Machine Learning with p53 Gene Somatic Mutations

Social/Environmental Injustice and the Correlation to Cancer Diagnoses and Causes

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

RUPAL SAINI

November 1, 2021

Technical Team Members:

N/A

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Dr. Sean Ferguson, Department of Engineering and Society

Dr. Nada Basit, Department of Computer Science

Introduction

When looking at cancer globally, it is the second leading cause of death. When it comes to low and middle income countries, it accounts for more than half of all deaths. Why are these areas with minorities suffering more than those that are more privileged? There is a system in place that makes groups that are underprivileged and underrepresented invisible to the standards of medical care being given to more privileged areas. As this issue continues to grow, researchers are looking for ways to put medical attention back on the groups who need it.

The goal of the technical Capstone project is to utilize machine learning clustering algorithms to classify groups of cells as cancerous or benign. The technical research project walks through this process with a database of p53 gene somatic mutations in human tumors and cell lines. This was done entirely individually. The motivation behind this research is to learn more about machine learning algorithms and create a foundational basis as the world progresses towards a primarily ML environment. This research is important because it can be used to determine if a person has cancer with a high accuracy without performing extremely invasive procedures. Further, without invasive procedures, this method of classification would allow those of lower socioeconomic groups to save money on medical fees and save time that would have gone to recovery.

The STS topic being explored is loosely related to the technical research project that was already completed. It discusses cancer and how it comes about in unfair manners when looking at different types of people. It further evaluates how human created environmental toxins overwhelmingly cause cancer in minority groups. The connection here is the discussion of cancer in both projects. This topic is important in that it starts the conversation on how to make minority groups visible when it comes to illnesses and how the illnesses are caused. It also allows for conversation around reform in the standards of environmental toxin disposal.

Technical Topic

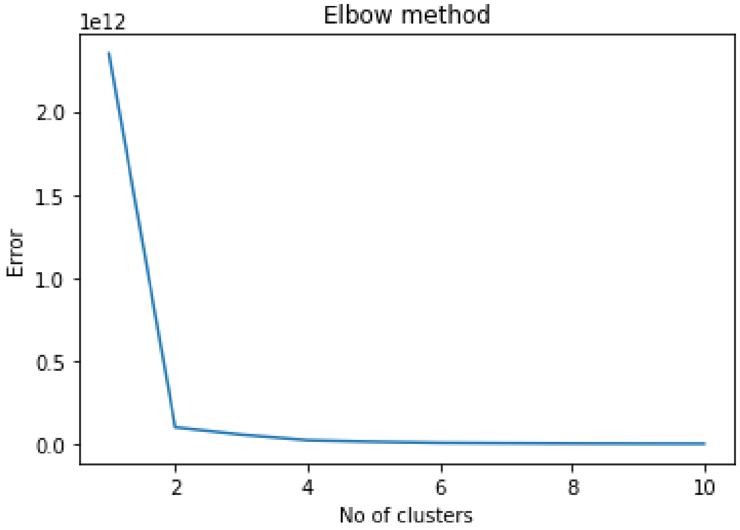
Clustering is a type of unsupervised learning that does not utilize the ground truth. It is a way to group unlabeled samples. Similarity is measured in the data and then clusters are formed where the data points in each collection have some like features that caused the algorithm to place them together. Using clustering, the ground truth of a dataset can be derived. This ground truth can in turn be used to make predictions based on the similar features. The following research walks through this process with a database of p53 gene somatic mutations in human tumors and cell lines. The motivation behind this research is to learn more about machine learning algorithms and create a foundational basis as the world progresses towards a primarily ML environment.

The first step of the technical research process was to determine which dataset to use for this research project. I was interested in bioinformatics, so it was ideal for me to use a dataset with some biological relationship. This ended up being a dataset on p53 cancer cells. Initially I had decided to use the p53 Mutants Data Set in the UCI Machine Learning Repository (Lathrop, 2010), however this dataset did not have descriptive attribute information which made it difficult to understand. I then moved to a Database of p53 gene somatic mutations in human tumors and cell lines (Hainut et al., 1997). This dataset, on the other hand, did not have a ground truth attribute which would describe whether or not the tumor was cancerous. This is where the idea of doing clustering instead of directly going to a machine learning algorithm came about.

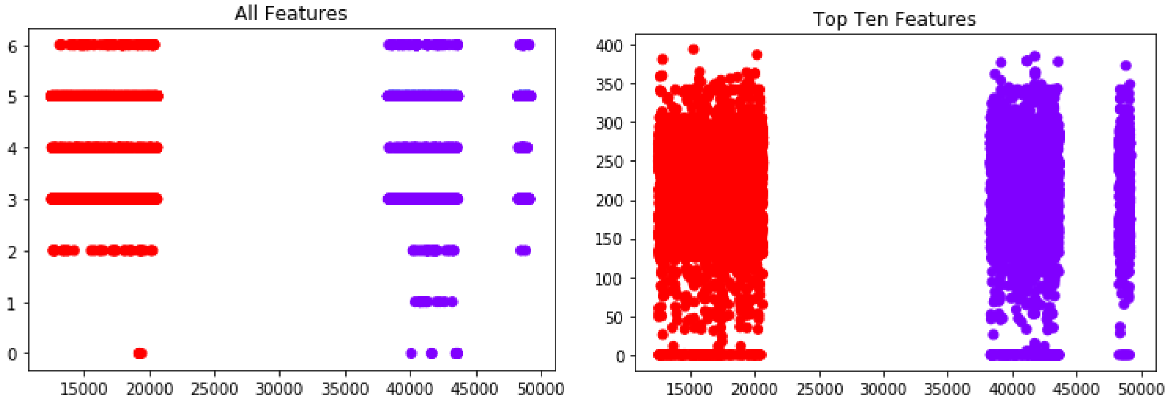
The most difficult part of this research project was formatting the data in a way that could be easily read by the feature selection and machine learning algorithm, also known as preprocessing. To format the data, a label encoder was utilized ("sklearn.preproccessing", n.d.). This normalized the labels and transformed non-numerical labels to numerical labels so that the algorithm could process the data. This was done to every attribute with non-numerical data.

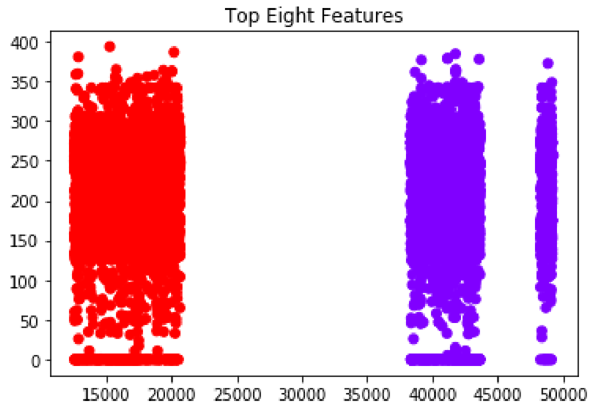
Next, feature selection (Brownlee, 2014) was performed. The method chosen was the Chi-Squared statistical test. This test belongs to a class of filter methods, a process to assign a score to each feature and then rank the features by that particular score (Paul, 2020). The Chi-Squared test was run on all of the attributes and the scores were printed out. The higher the score, the more important the attribute ("sklearn.feature_selection", n.d.). The following is a list of the features listed out from most important to least important: Mutation_ID (column 0), Morphology (column 16), Sub-topography (column 15), Topography (column 14), Topo_code (column 17), Putative stop (column 12), Mutant_codon (column 5), Codon (column 3), WT_codon (column 4), Description (column 6), WT _AA (column 8), Source (column 13), Mutant_AA (column 9), CpG (column 7), Frameshift (column 11), Splice (column 10), Location (column 2), Type (column 1).

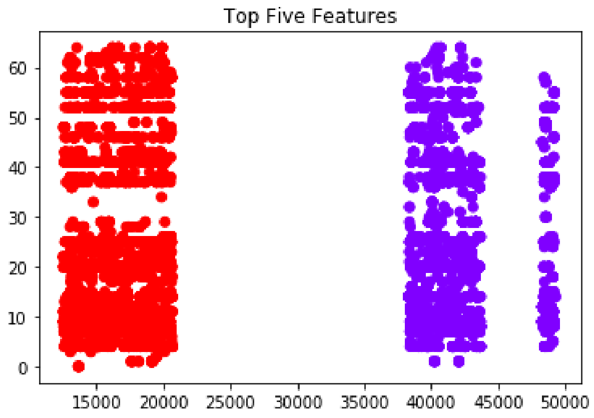
After feature selection, a clustering algorithm was researched. I decided to perform k-means clustering. To begin, the optimal number of clusters for the dataset needed to be determined. I knew that I wanted the dataset to be split into two clusters to represent cancerous and benign, however, I wasn't sure if that would be efficient. To get around this issue, the elbow method was performed (Dhiraj, 2019). This would plot a graph between the number of clusters and the error value corresponding to that. Wherever the shape of the elbow was formed, was determined to be the optimal number of clusters for the dataset. The elbow graph is shown below:



It is evident that the number of clusters the dataset should be split into, is two. This is in accordance with the prediction that I had made. From here, k-means clustering was carried out using a few lines of python code. This was done in a few ways - with all of the features, the top ten features, the top eight features, and the top five features. As stated earlier, the top features were determined through feature selection and the Chi-Square test. The following are the cluster graphs for each k-means algorithm carried out:



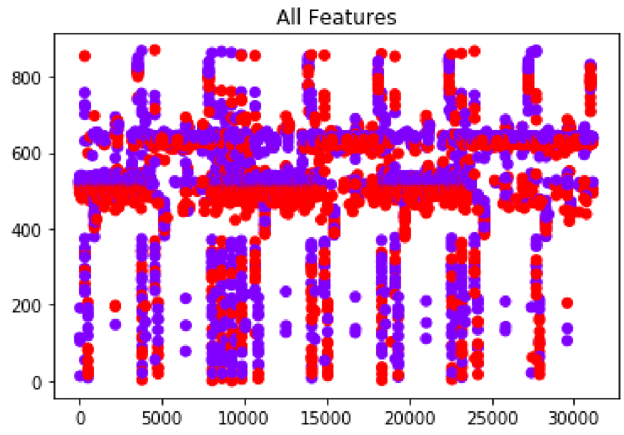




The last portion of this research assignment was to perform the Random Forest machine learning algorithm using the data from the clustering as the ground truth (Navlani, 2018). What I did is add another column to the data representing which cluster each sample belonged to. The ground truth in this case was taken from the k-means clustering results that used all features. The data was split into training and testing sets. 70% of the data was allotted for training and the remaining 30% was separated for testing. A gaussian classifier was made and the model was trained using the ground truth data that was already given. Then, predictions were made for the ground truth for the testing data, using the trained model. After this, the accuracy of the model was calculated. It came out to be a 1.0 accuracy, meaning the trained model perfectly predicted which cluster the testing data belonged to.

An issue that arose was the difficulty of understanding whether or not the Random Forest algorithm was working correctly. It was assumed that it wasn't because usually datasets do not give an accuracy as high as was found for the Database of p53 gene somatic mutations in human tumors and cell lines. To further determine if this was the case, Random Forest was performed a couple more times on other datasets. The first of which being the original p53 Mutants Data Set in the UCI Machine Learning Repository, as this included the ground truth already. Next, Random Forest was performed on the same p53 Mutants Data Set with clustering results as the ground truth. Finally, it was performed on the very well known Iris dataset (Fisher, 1988) to determine if the results of the previous findings were significant.

The process of performing the Random Forest algorithm on the p53 Mutants Data Set in the UCI Machine Learning Repository was very similar to what was done so far. First, feature selection was performed with the Chi Square test, then clustering, and finally the algorithm was used to make predictions. The top ten features of the UCI dataset turned out to be columns 2522, 2972, 3580, 4477, 4616, 4105, 4476, 2548, 4819, and 4098. As stated earlier, this dataset did not have in-depth attribute information, so the columns are being referred to by numbers instead of names. After this, clustering was performed to see where the data points would lie if there was no information on the ground truth given. The graph is below:



It can be seen that the clusters are somewhat overlapping. The purple points belong to one cluster, while the red points belong to another. It is not clear which cluster means that the p53 mutants are active or inactive. The results from the clustering were later used as a ground truth for the Random Forest algorithm. It was predicted that the accuracy of the algorithm would decrease due to the overlapping points. However, after using clustering as the ground truth, when 70% of the samples were trained and the other 30% were tested, the Random Forest algorithm still gave 98% accuracy in its prediction. When the given ground truth in the p53 Mutants Data Set was used, the accuracy of the prediction under the same constraints (70% trained and 30% were tested) was 99%.

Because these accuracies were so high, I needed to perform Random Forest on a known dataset with a known accuracy to determine whether the algorithm was working correctly. I decided to use the Iris dataset. With the Random Forest algorithm, it is known that this dataset has a 93% accuracy. After running Random Forest on my machine with the dataset using 70% of the samples for training and 30% for testing, the accuracy was indeed 93%.

This led me to believe that the Random Forest algorithm does work. It can also be concluded that the data in the Database of p53 gene somatic mutations in human tumors and cell lines is extremely well separated. This is the same idea for the p53 Mutants Data Set in the UCI Machine Learning Repository. Meaning, the algorithm itself is not faulty, it is simply the nature of the datasets which allowed the algorithm to make predictions at such a high accuracy. Therefore, it is important to note that all previous findings stated in this paper are valid and correct.

Throughout this research, I have learned a large amount of information on machine learning. I came into this project with very limited knowledge on what machine learning is, how it works, and why it is performed. Now I can answer these three questions and more with a very in-depth approach. By performing feature selection, clustering, and the Random Forest algorithm in my own environment, my understanding was furthered in a very significant way. Prior to this project, I had not even realized that many machine learning algorithms were already written and it is simply a few lines of code that need to be written to actually apply the algorithm to a dataset. Moreover, it is now evident that the hardest part of machine learning is modifying the dataset in a way that it is uniform. This in itself took me a couple of days to carry out. I am confident that I can use machine learning in the future whether it is in a job or a school project. I am very grateful to have had the opportunity to pursue this and learn more about a very interesting field of computer science with the mentorship of Professor Basit.

Other uses that may come of this research is using the same model to perform random forest on the different types of k-means clustering. Meaning, using feature selection and the ground truth that was found from k-means clustering of 10, 8, and 5 features, how does the accuracy of the model differ? I am also interested in using different machine learning algorithms such as logistic regression, association rule mining, and more to determine which is a more accurate predictor. Further, I can see what difference k-fold cross validation makes on the model.

STS Topic

Social injustice can come about in many different ways. In a world so greatly emphasizing equality, it is difficult not to see ways that the system put in place ostracizes different types of people, while empowering other ones.

There are some groups that are more greatly affected by harmful environmental variables that may lead to higher rates of cancer. **This contributes to the thesis of what groups of people take on the burdens of toxins leading to cancer and how certain groups are made invisible when it comes to medical diagnoses.**

The framework being looked at through this research project is that of risks and standards. This framework helps us understand how standards and risks benefit those of more powerful social groups, or the upper class, as compared to the underprivileged by treating the goals of the underprivileged as insignificant. In this case, there is an invisibility of risks because the standards to test for certain groups are nonexistent. This is agnotology, an intentional creation of ignorance.

There are health care disparities when it comes to people of color and lower-income families compared to others. Individuals that belong to these groups have a high risk of being uninsured and therefore unable to afford cancer treatment. With this lack of affordability comes poorer quality of care. For example, Black adults are more likely to have negative healthcare experiences than White adults. All of this combined plays into the fact that people of color have higher rates of illnesses than Whie people (Ndugga, 2021).

In particular, with respect to cancer in the early 2000's, African American women were found to be three times more likely than Whites to present advanced stages of Breast cancer. Further, Blacks with Colorectal cancer were more than 25% less likely to undergo major procedures as compared to their White counterparts. Another study showed that with lung cancer, almost 50 excess Black deaths (as compared to Whites) could be attributed to differences in surgery rates. It was concluded that in many of these cases, there was an absence of physician recommendation for surgery (Geiger, n.d.). Although these figures are somewhat outdated, the trend continues in today's time. This leads to a discussion of agnotology. When it comes to testing and diagnoses, physicians are intentionally creating ignorance. There are no standards that physicians are being held to, meaning people of color are suffering due to lack of testing and a decrease in quality care. This creates an invisibility of risks. For example, people of color are being diagnosed later on in the course of their illness, the risks are invisible up until that point due to the lack of standards addressing racial and ethnic disparities.

How these illnesses are caused is another problem. One example of this is in Canada's Sydney Tar Ponds, the home place of the indigineous Mi''kmaw people and a waste site for coal, tar, and cancer causing PCBs. Because of this, the cancer rate of the Mi''kmaw people was 45% higher than the average in Nova Scotia (Waldron, 2018). There was a lack of standards, allowing corporations to dump in a region with no pollution control. This greatly comes about because the indigineous people did not have a large say in the government. They were made invisible. Later on, the Mi''kmaw people were made visible through public effort, although it was through association. Meaning, the issue wasn't addressed due to the higher cancer rate, but it was acknowledged because lobsters in the harbor contained large amounts of toxic materials. After this, millions of dollars were put in to clean the Tar Ponds. It is interesting to see that the struggle of the indigenous people wasn't directly addressed.

Further, in Louisiana's "cancer Alley," also known as Donaldsonville, contaminants make it so that those who drink from the Mississippi River have a 2.1 times chance of getting rectal cancer and those who live within a mile of chemical facilities have a 4.5 times chance of getting lung cancer. The majority of people getting cancer in Donaldsonville from oil spills, dumping, shuck burning, chemical leaks, and more, are extremely low income, high poverty, and high illiteracy African Americans. Much of the toxin dumping in this area comes from the ease in which waste discharge applications are approved by the Environmental Control Commision. Residents of Donaldsonville have also noted that while trying to organize in opposition to the environmental condition, chemical industries have been known to buy out those that are protesting. Residents have no choice but to accept due to the poverty they live in (Singer, 2011). This is notable because it shows the power of corporations in Louisiana when it comes to the law. Standards are bent to accommodate the priorities of industries, while overlooking the residents of Donaldsonville due to their lack of resources, funding, education, and political power. Moreover, the industries and government have downplayed the effect of human made toxins on the health of residents in the “cancer Alley.” Sickness has been attributed to the lifestyle choices of residents, including eating habits, smoking, exercise, etc. It is evident that this divide has swayed residents into believing that the environmental risk they face is a natural feature, it is normal and there is no large danger to it. This has created additional ignorance in the community, making the underprivileged living in Donaldsonville even more invisible to risks they are facing due to a lack of standards (Singer, 2011).

With all of these issues in healthcare from environmental injustice to the unfair treatment of minorities, it is difficult to create change and combat the wrongs being carried out by more privileged groups. A very important way to do this is through education and increasing awareness. Many people lack knowledge in these areas and are therefore ignorant and part of the problem. By educating the public, bigger actors can take part and speak on behalf of those who are overlooked. Furthermore, problems where residents themselves are in denial of risk, like with “cancer Alley”, will be less prevalent. When it comes to medical disparities, education of physicians accomplishes the same thing. By making sure all physicians are aware of racial biases, they can then work towards mitigating prejudice when it comes to their work. More than this, it is important for the standards to be changed at the base. Through education, people will be able to band together and press the government and controlling corporations to put in the effort to push for better standards, or at least compensate the people affected by injustice. More specific to racial disparities in health care, education may not be enough. One solution to this is to push physicians to individuate - share power with the patient, develop a relationship, and see them as an individual (Penner et al., 2014). There should also be a very large emphasis on clear communication. Lastly, the issue comes back to knowledge. Information within health care systems should be aggregated to make clear what bias is really taking place. This information should be analyzed and released to the public, so that all parties know and can work towards a benchmark to decrease the bias.

Next Steps

Since the technical research capstone project was already completed, the next steps will apply only to the STS research topic. The following steps will be taken:

- Evaluate Prospectus after submission (look for problem areas)
- Complete a draft of the research paper by February/March 2022
- Submit the research paper by April/May 2022

Besides this tentative schedule, with my further research I will be looking into more solutions to the issues of environmental injustice and racial healthcare disparity. These will be elaborated on and will allow for a more concrete response/action to the research questions stated above. I may also choose to look into other illnesses besides cancer to see if there is a correlation between causes through human-made toxins. Some resources that can be used for this are in the References section below - (Geiger, n.d.), (Taylor & Francis, n.d.), and (Dillon, 2016).

References

Lathrop, R. H. (2010, February 9). *p53 Mutants Data Set*. UCI Machine Learning Repository: P53 mutants data set. Retrieved November 2, 2021, from https://archive.ics.uci.edu/ml/datasets/p53+Mutants.

Baronio, R., Ho, L., Hall, L., Salmon, K., Hatfield, G. W., Kaiser, P., & Lathrop, R. H. (2009, September 4). *Predicting positive p53 cancer rescue regions using most informative positive (MIP) active learning*. PLOS Computational Biology. Retrieved November 2, 2021, from https://journals.plos.org/ploscompbiol/article?id=10.1371%2Fjournal.pcbi.1000498.

Danziger, S. A., Zeng, J., Wang, Y., Brachmann, R. K., & Lathrop, R. H. (2007, July 1). *Choosing where to look next in a mutation sequence space: Active learning of informative p53 cancer rescue mutants*. Bioinformatics (Oxford, England). Retrieved November 2, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2811495/.

Danziger, S. A., Swamidass, S. J., Zeng, J., Dearth, L. R., Lu, Q., Chen, J. H., Cheng, J., Hoang, V. P., Saigo, H., Luo, R., Baldi, P., Brachmann, R. K., & Lathrop, R. H. (2009, September 22). *Functional census of mutation sequence spaces: The example of p53 cancer rescue mutants*. IEEE/ACM transactions on computational biology and bioinformatics. Retrieved November 2, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2748235/.

Brownlee, J. (2020, June 30). *Introduction to dimensionality reduction for machine learning*. Machine Learning Mastery. Retrieved November 2, 2021, from https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/.

Brownlee, J. (2014, October 6). *An introduction to feature selection*. Machine Learning Mastery. Retrieved November 2, 2021, from https://machinelearningmastery.com/an-introduction-to-feature-selection/.

Hainut, P., Soussi, T., Shomer, B., Hollstein, M., Greenblatt, M., Hovig, E., Harris, C., & Montesano, R. (1997, January 1). *Database of p53 gene somatic mutations in human tumors and cell lines: updated compilation and future prospects*. Oxford Academic. Retrieved November 2, 2021, from https://academic.oup.com/nar/article/25/1/151/1085590.

Paul, S. (2020, January 2). *Feature selection in python sklearn*. DataCamp Community. Retrieved November 2, 2021, from https://www.datacamp.com/community/tutorials/feature-selection-python.

Scikit Learn. (n.d.). *Sklearn.feature_selection.RFE*. scikit. Retrieved November 2, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html#sklearn.feature_selection.RFE.

K, D. (2019, October 31). *K-means clustering using sklearn and python*. Medium. Retrieved November 2, 2021, from https://heartbeat.fritz.ai/k-means-clustering-using-sklearn-and-python-4a054d67b187.

Navlani, A. (2018, May 16). *Sklearn Random Forest classifiers in Python*. DataCamp Community. Retrieved November 2, 2021, from https://www.datacamp.com/community/tutorials/random-forests-classifier-python.

Scikit Learn. (n.d.). *Sklearn.preprocessing.LabelEncoder*. scikit. Retrieved November 2, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html.

Fisher, R. A. (1988, July 1). *Iris Data Set*. UCI Machine Learning Repository: Iris data set. Retrieved November 2, 2021, from https://archive.ics.uci.edu/ml/datasets/iris.

What is cancer? National Cancer Institute. (n.d.). Retrieved October 18, 2021, from https://www.cancer.gov/about-cancer/understanding/what-is-cancer.

Innovation policy, structural inequality, and covid-19 - par.nsf.gov. (n.d.). Retrieved October 18, 2021, from https://par.nsf.gov/servlets/purl/10227226.

Kenny, K. E. (n.d.). *The biopolitics of global health: Life and death in neoliberal time - Katherine E. Kenny, 2015*. SAGE Journals. Retrieved October 18, 2021, from https://journals.sagepub.com/doi/full/10.1177/1440783314562313?casa_token=TRPZVf0vAuQAAAAA%3ADL8IgIk-jJ0LhfCJO5Mf6kK11cgBwnobx4AIPNOWP8PQfCH-A2dUkboyw7p3YzdTvMkEWQ4iJxDU.

Anderson, W. (n.d.). *Making Global Health History: The Postcolonial Worldliness of Biomedicine*. Validate user. Retrieved October 18, 2021, from https://academic.oup.com/shm/article/27/2/372/1708273.

Undone Science and counter-expertise: Fighting for justice in an Argentine community contaminated by pesticides. Taylor & Francis. (n.d.). Retrieved October 18, 2021, from https://www.tandfonline.com/doi/full/10.1080/09505431.2018.1533936.

Google. (n.d.). *The public shaping of medical research*. Google Books. Retrieved October 18, 2021, from https://books.google.com/books?hl=en&lr=&id=-z6cBQAAQBAJ&oi=fnd&pg=PP1&dq=%22science%2Band%2Btechnology%2Bstudies%22%2Bcancer%2Bresearch%2Bjustice&ots=suoZomZ3kg&sig=OrZeUd_-EcFITVMJvRyPASDLfEM#v=onepage&q=%22science%20and%20technology%20studies%22%20cancer%20research%20justice&f=false.

Ocran Mattila, P., Ahmad, R., Hasan, S. S., & Babar, Z.-U.-D. (1AD, January 1). Availability, affordability, access, and pricing of anti-cancer medicines in low- and middle-income countries: A systematic review of literature. Frontiers. Retrieved October 4, 2021, from https://www.frontiersin.org/articles/10.3389/fpubh.2021.628744/full.

Nambi Ndugga and Samantha Artiga Published: May 11, 2021. (2021, May 12). Disparities in health and health care: 5 key questions and answers. KFF. Retrieved October 18, 2021, from https://www.kff.org/racial-equity-and-health-policy/issue-brief/disparities-in-health-and-health-care-5-key-question-and-answers/.

Geiger, J. (n.d.). *Racial and ethnic disparities in diagnosis and treatment: A review of the evidence and a consideration of causes*. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Retrieved November 3, 2021, from https://www.ncbi.nlm.nih.gov/books/NBK220337/.

Singer, M. C. (2011, June). *Down cancer Alley: The Lived Experience of Health and Environmental Suffering in Louisiana's Chemical Corridor*. ResearchGate. Retrieved November 3, 2021, from https://www.researchgate.net/publication/51564478_Down_cancer_Alley_The_Lived_Experience_of_Health_and_Environmental_Suffering_in_Louisiana's_Chemical_Corridor.

Wolff, M. S., Britton, J. A., & Wilson, V. P. (2003). *Environmental Risk Factors for Breast Cancer among African-American Women*. EPA. Retrieved November 3, 2021, from https://cfpub.epa.gov/ncer_abstracts/index.cfm/fuseaction/display.files/fileID/13597.

Dillon, L. (2016, September). *Police Power and Particulate Matters: Environmental Justice and the Spatialities of In/Securities in U.S. Cities*. ResearchGate. Retrieved November 3, 2021, from https://www.researchgate.net/profile/Lindsey-Dillon/publication/318707771_Police_Power_and_Particulate_Matters_Environmental_Justice_and_the_Spatialities_of_InSecurities_in_US_Cities/links/5978a8b6a6fdcc30bdc2cbe1/Police-Power-and-Particulate-Matters-Environmental-Justice-and-the-Spatialities-of-In-Securities-in-US-Cities.pdf.

Penner, L. A., Blair, I. V., Albrecht, T. L., & Dovidio, J. F. (2014, October 1). *Reducing racial health care disparities: A Social Psychological Analysis*. Policy insights from the behavioral and brain sciences. Retrieved November 3, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4332703/.