R value: ",rvalue,"R^2: ", rva
print ("P value: ",pvalue, "Std error: ",stderr)
print ("R^2 Value is very low, thus linear regression is not a great choice")
print ("Optimal values (min. sum of squares) Altmetric Score: ",popt[0], "Stage
print ("Variance in Y - Values (Altmetric Scores): ",pcov[0,0])
print ("Covariance between Almetric Score and Stage of Cancer: ", pcov[1,0])
print ("Covariance score is postive, larger than R^2, thus indicating power reg
print ("Covariance Matrix Row 1", pcov[0])
print ("Covariance Matrix Row 2", pcov[1])
print ("Standard deviation (Sigma) of Altmetric Scores = ",perr[0])
```


Relationship between Altmetric Scores and Stage of Breast Cancer
450
400
100
80
35
30
25
20
15
10
5
0
Altmetric Score
1
2
3
4
Stage of Cancer
Linear Regression
Power Regression




```
Slope:  -6.723917672107885 Y - int:  25.28715400993614 R value:  -0.1265769956
5901473 R^2:  0.016021735830062233
P value:  0.17766251635443928 Std error:  4.957025291233932
R^2 Value is very low, thus linear regression is not a great choice
Optimal values (min. sum of squares) Altmetric Score:  21.134262589266086 Stag
e of Cancer:  1.9247412788518965
Variance in Y - Values (Altmetric Scores):  50.584731855121774
Covariance between Almetric Score and Stage of Cancer:  3.106389970302665
Covariance score is postive, larger than R^2, thus indicating power regression
is a good fit
Covariance Matrix Row 1 [50.58473186  3.10638997]
Covariance Matrix Row 2 [3.10638997 5.12180728]
Standard deviation (Sigma) of Altmetric Scores =  7.112294415666563
```

In [ ]:

Altmetric Attention Score	Title	Stage	Journal/Collection Title	Journal ISSNs	Authors at my Institution	Departments	Output Type	OA Status	Subjects (LoR)	Affiliations (GRID)	Funder	Publication Date	DOI	ISBN	National Clinical Trial ID	URI	PubMed ID	PubMedCentral ID	Handle.net IDs	ADS Bibcode	arXiv ID	RePEc ID	SSRN	URN	News mentions	Blog mentions	Policy mentions	Patent mentions	Twitter mentions	Peer review mentions	Weibo mentions	Facebook mentions	Wikipedia mentions
439	Pre-operative IRX-2	1					Clinical trial study record	FALSE				10/28/2016			NCT02950259										61	0	0	0	0	0	0	0	0
392	Cryoablation of Sma	1	ClinicalTrials.gov				Clinical trial study record	FALSE				11/6/2013			NCT01992250										95	0	0	0	6	0	0	2	0
91	A Phase III Random	1					Clinical trial study record	FALSE				4/9/2010			NCT01101438										10	2	0	0	4	0	0	0	0
69	Study Evaluating The	1					Clinical trial study record	FALSE				4/8/2009			NCT00878709										11	0	0	0	1	0	0	0	0
31	A Study Looking the	1					Clinical trial study record	FALSE				2/17/2015			NCT02400476										4	0	0	0	1	0	0	0	0
25	Aspirin in Preventing	2					Clinical trial study record	FALSE				10/5/2016			NCT02927249										0	0	0	0	52	0	0	11	1
22	Radiation Therapy (	2					Clinical trial study record	FALSE				2/7/2005			NCT00103181										2	1	0	0	2	0	0	0	0
21	Mindfulness Meditati	1					Clinical trial study record	FALSE				1/17/2017			NCT03025139										0	0	0	0	42	0	0	3	0
18	S1200: Acupuncture	1					Clinical trial study record	FALSE				2/14/2012			NCT01535066										2	0	0	0	3	0	0	0	0
17	A Phase 2, 2-Stage	2					Clinical trial study record	FALSE				1/9/2014			NCT02034916										1	1	0	0	4	0	0	0	0
17	Trial of SPI-2012 Va	1					Clinical trial study record	FALSE				11/2/2016			NCT02953340										2	0	0	0	3	0	0	0	0
14	Implantable Microde	1					Clinical trial study record	FALSE				8/5/2015			NCT02521363										2	0	0	0	0	0	0	0	0
14	Paclitaxel and Trastu	3					Clinical trial study record	FALSE				10/9/2008			NCT00770809										2	0	0	0	0	0	0	0	0
13	Lymph Node Remov	1					Clinical trial study record	FALSE				11/1/1999			NCT00003855										1	1	0	0	0	0	0	0	0
12	Pembrolizumab, Let	4					Clinical trial study record	FALSE				5/18/2016			NCT02778685										1	1	0	0	0	0	0	0	0
11	A Study Evaluating S	1					Clinical trial study record	FALSE				12/13/2013			NCT02032277										1	0	0	0	2	0	0	1	0
11	Early Surgery or Sta	4					Clinical trial study record	FALSE				11/16/2010			NCT01242800										1	0	0	0	4	0	0	0	0
11	Open Label Trial Eva	2					Clinical trial study record	FALSE				8/8/2018			NCT03620201										1	0	0	0	4	0	0	0	0
10	A Study Evaluating E	1					Clinical trial study record	FALSE				5/6/2014			NCT02132949										1	0	1	0	0	0	0	0	0
10	Fulvestrant and/or A	3					Clinical trial study record	FALSE				10/1/2013			NCT01953588										0	0	0	0	22	0	0	2	0
9	S0927: Omega-3-Fa	1					Clinical trial study record	FALSE				6/28/2011			NCT01385137										0	0	0	0	12	0	0	0	0
8	Higher Per Daily Trie	1					Clinical trial study record	FALSE				5/5/2011			NCT01349322										0	1	0	0	2	0	0	1	0
8	Hypofractionated Ra	2					Clinical trial study record	FALSE				1/30/2018			NCT03414970										0	0	0	0	22	0	0	4	0
8	Trial to Evaluate Ger	1					Clinical trial study record	FALSE				3/9/2008			NCT00636441										0	1	1	0	0	0	0	0	0
7	Clodronate With or V	2					Clinical trial study record	FALSE				2/2/2001			NCT00009945										1	0	0	0	0	0	0	0	0
7	Combination Chemo	1					Clinical trial study record	FALSE				4/6/2004			NCT00003012										1	0	0	0	0	0	0	0	0
7	Combination Chemo	3					Clinical trial study record	FALSE				2/6/2006			NCT00288002										1	0	0	0	0	0	0	0	0
7	Study of Pembrolizu	1					Clinical trial study record	FALSE				10/30/2018			NCT03725059										0	0	0	0	10	0	0	0	0
7	Tamoxifen Citrate or	4					Clinical trial study record	FALSE				1/18/2008			NCT00601900										1	0	0	0	0	0	0	0	0
7	Vaginal Testosteron	1					Clinical trial study record	FALSE				6/11/2008			NCT00698035										1	0	0	0	0	0	0	0	0
6	A Study of Ipatasertl	1					Clinical trial study record	FALSE				11/26/2014			NCT02301988										0	1	0	0	0	0	0	0	0
6	Combination Chemo	2					Clinical trial study record	FALSE				11/1/1999			NCT00003782										0	1	0	0	0	0	0	0	0
6	Ketogenic Diet and E	4					Clinical trial study record	FALSE				5/24/2018			NCT03535701										0	0	0	0	8	0	0	1	0
6	MRI and Mammogra	1					Clinical trial study record	FALSE				3/6/2013			NCT01805076										0	0	0	0	13	0	0	4	0
5	Occupational Therap	2					Clinical trial study record	FALSE				10/30/2018			NCT03723863										0	0	0	0	7	0	0	0	0
5	Radiation Therapy A	1					Clinical trial study record	FALSE				5/19/2003			NCT00005588										0	1	0	0	0	0	0	0	0
5	Role of Individualize	1					Clinical trial study record	FALSE				1/23/2018			NCT03407768										0	0	0	0	9	0	0	0	0
5	Vaccine Therapy in	4					Clinical trial study record	FALSE				6/3/2014			NCT02157051										0	0	0	0	8	0	0	1	0
5	Web-Based Decisio	2					Clinical trial study record	FALSE				12/5/2018			NCT03766009										0	0	0	0	12	0	0	4	0
4	Atezolizumab and C	4					Clinical trial study record	FALSE				6/22/2018			NCT03566485										0	0	0	0	7	0	0	0	0
4	BKM120 and Fulves	4					Clinical trial study record	FALSE				4/20/2011			NCT01339442										1	0	0	0	0	0	0	0	0
3	Acupuncture in Redu	3					Clinical trial study record	FALSE				4/23/2018			NCT03505671										0	0	0	0	8	0	0	0	0
3	Exercise Interventio	2					Clinical trial study record	FALSE				5/14/2018			NCT03523195										0	0	0	0	8	0	0	0	0
3	Fulvestrant and Palb	4					Clinical trial study record	FALSE				12/19/2017			NCT03377101										0	0	0	0	8	0	0	0	0
3	Hypofractionated Ra	1					Clinical trial study record	FALSE				11/17/2017			NCT03345420										0	0	0	0	5	0	0	0	0
3	Intra-Operative Radi	1					Clinical trial study record	FALSE				1/16/2014			NCT02040493										0	0	1	0	9	0	0	0	0
3	Intraoperative Radio	1					Clinical trial study record	FALSE				6/21/2010			NCT01189851										0	0	1	0	0	0	0	0	0
3	LDE225 in Treating	2					Clinical trial study record	FALSE				12/20/2012			NCT01757327										0	0	0	0	0	0	0	0	1
3	Neoadjuvant Letrozo	3					Clinical trial study record	FALSE				1/28/2019			NCT03819010										0	0	0	0	5	0	0	0	0
3	Pilot Study of Cyber	1					Clinical trial study record	FALSE				5/13/2019			NCT03946683										0	0	0	0	4	0	0	0	0
3	Tricribine Phosphat	3					Clinical trial study record	FALSE				9/13/2012			NCT01697293										0	0	0	0	0	0	0	0	2
3	Vaccine Therapy in	4					Clinical trial study record	FALSE				11/14/2008			NCT00791037										0	0	0	0	2	0	0	0	0
3	Weight Bearing Exe	2					Clinical trial study record	FALSE				10/16/2018			NCT03708055										0	0	0	0	5	0	0	0	0
2	Adjuvant Endocrine	1					Clinical trial study record	FALSE				12/3/2018			NCT03761420										0	0	0	0	3	0	0	0	0
2	Cardiac Rehabilitatio	2					Clinical trial study record	FALSE				3/7/2016			NCT03039140										0	0	0	0	2	0	0	0	0
2	Comparing a Single	1					Clinical trial study record	FALSE				9/10/2018			NCT03664687										0	0	0	0	2	0	0	0	0
2	Effect of Low Intens	1					Clinical trial study record	FALSE				10/19/2018			NCT03712813										0	0	0	0	3	0	0	0	0
2	Interactive Gentle Yo	2					Clinical trial study record	FALSE				12/23/2013			NCT02023008										0	0	0	0	8	0	0	4	0
2	Intraoperative Electr	1					Clinical trial study record	FALSE				10/6/2016			NCT02927912										0	0	0	0	2	0	0	0	0
2	Monitoring Plasma T	1					Clinical trial study record	FALSE				4/15/2016			NCT02743910										0	0	0	0	3	0	0	0	0
2	PC2 Treatment Arm	1					Clinical trial study record	FALSE				10/19/2017			NCT03314805										0	0	0	0	5	0	0	0	0
2	POWER-remote We	1					Clinical trial study record	FALSE				6/6/2013			NCT01871116										0	0	0	0	2	0	0	0	0
2	RAD 1802: Pilot Tria	1					Clinical trial study record	FALSE				8/23/2018			NCT03643861										0	0	0	0	3	0	0	0	0
2	Riluzole in Women V	2					Clinical trial study record	FALSE				5/15/2009			NCT00903214										0	0	0	0	1	0	0	1	0
2	Risk and Clinical Be	1					Clinical trial study record	FALSE				12/14/2017			NCT03373708										0	0	0	0	4	0	0	0	0
2	Safety Study of Nivo	3					Clinical trial study record	FALSE				12/3/2014			NCT02309177										0	0	0	0	2	0	0	1	0
2	Sensorimotor Rehab	1					Clinical trial study record	FALSE				6/25/2018			NCT03568526										0	0	0	0	4	0	0	0	0
2	Standard or Compre	1					Clinical trial study record	FALSE				6/3/2013			NCT01872975										0	0	0	0	7	0	0	0	0
2	The Effect of Simva	1					Clinical trial study record	FALSE				3/6/2018			NCT03454529										0	0	0	0	5	0	0	0	0
2	Vaccine Therapy in	3					Clinical trial study record	FALSE				5/16/2011			NCT01355393										0	0	0	0	4	0	0	0	0
2	Yoga or Educational	2					Clinical trial study record	FALSE				10/14/2009			NCT00994279										0	0	0	0	2	0	0	0	0

1	12-week Exercise In	1						Clinical trial study record	FALSE				5/8/2018		NCT03518957										0	0	0	0	1	0	0	0	0
1	A Pilot Study Examir	1						Clinical trial study record	FALSE				7/10/2018		NCT03581552										0	0	0	0	2	0	0	0	0
1	A Study to Investiga	1						Clinical trial study record	FALSE				6/23/2017		NCT03197935										0	0	0	0	1	0	0	0	0
1	Acupuncture in Trea	2						Clinical trial study record	FALSE				1/15/2010		NCT01050075										0	0	0	0	2	0	0	0	0
1	Adjuvant PVX-410 V	2						Clinical trial study record	FALSE				7/5/2016		NCT02826434										0	0	0	0	2	0	0	0	0
1	Akt Inhibitor MK-220	2						Clinical trial study record	FALSE				1/25/2013		NCT01776008										0	0	0	0	1	0	0	0	0
1	Alternative Dosing o	1						Clinical trial study record	FALSE				11/5/2015		NCT02598557										0	0	0	0	1	0	0	0	0
1	Atorvastatin in Treat	3						Clinical trial study record	FALSE				3/13/2019		NCT03872388										0	0	0	0	1	0	0	0	0
1	Biomarkers in Wom	3						Clinical trial study record	FALSE				4/23/2008		NCT00665457										0	0	0	0	2	0	0	0	0
1	CAELYX® as Adjuv	1						Clinical trial study record	FALSE				10/19/2018		NCT03712956										0	0	0	0	2	0	0	0	0
1	Caloric Restriction in	1						Clinical trial study record	FALSE				3/27/2013		NCT01819233										0	0	0	0	1	0	0	0	0
1	Carboplatin and Cor	2						Clinical trial study record	FALSE				3/21/2013		NCT01818063										0	0	0	0	1	0	0	0	0
1	Carboplatin and Pac	2						Clinical trial study record	FALSE				8/29/2016		NCT02883062										0	0	0	0	1	0	0	0	0
1	Carboplatin With or I	4						Clinical trial study record	FALSE				6/26/2017		NCT03206203										0	0	0	0	1	0	0	0	0
1	Cisplatin With or Wit	4						Clinical trial study record	FALSE				11/3/2015		NCT02595905										0	0	0	0	0	0	0	1	0
1	Clinical Information	4						Clinical trial study record	FALSE				11/9/2018		NCT03737695										0	0	0	0	1	0	0	0	0
1	Direct Tumor Microi	4						Clinical trial study record	FALSE				2/14/2018		NCT03432741										0	0	0	0	1	0	0	0	0
1	Efficacy of Letrozole	2						Clinical trial study record	FALSE				3/27/2015		NCT02400567										0	0	0	0	1	0	0	0	0
1	Endocrine Immunoti	1						Clinical trial study record	FALSE				10/25/2018		NCT03719495										0	0	0	0	2	0	0	0	0
1	Exercise Testing for	1						Clinical trial study record	FALSE				7/14/2016		NCT02836093										0	0	0	0	2	0	0	0	0
1	FES (18-(18F)-Fluo	1						Clinical trial study record	FALSE				11/1/2018		NCT03726931										0	0	0	0	1	0	0	0	0
1	Impact of 5-fraction	1						Clinical trial study record	FALSE				6/25/2018		NCT03568981										0	0	0	0	2	0	0	0	0
1	Intraoperative Radia	1	ClinicalTrials.gov					Clinical trial study record	FALSE				10/10/2014		NCT02266602										0	0	0	0	1	0	0	0	0
1	Intravenous Ascorbi	1						Clinical trial study record	FALSE				7/27/2015		NCT02521077										0	0	0	0	1	0	0	0	0
1	IORT Following Brea	1						Clinical trial study record	FALSE				5/25/2018		NCT03536897										0	0	0	0	2	0	0	0	0
1	KeraStat Skin Thera	1						Clinical trial study record	FALSE				6/20/2014		NCT02168179										0	0	0	0	1	0	0	0	0
1	Metformin and Orne	1						Clinical trial study record	FALSE				10/28/2014		NCT02278965										0	0	0	0	1	0	0	0	0
1	Metformin Hydrochl	2						Clinical trial study record	FALSE				9/24/2009		NCT00984490										0	0	0	0	1	0	0	0	0
1	Minocycline Hydrochl	2						Clinical trial study record	FALSE				7/28/2014		NCT02203552										0	0	0	0	2	0	0	0	0
1	Neo-adjuvant Therap	2						Clinical trial study record	FALSE				7/14/2011		NCT01394211										0	0	0	0	2	0	0	0	0
1	Neoadjuvant Study o	2						Clinical trial study record	FALSE				9/20/2016		NCT02907918										0	0	0	0	1	0	0	0	0
1	Oral CDB-4124 vs. E	2						Clinical trial study record	FALSE				2/24/2013		NCT01800422										0	0	0	0	2	0	0	2	0
1	Paclitaxel With Trast	1						Clinical trial study record	FALSE				4/9/2013		NCT01827163										0	0	0	0	1	0	0	0	0
1	Pancreatic Nutritiona	2						Clinical trial study record	FALSE				4/29/2015		NCT02432950										0	0	0	0	3	0	0	0	0
1	Psychosexual Interv	2						Clinical trial study record	FALSE				1/10/2013		NCT01764802										0	0	0	0	1	0	0	0	0
1	Radiation Therapy in	1						Clinical trial study record	FALSE				12/25/2008		NCT00814567										0	0	0	0	1	0	0	0	0
1	S1105: Test-Messag	1	ClinicalTrials.gov					Clinical trial study record	FALSE				1/21/2012		NCT01515800										0	0	0	0	2	0	0	0	0
1	S1202: Duloxetine fo	1						Clinical trial study record	FALSE				5/15/2012		NCT01598298										0	0	0	0	1	0	0	0	0
1	S1222 Trial (Everoli	4						Clinical trial study record	FALSE				5/14/2014		NCT02137837										0	0	0	0	2	0	0	0	0
1	Stress Test in Datec	2						Clinical trial study record	FALSE				4/23/2018		NCT03505736										0	0	0	0	3	0	0	0	0
1	Study of Adagloxad	1						Clinical trial study record	FALSE				6/19/2018		NCT03562637										0	0	0	0	1	0	0	0	0
1	Tissue Procurement	1						Clinical trial study record	FALSE				5/26/2015		NCT02455882										0	0	0	0	3	0	0	0	0
1	Trial to Compare the	3						Clinical trial study record	FALSE				5/9/2016		NCT02768714										0	0	0	0	2	0	0	0	0
1	Typhoid Vaccine in	2						Clinical trial study record	FALSE				4/9/2015		NCT02415387										0	0	0	0	1	0	0	0	0

CSCI 636 Final Project: Altmetric Score and Correlation to Severity(Stage) of Breast Cancer

By: Alexander Al-Taraireh, Brady Balk, and Ben Ford

Overview

- Introduction
- Motivation
- Objective
  - Initial hypothesis
  - Revised hypothesis
- Related Work
- Data
- Implementation/Methods
- Results
- Future Work
- Conclusion

Introduction

- Breast cancer is currently the most common cancer globally, accounting for 12.5% of all new annual cancer cases worldwide. About 13% (about 1 in 8) of U.S. women are going to develop invasive breast cancer in the course of their life. About 43,250 women in the U.S. are expected to die in 2022 from breast cancer.
    - [*Breast Cancer Facts and Statistics* . Breast cancer facts & statistics 2022. (n.d.). Retrieved December 4, 2022, from https://www.breastcancer.org/facts-statistics]

Motivation: Why This Topic?

- Cancer is a disease that affects so many people around the world, it is hard not to find a person who hasn't experienced it one way or another. So, we wanted to see if it was possible to predict the severity of a specific type of cancer via how many publications its research methods get.

Initial hypothesis

- Our initial hypothesis was that the higher the Altmetric score, the higher the severity of the breast cancer would be.
  - This is because we believed that the more a piece of research is published and shared, it would be because it is relating to a high severity of breast cancer.

Revised Hypothesis

- After cleaning and sorting the data, we found that in fact the opposite turned out to be true. That is the higher the Altmetric score, the lower the severity of the breast cancer.
  - This is believed to be because while in the early stages, it is more treatable/curable thus leading to more publications about potential life-saving treatments.

Related Work

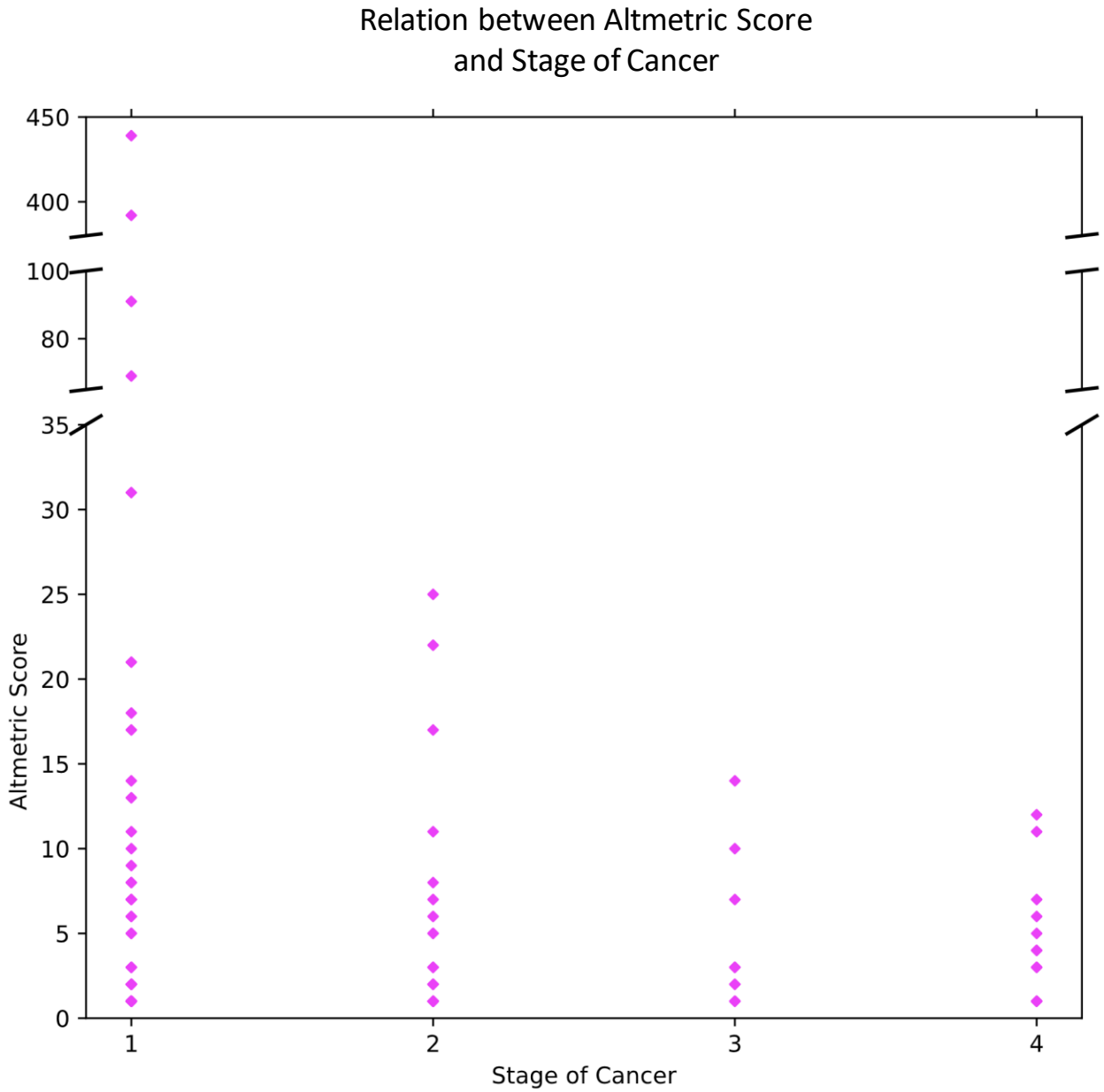
- In the hopes of shortening the time it takes to identify and qualify scientific articles new tools have been used for both data collection and quantifying their usage
  - The two most stand out tools are Google Trends and Altmetrics
  - Unfortunately, Google Trends utility was in gather data not in verifying the validity of scientific articles
- Alternatively, Altmetrics are sourced widely across internet platforms
  - A study wanted to determine if altmetrics could predict future citation counts in critical care medicine publications
  - Another study wanted to test if highly cited cancer articles would correlate positively with altmetric scores
- It was concluded that both Altmetrics and Google Trends were both good sources of data, but could not expedite the process of gathering citations for research papers

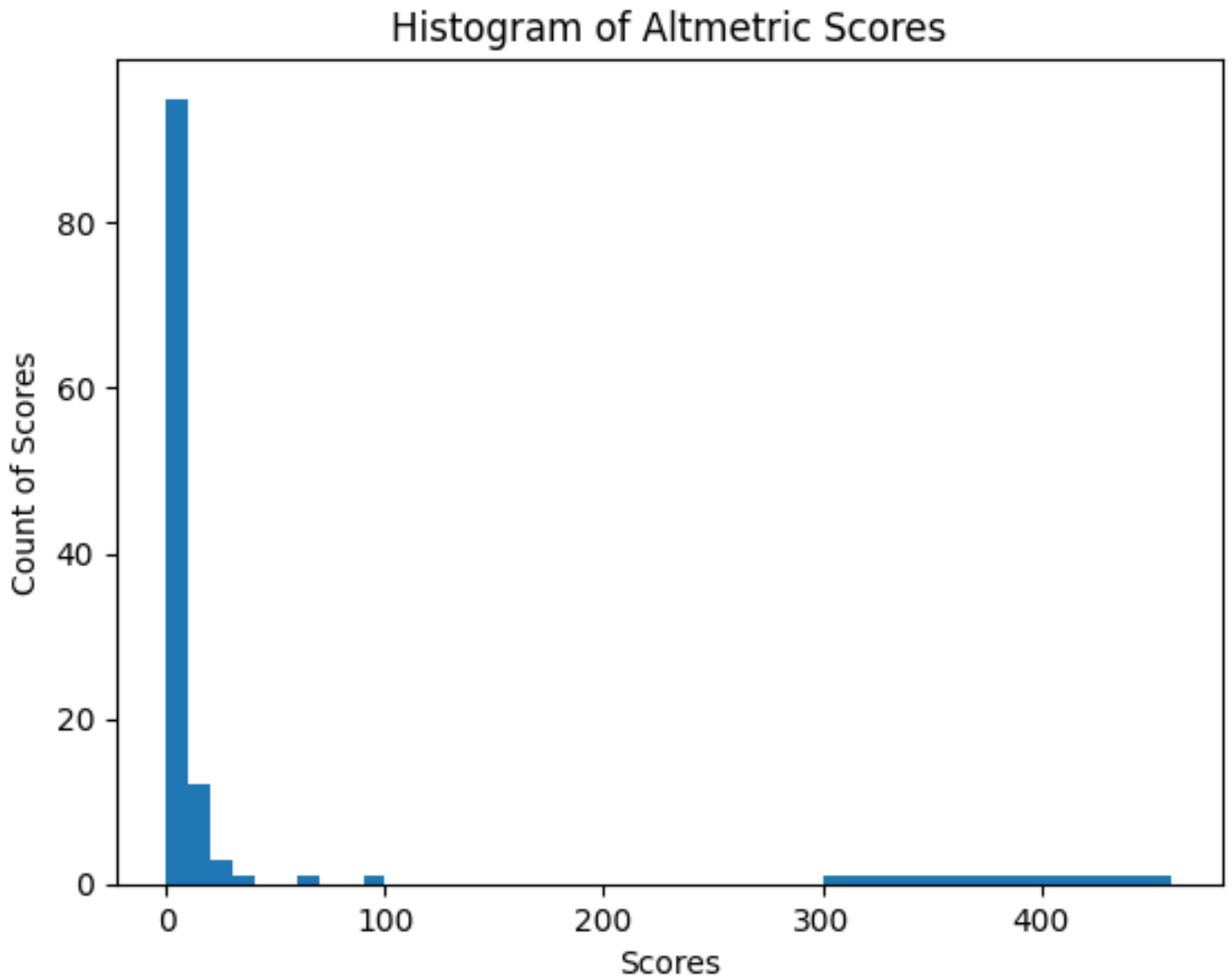
Data

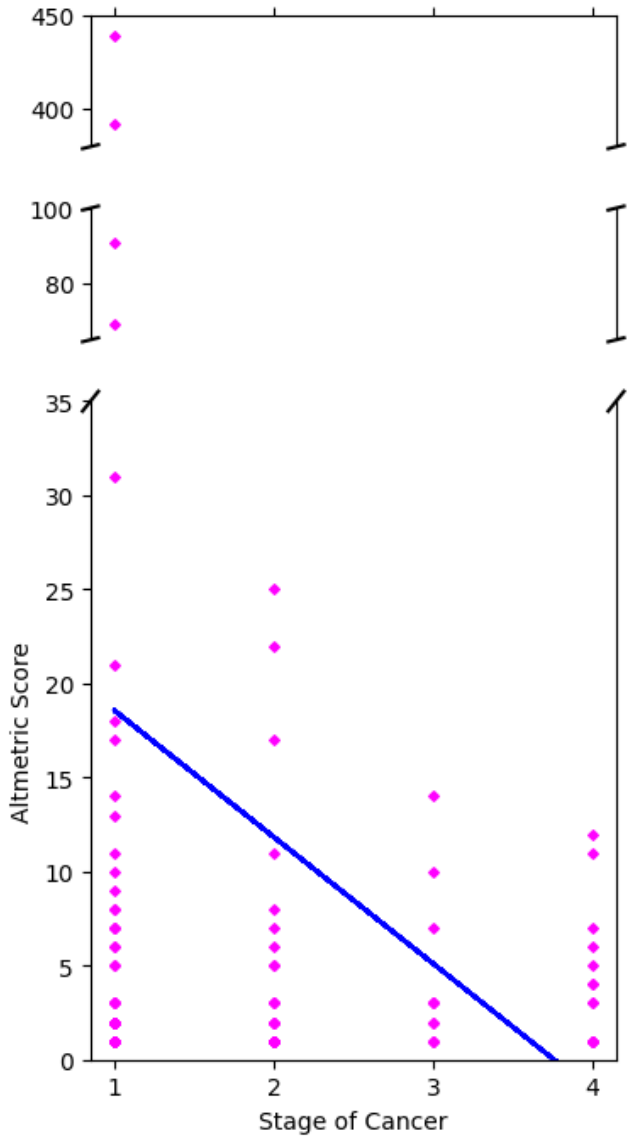
- Altmetric Clinical Trial Dataset
  - Sourced from: https://www.dropbox.com/s/8jjzdcbtgyx9lht/Dimensions-Clinical-Trial-2019-11-01_00-39-34.xlsx?dl=0
  - 50,330 unique clinical trials
  - 45 features
  - Used to find breast cancer trials
    - Then used the score and stage of cancer as defining features

Implementation/Methods

- Dataset: Altmetric score of research publications for clinical trials
  - Cleaned the data (preprocessing) to show only breast cancer publications and within that, ones that showed a specific stage in the title and valid score (features)
- Python tools: NumPy, Matplotlib, and Scikit-learn
  - NumPy to import data and to calculate some stats
  - Matplotlib to visually show our results
  - SciPy to use a linear regression model, power regression model, and stats calculations


Relation between Altmetric Score
and Stage of Cancer
Altmetric Score
Stage of Cancer
450
400
100
80
35
30
25
20
15
10
5
0
1
2
3
4


Histogram of Altmetric Scores
Count of Scores
0
20
40
60
80
0
100
200
300
400
Scores



As shown by the graph, the linear regression model was not a good fit for our data. It led to very low scores of accuracy, thus concluding that our data was nonlinear and would need a different type of model.

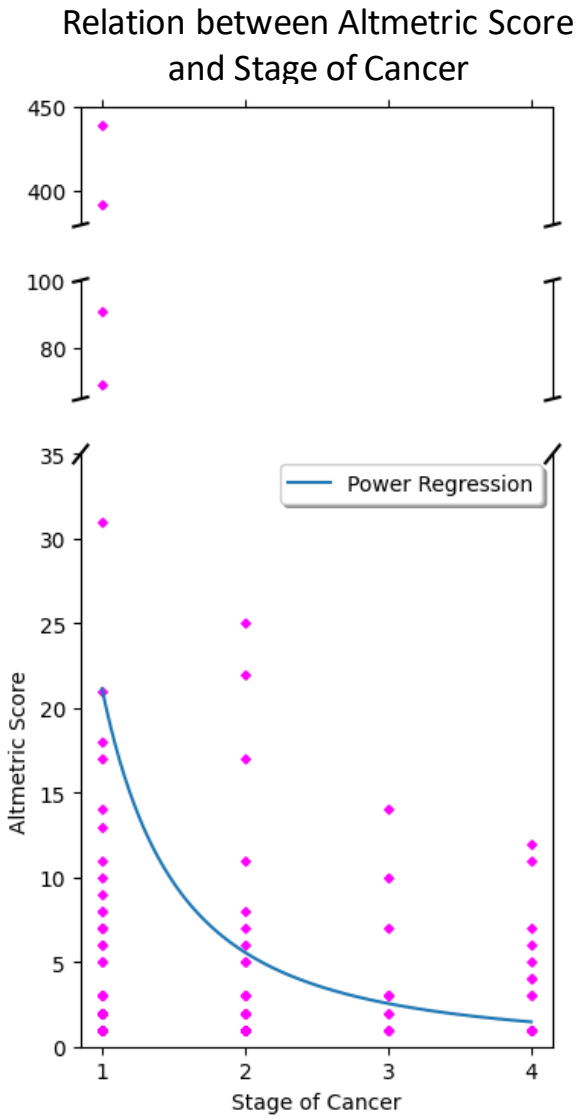
Linear regression statistics:
- Slope: -6.723917672107885
- Y - int: 25.28715400993614
- R value: -0.12657699565901473
- R^2: 0.016021735830062233
- P value: 0.17766251635443928
- Std error: 4.957025291233932

Conclusion: R^2 Value is very low, thus linear regression is not a great choice

Graph Details:
- Created using matplotlib.
- Power Regression via SciPy
- Graph Pattern:
Concave Up, Decreasing

Data Details:
- Altmetric Scores:
115 Values
24 Unique Values
1- 439
1(44), 2(18), 3(12). 4(2), 5(5), 6(4), 7(6), 8(3), 9, 10(2), 11(3), 12, 13, 14(2), 17(2), 18, 21, 22, 25, 31, 69, 91, 392, 439.
- Largest Outlier: 439 | 1

- Stage of Cancer:
4 Values
1-4



Optimal values (min. sum of squares)
- Altmetric Score: 21.134262589266086
- Stage of Cancer: 1.9247412788518965

Variance in Y - Values (Altmetric Scores): 50.584731855121774

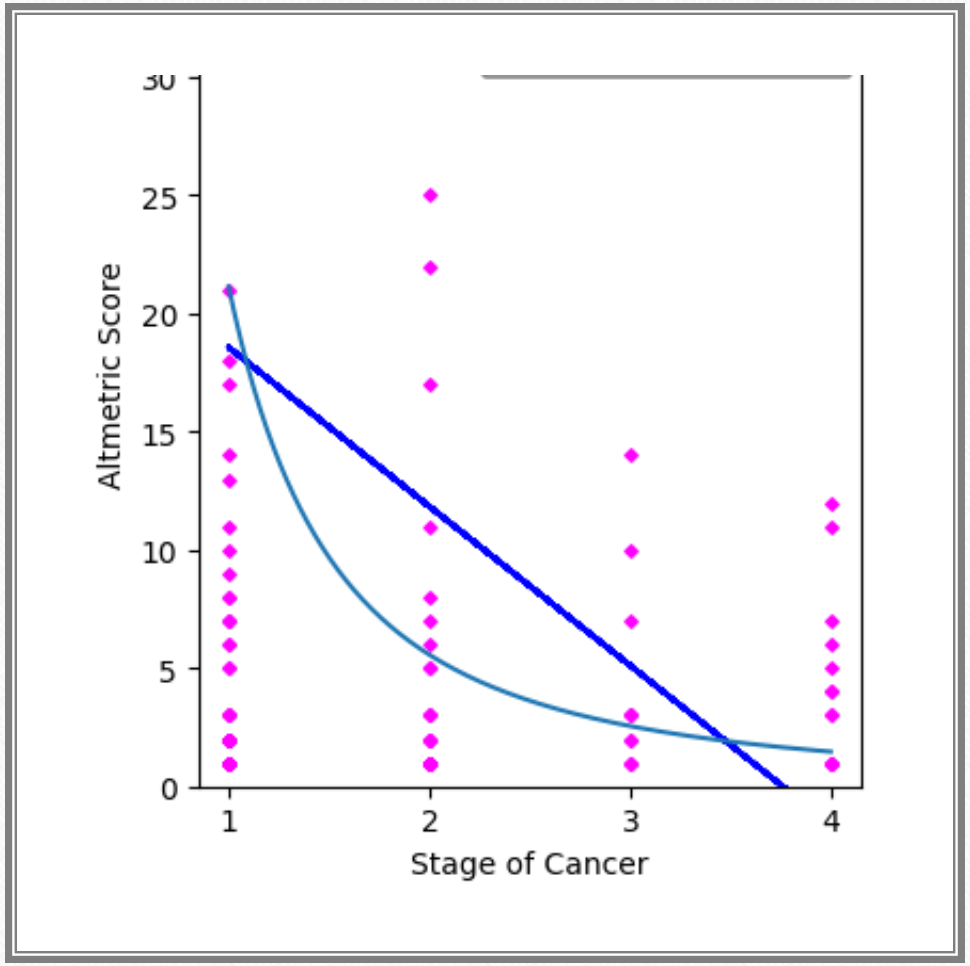
Covariance between Almetric Score and Stage of Cancer: 3.106389970302665

Covariance score is postive, larger than R^2, thus indicating power regression is a good fit

Covariance Matrix Row 1
[50.58473186 3.10638997]

Covariance Matrix Row 2
[3.10638997 5.12180728]

Standard deviation (Sigma) of Altmetric Scores = 7.112294415666563



Comparison of Two Models

- Largest outliers are omitted in this graph
- Highlighting the under fitting of linear regression and the weighted fitting of power regression
- Power regression line is pulled down because of volume of stage 1 scores and how high they are compared to the rest and then evens out as the higher stages have lower scores and less data points.

Explanation

- Linear regression model proved to be a poor fit for the data
  - Well underfit and low accuracy with an R^2 value of ~0.016
- Power regression model proved to be a good fit for the data
  - Sheer number of stage one trials with higher scores pulled the curve down (thus power regression) and it flattened out as the stage increased and there were fewer scores with high numbers.
  - Covariance of ~3.1 proves it is a better fit as this indicates a stronger relationship than a linear regression model can provide.

Future Work

- A continuation of our revised hypothesis will require more extensive testing with new data
- If our model is proven to be accurate, then the next step will be creating an algorithm that will automatically compile and grade papers based off their altmetric score
- The data the algorithm can be set to produce can be used as both a data set and a database for researchers looking for up to date information on breast cancer

Conclusion

- With our findings, we have reason to believe there are more papers covering breast cancer of a lower severity because of the treatability in the beginning stages of breast cancer. It has been proven that it is extremely valuable to spread awareness and encourage people to seek out a screening for cancer. Catching a tumor before it has metastasized has effectively improved the survival rates of breast cancer. This paired with the advancements of treatment options for cancer neutralizing the disease before it progresses to the later stages is more effective for preserving human life than focusing on advancing research and treatment options for the later stages of breast cancer.

Links:

Lehane, Daniel J, and Colin S Black. “Qualitative Research - Journals.sagepub.com.” *Sage Journals*, 2 Feb. 2020, https://journals.sagepub.com/doi/full/10.1177/14687941221132956.

Bayar, Bahattin, and Rıfat Peksöz. “Analysis of Trending Topics in Breast Cancer Articles from an Altmetric Perspective.” *Cureus*, 5 July 2022, https://www.cureus.com/articles/101281-analysis-of-trending-topics-in-breast-cancer-articles-from-an-altmetric-perspective.

Bcrf. “BCRF Announces 2021-22 Investment.” *Breast Cancer Research Foundation*, BCRF, 11 Feb. 2022, https://www.bcrf.org/blog/bcrf-2021-2022-research-investment/.

Bornmann, Lutz, and Robin Haunschild. “Do Altmetrics Correlate with the Quality of Papers? A Large-Scale Empirical Study Based on F1000prime Data.” *PloS One*, U.S. National Library of Medicine, 23 May 2018, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5965816/.

“Breast Cancer.” *World Health Organization*, World Health Organization, 26 Mar. 2021, https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

Celik, Emir, et al. “Comparison of Attention for Cancer Research on Social Media versus ...” *Research Gate*, Jan. 2020, https://www.researchgate.net/profile/Betul-Borku-Uysal/publication/340264974_Comparison_of_Attention_for_Cancer_Research_on_Social_Media_Versus_Academia_An_Altmetric_Score_Analysis/links/61ebb8c88d338833e3877961/Comparison-of-Attention-for-Cancer-Research-on-Social-Media-Versus-Academia-An-Altmetric-Score-Analysis.pdf.

Chou, Peter, et al. “Correlation between Cancer Research Trends and the Importance of Cancers Based on Mortality and Diagnosis Rates: An Analysis of Altmetric Data.” *STEM Fellowship Journal*, 23 Nov. 2017, https://journal.stemfellowship.org/doi/10.17975/sfj-2017-014.

Goksoy, Beslen, and Hilmi Bozkurt. “Social Attention of the Top 50 Scientific Articles on Gastric Cancer ...” *Research Gate*, Dec. 2020, https://www.researchgate.net/profile/Beslen-Goksoy/publication/346653939_Social_attention_of_the_top_50_scientific_articles_on_gastric_cancer_Bibliometric_and_altmetric_analysis/links/600cb23a45851553a0678de9/Social-attention-of-the-top-50-scientific-articles-on-gastric-cancer-Bibliometric-and-altmetric-analysis.pdf.

Hassona, Yazan, et al. “The Online Attention to Oral Cancer Research: An ... - Wiley Online Library.” *Wiley Online Library*, 29 Apr. 2019, https://onlinelibrary.wiley.com/doi/abs/10.1111/odi.13111.

Johnson, Bradley S., et al. “Using Google Trends and Twitter for Prostate Cancer Awareness: A Comparative Analysis of Prostate Cancer Awareness Month and Breast Cancer Awareness Month.” *Cureus*, 13 Feb. 2021, https://www.cureus.com/articles/47045-using-google-trends-and-twitter-for-prostate-cancer-awareness-a-comparative-analysis-of-prostate-cancer-awareness-month-and-breast-cancer-awareness-month.

Loach, Tamar, and Jonathan Adams. “Altmetric ‘Mentions’ and the Identification of Research Impact - NESTA.” *Nesta*, Feb. 2015, https://media.nesta.org.uk/documents/nesta_altmetrics_working_paper.pdf.

Mohamad, Mazlyfarina, and Hui Sin Kok. "Using Google Trends Data to Study Public Interest in Breast Cancer Screening in Malaysia." *Asian Pacific Journal of Cancer Prevention : APJCP*, U.S. National Library of Medicine, 25 May 2019, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6857874/.

Munnolli, Satish S, and Shamprasad M Pujar. "Scientometric Study of Indian Cancer Research Based on Scopus Database." *Taylor & Francis*, 20 Oct. 2017, https://www.tandfonline.com/doi/abs/10.1080/09737766.2017.1284729.

Pujar, Shamprasad. "The Impact of Indian Cancer Research in Social Media: A Study Using ..." *Research Gate*, Nov. 2015, https://www.researchgate.net/profile/Satish_Munnolli/publication/286921576_The_Impact_of_Indian_Cancer_Research_in_Social_Media_A_Study_using_Altmetric_explorer/links/566fb4ab08ae4d9a42572aa2/The-Impact-of-Indian-Cancer-Research-in-Social-Media-A-Study-using-Altmetric-explorer.pdf.

Rooney, Michael Kevin, et al. "Article Tools." *Journal of Clinical Oncology*, 2 June 2022, https://ascopubs.org/doi/abs/10.1200/JCO.2022.40.16_suppl.11014?role=tab.

Schootman, M, et al. "The Utility of Google Trends Data to Examine Interest in Cancer Screening." *BMJ Open*, British Medical Journal Publishing Group, 1 June 2015, https://bmjopen.bmj.com/content/5/6/e006678.

Taylor, Mike. "City University of New York." *Research Trends*, 6 Jan. 2013, https://academicworks.cuny.edu/cgi/viewcontent.cgi?article=1229&context=le_pubs.

Trasta, Anthi. “Where Does Public Funding for Cancer Research Go: Allocation of Research Funding for Cancer and COPD Is Not Always Proportional to Disease Burden.” *EMBO Reports*, U.S. National Library of Medicine, 19 Feb. 2018, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5836059/.

Vasconcellos-Silva, Paulo Roberto, et al. “Using Google Trends Data to Study Public Interest in Breast Cancer Screening in Brazil: Why Not A Pink February?” *JMIR Public Health and Surveillance*, U.S. National Library of Medicine, 6 Apr. 2017, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5399222/.

“What Are Altmetrics?” *Altmetric*, 17 Dec. 2018, https://www.altmetric.com/about-altmetrics/what-are-altmetrics/.

“What Is Breast Cancer?” *American Cancer Society*, 19 Nov. 2021, https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html.

“What Is Cancer?” *National Cancer Institute*, 5 May 2021, https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

Çelik, Emir, et al. “Comparison of Attention for Cancer Research on Social Media versus Academia: An Altmetric Score Analysis.” *DSpace@Biruni*, Akademi Doktorlar Yayınevi, 1 Jan. 1970, http://openaccess.biruni.edu.tr/xmlui/handle/20.500.12445/1255.

Breast Cancer Facts and Statistics . Breast cancer facts & statistics 2022. (n.d.). Retrieved December 4, 2022, from https://www.breastcancer.org/facts-statistics

Resources:

- Jupyter Notebook

- Python 3
    - Matplotlib
    - SciPy
    - NumPy
- Microsoft Office Suite (To create documents, presentation slides, and record slides)
- ILovePDF (To merge all deliverables aside from presentation movie into a single PDF)
- iMovie (To combine, edit, and export presentation movie)