COVID-19 US Nursing Home Forecasting and Visualizing

Introduction

The COVID-19 pandemic has wreaked havoc across the globe, and there are more than 100 million people that had been diagnosed infected within a short time. Currently, the number of COVID-19 cases and deaths among U.S. nursing home residents has been always above the average. Prediction of COVID-19 infection risks could help policy makers and nursing home owners respond to future epidemics swiftly. While various COVID-19 prediction models and research papers on nursing home risk factors are reported, no tools are available to predict the future outbreaks for nursing homes.

Problem Definition

To help nursing homes to act swiftly on future outbreaks and health policymakers to analyze the facility's responses to infections and mortality, our team plans to 1) build a machine learning model to predict short-term infection risk of nursing homes and 2) create a data visualization tool, i.e. an interactive collection of choropleth maps, to visualize both current statistics and predicted infection risk.

Literature Survey

Based on the available data and published reports, we have finished the literature survey mainly focused on the following three aspects:

1. Nursing home residents' data analytics and visualization.

Millions of nursing home data can be accessed from the government dataset. However, Federal data undercounts total COVID-19 cases and deaths in nursing homes due to the delays in federal data starting. Karen Shen et al. reported a statistical model using fuzzy-matching and geocoding methods, which could separate the nursing homes' COVID-19 cases and deaths data from the total Database. This method presented the nursing house non-reported cases and deaths with color gradient percentage by state mimic map, which is clear to read and analyze[1]. In addition, Tableau is a powerful tool for visualizing massive data and to create map-based visualization[2]. These researches have limitations because they only estimated the COVID-19 cases and deaths of reported data, but there is no correlation analysis and prediction for the nursing homes.

2. Infection and mortality risk factors.

It is important to understand how the transmission mechanisms within such facilities are different from the transmission in regular communities. Shamik Giri et al. reported key factors to include residents' and facility characteristics, staff-related factors, and external factors such as PPE shortage3, Asymptomatic transmission.[3] Debasree Das Gupta et al. studied the correlation between nursing home quality and mortality rate through the multivariable regression methodology, including many covariates, such as ownership, facility size, and staff shortages time[4]. Christopher J. Cronin et al. used a negative binomial model to study the correlation between nursing home quality and mortality rate. They also control for the following: under 65 years old, black, Hispanic, and on Medicaid, whether the home is for profit, and acuity index[5]. R. Tamara Konetzka et al. summarized the most important predictors of facilities having COVID-19 are larger bed size, facility location, facility racial composition, and staff numbers[6]. Idoia Beobide et al. concluded that people with advanced dementia are less likely to be infected, but they have a higher risk of COVID-19 mortality. People who take drugs such as antipsychotics would have a higher risk of mortality once infected[7]. Margaret M Sugg et al. reported nursing homes that have more COVID-19 cases when they are in high-density communities and counties with a more significant proportion of minority residents[8]. M. Keith Chen et al. concluded that share workers were an important factor in infections in nursing homes [9]. These papers used many tables to demonstrate the data, which is hard to follow.

3. COVID-19 prediction methods.

It's important to study the critical value of COVID-19 cases to provide a prediction model in nursing homes and provide an alarm if it exceeds the nursing house care capability. Christopher L.F. Sun et al.

built machine learning models to evaluate the risk of covid-19 infection in nursing homes using self-constructed data reported on April 2020[10]. Xin Wang et al. developed supervised machine-learning algorithms on multiple digital metrics including symptom search trends, population mobility, and vaccination coverage in the UK[11]. Megan Mun Li et al. used COVID-19 case number history, demographic characteristics, and social distancing policies both independently/interdependently to predict county-level cases[12]. Ben R. Craig et al. compared two approaches when forecasting the path of the COVID-19 pandemic: curve fitting versus structural[13]. Matheus et al. developed efficient short-term forecasting models that allow forecasting the number of future cases[14]. Haoran Dai et al. explored a long-term forecasting model and modeled it at the segment level for better accuracy[15]. Youssoufa Mohamadou et al. comprehensively reviewed the methods used in studies of the dynamics and early detection of COVID-19 via mathematical modeling and Artificial intelligence (AI)[16]. Ruifang Ma et al. combined LSTM and Markov model, which could improve the prediction accuracy effectively[17]. In summary, techniques used for COVID-19 prediction include ARIMA models, Logistic Regression, Neural Networks, XGBoost, SEIR models, LSTM models and GLEM models[18]. However, research on nursing homes level prediction is very limited due to delay in data reporting.

Proposed method

Our innovation includes:

- Reliable and high quality data source downloaded from CMS.gov, which follows federal reporting guidelines and is updated weekly;
- Nursing home level COVID-19 historical and predicted data visualization;
- Combined risk factors from various literatures for model prediction.
- Use Machine Learning Algorithm to predict the COVID-19 infection risks for each nursing home.

1. Risk factors from the literatures, Data availability and Data source

The risk factors reported from literatures can help us build a better model, which includes facility characteristics, staff-related factors, shortage, PPE, facility size, vaccination, Residents related factors, COVID treatment, County level data, facility location, high-density communities. Most of the risk factors data are available and downloaded from CMS government (COVID-19 nursing home level data and nursing home characteristic data, 752MB), LTCfocus.org (additional nursing home characteristics data, 7.1MB) and nytimes (county level COVID-19 data, 13MB). Here, the detailed data sources are described in the Appendix table.

2. Data pre-processing

Python and Pandas, NumPy, Scikit-learn libraries are used for data processing.

- Convert nursing home locations to latitude and longitude for map visualization:

To visualize each nursing home on map, latitude and longitude for each nursing home were retrieved by Google geocoding API using the full address of each nursing home.

- Combine datasets from different data sources:

Latitude and longitude of each nursing home are combined with Nursing home COVID-19 data, and then merged with nursing home characteristics data. Nursing homes with no characteristics data (possibly due to shutdown) were removed. Then combine with county level COVID-19 data (aggregated weekly from daily data).

- Deal with missing data:

For fixed values, forward fill them using the latest value available, then backward fill if still missing. For values which are not expected to move drastically in a short period of time, forward fill only. Skipping backward fill to avoid peeking into the future.

3. Algorithm selection

The programming language we used for modeling was python. We also used Google Colab as our modeling platform which enabled us to collaboratively work on the models.

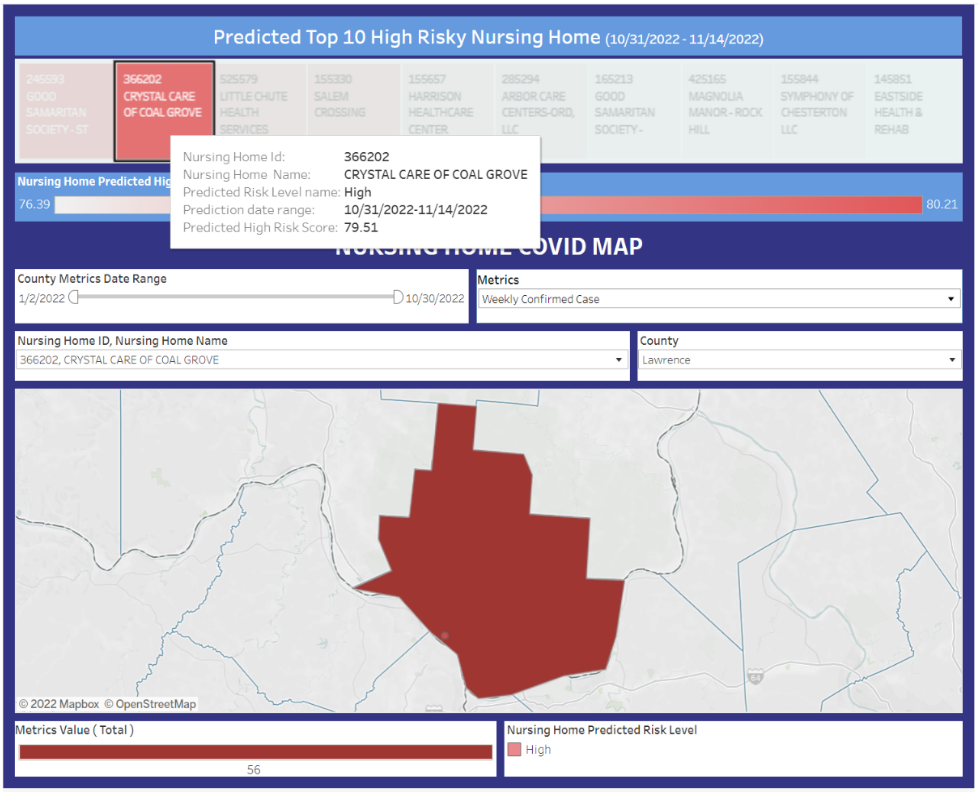
We used the LightGBM algorithm for modeling. LightGBM is a tree based gradient boosting framework which is well known for low memory usage and efficient training. We also tested the Long Short-Term Memory (LSTM) model. However, the model did not perform well likely due to: 1) too many 0 in the input target series which prevents the model from memorizing useful information in the input data; 2) limitation of ram and CPU usage in Google Colab given it is a free and public platform.

4. Visualization design

Data pre-processing generated the time series historical data portion for visualization plot, then model predicted result (next two weeks' risk level at nursing home level) merged back to historical data to produce a concise time series visual dataset for building visualization dashboard.

Tableau is used to build visualization, tableau data extract produced from visual data, and then we published our data and dashboard on tableau public, which could be accessed from this link https://public.tableau.com/app/profile/ruby1883/viz/team88-final-project/Dashboard1?publish=yes. Tableau workbook could be downloaded from this link too. Detail of the dashboard design is as below:

On the top of the dashboard, A tree map with sequential color is used to visualize the top 10 risky nursing homes from model predicted results. This could alert nursing homeowners to get better prepared. Tooltips will show up with more details (high risk score, risk level etc.) when hovering over these nursing homes. COVID-19 map at the bottom will be filtered if you click these top facilities, which allow users to have geographics view and investigate the map for more details.

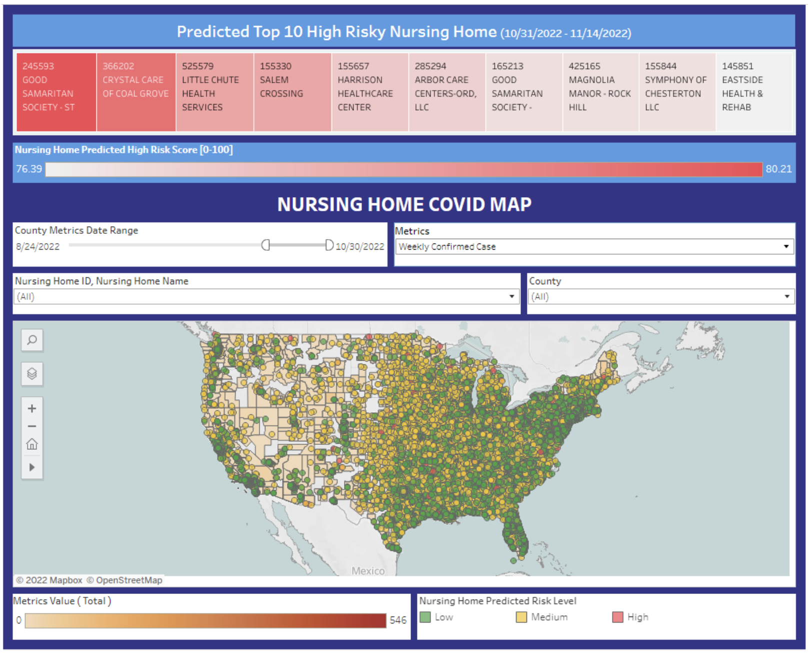


Predicted Top 10 High Risky Nursing Home (10/31/2022 - 11/14/2022)
245593 GOOD SAMARITAN SOCIETY - ST
366202 CRYSTAL CARE OF COAL GROVE
525579 LITTLE CHUTE HEALTH SERVICES
155330 SALEM CROSSING
155657 HARRISON HEALTHCARE CENTER
285294 ARBOR CARE CENTERS-ORD, LLC
165213 GOOD SAMARITAN SOCIETY -
425165 MAGNOLIA MANOR - ROCK HILL
155844 SYMPHONY OF CHESTERTON LLC
145851 EASTSIDE HEALTH & REHAB
Nursing Home Predicted Hig
76.39
80.21
NURSING HOME COVID MAP
© 2022 Mapbox © OpenStreetMap
Metrics Value (Total)
56

Nursing Home Id: 366202
Nursing Home Name: CRYSTAL CARE OF COAL GROVE
Predicted Risk Level name: High
Prediction date range: 10/31/2022-11/14/2022
Predicted High Risk Score: 79.51

County Metrics Date Range: 1/2/2022 – 10/30/2022
Metrics: Weekly Confirmed Case
Nursing Home ID, Nursing Home Name: 366202, CRYSTAL CARE OF COAL GROVE
County: Lawrence
Nursing Home Predicted Risk Level: 🟥 High

At the bottom of the dashboard, a nursing home COVID-19 map combined historical data and predicted risk level information. Several filters were added for users to filter the map, including nursing home id and name, county, date range, metrics (a metric swap filter controls metric used to color map, including confirmed case and death), and also predicted risk level at the bottom right could be used to highlight the map.


Predicted Top 10 High Risky Nursing Home (10/31/2022 - 11/14/2022)

245593 GOOD SAMARITAN SOCIETY - ST
366202 CRYSTAL CARE OF COAL GROVE
525579 LITTLE CHUTE HEALTH SERVICES
155330 SALEM CROSSING
155657 HARRISON HEALTHCARE CENTER
285294 ARBOR CARE CENTERS-ORD, LLC
165213 GOOD SAMARITAN SOCIETY -
425165 MAGNOLIA MANOR - ROCK HILL
155844 SYMPHONY OF CHESTERTON LLC
145851 EASTSIDE HEALTH & REHAB

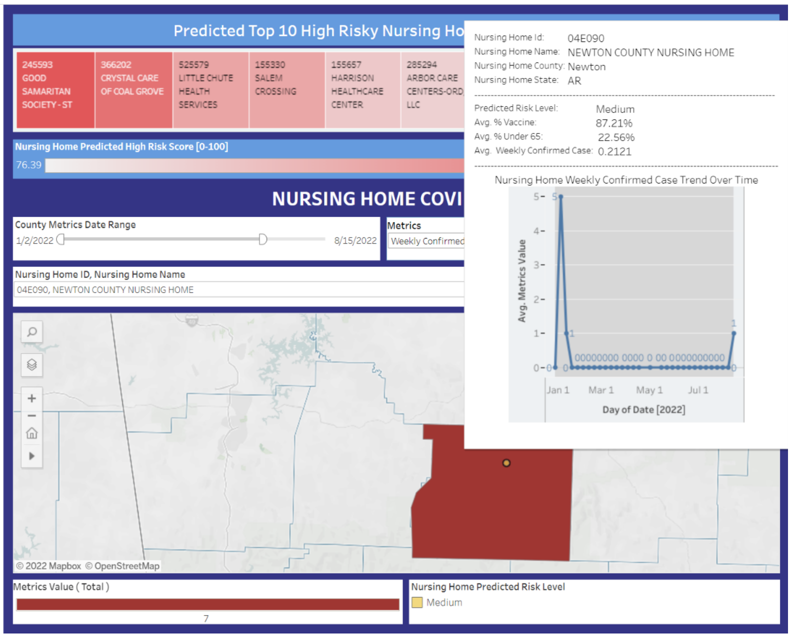
Nursing Home Predicted High Risk Score [0-100]: 76.39 – 80.21

NURSING HOME COVID MAP

County Metrics Date Range: 8/24/2022 – 10/30/2022
Metrics: Weekly Confirmed Case
Nursing Home ID, Nursing Home Name: (All)
County: (All)

Mexico
© 2022 Mapbox © OpenStreetMap
Metrics Value (Total)
0
546
Nursing Home Predicted Risk Level
🟩 Low
🟨 Medium
🟥 High

The COVID-19 map has two layers: bottom layer is a county map with sequential color of metrics selected; Top layer is a nursing home map with risk level color coded. When hovering over facility circles, tooltips will show up with detailed risk factor information and a trend plot of metrics selected.


Predicted Top 10 High Risky Nursing Ho

245593 GOOD SAMARITAN SOCIETY - ST
366202 CRYSTAL CARE OF COAL GROVE
525579 LITTLE CHUTE HEALTH SERVICES
155330 SALEM CROSSING
155657 HARRISON HEALTHCARE CENTER
285294 ARBOR CARE CENTERS-ORD LLC

Nursing Home Predicted High Risk Score [0-100]: 76.39

NURSING HOME COVI

County Metrics Date Range: 1/2/2022 – 8/15/2022
Metrics: Weekly Confirmed
Nursing Home ID, Nursing Home Name: 04E090, NEWTON COUNTY NURSING HOME

Nursing Home Id: 04E090
Nursing Home Name: NEWTON COUNTY NURSING HOME
Nursing Home County: Newton
Nursing Home State: AR

Predicted Risk Level: Medium
Avg. % Vaccine: 87.21%
Avg. % Under 65: 22.56%
Avg. Weekly Confirmed Case: 0.2121

Nursing Home Weekly Confirmed Case Trend Over Time
Avg. Metrics Value
0
1
2
3
4
5
Jan 1
Mar 1
May 1
Jul 1
Day of Date [2022]
© 2022 Mapbox © OpenStreetMap
Metrics Value (Total)
7
Nursing Home Predicted Risk Level
🟨 Medium

Experiments/Evaluation

List of questions to answer:

1. What should be the target prediction variable?
2. How to select features for the model? What features are the most important ones?
3. How to evaluate the model?
4. How to visualize the nursing home COVID-19 historical data and predicted data?

1. Model implementation and evaluation

- Framing the modeling problem

The goal of the model was to predict COVID-19 infection risk for each individual nursing home. After several modeling experiments, we concluded that classifying infection risk into different risk levels and training a classification model is better than training a regression model. We also decided to extend the forecasting period to 2 weeks to compensate for lagging in data publishing. Moreover, because the size of nursing homes differs a lot, the same number of confirmed cases can mean different risk levels for different nursing homes, e.g. 5 confirmed cases could be severe for a nursing home with 20 residents but negligible for a nursing home with 300 residents. Therefore, we decided to scale confirmed cases using the size of nursing homes (number of beds). Eventually the modeling target is framed as:

$$Risk\ Level\ Num_{t0,nh_i} = (weekly\ infect\ num_{t1,nh_i} + weekly\ infect\ num_{t2,nh_i})\ /\ num\ of\ beds_{nh_i}$$

$$Risk\ Level = \{\ 0,\ if\ Risk\ Level\ Num\ =<\ 0.01; \\ 1,\ if\ Risk\ Level\ Num\ =<\ 0.05\ and\ Risk\ Level\ Num\ >\ 0.01 \\ 2,\ if\ Risk\ Level\ Num\ >\ 0.05$$

- Feature selection

We started feature selection with the variable list gathered from literature review, then narrowed it down based on data availability. LightGBM model is capable of selecting important features, in fact it is widely used in industry for selecting important features to be fed into other machine learning models. Therefore, there is no need to further narrow down the feature list manually by the model developer.

- Feature engineering

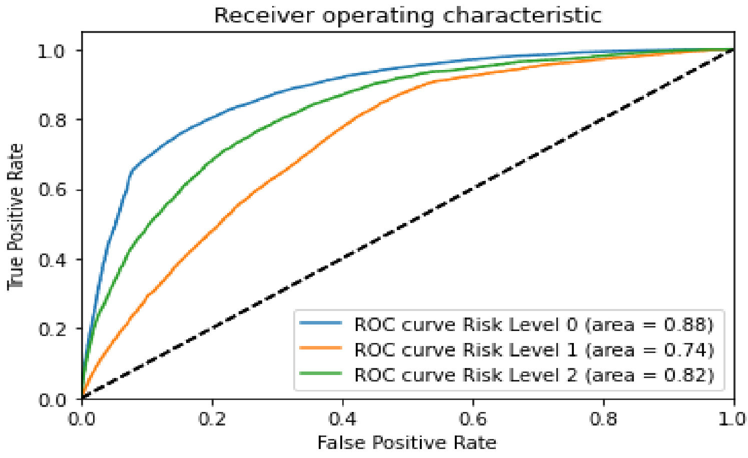
Besides the data pre-processing mentioned above, we further processed the data for modeling process. These include: 1) One-hot encoding categorical variables; 2) Logarithm transformation of right skewed numeric features[1]; 3) Lagging selected features to allow the model to see more historical data.

- Train the model

The development data is splitted into train set and test set. We used data after 2022-8-30 as the test data along with 5% nursing homes from random selection. As discussed above, we allow the model to see 4 weeks data of confirmed cases and 1 week data of other features to forecast the risk level of the following 2 weeks. We trained the model using gbdt booster and log loss for multi-class classification, with 200 boost rounds.

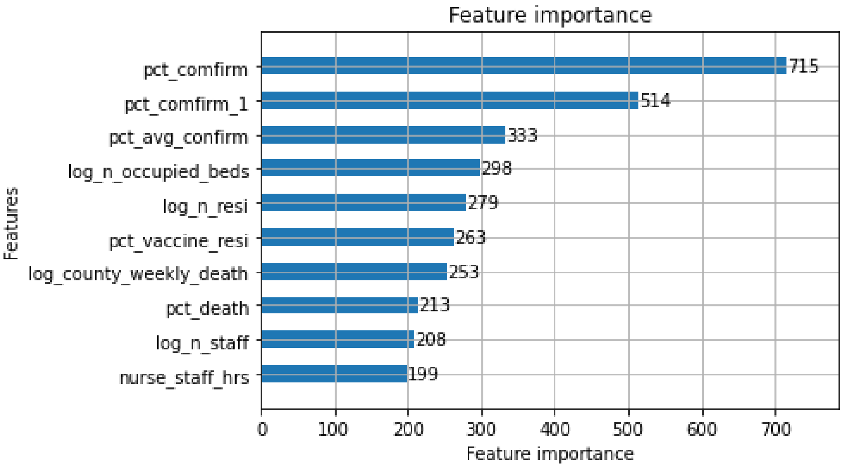
- Model evaluation

To better measure model performance we plotted its Receiver Operating Characteristic as presented below. The plot indicated that the model is good at identifying Risk level 0 (the least risky) nursing homes as well as Risk level 2 (the most risky) while performing less accurately for Risk level 1. The model has a weighted average area under ROC of 82.6%.



[1] LightGBM is a tree based model and is sensitive to the scale of features, which means there is no need to normalize the features. We decided to keep the log transformation for easy comparison with other candidate machine learning algorithms.

The model also ranked the features based on feature importance. Following are the top 10 important features. As we can see from the plot, the infection history of the nursing home is very important for forecasting future infection risk. The size of the nursing home, vaccination status and county level infection status are also very important.



2. Visualization evaluation:

Team members tested and provided user feedback to achieve the final visual dashboard structure.

To ensure visualization dashboard works properly, all filters in dashboard clicked and tested functional. visual demo.mp4 saved in the doc folder could be used as evidence.

Furthermore, to ensure dashboard convey correct information, below work has been done:

1) Top 10 risky nursing home and tooltip information validated;
2) Covid map county total metric value sampled to validate with different filter combinations;
3) Covid map nursing home predicted risk level, metric trend plot, additional risk factor information sampled to validate with different filter combinations.

Conclusions and discussion

Using the COVID-19 nursing home data, we built a machine learning model to predict the risk level in the next two weeks for each individual nursing home and the model performance is pretty good. Tableau was used to visualize the map of county level and nursing home level COVID-19 historical data and predicted data with multiple ways to filter and view the data. A tree map with sequential color was used to visualize the top 10 risky nursing homes from model predicted results. Nursing homeowners and health policymakers could quickly analyze the facility's responses from detailed risk factor information. Currently the Tableau dashboard is presenting infection risk prediction produced using static data retrieved from CMS website as of Oct 30 2020. We could further improve the tool to automatically download the newest CMS nursing home dataset and update the forecast regularly.

Nursing home infection risks were grouped into three levels: low/0, medium/1 and high/2. The model is good at identifying Risk level 0 (the least risky) nursing homes as well as Risk level 2 (the most risky) while performing less accurately for Risk level 1. We could further explore to enhance the performance of the model.

All team members have contributed a similar amount of effort.

References:

1. Shen, Karen, Lacey Loomer, Hannah Abrams, David C. Grabowski, and Ashvin Gandhi. "Estimates of COVID-19 cases and deaths among nursing home residents not reported in federal data." *JAMA network open* 4, no. 9 (2021): e2122885-e2122885.

2. Akhtar, Nikhat, Nazia Tabassum, Asif Perwej, and Yusuf Perwej. "Data analytics and visualization using Tableau utilitarian for COVID-19 (Coronavirus)." *Global Journal of Engineering and Technology Advances* (2020).

3. Giri, Shamik, Lee Minn Chenn, and Roman Romero-Ortuno. "Nursing homes during the COVID-19 pandemic: a scoping review of challenges and responses." *European Geriatric Medicine* 12, no. 6 (2021): 1127-1136.

4. Das Gupta, Debasree, Uma Kelekar, Sidney C. Turner, Anupam A. Sule, and Taya G. Jerman. "Interpreting COVID-19 deaths among nursing home residents in the US: The changing role of facility quality over time." *PloS one* 16, no. 9 (2021): e0256767.

5. Cronin, Christopher J., and William N. Evans. "Nursing home quality, COVID-19 deaths, and excess mortality." *Journal of Health Economics* 82 (2022): 102592.

6. Konetzka, R. Tamara, Elizabeth M. White, Alexander Pralea, David C. Grabowski, and Vincent Mor. "A systematic review of long‐term care facility characteristics associated with COVID‐19 outcomes." *Journal of the American Geriatrics Society* 69, no. 10 (2021): 2766-2777.

7. Telleria, Idoia Beobide, Alexander Ferro Uriguen, Esther Laso Lucas, Cinzia Sannino Menicucci, Maria Enriquez Barroso, and Adolfo López de Munain Arregui. "Epidemiology and risk factors associated with COVID-19 infection and mortality in nursing homes." *Atención Primaria* (2022): 102463.

8. Sugg, Margaret M., Trent J. Spaulding, Sandi J. Lane, Jennifer D. Runkle, Stella R. Harden, Adam Hege, and Lakshmi S. Iyer. "Mapping community-level determinants of COVID-19 transmission in nursing homes: A multi-scale approach." *Science of the Total Environment* 752 (2021): 141946.

9. Chen, M. Keith, Judith A. Chevalier, and Elisa F. Long. "Nursing home staff networks and COVID-19." *Proceedings of the National Academy of Sciences* 118, no. 1 (2021): e2015455118.

10. Sun, Christopher LF, Eugenio Zuccarelli, A. Zerhouni El Ghali, Jason Lee, James Muller, Karen M. Scott, Alida M. Lujan, and Retsef Levi. "Predicting coronavirus disease 2019 infection risk and related risk drivers in nursing homes: a machine learning approach." *Journal of the American Medical Directors Association* 21, no. 11 (2020): 1533-1538.

11. Wang, Xin, Yijia Dong, William David Thompson, Harish Nair, and You Li. "Short-term local predictions of COVID-19 in the United Kingdom using dynamic supervised machine learning algorithms." *Communications medicine* 2, no. 1 (2022): 1-8.

12. Li, Megan Mun, Anh Pham, and Tsung-Ting Kuo. "Predicting COVID-19 county-level case number trend by combining demographic characteristics and social distancing policies." *JAMIA open* 5, no. 3 (2022): ooac056.

13. Craig, Ben R., Tom Phelan, Jan-Peter Siedlarek, and Jared Steinberg. "Two approaches to predicting the path of the COVID-19 pandemic: is one better?." *Economic Commentary* 2021-10 (2021).

14. Ribeiro, Matheus Henrique Dal Molin, Ramon Gomes da Silva, Viviana Cocco Mariani, and Leandro dos Santos Coelho. "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil." *Chaos, Solitons & Fractals* 135 (2020): 109853.

15. Dai, Haoran, Wen Cao, Xiaochong Tong, Yunxing Yao, Feilin Peng, Jingwen Zhu, and Yuzhen Tian. "Global prediction model for COVID-19 pandemic with the characteristics of the multiple peaks and local fluctuations." *BMC Medical Research Methodology* 22, no. 1 (2022): 1-14.

16. Mohamadou, Youssoufa, Aminou Halidou, and Pascalin Tiam Kapen. "A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19." *Applied Intelligence* 50, no. 11 (2020): 3913-3925.

17. Ma, Ruifang, Xinqi Zheng, Peipei Wang, Haiyan Liu, and Chunxiao Zhang. "The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method." *Scientific Reports* 11, no. 1 (2021): 1-14.

18. Shakeel, Sheikh Muzaffar, Nithya Sathya Kumar, Pranita Pandurang Madalli, Rashmi Srinivasaiah, and Devappa Renuka Swamy. "COVID-19 prediction models: a systematic literature review." *Osong Public Health and Research Perspectives* 12, no. 4 (2021): 215.
19. Hochreiter, S., Ja1 4 rgen schmidhuber (1997).“long short-term memory”. *Neural Computation*, *9*(8).

Appendix:

Risk factors from the literatures	Data availability	Data source (describe the data)
facility characteristics	https://data.cms.gov/covid-19/covid-19-nursing-home-data/data	Provider name, address, city, state, zip code, phone number
	https://data.cms.gov/provider-data/dataset/4pq5-n9py	Ownership type, Overall/Health inspection rating, Long-stay/Short-stay QM rating, Number of Facility Reported Incidents/Substantiated Complaints/Citations from Infection Control Inspections/Fines, Total Amount of Fines in Dollars, Total Number of Penalties
staff-related factors	https://data.cms.gov/covid-19/covid-19-nursing-home-data/data	Staff weekly/total confirmed covid 19 Staff weekly/total covid 19 deaths
	https://data.cms.gov/provider-data/dataset/4pq5-n9py	Staffing rating, Reported Total Nurse Staffing Hours per Resident per Day, Total number of nurse staff hours per resident per day on the weekend, Total nursing staff turnover, Adjusted Total Nurse Staffing Hours per Resident per Day, Adjusted Weekend Total Nurse Staffing Hours per Resident per Day
shortage	https://data.cms.gov/covid-19/covid-19-nursing-home-data/data	Shortage of nursing/clinical/aides/other staff
PPE	https://data.cms.gov/covid-19/covid-19-nursing-home-data/data	n95 respirator /face masks /eye protection/gowns/ gloves no longer available in 7 days
facility size	https://data.cms.gov/covid-19/covid-19-nursing-home-data/data	Number of all beds Total number of occupied beds
vaccination	https://data.cms.gov/covid-19/covid-19-nursing-home-data/data	Number of Residents Not Vaccinated/Received COVID-19 Vaccine Dose 1/Doses 1,2/Doses 1,2, booster Before/After Positive Test
Residents related factors	https://ltcfocus.org/data (ltc-nursing home characteristics_2020.csv)	Residents acuindex2, Medicaid, Medicare, Black/Hispanic/White /< 65 Years Old/Female/CFS = 4/Alzheimers/Dementia (all admits), Average Age, Any Prior NH stay
treatment	https://data.cms.gov/covid-19/covid-19-nursing-home-data/data	Therapeutic with bamlanivimab/ casirivimab/imdevimab/etesevimab/sotrovimab
County level data	https://github.com/nytimes/covid-19-data	Cases, Deaths

facility location	https://data.cms.gov/covid-19/covid-19-nursing-home-data/data	Provider address
high-density communities	rural_urban_codes_2013.csv	Population, Metro/Nonmetro
Asymptomatic transmission	No	