

COVID-19 Infection Risk & Vaccine Allocation Models: AI/ML & SAS Viya for Nursing Homes

December 20, 2020

Alexandra Allen, Anzhi Mou, Zhaoyu Qiao, Harvir Virk Singh, Xiaoyu Zhu

Carnegie Mellon University, Heinz College, Graduate Capstone Project in Partnership with SAS Enterprise

CMU Faculty Advisor: Dr. Rema Padman, SAS Enterprise Client Advisor: Manuel Figallo

Authors & Advisors

Alexandra Allen

Project Manager / Healthcare Policy Advisor, MS Public Policy & Management

Anzhi Mou

Data Manager & Quality Assurance Engineer, MS Information Systems Management

Zhaoyu Qiao

Analytics Chief & Data Scientist, MS Healthcare Analytics and Information Technology

Harvir Virk Singh

Healthcare Data Analyst / Documentation Manager, MS Healthcare Analytics and Information Technology

Xiaoyu Zhu

Language Processing Engineer / Data Scientist, MS Information Systems Management

Rema Padman

Faculty Advisor, Carnegie Mellon University, Trustees Professor of Management Science & Healthcare Informatics at Heinz College

Manuel Figallo

Client Advisor, SAS Institute, Principal Systems Engineer at SAS Analytics Government Business Unit

Citation

Please use the following citation for this project when referenced in publications, presentations, or materials of any kind:

“Allen, Alexandra, Mou, Anzhi, Qiao, Zhaoyu, Virk, Harvir Singh, Zhu, Xiaoyu. (2020). COVID-19 Infection Risk & Vaccine Allocation Models: AI/ML & SAS Viya for Nursing Homes. Carnegie Mellon University - Heinz College graduate capstone project in partnership with SAS Enterprises. Faculty advisor Dr. Rema Padman.”

Executive Summary

Roughly 40% of COVID-19 deaths have been among residents and staff of nursing homes and other long-term care facilities in the US. Roughly 787,000 people have been infected resulting in at least 106,000 deaths as of December 4, 2020, with facilities with higher shares of Black and Hispanic residents experiencing more severe outbreaks and higher likelihood of death from COVID-19.¹ Vaccine distribution is now underway, but facilities still need supplies to reduce infection spread through the fall surge and beyond.

This project aims to identify nursing homes at high-risk of infection and mortality in order to recommend testing and personal protective equipment resources and facilitate equitable vaccine allocation. The team constructed machine learning models to predict nursing home infection and mortality rates based on several datasets containing nursing home COVID-19 infection and mortality data, facility quality ratings, and past inspection report features, as well as systemic factors such as resident demographics, pre-pandemic acuity, and the Social Vulnerability Indexes of their communities. The team also applied natural language processing techniques to nursing homes' textual inspection reports, demonstrating that they have some predictive potential, and providing a streamlined way for investigators to review past reports.

The models ultimately did not provide enough predictive capability that we felt confident staking a recommendation on their outputs, so the team constructed our interactive dashboard in SAS Viya/SAS Visual Analytics solely using the Group-Based Trajectory model that clusters historic infection trends (as opposed to a predictive model) and aligned each component with our 3 use cases of Targeting, Investigation, and Intervention.² This interactive dashboard combines all of the work we have done into a practical, useful tool to visualize complex problems and identify actionable solutions in a rapidly changing environment.

¹ The New York Times. (2020, June 27), About 38% of U.S. Coronavirus Deaths Are Linked to Nursing Homes, Retrieved December 18, 2020.
<https://www.nytimes.com/interactive/2020/us/coronavirus-nursing-homes.html>

² Jones, B. and Nagin, D., 2013. A Note on a Stata Plugin for Estimating Group-based Trajectory Models, *Sociological Methods & Research*, 42(4), pp.608-613.

Table of Contents

- I. Introduction
- II. Objectives
- III. Data Sources & Pipeline
- IV. Interview Insights
- V. Modeling Process & Results
- VI. Interactive Dashboard
- VII. Conclusions and Recommendations

Introduction: Description of Problem and Context

Since the outbreak of COVID-19 in the United States, roughly 40% of COVID-19 deaths have been among residents and staff of nursing homes and other long-term care facilities in the US. As of December 4, 2020, COVID-19 has infected more than 787,000 people at some 28,000 facilities, resulting in at least 106,000 deaths.³ Vaccine distribution is underway starting December 2020, but facilities still need supplies to reduce infection spread through the fall surge and beyond.

There are also stark disparities among nursing homes. A recent study from the Kaiser Family Foundation found that, among nursing homes with at least one case, those with 20% or more Black or Hispanic residents reported more severe outbreaks and were more likely to have at least one COVID-19 death than facilities with low shares of Black or Hispanic residents.⁴ The US healthcare system has long marginalized Black and Indigenous People and People of Color, and the Covid-19 pandemic is demonstrating the deadly results. We have considered it our responsibility to integrate Covid-19's disproportionate impact on nursing homes and residents of color into our analysis.

There is uncertainty about which nursing homes will have outbreaks and high COVID-19 infection rates in the near future, as well as which nursing homes will face the highest mortality rates if residents are exposed. This makes it difficult for Centers for Medicare & Medicaid Services (CMS) to effectively target and prioritize distribution of COVID-19 vaccine doses as well as testing supplies and personal protective equipment (PPE) to the highest-risk facilities. Vaccine allocation is to be determined by state governments, with distribution and dose administration to long-term care facilities managed by private partner pharmacies CVS and Walgreens.

This project is a *proof of concept* for approaching these issues using available data, tools, and technology to identify nursing homes at high-risk of infection and mortality in order to recommend priority distribution of testing supplies, PPE, and COVID-19 vaccines.

³The New York Times, (2020, June 27), About 38% of U.S. Coronavirus Deaths Are Linked to Nursing Homes, Retrieved December 18, 2020.
<https://www.nytimes.com/interactive/2020/us/coronavirus-nursing-homes.html>

⁴Chidambaram, P., Neuman, T., & Garfield, R. (2020, October 27). Racial and Ethnic Disparities in COVID-19 Cases and Deaths in Nursing Homes. Retrieved December 09, 2020, from <https://www.kff.org/coronavirus-covid-19/issue-brief/racial-and-ethnic-disparities-in-covid-19-cases-and-deaths-in-nursing-homes/>

Objectives

After carefully examining the project timeline and the students' collective skillset, the team decided on three major project goals to accomplish. These deliverables will be handed off to our client in the form of a data processing/Machine Learning pipeline and an interactive dashboard product powered by SAS Viya/SAS Visual Analytics.

1) Targeting

The team will use machine learning techniques to identify nursing homes at high risk of COVID-19 infection and mortality to support them with resources such as testing supplies and personal protective equipment.

2) Investigation

The team will extract, analyze and summarize information from past CMS inspection reports to summarize narratives for more efficient review by investigators, and explore potential issues in nursing homes related to COVID-19 to inform policy making.

3) Intervention

The team will design and implement a web dashboard component that can operationalize equitable vaccine allocation strategies.

Scope Summary

In order to reach the objectives and proof of concept, we combined eight datasets to create a machine learning model to target nursing home facilities most at risk from COVID-19 infection/mortality and an interactive dashboard to facilitate real-time decision making for vaccine and resource allocation, aligned with the use cases of Targeting, Investigation, and Intervention. The project was conducted over the course of 4 months in fall 2020. The team consulted with experts in their respective fields to address any limitations within the implementation framework.

Modeling was an iterative process. The team started with simple models in order to capture the effect of COVID-19 infections and mortality within nursing homes. However the team soon identified simple models failed to accurately predict the most at risk nursing homes. The models evolved to more complex ones but still did not provide enough predictive capability that we felt confident staking a recommendation on their outputs. Therefore, we constructed our interactive dashboard in SAS Viya/SAS Visual Analytics using the Group-Based Trajectory model that clusters historic infection trends (as opposed to a predictive model) and aligned each component with our 3 use cases.

Data Sources & Pipeline

1) CMS Facility COVID-19 (2020)

The Centers for Medicare and Medicaid Services COVID-19 Nursing Home dataset is a publicly available dataset. The dataset provided weekly data on COVID-19 in each nursing home, including infections and deaths among residents and staff, COVID-19 resident admissions, testing capacity and use of testing as surveillance and following suspected/confirmed infections, personal protective equipment supply, and staffing. The data was collected starting from May 24, 2020 through November 29, 2020 as of this paper's compilation. The lack of data before May 24 is a significant limitation of this dataset - we could not adequately assess patterns during the initial spring surge. In addition, much of the data collected was considered unreliable for longitudinal analysis (did not pass Quality Assurance Check) as nursing homes were adjusting to providing this level of data on a weekly basis. We followed CMS's data quality recommendation and separated out May 24 data in our modeling, feature engineering, and the dashboard. Some features were engineered using 6/21 as a reference date for this reason - see Notes on Dependent Variable section for details. Some missing values were also imputed - see Hybrid Modeling section for details.

Centers for Medicare & Medicaid Services (CMS), COVID-19 Nursing Home Dataset, (2020), <https://data.cms.gov/Special-Programs-Initiatives-COVID-19-Nursing-Home/COVID-19-Nursing-Home-Dataset/s2uc-8wxxp>

2) LTCFocus (2017)

The Brown School of Public Health Long Term Care Focus (LTCFocus) dataset is a publicly available dataset. The dataset contains information about resident demographics aggregated at the facility level from 2017, such as share of residents by race/ethnicity (White, Black, & Hispanic only), average resident acuity, and average resident age. The dataset also contained features allowing a review of any speciality care the facility may provide. The dataset was used for the model of this project as well as the dashboard deliverable. However, this dataset did have the limitation that the data available was from 2017 whereas the COVID-19 dataset was from 2020. Another slight limitation is that it did not contain any information that pertained to small numbers of residents (1-11 specifically) to maintain confidentiality of personal identifiers, per an arrangement with CMS. Such values were replaced with "LNE" in the original dataset. For our modeling process, we imputed those values based on the size of the facility (see Modeling Codebook for details). In the dashboard, we treated them as "missing".

Shaping Long Term Care in America Project at Brown University funded in part by the National Institute on Aging (1P01AG027296), <http://ltcfocus.org/1/about-us>

3) Provider Information (2020)

The Centers for Medicare and Medicaid Services Provider Information is a publicly available dataset. The Provider Information dataset includes information on active nursing home quality measures, type of ownership for the facility, staffing of specialized roles such as registered nurses and physical therapists, as well as staff ratings as an aggregate for each nursing home. The Provider Information dataset provided a comparison between each nursing home based on various rating scales. The dataset is used in the modeling and the dashboard components of the deliverables to add additional features to compare between nursing homes. There were missing numerical values in the original dataset. For modeling purposes, the team used a K Nearest Neighbour Imputation method to fill those missing values.

Centers for Medicare & Medicaid Services (CMS), Provider Information, (2019),
<https://data.cms.gov/provider-data/dataset/4pq5-n9py>

4) CMS Inspection Reports (2019)

The Centers for Medicare and Medicaid Services Inspection Text Report is a publicly available dataset. The dataset used for this project was from 2019 and covered inspection reports from the nursing homes who had an inspection during the 2019 cycle. The data consists of reports of any deficiencies a resident may experience and ranged from a low severity and occurrence to a high severity and occurrence. The deficiencies also ranged in a variety of categories from environmental deficiencies to those that impact quality of life and care. The dataset was used in modeling and the dashboard to provide more dimensions as to how facilities compare to one another in performance measures.

ProPublica. Nursing Home Inspect, Retrieved 12/3/2020,
<https://projects.propublica.org/nursing-homes/summary>

****Note:** For reasons unknown we are unable to find the exact dataset presented to us by our client. Our codebooks/pipeline incorporate the original copy provided to us by our client. However, we were able to locate a citation for inspection texts that are not a replica given to us by the client, but still have inspection reports.

5) New York Times COVID-19

The New York Times Community Infection and Mortality Rates dataset is a publicly available dataset. The dataset is updated frequently to reflect real time COVID-19 infections and mortalities throughout the community. The Community Infection and Mortality Rates dataset provided insight as to how the communities surrounding nursing home facilities were impacted by COVID-19 and how the infections and mortalities within the community impact nursing home residents. The Community Infection and Mortality Rates dataset

along with the COVID-19 Nursing Home dataset allowed our model to be timely and current with the ever changing surges and reductions in COVID-19 infections and mortalities.

The New York Times, (2020), Coronavirus (Covid-19) Data in the United States, Retrieved 11/15/2020, contained data from 6/5/2020 through 11/15/2020 at the time of download, from <https://github.com/nytimes/covid-19-data>.

6) COVID-19 Community Social Vulnerability Index (2020)

The Community COVID-19 Social Vulnerability Index is a publicly available dataset. This dataset is a powerful tool to examine which communities may be most vulnerable and why in terms of COVID-19 infections and mortalities. The Community COVID-19 Social Vulnerability Index measures the impact of COVID-19 at a county level and provides a more in depth analysis of equitable vaccine distribution than just the COVID-19 Nursing Home and Community Infection Mortality Rates datasets alone.

Surgo Ventures (2020), Bringing Greater Precision to the COVID-19 Response, Retrieved 11/15/2020, <https://precisionforcovid.org/>

**Note: For reasons unknown, this dataset is now offline. Our codebooks/pipeline incorporate the original copy downloaded from the above source into dataset creation, modeling, and the interactive dashboard.

7) Tribal Nursing Home & Assisted Living Facility Directory (2020)

The Tribal Nursing Home and Assisted Living Facility Directory is a public document that we turned into a small dataset. We included it in order to in some way account for COVID-19's disproportionate impact on indigenous communities in the US and provide a mechanism through which to prioritize them for vaccine distribution. This data was used exclusively for the interactive dashboard, it was not used for modeling.

Department of Health & Human Services, & Centers for Medicare & Medicaid Services, (2020, March 15), Tribal Nursing Home & Assisted Living Facility Directory, Retrieved December 19, 2020, from <https://www.cms.gov/files/document/2020-directory-tribal-nursing-homes-assisted-living-facilities.pdf>

8) Home Health, Hospice, SNF, IRF and LTCH Provider Table (2017)

The Centers for Medicare and Medicaid Services Home Health, Hospice, SNF, IRF and LTCH Provider Table is a publicly available dataset. It contains some demographic information on select CMS facilities. We used this dataset to augment the Tribal Nursing Home collaborative facilities list with additional facilities that recorded 10% or more Native

American/Alaska Native residents. This data was used exclusively for the interactive dashboard, it was not used for modeling.

Centers for Medicare & Medicaid Services (CMS), Home Health, Hospice, SNF, IRF and LTCH Provider Table, CY 2017, (2017).

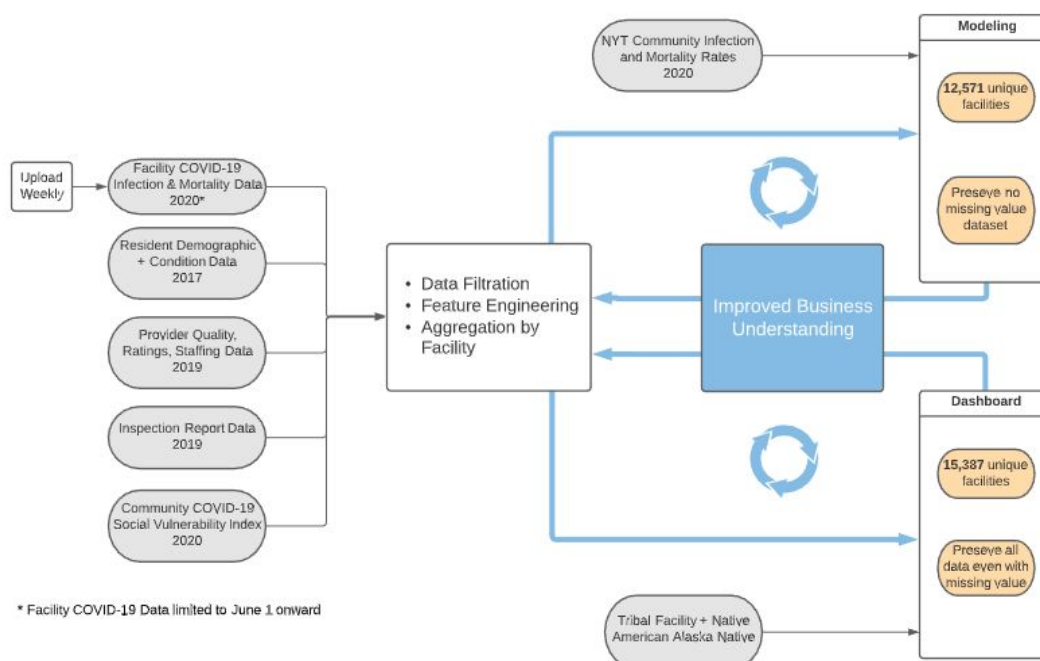
<https://data.cms.gov/Medicare-Hospice/Home-Health-Hospice-SNF-IRF-and-LTCH-Provider-Tabl/44n3-jbm8>

Data Processing Pipeline

Data sources 1 through 6 were joined by a combination of Federal Provider ID, 5-digit FIPS code (COVID-10 Social Vulnerability Index and NYT community infection and mortality), and were then processed, imputed, and feature engineered for our predictive modeling (additional details provided in modeling section). As we evaluated the model results, we developed better understanding of the data and progressed through multiple iterations.

The majority of the data processing and modeling was done using the open source program Python, augmented by Amazon Web Services (AWS) for the Natural Language Processing of inspection text data.

For the interactive dashboard, we preserved the original, un-imputed values from each dataset for accurate reporting purposes, and added data from sources 7 and 8 related to Tribal facilities and those with Native American/Alaska Native residents.



Interview Insights

- Michal Balass, Centers for Medicare and Medicaid Services:** Michal Balass provided the team with insights about the project's dependent variable as well as provided the team with the necessary resources to expand some limitations within the dataset. Michal's insight as an employee of CMS allowed the team to ask specific questions about life versus life years, best practice of adding a more in depth staff analysis, and the limitations of the demographic breakdown CMS provides within their datasets. She also provided insights on which features might be important for determining COVID-19 vaccine allocation, including number of residents in the facility, social vulnerability, and resident acuity.
- Nancy Zoints, Jewish Healthcare Foundation:** Nancy Zoints provided valuable information on the state of COVID-19 in Pennsylvania nursing homes, including the prevalence of infection among facilities regardless of CMS Quality Rating levels and the importance of staff as disease vectors. She also provided feedback on the important features that were likely *missing* from our dataset, including number of staff in facilities, number of staff working in multiple facilities, and cross-ward work protocols and their impact on infection spread. Finally, she emphasized to the team that vaccines alone cannot prevent infection in nursing homes, in part because of the lag between administration and full resistance. This prompted a renewed focus on the importance of PPE and testing capacity in facilities.
- Dr. Jonathan Caulkins, Carnegie Mellon University:** Jonathan Caulkins is a professor of Management Science at the Heinz College. Two students from this project are attending Multi Criteria Decision Making course and felt his expertise would add more depth in equitable vaccine allocation. The team presented their current findings and Professor Caulkins noted the current model is focused on an effective distribution of the COVID-19 vaccine with very little weights on equitable distribution of vaccine. Ultimately with his questions and criticism, we identified a path in our project to add a proof of concept for an equitable COVID-19 vaccine distribution within the dashboard.
- Rich Figallo, iCareNetwork:** Rich Figallo of iCareNetwork presented to our team early on in the project with a basic overview of Medicare, Medicaid, and skilled nursing facility concepts and terminology. He also participated in a one-on-one session where he provided feedback on manual feature selection based on domain knowledge, including which features from the LTCFocus and Provider Information datasets might be related to COVID-19 infection mortality and why.
- Ed Mortimer, Centers for Medicare and Medicaid Services (former):** Ed Mortimer provided valuable insight into CMS's decision-making process and how the agency approaches questions of equity.

Modeling Process and Results

This section describes the machine learning/predictive modeling process the team went through to identify nursing homes at high risk of COVID-19 infection and mortality.

Notes on Dependent Variables/Dashboard Variables

Over the course of this project, we experimented with multiple dependent “target” variables for our modeling, analysis, and interactive dashboard. See specific models for which dependent variables were used in each case. Each variable captures a different aspect of the complex COVID-19 dynamic in nursing homes. Any future analysis should carefully consider these when selecting target variables.

1. **Number of resident confirmed COVID-19 cases**: can capture number of residents affected by COVID-19, but may mask the impact in smaller facilities whose infections cannot compare to those of larger facilities by simple numbers.
2. **Average weekly infections per resident (# Occupied Beds)**: Can capture the prevalence of COVID-19 infections across facilities of all sizes, but may mask incidence of infection in large facilities (example: 10% infection rate in a facility with 500 residents is objectively more concerning than a 10% infection rate in a facility with 10 residents).
3. **Average weekly deaths per resident (# Occupied Beds)**: similar to number of resident confirmed COVID-19 cases variable above, but with respect to deaths.
4. **Cumulative case fatality ratio 1/1/2020 through most recent week** (total confirmed COVID-19 deaths divided by total confirmed COVID-19 cases and admissions): A limitation of this variable was that a lack of testing in the first several months of the pandemic likely depressed the number of confirmed infections in the denominator, effectively inflating the case fatality ratio.
5. **Cumulative case fatality ratio 6/21/2020 through most recent week** (total confirmed COVID-19 deaths divided by total confirmed COVID-19 cases and admissions) - designed to remedy the above issue with the “since 1/1/2020” version above. Selected 6/21/2020 as reference date because it had the lowest number of facilities whose data did not pass the CMS Quality Assurance Check (and therefore could not have this value calculated) in the month of June.

**Note that these variables only include resident infections and deaths. Unfortunately, we could not include staff in any of the ratio variables because CMS did not collect data on the number of staff in each facility, making it impossible to include them in the denominators.

Unstructured Text Data Modeling

For Natural Language Processing, we first cleaned and preprocess data. Next, we summarized the long narratives. There are two reasons: 1) It is too long to feed into the model. 2) We want to help readers refine the narratives so that they can quickly capture the key information without spending a long time reading the whole passage. Thirdly, we applied the topic modeling to choose an embedding method. Fourthly, we conducted a sentiment analysis to understand the value of the narratives. Finally, we applied CNN and BERT to predict the average mortality rate for each nursing home.

1) Text Pre-processing

- | | |
|---------------------------|---------------------------|
| 1. Case normalization | 6. Frequent words removal |
| 2. HTML Tag removal | 7. Rare words removal |
| 3. Special Symbol removal | 8. Tokenization |
| 4. Punctuation removal | 9. Stemming |
| 5. Stopwords removal | 10. Lemmatization |

2) Text Summarization

- a) **LSA** - LSA is designed to summarize the text and make it “readable” for machines. That is to say, punctuation words are removed, and the words are tokenized and lemmatize the words. For each text, we retained the top 3 important sentences.
- b) **TextRank** - This is for the use case of a dashboard. Therefore, we keep the original format of sentences. For each text, we retained the top 20% important sentences. Therefore, the number of sentences may vary depending on their original lengths.

3) Topic Modeling

The purpose of Topic Modeling was to determine whether Machine Learning could effectively match narrative text to the appropriate CMS Deficiency Category (including Residents Rights Deficiencies and Environmental Deficiencies).⁵ To categorize each long inspection text narrative into different deficiency topics, we applied several embedding methods to capture key information or characteristics among sentences so that they can be utilized in later stages of the prediction process. For example, when considering an inspection text, the hope is to classify the category of this inspection and using a specific algorithm to represent sentences as vectors. After clustering, we tried to find the 5 most similar inspection texts to test the performance. If more similar texts are in the same category, the embedding method is better. For Topic Modeling, we applied Doc2Vec, SentenceBERT, InferSent, and Universal Sentence Encoder. Among these methods, BERT and USE outperformed the other models (See Appendix A).

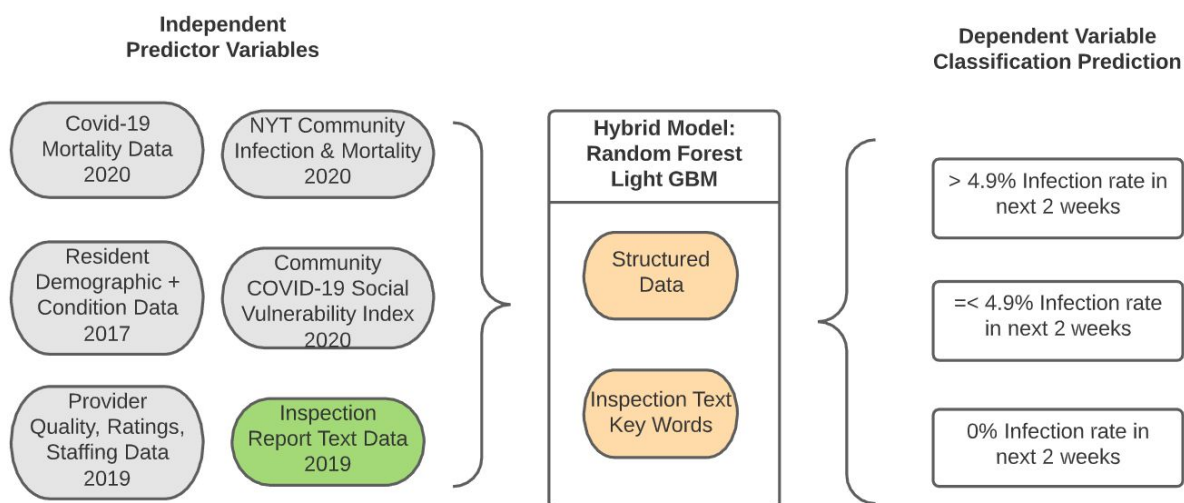
⁵“Health Deficiency,” Centers for Medicare and Medicaid Services, last modified November 01, 2020, <https://data.cms.gov/provider-data/dataset/r5ix-sfxw>

4) Mortality Prediction

- a) **Sentiment Analysis:** To figure out the value of text data, we conducted a sentiment analysis. We sent the inspection text into a “sentiment analyzer” to churn out a score. The positive score indicates that the nursing homes receives a mild complaint, and the negative score indicates the nursing homes received a serious complaint. We then calculated the correlation between the average sentiment score for each facility and the average mortality rate. The result -0.17 indicates that Sentiment score alone was not a strong predictor of COVID-19 mortality (slight negative correlation), but key words and sentences extracted through ML demonstrated some predictive value.
- b) **CNN:** Sentiment analysis told us that text data is useful, so we first tried Convolutional Neural Network (CNN) to make a prediction about Average weekly deaths per resident (# Occupied Beds). Although CNN is normally used in image setting, it can be applied in NLP as the sentence can be interpreted as a 1-dimensional image. The big advantage of CNN is it is fast given the limited budget and time in the mid-term. CNNs are also efficient in terms of representation and can learn good representations automatically, without needing to represent the whole vocabulary. The best accuracy for CNN is 62%.
- c) **BERT:** In the second phase of the project, we used AWS to run the NLP model. We implemented an advanced model - BERT based on the previous topic modeling experiment to predict Average weekly deaths per resident (# Occupied Beds). The BERT model’s key technical innovation is applying the bidirectional training of Transformer. BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. With BERT, the highest accuracy reached 72%.

BERT showed the best result in predicting deaths per number of residents. However, it is designed for natural language processing solely and could not process any structured features. Therefore, in the hybrid model, we utilized another method called TF-IDF to transform the text data so that we could combine both structured and unstructured information.

Hybrid Modeling - Combine Structured and Unstructured Data



*4.9% is average infection rate for the general population.

Hybrid Modeling: Structured Data

1) Data Preprocessing

- Data Filtration:** Using the Submitted and Quality Passed columns to filter out all records that meet both standards to ensure data quality.
- Imputation:** Within the CMS Facility COVID-19 (2020) dataset, for any bed quantity related columns, median values of the nursing home were used to impute missing values. For the duration related features related to Amount of Time to Obtain Test Results, mode values of the nursing home are used. For the rest of features within COVID-19 nursing home dataset, mean values of the nursing home are used. If missing value still exists, which indicates the nursing home has no data across all weeks, 0 is used to impute the cells.
- Feature Engineering:** To deal with the data collection and quality issue, we extract only the total case and death related features in the first week(May 24th 2020) as new independent variables. Also suspected but no test features respectively for residents and staff were generated to capture the testing capability of the nursing home.

- d) **Time Series Sliding Window & (In)Dependent Variable Construction:** A cut time was added every week for the dataset: Every 4 weeks before the cut time represented the independent variables for that week. Every 2 weeks after the cut time represented the dependent variable. By aggregating different window periods starting from June 7th 2020 to November 15th 2020, the processing method is designed to adapt training in every period.

The independent variables consisted of 4-week average of each time-series variable (for example, the community infections from the New York Times would be aggregated into 4 weeks worth of data for the cut time, and nursing home infection/mortality, testing capacity, and PPE supply varied over time), combined with the static facility conditions that did not change week to week, such as resident race/ethnicity percentage and provider quality rating.

The dependent variable is the prediction of the infection rate (Average weekly infections per resident, # Occupied Beds) aggregated over 2 weeks. This was done iteratively from the first week of the dataset to 2 weeks before the end date of the dataset to accurately capture the new independent variable and the new dependent variables with respect to the cut time.

Any infinity infection rate resulting from the missed occupied beds divisors were imputed using the max infection rate of all data. The numeric label was further grouped into three clusters using 0 and 0.049 as the cutoff. 0.049 is the general infection rate for US citizen which was calculated using the US total infections from the Centers of Disease Control and Prevention COVID-19 infections⁶ divided by the US population from the United States Census Bureau.⁷ The zero infection rate was invited as the first cutoff because any infection rate above zero indicates potential spread of COVID in the future. The general infection rate for US citizens was invited as the second cutoff with the regard that infection rate lower than general infection rate for a nursing home could be interpreted as medium risk.

- e) **Merging:** The initial merge involved CMS Facility COVID-19 (2020) ,LTCFocus (2017) and Provider Information (2020) using facility id, and returned facility level as each entry of the dataset. The next merge added the New York Times COVID-19 data and COVID-19 Community Social Vulnerability Index (2020) data, both joined on 5-digit FIPS code level.

⁶ "CDC COVID Data Tracker," Centers for Disease Control and Prevention, Updated December 20, 2020,

https://covid.cdc.gov/covid-data-tracker/#cases_casesper100klast7days

⁷ "U.S. and World Population Clock," United States Census Bureau, Updated December 20, 2020, <https://www.census.gov/popclock/>

2) Feature Selection

- a) **Removal Based on Domain Knowledge & Missing Value Percentage:** Based on health care and nursing home domain knowledge, 86 features including redundant features were removed (see Modeling Notebook for details). We also identified certain features which were essential to the project, such as resident % race/ethnicity, which we imputed based on the size of the facility (See Notebook for more details). After imputation of most features in the COVID-19 nursing home dataset, as well as selected features from the LTCFocus and Provider Information Datasets identified as important through domain knowledge (and at times with maximum 20% data missing), any remaining features with missing value percentage higher than 0% were removed.
- b) **Correlation Coefficient (VIF):** The Variance Inflation Factor (VIF) measures the severity of multicollinearity across features. VIF above 10 is considered as high collinearity. Based on the VIF test, 8 features were removed (See SAS Combined Model Classification Notebook for details).
- c) **Feature Importance:** The remaining 132 total features were fed into Random Forest and Light GBM, which generated a Feature importance index which we drew from to select the top 50 important features from structured data. The final lists of features were extracted from the aggregated feature importance rank of the two models. They were then fed into the combination model described below.

Hybrid Modeling: Unstructured Text Data TF-IDF

To incorporate both structured and unstructured, we applied TF-IDF to extract keywords. This is a preparation for creating a hybrid model. TF is a measure of how frequently a word occurs in a certain report, and IDF is a measure of how frequently this word occurs in all reports. By multiplying TF and IDF, we get the TF-IDF score.

We then applied TruncateSVD to reduce the dimension of vectors into 30 that are used to explain the location of words in the space.

Hybrid Modeling: Combination

Random Forest and Light GBM algorithms were applied in our modeling. 50 selected features from structured datasets and 30 features from unstructured datasets are combined to feed into this modeling. Before searching for the best parameter in each model, we designed a customized cross validation method using the last three periods of sliding windows. To be specific, the training process was validated using:

- [Independent variable time period] October 4th - October 30th data modeled on [dependent variable time period] November 1st - November 15th data;
- September 27th - October 24th data modeled on October 26th - November 8th data;
- September 20th - October 17th data modeled on October 18th - November 1st data.

We leveraged this customized cross validation method to best monitor the real data access situation.

Random Forest

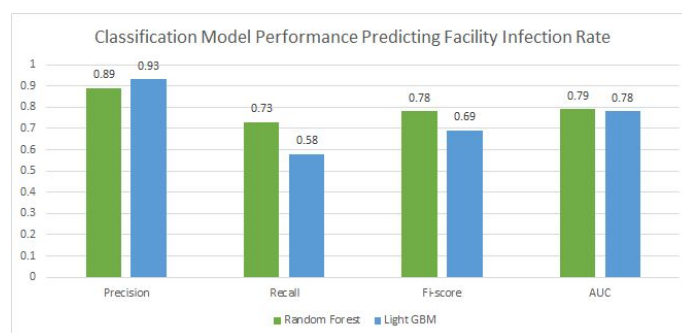
Random Forest is a tree based classification model that leverages ensemble to improve the model result. Using the gridsearch pipeline, the best parameter was reported as 'criterion': 'gini', 'max_depth': 15, 'max_features': 'log2', 'n_estimators': 200. Under this setting, the precision, recall, f1-score and AUC score of the model is 0.89, 0.73, 0.78 and 0.79 respectively.

Light GBM

Light GBM is another classification algorithm that learns data with a horizontal growing tree, which ensures a faster speed especially for big data. With the gridsearch pipeline, the best parameter was reported as 'criterion': 'gini', 'max_depth': 15, 'max_features': 'log2', 'n_estimators': 200. Under this setting, the precision, recall, f1-score and AUC score of the model is 0.93, 0.58, 0.69, and 0.78 respectively.

Final Model Selection - Light GBM

Comparing the random forest and Light GBM modeling results, it was clear from the performance results solely, that RandomForest gave us better results. However, after detailed comparison of the recall and specificity for each class, we found Light GBM provided us with 1.00 precision in the low risk group (zero infection rate).



		Predicted				
		Risk	Low	Middle	High	
Actual	Low	✓ 6752	2564	2332	0% Infection rate in next 2 weeks	
	Middle	! 0	✓ 285	230	≤ 4.9% Infection rate in next 2 weeks	
	High	! 0	199	✓ 208	> 4.9% Infection rate in next 2 weeks	

The above chart is the confusion matrix of our final selected model. Each row represents the number of facilities falling in the actual infection level category from Nov 08 to Nov 22 and each column represents the number of facilities predicted infection level. The three green colored cells with the check marks are the number of facilities we correctly classified. The two zero cells highlighted in orange indicate that the model was able to obtain 0 misclassification on the high or medium infection risk group into 0% infection rate low risk.

Since we do not want to underestimate the infection risk for any medium or high infection facilities, which could result in a delay of resource allocation and a further acceleration on the COVID-19 spread, zero misclassification on the predicted low category ensures no nursing home with over 0% infection rate would be ignored by our model.

Modeling Limitations and Future Steps

This modeling pipeline is a proof of concept, and provides a baseline for processing available nursing home data and inspection reports. However, after careful consideration, we do not recommend applying the results directly into resource allocation strategies due to the following limitations. If you are interested to build up from the existing framework, we recommend referring to the future steps we listed to consolidate the model.

Limitations:

- Imbalanced data
- Novel pandemic with limited historic data
- Not involving policy and human behavior
- Only considering short term effect of historic time series data on the current risk
- Important features from modeling still merit further examination

Future Steps:

- Consolidate label classification involving both infection and mortality to present the comprehensive risk
- Consider both long term and short term historical data impact
- Add local level policy decisions data
- Validate important features using diverse methods

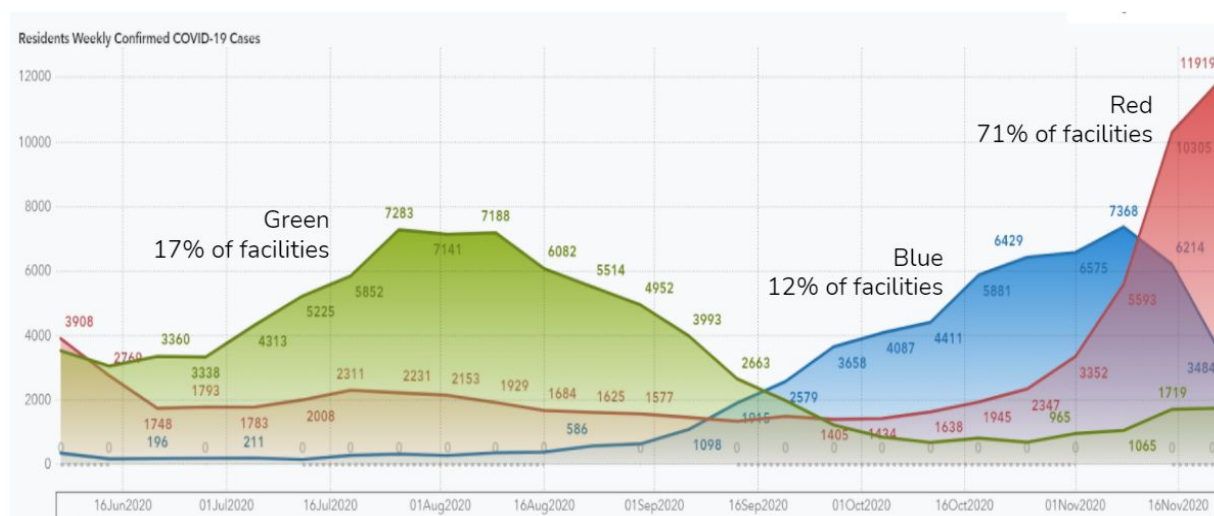
Group-Based Infection Trajectory Modeling

Although our hybrid predictive classification modeling results were not strong enough to stake a recommendation on, we were still able to use the COVID-19 nursing home data to capture historic progression of infections. Using a Group-Based Trajectory Model (GBTM)² developed by Daniel Nagin at Carnegie Mellon, we clustered all nursing homes into 3 infection trajectory groups based on their longitudinal infection counts from July - November 2020, using target dependent variable "Number of resident confirmed COVID-19 cases".

This TRAJ model revealed that facilities' resident infections peaked at different times. This modeling process is useful for distinguishing infection trends in order to target facilities currently experiencing high infection counts, and allows comparisons between states over time. The clusters were later used in the interactive dashboard.

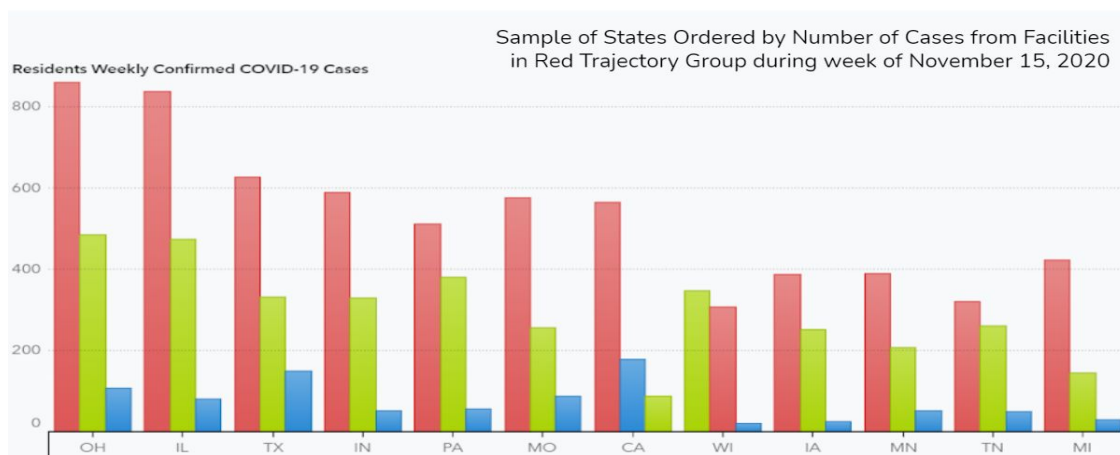
The green trend which makes up 17% of the facilities, peaked in mid-summer. The blue which makes up 12%, peaked in November but appeared to taper off. And finally, in the red group, which makes up 71%, the number of new cases each week was relatively stable from July onward, but starting around mid to late October, they appeared to experience a surge in infection cases still on an upward trajectory.

TRAJ Classification Shows Varied Past Infection Trends

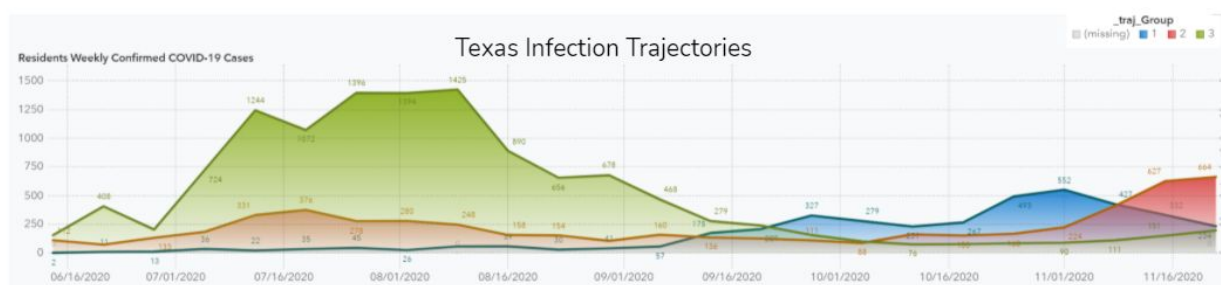
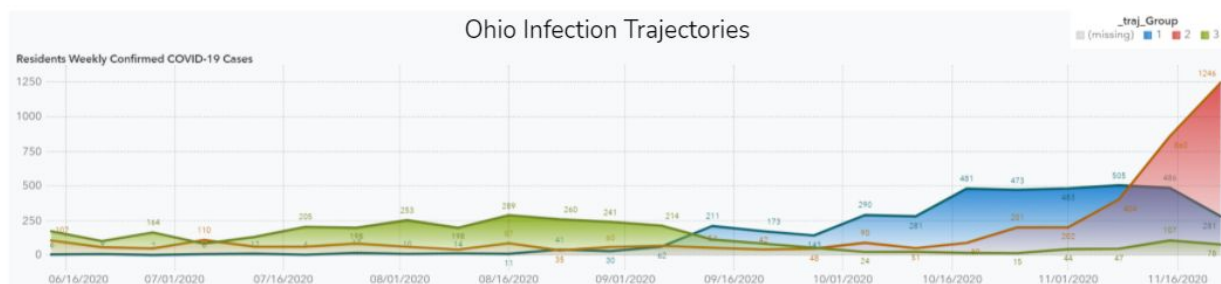


To investigate the red group further, we examined a 1-week snapshot of the number of infections in the week of November 15, 2020. By number of cases, the surge of infections in the red group is being driven by facilities in Ohio, Illinois, Texas, Indiana, Pennsylvania, Missouri, and California, followed by Wisconsin, Iowa, Minnesota, Tennessee, and Michigan. Some of these states have a large number of nursing homes to begin with, such as Texas and California, but it is noteworthy that even states like Iowa with relatively fewer facilities are contributing substantially.

Sample of States Driving High Case Count in Fall Surge



It is also important to note that these trends are not uniform within states. For example, Ohio's trajectory group mix shows most activity in the fall, whereas Texas's infection trajectories show a more significant spike in the summer followed by a noticeable but lower upward trend in the fall. Ultimately, this combination of GBTM modeling and interactive dashboard can facilitate comparison of infection trends over time between states.

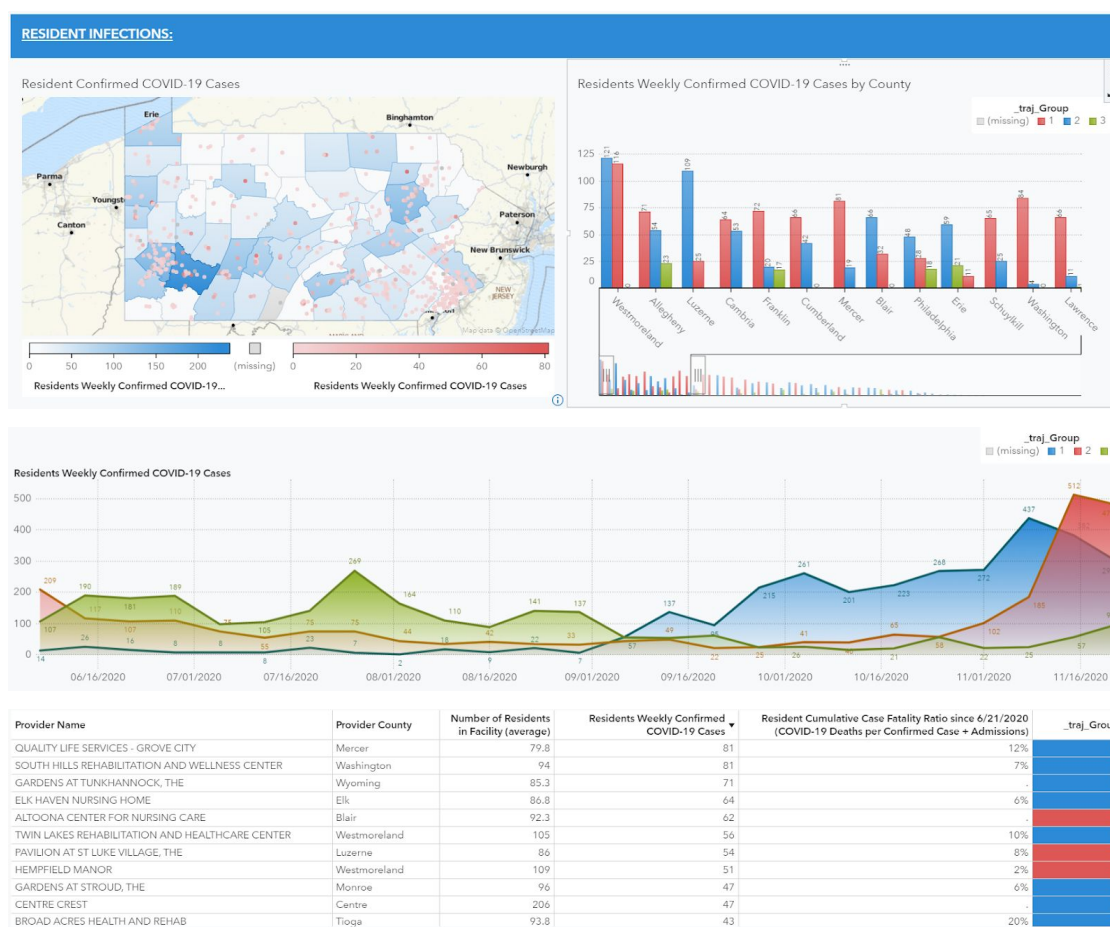


Interactive Dashboard

After completing the modeling process, we produced tangible deliverables related to each of the three use case objectives through our interactive dashboard. **Constructed using SAS Viya and SAS Visual Analytics, this dashboard combines all of the work we have done into a practical, useful tool to visualize complex problems and identify actionable solutions in a rapidly changing environment.**

Targeting:

Identify nursing homes with high infection rates and recommend resource supports such as testing supplies and PPE.



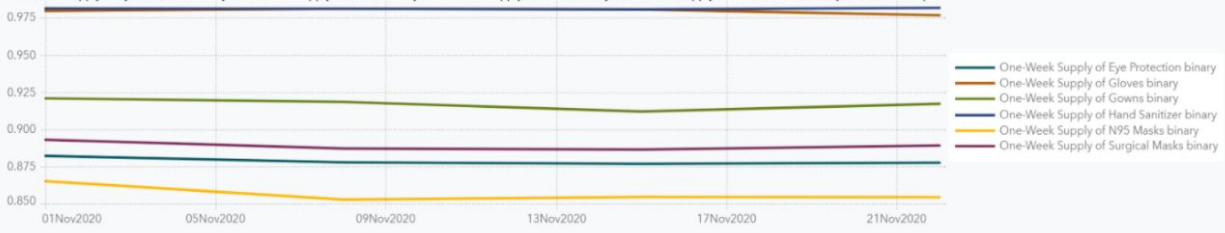
For the Targeting objective, we created a dashboard page that provided an overview of infection at the county and individual facility level. This allows state-level public health agencies and decision-makers to quickly identify which facilities/areas are experiencing high infection incidence so that they can target facilities needing resources to care for COVID-19 positive residents and prevent further spread. Facilities were grouped by Group-Based Trajectory Model⁸ clusters to aid trend comparison within states.

⁸ Jones, B. and Nagin, D., 2013, A Note on a Stata Plugin for Estimating Group-based Trajectory Models, Sociological Methods & Research, 42(4), pp.608-613.

PPE SUPPLY: Contact the facilities with less than 100% record of 1-week supply of PPE items in the last 4-6 weeks. Provide them with the needed resources.

One-Week Supply of Assorted PPE

One-Week Supply of Eye Protection binary / One-Week Supply of Gloves binary / One-Week Supply of Gowns binary / One-Week Supply of Hand Sanitizer binary / One-Week Supply of N95 Masks binary / One-Week Supply of Surgical Masks binary



Provider County	Provider Name	One-Week Supply of Eye Protection binary	One-Week Supply of Gloves binary	One-Week Supply of Gowns binary	One-Week Supply of Hand Sanitizer binary	One-Week Supply of N95 Masks binary	One-Week Supply of Surgical Masks binary	One-Week Supply of Ventilator Supplies binary
Montgomery	GARDEN SPRING NURSING AND REH...	.92	.92	.92	.92	.92	.92	1
Berks	LAUREL CENTER	.24	.96	.24	1	0.2	.24	1
Cumberland	FOX SUBACUTE AT MECHANICSBURG	.96	.96	.96	.96	.96	.96	1
Lycoming	MUNCY PLACE	1	1	.96	1	0	1	1
Montgomery	ARISTACARE AT MEADOW SPRINGS	.76	1	.28	1	.16	.96	1
Delaware	AVENTURA AT PROSPECT	1	1	1	1	1	1	1
Butler	ST JOHN SPECIALTY CARE CENTER	1	1	1	1	1	1	1
Lackawanna	ALLIED SERVICES SKILLED NURSING...	1	1	0.9	.95	.95	1	1

Example Interpretation:

1 = One-week supply of item 100% of the time within selected weeks.

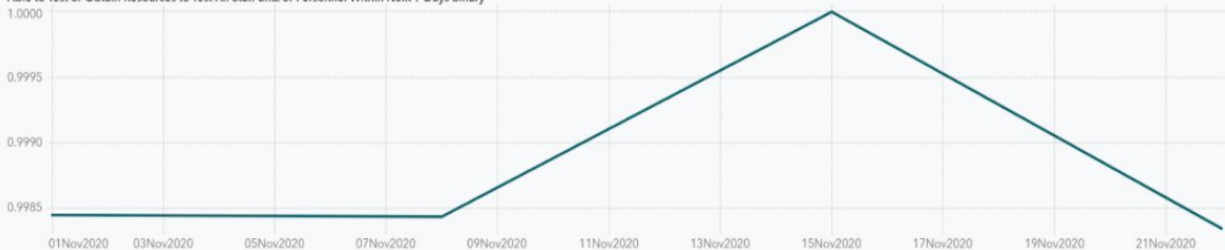
0.85 = One-week supply of item 85% of the time within selected weeks.

For allocation of infection prevention resources we included an element on personal protective equipment (PPE) supply. This graphic allows us to see which percentage of facilities in a given state have a one week supply of each PPE item. Then it highlights which specific facilities have lacked supply of each PPE item what percent of the time.

TESTING CAPACITY: Contact facilities unable to test all staff 100% of the time in the last 4-6 weeks. Address their reasons for not testing and provide them necessary resources.

Able to Test or Obtain Resources to Test All Staff and/or Personnel Within Next 7 Days binary by Week

Able to Test or Obtain Resources to Test All Staff and/or Personnel Within Next 7 Days binary



Provider Name	Provider County	Able to Test or Obtain Resources to Test All Staff and/or Personnel Within Next 7 Days binary	Reason for Not Testing Staff and/or Personnel - Lack of Access to Tira...	Reason for Not Testing Staff and/or Personnel - Lack of Access to Lab...	Reason for Not Testing Staff and/or Personnel - Lack of PPE for Perso...	Reason for Not Testing Staff and/or Personnel - Lack of Supplies binary	Reason for Not Testing Staff and/or Personnel - Other binary	Reason for Not Testing Staff and/or Personnel - Uncertainty About Re...
GINO J MERU VETERANS CENTER	Lackawanna	0.00	0.00	0.00	0.00	1.00	0.00	0.00

Able to Test ... All Staff in Next 7 Days:

0.0 = Able to test all staff in the next 7 days 0% of selected weeks.

0.75 = Able to test all staff in the next 7 days 75% of selected weeks.

Reason for Not Testing:

0.0 = This was not a reason for not testing during selected weeks.

0.75 = This was the reason for not testing staff in 75% of selected weeks.

1.0 = This was the reason for not testing in 100% of the selected weeks.

We included an element on staff testing capacity as staff are the primary vectors of infection from communities to residents. This graphic allows us to see what percentage of facilities in a state have been able to test all of their staff in the next 7 days. It then highlights which specific facilities have not been able to test all of their staff and why. Public health officials can then provide those facilities the resources they need.

DATA QUALITY ASSURANCE: Contact the facilities with less than 100% record of Passing Quality Assurance Check in the last 4-6 weeks and review submission protocols.

Passed Quality Assurance Check binary, Submitted Data binary by Week

Passed Quality Assurance Check binary / Submitted Data binary

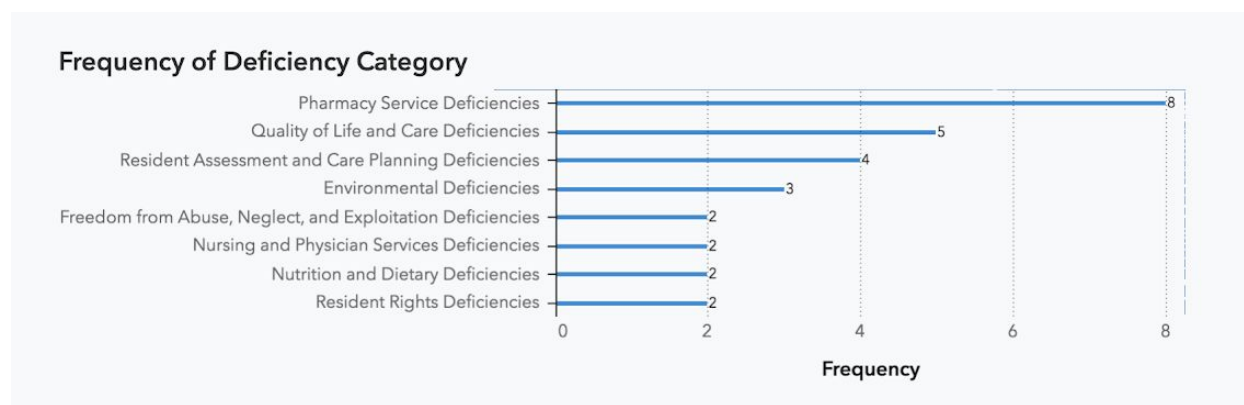
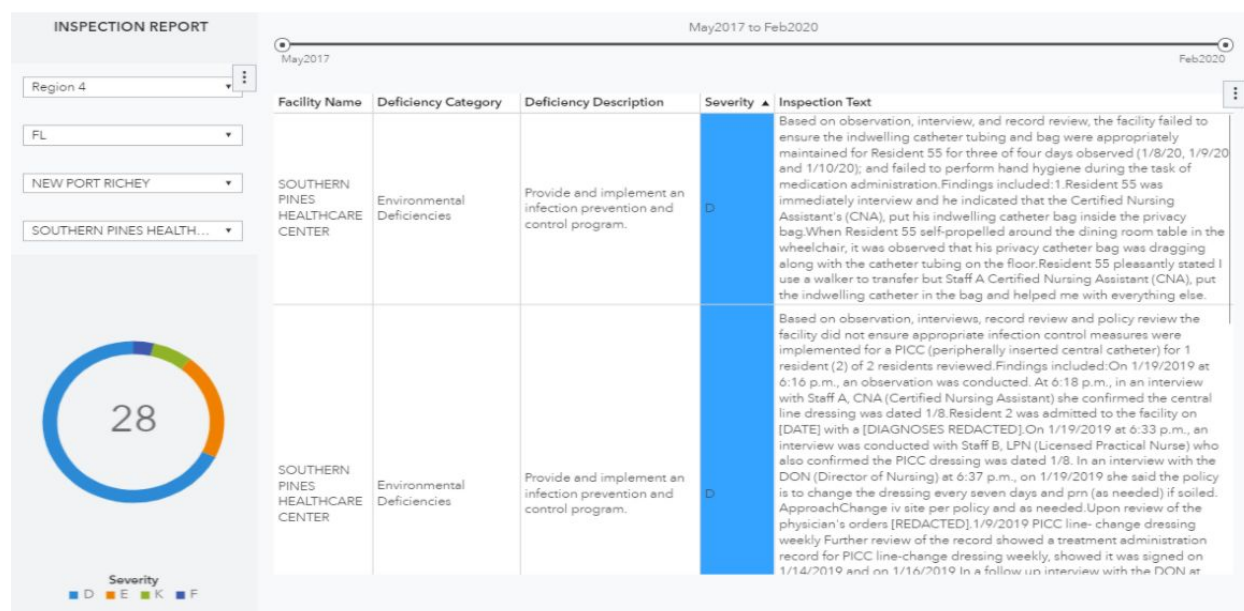


Provider County	Provider Name	Submitted Data binary	Passed Quality Assurance Check binary ▲
Carbon	SUMMIT AT BLUE MOUNTAIN NURSING...	100%	0%
Westmoreland	GREENSBURG CARE CENTER	25%	0%
Cumberland	CUMBERLAND CROSSINGS RETIREMEN...	100%	0%
Luzerne	GARDENS AT EAST MOUNTAIN, THE	100%	0%
Fulton	FULTON COUNTY MEDICAL CENTER	100%	0%
Luzerne	TIMBER RIDGE HEALTH CENTER	100%	0%
Crawford	EMBASSY OF PARK AVENUE	100%	25%
Centre	VILLAGE AT PENN STATE, THE	100%	25%
Butler	QUALITY LIFE SERVICES - CHICORA	100%	25%
Cumberland	BETHANY VILLAGE RETIREMENT CENTER	100%	25%
York	SPIRITRUST LUTHERAN THE VILLAGE A...	100%	25%
Centre	EMBASSY OF HEARTSIDE	100%	25%
Allegheny	BRIDGEVILLE REHABILITATION & CARE...	100%	50%
Burks	PHOEBE RICH AND HCC	100%	50%

The dashboard also provides a data quality assurance control section. The graph shows what percent of facilities in the state have either failed to submit data or submitted data that did not pass the Quality Assurance Check in selected weeks. This means CMS has flagged and removed their data due to inconsistencies, and we do not have their infection rates, number of deaths, and other important information. It then highlights the recent quality records of specific facilities with this issue so that public health officials can reach out to those facilities and correct their data and ensure that their residents and facility conditions are accounted for.

Investigation:

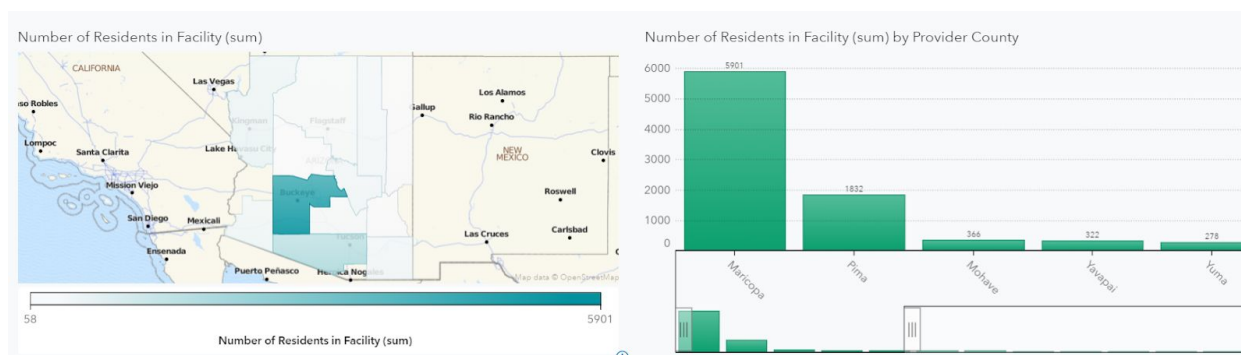
Summarize nursing home inspection report narratives for more efficient review by investigators.



For the Investigation use case, we created an inspection report dashboard page to summarize inspection report text and visualize Deficiency Tags and the scope and severity of deficiencies. The reports include descriptions of facility conditions, experiences of residents, and investigations into allegations of abuse and neglect, among other things. There are some useful features such as severity score and deficiency categories. The table listed the facility name, deficiency category, severity score and summarized texts. The pie plot showed the distribution of severity score of that specific nursing home. The bar plot showed the distribution of nine categories in a descending order. Users can choose the region, state, city and specific nursing home they are interested in, with a slider to specify the date range. This dashboard page enabled analysts to quickly understand the key idea of the inspection report and the overall deficiency status of the facility.

Intervention:

Operationalize potential equitable vaccine allocation strategy in interactive dashboard.



Tribal Facilities & Non-Continental Facilities

Facilities in the CMS Tribal Nursing Home Collaborative for which COVID-19 data was reported; Select CMS facilities with >10% Native American/Alaska Native Residents for which COVID-19 data was reported. Facilities in HI, AK, PR, and GU.

Tribal Facilities and Select CMS Facilities with >10% Native American/Alaska Native Residents for which COVID-19 data was reported.

IMPORTANT NOTES:

(1) As sovereign nations, each tribe has the right to choose which entity they will receive vaccine doses from, including states or the Indian Health Services (IHS). Not every tribal nursing home/elder care facility is accounted for by CMS or in this dashboard. Vaccine allocation decision-makers must consult with individual tribes to ensure tribal long-term care facilities are included and prioritized.

See the IHS COVID-19 Pandemic Vaccine Plan November 2020 for more details:

https://www.ihs.gov/sites/newsroom/themes/responsive2017/display_objects/documents/2020_Letters/Enclosure_DTLL_DUIOLI_11182020.pdf

Provider Name	Provider State (1)	Number of Residents in Facility (sum)	Tribal Nursing Home	Percent American Indian or Alaska Native Beneficiaries	Resident Weekly Infection Rate (sum)	Resident Cumulative Case Fatality Ratio since 6/21/2020 (COVID-19 Deaths per Confir...
THE PEAKS HEALTH & REHABILITATION	AZ	31	N	39%	0%	-
SOUTH MOUNTAIN POST ACUTE	AZ	94	N	18%	0%	4%
ALLEGIAN HEALTHCARE OF PHOENIX, LLC	AZ	80	N	11%	0%	0%
CATALINA POST ACUTE AND REHABILITATION	AZ	79	N	11%	15%	2%
CHANDLER POST ACUTE AND REHABILITATION	AZ	94	N	15%	20%	9%
CAMELBACK POST ACUTE AND REHABILITATION	AZ	79	N	12%	0%	18%

Facilities Ranked by Number of Residents, Share of Black/Hispanic residents, Resident Acuity, and Community COVID-19 Social Vulnerability.

Primary rank facilities by Number of Residents highest to lowest with % Black/Hispanic residents as tie-breaker highest to lowest.

Among facilities missing % Black/Hispanic, Average Resident Acuity as tie-breaker highest to lowest.


Among facilities for which we lack any demographic or acuity data, ranked by Community COVID-19 Social Vulnerability Index highest to lowest.

Provider Name	Provider State (1)	Number of Residents in Facility (average)	2017 % Black or Hispanic Residents	2017 Average Acuity Index *	County Social Vulnerability Index Score (0-1)
DEVON GABLES REHABILITATION CENTER	AZ	186	21%	12.978	0.391
MONTECITO POST ACUTE CARE AND REHABILITATION	AZ	184	-	13.578	0.287
IMMANUEL CAMPUS OF CARE	AZ	156	19%	11.973	0.287
NORTH MOUNTAIN MEDICAL AND REHABILITATION CENTER	AZ	147	22%	17.439	0.287
BELLA VITA HEALTH AND REHABILITATION CENTER	AZ	147	-	12.391	0.287
HORIZON POST ACUTE AND REHABILITATION CENTER	AZ	137	35%	12.356	0.287
PLAZA HEALTHCARE	AZ	135	23%	17.449	0.287
PALM VALLEY REHAB & CARE CTR	AZ	133	39%	13.121	0.287
LIFE CARE CENTER OF NORTH GLENDALE	AZ	125	-	12.739	0.287
ESTRELLA CENTER	AZ	123	43%	13.069	0.287

For the final use case, Intervention, we created a dashboard page that captures the vaccine dose needs of each county/facility, and can rank order facilities for equitable vaccine allocation in the event of limited doses available for nursing home residents.

It first maps the number of nursing home residents in each county who may be able to receive a vaccine. It can then order facilities for vaccine allocation based on assorted equity based attributes. For example, the National Academy of Sciences Framework for Equitable Allocation of Vaccine for the Novel Coronavirus lists saving as many lives as possible as a priority, but also emphasizes the importance of prioritizing high-risk groups.⁹ To reflect this,

⁹ "A Framework for Equitable Allocation of Vaccine for the Novel Coronavirus," The National Academies of Sciences Engineering Medicine, Accessed on December 20, 2020. <https://www.nationalacademies.org/our-work/a-framework-for-equitable-allocation-of-vaccine-for-the-novel-coronavirus>



our dashboard can rank order facilities based first by facility size, and among facilities of similar size, allocate vaccines first to facilities with higher shares of Black and Hispanic residents (which we know are at higher risk for infections and deaths).¹⁰

To our knowledge, this would be the only publicly available dataset/tool that links COVID-19 in nursing homes with shares of Black and Hispanic residents, community social vulnerability, and any markers of Tribal affiliation or indigenous resident shares at the national level. Given the disproportionate impact of COVID-19 on these groups, this information is crucial for making equitable allocation decisions.

This flexible dashboard allows state officials to incorporate equity-focused components into their vaccine allocation strategy and assess tradeoffs between one approach in a simple fashion that can be easily explained to stakeholders. The dataset included in this section also allows for more complex Multi-Criteria Decision Making approaches, although they were not incorporated here.

¹⁰ Chidambaram, P., Neuman, T., & Garfield, R. (2020, October 27). Racial and Ethnic Disparities in COVID-19 Cases and Deaths in Nursing Homes. Retrieved December 09, 2020, from <https://www.kff.org/coronavirus-covid-19/issue-brief/racial-and-ethnic-disparities-in-covid-19-cases-and-deaths-in-nursing-homes/>

Conclusions and Recommendations

Modeling

The team was able to combine comprehensive structured data on nursing homes with unstructured text from past inspection reports to provide a robust data pipeline and an early look at predicting nursing home infection risk. The data pipeline we have built is agile and highly automated. It can be applied over many months into the future of the pandemic in the United States. We recommend that our client should take advantage of this data pipeline and use it in the future for similar data analysis.

Unstructured Text Data Modeling was an effective and efficient way to extract key information from long inspection reports and turn unstructured data into summarized tabular format with categories. It eliminates the huge labor cost of manually reviewing inspection reports and leverages the value of textual data with natural language processing techniques. Policy makers should use NLP to review inspection reports more efficiently in the future to understand the unique situation at each facility.

Hybrid Modeling that combined structured data and inspection reports had limited predictive power on future infection. At least, it was not strong enough to be used as a targeting tool for identifying potential high risk facilities. One main reason for the suboptimal result is that the model does not capture local level policy that could potentially influence infection spread. It also does not include any human interaction information between residents, staff as well as other personnel in the nursing homes. These factors have a high impact on the virus spread and should be investigated. We recommend that going forward, these factors should be incorporated into the model for best predictive outcome. In specific, the model should capture state/local policy decisions that could influence infection. We also suggest that officials should start to collect data on infection prevention protocol and human interactions to improve the model accuracy. Also, We recognize that COVID-19 is a new virus and there is very limited data available to represent the full story. It is hard to make accurate predictions when we have few ideas of what the pattern should look like.

The Group-Based Trajectory Model was an unique way of clustering nursing homes based on their past infection trend. It made a strong point that facilities experience surges at different times. It could be in the past, right now or even in the future. For this reason, this clustering model could potentially be a methodology to assign different levels of risk.

Dashboard

The team is grateful that our client exposed us to SAS Viya, a great visualization and analytics dashboard tool that allows us to summarize and showcase our deliverables in the most compelling way. We were able to leverage the dashboard and accomplish our three main project objectives: 1. We used the dashboard as a tool to target facilities with high infection or in need of testing/PPE resources. 2. We used the dashboard to display highly summarized information from inspection reports for policy review and investigation. 3. We used the dashboard to produce a ranked list of facilities for vaccine distribution priority decision based on criteria like facility size, share of minority population.

The team firmly believes that the dashboard is the best way to summarize all essential information in such a complex pandemic situation. Once implemented by our client, it can be used to inform operational decisions, policy decisions as well as vaccine allocation decisions.

Acknowledgements

Rema Padman, Carnegie Mellon University

Manuel Figallo, SAS Institute

Ed Mortimer, Center for Medicare and Medicaid Services

Michal Balass, Center for Medicare and Medicaid Services

Nancy Zoints, Jewish Healthcare Foundation

Rich Figallo, iCareNetwork

Jonathan Caulkins, Carnegie Mellon University

Appendix A: Topic Modeling

Doc2Vec

By using Doc2Vec, each corpus is represented as a TaggedDocument which contains a list of words and a tag associated with it. We further cluster inspection texts and test on a new inspection text to find the similar inspection texts. However, this algorithm is not suitable for our dataset because of the low accuracy in the test dataset.

SentenceBERT

Sentence-BERT uses a Siamese network and transformer architecture to encode the sentence. we first loaded the pre-trained BERT model, and displayed the sentences as vectors. Then we retrieved the similar sentences for test data using cosine similarity. This algorithm demonstrates high efficiency and accuracy.

InferSent

InferSent uses GloVe vectors for pre-trained word embedding. Similar to SentenceBERT, we first downloaded the pre-trained word vectors and saved in the working directory. Then we built the vocabulary from our dataset and tested the performance in our test set. However, this method turned out to be the slowest and least efficient method.

Universal Sentence Encoder

Universal Sentence Encoder has two types of encoder, transformer and deep average network (DAN). we first tokenized the sentences and got each sentence converted to a 512-dimensional vector. Then, we try to classify each inspection text. This method also demonstrated outstanding performance.

Appendix B: Additional Dashboard Country Overview

These graphics were provided in the dashboard's introductory page alongside the two Group-Based Multi Trajectory classification figures shown in this report.

