SMU Data Science Review

Volume 6
Number 1 *Spring 2022*

Article 7

2022

Using Hospital Bed Capacity Prediction During COVID-19 to Determine Feature Importance

Helene Barrera
Southern Methodist University, hbarrera@smu.edu

Justin Ehly
Southern Methodist University, jehly@smu.edu

Blake Freeman
Southern Methodist University, blakef@smu.edu

Chris Papesh
University of Nevada, Las Vegas, chris.papesh@unlv.edu

Brad Blanchard
Southern Methodist University, bablanchard@smu.edu

Using Hospital Bed Capacity Prediction During COVID-19 to Determine Feature Importance

Helene Barrera[1], Justin Ehly[1], Blake Freeman[1], Brad Blanchard[2] Chris Papesh[3]
[1] Master of Science in Data Science, Southern Methodist University,
[2] Adjunct Lecturer, Master of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA
[3] School of Public Health, University of Nevada, Las Vegas, 4700 S. Maryland Pkwy, Suite 335,
Las Vegas, NV 89119
{hbarrera, jehly, blakef, bablanchard} @smu.edu
chris.papesh@unlv.edu

Abstract. The COVID-19 pandemic has exacerbated existing hospital capacity limitations in the United States, causing hospitals in certain regions to hit maximum capacity. The purpose of this study is to investigate key features of COVID-19 related admissions to help create a higher level of public understanding and help guide healthcare management professionals and governments when considering preventive measures. The introduction of preventative measures and new regulations during the pandemic have led to the generation of multiple types of models and feature selection methods in the field of Machine Learning that are increasingly complicated. This study focuses on the exploration of feature selection through building multiple models, one simple linear model and one decision tree model for prediction on inpatient hospitalization rates. This will result in a highly interpretable model that can be more readily understood and easily used.

1 Introduction

With the Coronavirus pandemic (COVID-19) that begun in late 2019, scientists from many disciplines have been scrambling to accurately model the impact and scope of the pandemic. The reality of COVID-19 has had far-reaching consequences on everything from healthcare systems to work-from-home business practices to supply chain management. A particular area of alarm is the increasing multitude of hospitals reaching maximum capacity for inpatient beds and ventilators, even following the introduction of a vaccine. Understanding the factors at play behind these capacity issues

is imperative to proper management of this crucial resource, which should result in fewer preventable deaths.

The inpatient and ICU capacity issue is not a brand new one, even in times when there is no ongoing public health crisis such as COVID-19. Hospitals have historically been slow to adjust and adopt to changes in demand caused by normal factors such as changing population demographics, prices, and insurance coverage for various services, and even change in market share from other hospitals shutting down [1]. Additionally, hospital capacity is not limited only by the number of beds and physical space, but also by the number of healthcare professionals who are qualified to issue that level of care.

Traditionally, hospitals aim for around an 85% occupancy rate, although in recent years there has been concern in the United States about excess beds causing healthcare costs to rise [2]. According to the New York Times, "experts say maintaining existing standards of care for the sickest patients may be difficult or impossible at hospitals with more than 95 percent occupancy" [3]. It is important to balance being cost-efficient with having enough extra beds to provide care for unpredictable disasters such as a worldwide pandemic. There are many existing models and methods for calculating the optimal number of beds a hospital should have, given a certain set of factors, but the methods are not standardized and vary by region and country [4]. A more complex model for hospital management may account for additional beds or contingency plans needed due to disease spread. Additionally, public health policy can also be used to manage the total number of hospital admissions from COVID-19 to avoid reaching maximum bed capacity.

As stated previously, capacity is not limited by the number of beds alone – and COVID-19 has brought existing healthcare supply chain management issues into sharp focus. PPE (Personal Protection Equipment) is vital for healthcare professionals to perform their jobs in a safe manner for themselves and their patients. However, due to panic buying and hoarding, necessities like gloves and masks can become expensive during a national crisis and can take months to arrive, putting healthcare workers at risk and making it more difficult to fully staff hospitals [5]. The COVID-19 pandemic has also shown that the United States needs to reconsider its reliance on products produced in other countries and supplies warehoused offshore, due to the potential of border closures in an emergency [6]. Although a national emergency healthcare supply system is in place, known as the Strategic National Stockpile (SNS), as well as individual state stockpiles, those are meant to be supplemental and they are not necessarily as modern and reliable as they could be [7, 8].

Lastly, capacity in hospitals is limited by qualified healthcare professionals. Like the other factors that led to hospital bed shortages during COVID-19, a shortage of registered nurses is not new or unique. In fact, the current shortage was forecasted to last through 2030 as early as 2012 using projected changes in demographics [9]. Shortages can be caused by an aging workforce, lack of nursing students entering the workforce, and an aging general population that requires more care [10]. Shortage of personnel is particularly challenging to the healthcare system because contingency plans for increasing personnel shortages involve solutions like bringing workers out of

retirement and recruiting educators, students, and even non-clinical staff, none of which are ideal options [11].

All these factors lead to the reality that the capacity of hospitals in the United States was already limited prior to COVID-19 and is now heavily strained (in some regions more than others). Public policy measures including mask mandates, social distancing, vaccines, travel bans, and stay-at-home orders have been shown to alleviate daily reported COVID-19 cases [12]. If governments and healthcare management professionals can forecast bed shortages taking COVID-19 related admissions into account, they can potentially draw on such public policy measures to prevent hospital systems from hitting full capacity and being forced to triage and turn patients away.

While there are many models for predicting and analyzing COVID-19 infection and death rates, this study looks to utilize simplified and highly interpretable models to identify feature importance relating to COVID-19 inpatient hospitalizations. By identifying these key features, the expectation for this model is to act as a high-level litmus test for use by healthcare management professionals and local governments to monitor and decide what measures would be the most effective to put in place to reduce the impact on hospital capacity. Although the study will be conducted using data obtained in the United States and focused around COVID-19 inpatient hospitalizations, it may have indications for other countries and future airborne viral outbreaks, and it may be useful for planning future ICU and specialized facilities (negative pressure environments) and warehousing models.

2 Literature Review

Understanding the ongoing COVID-19 pandemic and preparing for future pandemics on a global scale is paramount to protecting the modern world's peoples and economies. In 2020 and 2021, there were global lockdowns and supply shortages across all industries whose impacts are still affecting global markets today. In the US, hospitals are being tested as the influx of COVID-19 patients pushes the capacity of inpatient and ICU wards to the limit. It is not only the physical number of beds that is responsible for the capacity issues – ICU wards need medical professionals with certain skill sets to run the specialized equipment – and the equipment itself could be in short supply as global supply chain networks break down [13]. For example, specialized medical equipment could be sitting off the coast of Long Beach on a ship from Asia, unable to be unloaded and used. Medical professionals with the training to work in ICU units may have fallen sick with COVID-19 themselves or experienced burnout and left the medical field altogether.

Reviewing literature pertaining to current COVID-19 prediction models to build on those develop a sense of what the most important predictive features are will help guide this research This study also reviews feature selection methods that are currently used with COVID-19 models to build a further understanding of key features. Further, the scope of this research focuses on the United States and hypothesizes that by creating a mean of the top 3 performing states with the bottom 3 performing states in terms of

COVID-19 mitigation as well as un-recovered COVID-19 patients, the best features will be established.

2.1 Existing COVID-19 Data Models and Calculators

There is ample and ongoing research into data models to predict infection rates, supply and demand for person protection devices, hospital beds and more [14]. The most common epidemiology models observed in research for this paper are the Susceptible-Infection-Recovery (SIR) with variants such as Susceptible-Exposed-Infection-Recovery (SEIR) and the Susceptible-Laten-Infected-Recovered (SLIR) models [15]. There are also three exceedingly popular surge calculators that are widely used to estimate the need for hospital supplies et al [16] and they are CHIME (The COVID-19 Hospital Impact Model for Epidemics) from Penn Medicine [17], WHO COVID-19 ESFT (essential supplies forecasting tool) [18] and COVID-19 AUBMC [19].

SIR models are over a century old and are widely used to predict the effects of disease [20]. They are a class of compartmental models where the population moves from the front compartment to the end in a linear fashion, for example a person or population is susceptible to a disease, then infected and eventually recovered or removed. The variants introduce additional compartments such as exposed, laten or in some instances, susceptible again.

Interesting work done with SIR models involves reducing them to basic logistic regression equations, as was done by Eugene B. Postnikov (the Verhulst equation) [21]. Postnikov concluded that overall, the reduction to a logistic equation on a country-by-country basis provided accurate predictions of infected people with exceptions noted for countries that utilized highly mitigated strategies to reduce infection rates among their populations. This study contributes well to this research where the goal is to extract feature importance from highly interpretable models such as logistic regression.

Another method of predicting new cases of COVID-19 was presented by Deschepper et al. The research team utilized an "…additive Poisson model with a penalized regression spline for calendar time," (Deschepper et al 2021, p. 4) for prediction and then fed those predictions into a multistate model for further analysis [14]. Since Poisson is a generalized linear model, it is also highly interpretable and provides additional context into just how simple a model may be to predict COVID-19 and related parameters.

An additional take in data modeling sought to explore non-pharmaceutical interventions (NPI) to control or reduce the spread of COVID-19 utilizing a partial differential equation (PDE) based SLIR model [22]. The model was fit and validated against actual COVID-19 data over a specific amount on time in Ohio (30-days testing data, 15-days prediction data) using NPI controls in the form of social distancing and mask mandates. This allowed the team to analyze any additional efficiencies produced in the predictions of how the virus would spread. The results proved that the NPI controls in fact would reduce the spread of the virus and the authors felt the model provided a benchmark for further investigation using data from other than Ohio.

While SIR models prove to be a standard go-to for predicting infection rates and/or viral distribution among populations, there is research that suggests combing SIR

models with other tools provides a more robust framework that can be inclusive of mitigating factors such as, “time-dependent human behavior.” This was presented by Arazi et al. [23] where an SIR model was appended with an optimal level of social distancing to build a utility function. While their model was kept relatively simple, it does provide for additional complexity in iterative versions to perhaps include factors beyond just social distancing. The final model allows for predicting pandemic waves.

2.2 Existing Methods of Feature Selection

Like many other subjects in information science, COVID-19 has an inherent problem of dimensionality that affects machine learning models due to a multitude of redundant or irrelevant features [24],[25]. This study looks at certain feature selection practices that have been applied to COVID-19 such as Hyper Learning Binary Dragonfly Algorithm (HLBDA) [24], Matrix Factorization-based methods [25] and utilization of Random Forest, Gradient Boosting and XGBoost [26].

Hyper Learning Binary Dragonfly Algorithm (HLBDA) is continuation and expansion of the Binary Dragonfly Algorithm (BDA), which looks for personal best and personal worst solutions on an individual level. The study on HLBDA compared eight different feature selection methods with five different evaluating metrics with a repetition of 20 times. Based on the findings the study HLBDA performance was superior in most cases. The study continues to state that when applied COVID-19 models HLBDA showed excellent performance in predicting patient health [24].

The Matrix Factorization-based methods study focused on comparing Matrix Factorization Feature Selection (MFFS), Maximum Projection and Minimum Redundancy (MPMR), Sparse and Low-redundant Subspace learning-based Dual-graph Regularized Robust (SLSDR) and Regularized Matrix Factorization Feature Selection (RMFFS). The study used Clustering Accuracy (ACC) and Normalized Mutual Information (NMI) as the evaluation metrics. Based on findings in the study initial SLSDR out-performed other methods on another dataset, however when applied to the COVID-19 data there was no clear superior model for feature selection when comparing models as the number of features being selected increased [25].

The study that utilized Random Forest along with Gradient Boosting and XGBoost, was done to calculate the reproduction rate of Corona virus. Random Forest was used along with Gradient Boosting and XGBoost to create features that reduced impurity as well as identified and boosted the weak learners. The study utilized this method to try to increase the generalization that could be derived from this model. This study used Root mean square error, mean absolute error, Determination coefficient, Relative absolute error, and Root relative squared error as performance metrics in evaluating the performance of features. In the end it was determined in the results of this study that the methods used showed a decrease in the prediction error rate for the feature selection than other tested models [26].

2.3 Localization of Data Models

While generalized data models all draw from the same basic concept of the SIR model, the application of those models to localized populations have the propensity to show that different features exhibit different importance [27, 28]. For example, people over 60 living in a densely populated area with a significant percentage of the population working and/or going to school full-time (on-site) have a higher risk of infection or severe illness than people under 60 in the same conditions [27]. Likewise, rural populations with low high-school completion rates, denser minority populations and people with existing or multiple existing medical conditions tend to be at higher risk for COVID-19 infection than more affluent, less diverse populations [28].

Another set of features that was studied on a smaller scale, was the effect of human behavior using a movement model on the spread of disease, specifically relating to commuting from home to the office/classroom and controlled interpersonal contact within those environments [29]. If people use mass transit, then there is the notion that randomness of contact can induce viral spread. However, if people have their work schedules set up to rotate personnel to avoid interpersonal contact and reduce the randomness in mass transit (similar people should be commuting together on a regular basis), then there is a good chance at mitigating pathogen spread.

3 Data

There are multiple sources of data for this analysis. The first data set provides information on government stringency measures in response to COVID-19 globally on a daily basis. The second data set quantified domestic patient impact and hospital capacity on a weekly basis. The third data set was from the Immunization Action Coalition and provided dates of when different COVID-19 vaccinations were approved. The fourth and fifth data sets were gathered from the CDC website and provided flu season dates as well as daily COVID-19 vaccination doses administered by state and percentage of population vaccinated.**3.1 Oxford COVID-19 Government Response Tracker**

The Oxford COVID-19 Government Response Tracker (OxCGrT) was introduced by the University of Oxford starting with January 1, 2020 [30]. The tracker tracks global government response measures across 23 features that are scales or raw dollar amounts based on specific government actions to provide a unified way for anyone interested in understanding individual government responses to the pandemic.

The data is collected by volunteers globally and continuously updated for public consumption. The 23 features or indicators are grouped as follows: (8) indicators about information on government containment and closure policies, (4) indicators on economic policies both internally such as supporting citizens and externally such as foreign aid, (8) indicators on health system policies that includes vaccinations and testing and finally (3) indicators on vaccine policies specifically. Additionally, some

indicators have a binary weight attached called a Flag that corresponds to the scope of the indicator such as geography or individual vs government funding of vaccinations.

The indicators and associated flags are then used to compute sub-indexes which are then used to produce policy indices. There are four policy category indices which is a combination of policy's, the indices are as follows: overall government response index that documents how governmental response and uses all containment, economic and health system indicators, containment and health index combining containment and health system indicators, stringency index that combines all containment and the health care indicator pertaining to public information campaigns and finally, economic index that combine the economic indicators.

3.2 COVID-19 Reported Patient Impact and Hospital Capacity by Facility

The COVID-19 Reported Patient Impact and Hospital Capacity by Facility is a publicly available dataset from HealthData.gov [31]. The dataset is built from facility-level data and aggregated on a weekly basis with weeks beginning on Friday and ending on data Thursday.

Hospital data comes from all facilities registered with Centers for Medicare and Medicaid Services (CMS) as of June 1, 2020. The data also includes non-CMS facilities that began reporting on June 15, 2020. The data does not include Veterans Affairs (VA), Defense Health Agencies (DHA), Indian Health Services (IHS), religious non-medical hospitals, psychiatric, or rehabilitation facilities. The main data sources are the US Department of Health and Human Services (HHS) TeleTracking and information reported to HHS Protect by state health departments on behalf of their healthcare facilities.

3.3 Immunization Action Coalition Website Data

The Immunization Action Coalition website [32] provides a comprehensive list of press releases from the ACIP (Advisory Committee on Immunization Practices, CDC (Centers for Disease Control and Prevention) and FDA (Food and Drug Administration) organized by date. This provided dates for official FDA authorization of different COVID-19 vaccinations.

3.4 CDC Website Data

The CDC or Centers for Disease Control and Prevention website [33] provided data on the months of Flu Season as well as [34] daily COVID-19 vaccination doses by state broken down by first dose and fully vaccinated people in the United States.

3.5 Merged Dataset

To combine the datasets, the Oxford data was reduced from daily data to weekly data with weeks starting on Fridays and ending on Thursdays and dates ranging from

Friday, July 10, 2020, to Thursday, September 16, 2021. The data was further reduced to only included the 50 states and Washington, D.C. in the United States.

The patient and hospital data were reduced to just include dates from July 10, 2020, to September 16, 2021, and only utilized the 50 states and Washington, D.C. in the United States.

The datasets were combined using the state and reporting week as joining keys allowing for a dataset that includes both government stringency measures and the resulting patient impact and hospital capacity.

Next the Pfizer vaccine authorization and Flu Season data were added to the merged data set as binary variables with a zero (0) representing a no or not in season and a one (1) denoting a yes or in season and the weekly dates were used as the join key. It should be noted that at the time of the data source creation, the Pfizer vaccine authorization was the only date available. As the project moved over time, adding the additional Moderna and J&J vaccine dates of authorization were overlooked.

Last, the first dose and completed vaccination daily information was combined into weekly information starting on Friday, July 10, 2020, and ending on Thursday, September 16, 2021, as the total number of doses given to date and total percent of population using the state and week date as join keys.

3.6 Additional Data

Additional data that might be useful is benchmark data on hospital occupancy and supply and staffing shortages prior to COVID-19.

4 Methods

4.1 Evaluations Metrics

Since the goal of the project is to utilize highly interpretable models, r-squared and adjusted r-squared are the best metrics to determine how well a model is representing the data. R-squared is calculated by dividing the variation of the target or dependent variable by the variation in the independent variables (or the features). Adjusted r-square penalizes the r-square value based on the number of features utilized in the model.

4.2 Linear Regression Model

The feature selection method utilizing linear regression that is used in the modeling stems from using relative importance metrics estimates.[35] This approach uses a linear regression model that predicts future COVID-19 hospitalizations. From this model the features were sorted by importance's based on the effect of each coefficient. This was to simplify the results and make them explainable with the hopes of finding features that could be interpreted into actionable items.

4.3 Random Forest Regressor

While linear regression is highly interpretable, it does not offer any tuning parameters that are available with more complex models. Random Forest Regressor models from Sci-kit Learn offer seventeen tuning parameters and to find the best combination of those a grid search was used to test 1,620 different combinations of tuning parameters. The tuning parameter values were randomly chosen and included the default parameter values. The best r-squared score determined the best tuned model and that was compared to a base-line model to ensure there was an increase in the model's ability to represent the data.

Once final models are determined, a second pass will be performed using a sliding window of time to see how well the best performing model generalizes the data using six training weeks and two predicted weeks.

4.4 Exploratory Data Analysis (EDA)

Using the merged data set described in Section 3.5, exploratory data analysis (EDA) was performed. EDA provides researchers the ability to become intimately familiar with the data, explore relationships between and among distinctive features and apply any material information provided by subject matter experts/ advisors.

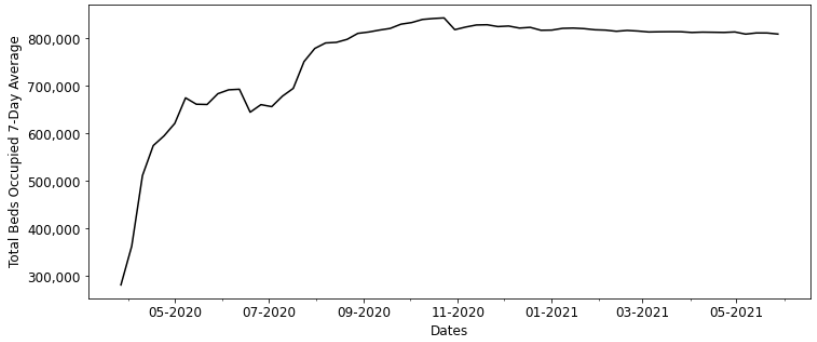

Fig. 1. Total 7-day average of all staffed inpatient and outpatient beds in reporting hospitals, including inpatient and outpatient services for ICU, ED, and observation. This figure shows the dramatic increase from the time hospital records were recorded across all states, Jan. 24, 2020, through Nov. 2020 and a steady plateau since.

The EDA process was dynamic and consisted of multiple iterations over the modeling process that resulted in additional changes and expansion of the dataset. This stemmed from working with subject matter experts that have first-hand knowledge of medical practices and how to conduct medical research. Initial feature extraction underwent a correlation matrix to visualize and identify where there were features that were correlated with each other more than ninety percent as shown in in Figure 2.

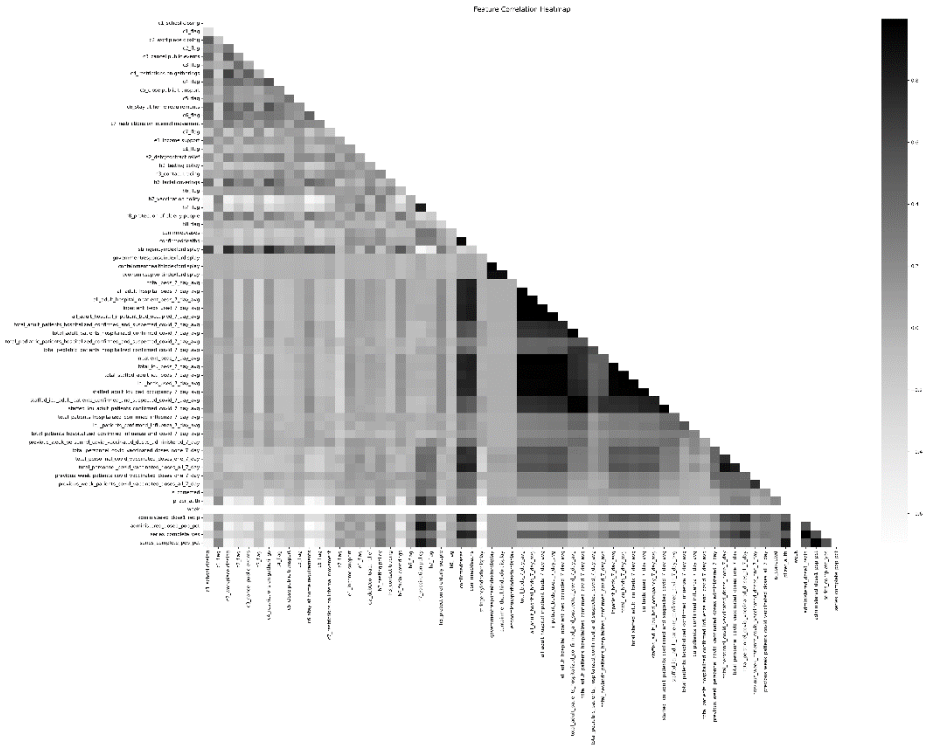

Fig. 2. Correlation Matrix Heatmap of the full dataset to determine where there were features that could be removed due to correlation above 90%.

Another process of elimination from the data set was how much variability existed within each discrete feature from the Oxford data because features with a lot of variability tend to be emphasized by the model compared to features with low variability as was reflected in the low weight scores in the random forest regressor. Removing variables that either did not change or changed very little or had minor impact on the overall change of other variables in the model were deemed unnecessary and removed. This was done visually using heatmaps such as the one shown in Figure 3 below where we can see that restrictions on gatherings have a lot more visual variance than that of public information campaigns in Figure 4, which have no variance. This was performed on all the Oxford data.

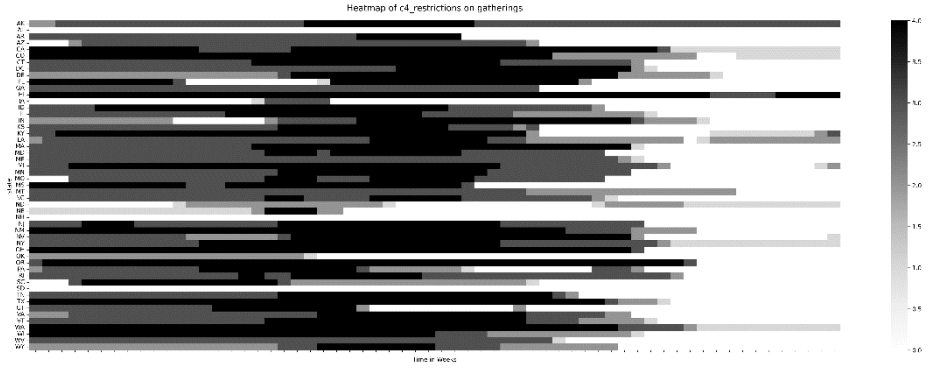

Fig. 3. Heatmap of data by state and week of Government Restrictions on Public Gatherings that were rated based on either no data or data based on a tiered approach (none, more than 1,000 people, 101-1 ,000 people, 11-100 people, less than 10 people).

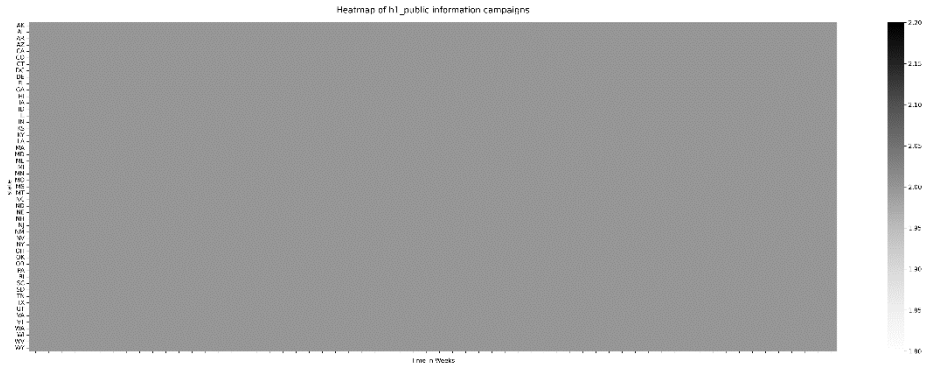

Fig. 4. Heatmap of data by state and week of Public Information Campaigns that were rated based on no data or data based on a tiered approach (none, public officials warning people about COVID-19, media campaign about COVID-19).

Healthcare.gov data was largely reduced due to high correlations, a thorough analysis of the description of each feature and how it contributed to the scope of the project. From this process most of the healthcare.gov data was deemed too highly correlated and was reduced to just 4 features that provided guidance on total hospital bed occupancy and COVID-19 hospital bed occupancy as illustrated in Figure 5.

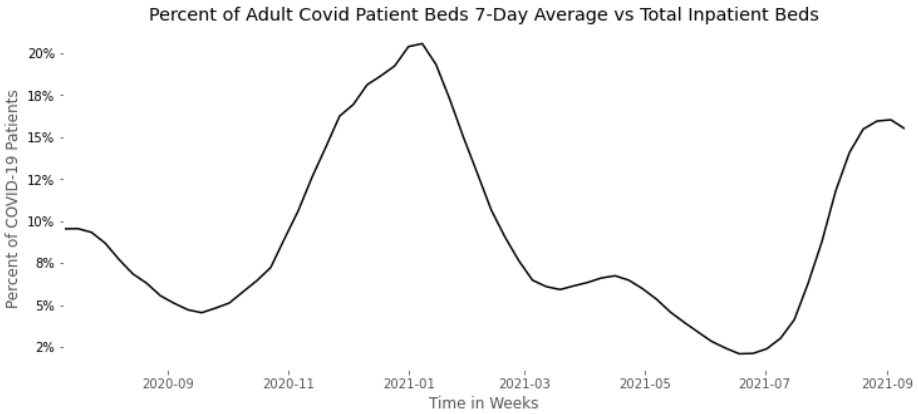

Fig. 5. Chart showing the percentage of occupied inpatient hospital beds by adult COVID-19 patients within the fifty states and Washington, D.C out of total staffed inpatient beds, based on a 7-day average from the healthcare.gov data.

The final piece of the data set was also included on the advice of our advisors and that is the number of vaccinated people by week based on the CDC data. Initially we pulled data for initial doses given and complete vaccinations, but found those values were correlated greater than ninety percent so only the weekly number of initial dose data was utilized.

5 Results

5.1 Models

The linear regression model which was the initial model used preformed with an r-squared score 0.685 which implies that only that it able to capture approximately 69% of the variance in the data. The coefficients for the linear regression model are in Table 1 with the general formula is written the standard way below:

$y = b_0 + b_1 x \ldots b_p x$ where $b_0 = 1340.061$

Features	Coef
Inpatient Beds 7 Day Avg	1386.067
Total Pediatric Patients Hospitalized Confirmed	760.0623
H7 Flag (Free vs paid vaccination policies)	549.444
Administered Dose1 Receipt	-459.748
H7 Vaccination Policy	-338.629
C1 Flag	-134.79
Stringency Index	92.15131

C5 Close Public Transport	-90.0622
Is Corrected	71.23588
H6 Facial Coverings	-67.3652
H8 Flag	-60.9037
E2 Debt/Contract Relief	-58.4908
H2 Testing Policy	-57.0846
C1 School Closing	-52.9401
C3 Flag	-51.6384
C7 Flag	-48.6757
C6 Flag	-48.2369
C4 Flag	48.00553
H6 Flag	-46.868
C4 Restrictions On Gatherings	46.72329
C2 Workplace Closing	-41.1985
C6 Stay At Home Requirements	38.70962
E1 Flag	-34.2403
Government Response Index	31.07531
H3 Contact Tracing	30.02181
E1 Income Support	29.01078
Pfizer Auth	23.80163
C7 Restrictions On Internal Movement	18.92049
C2 Flag	-11.1821
Total Personnel Covid Vaccinated Doses None 7 Day	-9.04379
C3 Cancel Public Events	5.719272
H8 Protection Of Elderly People	5.188044
C5 Flag	-4.24398
Week	0

Table 1. Chart showing the percentage of occupied hospital beds by adult COVID-19 patients within the fifty states, plus Washington, D.C.

To interpret this model further in relation to feature importance. It would stand, that prior cases of COIVD-19, total beds, percent of vaccinations distributed, and vaccination policies have great effects in the linear model. Since the model only

produces a 0.685 r-squared score other models were tested to see if better results could be achieved.

The base line random forest regressor was based on the default parameters and yielded an r-square of 0.848, considerably better than the linear regression model. The tuned random forest regression model performed with a r-squared score of 0.857 effectively capturing approximately 86% of the variability present in the data. Table 2 shows the model parameter tuning and results.

Tuned Random Forest Regressor Model	
model	RandomForestRegressor
target	total_adult_patients_hospitalized_confirmed_co...
bootstrap	TRUE
ccp_alpha	0
criterion	absolute_error
max_depth	5
max_features	auto
max_leaf_nodes	None
max_samples	None
min_impurity_decrease	0
min_samples_leaf	2
min_samples_split	2
min_weight_fraction_leaf	0
n_estimators	100
n_jobs	-1
oob_score	FALSE
random_state	42
verbose	0
warm_start	FALSE
sd	2046.370535
var	4187632.365
mean	1021.95213
r2_score	0.857266
adjusted_r2	0.85519

Table 2. The parameters and model statistics from the tuned RandomForestRegressor model

The tuned random forest regression model (r-square = 0.857) outperformed the linear regression model (r-square = 0.685). There was also significant difference in the

predictions from the linear model (mean = 695.031, standard deviation = 1582.712) and the tuned random forest model (mean = 1021.952, standard deviation = 2046.371); t(1484.182) = 3.5496, p-value = 0.0004, 95% confidence interval [146.26 , 507.58], Cohen-d = 0.179 as shown in Table 3.

T-test Results	
T	3.459631
CI95%	[146.26, 507.58]
Cohen's-D	0.178601
BF10	28.358
Power	0.943801

Table 3. T-test statistics comparing the predictions from the Linear Model to the tuned RandomForestRegressor model

To ensure the random forest regressor was generalizing on the data, a sliding window of 6 training weeks to predict 2 future weeks was implemented where we saw model performance drop by 10% to a mean r-squared value of 0.745, mean adjusted r-square value of 0.742, that is still out-performing the basic linear regression model.

5.2 Results

Evaluating the results of the random forest regressor in the scope of the research, what government actions were the most correlated with COVID-19 hospitalizations, the three most highly correlated individual government actions are as follows: stay at home requirements (weight = 1.028), school closings (weight = 0.844 and testing policy (weight = 0.627). Combined these 3 policies only make up 2.5% of the all the weights for the model. The overall largest weights were the more directly correlated hospital data itself followed by the counts for the first vaccine dose and then the overall government index and the stringency index. We do see some of the flags appearing in the top of the weights, but we attribute that to a higher degree of variability as discussed in section 5.1.

6 Discussion

6.1 Discussion

The results of the study indicate that vaccinations and vaccination policies that made vaccines more available were the most highly correlated attributes in reducing hospitalizations due to COVID-19 within the scope of this study. Government policies targeted at reducing the spread of the disease also appeared to be the most highly correlated attributes to a reduction in hospitalizations. However, considering the enormous impact the policies could have on the economy and society, the correlation of the policies on hospitalizations was somewhat small. There are a few reasons why that may be, some of which are due to limitations of this study.

While the vaccines were not in line with traditional vaccines that underwent years of study and results, they did provide the highest level of correlation with the reduction of COVID-19 hospitalization. This vaccination was able to achieve this rapid development though benefiting from past research, recruiting volunteers, getting fast results, and jumping the line for approval. [40]. The vaccines also initiated divisive political debate. However, whether people were for or against them, the research shows they were the most highly correlated with reducing the disease's effect on hospitalizations due to COVID-19.

The findings in this study indicate that government isolation measures when evaluated on a combined basis were correlated with a reduction of COVID-19 hospitalizations. While we the researchers had to self-isolate in accordance, and it had psychological side effects such as loneliness and cabin fever, the end result was that the citizenry as a whole benefitted by not being hospitalized or dying from COVID-19.

When reviewing this study, government isolation measures that are a combination of the isolation policies and the vaccination itself had an effect in reducing COVID-19 hospitalizations. The recommendation based on this study is if a similar variant or COVID-19 like virus were to spread once again it would be effective to implement these government isolation policies, while trying to expedite a vaccination to have a reduction in COVID-19.

To think of these results in a broader context, we saw the establishment of advanced timetables in vaccine development coupled with political weaponizing of uncertainty in science. To prepare for the next outbreak, there needs to be a more coherent distribution of information to encourage trust in science.

6.2 Possible Steps to be taken

This study utilized past COVID-19 hospitalizations to predict current COVID-19 hospitalizations along how much of the percentage of the population had an administered dose. This can develop insight into where hospitals can try to further expand facilities based on the current outbreak. In addition, if another pandemic is transmitted or spreads similarly to COVID-19 similar recommendations or even harsher regulations around vaccinations could be put into place to try to prevent COVID-19 hospitalizations. Although our study uncovered little impact from most government policies, it does point to vaccine policy being effective in reducing hospitalizations.

6.3 Ethical Implications

While this study does not seek to directly examine the topic of ethics, there are several ethical issues surrounding the topics of COVID-19, vaccines, and government responses to the pandemic.

These issues have become political topics, and many people, including researchers, may hold pre-existing political beliefs around these issues. Ethical issues can become apparent when researchers either consciously or not impart their beliefs into how a study is run or which model is chosen in a way that is biased to reflect their beliefs and

disprove others. This practice of bad science can have further ethical implications on public trust and support of scientific studies.

On the other side of the issue is that even when a study is conducted ethically it may have a negative reception from the general public – either because the results go against deeply rooted beliefs, or because the science behind the results is not well understood.

The paper written by Kreps, & Kriner, D. L. warns of some of the relevant ethical implications specifically for COVID-19 and public health research. The study shows that the way scientific studies are communicated to the public matters, and that introducing uncertainty surrounding model results can sew distrust as to the validity of the study [36]. This is of interest to our study, since the results, while statistically significant, leave some accuracy to be desired. The research performed here intended to find correlations between specific government actions tracked by Oxford and hospital bed occupancy across the USA. The tuned random forest regressor model while accurate by statical standards, does have inaccuracy that could impact the way the general public views and understands this study.

This paper shows that vaccinations and social isolation have the highest correlations with the number of COVID-19 hospital bed occupancies. The impact of other government policies is unclear from the study, but that doesn't necessarily mean that those policies are ineffective or overreaching policies. Our study used two fairly simple models and the factors at play with the interaction of government policies on COVID-19 spread and hospitalizations are extremely complex. The lack of strong and clear results in these areas could have an impact on public trust in COVID-19 studies in general which is of ethical concern.

6.3 Challenges and Limitations

A major limitation of this study was the subjective nature of the Oxford data set which provided most of our features. The public data was gathered and entered by volunteers at a global scale. The data were somewhat subjective as it was not always clear cut which category and level a policy may fall under, and often the realities of implementation, enforcement, and compliance were very different from the written language of the policy.

Meanwhile, the healthdata.gov data is derived from HHS TeleTracking and self-reported by state health departments. CDC COVID-19 vaccination administration data is collected from immunization information systems (IISs), Vaccine Administration Management Systems (VAMS) and from the COVID-19 Data Clearing House.

The data utilized could have been more comprehensive to include additional data points to provide more correlation between government responses to COVID-19 and hospitalization resulting from critical COVID-19 infections. The data could have also been more geographically granular to go below the state level to the county, city and smaller geographic levels which could have displayed more individual variances and drawn a bigger picture. This would have also allowed for more tailored actionable items based on the feature selection process.

6.4 Possible Applications for Future Works

Through modeling this dataset and looking for the features that were most highly correlated to the target of adults with COVID-19 occupying hospital beds there are some key takeaways that can be applied toward future works.

Vaccinations, the number of people receiving at least the initial dose of a COVID-19 vaccine is the highest correlated feature that was not derived from the healthcare.gov data. That suggests that vaccinations were affecting the outcome of the number of adults patients occupying beds due to COVID-19, by adding additional data over a longer period of time it is possible that feature would become more highly correlated.

Vaccinations can also be tied directly back to government actions tracked by Oxford and in fact, the h1_vaccination_policy was the 8th highest ranked weight in the tuned model suggesting that government actions were beneficial in mitigating COVID-19 hospitalizations.

Suppressing social interactions is also suggested by the model to be correlated with reducing stress on the healthcare system relating to COVID-19 judging from the fact that out of 34 overall features in the model stay at home requirements, school closing, canceling of public events and restrictions on gatherings ranked sixth, ninth, eleventh and thirteenth. While closing of public transportation was not ranked high in the feature list, it is estimated that about five percent of the American workforce relied on public transportation in 2019 based on the American Community Survey (ACS) [37]. The percentage of commuters relying on public transportation varies widely from region to region where in very large cities such as New York and San Francisco saw a many as a third of the workforce utilize public transit. If this study were carried out at a more granular level, that feature may play a more prominent role in the model depending on the geographical focus of the research.

Facial coverings did not appear to play a more prominent role in the model with a ranking of twenty-second of thirty-four in importance. That finding is interesting because there was such a push from governments and school boards to require face coverings that often was juxtaposed with members of the community vehemently opposing them, often putting local business at odds with customers [38] and school boards at odds with parents [39].

Overall, the modeling of the covid data gathered for this corpus of research provided interesting insights into which government actions, vaccinations and reducing face to face human interaction, may be the most important for community leaders to focus on in the event another pandemic strikes.

7 Conclusion

The model's findings showed that utilizing past hospitalizations as a feature to predict future COVID-19 hospitalizations, the total inpatient beds available, and vaccination policy to have high feature importance. At the same time, it should be noted that limiting social interactions also provided relevance to adversely impact the effects

of the pandemic. Based on the modeled data, vaccination policy had the highest correlation while a combined approach to limiting social interactions was the second most correlated policy to COVID-19 hospitalizations of adults on a state-by-state basis.

References

[1.] Bazzoli, G. J., Brewster, L. R., May, J. H., & Kuo, S. (2006). The transition from excess capacity to strained capacity in U.S. hospitals. The Milbank quarterly, 84(2), 273–304. https://doi.org/10.1111/j.1468-0009.2006.00448.x

[2.] Ravaghi, H., Alidoost, S., Mannion, R., & Bélorgeot, V. D. (2020). Models and methods for determining the optimal number of beds in hospitals and regions: A systematic scoping review. BMC Health Services Research, 20(1). https://doi.org/10.1186/s12913-020-5023-z

[3.] Conlen, M., Keefe, J., Sun, A., Leatherby, L., & Smart, C. (2021, June 9). How Full Are Hospital I.C.U.s Near You? The New York Times. Retrieved November 8, 2021, from https://www.nytimes.com/interactive/2020/us/covid-hospitals-near-you.html

[4.] Green. (2002). How Many Hospital Beds? Inquiry (Chicago), 39(4), 400–412. https://doi.org/10.5034/inquiryjrnl_39.4.400

[5.] World Health Organization. (2020, March 3). Shortage of personal protective equipment endangering health workers worldwide. WHO. Retrieved November 8, 2021, from https://www.who.int/news/item/03-03-2020-shortage-of-personal-protective-equipment-endangering-health-workers-worldwide

[6.] DeVore, S. (2020, September 30). Surviving The Waves Of A Pandemic Storm: How To Fix The Supply Chain Flaws Exposed By COVID-19. HealthAffairs. Retrieved November 8, 2021, from https://www.healthaffairs.org/do/10.1377/hblog20200928.305253/full/

[7.] Association of State and Territorial Health Officials. (2011). Strategic National Stockpile Fact Sheet | State Public Health | ASTHO. ASTHO. Retrieved November 8, 2021, from https://www.astho.org/Programs/Preparedness/Public-Health-Emergency-Law/Emergency-Use-Authorization-Toolkit/Strategic-National-Stockpile-Fact-Sheet/

[8.] North Carolina State University. (2020, November 12). The Strategic Stockpile failed; experts propose new approach to emergency preparedness. ScienceDaily. Retrieved November 8, 2021, from https://www.sciencedaily.com/releases/2020/11/201112120500.htm

[9.] Juraschek, S. P., Zhang, X., Ranganathan, V., & Lin, V. W. (2012). United States Registered Nurse Workforce Report Card and Shortage Forecast. American Journal of Medical Quality, 27(3), 241–249. https://doi.org/10.1177/1062860611416634

[10.] Themes, U. (2016, August 7). Confronting the Nursing Shortage. Nurse Key. Retrieved November 8, 2021, from https://nursekey.com/confronting-the-nursing-shortage/

[11.] Boyle, P. (2021, September 7). Hospitals innovate amid dire nursing shortages. AAMC. Retrieved November 8, 2021, from https://www.aamc.org/news-insights/hospitals-innovate-amid-dire-nursing-shortages

[12.] Chung, H.W., Apio, C., Goo, T. et al. (2021, October 14). Effects of government policies on the spread of COVID-19 worldwide. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-99368-9

[13.] Ngo, M. (2021, October 28). Program to Lend Billions to Aid California's Supply-Chain Infrastructure—The New York Times. https://www.nytimes.com/2021/10/28/us/politics/california-ports-supply-chain.html

[14.] Deschepper, M., Eeckloo, K., Malfait, S., Benoit, D., Callens, S., & Vansteelandt, S. (2021). Prediction of hospital bed capacity during the COVID- 19 pandemic. BMC Health Services Research, 21(1), 468–468. https://doi.org/10.1186/s12913-021-06492-3

[15.] Weller, Jason & Dimitoglou, George. (2009). Survey of Models, Methods and Techniques for Computational Epidemiology. 99-105.

[16.] Kamar, A., Maalouf, N., Hitti, E., El Eid, G., Isma'eel, H., & Elhajj, I. H. (2021). Challenge of forecasting demand of medical resources and supplies during a pandemic: A comparative evaluation of three surge calculators for COVID-19. Epidemiology and Infection, 149, e51–e51. https://doi.org/10.1017/S095026882100025X

[17.] Announcing CHIME, A tool for COVID-19 capacity planning. (n.d.). Retrieved October 16, 2021, from http://predictivehealthcare.pennmedicine.org/2020/03/14/accouncing-chime.html

[18.] WHO COVID-19 essential supplies forecasting tool (COVID-ESFT). (n.d.). Retrieved October 16, 2021, from https://www.who.int/publications/i/item/WHO-2019-nCoV-Tools-Essential-forecasting-2021-1

[19.] AUBMC | COVID-19. (n.d.). Retrieved October 16, 2021, from https://aubmc.org.lb/COVID-19/Pages/covid-19.html

[20.] Compartmental models in epidemiology. (2021, October 14). In *Wikipedia*. https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology

[21.] Postnikov, E. B. (2020). Estimation of COVID-19 dynamics "on a back-of-envelope": Does the simplest SIR model provide quantitative parameters and predictions? Chaos, Solitons & Fractals, 135, 109841. https://doi.org/10.1016/j.chaos.2020.109841

[22.] Majid, F., Gray, M., Deshpande, A. M., Ramakrishnan, S., Kumar, M., & Ehrlich, S. (2021). Non-Pharmaceutical Interventions as Controls to mitigate the spread of epidemics: An analysis using a spatiotemporal PDE model and C O V I D –19 data. ISA Transactions, S001905782100121X. https://doi.org/10.1016/j.isatra.2021.02.038

[23.] Arazi, R., & Feigel, A. (2021). Discontinuous transitions of social distancing in the SIR model. Physica A: Statistical Mechanics and Its Applications, 566, 125632. https://doi.org/10.1016/j.physa.2020.125632

[24.] Too, J., & Mirjalili, S. (2021). A Hyper Learning Binary Dragonfly Algorithm for Feature Selection: A COVID-19 Case Study. Knowledge-Based Systems, 212. https://doi.org/10.1016/j.knosys.2020.106553

[25.] Decoding Clinical Biomarker Space of COVID-19: Exploring Matrix Factorization-based Feature Selection Methods. (2021). In Medical Letter on the CDC & FDA (p. 112–). NewsRX LLC.

[26.] Kaliappan, Srinivasan, K., Qaisar, S. M., Sundararajan, K., Chang, C.-Y., & C, S. (2021). Performance Evaluation of Regression Models for the Prediction of the COVID-19 Reproduction Rate. Frontiers in Public Health, 9, 729795–729795. https://doi.org/10.3389/fpubh.2021.729795

[27.] Stedman, M., Lunt, M., Davies, M., Gibson, M., & Heald, A. (2020). COVID-19: Generate and apply local modelled transmission and morbidity effects to provide an estimate of the variation in overall relative healthcare resource impact at general practice granularity. International Journal of Clinical Practice, 74(9). https://doi.org/10.1111/ijcp.13533

[28.] Li, D., Gaynor, S. M., Quick, C., Chen, J. T., Stephenson, B. J. K., Coull, B. A., & Lin, X. (2021). Identifying US County-level characteristics associated with high COVID-19 burden. BMC Public Health, 21(1), 1007. https://doi.org/10.1186/s12889-021-11060-9

[29.] Shaw, A. K., White, L. A., Michalska-Smith, M., Borer, E. T., Craft, M. E., Seabloom, E. W., Snell-Rood, E. C., & Travisano, M. (2021). Lessons from movement ecology for the return to work: Modeling contacts and the spread of COVID-19. PLOS ONE, 16(1), e0242955. https://doi.org/10.1371/journal.pone.0242955

[30.] Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, and Helen Tatlow. (2021). "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)." Nature Human Behaviour. https://doi.org/10.1038/s41562-021-01079-8

[31.] COVID-19 Reported Patient Impact and Hospital Capacity by Facility—RAW | HealthData.gov. (2021, November 1). https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/uqq2-txqb

[32.] Official Vaccine Releases: View All by Date 2021. (n.d.). https://www.immunize.org/newreleases/

[33.] Flu Season | CDC. (n.d.). https://www.cdc.gov.flu/about/season/flu-season.htm

[34.] COVID-19 Vaccinations in the United States,Jurisdiction | Data | Centers for Disease Control and Prevention, (n.d.). https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc

[35.] Groemping. (2007). Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. The American Statistician, 61(2), 139–147. https://doi.org/10.1198/000313007X188252

[36.] Kreps, & Kriner, D. L. (2020). Model uncertainty, political contestation, and public trust in science: Evidence from the COVID-19 pandemic. Science Advances, 6(43). https://doi.org/10.1126/sciadv.abd4563

[37.] Burrows, M., Burd, C., & McKenzie, B. (n.d.). Commuting by Public Transportation in the United States: 2019. 11.

[38.] Texas businesses requiring masks face backlash, threats | The Texas Tribune. (n.d.). Retrieved February 21, 2022, from https://www.texastribune.org/2021/03/07/texas-businesses-masks-threats/

[39.] Upset parents sue Loudoun County School Board for continuing mask requirement. (n.d.). Retrieved February 21, 2022, from https://www.fox5dc.com/news/upset-parents-sue-loudoun-county-school-board-for-continue-its-mask-requirement

[40.] Travis, K. (2021, June 29). How COVID-19 vaccines were made so quickly without cutting corners. Science News. https://www.sciencenews.org/article/covid-coronavirus-vaccine-development-speed