JAMDA 21 (2020) 1533–1538

JAMDA

journal homepage: www.jamda.com

Original Study

Predicting Coronavirus Disease 2019 Infection Risk and Related Risk Drivers in Nursing Homes: A Machine Learning Approach

Christopher L.F. Sun PhD [a,b], Eugenio Zuccarelli MBAn, MSc [a,c], El Ghali A. Zerhouni MBAn [a,c], Jason Lee MBA, MS [a,d], James Muller BEc [e], Karen M. Scott MPA [a], Alida M. Lujan MPA, MBA [a], Retsef Levi PhD [a,*]

[a] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA
[b] Healthcare Systems Engineering, Massachusetts General Hospital, Boston, MA, USA
[c] Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA
[d] School of Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[e] Muller Consulting and Data Analytics, LLC, Washington, DC, USA

Keywords:
Nursing homes
COVID-19
infection prevention
health policy
long-term care facility
risk modeling
machine-learning

A B S T R A C T

Objective: Inform coronavirus disease 2019 (COVID-19) infection prevention measures by identifying and assessing risk and possible vectors of infection in nursing homes (NHs) using a machine-learning approach.
Design: This retrospective cohort study used a gradient boosting algorithm to evaluate risk of COVID-19 infection (ie, presence of at least 1 confirmed COVID-19 resident) in NHs.
Setting and Participants: The model was trained on outcomes from 1146 NHs in Massachusetts, Georgia, and New Jersey, reporting COVID-19 case data on April 20, 2020. Risk indices generated from the model using data from May 4 were prospectively validated against outcomes reported on May 11 from 1021 NHs in California.
Methods: Model features, pertaining to facility and community characteristics, were obtained from a self-constructed dataset based on multiple public and private sources. The model was assessed via out-of-sample area under the receiver operating characteristic curve (AUC), sensitivity, and specificity in the training (via 10-fold cross-validation) and validation datasets.
Results: The mean AUC, sensitivity, and specificity of the model over 10-fold cross-validation were 0.729 [95% confidence interval (CI) 0.690–0.767], 0.670 (95% CI 0.477–0.862), and 0.611 (95% CI 0.412–0.809), respectively. Prospective out-of-sample validation yielded similar performance measures (AUC 0.721; sensitivity 0.622; specificity 0.713). The strongest predictors of COVID-19 infection were identified as the NH's county's infection rate and the number of separate units in the NH; other predictors included the county's population density, historical Centers of Medicare and Medicaid Services cited health deficiencies, and the NH's resident density (in persons per 1000 square feet). In addition, the NH's historical percentage of non-Hispanic white residents was identified as a protective factor.
Conclusions and Implications: A machine-learning model can help quantify and predict NH infection risk. The identified risk factors support the early identification and management of presymptomatic and asymptomatic individuals (eg, staff) entering the NH from the surrounding community and the development of financially sustainable staff testing initiatives in preventing COVID-19 infection.

Long-term care facilities (LTCFs) have emerged as critical epicenters of coronavirus disease 2019 (COVID-19) outbreaks and are associated with approximately 1 in 10 COVID-19 cases and 1 in 3 COVID-19 fatalities in the United States.[1] Among LTCFs, nursing homes (NHs) have been shown to have high-risk populations that are particularly vulnerable to COVID-19 infection and poor subsequent outcomes.[2,3] Rapid COVID-19 transmission within NHs stress the need for proactive measures preventing infection and facility spread.[4–7] However,

Dr Sun is a recipient of a Canadian Institutes of Health Research Fellowship.
The authors declare no conflicts of interest.
* Address correspondence to Retsef Levi, PhD, Sloan School of Management, Massachusetts Institute of Technology, 100 Main St, Room E62-562, Cambridge, MA 02142-1347.
E-mail address: retsef@mit.edu (R. Levi).

https://doi.org/10.1016/j.jamda.2020.08.030
1525-8610/

developing effective policies and interventions is challenging because of a lack of both accurate data sources as well as data-driven analyses regarding infection vectors.[8]

This study describes the development and implications of a machine-learning model, trained on NH COVID-19 outcome data from multiple US states, to assess risk of COVID-19 infection and identify associated risk factors and possible infection introduction mechanisms.

Methods

Study Setting and Population

This study included public NH COVID-19 facility-level case data reported by state and local departments of health across the United States, which were collected to create a binary outcome variable for whether there was at least 1 resident infection in the facility. The model was trained on COVID-19 outcomes reported on April 20, 2020 from 1146 NHs in Massachusetts, Georgia, and New Jersey, and prospectively validated out-of-sample against outcomes reported on May 11, 2020 from 1021 NHs in California. These states had relatively comprehensive reporting and testing capacity at the time of outcome collection (see Supplementary Material, Data Sources and Model Inputs section for details).

Data Sources

Predictive features were created from a self-constructed dataset, integrating public and private sources from organizations including the Centers of Medicare and Medicaid Services (CMS), Long-Term Care Focus, and National Investment Center for Seniors Housing and Care, covering 15,300 federally certified US NHs and their surrounding communities. The dataset includes information on each NH's physical infrastructure, number of units, and historical financial, managerial, resident, staffing, and quality-of-care characteristics. In addition, the dataset includes NH community information (eg, COVID-19 infection rates on the day of NH reporting, social distancing, and population characteristics and mobility measures). See the Supplementary Table 1 for details on the dataset and model features.

Model Development

The model used a tree-based gradient boosting algorithm,[9] predicting a binary classification outcome by facility that signifies at least 1 resident COVID-19 infection. In other words, the model generates a risk index associated with the likelihood of COVID-19 infection at the NH. Prior to training, highly correlated predictors with a variance inflation factor greater than 10 were removed. Then, the remaining predictors were assessed for predictive stability by determining the number of times each predictor was selected by the model following L1 regularization during 10-fold cross-validation (CV). Features selected at least 7 times out of the 10 folds were considered stable predictors.

The model inputs were restricted to the identified stable predictors and hyperparameter tuning and out-of-sample performance assessments were conducted via 10-fold CV. The mean out-of-sample area under the receiver operating characteristic curve (AUC), sensitivity, and specificity, along with the associated 95% confidence intervals (CIs), were calculated over the 10-folds. The final tuned model was fit over the entire training dataset and was used to calculate the stable predictors' feature importance and predict updated risk indices for model validation. A logistic regression model and 2-layer feedforward neural network were also developed using the identified stable predictors and assessed via the same methods, serving as benchmark predictive models for comparison. Although a neural network model is not easily interpretable, the model was used as a benchmark for comparison due its strong predictive capabilities. Model development and feature importance evaluation is further described in the Supplementary Material (Model Development and Feature Importance Evaluation section).

Model Validation

To assess the model's prognostic ability and generalizability, risk indices from May 4 were prospectively validated against NH outcomes reported on May 11 from California. The predicted risk indices were categorized as binary outcomes using an optimal threshold value (selected as the value that minimizes the difference between the model's sensitivity (true positive rate) and specificity (true negative rate) across the entire training dataset). Risk indices above the threshold value were predicted as infected NHs. The differences in the predictive characteristics between the training and validation datasets were compared using the Mann–Whitney U test for continuous variables, and the χ^2 test for binary variables.

In addition, reported outcomes from 7660 LTCFs[1] on May 11 were used to calculate the Pearson correlation coefficient between each state's median NH risk index (ie, the median risk index, from May 4, across all NHs that are in our dataset in the state) and each state's LTCF related COVID-19 infection and death rates. The benchmark logistic regression and neural network models were also validated in the same manner, and the performance of the 3 models were compared. Model validation is further detailed in the Supplementary Material (Prospective Out-of-Sample Model Validation section).

Results

Table 1 summarizes and compares the characteristics the NHs used to train and validate the model. The training set included 1146 NHs that reported COVID-19 cases on April 20 (60.3% reported at least one resident COVID-19 case). The validation set included 1021 NHs (20.5% reported at least 1 resident COVID-19 case) reporting on May 11. The NH characteristics in the validation set was significantly different from the training set, indicating the validation set is suitable to rigorously assess the model's out-of-sample predictive performance and generalizability to unseen data.

Overall, 7 out of 41 inputted features were identified as predictors of infection (Figure 1). The NH's county's infection rate and number of units were the strongest predictors of risk and positively associated with increased infection risk. The other predictors of infection include the NH's county's population density, CMS cited health deficiencies, and resident and staff densities, which were positively associated with infection risk, as well as the percent of non-Hispanic White residents, which was negatively associated with infection risk (Supplementary Figure 1). The gradient boosting model's mean out-of-sample AUC, sensitivity, and specificity from 10-fold CV over the training set were 0.729 (95% CI 0.690–0.767), 0.670 (95% CI 0.477–0.862), and 0.611 (95% CI 0.412–0.809), respectively.

The model had an AUC of 0.721 (sensitivity 0.622; specificity 0.713) when prospectively compared against California NHs with reported outcomes from May 11. The optimal threshold value used to form binary outcome classifications from predicted risk indices was 0.618. Table 2 shows LTCF related case and death rates from May 11 with the model's median risk indices by state. The correlation was statistically significant for both case ($R = 0.859$; $P < .001$) and death ($R = 0.856$; $P < .001$) rates.

Compared with the benchmark models, logistic regression and neural network (Table 3), the gradient boosting model demonstrated stronger prognostic ability and higher correlation to LTCF case and death rates by state. The gradient boosting model had higher mean

Table 1
The Summary and Comparison of the Predictive Characteristics of the NH in the Model's Training and Validation Sets

Identified Predictive Features	Training Set* (n = 1146)	Prospective Validation Set† (n = 1021)	*P* Value
Cumulative number of positive COVID-19 cases per capita in the facility's county on the day of NH COVID-19 case reporting (confirmed cases per 100,000 people), median (IQR)	478.1 (182.0–730.7)	112.5 (79.5–244.7)	<.001
Total number of beds at the facility, median (IQR)	122 (94–167)	99 (74–140)	<.001
Population density of the facility's county (population per square mile), median (IQR)	1027.0 (420.6–2033.7)	1613.3 (343.8–2508.6)	<.05
Number of health deficiencies at the facility as defined by the CMS, median (IQR)	12 (7–19)	35 (23–50)	<.001
Percent of NH residents who were non-Hispanic white prior to the COVID-19 outbreak, median (IQR)	83.6 (62.0–94.2)	59.6 (42.9–78.5)	<.001
Number of patients per 1000 square feet in the facility prior to the COVID-19 outbreak, median (IQR)	2.95 (2.15–3.77)	4.83 (3.75–5.68)	<.001
Number of clinical workers per 1000 square feet in the facility prior to the COVID-19 outbreak, median (IQR)	1.06 (0.78–1.33)	1.90 (1.52–2.32)	<.001
Positive COVID-19 resident case in NH, No. (%)	722 (63.0)	209 (20.5)	<.001

IQR, interquartile range.
Significant differences in the characteristics between the 2 sets were found. A strong predictive performance across a validation set population that is significantly different from its training set population suggests the model will be generalizable to different populations. *P* values from Mann–Whitney U and χ^2 tests, as appropriate, comparing the differences in the characteristics are shown.
*NHs from Massachusetts, Georgia, and New Jersey with outcomes reported on April 20, 2020.
†NHs from California with outcomes reported on May 11, 2020.

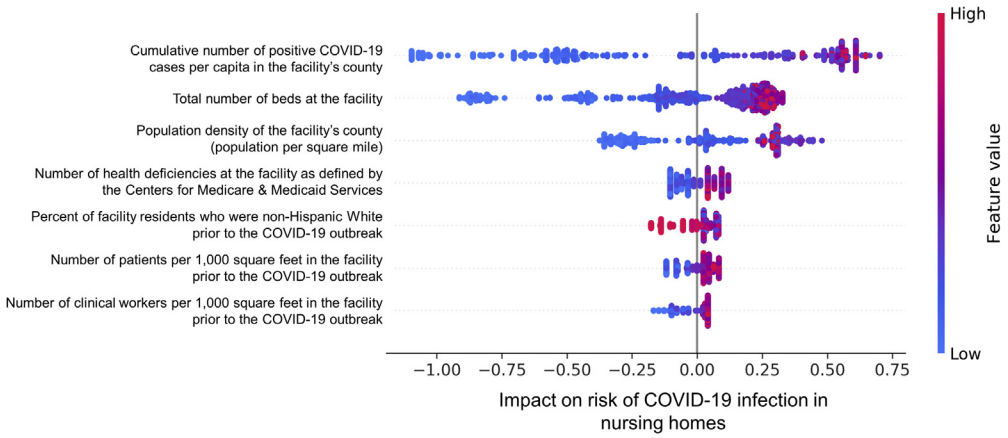


Fig. 1. Feature importance and impact on risk of COVID-19 infection in NHs from the gradient boosting model. The NH's county's COVID-19 infection rate and size had the largest impact on infection risk (features are in descending order from highest to lowest importance). In the figure, each dot represents a NH that the model has been trained on. For each NH, a high feature value corresponds to the color red, and a low feature value corresponds to the color blue. The horizontal axis shows whether the effect of the feature value is associated with a higher or lower risk of NG infection.

out-of-sample AUC, sensitivity, and specificity compared with the logistic regression and neural network models from 10-fold CV over the training set (Table 3). In the validation set, the logistic regression model had a lower AUC (0.689) compared with the gradient boosting model and a large difference in sensitivity (0.914) and specificity (0.233), indicating its poor predictive power (overestimating the number of infected NH's) and generalizability. Similarly, the neural network had a lower AUC (0.707) compared with the gradient boosting model and a large discrepancy in sensitivity (0.904) and specificity (0.308) across the validation set, also indicating overestimation of infected NH's and poor model generalizability. The optimal threshold value used to form binary outcome classifications from the logistic regression and neural network model predictions were 0.609 and 0.640, respectively. The correlation between the gradient boosting model's median risk index and LTCF outcome rates by state was stronger compared with both the logistic regression and neural network models for both state case rates, R = 0.384 ($P < .05$) and R = 0.731 ($P < .001$), respectively, as well as state death rates, R = 0.335 ($P < .05$) and R = 0.705 ($P < .001$), respectively. The benchmark logistic regression model is further detailed in the Supplementary Table 2 for the interested reader.

Discussion

Predicting COVID-19 outbreaks in senior care facilities has been a challenge for policymakers and nursing home operators who prioritize the allocation of various critical resources (eg, personal protection equipment (PPE), training, audits, testing) to prevent and mitigate outbreak and its consequences.[10–12] For example, previous studies have mixed results on the relationship between standard LTCF ratings, such as the CMS 5-star overall and health inspection ratings, and increased risk of infection.[11–16] This study describes the development and application of a machine-learning gradient boosting model to quantify complex predictive relationships between NH COVID-19 infection risk and granular NH characteristics that were decomposed from traditional aggregated NH measures, highlighting factors contributing to NH infection during the initial COVID-19 outbreak phase (March/April 2020). The model demonstrated moderate

Table 2
Predicted NH Risk from the Gradient Boosting Model and LTCF Related COVID-19 Case and Death Rates Reported on May 11 by State

State Ranking Based on Predicted NH Risk Index (as of May 4, 2020)	State	Predicted on May 4, 2020	Reported on May 11, 2020	Reported LTCF Related Deaths per 1000 Beds (Relative Rank)
		Median Predicted NH Risk Indices (IQR)	Reported LTCF Related Cases per 1000 Beds (Relative Rank)	
1	New Jersey	78.7 (67.1–82.7)	500.9 (1)	92.7 (1)
2	Massachusetts	74.8 (65.1–81.2)	365.3 (2)	66.9 (2)
3	Connecticut	71.3 (50.6–78.3)	251.5 (3)	63.3 (3)
4	New York	66.1 (44.4–83.5)	No reporting data	47.8 (4)
5	Maryland	65.7 (49.8–72.2)	226.3 (5)	28.8 (7)
6	Rhode Island	63.1 (54.6–70.8)	230.1 (4)	37.9 (5)
7	Delaware	62.6 (58.5–72.2)	91.9 (14)	28.0 (8)
8	Louisiana	58.0 (42.9–74.5)	112.3 (11)	23.3 (10)
9	California	54.0 (39.9–63.7)	82.7 (15)	8.4 (21)
10	Pennsylvania	51.0 (37.8–68.1)	152.5 (8)	29.0 (6)
11	Florida	51.0 (39.7–59.6)	66.9 (17)	8.6 (19)
12	Virginia	49.6 (37.7–63.0)	115.1 (10)	15.2 (14)
13	Michigan	45.8 (34.0–70.1)	99.7 (13)	4.6 (27)
14	Illinois	45.1 (33.9–73.9)	127.5 (9)	17.4 (12)
15	Colorado	44.8 (35.6–59.9)	184.7 (6)	26.8 (9)
16	Washington	44.8 (36.6–51.7)	51.0 (26)	3.8 (32)
17	Georgia	44.8 (36.7–62.6)	158.9 (7)	17.5 (11)
18	Nevada	44.8 (40.7–48.6)	102.3 (12)	8.6 (20)
19	Utah	43.7 (31.5–53.8)	10.3 (40)	2.0 (37)
20	Mississippi	43.6 (37.7–52.6)	66.4 (18)	10.5 (17)
21	Indiana	43.5 (36.2–52.4)	59.1 (20)	11.4 (16)
22	Alabama	41.5 (34.1–47.3)	62.2 (19)	1.0 (40)
23	Ohio	40.7 (31.2–56.0)	54.2 (24)	3.2 (34)
24	Texas	40.5 (31.5–56.0)	10.1 (41)	3.6 (33)
25	South Carolina	40.0 (33.3–44.8)	54.4 (23)	5.4 (25)
26	Nebraska	39.8 (31.5–45.3)	5.2 (45)	0.1 (45)
27	Hawaii	38.6 (31.5–45.4)	0.7 (49)	No reporting data
28	New Mexico	37.7 (31.5–47.3)	5.7 (42)	2.2 (36)
29	North Carolina	37.7 (31.5–46.2)	56.4 (21)	7.3 (22)
30	Arizona	37.7 (31.9–44.0)	76.0 (16)	12.7 (15)
31	Alaska	36.6 (31.5–42.0)	3.9 (46)	No reporting data
32	New Hampshire	33.9 (26.6–39.8)	19.3 (38)	1.7 (38)
33	Vermont	33.7 (31.5–38.7)	55.8 (22)	9.6 (18)
34	Tennessee	33.5 (28.4–44.8)	22.6 (36)	2.4 (35)
35	Kentucky	33.3 (28.0–43.6)	53.3 (25)	6.7 (23)
36	Oklahoma	33.2 (28.5–44.8)	35.8 (32)	4.3 (29)
37	Arkansas	33.1 (29.8–40.5)	17.8 (39)	1.4 (39)
38	Iowa	32.4 (28.0–41.5)	36.8 (31)	0.5 (42)
39	Missouri	32.1 (28.0–44.8)	2.4 (48)	0.2 (43)
40	Idaho	31.6 (29.6–40.5)	29.1 (34)	4.7 (26)
41	Minnesota	31.5 (28.0–46.4)	48.7 (27)	16.8 (13)
42	Kansas	31.5 (28.0–40.7)	26.2 (35)	4.2 (30)
43	Wyoming	31.5 (28.0–37.7)	5.4 (43)	No reporting data
44	West Virginia	31.5 (29.7–37.7)	30.7 (33)	4.0 (31)
45	Oregon	31.5 (31.3–50.8)	39.9 (29)	6.3 (24)
46	Montana	29.8 (28.0–34.4)	5.4 (44)	0.9 (41)
47	North Dakota	29.1 (27.6–39.5)	44.0 (28)	No reporting data
48	Maine	28.9 (25.5–35.1)	37.7 (30)	4.4 (28)
49	South Dakota	28.5 (25.3–36.6)	2.8 (47)	No reporting data
50	Wisconsin	28.0 (25.3–37.9)	22.1 (37)	0.2 (44)

IQR, interquartile range.
States were ranked in descending order based on the state's median, 75th percentile and 25th percentile risk index as of May 4, 2020.

predictive power and strong association with NH and LTCF outcomes across the United States, suggesting the value of the identified risk factors in predicting which NHs are most susceptible to infection introduction. The gradient boosting approach outperformed logistic regression and neural network benchmark models, further demonstrating its ability in providing insights to inform healthcare policies to prevent COVID-19 infection.

The identified risk factors provide data-driven support of hypotheses regarding 2 primary infection mechanisms: (1) introduction via presymptomatic and asymptomatic individuals from the surrounding community, and (2) intra-facility transmission following initial exposure.[4,5,7,14] Opportunities for infection introduction increase with the number and frequency of individuals entering the NH from the surrounding community. Intra-facility transmission following exposure from the outside community appears to increase with staff and resident density, suggestive of greater interaction within the NH. In addition, historical CMS-cited health deficiencies could indicate poor safety culture, inappropriate infection control practices, and lack of financial resources to implement appropriate safety measures,[17,18] all of which may impact both infection introduction and spread.[17,19,20] Lastly, a higher percent of non-Hispanic white residents was associated with lower risk of infection, consistent with the racial disparities of COVID-19 infection risk, as well as social and structural determinants of health, affecting both the general public[21–25] and the geriatric and NH community.[10–12,26] As lower long-term[18,27] and post-acute care quality,[28] as well as more limited financial resources[29] have been found in NHs with a higher percentage of minority residents, these results further suggest poor infection

Table 3
The Gradient Boosting Model's Performance and Correlation to LTCF Related COVID-19 Case and Death Rates by State Compared with the Performance of the Benchmark Logistic Regression and Neural Network Models

Dataset	Metric of Interest	Gradient Boosting Model	Benchmark Logistic Regression Model	Benchmark Neural Network Model
Training set* (via 10-fold cross validation)	AUC, mean (95% CI)	0.729 (0.690–0.767)	0.653 (0.599–0.706)	0.696 (0.657–0.734)
	Sensitivity, mean (95% CI)	0.670 (0.477–0.862)	0.610 (0.483–0.738)	0.664 (0.484–0.843)
	Specificity, mean (95% CI)	0.611 (0.412–0.809)	0.592 (0.450–0.733)	0.585 (0.410–0.760)
Prospective validation set†	AUC	0.721	0.689	0.707
	Sensitivity	0.622	0.914	0.904
	Specificity	0.713	0.233	0.308
State LTCF outcome rates‡	Correlation between median risk index and state LTCF case rates by state, Pearson correlation coefficient	0.859	0.384	0.731
	Correlation between median risk index and LTCF deaths rates by state, Pearson correlation coefficient	0.856	0.335	0.705

*NHs from Massachusetts, Georgia, and New Jersey with outcomes reported on April 20, 2020.
†NHs from California with outcomes reported on May 11, 2020.
‡LTCF-related COVID-19 case and death rates reported on May 11th by states across the United States.

control practice and limited access to infection control resources may impact COVID-19 introduction. These factors help inform policy priorities that have emerged in NH COVID-19 management: staff and resident testing; positive workforce practices; PPE availability and proper use; financial relief for NHs; and development of high-quality facility level COVID-19 databases.

The importance of community transmission supports evidence that early identification and management of presymptomatic and asymptomatic individuals, particularly staff who frequently enter and exit the facility, can be effective in infection introduction.[4,5,30–32] The role that presymptomatic and asymptomatic individuals play in transmission underscores the importance of frequent surveillance testing of staff as a preferable policy to symptom screening in effectively preventing infection introduction.[33] Staff have not been prioritized in many NH testing strategies, which have been extremely inconsistent across states.[34,35] Less than one-half of the states were reporting COVID-19 cases in staff during the initial COVID-19 outbreak in April,[36] some of which did not even perform staff testing following a NH outbreak.[37] The development of a state or federally supported surveillance testing approach for staff during the COVID-19 pandemic is essential to sustain effective infection prevention practices in NHs; most important, is securing funding sources and providing operational capacity — such questions have been raised in many states and most notably in New York which has recently mandated regular testing of workers.[38,39] In addition to testing, state-supported workforce policies could help address staffing shortages, facilitate effective organizational communication, and provide paid sick leave as COVID-positive workers are identified.

Once a facility has at least 1 COVID-19 case, the relevant mechanism for infection to consider is intrafacility transmission among residents and staff.[4,5,7] The positive association of risk to resident and staff density supports interventions that minimize staff transitions across parts of the facility, and that limit unnecessary in-person interactions with residents. In the short term, intrafacility infection spread may be lower in facilities with reduced occupancy rates as a result of the first wave of the COVID outbreak. At the same time, reduced occupancy has a significant financial impact on facilities, particularly from decreased Medicare revenue associated with low post-acute care referrals and increased patient management costs.[3] At the federal level, short-term policies that bring Medicaid payments in line with Medicare payments per head and eliminate low occupancy penalization should be considered to provide financial stability to facilities while at the same time reducing the risk of intra-facility spread. And finally, immediate actions to increase available PPE for NH staff, which have been in shortage,[3,8,31,35,40] are also essential, as unprotected and asymptomatic staff are likely primary vectors accelerating infection spread.

The challenge of improving COVID-19 outcomes in NHs and compliance to infection policies emphasize the continuing role of data analytics and advanced modeling techniques to inform NH response. Risk indices, such as the one generated by our model, can be used by policymakers to prioritize certain facilities for enhanced support, as well as reveal critical support needs. Moreover, predictive risk models can be instrumental in informing the relaxation and tightening of NH visitor policies. Diligence around identifying risk factors and drivers of infection will remain critical through future COVID-19 recovery phases.

Maintaining quality, up-to-date facility level data will help inform data-driven analysis in the dynamically changing NH and COVID-19 landscape. Moving forward, health organizations including CMS and Centers for Disease Control and Prevention would benefit from developing high-quality national datasets to inform infection and control policies. Along with this, frequent assessment of NH characteristics that are relevant to informing decision-makers should be conducted to support analysis of suspected infection mechanisms. For example, the inflow and outflow of residents and workforce, staffing levels, and workforce status within NHs have been points of interest[4,5] that have not been reported in public datasets.

Applying these modeling techniques to inform targeted interventions may also improve COVID-19 outcomes in other institutional settings, such as homeless shelters and correctional facilities, that have experienced rapid intra-facility transmissions.[30,41] The strong correlation between state median risk indices and LTCF case and death rates can be explained as most infections and deaths occurred in NHs, but also could indicate the relevance of the risk factors to such settings.

This study has several limitations. First, NH COVID-19 outcomes were inconsistently reported across states and could underestimate actual infection and fatality rates. Partial testing of NHs could also result in underestimations of outcomes. To mitigate this risk, training data was collected from states with relatively higher testing levels, and better data quality. Second, while the model performed strongly when validated on a state with significantly different characteristics from the training states, model performance could still be inconsistent across different geographic areas. Lastly, model predictors describing NHs were developed from historical reports, such as those from the 2017 Long-Term Care Focus database and may not reflect real-time NH characteristics.

Conclusions and Implications

A machine-learning gradient boosting model can describe and predict the risk of COVID-19 outbreak in NHs, providing data-driven support for NH infection control policies, strategies for the prioritization of resources to high-risk NHs, and the relaxation and restriction of NH visitor policies. The prevalence of COVID-19 infections in a NH's surrounding community and a NH's size were identified as the primary risk factors associated with NH infection, suggesting that the introduction of infection from the outside community as a likely infection mechanism. Developing financially sustainable testing and screening approaches to identify presymptomatic and asymptomatic individuals entering a NH are critical to preventing and controlling COVID-19 outbreaks in these settings.

Acknowledgments

We thank Lily Bailey (Massachusetts Institute of Technology) for her contributions to data acquisition and processing. We also thank Simon Johnson and Kate Kellogg (Massachusetts Institute of Technology) for multiple discussions related to infection risk in NHs. We thank the National Investment Center for Seniors Housing and Care and its associated NIC MAP data service, for their contributions in data acquisition and interpretation. We thank Massachusetts Senior Care Association for their contributions as field experts providing insight and interpretation.

References

1. One-Third of All U.S. Coronavirus Deaths Are Nursing Home Residents or Workers. Available at: https://www.nytimes.com/interactive/2020/05/09/us/coronavirus-cases-nursing-homes-us.html. Accessed May 19, 2020.
2. Barnett ML, Grabowski DC. Nursing homes are ground zero for COVID-19 pandemic. JAMA Health Forum 2020;1:e200369.
3. Grabowski DC, Mor V. Nursing home care in crisis in the wake of COVID-19. JAMA 2020;324:23–24.
4. 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;382:2005–2011.
5. Arons MM, Hatfield KM, Reddy SC, et al. Presymptomatic SARS-CoV-2 Infections and Transmission in a Skilled Nursing Facility. N Engl J Med 2020;382:2005–2011.
6. Rozzini R. The COVID grim reaper. J Am Med Dir Assoc 2020;21:994.
7. Goldberg SA, Pu CT, Thompson RW, et al. Asymptomatic spread of COVID-19 in 97 patients at a skilled nursing facility. J Am Med Dir Assoc 2020;21:980–981.
8. Pillemer K, Subramanian L, Hupert N. The importance of long-term care populations in models of COVID-19. JAMA 2020;324:25–26.
9. Chen T, Guestrin C. Xgboost: A scalable tree boosting system, Proc 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016. p. 785–794.
10. Unruh MA, Yun H, Zhang Y, et al. Nursing home characteristics associated with COVID-19 deaths in Connecticut, New Jersey, and New York. J Am Med Dir Assoc 2020;21:1001–1003.
11. White EM, Kosar CM, Feifer RA, et al. Variation in SARS-CoV-2 prevalence in US skilled nursing facilities. J Am Geriatr Soc 2020:1–7.
12. He M, Li Y, Fang F. Is there a Link between nursing home reported quality and COVID-19 cases? Evidence from California Skilled Nursing Facilities. J Am Med Dir Assoc 2020;21:905–908.
13. Testimony of R. Tamara Konetzka - Caring for Seniors Amid the COVID-19 Crisis United States Senate Special Committee on Aging. Available at: https://www.aging.senate.gov/hearings/caring-for-seniors-amid-the-covid-19-crisis. Accessed June 16, 2020.
14. Special Grand Rounds. COVID-19 in Nursing Homes: Pragmatic Research Responses to the Crisis. Available at: https://impactcollaboratory.org/special-grand-rounds-covid-19-in-nursing-homes-pragmatic-research-responses-to-the-crisis/. Accessed May 19, 2020.
15. Abrams HR, Loomer L, Gandhi A, et al. Characteristics of US nursing homes with COVID-19 Cases. J Am Geriatr Soc 2020;68:1653–1656.
16. Li Y, Temkin-Greener H, Gao S. COVID-19 infections and deaths among Connecticut nursing home residents: facility correlates. J Am Geriatr Soc 2020;68.1899–190.
17. Castle NG. Nurse Aides' ratings of the resident safety culture in nursing homes. Int Jr Qual Health Care 2006;18:370–376.
18. Mor V, Zinn J, Angelelli J, et al. Driven to tiers: Socioeconomic and racial disparities in the quality of nursing home care. Milbank Q 2004;82:227–256.
19. Li Y, Cen X, Cai X, et al. Perceived patient safety culture in nursing homes associated with "nursing home compare" performance indicators. Med Care 2019;57:641–647.
20. Cohen CC, Engberg J, Herzig CT, et al. Nursing homes in states with infection control training or infection reporting have reduced infection control deficiency citations. Infect Control Hosp Epidemiol 2015;36:1475–1476.
21. Tai DBG, Shah A, Doubeni CA, et al. The disproportionate impact of COVID-19 on racial and ethnic minorities in the United States. Clin Infect Dis 2020:ciaa815.
22. Yancy CW. COVID-19 and African Americans. JAMA 2020;23:1891–1892.
23. Holtgrave DR, Barranco MA, Tesoriero JM, et al. Assessing racial and ethnic disparities using a COVID-19 outcomes continuum for New York State. Ann Epidemiol 2020;48:9–14.
24. Hooper MW, Nápoles AM, Pérez-Stabl EJ. COVID-19 and racial/ethnic disparities. JAMA 2020;323:2466–2467.
25. Price-Haywood EG, Burton J, Fort D, et al. Hospitalization and mortality among black patients and white patients with Covid-19. N Engl J Med 2020;382:2534–2543.
26. Shippee TP, Akosionu O, Ng W, et al. COVID-19 pandemic: Exacerbating racial/ethnic disparities in long-term services and supports. J Aging Soc Pol 2020;32:323–333.
27. Campbell LJ, Cai X, Gao S, et al. Racial/ethnic disparities in nursing home quality of life deficiencies, 2001 to 2011. Gerontol Geriatr Med 2016;2:1–9.
28. Zuckerman RB, Wu S, Chen LM, et al. The five-star skilled nursing facility rating system and care of disadvantaged populations. J Am Geriatr Soc 2019;67:108–114.
29. Li Y, Harrington C, Mukamel DB, et al. Nurse staffing hours at nursing homes with high concentrations of minority residents, 2001–11. Health Affairs 2015;34:2129–2137.
30. Gandhi M, Yokoe DS, Havlir DV. Asymptomatic transmission, the Achilles' heel of current strategies to control Covid-19. N Engl J Med 2020;382:2158–2160.
31. Reducing COVID-19 Deaths In Nursing Homes: Call To Action. Available at: https://www.healthaffairs.org/do/10.1377/hblog20200522.474405/full/. Accessed June 16, 2020.
32. Van Houtven CH, DePasquale N, Coe NB. Essential long-term care workers commonly hold second jobs and double-or triple-duty caregiving roles. J Am Geriatr Soc 2020;68:1657–1660.
33. Key Strategies to Prepare for COVID-19 in Long-Term Care Facilities (LTCFs). Available at: https://www.cdc.gov/coronavirus/2019-ncov/hcp/long-term-care-strategies.html. Accessed May 19, 2020.
34. As Deaths Mount, Coronavirus Testing Remains Wildly Inconsistent In Long-Term Care. Available at: https://khn.org/news/as-deaths-mount-coronavirus-testing-remains-wildly-inconsistent-in-long-term-care/. Accessed May 19, 2020.
35. Quigley DD, Dick A, Agarwal M, et al. COVID-19 Preparedness in Nursing Homes in the Midst of the Pandemic. J Am Geriatr Soc 2020;68:1164–1166.
36. State Reporting of Cases and Deaths Due to COVID-19 in Long-Term Care Facilities. Available at: https://www.kff.org/coronavirus-covid-19/issue-brief/state-reporting-of-cases-and-deaths-due-to-covid-19-in-long-term-care-facilities/. Accessed May 19, 2020.
37. As some states race for mass testing in nursing homes, others lag behind. Available at: https://abcnews.go.com/Health/states-race-mass-testing-nursing-homes-lag/story?id=70454739. Accessed May 19, 2020.
38. White House goal on testing nursing homes unmet. Available at: https://apnews.com/681479be1d9f1a0fea4d64b1a44d5b9e. Accessed June 16, 2020.
39. Testing Nursing Home Workers Can Help Stop Coronavirus. But Who Should Pay? Available at: https://www.nytimes.com/2020/06/09/health/testing-coronavirus-nursing-homes-workers.html. Accessed June 16, 2020.
40. Trabucchi M, De Leo D. Nursing homes or besieged castles: COVID-19 in northern Italy. Lancet Psychiatry 2020;7:387–388.
41. Baggett TP, Keyes H, Sporn N, et al. Prevalence of SARS-CoV-2 infection in residents of a large homeless shelter in Boston. JAMA 2020;323:2191–2192.

Supplementary Appendix
Supplemental Methods

Data Sources and Model Inputs

The data used in the study, and to train the machine-learning model, was a unique self-constructed dataset, describing 15,300 federally certified NHs and their surrounding community across the United States. This dataset integrated information collected from public sources as well as provided by industry sources. Facility information from each source was merged based on keys generated from the facility's name, address, and, if available, the CMS certification number.

Outcomes

Data pertaining to NH outcomes, describing if there was at least 1 resident COVID-19 infection at an NH, was collected from public State Department of Health NH and LTCF COVID-19 reports.[1–5] The states included in the study (Massachusetts, Georgia, New Jersey, and California), were selected based on the availability of facility level NH infection data (as a limited number of states were reporting during the study period), number of daily tests per capita,[6] number of reporting NHs, and granularity of reporting COVID-19 related data (ie, if information on the NH's census, resident deaths, and staff outcomes were reported, in addition to resident infections). This selection criteria were used to help mitigate the impact of reporting bias stemming from variations in each state's and NH's willingness and ability to both test for infection and report outcomes. Such variations and limitations in testing and reporting behaviors can result in underestimated numbers of infected NHs, artificially lowering measured NH infection risk.

Predictors

The data included in our dataset pertaining to NH characteristics were from Muller Consulting and Data Analytics (MCDA),[7] OMEGA Healthcare Investors,[8] the National Investment Center for Seniors Housing and Care (NIC),[9] Walk Score,[10] Centers for Disease Control and Prevention,[11] Long-term Care Focus (LTCFocus),[12] US Census,[13] and CMS.[14] Data from CMS includes, Nursing Home Compare data, which pulls from the CMS Provider Certification database, Health Inspection database, Payroll-Based Journal system, the Minimum Data Set national database, and Medicare claims data. Data on the NH's surrounding community, (ie, the facility's county and zip code area), was collected from the New York Times' COVID Tracker[15] and provided by Claritas Inc.[16] and SafeGraph.[17] The MCDA features describing NH characteristics were developed using CMS claims and staffing data as of 2019 Q3, as well as provider certification, health inspections, and cost report data as of March 2020. The LTCFocus features describe NH characteristics as of April 2017.

The specific variables used in the machine-learning model are shown in Supplementary Table 1. These predictors were selected based on a priori knowledge and suggestions from field experts. Predictors with missing variable values (Supplementary Table 1) were imputed using a k-nearest neighbors approach based on all predictor variables of the NHs using the Python package fancyimpute.[18] NHs with missing data on all of the predictor variables were excluded from the study. A total of 1146 of 1242 (92.3%) and 1021 of 1254 (81.4%) of reporting NHs were included in the training and testing sets, respectively.

Model Development and Feature Importance Evaluation

Model development

We used a binary classification gradient boosting model to predict the probability of the presence (binary outcome of 1) of at least 1 positive resident COVID-19 case in the NH. A binary metric was used as a single confirmed COVID-19 case is likely indicative of a compromised facility because of rapid intrafacility transmission and the presence of undetected asymptomatic and presymptomatic individuals.[19,20]

Following feature selection, the model's hyperparameter and out-of-sample performance were tuned and calculated, respectively, through 10-fold CV. Specifically, we first divided the whole training set into 10 stratified folds, creating 10 different splits where one fold was selected as the CV test set and the 9 others as the CV training set. Then, we applied an internal 10-fold CV, with no repetition, only on the formed CV training set, while restricting the model inputs to only be the identified stable predictors that were selected via feature selection methods described in the main document. For the internal 10-fold CV, the model's hyper-parameters were optimized via the Python sklearn cross-validated grid search function[21] with the area under the receiver operating characteristic curve (AUC) as the performance metric to maximize. After hyper-parameter tuning, the optimal threshold value, which categorizes predicted probabilities as binary outcomes (a probability above the threshold indicates the predicted presence of at least one positive resident COVID-19 case), was selected as the value that minimizes the difference between the model's in-sample sensitivity and specificity. This process was repeated for each of the 10 training-testing set splits from the external folds. The mean out-of-sample AUC, sensitivity (true positive rate) and specificity (true negative rate), and their 95% CIs were calculated over the external 10-fold CV testing folds. Logistic regression and neural network models were also developed and assessed using the same methods and identified stable predictors, serving as benchmark models for comparison.

Feature importance

Following model hyperparameter tuning, the model was fit over the entire training dataset to evaluate the feature importance of the identified stable predictors of risk. Specifically, we first used the Python SHapley Additive exPlanations package[22] to calculate and visualize the impact of the predictors on infection risk (Figure 1). We then assessed the relationship between each identified stable predictor and estimated infection risk levels by generating risk levels for each of the 15,300 NHs in our dataset, using the trained model, while varying the value of the specific predictor variable of interest. For each NH, the predictors other than the predictor of interest being varied retained their actual values (the NH's county's infection rate used was the rate reported on April 20, 2020). The median, 25th and 75th percentiles, and 5th and 95th percentiles of the NH risk levels across the 15,300 NHs, while varying the predictor values, were calculated.

Prospective Out-of-Sample Model Validation

For prospective validation, we fit the model over the entire training dataset (composed of NH outcomes from Massachusetts, Georgia, and New Jersey reported on April 20, 2020), using only the identified stable predictors, and generated new predictions with predictor variables from May 4, 2020. The predicted probabilities were converted to binary outcomes using an optimal threshold value calculated, as previously described, over the entire training dataset from April 20.

The predictions were prospectively compared with outcomes reported a week later on May 11, 2020, in 2 ways. We first calculated the AUC, sensitivity, and specificity of the predictions against outcomes from 1021 NHs in California. We then compared the predictions to COVID-19 case and death rates from 7660 LTCFs across the United States. Specifically, we calculated the Pearson correlation coefficient between each state's median NH risk score, across all the state's NHs in our dataset, and each state's LTCF related COVID-19 infection and death rates. The benchmark logistic regression and neural network models were also validated as described for comparison to the machine-learning model.

The LTCF related infection and death rates used for validation were calculated by dividing the sum of the reported number of LTCF related outcomes by the sum of bed counts from the 15,300 federally certified NHs in our dataset, by state. While normalizing LTCF related outcomes by NH beds increases calculated case and death rates compared to normalizing using LTCF bed counts, given the available data, we believe using NH beds results in a robust measure for validation for the following 2 reasons. First, to the best of our knowledge, there is no comprehensive LTCF database that contains information on assisted-living and independent-living facilities across all states. Including incomplete assisted-living and independent-living data per state can lead to overestimates of case and death rates for states with missing data. In other words, states with missing bed counts data would have artificially increased outcome rates compared with states without missing bed count data, as the incomplete bed count denominator would be smaller than complete bed counts. In contrast, the federal NH data we used is regulated and consistently reported across all US states, providing an accurate lower bound of the number of LTCF beds per state. Second, the reported cases and deaths are likely underestimates of the actual COVID-19 related outcomes in LTCFs, counteracting potential overestimation when normalizing by NH beds.

Supplemental Results

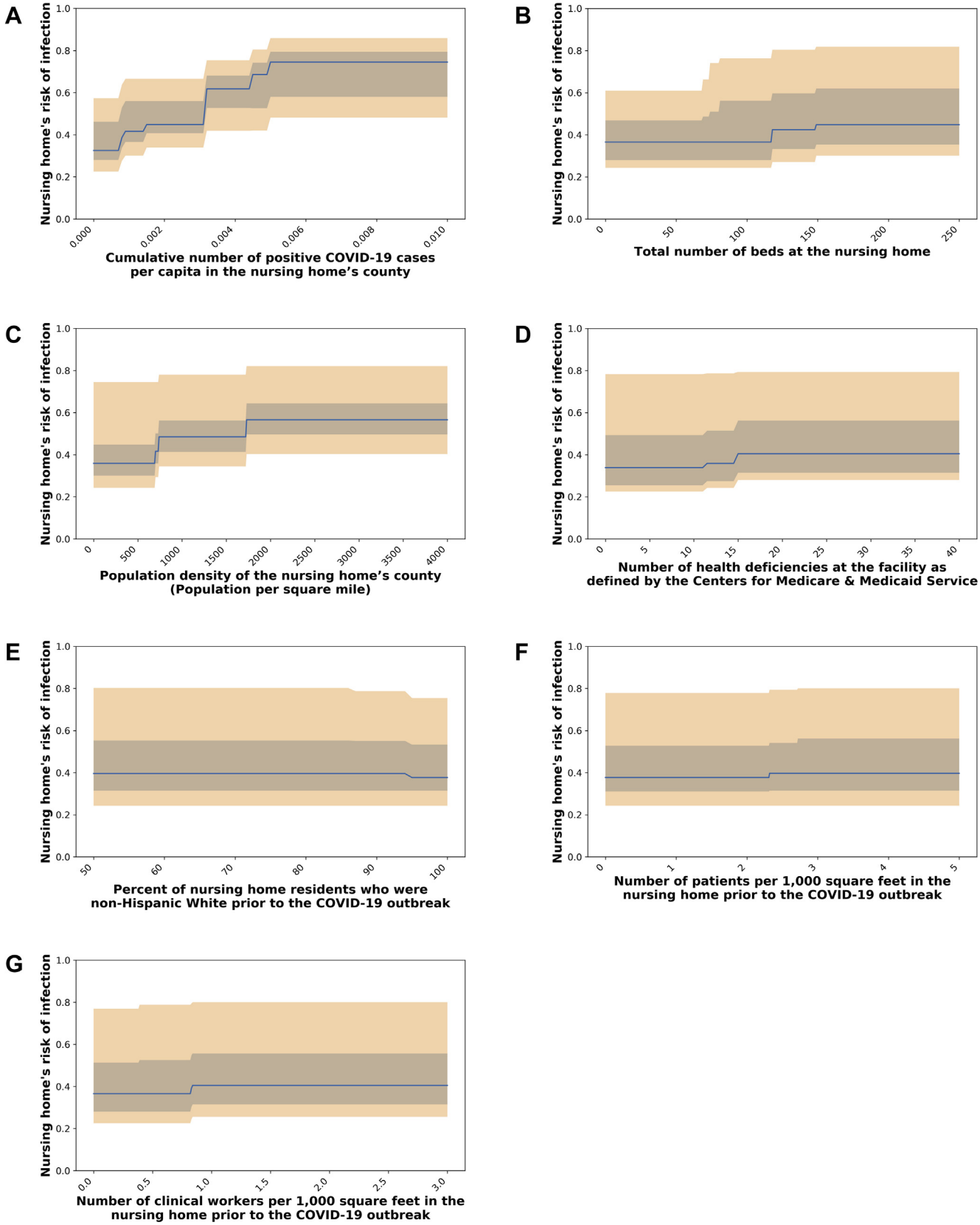
Impact of Identified Predictors on Infection Risk

The estimated change in NH infection risk when varying the values of each of the 6 identified key predictors is shown in Supplementary Figure 1. The NH's county's infection rate and size had the largest impact on infection risk.

Benchmark Logistic Regression Model

The odds ratios of the logistic regression model based on the full training dataset are shown in Supplementary Table 2. The relationships identified using the logistic regression model should be interpreted with caution because of its poor predictive performance (Table 3), namely, the overestimation of the number of infected nursing homes in the validation set and generating risk indices with poor correlation to state level LTCF infection and death rates.

The relationships between the predictor variables and COVID-19 infection found using logistic regression were not all aligned with those identified by the gradient boosting model (Figure 1). Consistent with the gradient boosting model, the logistic regression model found that the NH's county's infection rate, number of units, CMS-cited health deficiencies, and clinical worker density were significantly associated with COVID-19 infection. However, both the population density of the facility's county (population per square mile) and percent of nursing home residents who were non-Hispanic white variables were not significant predictors in the logistic regression model, despite being the third and fifth most impactful predictors, respectively, according to the machine learning gradient boosting model (Figure 1). The logistic regression model also did not find a significant association between the NH's patient density and COVID-19 infection.



Supplementary Fig. 1. Predictive feature's impact , shown in subfigures (A–G), on estimated NH risk of COVID-19 infection. The median (*blue line*), 25th and 75th percentiles (*gray band*), and 5th and 95th percentiles (*orange band*) of the infection risk levels generated by the trained model are shown across 15,300 NHs in the United States.

Supplementary Table 1
Overview of the Data Inputs of the NH Risk Model

Data Category	Variable Description	Data Source	Number of NH in Training Set without Missing Feature Values (n = 1146), n (%)
Facility's community characteristics	Cumulative number of positive COVID-19 cases per capita in the NH's county on the day of NH COVID-19 case reporting	NYT COVID Tracker	1133 (98.9)
	Estimated poverty score of the NH's county (Based on: Household income)	CDC	1133 (98.9)
	Overall comorbidity score of the NH's county (Based on: Obesity, diabetes, hypertension, cardiovascular characteristics from 2018)	CDC	1133 (98.9)
	Overall percentile ranking of social vulnerability index of the NH's county (Based on: Socioeconomic, household composition, minority status/language, and housing type/transportation characteristics)	CDC	1133 (98.9)
	Percent of facility's county who are non-Hispanic White	US Census	1133 (98.9)
	Percentage of family households in the NH's county	Claritas	1133 (98.9)
	Population density of the NH's county (population per square mile)	Claritas	1133 (98.9)
Community social distancing and population mobility characteristics	Proportion of Safe Graph tracked devices traveling less than 8000 meters per day out of all tracked devices in the facility's Zip code of the week of NH COVID-19 case reporting	Safe Graph	1096 (95.6)
	Proportion of Safe Graph tracked devices traveling more than 50,000 meters per day out of all tracked devices in the facility's zip code of the week of NH COVID-19 case reporting	Safe Graph	1096 (95.6)
	Proportion of Safe Graph tracked devices exhibiting full-time employment behavior in the zip code of the week of NH COVID-19 case reporting	Safe Graph	1101 (96.1)
	Proportion of Safe Graph tracked devices traveling less than 8000 meters per day out of all tracked devices in the facility's county of the week of NH COVID-19 case reporting	Safe Graph	1133 (98.9)
	Proportion of Safe Graph tracked devices traveling more than 50,000 meters per day out of all tracked devices in the facility's county of the week of NH COVID-19 case reporting	Safe Graph	1133 (98.9)
	Proportion of Safe Graph tracked devices exhibiting full-time employment behavior in the county of the week of NH COVID-19 case reporting	Safe Graph	1133 (98.9)
	Inflow of Safe Graph tracked devices to the facility's county of the week of NH COVID-19 case reporting	Safe Graph	1131 (98.7)
	Outflow of Safe Graph tracked devices from the facility's county of the week of NH COVID-19 case reporting	Safe Graph	1132 (98.8)
	Percentage of county's population taking public transportation to work	Claritas	1133 (98.9)

(*continued on next page*)

Supplementary Table 1 (*continued*)

Data Category	Variable Description	Data Source	Number of NH in Training Set without Missing Feature Values (n = 1146), n (%)
Facility characteristics	Age of the nursing home in years	NIC	750 (65.4)
	Standardized hourly cost per clinical worker excluding certified nursing assistants (per patient per day)*,†	MCDA	1090 (95.1)
	Standardized hourly cost per clinical worker including certified nursing assistant (per patient per day)*,†	MCDA	1090 (95.1)
	Walk Score measures walkability on a scale from 0–100 based on walking routes to destinations such as grocery stores, schools, parks, restaurants, and retail	Walk Score	760 (66.3)
	Number of clinical workers per 1000 square feet in the facility prior to the COVID-19 outbreak*	MCDA	999 (87.2)
	Number of health deficiencies at the facility as defined by the Centers for Medicare and Medicaid Service since 2014	CMS	977 (85.3)
	Number of patients per 1000 square feet in the facility prior to the COVID-19 outbreak	MCDA	999 (87.2)
	Overall infection control process and performance index	MCDA	974 (85.0)
	Rate of influenza vaccination for long stay residents	MCDA	1065 (92.9)
	Rate of influenza vaccination for short stay residents	MCDA	1078 (94.1)
	Rate of pneumococcal vaccination for long stay residents	MCDA	1065 (92.9)
	Rate of pneumococcal vaccination for short stay residents	MCDA	1080 (94.2)
	Rate of rehospitalizations of residents due to infection	MCDA	1091 (95.2)
	Index based on CMS health inspection citations related to infection control measures (citations weighted according to scope, severity, and recency)	MCDA	849 (74.1)
	Index based on CMS health inspection citations related to laboratory processes (citations weighted according to scope, severity, and recency)	MCDA	739 (64.5)
	Index based on CMS health inspection citations related to managerial processes (citations weighted according to scope, severity, and recency)	MCDA	966 (84.3)
	Index based on CMS health inspection citations related to physical environment (citations weighted according to scope, severity, and recency)	MCDA	739 (64.5)
	Total number of beds at the facility	CMS, NIC	1146 (100.0)
	Total number of beds for Nursing Care	NIC	768 (67.0)
	Total number of units for assisted living	NIC	768 (67.0)
	Total number of units for independent living	NIC	768 (67.0)
	Total number of units for memory care	NIC	768 (67.0)
	Percent of facility residents who were non-Hispanic white prior to the COVID-19 outbreak	LTCFocus	961 (83.9)
	Percent of facility residents whose primary support was Medicare prior to the COVID-19 outbreak	LTCFocus	961 (83.9)
	Percent of facility residents whose primary support was Medicaid prior to the COVID-19 outbreak	LTCFocus	961 (83.9)
Facility outcomes	Presence of at least one resident COVID-19 case	State Departments of health	1146 (100.0)

CDC, Centers for Disease Control and Prevention; LTCFocus, Long-term Care Focus; MCDA, Muller Consulting and Data Analytics; NIC, National Investment Center for Seniors Housing and Care; NYT, New York Times.

*Clinical worker defined as registered nurses, licensed practical nurses, certified nursing assistants, nursing aides, medical aides/technicians, nursing home administrators, medical directors, physicians, physician assistants, nurse practitioners, clinical nurse specialists, pharmacists, dieticians, feeding assistants, occupational therapists, occupational therapy assistants, occupational therapy aides, physical therapists, physical therapist assistants, physical therapist aides, respiratory therapists, respiratory therapy technicians, speech/language pathologists, therapeutic recreation specialists, qualified activities professionals, other activities staff, qualified social workers, other social workers, mental health service workers.

†Standardized costs based on average hours worked multiplied by Bureau of Labor Statistics national wage rate estimates for respective occupations in skilled nursing facilities.

Supplementary Table 2
Odds Ratios for Benchmark Multivariate Logistic Regression Model Based on the Training Set Data (n = 1146)

Variables	Odds Ratio (95% CI)	*P* Values
Cumulative number of positive COVID-19 cases per capita in the facility's county on the day of NH COVID-19 case reporting (confirmed cases per 100,000 people)	75.7×10^{90} (57.8×10^{67}–99.0×10^{113})*	<.001
Total number of beds at the facility	1.003 (1.001–1.005)*	<.001
Population density of the facility's county (population per square mile)	1.0000 (0.9999–1.0001)	.782
Number of health deficiencies at the facility as defined by the Centers for Medicare and Medicaid Services	1.029 (1.015–1.044)*	<.001
Percent of nursing home residents who were non-Hispanic white prior to the COVID-19 outbreak	0.997 (0.991–1.004)	.456
Number of patients per 1000 square feet in the facility prior to the COVID-19 outbreak	0.844 (0.653–1.091)	.195
Number of clinical workers per 1000 square feet in the facility prior to the COVID-19 outbreak	2.528 (1.185–5.390)*	<.05
Intercept	0.204 (0.096–0.435)*	<.001

*$P < .05$.

References

1. Georgia Department of Community Health. The Georgia Department of Community Health. Available at: https://dch.georgia.gov/. Accessed May 19, 2020.
2. New Jersey COVID-19 Information Hub. New Jersey Department of Public Health. Available at: https://covid19.nj.gov/#live-updates. Accessed May 19, 2020.
3. Archive of COVID-19 cases in Massachusetts | Mass.gov. Massachusetts Department of Public Health. Available at: https://www.mass.gov/info-details/archive-of-covid-19-cases-in-massachusetts. Accessed May 19, 2020.
4. SNFsCOVID_19. California Department of Public Health. Available at: https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/SNFsCOVID_19.aspx. Accessed May 19, 2020.
5. One-Third of All U.S. Coronavirus Deaths Are Nursing Home Residents or Workers. New York Times. Available at: https://www.nytimes.com/interactive/2020/05/09/us/coronavirus-cases-nursing-homes-us.html. Accessed May 19, 2020.
6. Coronavirus Testing Needs to Triple Before the U.S. Can Reopen, Experts Say. New York Times. Available at: https://www.nytimes.com/interactive/2020/04/17/us/coronavirus-testing-states.html. Accessed May 19, 2020.
7. Muller Consulting and Data Analytics. Available at: http://www.mcdaintel.com/. Accessed May 19, 2020.
8. Omega Healthcare Investors, Inc. Available at: http://www.omegahealthcare.com/. Accessed May 19, 2020.
9. Senior Housing Investment | NIC. h. Available at: ttps://www.nic.org/. Accessed May 19, 2020.
10. Walk Score. Available at: https://www.walkscore.com/. Accessed May 19, 2020.
11. Atlas of Heart Disease and Stroke. Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/dhdsp/maps/atlas/index.htm. Accessed May 19, 2020.
12. Long-Term Care: Facts on Care in the US. Brown University - Long-term Care Focus. Available at: http://ltcfocus.org/. Accessed July 29, 2020.
13. County Population by Characteristics: 2010-2019. United States Census Bureau. Available at: https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html. Accessed July 29, 2020.
14. CMS Homepage | CMS. Centers for Medicare and Medicaid Services. Available at: https://www.cms.gov/. Accessed May 19, 2020.
15. nytimes/covid-19-data: An ongoing repository of data on coronavirus cases and deaths in the U.S. New York Times. Available at: https://github.com/nytimes/covid-19-data. Accessed May 19, 2020.
16. Claritas | Custom Targeting and Audience Segments. Available at: https://www.claritas.com/. Accessed May 19, 2020.
17. SafeGraph | POI Data, Business Listings, and Foot-Traffic Data. Available at: https://www.safegraph.com/. Accessed May 19, 2020.
18. fancyimpute · PyPI. PyPI. Available at: https://pypi.org/project/fancyimpute/. Accessed May 19, 2020.
19. Arons MM, Hatfield KM, Reddy SC, et al. Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. N Engl J Med 2020;382:2081–2090.
20. 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;382:2005–2011.
21. sklearn.model_selection.GridSearchCV — scikit-learn 0.23.1 documentation. Sklearn. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. Accessed May 19, 2020.
22. Cohen SB, Ruppin E, Dror G. Feature Selection Based on the Shapley Value. Paper presented at IJCAI 2005.