Improving Heart Disease Outcomes in the United States

Colin Beveridge, Ben Grudt, Grace Seiler, Hang Zhang Cao

February 18, 2022

Introduction: In Chapter 4 of Malcolm Gladwell's bestselling book *Blink: The Power of Thinking Without Thinking,* Gladwell discusses the struggles of Cook County Hospital, a public hospital near downtown Chicago, to efficiently diagnose patients experiencing symptoms of a heart attack. Gladwell elaborates on just how difficult this problem is, and how expensive and dangerous mistakes can be.

Doctors are constantly asked to toe the line between caution and pragmatism. On one hand, the average cost of a night in the hospital for a patient in the United States is $11,700 (Fay). The average cardiac stay requires around three nights. Additionally, it can be dangerous to take up a bed in a crowded hospital when someone experiencing a serious issue may need it more. This problem has gotten worse over the years because the medical community has done such a good job of educating the public on the symptoms of heart attacks. Most people rush to the hospital at the first sign of chest pain. These days, only about 10% of patients complaining of chest pain are having a heart attack (Gladwell, 131). Of course, there are also massive costs and dangers to not being cautious enough. Obviously, sending a patient home when they are in the throes of a heart attack is close to the worst-case scenario for any hospital emergency department. Compounding this danger is the monetary threat of malpractice lawsuits for the doctor and hospital.

The real surefire test to confirm that someone is having a heart attack takes hours that the doctor making the diagnosis does not have. Instead, doctors rely on several different pieces of information about the patient, including their age, blood pressure, medical history, EKG reading and more. The problem with this strategy is that sometimes doctors are wrong (Gladwell, 129).

This is a serious problem that most hospitals deal with in some form, but Cook County is special because in the late 1990's their hand was forced, being a public hospital in a crowded metropolitan area, to find a more efficient and effective way of making these decisions. In 1996, Brendan Reilly became the chairman of their Department of Medicine and ushered in a

project to use the work of Lee Goldman from the 1970's. Goldman had dug into patient cases to try to determine an algorithm for predicting whether or not a patient was having a heart attack and came up with an effective decision tree. From 1996 to 1997, Reilly compared doctors to the algorithm, and the results were not close. Overall, the algorithm was about 70% better at recognizing the patients that were not having a heart attack. As for the patients that were having a serious heart attack, doctors guessed correctly between 75% and 89% of the time, while the algorithm diagnosed correctly better than 95% of the time (Gladwell, 136).

This case study makes three points abundantly clear. The first is that the problem of diagnosing heart attacks accurately is incredibly important. According to the Centers for Disease Control and Prevention, heart disease is the leading cause of death in the United States, with about 659,000 people dying from heart disease each year (National Center for Chronic Disease Prevention and Health Promotion). Every correct diagnosis equals lives and money saved for all parties. The second is that doctors usually are not very good at these diagnoses, or at the very least there is room for improvement. Finally, the Cook County Hospital case study suggests that some form of machine learning classifier can improve upon doctor performance. The combination of these three factors yielded the idea for this project.

Project Goal: With these three points in mind, the group decided to pursue a project that investigated the application of machine learning to this important problem. The final project idea is to serve a hypothetical client, with the goal of improving accurate diagnoses for patients experiencing chest pain or wanting to get their heart examined, while minimizing costs. The client will establish clinics in locations recommended by the group, with the intention of serving the largest number of patients in need of these services.

The group is specifically looking for high-risk individuals that would benefit from a clinic that provides a low-cost intermediate option to seeking emergency room services. These groups have been identified as: low income and uninsured people, men and people aged 65+, and people who smoke or who are obese. These groups were chosen based on their need for low-cost medical services and their increased risk for developing heart disease (Mayo Clinic).

Once these clinics are placed in optimal locations to help the most people, their goal will be to accurately diagnose patients, and refer the patients most likely to be experiencing some form of heart disease to the nearest hospital or doctor for more extensive medical treatment. Additionally, a dashboard will be created that aims to educate the public on heart disease, risk factors, and insurance information in their area. It will provide links to sites that direct the user to information on improving heart health and obtaining insurance.

Limitations: Due to the scope of the course and this project, the research in this report is rather limited. Several factors limited the ability of the group to obtain accurate machine learning results and to select locations with the most need. To extend the results of this project to real world applications, the group would need access to more comprehensive patient data and more current and specific hospital data.

The first limitation of this project is that the total patient data available to develop a machine learning model was only about 900 records. Splitting this dataset into 75% training and 25% testing, the model trained on less than 700 records. In addition, there were only 11 features in the dataset. More patient records and more information per record would likely have led to a better model.

Another limitation on the patient data is the lack of clarity of the target column of the classifier. The dataset describes the features as being important to predicting possible "heart disease," not specifically heart attacks. This was not the original goal of the project and represents a limitation of the model due to the lack of real emergency room data. However, the primary objective of the client is to improve overall heart health in the US. The clinics will do that by allowing people a cheap way to get their heart checked, and then giving them a recommendation about seeking more extensive medical help. Though the clinics are not technically classifying heart attacks as Gladwell describes, they are nonetheless providing a very important medical service to the public. If someone is experiencing chest pain but is classified as negative for heart disease according to the model, it is unlikely that they are experiencing a heart attack. If a patient with no current issues except some long-term concerns about their health shows up, then they can be told whether they need to see a doctor based on the results

of the classifier. The current model serves more as a future risk predictor for complications from heart disease, which of course very often includes a heart attack.

The group was also limited by the ambiguity in deciding which cities or metropolitan areas would get a clinic. The best the group could do was to consider states and metropolitan areas that were most economically in need and had very prevalent risk factors, and then create a priority list from that data. Ideally, real time hospital data about wait times in emergency rooms for heart attack patients would allow the group to narrow down the search even further. Unfortunately, the group did not have access to data that specific. The group had to settle for a less exact method of selecting locations with the most need. However, this is not a huge problem, mainly because these clinics are very likely to help a lot of people regardless of where they are placed. Even states with very strong healthcare systems have use for this service, as no hospital handles this issue perfectly.

Regional Focus: The group began its investigation into ideal locations for clinics by looking for certain at-risk economic and demographic groups and behavioral risk factors at the state level. The classifications have been identified as: low income and uninsured people, men and people aged 65+, and people who smoke as well as obese people. For the economic and demographic classifications, these categories are based on the populations of these classifications by state. The group recognizes that states with high overall populations are more likely to have high populations of the relevant classifications. It was decided that the clinics are meant to serve a high number of people, so the population of an area should play some role in the decision. Therefore, this data remains relevant to the analysis.

Figure 1 shows one such example of a map used to determine locations with a high population of uninsured people. The population sizes were found using Census data. The state in red has the highest population of uninsured people, followed by gradients of blue for states with high numbers of uninsured people. States below a certain threshold of uninsured people are marked in grey.

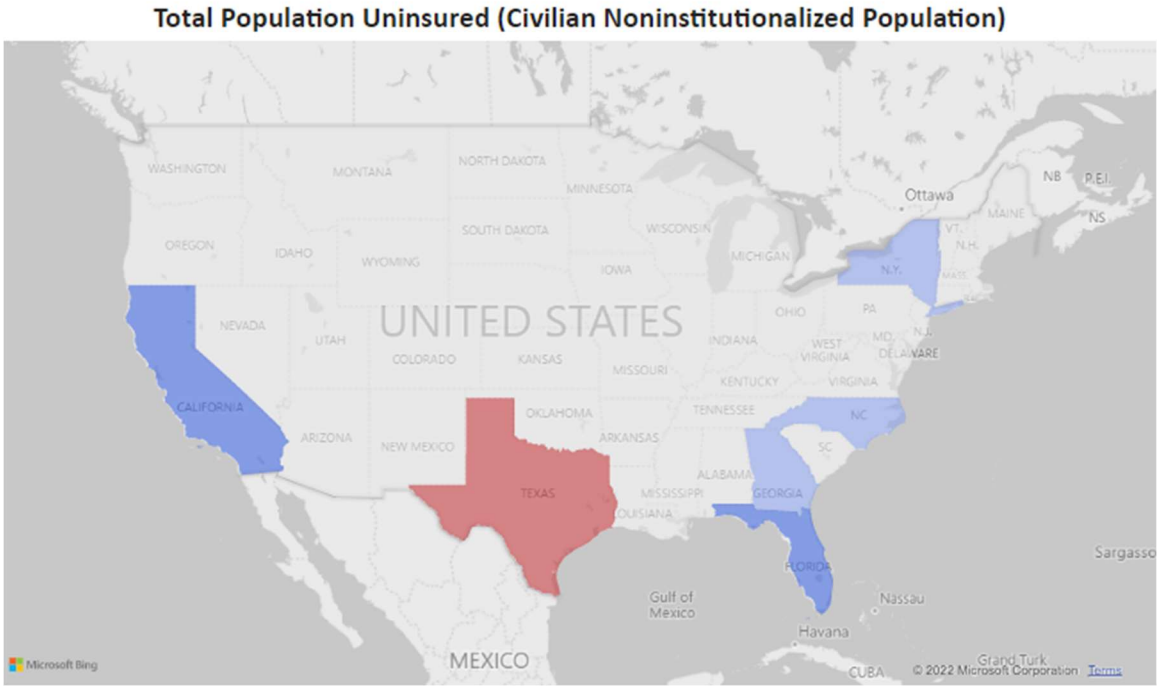

Figure 1: Map highlighting states that have high populations of uninsured people. Color break-down: states in red have the highest population of uninsured people with 4 million or more, dark blue states have 2 million-4 million, light blue states have 1 million-2 million, and gray states have 0-1 million uninsured people.

The group used Census data to repeat the above process for states with high populations of low-income people, males, and people aged 64 years and above. The data was additionally used to explore combinations of these demographics, including uninsured males and uninsured people aged 64 years and older. Figure 2 shows the group's findings across all economic and demographic groups.

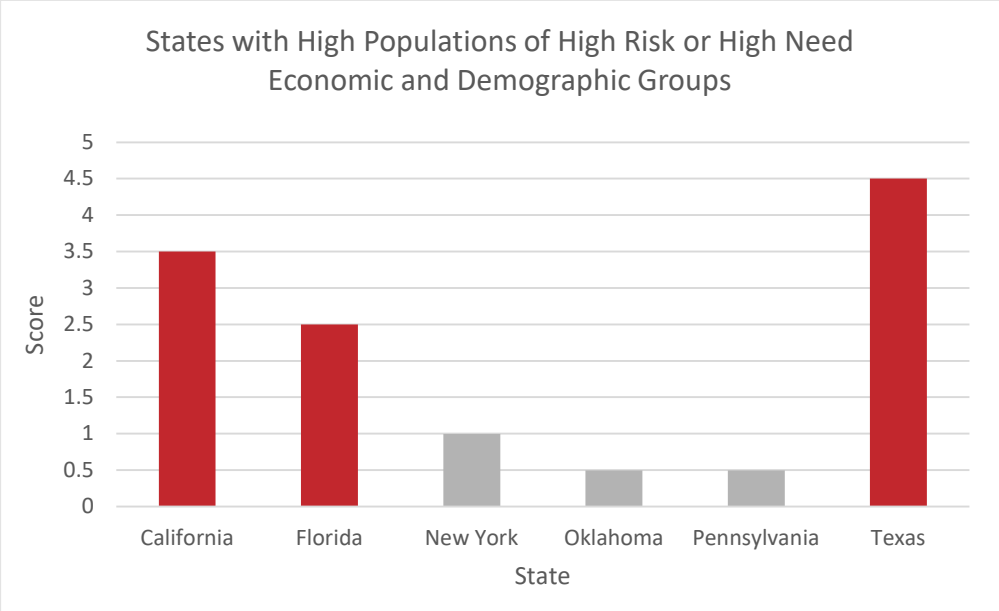

Figure 2: State scores based on the various economic or demographic classifications identified by the group as high risk or high need. Scores are given based on the category they fall into for each individual classification, with states in the red category given a score of 1 for that group and states in the dark blue category given a score of 0.5.

As you can see, Texas, California, and Florida had high populations of high need or high risk economic and demographic groups. As discussed, this was expected, as these states have a high total population.

The group followed a similar process using the CDC Behavioral Risk Factor data. This CDC data sampled groups of people and found the prevalence of each Behavioral Risk Factor as a percentage of the sample group.

Figure 3 shows the map used to determine states with a high prevalence of smoking among adults. For the CDC data, a percentage of the highest value is used to determine high-risk states rather than the total population. In Figure 3, West Virginia is the state with the highest prevalence of current smoking among adults, with 27 percent of the sample population reporting current smoking status. The highest risk state is reported in red, followed by gradients of blue for decreasingly high-risk states. Lower risk states are reported in grey.

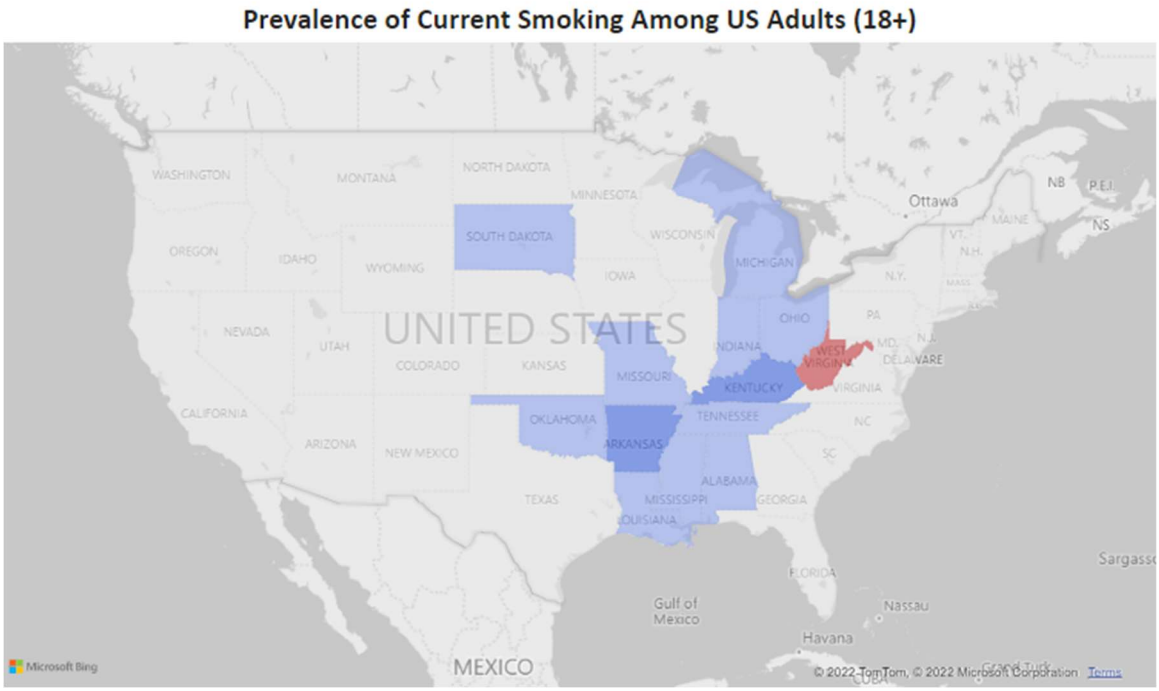

Figure 3: Map highlighting states that have a high prevalence of current smoking status among adults. Color break-down: 100-90 percent of the highest value, dark blue is 90-80 percent, light blue is 80-60 percent, and gray is 60-0 percent.

The CDC Behavioral Risk Factor data was also used to determine prevalence of obesity, physical inactivity, and major cardiovascular disease among US adults (18+). Figure 4 shows the group's findings across all behavioral risk factors.

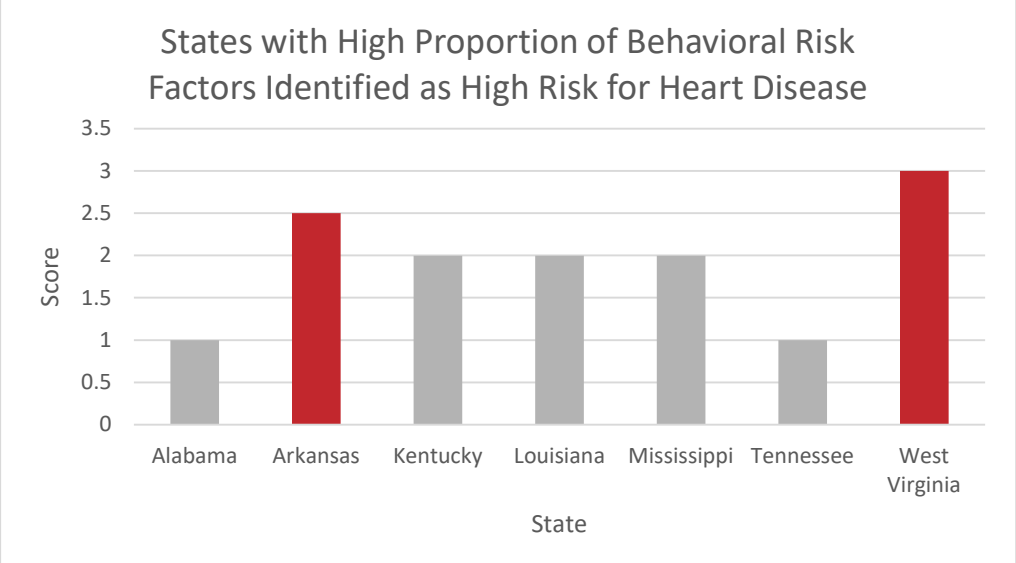

Figure 4: State scores based on the various behavioral risk factors identified as high risk for heart disease. Scores are given based on the category they fall into for each risk factor, with states in the red category given a score of 1 for that group and states in the dark blue category given a score of 0.5.

As you can see, West Virginia and Arkansas had the highest score of behavioral risk factors identified as high risk for heart disease.

Local Focus: Next the group investigated metropolitan areas with similar classifications to those used at the state level. Using another CDC Behavioral Risk Factor dataset, the group once again identified high need or high-risk classifications that would benefit most from the accessibility of an intermediate heart disease clinic.

Figure 5 is a map of adults in the US with no kind of health care coverage, used by the group to determine economic (healthcare-centric) need. For the metropolitan CDC data, the group followed the same procedure as the state-based CDC data for analyzing categories. The highest risk metropolitan area is reported in red, followed by gradients of blue for decreasingly high-risk areas. Lower risk areas are reported in grey.

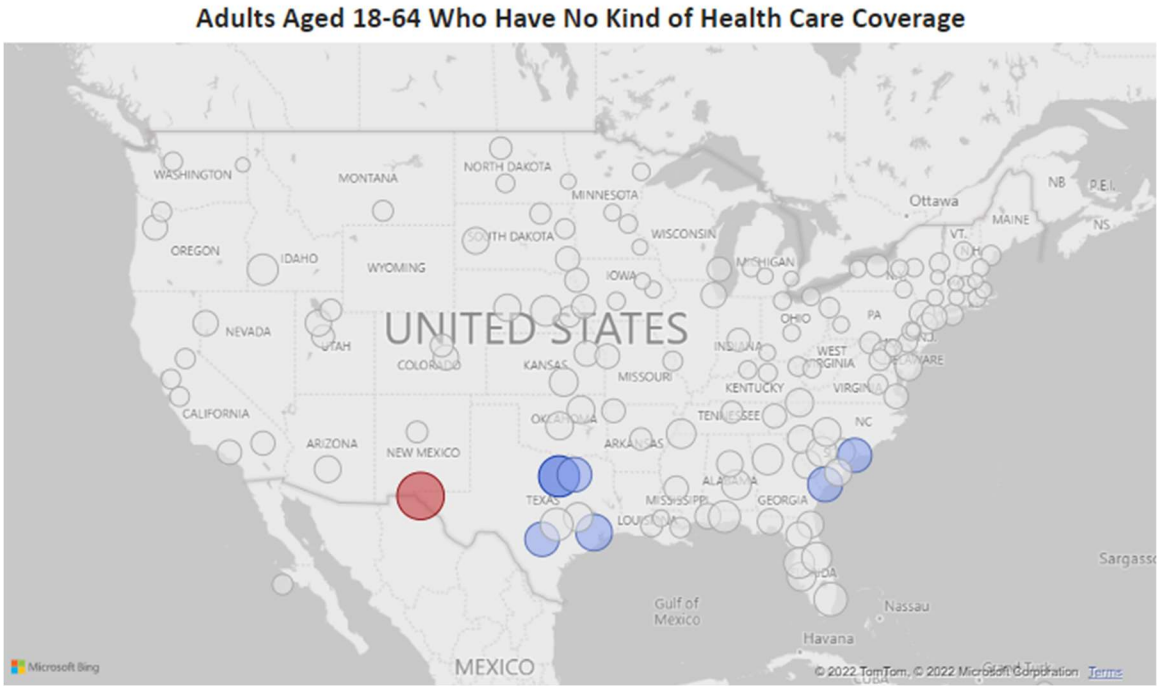

Figure 5: Map highlighting metropolitan areas that have a high proportion of adults without healthcare coverage. Color break-down: 100-90 percent of the highest value, dark blue is 90-80 percent, light blue is 80-60 percent, and gray is 60-0 percent.

The metropolitan level CDC Behavioral Risk Factor data did not include the same fields that the state level Census and CDC Behavioral Risk Factor datasets included. However, the fields were similar enough to analyze the same trends at both levels. The group analyzed the state and

metropolitan level datasets independently, using both to draw conclusions about the areas in need of clinics.

In addition to adults with no kind of healthcare coverage, the group also analyzed which metropolitan areas had a high proportion of respondents answering yes to the question: was there a time in the past 12 months when you needed to see a doctor but could not because of cost? The results identifying metropolitan areas with high need can be found in Figure 6.

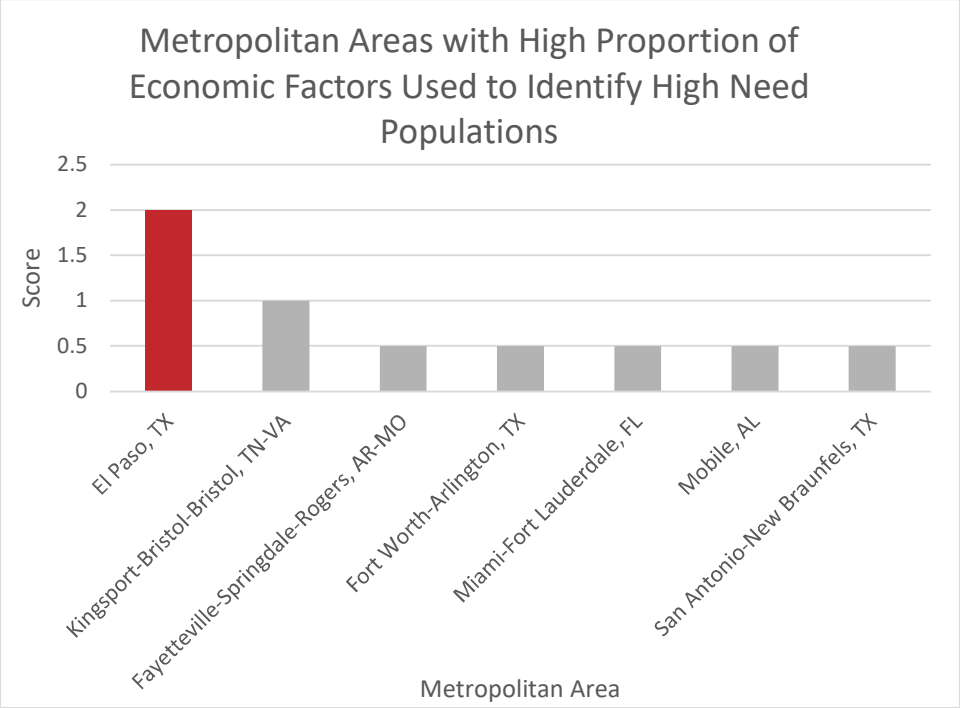

Figure 6: Metropolitan area scores based on economic factors used to identify high need populations. Scores are given based on the category they fall into for each risk factor, with states in the red category given a score of 1 for that group and states in the dark blue category given a score of 0.5.

El Paso, Texas is identified as being the most in need economically of a potential clinic. This observation aligns with the group's state-based investigation of economic and demographic high need and high-risk groups, in which Texas also scored the highest.

The group next analyzed behavioral risk factors that were determined to indicate high risk for heart disease. Figure 7 is a map of people who smoke every day or some days, used by the group to determine metropolitan areas that have high behavioral health risk factors. For the metropolitan CDC data, the group followed the same procedure as the state level CDC data for

analyzing categories. The highest risk metropolitan area is reported in red, followed by gradients of blue for decreasingly high-risk areas. Lower risk areas are reported in grey.

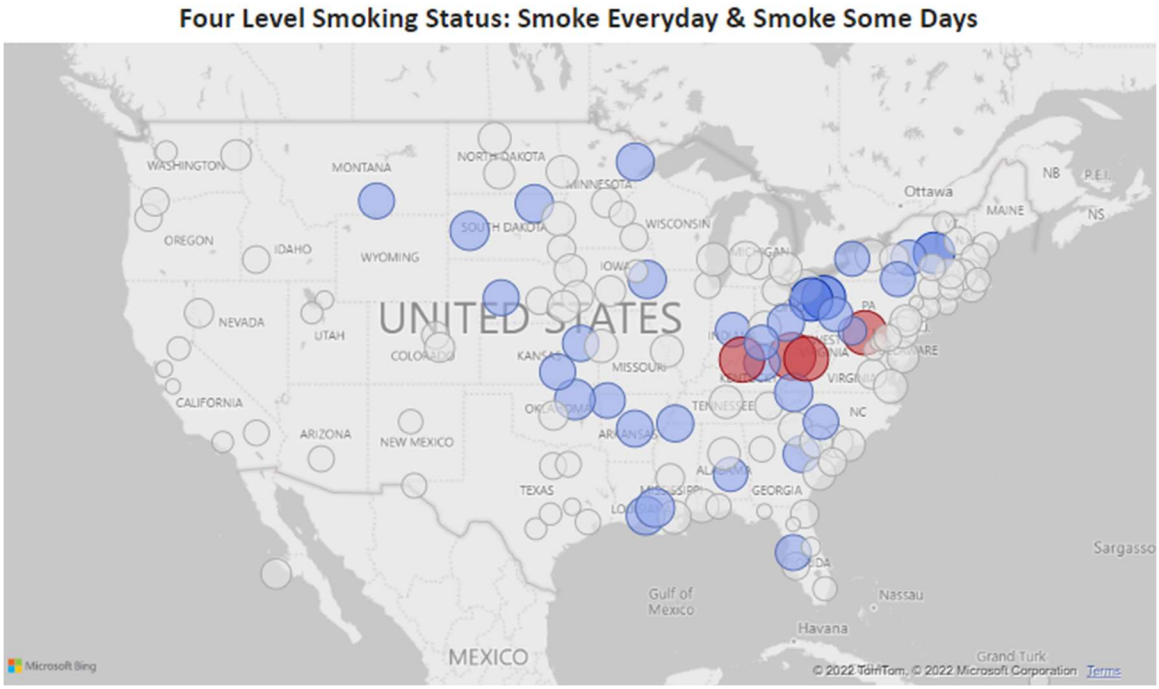

Figure 7: Map highlighting metropolitan areas that have a high proportion of adults who smoke at least some days. Color break-down: 100-90 percent of the highest value, dark blue is 90-80 percent, light blue is 80-60 percent, and gray is 60-0 percent.

The group used CDC data to repeat the above process for metropolitan areas with high percentages of obese people and people who have reported having coronary heart disease. Figure 8 shows the group's findings across these behavioral risk factor groups.

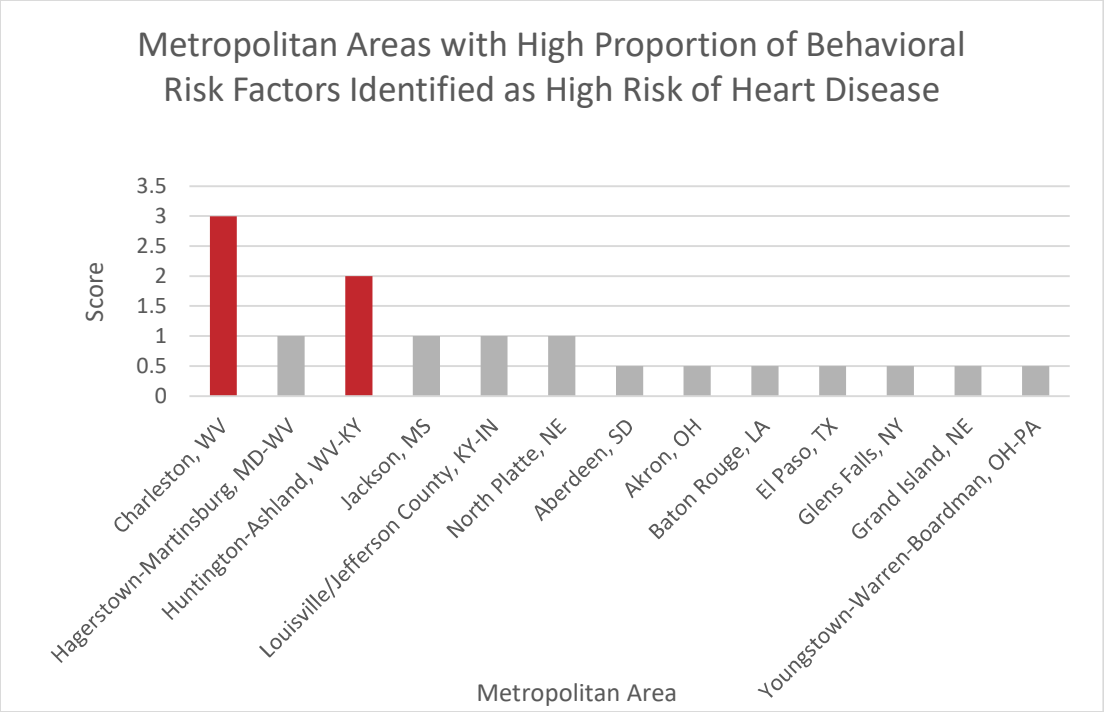

Figure 8: Metropolitan area scores based on behavioral risk factors used to identify populations at high risk for heart disease. Scores are given based on the category they fall into for each risk factor, with states in the red category given a score of 1 for that group and states in the dark blue category given a score of 0.5.

Two metropolitan areas were identified by the group as being highest risk for developing heart disease in the state of West Virginia. This once again aligns with the results of the state-level investigation.

Machine Learning: The group used the Heart Failure Prediction Dataset by fedesoriano from Kaggle to generate the machine learning (ML) model (Kaggle). This dataset is a combination of 5 different independent heart datasets from the Cleveland, Hungary, Switzerland, Long Beach, and Stalog datasets. The Heart Failure Prediction Dataset contains attributes of patients, which are useful for diagnosing heart disease. These attributes are age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting electrocardiogram results, maximum heart rate, exercise-induced angina, old peak, slope of the peak exercise ST segment. The dataset also indicates the presence of heart disease in each patient, allowing the group to use supervised machine learning models.

Prior to deciding on implementing any ML models, the group first examined the dataset. The outcome is either 0 or 1, heart disease or no heart disease, and thus the group identified it to be a binary classification problem. Next, the group split up the dataset into 75% training data and 25% testing data. After splitting, the group looked for any missing values or outliers within the training dataset. The group found that out of 688 rows of data, 128 of them had the value 0 for the Cholesterol column. The RestingBP column also had 0 as a value in some rows. Additionally, certain datapoints in the Cholesterol column had values greater than 500 mg/dL, which is exceedingly high considering the normal range of cholesterol is < 200 mg/dL. However, after more research on this topic, it was found that it is possible for individuals to have a cholesterol level greater than 500 mg/dL. As a result, these seemingly high values of cholesterol level are not treated as outliers in the machine learning model.

The 0 values in the Cholesterol and RestingBP columns are addressed by imputing with the mean. The function SimpleImputer is not used in this case. Instead, a custom imputation method is developed. Research suggests that individuals of different genders are likely to develop different types of heart diseases at different ages (Maas and Appelman). As a result, the group decided to impute the missing values based on the mean of the nearest neighbors. The group defines nearest neighbors as individuals with the same gender who are within an age range of plus or minus 10 years. This custom imputation method improves the accuracy and precision by approximately 0.5% and 1%, respectively. Other feature engineering techniques were also applied to this dataset, such as data normalization using sklearn's StandardScaler function (Pedregosa et al), and data balancing using the SMOTE module in imbalanced-learn (Lemaitre et al). Both techniques can greatly improve the performance of some of the tested ML models.

The group conducted a preliminary examination of the important features by plotting the density distribution of columns containing numeric values using the seaborn module (Waskom), as shown below.

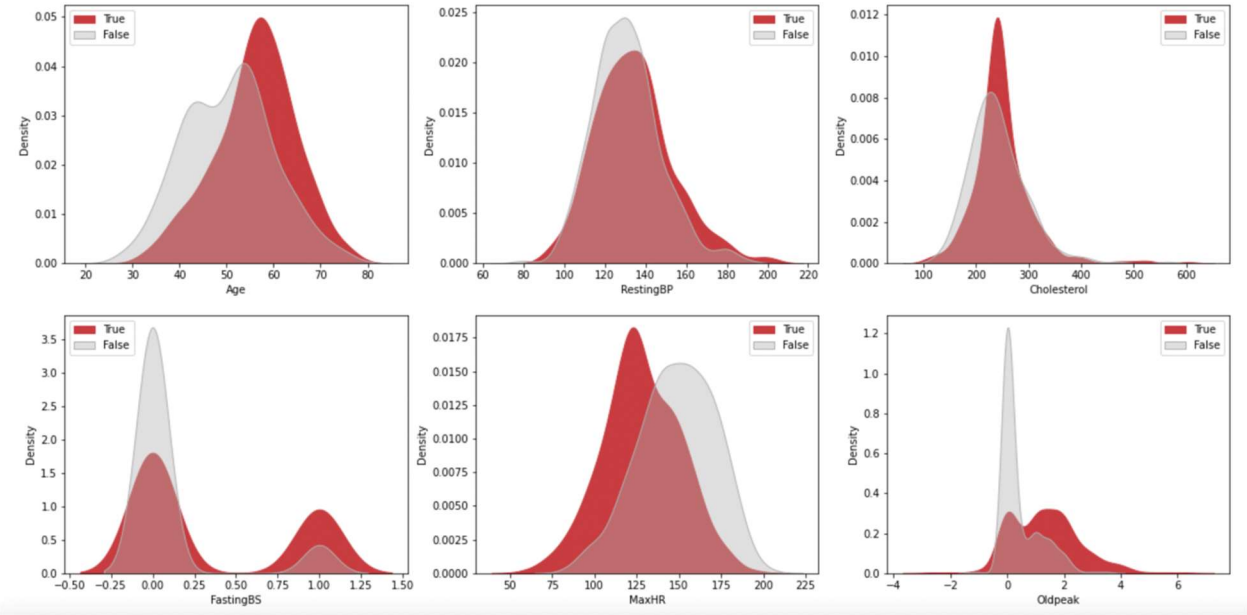

Figure 9: Density distribution plots of numeric data used by the group for feature importance examination.

Individuals with heart disease are labeled as True while individuals without heart disease are labeled as False. The degree of overlap between True and False reflects the importance of the feature in diagnosing heart disease. For example, the density plot of RestingBP is largely overlapped, which is an indication of irrelevance in feature importance. From the density distribution plots, the group identified MaxHR, Oldpeak, and Age to be the three most important features in diagnosing heart disease for the ML model.

The data given in the dataset allows supervised classification training, thus the group decided to compare the performance of seven different ML models: logistic regression, K-nearest neighbor classification, decision tree, random forest classification, Naïve Bayes classification on Bernoulli and Gaussian distribution, and support vector classification (SVC). The performance of these models is evaluated using K-fold cross validation with cv = 5. The models are evaluated based on different scoring metrics, such as accuracy, precision, recall, F1, and the ROC AUC score, as shown below in Figure 10.

	Model	Accuracy	Precision	Recall	F1	ROC
1	LogisticRegression	0.8644	0.8650	0.8680	0.8656	0.9349
2	KNeighborsClassifier	0.8657	0.8530	0.8888	0.8695	0.9172
3	DecisionTreeClassifier	0.8075	0.8030	0.8165	0.8087	0.8075
4	RandomForestClassifier	0.8774	0.8639	0.8990	0.8804	0.9297
5	BernoulliNB	0.8618	0.8643	0.8604	0.8621	0.9257
6	GaussianNB	0.8618	0.8602	0.8681	0.8632	0.9156
7	SVC	0.8799	0.8717	0.8965	0.8830	0.9308

Figure 10: Cross validation scores of each machine learning model tested.

The Decision Tree Classifier scored the lowest amongst all models, with the remaining models having near-identical scores for each category. The group decided to adopt SVC as the machine learning model since it has the highest average scoring and is suitable for binary classification problems with small datasets.

Hyperparameter tuning of the SVC model was attempted using GridSearchCV function of the sklearn module (Pedregosa et al). The major hyperparameters c, gamma, and kernel were tuned and the GridSearchCV returns the optimum values to be c = 10, gamma = 0.05, and kernel = “poly”. The group then evaluated both the base SVC model and the tuned SVC model based on the confusion matrix and Matthew's correlation. Interestingly, the tuned model shows a slight decrease across parameters such as accuracy, precision, sensitivity, and Matthew's correlation. Even though the tuned SVC model yields two extra records of True Positive diagnosis compared to that of the base model, the model also yields five extra records of False Negative diagnosis. Since the goal of this project is to develop a model that can accurately diagnose potential heart disease, the negative effect of false negative outweighs the benefit of the true positives. Thus, the group decided to adopt the base SVC model for prediction.

Recommendations: To rank metropolitan areas by their score, the group used a simple method to calculate a metric that indicates the need of a given area multiplied by their population. The goal of the client is to help the most patients possible. Therefore, population size must play a role in the decision. The metric was calculated by taking the score of the area given above in the bar graphs, multiplying by the population of the area, and then dividing by 1,000,000. The results are in Table 1.

Metropolitan Area	Metric
Fort Worth-Arlington, TX	3.85
Miami-Fort Lauderdale, FL	3.00
El Paso, TX	2.17
San Antonio-New Braunfels, TX	1.30
Louisville/Jefferson County, KY-IN	1.26
Huntington-Ashland, WV-KY	0.72
Charleston, WV	0.62
Jackson, MS	0.59
Baton Rouge, LA	0.44
Akron, OH	0.35

Table 1: Metropolitan Area ranked by metric.

Table 1 represents a list of the final metropolitan areas that the group recommends to the client for clinic sites. The metric was designed to target the areas with the highest populations that showed up somewhere on the highest risk list for any of the risk factors. The three areas with the most risk factors represented were Charleston, Huntington-Ashland, and El Paso. However, these areas are not first on the list of recommendations because the metric is population weighted. El Paso is still the third priority recommendation, and the West Virginia areas still make the top ten list. But they are not highest priority because they are not as populous as Fort Wort, TX or Miami-Fort Lauderdale, FL. The client can use this list as a guide for the allocation of funds, which will achieve the client's original goal of helping the maximum number of possible patients in need of these services.

The machine learning model developed for this project, though limited, remains reasonably accurate at diagnosing heart disease. The model does not exactly fit the description of the work done by Reilly in the introduction to this report, and the model is not quite as effective as the algorithm used in that research. However, given the limitations of the project and the general goals of the client, the model remains very successful despite its shortcomings. The model diagnoses heart disease in patients with an accuracy of around 90%. It can provide a very important service to high-risk communities with large uninsured populations. From here, the client can select ideal locations to place clinics, which will implement the finalized model to assist the local populations.

Works Cited:

American Heart Association. (2022). *American Heart Association | To be a relentless force for a world of longer, healthier lives*. Www.Heart.Org. Retrieved February 11, 2022, from https://www.heart.org/en/

Behavioral Risk Factor Surveillance System (BRFSS) - National Cardiovascular Disease Surveillance Data | Chronic Disease and Health Promotion Data & Indicators. (2021, June 24). [Behavioral risk factors survey by state.]. Centers for Disease Control. https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Behavioral-Risk-FactorSurveillance-System-BRFSS-N/ikwk-8git

Fay, B. (2021, October 12). *Hospital and Surgery Costs – Paying for Medical Treatment*. Debt.Org. Retrieved February 11, 2022, from https://www.debt.org/medical/hospital-surgery-costs/

Gladwell, M. (2005). *Blink: The Power of Thinking Without Thinking*. Little, Brown and Company.

Harris Et Al., C. (2020). *Numpy* (1.19.2) [Numpy is the primary array programming library for the python language.]. Nature. https://www.nature.com/articles/s41586-020-2649-2#citeas

Hunter, J. (2007). *Matplotlib* (3.4.2) [Matplotlib is a 2D graphics package used for Python for application development, interactive scripting, and publication-quality image generation across user interfaces and operating systems]. IEEE. https://ieeexplore.ieee.org/document/4160265/authors#authors

Kaggle. (2021, September 10). *Heart Failure Prediction Dataset* [Data provides information on 11 different risk factors and indicators of heart disease, allowing for an ML Prediction Model]. https://www.kaggle.com/fedesoriano/heart-failure-prediction

Lemaitre Et Al., G. (2014). *Imblearn* (0.9.0) [Imbalanced-learn is an open-source python toolbox aiming at providing a wide range of methods to cope with the problem of imbalanced dataset]. Journal of Machine Learning Resources. https://jmlr.org/papers/v18/16-365.html

Mayo Clinic. (2021, February 9). *Heart disease - Symptoms and causes*. Retrieved February 2, 2022, from https://www.mayoclinic.org/diseases-conditions/heart-disease/symptomscauses/syc-20353118

Maas, A.H.E.M., Appelman, Y.E.A. (2010). Gender differences in coronary heart disease. *Netherlands heart journal: monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation*, 18(12), 598-602. doi: 10.1007/s12471-010-0841-y

McKinney, W. (2010). *Pandas* (1.2.4) [Python library of data structures and statistical tools]. 9th Python in Science Conference. https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

National Center for Chronic Disease Prevention and Health Promotion, Division for Heart Disease and Stroke Prevention. (2022, January 28). *Heart Disease | cdc.gov*. Centers for Disease Control and Prevention. Retrieved February 11, 2022, from https://www.cdc.gov/heartdisease/index.htm#:%7E:text=Heart%20disease%20is%20the%20leading,can%20lead%20to%20heart%20attack

National Center for Chronic Disease Prevention and Health Promotion, Division for Heart Disease and Stroke Prevention. (2021, September 27). *Heart Disease Resources | cdc.gov*. Centers for Disease Control and Prevention. Retrieved February 11, 2022, from https://www.cdc.gov/heartdisease/about.html

Pedregosa Et Al. (2011). *Scikit learn* (1.0.2) [Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems.]. Journal of Machine Learning Research. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Preidt, R. (2021, December 17). *Many ER Patients With Chest Pain Can Be Sent Home, Study Finds*. Consumer Health News | HealthDay. Retrieved February 11, 2022, from https://consumer.healthday.com/general-health-information-16/emergencies-and-first-aid-news-227/many-er-patients-with-chest-pain-can-be-sent-home-study-finds-699484.html

Waskom, M. (2021). *Seaborn* (0.11.1) [Seaborn is a library for making statistical graphics in Python. It provides a high-level interface to matplotlib and integrates closely with pandas data structures.]. Journal of Open Source Software. https://joss.theoj.org/papers/10.21105/joss.03021