Predicting Inpatient and Outpatient Hospitalization Rates in U.S. Counties with Alternative Data Sources

Carter Perkins
Courant Institute of Mathematical Sciences
Computer Science Department
New York University
New York City, U.S.A.
csp9835@nyu.edu

Abhishek Ranjane
Courant Institute of Mathematical Sciences
Computer Science Department
New York University
New York City, U.S.A.
arr7980@nyu.edu

***Abstract*—Understanding hospital usage in each US county is imperative for identifying characteristics of underserved communities and developing strategies to increase community well-being. Numerous studies have found alcohol consumption and high air pollution can have adverse effects on a population's overall health, and the usage of these substances is also affected by socioeconomic factors such as household income. In this project, we hypothesize that alcohol consumption and air pollution rates along with socioeconomic factors can be key components in predicting hospital usage by using the number of inpatient days and outpatient visits per capita as regression target variables. The purpose of this project is to use data from alternative data sources for air pollution, alcohol consumption, and socioeconomic factors to predict health markers for each county while analyzing the correlation between these factors. By analyzing data on the county level, we aim to derive regional patterns that may give insight into over and under-burdened populations that may be potentially lost by generalizing to a state or national level. We find that the best features are measures of demographics, healthcare facility characteristics, and alcohol consumption. Additionally, the latitude and longitude coordinates of a county has predictive performance which further emphasizes regional importance for predicting hospital rates. Our highest performing model, XGBoost Regressor, achieved 0.719 R^2 and 0.219 R^2 for inpatient and outpatient prediction respectively. We find that alternative data sources, particularly alcohol consumption, are beneficial for predicting county-level hospital rates.**

***Index Terms*—counties, alcohol, socioeconomic, air pollution, hospital, predictive analytics**

I. Introduction

Inefficiencies in hospital resource planning and strategy are one of the largest sources of healthcare costs in the United States. Triage systems frequently can be backed up due to under-staffing, surging medical cases, and lack of beds which directly affect the level of care patients will receive. Therefore, reducing healthcare costs is critical for providing better healthcare as approximately 25% of annual United States healthcare spending is wasted to factors such as low-value care, administrative complexity, and fraud [1]. As access to healthcare directly affects every individual in the United States, it is pertinent for health providers and legislative bodies to have an understanding of the factors that influence inpatient and outpatient hospitalization rates across the United States so they can mitigate healthcare costs to better serve their communities.

In this paper, we research the effect of using alternative data sources, such as air pollution, alcohol consumption, and demographic information, to predict inpatient days per capita and outpatient visits per capita in U.S. counties from 2006-2016. By associating healthcare at the county level, we are able to create a localized understanding of healthcare across the United States and potentially find correlations in different regions of the country. We plan to investigate the following research questions:

- **RQ1** Can alternative data sources be used for predicting inpatient and outpatient hospitalization rates?
- **RQ2** Which set of features are most important for predicting inpatient and outpatient hospitalization rates, and how much do these two sets differ?

II. Related Work

Previous work has utilized cohort studies of U.S. health plan claims for predicting inpatient hospitalizations [2], but have not explored regional populations as separate entities. There is powerful evidence, covered by multiple studies, that indicates the negative effects of alcohol on health factors. The negative effects include degrading cognitive function [3], declining skeletal system [4], effect on the digestive system [5] (specifically the gastrointestinal tract), and effect on the respiratory [6] and circulatory system [7].

Therefore, alcohol is negatively correlated with health measures collected at hospitals. A study done by Reuter et al. [8], shows that there was a dramatic drop in patient numbers when alcohol sales were prohibited in African states in 2019.

Similar to alcohol, air pollution also has a high impact on health measures. The research done by Manan et al. [9], indicates that air pollution levels, such as ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), and sulpher dioxide

(SO_2), are associated with an increasing trend of hospital admission.

Although big data analytics has been used for U.S hospitals [10], our project aims to use the strong correlation between alcohol, air pollution, and their negative effects to make predictions. We use alcohol, demographic information, and air pollution to make predictions for hospital data, such as the number of patients, on a county level.

III. Data Understanding

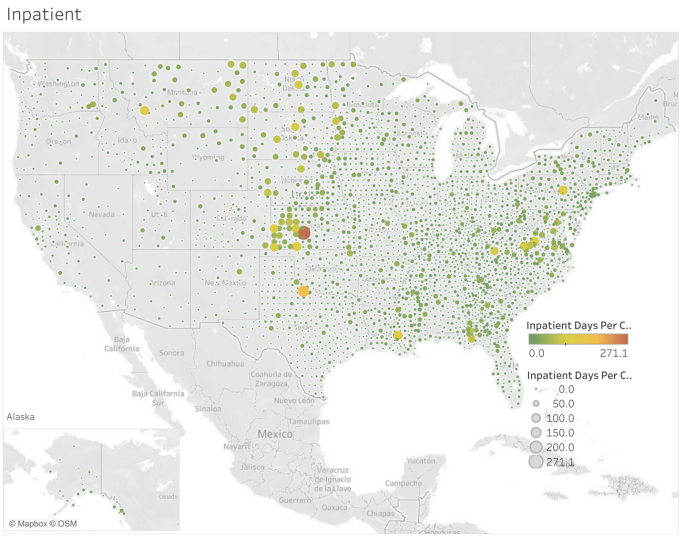

Fig. 1. A heat map depicting the variation of the number of inpatients per capita for county data.

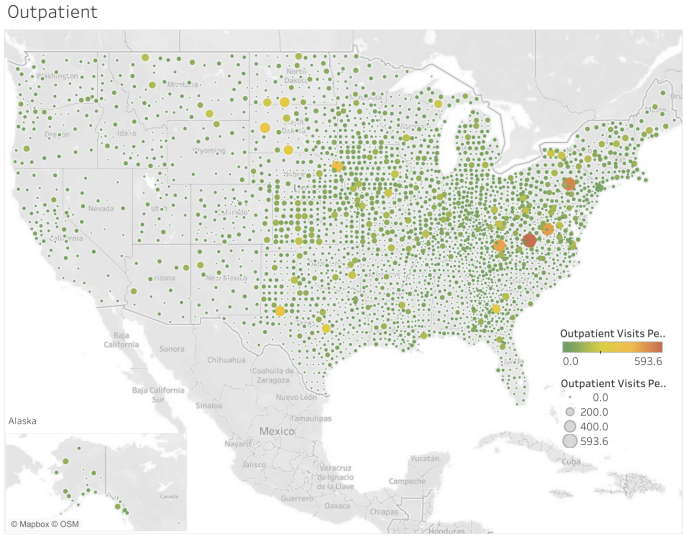

Fig. 2. A heat map depicting the variation of the number of outpatients per capita for county data.

A. AHRF Dataset

The Area Health Resources Files (AHRF) dataset contains county-level demographics, insurance coverage, healthcare supply and demand, and healthcare infrastructure information per year [11]. The dataset also contains opioid and prescription information from 2006-2014, but we omit these features in favor of extending the year coverage to 2006-2016.

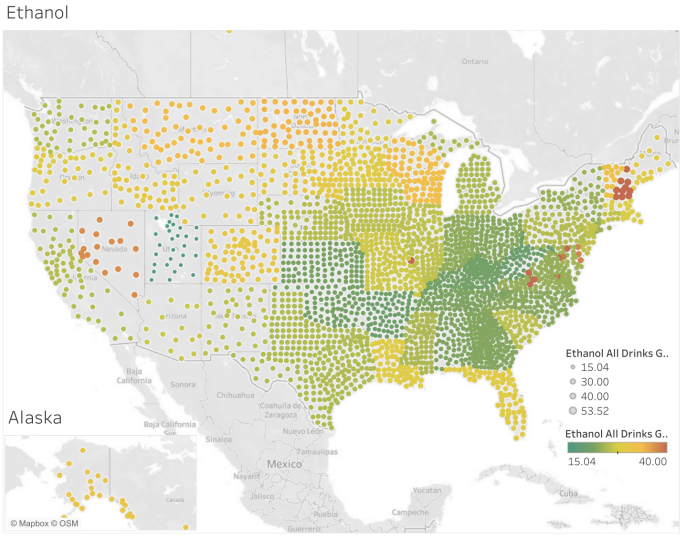

Fig. 3. A heat map depicting the variation of the amount of ethanol consumed per capita for county data.

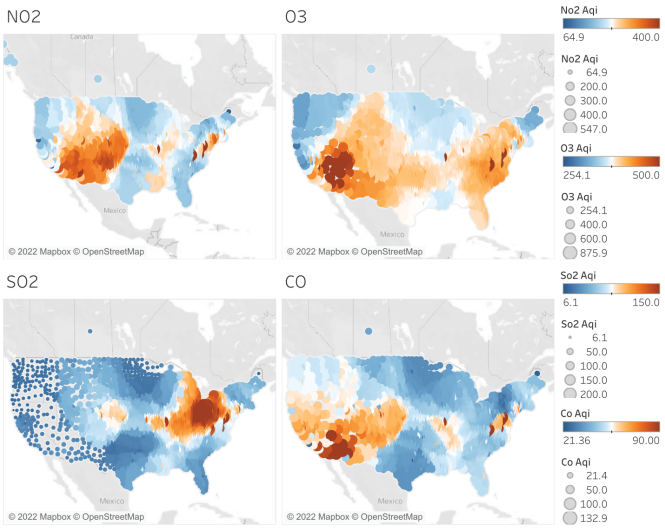

Fig. 4. A heat map depicting the AQI for CO, O_3, SO_2 and NO_2 for county data.

B. Alcohol Consumption Dataset

The alcohol consumption dataset contains state-level, per capita measurements of ethanol, beer, wine, and liquor consumption from 1977-2018 [12]. As this data is collected on the state-level, we replicate these columns for all counties in a given state.

C. Air Pollution Dataset

The time series air pollution dataset contains air sensory information measuring four primary pollutants (i.e., Nitrogen

Dioxide, Sulphur Dioxide, Carbon Monoxide, and Ozone) in United States counties from 2000-2016[1]. The air pollution data is at a county level, but has missing counties compared to the AHRF dataset.

D. U.S. County Latitude and Longitude

The dataset contains latitude and longitude coordinates for U.S. counties[2].

E. Data Exploration

Analyzing the Figs. 1 to 4, we can see that alternative data sources, ethanol, and air pollution, have a very high correlation to inpatient and outpatient data. The most important findings are as follows:

- The **Central region** consisting of Counties from Texas, Kansas, North and South Dakota, seem to have counties with the highest number of inpatients and outpatients per capita. These areas are also associated with the highest levels of ethanol consumption and air pollution.
  - **Pawnee county** in Kansas, has the highest inpatients per capita. We can observe that the area is one of the most affected by air pollution by CO, O_3, and NO_2.
- **South West region**, including Navada, Utah, Arizona and California, do not seem to follow the hypothesis. They have lower to mid level ethanol consumption, but very high levels of CO, O_3, SO_2, and NO_2. These areas also have low inpatient and outpatient numbers.
- **Eastern region**, such as counties in Pennsylvania, New York, New Jersey, Massachusetts, etc., also show a positive correlation. These areas have low numbers of inpatients and outpatients, and also have a low number of Ethanol consumption and air pollution.
  - **Montour**, in Penssylvania has a very high number of inpatients (122.4) per capita and outpatients(533.8) per capita. We can observe that the O_3, SO_2, and NO_2 levels are higher in the area.

IV. Data Preparation

A. Geographical Imputation Strategy

As our datasets are regionally dependent, we leverage this locality to impute our missing features. Typical distance metrics, such as Euclidean and Minkowski distance, are not appropriate as latitude and longitude coordinates are not in vector space as seen in Fig. 5 — in reference to this page[3]. We instead recognize that since latitude and longitude coordinates are angles on Earth, a roughly spherical surface, then we can leverage this fact to use a distance metric called the haversine formula[4] to determine distances between U.S. counties from their latitudinal and longitudinal positions. The haversine formula is rooted in navigation and finds the distance between two latitude and longitude coordinates, $p_0 = (\phi_0, \lambda_0)$ and

[1]https://data.world/data-society/us-air-pollution-data
[2]https://simplemaps.com/data/us-counties
[3]https://en.wikipedia.org/wiki/Geographic_coordinate_system
[4]https://en.wikipedia.org/wiki/Haversine_formula

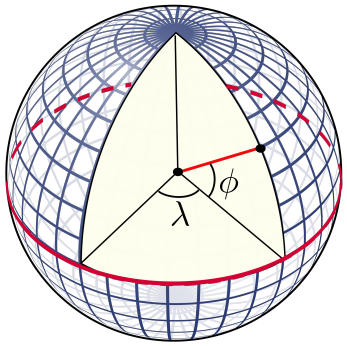

Fig. 5. An overview of how latitude ϕ and longitude λ coordinates are used to point to positions on a spherical surface.

$p_1 = (\phi_1, \lambda_1)$, by the length of the arc on the surface of the sphere between them. The haversine distance d is defined as:

$$d = 2r \arcsin \sqrt{h} \tag{1}$$

where r is the radius of the sphere and h is the haversine formula given by

$$h = \sin^2\left(\frac{\phi_1^{\text{rad}} - \phi_0^{\text{rad}}}{2}\right) + \cos\phi_0^{\text{rad}} \cdot \cos\phi_1^{\text{rad}} \cdot \sin^2\left(\frac{\lambda_1^{\text{rad}} - \lambda_0^{\text{rad}}}{2}\right)$$

with $\phi_i^{\text{rad}} = \frac{\pi}{180}\phi_i$ and $\lambda_i^{\text{rad}} = \frac{\pi}{180}\lambda_i$ being the latitude and longitude coordinates in radians respectively.

For imputation, we first find the longitude and latitude coordinates for every county using the U.S. County Latitude and Longitude dataset. Then, we iterate over each year and every county to locate either counties with a few missing features or counties missing entirely. For any unobserved feature in a county, we find the $k = 3$ closest counties that have observed the feature, given by the haversine distance, and average their values to impute the missing value. It is important to note that our missing values reside only in the air pollution dataset and are all continuous. Therefore, we believe that we can impute these values with sufficient accuracy.

B. Dataset Preprocessing

For preprocessing the AHRF dataset, we drop all correlated features that are dependent on other features, and features that contain uninterpretable processing codes. In the air pollution dataset, we aggregate the air sensor measurements and average them in order to properly merge with the other datasets. The air pollution data is aggregated for each year to match the other data. The alcohol consumption dataset contained alcohol data at the state level. Thus, this data was generalized to all the counties within a state. Finally, we manually standardize the county names in each dataset to match the format of the county latitude and longitude table so we can join them together.

C. Dataset Merging

For each of our three individual datasets, we merge them on the shared state, county, and year columns. We encode the state column using one-hot encoding, and add the state abbreviation

as a prefix to the county column before also one-hot encoding it to remove false relations between shared county names in different states. The merged dataset also contains a non-uniform number of observed counties across each dataset and even in different years of a given dataset. We use the union of counties in the merged dataset to maximize the number of rows in the final data set, and use our geographical imputation strategy for filling in the missing values.

Next, we drop highly correlated features by using the absolute value of the Pearson correlation coefficient r between all pairs of features:

$$r = \left| \frac{\sum (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum (p_i - \bar{p})^2 (q_i - \bar{q})^2}} \right| \tag{2}$$

where p_i, q_i are features of the i-th sample and $\bar{p}, \bar{q}$ are the mean of the features across the entire dataset. We construct an upper triangular correlation matrix from the N features in our dataset and drop all features with $r \geq 0.95$. Then, we randomly split the merged dataset into 80% train and 20% test.

After getting our dataset splits, we use min-max scaling to standardize the scale of our features. For each feature X_i, we scale it to a new value X_i' by computing the following:

$$X_i' = \frac{X_i - X_{min}}{X_{max} - X_{min}} \cdot (X_{max} - X_{min}) + X_{min} \tag{3}$$

where X_{min} is the smallest value of the feature column in the training dataset split and X_{max} is the largest value of the feature column in the training dataset split. Note that we compute the minimum and maximum values from the training data in Eq. (3), but we scale the features in both the training and test dataset set splits to avoid data leakage.

We establish our data as a simple regression problem rather than a time series forecasting problem due to a lack of county level data — if we were to establish a forecast for every U.S. county, then we would have 3140 datasets with 11 samples each. Instead, we aim to let a model learn from shared characteristics of U.S. counties with the intuition this will help predictive performance.

D. Feature Selection

Prior to model fitting, we rank each feature by calculating its information gain for predicting both the inpatient and outpatient continuous rates. Information gain is a metric used primarily in decision trees that measure the signal of one random variable for observing another random variable

$$IG(S) = \mathbb{E}[-\log p(A)] - \mathbb{E}[-\log p(A|S))] \tag{4}$$

with feature S and target column A. We then create two new datasets for each of the regression tasks by selecting the the top 10% of features, or the 90th percentile, according to each task's feature ranking. To avoid data leakage we only rank the features based on the train data and then apply the set of selected features to both the train and test datasets for each task, and a complete summary of the top ranked features can be found in Figs. 6 and 7.

Rank	Feature	Info Gain
1	# of licensed nursing home beds	0.273
2	# of male medical doctors	0.246
3	# of nurse practitioners	0.237
4	# of dually eligible medicare and medicaid	0.214
5	population estimate	0.204
6	veteran population estimate	0.180
7	# of home health agencies	0.175
8	asian male population estimate	0.155
9	# of hospices	0.155
10	hispanic male population estimate	0.136
11	black male population estimate	0.122
12	# of long-term hospital beds	0.118
13	# of short-term hospital beds	0.106
14	ethanol beer gallons per capita	0.094
15	ethanol wine gallons per capita	0.086
16	health profession shortage area whole county	0.086
17	rural county	0.084
18	% of population over age 25 with 4 year college degree	0.083
19	ethanol spirit gallons per capita	0.080
20	% of population black	0.075
21	longitude	0.073
22	SO_2 mean	0.072
23	ethanol all drinks gallons per capita	0.070
24	nonmetro county	0.063
25	O_3 max hour	0.053

Fig. 6. A summary of the top 25 ranked inpatient features along with their information gain. We find that demographics and healthcare facility characteristics are the most informative features which are closely followed by alcohol consumption ones.

Ranking	Feature	Info Gain
1	# of male medical doctors	0.162
2	veteran population estimate	0.141
3	population estimate	0.136
4	ethanol wine gallons per capita	0.131
5	# of nurse practitioners	0.122
6	ethanol all drinks gallons per capita	0.118
7	ethanol spirit gallons per capita	0.118
8	# of home health agencies	0.117
9	# of hospices	0.116
10	# of licensed nursing home beds	0.115
11	# of dually eligible medicare and medicaid	0.114
12	ethanol beer gallons per capita	0.107
13	asian male population estimate	0.102
14	health profession shortage area whole county	0.075
15	black male population estimate	0.086
16	hispanic male population estimate	0.084
17	% of population black	0.068
18	latitude	0.067
19	health profession shortage area part of county	0.063
20	CO max hour	0.053
21	actual per capita medicare cost	0.052
22	% of population over age 25 with 4 year college	0.052
23	% rural county	0.052
24	% of population under 65	0.050
25	SO_2 max hour	0.045

Fig. 7. A summary of the 25 ranked outpatient features along with their information gain. We find that alcohol consumption features are very informative along with healthcare facility characteristics and demographics.

We also utilize SVD for understanding the hidden concepts in our merged dataset and found that there is one singular value that retains 99% of the total energy in the singular value matrix Σ. We avoid using this to reduce the dimensionality of our dataset, however, because we do not want to reduce our data to a single dimension.

V. Regression Models

A. Linear Regression

Linear regression is a well-known model for predicting a continuous target variable from a linear combination of feature variables. The model is defined as $y_i = \boldsymbol{x_i^\intercal}\boldsymbol{\beta} + \epsilon_i$ where y_i is the continuous target variable, $\boldsymbol{x_i^\intercal}$ is the vector of numerical features, $\boldsymbol{\beta}$ is the vector of coefficients learned by the model, and ϵ_i is the irreducible noise for sample i. We aim to reduce the mean squared error (MSE) for linear regression in order to learn the coefficient vector which is defined as:

$$\mathcal{L}_{mse} = \frac{1}{N}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2 \tag{5}$$

where N is the the number of samples in our dataset split, y is the observed or ground truth values, and $\hat{y}_i$ is the predicted values from our regression model.

B. Lasso Regression

Lasso regression is an extension of linear regression that aims to improve generalizability by preventing the model from overfitting the training data. We extend Eq. (5) with a L_1 regularization penalty:

$$\mathcal{L}_{lasso} = \mathcal{L}_{mse} + \lambda\|\boldsymbol{\beta}\|_1 \tag{6}$$

where λ is a hyperparameter coefficient controlling the penalization strength.

C. Ridge Regression

Similarly, ridge regression extends Eq. (5) with a L_2 regularization penalty:

$$\mathcal{L}_{ridge} = \mathcal{L}_{mse} + \lambda\|\boldsymbol{\beta}\|_2 \tag{7}$$

where λ is the same hyperparameter from Eq. (6).

D. Elastic Net Regression

The Elastic Net regression model improves on linear regression by utilizing both L_1 and L_2 regularization as a penalty:

$$\mathcal{L}_{elastic} = \mathcal{L}_{mse} + \frac{\lambda(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2 + \lambda\alpha\|\boldsymbol{\beta}\|_1 \tag{8}$$

where λ is the regularization penalization strength, and α controls the weight of importance between the L_1 and L_2 penalty.

E. k-NN Regression

Unlike the linear regression-based models, k-NN regression is a non-parametric machine learning model that uses the inverse distance of the k nearest feature vectors as a weighted average for predicting a continuous target value. We use $\|\boldsymbol{p} - \boldsymbol{q}\|_2$ for calculating distance between two feature vectors $\boldsymbol{p}$ and $\boldsymbol{q}$.

F. Random Forest Regression

Random forest regression is a type of ensemble machine learning method that combines decision tree models such that each tree is trained on a different subset of the training data. During the training process, the model randomly selects a subset of the features to consider when making predictions at each node in the tree. When making predictions, the random forest model aggregates the predictions of all of the decision trees, and the final prediction is the average of the predictions of all the trees. This can lead to more accurate predictions than a single decision tree, as the random forest model is able to capture a wider range of relationships in the data. In our case, the decision trees are used to predict the continuous target variables for inpatient and outpatient rates.

G. XGBoost Regression

Similarily to Random Forest Regression, XGBoost Regression works by building an ensemble of decision tree models and it uses a more efficient method for training the decision trees, called gradient boosting, which involves training each tree on the residual errors of the previous tree. This allows the model to learn complex non-linear relationships in the data and make more accurate predictions, and is very popular for tabular datasets.

VI. Experiments

A. Metrics

We utilize two different metrics for evaluating our regression models. We use MSE defined in Eq. (5) to assess the error of each estimator where an MSE of zero indicates the estimator perfectly predicts our target values and a MSE greater than zero indicates imperfect predictions with a larger MSE being worse. Additionally, we record the coefficient of determination (R^2) for each model. This metric measures the proportion of the variance in the target that is predictable from the features. An R^2 of one means the model perfectly measures the variance of the target, an R^2 of zero means the model performs no better than just guessing the mean of the target values, and an R^2 of less than zero means the model performs worse than guessing the mean of the target values where a larger negative values is a worse estimator.

B. Modelling

In each experiment, we choose a regression-based model and a feature selection threshold to train with 10-fold cross validation for hyperparameter tuning on just the train data. Then we train and evaluate our models on using split validation with the train and test datasets, and we record the MSE and R^2 on the test data set as seen in Fig. 8. We find that the models predict inpatient days per capita much better than outpatient visits per capita with tree-based models performing the best. Another notable finding is that having a stricter feature selection threshold resulted in the best models. This point indicates that the 99th percentile of feature (i.e., the top 1% of features or top 32 ranked features by information gain) are the most critical for predictions.

	Threshold	Inpatient		Outpatient	
		MSE	R^2	MSE	R^2
Linear Regression	0.90	$7.039 \cdot 10^{24}$	$-6.000 \cdot 10^{24}$	$7.620 \cdot 10^{26}$	$-6.677 \cdot 10^{25}$
	0.95	$6.154 \cdot 10^{17}$	$-5.246 \cdot 10^{17}$	$2.152 \cdot 10^{24}$	$-1.886 \cdot 10^{23}$
	0.99	0.970	0.174	9.946	0.129
Ridge Regression	0.90	0.959	0.182	9.964	0.127
	0.95	0.959	0.182	9.907	0.132
	0.99	0.979	0.166	10.083	0.117
Lasso Regression	0.90	1.174	-0.001	11.424	-0.001
	0.95	1.174	-0.001	11.424	-0.001
	0.99	1.174	-0.001	11.424	-0.001
Elastic Net Regression	0.90	1.174	-0.001	11.424	-0.001
	0.95	1.174	-0.001	11.424	-0.001
	0.99	1.174	-0.001	11.424	-0.001
k-NN Regression	0.90	1.144	0.024	12.196	-0.069
	0.95	1.136	0.032	12.196	-0.069
	0.99	1.164	0.008	12.329	-0.080
Random Forest Regression	0.90	0.420	0.642	10.406	0.088
	0.95	0.432	0.632	10.557	0.075
	0.99	0.399	0.660	9.621	0.157
XGBoost Regressor	0.90	0.333	0.716	9.382	0.178
	0.95	0.336	0.713	9.310	0.184
	0.99	**0.330**	**0.719**	**8.919**	**0.219**

Fig. 8. Experimental results of the seven regression models we utilized in our experiments along with their corresponding feature selection percentile threshold, and mean squared error (MSE) and R^2 for both the inpatient and outpatient target values.

VII. Discussion

We observe several things from our experimental results. Primarily, we notice that inpatient achieves about 50% higher R^2 in the best performing model, XGBoost Regressor, which we believe is a result of a combination of the following:

- Inpatients are subject to more severe diseases that require constant hospital monitoring — such as lung cancer from air pollution or alcoholic hepatitis. Therefore, datasets that capture predictors of severe disease will make it easier to predict inpatient target values.
- Inpatients are correlated with health facility characteristics. For instance, counties with trauma centers will service a higher number of long term injuries so we assume models can extrapolate these features such that these counties have a higher inpatient visitation rate than counties without as establish health facilities.
- Outpatients have a much larger domain of potential causes and reasons to visit which is harder to capture in datasets. Although we notice that alcohol consumption features, even at state level, are beneficial as evidenced by their relative feature ranking in Fig. 7. Our intuition here is that alcohol is a common cause of a lot of outpatient visits due to its accessibility and cultural significance in the United States.

During feature selection, we found out the ranking of the features based on information gain to answer **RQ2**. It seems that demographic and hospital information had the highest rankings with alcohol also being a strong feature (more so in outpatient prediction). We notice that air pollution is not ranked highly at all, which slightly refutes **RQ1** that air pollution can be beneficial for predicting our target values.

VIII. Conclusion

In this paper, we outline the importance of predicting hospitalization usage rates for combating inefficient and inadequate healthcare in the United States. We discuss our geographical imputation strategy for filling in missing data that is geographically dependent where nuances can be lost by taking simple aggregations over the entire population of data. From our information gain feature ranking, we find that population demographic and health facility characteristic features are the most informative for making predictions for inpatient and outpatient hospitalization rates. Additionally, the latitude and longitude coordinates of counties, and the total gallons of alcohol consumed of states also capture some of the signal for the target regression variables. The county air pollution rates do not seem to contribute to the hospitalization rates consistently. Our experimentation shows that XGBoost regression model outperforms linear regression based models with Random Forest regression being a close second. We also conclude that predicting inpatients using these features is more advantageous than predicting outpatients, because of the non-specificity of outpatient features and high informativeness of features for inpatient prediction.

References

[1] W. H. Shrank, T. L. Rogstad, and N. Parekh, "Waste in the US health care system," *JAMA*, vol. 322, no. 15, p. 1501, Oct. 2019. [Online]. Available: https://doi.org/10.1001/jama.2019.13978

[2] K. W. Lemke, J. P. Weiner, and J. M. Clark, "Development and validation of a model for predicting inpatient hospitalization," *Medical Care*, vol. 50, no. 2, pp. 131–139, Feb. 2012. [Online]. Available: https://doi.org/10.1097/mlr.0b013e3182353ceb

[3] S. E. Brennan, S. McDonald, M. J. Page, J. Reid, S. Ward, A. B. Forbes, and J. E. McKenzie, "Long-term effects of alcohol consumption on cognitive function: a systematic review and dose-response analysis of evidence published between 2007 and 2018," *Systematic Reviews*, vol. 9, no. 1, Feb. 2020. [Online]. Available: https://doi.org/10.1186/s13643-019-1220-4

[4] J. N. Turk, E. R. Zahavi, A. E. Gorman, K. Murray, M. A. Turk, and D. J. Veale, "Exploring the effect of alcohol on disease activity and outcomes in rheumatoid arthritis through systematic review and meta-analysis," *Scientific Reports*, vol. 11, no. 1, May 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-89618-1

[5] L. Bujanda, "The effects of alcohol consumption upon the gastrointestinal tract," *American Journal of Gastroenterology*, vol. 95, no. 12, pp. 3374–3382, Dec. 2000. [Online]. Available: https://doi.org/10.1111/j.1572-0241.2000.03347.x

[6] E. Simou, J. Leonardi-Bee, and J. Britton, "The effect of alcohol consumption on the risk of ARDS," *Chest*, vol. 154, no. 1, pp. 58–68, Jul. 2018. [Online]. Available: https://doi.org/10.1016/j.chest.2017.11.041

[7] D. Metro, F. Corallo, F. Fedele, M. Buda, L. Manasseri, V. L. Buono, A. Quartarone, and L. Bonanno, "Effects of alcohol consumption on oxidative stress in a sample of patients recruited in a dietary center in a southern university hospital: A retrospective study," *Medicina*, vol. 58, no. 11, p. 1670, Nov. 2022. [Online]. Available: https://doi.org/10.3390/medicina58111670

[8] H. Reuter, L. S. Jenkins, M. D. Jong, S. Reid, and M. Vonk, "Prohibiting alcohol sales during the coronavirus disease 2019 pandemic has positive effects on health services in south africa," *African Journal of Primary Health Care & Family Medicine*, vol. 12, no. 1, Jul. 2020. [Online]. Available: https://doi.org/10.4102/phcfm.v12i1.2528

[9] N. A. Manan, A. N. Aizuddin, and R. Hod, "Effect of air pollution and hospital admission: A systematic review," *Annals of Global Health*, vol. 84, no. 4, p. 670, Nov. 2018. [Online]. Available: https://doi.org/10.29024/aogh.2376

[10] N. S. Downing, A. Cloninger, A. K. Venkatesh, A. Hsieh, E. E. Drye, R. R. Coifman, and H. M. Krumholz, "Describing the performance of u.s. hospitals by applying big data analytics," *PLOS ONE*, vol. 12, no. 6, p. e0179603, Jun. 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0179603

[11] K. N. Griffith, Y. Feyman, S. G. Auty, E. L. Crable, and T. W. Levengood, "County-level data on u.s. opioid distributions, demographics, healthcare supply, and healthcare access," *Data*

in Brief, vol. 35, p. 106779, Apr. 2021. [Online]. Available: https://doi.org/10.1016/j.dib.2021.106779

[12] J. Kaplan, "Apparent per capita alcohol consumption: National, state, and regional trends 1977-2018," 2020. [Online]. Available: https://www.openicpsr.org/openicpsr/project/105583/version/V4/view