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

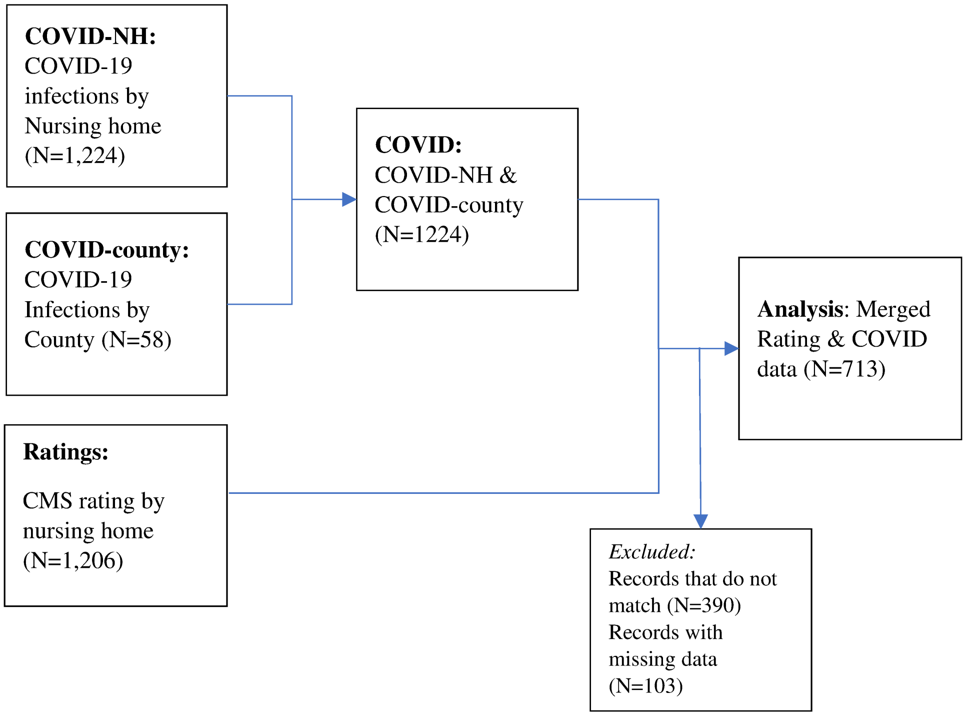
To test the robustness of our results and as a means of sensitivity analysis, we have replicated our main analysis using Bivariate and Double Poisson methods. The difference between these two methods and those reported in Table 2 of the main manuscript is these models do not assume an excess zero generating 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

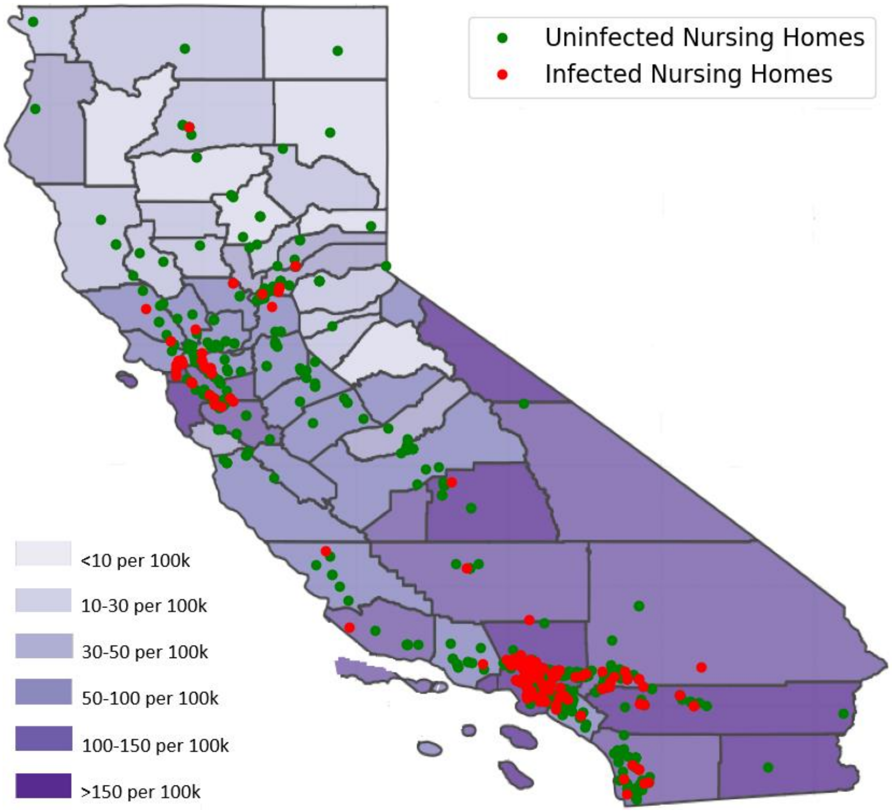
Figure S1. Study population and analysis sample


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.

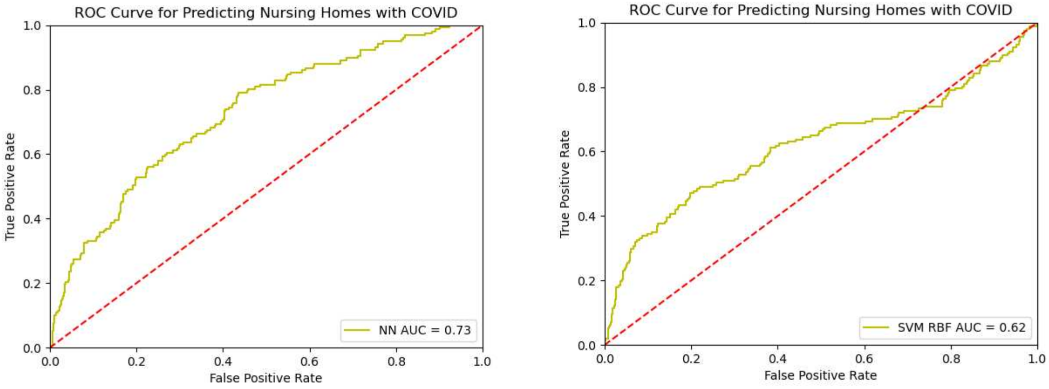
We first merged *COVID-NH* and *COVID-county* data for all 1224 rows (0 record lost). We then merged the resulting data (*COVID*) with *ratings* data which resulted in 713 rows. 390 records were lost due to mismatch between the names of the facilities in the two datasets, and 103 records were lost for 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



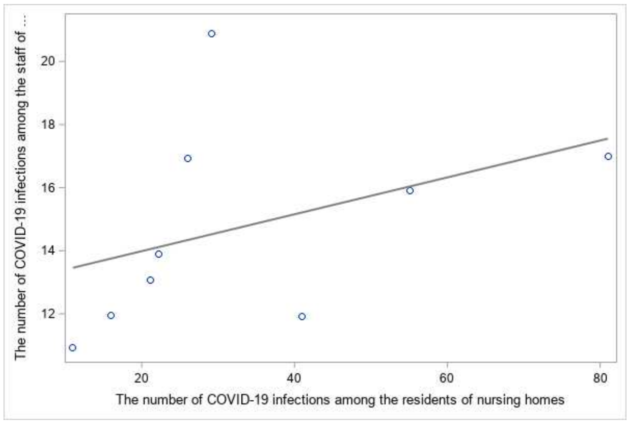
Note: The figure presents the spread of COVID-19 infection among California nursing homes as of May 1st, 2020

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%)