Open Access Full Text Article
RESEARCH LETTER

Identification and Comparison of Patient Characteristics for Those Hospitalized with COVID-19 versus Influenza Using Machine Learning in a Commercially Insured US Population

Xiaoxue Chen[1]
Zhi Wang[1]
Samantha G Bromfield[1]
Andrea DeVries[1]
David Pryor[2]
Vincent Willey[3]

[1]Anthem Translational Research, HealthCore Inc., Wilmington, DE, USA; [2]Clinical Strategy and Innovation, Anthem, Inc., Indianapolis, IN, USA; [3]Scientific Affairs, HealthCore Inc., Wilmington, DE, USA

Correspondence: Xiaoxue Chen
HealthCore, Inc., 123 Justison Street,
Suite 200, Wilmington, DE, 19801, USA
Tel +1 302-230-2090
Fax +1 302-230-2020
Email xchen@healthcore.com

Background

The novel severe acute respiratory syndrome coronavirus 2, the virus that causes coronavirus disease 2019 (COVID–19), continues to spread in the US through the 2020–2021 influenza season and beyond. Approaches to identify those most at risk for poor outcomes for the two viral infections are needed for future planning. As influenza is a well-known respiratory disease sharing some similarities to COVID-19, such comparison will aid physicians and health systems to predict disease trajectory and allocate health resources most efficiently.

A retrospective cohort study using a French national administrative database found that patients hospitalized with COVID-19 were more frequently obese or overweight, diabetic, and hypertensive.[1] Patients hospitalized with influenza more frequently had heart failure, chronic respiratory disease, and cirrhosis.[1] Similar observations were reported in an international network study that included US, South Korea, and Spain.[2] While this information provides useful context to the current understanding of characteristics of patients hospitalized with COVID-19 in several countries, understanding of the overall risk profile for the two viral infections is lacking in a broad US population.

Advanced modelling, machine learning, and artificial intelligence (AI) techniques have been employed to detect, diagnose, evaluate, and prioritize treatment for COVID-19. Examples include laboratory examination frameworks to prioritize patients with COVID-19, AI techniques in the detection and classification of COVID-19 medical images, and models to predict the spread of disease. An increasing number of severe COVID-19 outcome risk assessment studies have found that demographic factors, comorbidities, radiographic findings, and laboratory markers may individually or collectively predict worse outcomes.[3] To deepen the understanding of COVID-19, additional knowledge of the interplay between patient demographic characteristics, socioeconomic status, and medical history as well as a comparison with influenza is needed.

Therefore, the aim of this study was to comprehensively compare the demographic, socioeconomic and clinical characteristics of patients hospitalized with COVID-19 versus influenza using machine learning techniques within a large, geographically diverse US commercially insured population.

Methods

This retrospective study analyzed administrative claims and prior authorization data from a large commercially insured population drawn from employer-based insurance, individual insurance, and Medicare Advantage. This study was conducted in full compliance with relevant provisions of the Health Insurance Portability and Accountability Act. Only deidentified data were used, and the study was determined to be exempt from review by the WCG institutional review board.

Data Statement

The dataset was derived from the HealthCore Integrated Research Environment (HIRE). The HIRE contains medical and pharmacy claims from 14 commercial health plans geographically dispersed across the United States. We do not have permission to grant public access to the dataset.

Study Design and Patients

Patients hospitalized with COVID-19 were identified between 02/01/20 and 06/30/20 using health plan prior authorization data, and patients hospitalized with influenza were identified during the previous influenza season (09/01/18–05/30/19) using claims data. Patients were required to have at least six months continuous medical and pharmacy coverage prior to the admission date.

Patients' demographic characteristics, area-level socioeconomic status, medical conditions and medication history were observed for a minimum of six months prior to and on the admission date (with the exception of the last 4 months for recent medication exposure). Demographic characteristics included age, sex, region, payer type, and zip code level urban/rural geographic classification. Area-level socioeconomic status was estimated at block group level from the American Community Survey (ACS) and included household crowding, percentage below the federal poverty line, median household income, and socioeconomic index. We first included conditions that were previously associated with increased risks of COVID-19 severe outcomes. To identify additional patient characteristics that are associated with hospitalization with COVID-19 or influenza, patient diagnosis and procedures from the claim history were clustered into clinically meaningful categories from the International Classification of Diseases, 10th Revision, codes by using the Healthcare Cost and Utilization Project's Clinical Classification Software (CCS). Patient medication history was collapsed at the Generic Product Identifier (GPI) 4-digit level and patients' overall medication usage was captured using GPI 8 medication count. We constructed a total of 1394 covariates in the initial stage. All-zero and near-zero covariates, and highly correlated covariate pairs were removed before modelling resulting in 499 covariates included in the model.

Statistical Analyses

We selected the approach of gradient boosting models (GBM) to evaluate the likelihood of COVID-19 hospitalization compared to influenza hospitalization, a classification problem. GBM is widely known as one of the best performing machine learning algorithms for classification. GBM provides robust prediction results through an iterative learning procedure that consecutively fits new trees to the residual of the trees that preceded it. The data were split into a training dataset and testing dataset in a 75% versus 25% ratio. We conducted 10-fold cross validation in the training dataset to tune the parameter and address potential overfitting. Model performance was assessed using area under the curve (AUC). For model interpretation, variable importance and marginal effect estimates from partial dependence plots (PDP) were provided for the 20 most influential predictors.[4,5] All statistical analyses were performed using R version 3.63 and H2O.

Results

A total of 14,373 individuals hospitalized with COVID-19 and 8698 hospitalized with influenza were included in the study. Mean age was 59 years (SD=17.6) for the COVID-19 group and 56 years (SD=25.0) for the influenza group. Males represented 72.6% and 64.4% of the COVID-19 and influenza group, respectively. Patients hospitalized with COVID–19 were more widely distributed across age categories relative to those hospitalized with influenza, which were concentrated among the very old and the very young. The GBM yielded acceptable accuracy (AUC=0.77; 95% CI, 0.75–0.79). The twenty most influential predictors are displayed in Figure 1 and their marginal effect estimates are in Table 1. All age groups above 19 years old showed an increase in the predicted likelihood of hospitalization with COVID-19, with those 45 to 64 years of age displaying the greatest increase (10%). Male sex was associated with a 5% increase in the predicted likelihood of hospitalization with COVID-19. Socioeconomic factors such as poverty and household

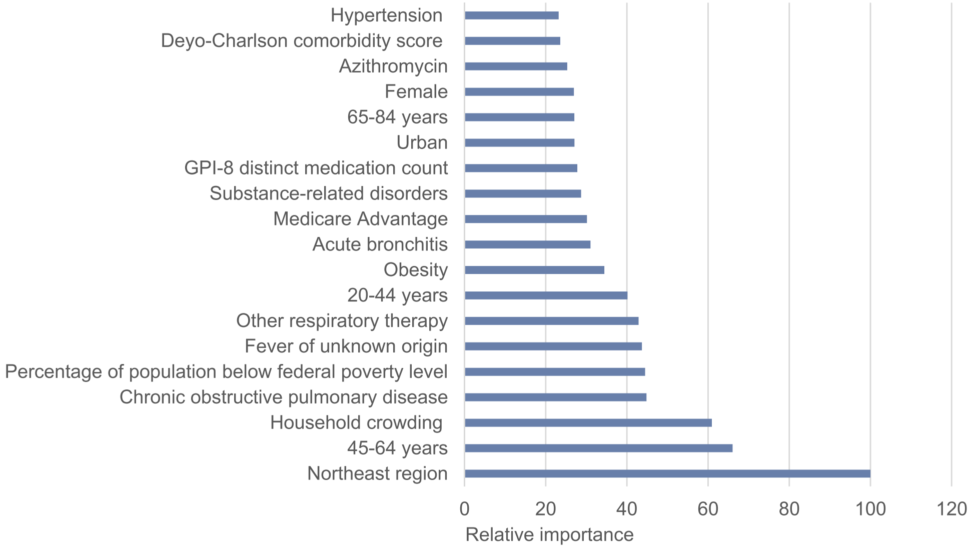

Figure 1 Importance ranking of the twenty most influential predictors from the gradient boosting model.

Notes: Relative importance values range from 0 to 100 and represent the proportional contribution of a covariate in predicting the likelihood of hospitalization with COVID–19 as opposed to hospitalization with influenza. Household crowding and percentage of population below federal poverty level were estimated from the American Community Survey. The reference group for age is 0–19 years; the reference for urban/rural classification is rural; the reference group for region is Midwest.

Abbreviation: GPI, generic product identifier.

crowding were positively associated with an increased predicted likelihood of COVID-19 hospitalization, as was living in an urban environment. Metabolic conditions such as obesity and hypertension were associated with a 5% and 2% increase in the predicted likelihood of hospitalization with COVID-19. Respiratory conditions such as COPD and acute bronchitis were associated with an 8% and 7% increase in the predicted likelihood of hospitalization with influenza, respectively. Medication use (as indicated by GPI-8 distinct medication count) was less common in patients hospitalized with COVID-19. Azithromycin medication use was associated with a 6% increase in the predicted likelihood of COVID-19-related admissions, while glucocorticosteroid use was associated with a 4% increase in the predicted likelihood of influenza-related admissions.

Discussion

In this study, we performed machine learning to identify patient characteristics associated with hospital admission with COVID-19 versus influenza using demographic, socio-economic and clinical characteristics. The study highlighted a few important findings with implications in clinical practice and care delivery.

Concordant with previous findings,[1,2] we found patients hospitalized with COVID-19 were more likely to be male, have higher rates of obesity and hypertension and lower rates of COPD compared to patients hospitalized with influenza. The age distribution of COVID-19 and influenza groups were found to be similar to Piroth et al's age distributions since both studies included children and young adolescents.

Our study adds significantly to the literature by quantifying the association between individual patient characteristics and COVID-19 hospitalization in the context of influenza hospitalization in a broad US population. Age was the strongest risk factor for COVID-19 hospitalization besides Northeast region, with the latter being more of an artifact of the geographic origin of the COVID-19 outbreak in the US. Additionally, age was strongly associated with many comorbidities that increased the risk of COVID-19 severe outcomes; however, our model suggested the impact of age was not based solely on comorbidities. Recent data describing the pathology and molecular change in COVID-19 patients indicated immunosenescence and inflammaging are major drivers of the higher mortality rates in older patients.[6] Obesity was another major risk factor for severe COVID-19 outcomes in our study, consistent with findings from others. Obesity, known to stimulate low-grade inflammation and predispose patients to cytokine storms, was associated with lower survival in COVID-19 infections.[6] As observed in our study, metabolic conditions (eg, obesity, hypertension) exposed patients to greater risks of severe COVID-19

Table 1 Effect Estimates of the Twenty Most Influential Predictors from the Gradient Boosting Model

Covariates	Change in Predicted Probability of Hospitalization with COVID–19
Northeast region	11%
45–64 years	10%
Household crowding	Non-linear positive association (estimate:11%)
Chronic obstructive pulmonary disease	−8%
Percentage of population below federal poverty level	Non–linear positive association (estimate:6%)
Fever of unknown origin	7%
Other respiratory therapy	−8%
20–44 years	6%
Obesity	5%
Acute bronchitis	−7%
Medicare Advantage	3%
Substance-related disorders	−7%
GPI-8 distinct medication count	Non–linear negative association (estimate: −2%)
Urban	7%
65–84 years	1%
Female	−5%
Azithromycin	6%
Deyo-Charlson comorbidity score	Non-linear association (estimate:0.4%)
Hypertension	2%
Glucocorticosteroids	−4%

Notes: The marginal effect estimates of a single covariate on the predicted likelihood of hospitalization with COVID-19 were obtained using partial dependence plot. For a continuous covariate, the marginal effect estimate reported the change in predicted probability of hospitalization with COVID-19 between the lowest values of the covariate and the highest values of the covariate. A positive value indicated the increased predicted probability of hospitalization with COVID-19; while a negative value indicated the increased predicted probability of hospitalization with influenza.

outcomes as opposed to influenza; while respiratory conditions (eg, COPD, acute bronchitis) were associated with less risks to COVID-19 as opposed to influenza patients.

A significant strength of our study is the inclusion of several social constructs as a proxy for social disparities, and these socioeconomic factors were associated with a significant increase in the risk of COVID-19 hospitalization. This information was obtained at the block group level from ACS, which depicts the characteristics of a division of the census tract that contains between 600 and 3000 people, the most granular level for area-level estimates. As our study considered various categories of patient characteristics conjointly in the model, we were able to show the incremental increases in COVID-19 risk due to socioeconomic factors after controlling for comorbid conditions. Our results suggested socioeconomic factors negatively impacted COVID-19 infections in a more direct manner and not just by worsening an individual's overall health status. Contrary to our study, Piroth et al did not find such association between social deprivation score and COVID-19 hospitalization likely due to the different payment mechanism of France's national health insurance system.[1] Our study included self-reported race as well. While prior literature has reported that Blacks and Hispanics have been disproportionally affected by COVID-19,[7] our model did not identify race among the most influential factors. This may be due to several factors, including only 30% of our study cohort had race data. Additionally, the effects of socioeconomic factors associated with race such as poverty and crowding may be more influential on COVID-19 outcomes than race in itself as other studies have not often included these socioeconomic factors.[7]

Conflicting information regarding the benefits of hydroxychloroquine/chloroquine and azithromycin in COVID-19 treatment and potential concerns for drugs, such as angiotensin-converting enzyme (ACE) inhibitors and angiotensin receptor blockers (ARBs), have complicated care during the pandemic. Our study found azithromycin was more commonly used prior to COVID-19 hospitalization relative to influenza hospitalization, which suggests azithromycin may have been used in an attempt to treat patients with COVID-19 early in the course of the disease with limited benefit. Also, prior hydroxychloroquine/chloroquine use was not more commonly observed in COVID-19 hospitalization relative to influenza hospitalization. Glucocorticosteroids, recently shown to be effective in treating patients hospitalized with COVID-19, were more commonly used prior to influenza hospitalization, most likely as an anti-inflammatory treatment. While there was debate on the potential risks and benefits of ACE inhibitors and ARBs in the context of COVID-19, these medications did not appear to be influential in our study.

Our study does have several limitations. As this was a retrospective study of patients hospitalized with COVID-19 versus influenza, we studied associations rather than causation. This study used routinely collected claims data and lacked detailed lab results and other clinical information at the time of the hospital admission. Claims data may be subject to accuracy or completeness data quality issues. Our study is a snapshot of COVID-19 patients up to

June 2020, and the changing patient characteristics, evolving clinical practice, and resource constraints as the pandemic evolves may influence associated risk factors in future analyses.

Conclusion

This study used machine learning techniques to depict patient risk profiles for COVID-19 hospitalization in the context of influenza within this large, geographically broad US population. Older age, male sex, metabolic disease history, and poorer socioeconomic status were major risk factors for COVID-19 hospitalizations, while a respiratory disease history was more associated with influenza hospitalizations. These findings can assist clinicians and health systems to target their outreach and treat the most susceptible patients more efficiently. Additionally, identifying at-risk patients for COVID-19 versus influenza hospitalization will help inform prevention strategies such as targeted vaccine administration approaches.

Role of the Sponsor

The sponsor had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation of the manuscript; or the decision to submit the manuscript for publication.

Disclaimer

The interpretation and reporting of these data are the sole responsibility of the authors.

Funding

This study was funded by Anthem, Inc.

Disclosure

Chen, Wang, Bromfield, Willey and DeVries reported being a full-time employee of HealthCore. Dr. Pryor reported being a full-time employee of Anthem. No other disclosures were reported.

References

1. Piroth L, Cottenet J, Mariet A-S, et al. Comparison of the characteristics, morbidity, and mortality of COVID-19 and seasonal influenza: a nationwide, population-based retrospective cohort study. *Lancet Respir Med*. 2021;9(3):251–259. doi:10.1016/S2213-2600(20)30527-0
2. Burn E, You SC, Sena AG, et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. *Nat Commun*. 2020;11(1):5009. doi:10.1038/s41467-020-18849-z
3. Gallo Marin B, Aghagoli G, Lavine K, et al. Predictors of COVID-19 severity: a literature review. *Rev Med Virol*. 2021;31(1):e2146. doi:10.1002/rmv.2146
4. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak*. 2019;19(1):146. doi:10.1186/s12911-019-0874-0
5. Friedman JH, Meulman JJ. Multiple additive regression trees with application in epidemiology. *Stat Med*. 2003;22(9):1365–1381. doi:10.1002/sim.1501
6. Mueller AL, McNamara MS, Sinclair DA. Why does COVID-19 disproportionately affect older people? *Aging (Albany NY)*. 2020;12(10):9959–9981. doi:10.18632/aging.103344
7. Khazanchi R, Evans CT, Marcelin JR. Racism, not race, drives inequity across the COVID-19 continuum. *JAMA Network Open*. 2020;3(9):e2019933–e2019933. doi:10.1001/jamanetworkopen.2020.19933