Introduction

As of today, COVID-19 has caused over 500,000 deaths in the United States, which continues to rise, and has caused many to lose their jobs, making a serious social and economic impact on the global community. Vaccines, now more than ever, are the key to stopping the pandemic, and vaccine strategy studies, especially with limited vaccine supplies, not only concerns the government policy makers but also the public. Currently, each state in the United States develops its own vaccine distribution program, based on CDC guidelines, that focus on front-line workers, essential workers, and high-age population. However, with limited vaccine supplies and order of vaccines on a first-come first-serve basis, states that have a higher high-risk population may not get enough vaccines while some lower risk states may have more than needed.

Problem Definition

As such, we are determining an optimized way to distribute the COVID-19 vaccines within the United States to minimize the total number of deaths. We are stratifying the population by risk, determined by historical COVID-19 death rates among different demographic groups. Using demographic data about each state, we are able to determine the risk to that state and project the number of deaths in an unvaccinated population and can allocate the existing vaccine supply accordingly.

Literature Summaries:

Our project topic focuses on building efficient modeling frameworks to optimize COVID-19 vaccine distribution. To effectively understand this topic, we broke our literature research into two main groups: Vaccination Campaigns, and Mathematics and Modeling.

We first began by understanding our goal; efficient vaccine distribution. What does this entail and how do we calculate this to the scale of the United States? The most notable studies explain the basic concept of achieving the “herd immunity threshold”[1], the immunity of an individual on the scale of an entire population, during COVID-19. Ultimately, there are two options to achieve herd immunity: mass vaccination campaign or the natural immunization of populations over time. It is predicted that without a vaccine campaign, herd immunity is not an achievable goal overall. Also, given re-infection data, it is expected that even with herd immunity, short-term immunity and annual outbreaks will occur, although person(s) would likely build immune control to the virus and limit disease pathology, which would decrease the clinical severity of subsequent infections. We plan to leverage these understandings, in addition to their division of demographics[2] (i.e. activity, age-group, social activity, or preventative measures) in our models to further optimize a distribution plan.

Unfortunately, these articles primarily focus on herd immunity in the abstract. While they do give some context into the COVID-19 pandemic, they did not all provide strategy or statistical modeling on how to achieve herd immunity in specific populations or countries. Also, most articles were written in mid-2020 without real vaccine data. We will use real and more up-to-date data across vaccination, population, and COVID-19 statistics for an efficient distribution plan.

The second grouping that we focused on was Mathematics and Modeling. We explored what analysts and researchers have already accomplished in this area, how they modeled to achieve their outcomes, and where we can make improvements for our project. These studies dive into the application of data analytics to COVID-19 like creating selection criteria models[3], like SEIR[4][5] and machine learning[6], ranking the most eligible groups[7] to receive the COVID-19 vaccine first, mathematical perspectives to estimate the percentage of decay in infections and mortality rates under alternative vaccine rollout speeds[8], vaccine distribution models[9][10], dosage trials[11], providing options for modeling when vaccine data is limited[12], and historic optimal influenza vaccine allocation for five outcome measures: death, infections,

years of life lost, contingent valuation, and economic costs[13]. Overall, these papers give us insight into various mathematical and statistical approaches that we can leverage when building our optimization model. We plan to specifically leverage regression, random forest, and naive bayes modeling techniques based on these readings for our initial death rate probability calculations. Then use machine learning tactics to compare our distribution plan against the current vaccination schedule.

Areas that these articles fell short are primarily due to outdated data. For example, most of the studies were conducted before the COVID-19 vaccinations were FDA approved, PCR test requirements were created, or the known availability of vaccines today. Although we will work with incomplete data, we will use the most recent data, like monthly updated COVID-19 statistics[A] in our prediction models to obtain more meaningful results for the near future.

Method

The stratification strategy mentioned in the literature survey focuses on factors such as age, years of life lost, and economic factors like cost of society and productivity but did not examine the impact of sex or race. Therefore we would like to include those as factors in our analysis. Also, the models used are mostly mathematical models like the epidemic SEIR model that is differential equation based, simulation based, or time series based models like ARIMA. We would like to approach this from a different perspective using factor based models.

Our approach is the following: first, we estimate the probabilities of deaths for all demographic groups based on the combination of age-group, sex, and ethnicity. Second, we rate each state's risk level by calculating the expected deaths for each demographic group in a given state using the data obtained above with the state's population data. We expect to allocate more vaccines to states that have a higher expected number of deaths. Lastly, we will develop a model to examine the relationship between vaccine allocation and COVID-19 deaths and apply the data we collected in prior steps to determine a better vaccine allocation program that aims to reduce the total number of deaths.

Overall, we believe our approach provides a better and more accurate way to identify higher risk groups compared to the current practice that only uses age-factor as a priority metric to determine who should be vaccinated first for the non-frontline population and general public.

Step 1: Modeling the probability of death by demographic groups.

We determine the demographic groups based on the features sex, ethnicity and age-group from the CDC's surveillance dataset[A] and aggregate the death status based on the combination of these features. We used an API to pull the 20M row dataset and filter out all the records that have "missing", "N/A", or "unknown" values in the features that we used. The filtered dataset has more than 630K records. We then apply logistic regression, random forest, and naive bayes models to estimate the probability of death of each group combination of sex, ethnicity, and age-group. We select 20% of the data as the test set and split the rest of the data by 60% for training and 20% of data for validation and parameters tuning. We use testing accuracy and AUC as benchmarks to compare the performance and select the best model [Figure 1]. Random forest has slightly higher testing accuracy of 95.51%, while logistic regression and naive bayes have the testing accuracy values of 95.45% and 95.49%, respectively. The AUC values of random forest, naive bayes and logistic regression are 0.9185, 0.9095 and 0.9155, respectively. Since random forest performs better, we use it to estimate the probability of death for each demographic group. Each tree in the random forest predicts the death probability of a demographic group as the fraction of samples of the deaths in the leaf node in which the demographic group is located. The random forest yields a single probability value for each demographic group by averaging the predicted death probabilities of all trees.

Step 2: Assessing the risk of each state based on its demographic composition.

Using the probability of death from Step 1, we assess each state's expected death probability based on its population of demographic groups. We use the US Census Bureau's State Population by Characteristics[B] dataset to obtain and group the population and demographic data. Then we calculate the expected percentage of deaths of each demographic group in each state by multiplying the population in each group with its associated probability of death. Then rank the states by expected percentage of death to determine the state's risk. At this stage, the results show the expected probability of death without any vaccination and mitigation methods like social distancing and mandatory masks wearing.

There are some limitations using the Census Bureau's dataset, as the most up to date population estimation is 2019, which is not preferable. Population may have significant changes in the year of 2020 especially during the COVID-19 outbreak period when people moved from one state to another to avoid living in hotspot areas. We are not able to obtain a more up-to-date dataset, but we anticipate the model will be as effective if we have 2020 data available.

Step 3: Evaluation - Comparing the projected results with the current approach

Our proposed approach is to prioritize states that have the highest risk. We defined risk as having the highest estimated number of deaths based on the results we obtained from step 2. If a state has a greater population of high risk demographic groups, we expect to see the state will have a higher expected number of deaths and therefore will gain higher priority in the vaccine allocation.

To evaluate whether the vaccination allocation strategy based on demographics is a better approach, we examine the relationship between the vaccine allocated in each state under the current allocation program using the daily vaccine distributed (cumulative) in the US vaccination dataset[D] and the actual cumulative number of COVID deaths from the Provisional COVID-19 Death Counts by Sex, Age, and State dataset[A]. We merged the two dataset by week since the state vaccine allocation data is available (Jan 23, 2021) and applied the data to the random forest model to look at the relationship. We define the features as the number of vaccines distributed to each state, and the response as cumulative COVID-19 death since Jan 23, 2021. Each row represents a week of data and its ending date will be indicated by the date column. We measure the performance of the random forest model with R-squared value. After we split the data into training and test sets, the trained model has the R-squared value 0.97 on the test set. Thus, our random forest model yields a good estimation for the cumulative number of COVID-19 deaths.

After we built the model, we calculated the number of vaccines allocated to each state based on the total amount of vaccines distributed in the US multiplied the weighted average of the estimated number of deaths in each state from the result we obtained in step 2. We then fit the data with the random forest model built. For each week, we proposed a vaccine allocation plan for each state, and the random forest model estimates the number of COVID-19 deaths in the US under this plan.

With the proposed vaccine allocation plan, the cumulative number of COVID-19 deaths remains less than the actual cumulative number of COVID-19 deaths starting from week 3. Between the dates Jan 23, 2021 and Feb 06, 2021, there were 57,348 COVID-19 deaths. If the vaccines were distributed as we proposed, then there would be approximately 53,785 COVID-19 deaths between these dates. From Jan 23, 2021 to Apr 10, 2021, 112,293 people lost their lives from COVID-19 in the US. The random forest model estimates that this number would be 109,276 under the proposed vaccine allocation plan. This estimation shows that if the vaccines were distributed to the states according to their priority level we found, then there would be a decrease in the number of COVID-19 deaths.

Our proposed approach shows 3017 less deaths as of Apr 10, 2021, which is about

2.7% more effective compared to the current approach. We believe there are a couple of factors: first we see from the result of step 1 that age-group is the most important factor in relationship with estimated COVID-19 death [Figure 2]. The current CDC guideline also prioritizes the older age population when administering the vaccine. Our result shows we can have some further improvement to the current CDC plan if we further differentiate and prioritize population based on sex and ethnicity.

We still see some potential improvement of the model if we are able to obtain more data for the model. First, we are only able to obtain weekly COVID-19 death data for the states, state vaccine allocation data are only available starting Jan 23, 2021. The common ground between the two dataset gives us only about 13 weeks worth of observations. And since we are using the states as variables. The random forest model has more variables than it does have observations. We remedy this by splitting the observations into a training set (80%) and a test set (20%). While we see good results, because of the limited observations and large number variables, we still think the model is likely to overfit to the current vaccine allocation plan.

Visualization

We made 3 interactive visualizations to help the users better understand the result of our findings (refer to visualization link in the appendix).

There are 3 tabs, the “Distribution Priority” tab shows a choropleth map of the US and displays the estimated death percentage based on the 2019 population estimate and risk rank of the state among the states in the U.S when the user hovers over the state.

The “Comparison” tab is a line chart that compares the result we obtained from step 3 and the current vaccine allocation plan.

The “Individual Risk” tab shows the estimated death rate across age groups based on the model, ethnicity, and gender which the users are able to select. It allows the user to see under the 3 models we explore, how the estimated death rate will change based on gender and ethnicity groups.

Innovations

First, we are coordinating the distribution at the federal level, as opposed to letting each state request the number it thinks it needs. This will be an improvement because we can ensure that areas with higher risk populations receive more doses, and that excess doses are not stockpiled unused, and thus go to waste. Second, we are using a factor-based model to determine which states have a higher risk, using variables like ethnicity, sex, and age. The primary considerations in federal guidelines have been age and medical conditions. Using factor-based models with variables drawn from CDC data about COVID-19 deaths allows us to label risk groups more accurately. Third, for each given state, we can also identify the populations which are more at risk, which allows states to target them. This is accomplished using a model that is capable of predicting the likelihood of death if a given individual contracts COVID-19.

Discussions and Conclusion

We have several observations when we are exploring the dataset and developing our approaches. From the demographic groups probability rates we deduced using the surveillance dataset[Figure 2] we can see certain demographic groups exhibit higher probability of death than others. For example, referring to figure 2, we see males (orange bars) have a higher death than females (blue bars) across all age-groups and races. And certain ethnic groups have higher overall death probability than other ethnic groups. When we ranked the states with the highest risk of death by COVID-19 [Figure 3]. Assuming there are no additional safety measures, and every person contracts COVID-19, Florida has the highest risk and can expect

6.98% death (~3M people). The most at-risk group across all demographics are Hispanic/Latino men, who are 80+ years old (55% death risk)[Figure 3]. This age group is in line with the CDC's vaccination guidance, as those older than 65 years we're in the first group to be vaccinated after healthcare workers.

These observations support our hypothesis that it will be more effective for the US to allocate vaccines first to states that have a greater high risk population.

Yet, we are unable to capture the most up to date population statistics to project the risk of the states. Therefore we use the 2019 population and demographic estimates. While this is a good baseline of the population distribution of the US, this can also potentially lead to inaccurate results as we see there are a lot of people moving from one state to another during the COVID-19 period. Also, we made some unrealistic assumptions in step 2. Because we are unable to collect data of confirmed cases based on demographics and states, we assume 100% COVID-19 confirm rates when we calculate the estimated death of each population group in each state.

Lastly, from the comparison of cumulative COVID-19 deaths between the current approach and the proposed approach, we see that our approach does perform slightly better. The model we built may be more biased towards the training set, which is the current approach, due to having more variables than observations. We attempted several other methods like grouping the states into risk groups but we still think having the states as individual variables may be better for us to compare the result to the current plan.

Overall, prioritizing vaccine allocation based on demographics is a better approach. We see the CDC guideline suggested to prioritize the vaccines to the general public based on age group. We proposed, in addition to age, we should further stratify the population based on gender and ethnicity as we found these factors do impact the death rate. Because of this, we rank the states' risk based on the estimated deaths and prioritize the vaccine distribution to the higher risk states. We see that our proposed approach performs better than the current approach.

Distribution of team effort:

Every team member has contributed equally in our project efforts.

Appendix:

Figures and Visualizations:

Figure 1: Table Comparison of Models for Classifying the Deaths and the Survivals

Model	AUC	Testing Accuracy
Random Forest	**0.9185**	**95.51%**
Logistic Regression	0.9155	95.45%
Naive Bayes	0.9095	95.49%

Figure 2: Death Probability by Demographic Groups Bar Chart

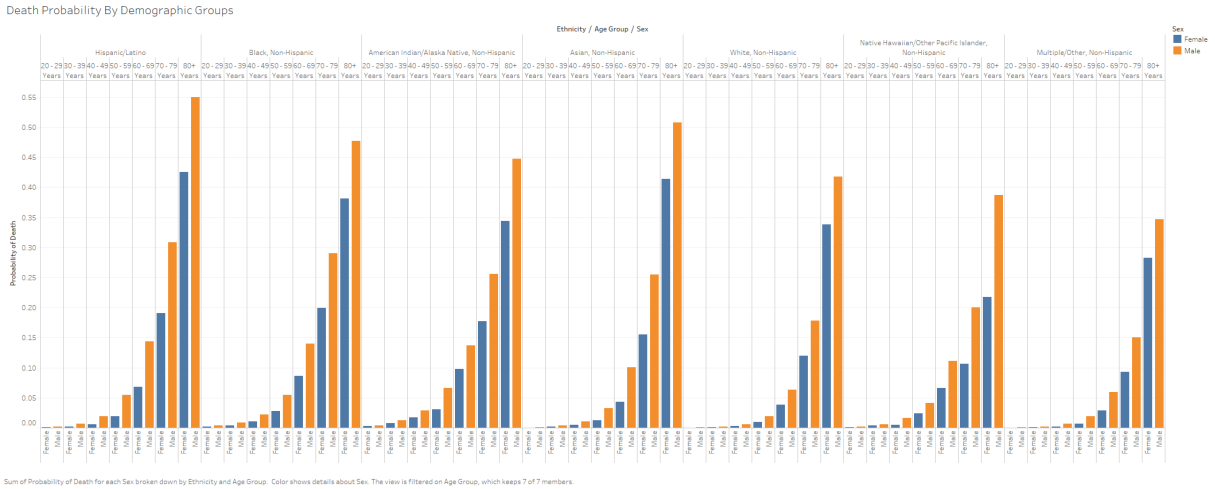
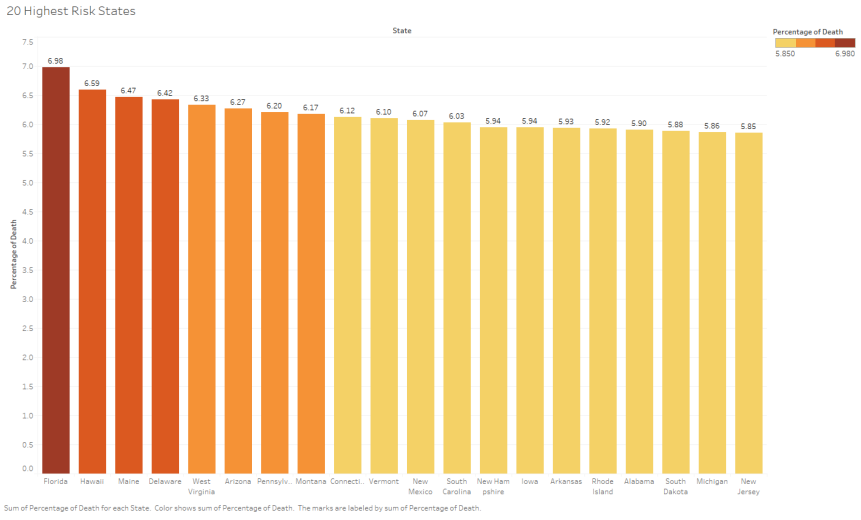


Figure 3: 20 Highest Risk States Ranked by Estimated Percentage Number of Death



Visualization Link: http://206.189.202.125:3000/

Datasets:

[C] “Age, Sex, Race, and Hispanic Origin - 6 Race Groups (5 Race Alone Groups and One Multiple Race Group) (SC-EST2019-ALLDATA6).” *State Population by Characteristics: 2010-2019*, United States Census, 22 June 2020, www.census.gov/data/tables/time-series/demo/popest/2010s-state-detail.html#par_textimage_785300169.

[B] “COVID-19 Case Surveillance Public Use Data.” *Centers for Disease Control and Prevention Data Catalog*, Surveillance Review and Response Group, 28 Feb. 2021, data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf.

[A]“Provisional COVID-19 Death Counts by Sex, Age, and State.” *Centers for Disease Control and Prevention Data Catalog*, National Center for Health Statistics, 28 Feb. 2021, data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku.

[D] “US State Vaccinations.” *COVID-19 Data*, Our World In Data, 24 Mar. 2021, github.com/owid/covid-19-data/tree/master/public/data/vaccinations.

Bibliography:

[5] Bartsch, Sarah M., et al. “Vaccine Efficacy Needed for a COVID-19 Coronavirus Vaccine to Prevent or Stop an Epidemic as the Sole Intervention.” *American Journal of Preventive Medicine*, vol. 59, no. 4, 2020, pp. 493–503., doi:10.1016/j.amepre.2020.06.011.

[7] Britton, Tom, et al. “A Mathematical Model Reveals the Influence of Population Heterogeneity on Herd Immunity to SARS-CoV-2.” Science, vol. 369, no. 6505, 2020, pp. 846–849., doi:10.1126/science.abc6810.

[2] Britton, Tom, et al. “The Disease-Induced Herd Immunity Level for Covid-19 Is Substantially Lower than the Classical Herd Immunity Level.” 2020, doi:10.1101/2020.05.06.20093336.

[8] Bubar, Kate M., et al. “Model-Informed COVID-19 Vaccine Prioritization Strategies by Age and Serostatus.” 2020, doi:10.1101/2020.09.08.20190629.

[9] Grauer, Jens, et al. “Strategic Spatiotemporal Vaccine Distribution Increases the Survival Rate in an Infectious Disease like Covid-19.” *Scientific Reports*, vol. 10, no. 1, 2020, doi:10.1038/s41598-020-78447-3.

[3] Hezam, Ibrahim M., et al. “COVID-19 Vaccine: A Neutrosophic MCDM Approach for Determining the Priority Groups.” *Results in Physics*, vol. 20, 2021, p. 103654., doi:10.1016/j.rinp.2020.103654.

[10] Koff, Wayne. “Faculty Opinions Recommendation of Ensemble Forecast Modeling for the Design of COVID-19 Vaccine Efficacy Trials.” *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature*, 2020, doi:10.3410/f.738716175.793579430.

[4] Kohli, Michele, et al. “The Potential Public Health and Economic Value of a Hypothetical COVID-19 Vaccine in the United States: Use of Cost-Effectiveness Modeling to Inform Vaccination Prioritization.” *Vaccine*, vol. 39, no. 7, 2021, pp. 1157–1164., doi:10.1016/j.vaccine.2020.12.078.

[11] Lee, Bruce Y., et al. “Single versus Multi-Dose Vaccine Vials: An Economic Computational Model.” *Vaccine*, vol. 28, no. 32, 2010, pp. 5292–5300., doi:10.1016/j.vaccine.2010.05.048.

Loomba, S., et al. “Measuring the Impact of Exposure to COVID-19 Vaccine Misinformation on Vaccine Intent in the UK and US.” 2020, doi:10.1101/2020.10.22.20217513.

[13] Medlock, J., and A. P. Galvani. “Optimizing Influenza Vaccine Distribution.” *Science*, vol. 325, no. 5948, 2009, pp. 1705–1708., doi:10.1126/science.1175570.

[6] Mondal, M. Rubaiyat Hossain, et al. “Data Analytics for Novel Coronavirus Disease.” *Informatics in Medicine Unlocked*, vol. 20, 2020, p. 100374., doi:10.1016/j.imu.2020.100374.

[1] Randolph, Haley E., and Luis B. Barreiro. “Herd Immunity: Understanding COVID-19.” *Immunity*, vol. 52, no. 5, 19 May 2020, pp. 737–741., doi:10.1016/j.immuni.2020.04.012.

[12] Zhang, Yao, and B. Aditya Prakash. “Scalable Vaccine Distribution in Large Graphs given

Uncertain Data.” *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, doi:10.1145/2661829.2662088.