1-6-2021

Compress the Curve: A Cross-Sectional Study of Variations in COVID-19 Infections Across California Nursing Homes

Ram Gopal

Xu Han

Niam Yaraghi

BMJ Open

Compress the curve: a cross-sectional study of variations in COVID-19 infections across California nursing homes

Ram Gopal ,[1] Xu Han,[2] Niam Yaraghi [3,4]

To cite: Gopal R, Han X, Yaraghi N. Compress the curve: a cross-sectional study of variations in COVID-19 infections across California nursing homes. *BMJ Open* 2021;**11**:e042804. doi:10.1136/bmjopen-2020-042804

► Prepublication history and additional material for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-042804).

RG, XH and NY contributed equally.

Received 17 July 2020
Revised 05 December 2020
Accepted 07 December 2020

Check for updates

[1]Warwick Business School, University of Warwick, Coventry, UK
[2]Gabelli School of Business, Fordham University, New York, New York, USA
[3]Miami Herbert Business School, University of Miami, Coral Gables, Florida, USA
[4]Governance Studies, The Brookings Institution, Washington, DC, USA

Correspondence to
Dr Niam Yaraghi;
niamyaraghi@miami.edu

ABSTRACT

Objective Nursing homes' residents and staff constitute the largest proportion of the fatalities associated with COVID-19 epidemic. Although there is a significant variation in COVID-19 outbreaks among the US nursing homes, we still do not know why such outbreaks are larger and more likely in some nursing homes than others. This research aims to understand why some nursing homes are more susceptible to larger COVID-19 outbreaks.

Design Observational study of all nursing homes in the state of California until 1 May 2020.

Setting The state of California.

Participants 713 long-term care facilities in the state of California that participate in public reporting of COVID-19 infections as of 1 May 2020 and their infections data could be matched with data on ratings and governance features of nursing homes provided by Centers for Medicare & Medicaid Services (CMS).

Main outcome measure The number of reported COVID-19 infections among staff and residents.

Results Study sample included 713 nursing homes. The size of outbreaks among residents in for-profit nursing homes is 12.7 times larger than their non-profit counterparts (log count=2.54; 95% CI, 1.97 to 3.11; p<0.001). Higher ratings in CMS-reported health inspections are associated with lower number of infections among both staff (log count=−0.19; 95% CI, −0.37 to −0.01; p=0.05) and residents (log count=−0.20; 95% CI, −0.27 to −0.14; p<0.001). Nursing homes with higher discrepancy between their CMS-reported and self-reported ratings have higher number of infections among their staff (log count=0.41; 95% CI, 0.31 to 0.51; p<0.001) and residents (log count=0.13; 95% CI, 0.08 to 0.18; p<0.001).

Conclusions The size of COVID-19 outbreaks in nursing homes is associated with their ratings and governance features. To prepare for the possible next waves of COVID-19 epidemic, policy makers should use these insights to identify the nursing homes who are more likely to experience large outbreaks.

Strengths and limitations of this study

- ► A bivariate Poisson model is employed to better capture the interdependencies of COVID-19 cases between staff and residents.
- ► Predictive models are developed to identify nursing homes with the highest chance of experiencing COVID-19 outbreaks.
- ► Data analysed are only from California.
- ► The dataset on nursing homes' features is based on the year 2017.
- ► The number of COVID-19 cases reported by nursing homes may be subject to under-reporting.

INTRODUCTION

Nursing homes have been most severely impacted by the COVID-19 pandemic owing to the advanced age and high number of comorbidities of their residents.[1 2] In Europe, as much as 57% of all deaths related to COVID-19 were at such facilities.[3] In the USA, nursing homes' residents and staff account for 34% of all COVID-19 fatalities.[4] Infection prevention and control at nursing homes and long-term facilities has therefore become a priority in managing the epidemic.[5 6]

Given the considerable variation in the prevalence and size of the COVID-19 outbreaks at nursing homes, the objective of this research is (1) to understand why some nursing homes are more susceptible to COVID-19 outbreaks, and (2) to develop predictive models that can identify such nursing homes so that they could be prioritised in efforts to prevent and contain next waves of the epidemic.[7 8]

METHODS

Patient and public involvement

Patients had no influence on the research questions or outcomes of this research. No patients were involved in the design of this study. We used blind patient files; therefore, no patient recruitment took place. We only used data on the aggregated number of patients with COVID-19 and staff in the nursing homes as reported by the state of California and therefore no personal information of patients was used in this study. Given the nature of removing all personal information, there is no requirement to disseminate the information to patients.

Data sources and study variables

We collected data from various publicly available sources. The New York Times aggregates and provides data on COVID-19 cases per county.[9] California Department of Public Health (CDPH) provides data on the number of confirmed COVID-19 infections among staff and residents of nursing homes in the state.[10] Centers for Medicare & Medicaid services (CMS) provides data on nursing home characteristics, including their self-reported ratings and CMS health inspections.[11] A description of this data is provided in the next section. Applying the methods suggested by Han *et al*,[12] we identified the nursing homes with significant discrepancies between their self-reported measures and independent CMS inspections for a consecutive 5-year period. We aggregated the results and used the number of years a nursing home is predicted to be a likely inflator as the overall inflation score for a nursing home. Therefore, an honest nursing home will have an inflation score of 0 while an inflating nursing home can have an inflation score between 1 and 5, with 5 being the most severe. In our dataset, 19.25% of nursing homes were inflating their scores and some of these had a score of 5 indicating that they inflated their scores in all 5 years.

These methods rely on data that are only available for nursing homes in California and therefore, the scope of this study is also limited to nursing homes in California. After cleaning and merging the abovementioned data sources, we analysed a final dataset consisting of 713 nursing homes in California. Details of the data cleaning and merging process are presented in online supplemental appendix 1.

We examined the following outcomes in this study: whether a nursing home has at least one COVID-19 infection among its residents or staff, the number of confirmed COVID-19 infections among its residents and the number of confirmed infections among its staff. We also calculated a fourth outcome that indicates the large outbreaks as the ones in which more than 10 members of staff or residents were infected with COVID-19. This threshold translates to approximately 95th percentile of the number of infected staff. Given that more residents are infected than staff, this threshold translates to 75th percentile of the number of residents.

The independent variables describe the severity of the COVID-19 outbreak in the surrounding area of a nursing home, its governance characteristics, as well as its ratings on quality, staffing and CMS inspections. Table 1 provides detailed description of the study variables. Note that while almost all nursing homes have resident councils, only 20% of nursing homes have existing family councils. We included the existence of family council as a binary variable in our analysis with the contention that it may imply closer coordination and higher engagement with the families of the residents.

Description of CMS' nursing home compare system

The CMS nursing home rating data consist of basic information about nursing facilities such as name, address, phone number and so on, as well as some key features used in our analysis, such as the number of certified beds, whether the nursing home is for-profit or non-profit, whether the nursing home has a family council and so on.

The CMS nursing home rating data serve the CMS nursing home compare system, in which nursing home ratings are generated based on three domains: inspection, staffing and quality measures. The inspection is conducted and reported by CMS-certified inspectors annually. The other two domains are self-reported by nursing homes. The annual inspection investigates areas such as medication management, nursing home administration, environment, food service and residents' rights and quality of life. The staffing domain is evaluated based on the self-reported CMS Certification and Survey Provider Enhanced Reports staffing data. The two measures used are the total nursing hours and registered nursing hours and are adjusted for case-mix based on the resource utility group case-mix system derived from the minimum data set. The staffing star rating is then updated by the end of the quarter when raw data are collected. Note that with more recent changes, the staffing data reported by nursing homes are subject to validation with nursing homes' payroll data reported through payroll-based journal. The quality measure rating uses quality measurement criteria, which covers both long-stay terms and short-stay terms. The quality measure star rating is updated by the end of each quarter by using the results from three most recent quarters.

To calculate the star ratings, CMS first assigns an initial star rating to all nursing homes based on their annual inspection results. Nursing homes are then assigned star ratings for the staffing and quality measures domains. The overall star rating is then calculated by considering the inspection rating as the baseline, increasing or decreasing by one star if any self-reported domain satisfies the conditions stated as follows. Both 4 and 5 stars in staffing rating are qualified for obtaining additional overall star rating, while only 5 stars in quality measure is qualified. Additional conditions apply to nursing homes whose inspection ratings are only 1 star, and for nursing homes which are in the CMS's special focus facility programme. The overall star rating is lowered by one star if any self-reported domain is 1 star. The overall star rating cannot be more than 5 stars or less than 1 star. Detailed data from CMS on nursing homes are available online.[13]

Statistical analysis

To answer the first research question and understand why some nursing homes are more susceptible to COVID-19 outbreaks, we applied Zero Inflated Bivariate Poisson (ZIBP) regression. The model allows us to examine the effects of nursing homes' ratings, governance features and their surroundings on the likelihood and size of their COVID-19 outbreaks. Econometric details of the model

Table 1 Sources and descriptions of the study variables

Variable	Description	Source	Mean	SD	Min	Max
Outcomes						
Nursing home infected	Indicates if the nursing home has at least one confirmed case of COVID-19 infection among its staff or residents	CDPH	0.23	0.42	0	1
Confirmed residents	The number of COVID-19 infections among the residents of nursing homes	CDPH	1.91	7.88	0	81
Confirmed staff	The number of COVID-19 infections among the staff of nursing homes	CDPH	0.41	2.19	0	26
Large outbreak	Among those nursing homes with at least one infection, indicates if the number of infected staff or residents is more than 10 infections	Authors’ calculation	0.31	0.46	0	1
Severity of COVID-19 epidemic in the surrounding area						
County infections per 100K	The rate of COVID-19 infections per 100 000 residents in the county in which the nursing home is located as of 1 May 2020	New York Times	143.42	80.07	0	259.8
Governance features						
For profit	Indicates if the nursing home has a for-profit status	CMS	0.86	0.35	0	1
Family council	Indicates if a family council for the residents exists in the nursing home	CMS	0.2	0.4	0	1
Certified beds	The number of beds certified to provide care to Medicare and Medicaid beneficiaries	CMS	98.89	54.77	14	769
Occupancy rate	The ratio of residents to the total number of certified beds	Authors’ calculation	0.87	0.12	0.14	1
Inflation score	Counts the number of years in which a significant discrepancy was observed between the self-reported quality measures and CMS-reported health inspections	Authors’ calculation	0.32	0.81	0	5
Ratings						
Quality rating	Self-reported indicator of quality of services as of 2017	CMS	4.59	0.87	0	5
Staffing rating	Self-reported measure of staffing hours as of 2017. This is based on a combination of registered nurse hours per resident day and the total nursing hours per resident day	CMS	3.41	1.13	0	5
Health inspection rating	CMS-reported indicator of health inspections ratings as of 2017	CMS	2.88	1.29	1	5

CDPH, California Department of Public Health; CMS, Centers for Medicare & Medicaid Services.

are provided by Walhin.[14] Conventional Poisson models are suitable for modelling count data, while the zero inflated variation of Poisson model is more suitable for modelling count data with excess zeros, especially when excess zeros are generated by a separate processes that could be modelled separately. This leads to a framework that consists of a logit model for estimating the excess zeros in addition to a Poisson count model. ZIBP model is an extension of zero inflated Poisson model and is best suited for situations in which the count data with excess zeros are generated for two outcomes that may be correlated. In cases where the outcome variables are independent, the model reduces to the product of two independent zero inflated Poisson regression models, referred to as Zero Inflated Double Poisson model. in our setting, the two count variables are the number of COVID-19 infections among staff, and residents. These counts include excess zeros since many nursing homes reported no COVID-19 cases, primarily because they are located in areas where at the time of the data collection, had not yet experienced significant surges in COVID-19 cases. These two counts are also correlated since they both happen at the same nursing home and the factors that give rise to them are common at the nursing home level.

Intuitively, we assume that the number of zero's in the count of infected staff and residents is generated either because the nursing home was in an area that was less infected by the COVID-19 or because it implemented successful prevention procedures to protect its staff and residents. Moreover, we assume that in a nursing home, the number of infected staff covaries with the number of infected residents since they can infect each other and since common infection prevention and control policies apply to both groups. Taking this interdependency into account also alleviates the concerns over the possible impact of omitted variables in our model. In this context, because of the close proximity of residents and staff, the same variables that could affect the number of infections among one group would most likely also impact the number of infections among the other group. The covariance coefficient captures this interdependency in outcomes. As a sensitivity analysis, we also report the results of Zero Inflated Double Poisson regression. In this model, the counts of infections among staff and residents are assumed to be independent from each other. We use NLMIXED procedure in SAS software to estimate our models.[15 16] Note that we have provided access to both the data and the SAS code for this analysis.[17 18]

To answer the second research question and identify the nursing homes with the highest risk of COVID-19 outbreaks, we used our models to predict the probability of experiencing an infection and compared their performance with common machine learning techniques, namely neural networks (NN) and support vector machine with radial basis function (SVM-RBF) kernel. Since our problem has a highly nonlinear structure, advanced machine learning models such as NN and SVM that do not rely on data structure assumptions may provide a flexible and desired solution. The target variable in each model is equal to 1 if at least one patient or staff reported to be infected. The prediction features include nursing home governance features such as occupancy rate, number of certified beds, whether a family council presents, whether the nursing home is for profit or not and inflation score evaluated from past years. The nursing homes' health inspection rating, staffing rating and quality rating are also included. The machine learning models are implemented in Python V.3.7 with 70% data training and 30% data testing. The entire dataset is used to plot the lift chart. We also measured the performance of our models in predicting the nursing homes with highest risks of experiencing large outbreaks with more than 10 infections.

RESULTS

Study sample

During the data cleaning and merging process, 493 nursing homes were eliminated from our final sample, either because their names were not matching across different datasets, or their ratings information is not available from CMS, or because their COVID-19 infections are not reported by CDPH. To ensure that the final sample is random and our results are not biased, we compared the eliminated nursing homes with the ones in the study sample. The results of two sample t-tests and logistic regression are presented in online supplemental appendix 1. None of the observed governance factors affect the chance of being included in the sample. Among the remaining variables, while the difference with regards to quality ratings and county infections per 100 000 is statistically significant between the two groups, their magnitude is small and serve to make our estimates more conservative.

Study sample included 713 nursing homes in California. As reported in table 1, as of 1 May 2020, 23% of the study sample reported at least one COVID-19 infection among either their staff or residents. Of those, 31% experienced large outbreaks with more than 10 infections among either their staff or residents. The geographic spread of COVID-19 infections in California nursing homes is graphically presented in the online supplemental appendix 1.

Preventing COVID-19 infections

According to the model selection criteria reported in table 2, the ZIBP model provides a better fit as its Akaike information criterion (AIC), Bayesian information criterion (BIC), and −2log likelihood are all smaller than those of Zero Inflated Double Poisson model. We therefore report the estimates of the ZIBP model in the text. The coefficients in the first panel of table 2 represent how the log odds of experiencing an infection changes with one unit of increase in the corresponding predictor. As reported in the first panel of table 2, the only variables

Table 2 Effects of study variables on the likelihood and the size of COVID-19 outbreaks

	Zero Inflated Bivariate Poisson Model			Zero Inflated Double Poisson Model		
Parameter	Estimate	95% CI	P Value	Estimate	95% CI	P Value
Nursing home (likelihood of nursing home getting at least one COVID-19 infection)						
Intercept	−2.34	−4.41 to −0.28	0.03	−1.76	−3.75 to 0.24	0.08
County infections per 100K	0.01	0.01 to 0.02	<0.001	0.01	0.01 to 0.02	<0.001
For profit	−0.36	−0.94 to 0.22	0.22	−0.27	−0.85 to 0.31	0.36
Family council	0.19	−0.28 to 0.64	0.44	0.21	−0.26 to 0.67	0.38
Certified beds	0.01	0.01 to 0.02	0.01	0.01	0.01 to 0.02	0.01
Occupancy rate	−0.2	−1.99 to 1.59	0.83	−0.98	−2.69 to 0.74	0.26
Inspection rating	−0.02	−0.19 to 0.17	0.9	−0.02	−0.19 to 0.17	0.90
Quality rating	−0.14	−0.36 to 0.1	0.26	−0.13	−0.35 to 0.1	0.27
Staffing rating	0.01	−0.17 to 0.18	0.97	−0.01	−0.18 to 0.17	0.96
Inflation score	0.06	−0.18 to 0.28	0.67	0.06	−0.17 to 0.29	0.61
Infected staff (number of staff with confirmed COVID-19 infections)						
Intercept	0.21	−2.11 to 2.52	0.87	−0.43	−2.1 to 1.25	0.63
County infections per 100K	−0.01	−0.01 to 0.01	0.23	−0.01	−0.01 to 0.01	0.11
For profit	−0.21	−0.78 to 0.37	0.49	−0.16	−0.55 to 0.24	0.44
Family council	−0.04	−0.54 to 0.46	0.89	0.19	−0.12 to 0.49	0.24
Certified beds	0.01	0.01 to 0.01	<0.001	0.01	0.01 to 0.01	0.02
Occupancy rate	−2.39	−4.3 to −0.47	0.02	−1.11	−2.53 to 0.32	0.13
Inspection rating	−0.19	−0.37 to −0.01	0.05	−0.16	−0.28 to −0.03	0.02
Quality rating	0.4	0.13 to 0.67	0.01	0.33	0.15 to 0.52	<0.001
Staffing rating	0.11	−0.07 to 0.28	0.23	0.25	0.12 to 0.37	<0.001
Inflation score	0.41	0.31 to 0.51	<0.001	0.27	0.19 to 0.35	<0.001
Infected residents (number of residents with confirmed COVID-19 infections)						
Intercept	1.36	0.36 to 2.35	0.01	1.69	0.84 to 2.55	<0.001
County infections per 100K	−0.01	−0.01 to −0.01	<0.001	−0.01	−0.01 to −0.01	<0.001
For profit	2.54	1.97 to 3.11	<0.001	1.88	1.51 to 2.26	<0.001
Family council	0.07	−0.09 to 0.21	0.4	0.1	−0.04 to 0.24	0.15
Certified beds	0.01	0.01 to 0.01	0.04	0.01	−0.01 to 0.01	0.13
Occupancy rate	−0.24	−1.01 to 0.54	0.55	−0.15	−0.88 to 0.6	0.71
Inspection rating	−0.2	−0.27 to −0.14	<0.001	−0.2	−0.26 to −0.14	<0.001
Quality rating	0.13	0.05 to 0.21	0.01	0.15	0.08 to 0.23	<0.001
Staffing rating	−0.26	−0.31 to −0.2	<0.001	−0.2	−0.25 to −0.15	<0.001
Inflation score	0.13	0.08 to 0.18	<0.001	0.11	0.06 to 0.16	<0.001
Covariance	0.69	0.54 to 0.87	0.01			
Fit statistics						
−2 log likelihood	4422.7			4561.7		
AIC	4484.7			4621.7		
BIC	4626.4			4758.8		

Note: The coefficients in the first panel represent how the log odds of experiencing an infection changes with one unit of increase in the corresponding predictor. The coefficients in the second and third panels represent how the expected log count of the infections changes for each unit increase in the corresponding predictor.
AIC, akaike information criterion; BIC, bayesian information criterion.

with statistically significant impact on the chance of COVID-19 outbreaks at nursing homes are their size and the rate of infections per 100 000 residents at the county in which they are located. For both variables, a one-unit of increase is associated with a 1% increase in the odds of experiencing at least one COVID-19 infection.

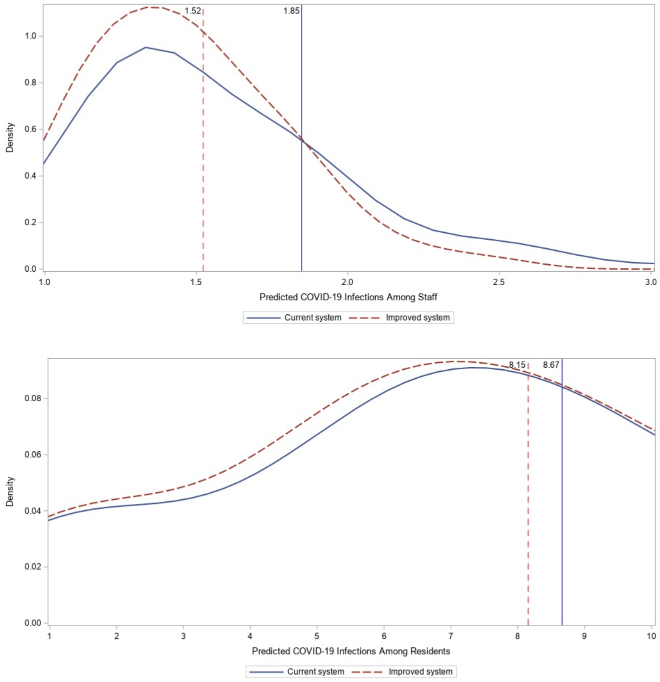


Figure 1 Impact of improved rating system on infection density curves. Note: the blue (solid) curve represents the density of predicted number of infections under current rating system while the red (dashed) curve shows the density of counterfactual number of infections had there been no discrepancy between self-reported and CMS-reported ratings. The vertical blue and red lines show the average number of predicted infections with and without discrepancy in ratings.

Controlling COVID-19 outbreaks

The coefficients in the second and third panels of table 2 represent how the expected log count of the infections changes for each unit increase in the corresponding predictor.

As reported in the second and third panels of table 2, the expected rate of infections among both staff and residents increases with the size of the nursing home. This indicates that the severity of COVID-19 epidemic in the surrounding area increases the chance of experiencing at least one infection at the nursing homes.

While the size of outbreaks among residents is about 12.7 times higher in for-profit nursing homes, the size of outbreak among staff in for-profit nursing homes is not statistically different from non-profit ones. This is in line with prior empirical research that has repeatedly shown that for-profit nursing homes are inferior in many aspects of care quality.[19–22]

Occupancy rate, which represents the ratio of the number of patients to the number of certified beds of a nursing home, is associated with a lower rate of infections among staff such that a 1% increase in occupancy rate decreases the expected count of infections among staff by 2.4%.

Among the three different ratings, the CMS-reported health inspection rating is associated with a sizeable decrease in the number of infections among both staff and residents. One unit of increase in CMS-reported health inspection ratings is associated with a 17% and 18% decrease in the expected number of infections in staff and residents, respectively. A one-unit improvement in staffing rating is associated with a 23% decrease in the number of infections among residents. Note that better staff rating is highly dependent on higher ratio of staff to residents and the higher number of staff per resident would allow nursing homes to control infections more efficiency among their residents. While the observed associations between ratings on health inspections and staffing with the number of infected staff and residents were expected, the association between self-reported quality ratings and the number of infections is the opposite of our expectations. One unit of increase in self-reported quality ratings is associated with, respectively, 49% and 14% increase in infections among staff and residents. This finding is aligned with the emerging stream of research that shows nursing homes embellish their self-reported quality ratings and therefore these ratings may not always indicate better quality of care for residents.[12 23–26] Our final variable, inflation score, quantifies the discrepancy between the self- and CMS-reported ratings. The higher the discrepancy, the more likely it is that the nursing home is overstating their quality measures. With a one-unit increase in such discrepancy, the expected number of infections among staff and residents increases by 51% and 14%, respectively.

Improving the quality reporting system

CMS could solve these discrepancies and improve the reporting process by implementing better inspection and auditing strategies.[27] Figure 1 shows how the number of infections among staff and residents could be compressed had the self-reported quality measures by nursing homes were truly reflecting their quality of care.

Given the importance of ratings for nursing homes,[28] with a reliable rating system with no discrepancy between self-reported and CMS-reported measures, nursing homes would strive to elevate their ratings through actual improvements in their quality of care. As shown in the upper panel of figure 1, compared with the current system, lower number of predicted infections among staff would have been more frequent under an improved rating system such that predicted average number of infections among staff would have decreased from 1.85 to 1.52, which is equal to 17.6% fewer total infections across the staff of all nursing homes. As shown in the lower panel of figure 1, the same effect is observed for nursing home residents. Had self-reported quality ratings were truly reflecting the quality of care, the expected number of infections among residents of nursing homes would have reduced from 8.67 to 8.15 which is equal to 5.8% fewer total infections across the residents of all nursing homes.

Finally, the sizeable covariance estimate (0.68; 95% CI 0.54 to 0.87; p=0.1) indicates that the number of infected staff is not independent from the number of infected

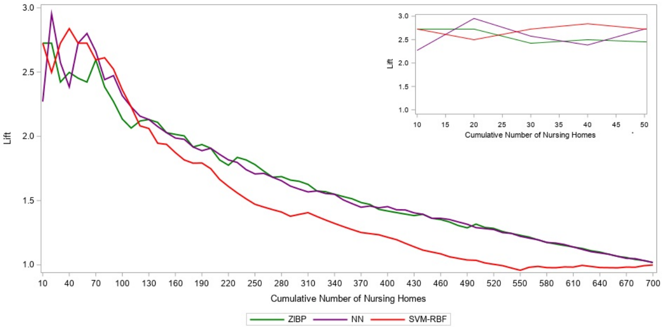


Figure 2 Comparison of performance of ZIBP, NN and SVM-RBF models in predicting at least one infection. Note: the first 50 nursing homes are zoomed in at the top right corner of the figure. The lift of ZIBP model is presented in green, while the lifts of NN and SVM-RBF are presented with purple and red lines, respectively. NN, neural network; SVM-RBF, support vector machine with radial basis function; ZIBP, Zero Inflated Bivariate Poisson.

residents. This observation empirically confirms our expectation of dependency between the count of infections in staff and residents such that nursing homes with high number of infected staff also have high number of infected residents. This finding was expected as residents and staff are in close contact with each other and once infections occur among the members of one group, it would be very difficult to prevent them in the other group. More importantly, common infection control procedures implemented by nursing homes would apply to both groups and prevent infections among both groups. Note that as discussed earlier, according to all the model selection criteria, the ZIBP performs better than its competitors. This is not surprising since it has the advantage of modelling and adjusting for the correlation between the count of infections among staff and residents. In the online supplemental appendix 1, we provide further empirical details on the correlation between the number of infections among residents and staff.

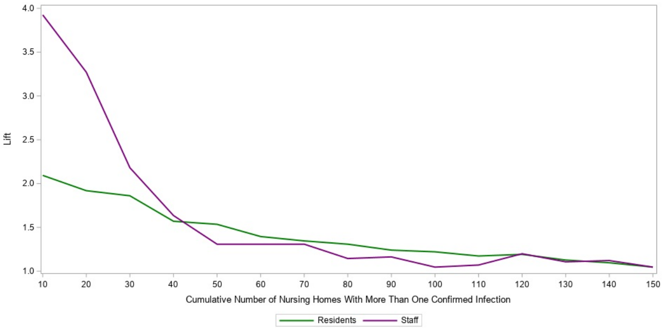


Figure 3 Performance of ZIBP model for predicting large outbreaks (more than 10 infections) among staff and residents. Note: the lifts of the ZIBP model for identifying large outbreaks among residents and staff are presented, respectively, by the green and purple lines. ZIBP, Zero Inflated Bivariate Poisson.

Identifying nursing homes with highest chance of COVID-19 infections and outbreaks

Figure 2 compares the lift of the ZIBP model with those of NN and SVM-RBF. We use lift as a measure for the ability of the model at predicting or classifying cases with respect to random selection. Lift shows how much better our model works compared with a random selection model. The first 50 nursing homes are zoomed in at the top right corner of the figure. The ZIBP model's performance is comparable with the common NN and SVM-RBF methods. For the first 50 nursing homes, the rate of true positives of ZIBP model is between 2.45 and 2.73 times higher than that of a random selection model. The area under the curve for ZIBP, NN and SVM-RBF models is respectively 0.68, 0.73 and 0.62.

Figure 3 presents the lifts of the ZIBP model in identifying the nursing homes with large COVID-19 outbreaks among those that have confirmed at least 10 infections. For the first 50 nursing homes, ZIBP correctly identifies nursing homes with large outbreaks among staff between 1.3 and 3.9 times better than a random selection model. The model's performance for predicting large outbreaks among residents for the first 50 nursing homes is 1.5–2.1 times better than a random selection model.

DISCUSSION

Staff and residents of nursing homes constitute the largest demographic of COVID-19 fatalities in the USA. However, nursing homes have not been uniformly impacted by the epidemic; some have not experienced even a single infection while some others have been devastated by COVID-19 fatalities. To prepare for the possible next waves of the epidemic, it is critical to uncover the underlying reason of such variation and to explore the nursing homes' features that are associated with higher chance and size of outbreaks.

The aim of this research was to understand how publicly available data on nursing homes can explain the significant variation in the chance and size of COVID-19 infections at nursing homes, and to also develop predictive models that can identify the nursing homes with the highest chance and size of outbreaks.

Our results indicate that COVID-19 outbreaks are more likely to happen at larger nursing homes and those with higher rate of COVID-19 infections in the surrounding area. These factors have been shown to be associated with higher probability of experiencing infections by other researchers as well.[29]

Those with better staffing and health inspection ratings are more successful in controlling the outbreaks. The association between staffing levels and likelihood of having COVID-19 infections among both staff and residents has been reported by other researchers as well.[30] Interestingly, higher self-reported quality ratings are associated with larger size of outbreaks. This counterintuitive result could further evidence that nursing homes exaggerate their self-reported quality measures. Higher discrepancy between self-reported measures and CMS-reported health inspections was associated with larger COVID-19 outbreaks.

The size of the outbreaks among residents is significantly higher in for-profit nursing homes which have been previously shown to also be of poorer quality in various aspects of care.[19–22]

There is a complex relationship between the main variables in our models. For-profit Nursing Homes generally have lower nurse staffing, more deficiencies, larger in size and have a greater likelihood of inflating their ratings.[31 32] It is therefore not surprising that they were found to be more likely to have larger numbers of COVID-19 infected residents and staff.

The model developed in this research can correctly identify the nursing homes that are more likely to experience an infection or are at the highest risk of an outbreak.

The insights of this research help policy makers to identify the nursing homes with the highest probability and size of COVID-19 outbreaks. This will allow them to prioritise such nursing homes in their efforts to control the epidemic. Such efforts could entail devoting more resources towards nursing homes with significantly higher risk or when feasible, temporarily transferring patients to different nursing homes to control the spread of the virus.

Our results show that our ZIBP model outperforms SVM and that the predictive ability of the NN is only modestly better than ZIBP model. That is, the application and comparison of these machine learning models with the results of the ZIBP model confirms that not only the ZIBP model can explain the relationship between various independent variables and COVID-19 infections at nursing homes, but it also offers competitive predictive performance.

An important takeaway from this research is the importance of data collection and transparency. Our research was made possible because of the availability of key information on COVID-19 infections in nursing homes in the USA and publicly available data such as ownership, size, staffing and key performance measures. Access to such data is invaluable in both understanding and taking preventive action to curb the COVID-19 infections in nursing homes. As such we hope that other industrialised nations take necessary steps to collect and disseminate such information to protect and safeguard the vulnerable residents in long-term care facilities.

This work leaves several areas for future research. First, given the variation in testing at different nursing homes, the number of confirmed infections may be undercounting the actual number of infections and therefore a more reliable measure would be the number of fatalities associated with COVID-19. Second, should temporal data become available, researchers can study growth curves of infections or deaths among staff and residents and examine their interlinked effects on each other. Third, should national data become available, we can test our contentions using a much larger sample at the national level. This would increase the external validity and generalisability of our findings. Finally, when data from other states and other time become available, we can include a spatial random effect in the model to account for spatial dependencies between the infections at different nursing homes.

One of the limitations of the study is that its data on nursing homes' features are collected in 2017 which is over 2 years prior to the outbreak. Although more recent data were available on the time of the study, the variable 'inflation score' had to be adopted from the 2017 data. We should also note that 86% of California nursing homes are for-profit and these nursing homes were probably more likely to under-report their infection rates and deaths than other nursing homes for fear of losing residents and revenue.[33]

Twitter Ram Gopal @ramdgopal

Contributors RG and NY designed the study. RG, XH and NY had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analyses. RG, XH and NY analysed the data. RG and NY interpreted the data. NY drafted the manuscript. NY and RG critically revised the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Map disclaimer The depiction of boundaries on this map does not imply the expression of any opinion whatsoever on the part of BMJ (or any member of its group) concerning the legal status of any country, territory, jurisdiction or area or of its authorities. This map is provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient consent for publication Not required.

Data availability statement All data and software code used in this research are made publicly available by the authors and their sources have been cited in the manuscript.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Ram Gopal http://orcid.org/0000-0003-4241-9355
Niam Yaraghi http://orcid.org/0000-0003-3497-0251

REFERENCES

1 McMichael TM, Currie DW, Clark S, *et al*. Epidemiology of Covid-19 in a long-term care facility in King County, Washington. *N Engl J Med* 2020;382:2005–11.

2 Arentz M, Yim E, Klaff L, *et al*. Characteristics and outcomes of 21 critically ill patients with COVID-19 in Washington state. *JAMA* 2020;323:1612–4.

3 Comas-Herrera A, Zalakain J, Lemmon E, *et al*. Mortality associated with COVID-19 in care homes: international evidence. *LTCcovid. org, International Long-Term Care Policy Network, CPEC-LSE* 2020

https://alzheimeriberoamerica.org/wp-content/uploads/2020/04/Mortality-associated-with-COVID-12-April-3.pdf
4 Yourish K, KKR L, Ivory D, *et al*. One-Third of all U.S. coronavirus deaths are nursing home residents or workers. the new York times, 2020. Available: https://www.nytimes.com/interactive/2020/05/09/us/coronavirus-cases-nursing-homes-us.html
5 Bedford J, Enria D, Giesecke J, *et al*. COVID-19: towards controlling of a pandemic. *Lancet* 2020;395:1015–8.
6 Adalja AA, Toner E, Inglesby TV. Priorities for the US health community responding to COVID-19. *JAMA* 2020;323:1343–4.
7 Xu S, Li Y. Beware of the second wave of COVID-19. *Lancet* 2020;395:1321–2.
8 Leung K, Wu JT, Liu D, *et al*. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: a modelling impact assessment. *Lancet* 2020;395:1382–93.
9 The New York Times. California coronavirus map and case count. the New York Times, 2020. Available: https://www.nytimes.com/interactive/2020/us/california-coronavirus-cases.html
10 California Department of Public Health. Skilled nursing facilities: COVID-19, 2020. Available: https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/SNFsCOVID_19.aspx
11 Archived Datasets |. Data.Medicare.gov. Data.Medicare.Gov. accessed may 23, 2020. Available: https://data.medicare.gov/data/archives/nursing-home-compare
12 Han X, Yaraghi N, Gopal R. Winning at All Costs: Analysis of Inflation in Nursing Homes' Rating System. *Prod Oper Manag* 2018;27:215–33.
13 Calgary O. Archived datasets | Data.Medicare.gov. Data.Medicare.Gov, 2020. Available: https://data.medicare.gov/data/archives/nursing-home-compare
14 Walhin JF. Bivariate ZIP models. *Biom J* 2001;43:147–60.
15 AlMuhayfith FE, Alzaid AA, Omair MA. On bivariate poisson regression models. *J King Saud Univ-Sci* 2016;28:178–89.
16 SAS Institute. Base SAS 9.4 procedures guide. SAS Institute. 2015.
17 SAS code for BMJ. figshare.
18 Data for COVID-19 in California nursing homes 2020.
19 Hillmer MP, Wodchis WP, Gill SS, *et al*. Nursing home profit status and quality of care: is there any evidence of an association? *Med Care Res Rev* 2005;62:139–66.
20 Comondore VR, Devereaux PJ, Zhou Q, *et al*. Quality of care in for-profit and not-for-profit nursing homes: systematic review and meta-analysis. *BMJ* 2009;339:b2732.
21 Harrington C, Woolhandler S, Mullan J, *et al*. Does investor ownership of nursing homes compromise the quality of care? *Am J Public Health* 2001;91:1452–5.
22 Amirkhanyan AA, Kim HJ, Lambright KT. Does the public sector outperform the nonprofit and for-profit sectors? Evidence from a national panel study on nursing home quality and access. *J Policy Anal Manage* 2008;27:326–53.
23 Johari K, Kellogg C, Vazquez K, *et al*. Ratings game: an analysis of nursing home compare and Yelp ratings. *BMJ Qual Saf* 2018;27:619–24.
24 Neuman MD, Wirtalla C, Werner RM. Association between skilled nursing facility quality indicators and hospital readmissions. *JAMA* 2014;312:1542–51.
25 Sanghavi P, Pan S, Caudry D. Assessment of nursing home reporting of major injury falls for quality measurement on nursing home compare. *Health Serv Res* 2020;55:201–10.
26 Fuller RL, Goldfield NI, Hughes JS, *et al*. Nursing home compare StAR rankings and the variation in potentially preventable emergency department visits and hospital admissions. *Popul Health Manag* 2019;22:144–52.
27 Han X, Yaraghi N, Gopal R. Catching them red-handed: optimizing the nursing homes' rating system. *ACM Trans Manag Inf Syst TMIS* 2019;10:1–26.
28 Werner RM, Konetzka RT, Polsky D. Changes in consumer demand following public reporting of summary quality ratings: an evaluation in nursing homes. *Health Serv Res* 2016;51 Suppl 2:1291–309.
29 Abrams HR, Loomer L, Gandhi A, *et al*. Characteristics of U.S. nursing homes with COVID-19 cases. *J Am Geriatr Soc* 2020;68:1653–6.
30 Harrington C, Ross L, Chapman S, *et al*. Nurse staffing and coronavirus infections in California nursing homes. *Policy Polit Nurs Pract* 2020;21:174–86.
31 McGregor MJ, Harrington C. COVID-19 and long-term care facilities: does ownership matter? *CMAJ* 2020;192:E961–2.
32 Harrington C, Olney B, Carrillo H, *et al*. Nurse staffing and deficiencies in the largest for-profit nursing home chains and chains owned by private equity companies. *Health Serv Res* 2012;47:106–28.
33 Harrington C, Pollock AM, Sutaria S. *Privatization of nursing homes in the United Kingdom and the United States. in: the privatization of care, the case of nursing homes*. Abingdon: Routledge, 2019: 51–67.

Supplementary Appendix *for*

Compress the Curve: Variations in COVID-19 Infections Across California Nursing Homes

Table of Contents

Missing Observations

Data cleaning process is presented in Figure S1. 493 nursing homes were excluded from the study sample either due to the mismatch between their names across multiple datasets or because their COVID-19 infection data were not available in CDPH reports. To examine if the excluded nursing homes are similar to those included in the study sample, we conducted two logistic regression with the dependent variables set to be 1 to indicate if a record is included in the study sample and 0 otherwise. In the first logistic regression we only include governance features as independent variables, while in the second logistic regression we include all the features.

As reported in Table S1, both regression results show that none of the governance features are statistically significant, which indicates that the included records have no selection bias on governance features. Amongst the remaining variables, quality rating and county infections per 100k are significant are statistically significant yet the difference between the two groups is not substantial, as reported in Table S2. Further, the differences in these two variables across the two groups make our estimates more conservative.

Machine learning Techniques

We then apply machine learning techniques to predict the COVID-19 infection in nursing homes and compare the results with our model. In view that our problem has a highly nonlinear structure, advanced machine learning models that do not rely on data structure assumptions may provide a flexible and desired solution. We predict the nursing home level COVID-19 infection situation by using Neural Networks (NN) and Support Vector Machines (SVM) with RBF kernel function. Variable *NH* is used as the target variable in each model, and is equal to 1 if at least one patient or staff reported to be infected. The prediction features include nursing home governance features such as occupancy rate, number of certified beds, whether a family council presents, whether the nursing home is for profit or not, and inflation score evaluated from past years. The nursing homes' health inspection rating, staffing rating and quality rating are also included in our prediction model. To capture the severity of COVID-19 epidemic in the surrounding area, we also incorporate county level COVID-19 infections per 100K population.

Bivariate and Double Poisson Estimates

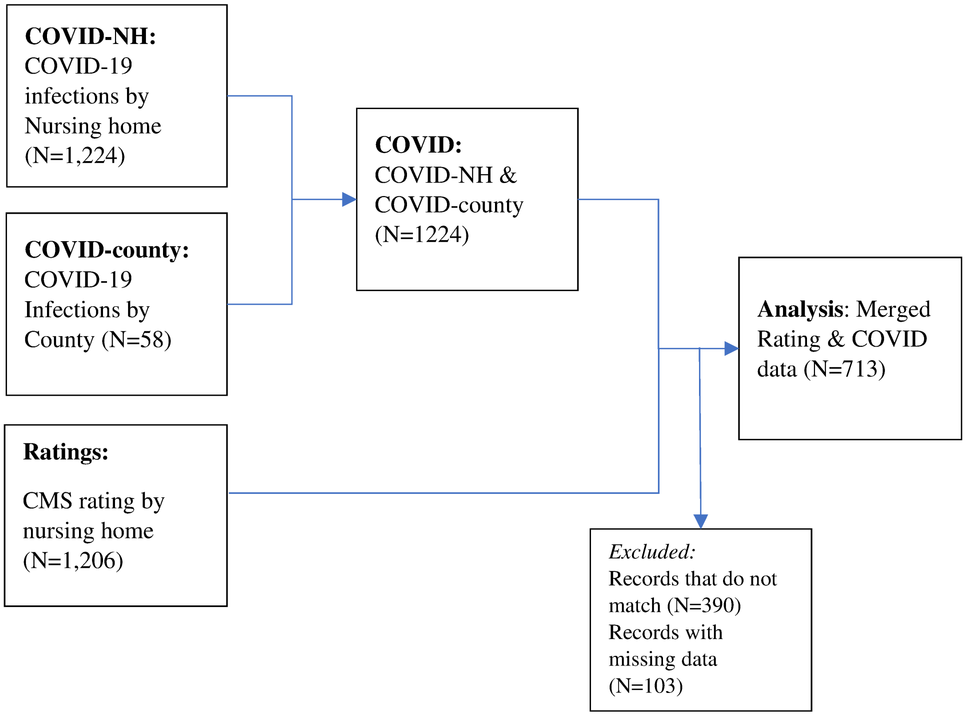
To test the robustness of our results and as a means of sensitivity analysis, we have replicated our main analysis using Bivariate and Double Poisson methods. The difference between these two methods and those reported in Table 2 of the main manuscript is these models do not assume an excess zero generating process and consider the outcome as a result of only two Poisson processes. In the Bivariate Poisson analysis, we assume that there is a correlation between the processes that give rise to the count of infections among staff and residents, while in the Double Poisson Regression, we assume independence between these two processes. The results are presented in Table S3. In comparison with the main results presented in the main table, the coefficients with larger sizes remain significant and close to their original estimates, while the smaller coefficients are not consistent with their original estimates. This is due to the fact that our dataset has significant excess zeros since most nursing homes had not reported infections many infections among either their staff or residents at the time of the study and therefore a zero inflated version of the Poisson models will be more appropriate for this setting.

Correlation Between Infections Among Staff and Residents

To better examine the correlation between infections among staff and residents, we report the number and percentage of nursing homes with and without infections among their staff and residents in Table S6. We can observe that 91.75% of nursing homes with no infections among their residents also experienced no infections among their staff. Similarly, 54.21% of nursing homes that had at least one infection among their residents, also had at least one infection among their staff. In Figure S4, we show the scatter plot of number of infections among staff and residents for only those nursing homes that experienced a large outbreak among both their staff and residents. There is a clear correlation between the number of infections among staff and residents.

Figures

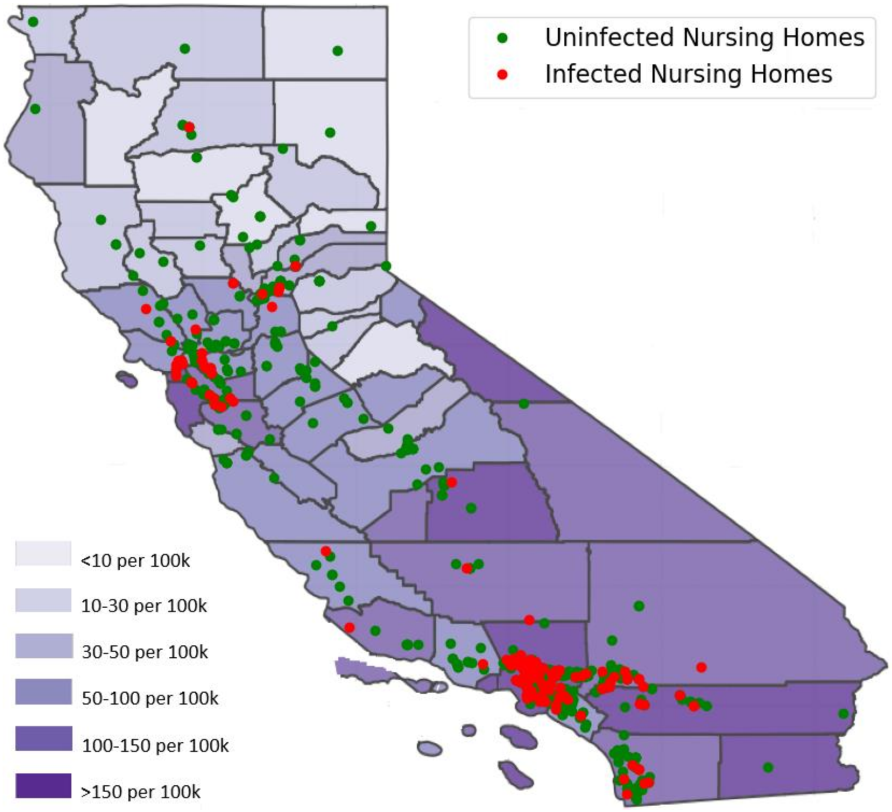
Figure S1. Study population and analysis sample


COVID-NH: COVID-19 infections by Nursing home (N=1,224)
COVID-county: COVID-19 Infections by County (N=58)
Ratings: CMS rating by nursing home (N=1,206)
COVID: COVID-NH & COVID-county (N=1224)
Analysis: Merged Rating & COVID data (N=713)
Excluded: Records that do not match (N=390) Records with missing data (N=103)

Note: Original CMS Rating for year 2017 data (*ratings*) include 1206 nursing homes. Original CA COVID-19 Infection by county (*COVID-county*) data as of April 30th, 2020 include on 58 counties Original COVID-19 CA Infections by nursing homes (*COVID-NH*) data as of April 30th, 2020 include 1224 nursing homes.

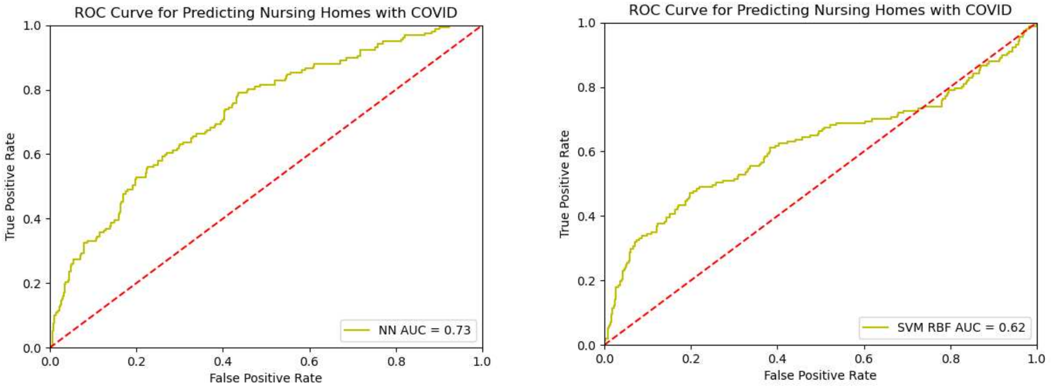
We first merged *COVID-NH* and *COVID-county* data for all 1224 rows (0 record lost). We then merged the resulting data (*COVID*) with *ratings* data which resulted in 713 rows. 390 records were lost due to mismatch between the names of the facilities in the two datasets, and 103 records were lost for those nursing homes that did not report COVID 19 infection data or their ratings information is missing.

Figure S2: Spread of COVID-19 Infection Among California Nursing Homes



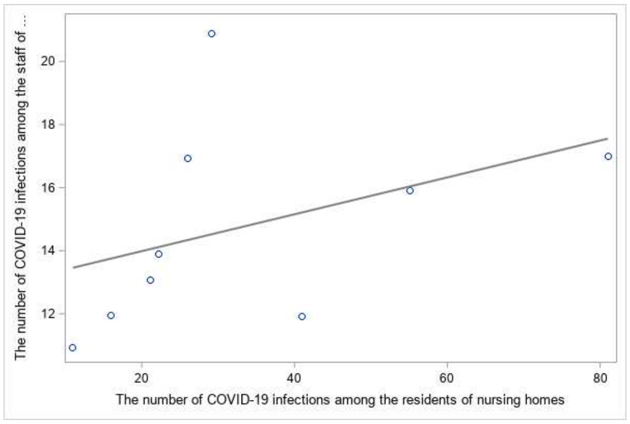
Note: The figure presents the spread of COVID-19 infection among California nursing homes as of May 1st, 2020

Figure S3: Receiver Operator Characteristic (ROC) Curves for Predicting at Least One Infection in Nursing Homes



Note: ROC for Nursing Home (NH) COVID-19 prediction using Neural Networks (NN), SVM with RBF kernel. The AUC is reported for each model: NN=0.73, SVM-RBF (default)=0.62

Figure S4: Scatter plot of number of infections among staff and residents for those nursing homes that have experienced large outbreaks amongst both their staff and resident populations



Tables

Table S1: Logistic Regression Results for Estimating the Effects of Nursing Homes' Features on Odds of Being Included in the Study Sample

	Validation with Governance Features Only (Included vs. Excluded Records)			Validation with All Features (Included vs. Excluded Records)		
Parameter	Estimate	(95% CI)	P Value	Estimate	(95% CI)	P Value
Constant	0.1	(-0.72 to 0.92)	0.81	-0.66	(-2.09 to 0.76)	0.36
For profit	0.25	(-0.08 to 0.58)	0.14	0.29	(-0.1 to 0.68)	0.14
Family council	-0.19	(-0.49 to 0.12)	0.23	-0.07	(-0.4 to 0.26)	0.68
Certified beds	-0.0004	(-0.003 to 0.002)	0.71	-0.0008	(-0.003 to 0.002)	0.52
Occupancy rate	0.61	(-0.3 to 1.52)	0.19	0.56	(-0.62 to 1.74)	0.35
Inflation score	-0.04	(-0.2 to 0.12)	0.6	-0.03	(-0.2 to 0.14)	0.75
Quality rating				0.21	(0.07 to 0.36)	0.004
Staffing rating				0.002	(-0.14 to 0.14)	0.97
Health inspection rating				0.08	(-0.04 to 0.19)	0.21
County infections per 100K				-0.002	(-0.004 to -0.0007)	0.004

Note: Coefficients represent how the log odds of the dependent variable changes with one unit increase in the corresponding predictor

Table S2: Results of Two-Sample t-Test for Equality of the Means of the Excluded and Included Observations

Features	Excluded Records*	Included Records*	P Value**
For profit	0.82	0.86	0.11
Family council	0.21	0.18	0.21
Certified beds	99.6	98.0	0.65
Occupancy rate	0.85	0.86	0.14
Inflation score	0.32	0.31	0.83
Quality rating	4.43	4.57	0.01
Staffing rating	3.49	3.49	0.93
Health inspection rating	2.66	2.86	0.01
County infections per 100K	159.36	143.88	0.003

Note: *: Reports the average value of features.

**:P values are for two-tailed t-tests of the equality of the two means.

Table S3: Replication of the main analysis results using Bivariate and Poisson Regression Models

	Bivariate Poisson Model			Double Poisson Model		
Parameter	Estimate	(95% CI)	P Value	Estimate	(95% CI)	P Value
Infected Staff (number of staff with confirmed COVID-19 infections)						
Intercept	-3.9	(-5.97 to -1.83)	0.01	-3.29	(-4.7 to -1.88)	<.001
County infections per 100K	0.01	(0.01 to 0.01)	<.001	0.01	(0.01 to 0.01)	<.001
For profit	0.33	(-0.28 to 0.93)	0.3	0.01	(-0.37 to 0.39)	0.97
Family council	-0.08	(-0.59 to 0.43)	0.77	0.18	(-0.1 to 0.46)	0.21
Certified beds	0.01	(0.01 to 0.01)	<.001	0.01	(0.01 to 0.01)	<.001
Occupancy rate	-2.5	(-4.05 to -0.95)	0.01	-0.89	(-2.02 to 0.24)	0.13
Inspection rating	0.1	(-0.1 to 0.28)	0.35	-0.12	(-0.23 to -0.01)	0.05
Quality rating	0.25	(-0.05 to 0.54)	0.11	0.21	(0.03 to 0.39)	0.03
Staffing rating	0.12	(-0.06 to 0.29)	0.19	0.26	(0.14 to 0.38)	<.001
Inflation score	0.49	(0.39 to 0.59)	<.001	0.31	(0.23 to 0.39)	<.001
Infected Residents (number of residents with confirmed COVID-19 infections)						
Intercept	-2.1	(-3.01 to -1.19)	<.001	-1.46	(-2.2 to -0.71)	0.01
County infections per 100K	0.01	(0.01 to 0.01)	<.001	0.01	(0.01 to 0.01)	<.001
For profit	2.71	(2.12 to 3.31)	<.001	1.89	(1.5 to 2.28)	<.001
Family council	0.16	(0.02 to 0.3)	0.03	0.19	(0.06 to 0.31)	0.01
Certified beds	0.01	(0.01 to 0.01)	<.001	0.01	(0.01 to 0.01)	<.001
Occupancy rate	-0.08	(-0.66 to 0.51)	0.82	0.02	(-0.54 to 0.57)	0.96
Inspection rating	-0.2	(-0.25 to -0.14)	<.001	-0.21	(-0.26 to -0.16)	<.001
Quality rating	0.05	(-0.03 to 0.13)	0.2	0.08	(-0.01 to 0.15)	0.06
Staffing rating	-0.22	(-0.27 to -0.17)	<.001	-0.15	(-0.2 to -0.11)	<.001
Inflation score	0.13	(0.08 to 0.18)	<.001	0.13	(0.08 to 0.17)	<.001
Covariance	0.21	(0.18 to 0.25)	<.001			
Fit Statistics						
-2 log likelihood		8011.7			8468.6	
AIC		8053.7			8508.6	
BIC		8149.7			8600.0	

Table S4: Confusion Matrix for SVM-RBF

		ACTUAL CLASS	
		0	1
PREDICTED CLASS	0	142	2
	1	47	7

Table S5: Confusion Matrix for NN

		ACTUAL CLASS	
		0	1
PREDICTED CLASS	0	137	7
	1	37	17

Table S6: Distribution of Infections Among Staff and Residents

		INFECTIONS AMONG STAFF (%)	
		0	>=1
INFECTIONS AMONG RESIDENTS (%)	0	556 (91.75%)	50 (8.25%)
	>=1	49 (45.79%)	58 (54.21%)