A Multi-Factor Risk Model for Severe Covid-19,

Vaccine Prioritization and Monitoring Based on a 16 Million Medicare Cohort

AUTHORS

Bettina Experton, M.D., M.P.H., Hassan A. Tetteh, M.D., Nicole Lurie, M.D., M.S.P.H., Peter Walker, Ph.D., Colin J. Carroll, M.S., Adrien Elena Ph.D., Christopher S. Hein, Ph.D., Blake Schwendiman B.S., Christopher R. Burrow, M.D.

Author Affiliations

From the Project Salus healthcare analytics group at Humetrix (B.E., A.E, C.S.H, B.S., C.R.B), Del Mar, CA, the Department of Defense Joint Artificial Intelligence Center (JAIC) Warfighter Health Mission Team (H.A.T., P.W.), Washington, D.C., the Coalition for Epidemic Preparedness Innovation (CEPI), Oslo, and Harvard Medical School (N.L.), Boston, MA, and the Johns Hopkins University Applied Physics Laboratory (C.J.C.), Baltimore, MD.

ABSTRACT

Background:
Public Health interventions to slow the spread of the Covid-19 pandemic focus on protecting individuals at risk for severe disease who will first receive the initially available limited doses of Covid-19 vaccines. Existing risk models for severe Covid-19 and the equitable allocation of vaccines, lack needed integration of both socio-demographic and clinical risk factors.

Methods:
We present an integrated multi-factor risk model for severe Covid-19 combining demographic and clinical data extracted from de-identified Medicare claims for a cohort of 16 million Medicare beneficiaries with over 900,000 Covid-19 cases, and socio-economic data at the county and zip code level from the CDC Social Vulnerability Index. The model based on both logistic regression and random forest and its associated digital maps were developed as part of Project Salus of the Department of Defense Joint Artificial Intelligence Center, for use by military personnel in their supply logistics mission to support the national response to the Covid-19 pandemic.

Results:
The model affirms ethnicity (North American Native: OR 2.01; 95% CI 1.85 - 2.17; Black: OR 1.54; 95% CI 1.51 - 1.57; FI 0.04) as leading factors for hospitalization risk, and age over 85 with the highest association with death (OR 5.73; 95% CI 5.44 - 6.05; FI 0.24). ESRD (OR 2.45; 95% CI 2.35 - 2.56; FI 0.05), prior hospitalization (OR 1.74; 95% CI 1.72 - 1.77; FI 0.15), chronic kidney disease (FI 0.04), morbid obesity (OR 1.57; 95% CI 1.53 - 1.60), pulmonary fibrosis or pulmonary hypertension (OR 1.53; 95% CI 1.49 - 1.57), and CHF (FI 0.04) are leading clinical risk factors for Covid-19 hospitalizations. However, the model reveals low risk for COPD (OR 1.17; 95% CI 1.15 - 1.19), minimal or no risk for diabetes (hospitalization OR 1.02; 95% CI 1.01 - 1.04), and demonstrates an association of preventive behaviors with prior use preventive screenings (OR 0.64; 95% CI 0.62 - 0.66), prior influenza (OR 0.72; 95% CI 0.70 - 0.73) and pneumococcal (OR 0.84; 95% CI 0.81 - 0.87) immunizations with survival.

Conclusion:

This multi-factor risk model concurrently affirmed by different analytical approaches can be applied for use by national and local health authorities for pandemic response, optimizing prioritization of Covid-19 vaccine allocation, and for monitoring vaccine safety and efficacy.

INTRODUCTION:

Over eleven million people in the United States have been infected with SARS-CoV-2, to the date of this publication, with more than 245,000 deaths of whom 80% are over the age of 65.[1,2] Public health interventions to slow the spread of the pandemic have focused on protecting individuals most at risk for exposure to SARS-CoV-2, or at risk for severe Covid-19. The Centers for Disease Control (CDC) has identified people of color, individuals with certain comorbidities and those over the age of 65 at highest risk.[3]

This risk-based approach to pandemic response is now directing Covid-19 vaccine prioritization to allocate the first available vaccine doses to individuals most at risk. All Covid-19 vaccine prioritization frameworks reviewed by the Advisory Committee for Immunization Practices (ACIP) are risk-driven with a time-phased approach.[4] The National Academies of Science, Engineering and Medicine recommend a four-phase equitable plan determined by risk categories.[5] The "jumpstart phase 1a" allocates the first vaccine doses to front line healthcare professionals and emergency responders, followed by phase 1b for high risk individuals with comorbidities or older individuals living in congregate or overcrowded settings. In addition, the National Academies recommend special efforts to deliver vaccines to residents of high-vulnerability areas by using the CDC's Social Vulnerability Index (SVI)[6] or the Covid-19 Community Vulnerability Index (CCVI).[7]

Multiple-factor risk categorization is needed for vaccine allocation schemes to simultaneously include individual demographics, comorbidities, and socioeconomic characteristics. Integration of these various data should also consider overlap of risk categories, as the CDC highlighted during the ACIP October 2020 meeting.[8] This is important when considering that more than half of the 53 million Americans over age 65 suffer from two or more chronic conditions,[9] qualifying them for phase 1b vaccine allocation.

At a population level, both clinical and sociodemographic risk need to be considered in tandem. The current CDC listing of risk factors for severe Covid-19[10] is derived from single hospital-based studies with limited sample sizes,[11,12] or hospital reporting, both of which lack nationwide representation.[13] The Center for Medicare and Medicaid Services (CMS) issues monthly Medicare Covid-19 Data Snapshots that include national demographic characteristics and prevalence of common chronic conditions of hospitalized fee-for-service (FFS) Medicare beneficiaries, but lack more detailed diagnosis, medication, and procedure data as well as socio-economic data needed for identifying risk factors for severe Covid-19 at the individual and population levels.[14] There are no analyses that fully support operationalization of the National Academies recommendation to use both clinical and SVI or CCVI risk factor data. The CCVI provides an integrated risk model of both CDC identified risk factors for severe Covid-19 and SVI data, at the state and county, but not at the local zip code level. This and other models of this type, such as the Pandemic Vulnerability Index,[15] have further limitations. They offer population (not individual) based risk data, include limited clinical data with certain chronic conditions with their geographic distributions, and use dissociated geographic data sources (national data for comorbidities, and census tract level data for SVI socio-economic risk factors).

This study integrates multiple factors in its risk models for severe Covid-19 in the Medicare population, based on de-identified Medicare claim data from which detailed clinical and past medical history data were extracted and the CDC SVI dataset. The risk models were developed with the use of a dual analysis approach incorporating both logistic regression and machine leaning random forest analyses. The dependent and independent regression variables and Features of Importance in the machine learning analyses were in part based on our prior research on the frail elderly and Medicare population[16] and

extracted by our data analysis capabilities developed as part of our Humetrix's CMS approved iBlueButton technology platform deployed since 2011.[17] The resulting risk models from this study are both individual and population based and were used to create an interactive dashboard currently hosted on a secure government intranet that enables visualization of calculated population risk at national, county and zip code levels.

The risk models and associated mapping were developed as part of a Department of Defense (DoD) Joint Artificial Intelligence Center (JAIC) Covid-19 related data analytics project named Salus (the Roman goddess for safety and health), to provide predictive visualization tools to the National Guard and other military personnel in their specific mission to assist in resourcing Covid-19-related healthcare and other vital sector supply and personnel assets, especially serving local hospitals and health departments affected by the pandemic.[18]

Following on the National Academies recommendation for use of existing systems across all levels of government, we discuss the possible use of the Salus Medicare risk models and its derived dashboard as a tool for supporting equitable allocation, distribution of Covid-19 vaccine and vaccine monitoring, as well as for Covid-19 surveillance in the higher risk Medicare population.

METHODS:

Study design and data sources

The Salus Medicare risk models are based on an observational study of all Medicare FFS beneficiaries who since January 1, 2020, either had a Covid-19 test or diagnosis, or for any medical reason had an emergency department, urgent care, or telehealth visit, or were hospitalized. The CMS Office of Enterprise Data and Analytics provides for this cohort weekly Medicare claim outputs from the CMS Chronic Condition Warehouse with Project Salus specified outpatient and inpatient institutional, Part B professional, skilled nursing facility, and hospice claims extending back to October 1, 2019, and Part D claims starting January 1, 2020. De-identified claims are received in the secure ECS Federal Secure Unclassified Network (SUNet) provided by the JAIC for processing and analysis by Salus partner Humetrix to build risk models for severe Covid-19. In addition to the clinical and demographic data generated from Medicare claims, the models used selected socio-economic variables from the CDC SVI.

Data processing

The Humetrix Enterprise Platform with its claim and other health data analytics software provides weekly data processing of over 100 million CMS de-identified Medicare claim records in its secure JAIC SUNet enclave, including data upload into a relational database to generate Covid-19 outcome datasets for: Covid-19 confirmed cases and related hospitalizations, ICU and ventilator use, outpatient care only, and death. Medicare claim-generated clinical, demographic and outcome variables are then merged with CDC SVI variables for the residential zip codes of individuals in the cohort.

Independent variables

The independent variables included in our severe Covid-19 risk models are: beneficiary age, sex, ethnicity, insurance coverage and residential zip code, prior health care utilization (prior hospitalization(s), skilled nursing home admissions, etc.) as a measure for disease severity and frailty, the individual's comorbidities with in addition to the CMS chronic condition flags, Humetrix compiled diagnostic categories using specific ICD-10 code algorithms, medications grouped by pharmaceutical class, vaccinations before Covid-19 diagnosis, and other variables starting October 1, 2019 (see Supplemental Methods in the Appendix). Socio-economic variables (e.g., income quartile, education,

residential density, and other factors) were defined at the individual residential zip code level, after conversion from the census track based data found in the CDC Social Vulnerability Index. A list of all the variables analyzed is found in the Supplemental Appendix.

Statistical analysis, variable selection and risk model

To determine significant predictors of Covid-19 related hospitalization and death outcomes, we used logistic regressions (R statistical software, version 3.6 with rms, glmnet and pROC packages).[19-22] For these regressions, we defined the following groups of Covid-19 cases: those who received outpatient care only (defined as cases that did not require hospitalization or didn't die at least thirty days after diagnosis), hospitalizations attributed to Covid-19, and deaths (defined as cases who died of SARS-CoV-2 infection within 60 days of diagnosis).

The logistic regression models were built on 60% training and 40% validation sets obtained by random selection of Covid-19 confirmed cases. We used a 50:50 ratio of Covid-19 cases who required either hospitalization, versus those who only received outpatient care for training the hospitalization model, and a 50:50 ratio for Covid-19 survivors versus deceased cases, for training the death model. Determination of correlation coefficients between independent variables, as well as lasso regression, were used to eliminate independent variables which demonstrated significant correlations and collinearities in both models. A stepwise backward variable selection procedure based on the Akaike Information Criterion (AIC) was also used to remove non-significant variables, verifying that independent variables discarded by the lasso regression corresponded with variables identified by the stepwise backward selection procedure. Computation of the 95% confidence intervals for the coefficient estimates were generated by bootstrapping (2000 repetitions) on the training set.

Model performance was measured by the Area Under the Receiver Operating Characteristics (AUROC) curve value. The hospitalization risk model AUROC was computed on a validation set composed of cases that were not used in the training set, with adjustments to give a 60:40 ratio of outpatient to hospitalized cases (which is the observed ratio for hospitalization in the Covid-19 study population). For the death model AUROC we used an 85:15 ratio (which is the observed case fatality ratio in the Covid-19 study population).

To ascertain which variables were the most important in determining severe Covid-19 outcomes of hospitalization and death, we performed a random forest and computed Feature Importance (Python, scikit-learn version 0.22.1 with RandomForestClassifier and GridSearchCV packages). The data sampling procedure, variable definition, feature engineering, and patient outcome definitions were identical to those described above for Logistic Regression. Trees were build using bootstrapping with balanced subsamples. Parameters specifying the maximum depth of the tree and the number of estimators (trees in the forest) were optimized by cross-validated grid-search in the training set. Model performance was measured with AUROC using the validation set.

Logistic regression coefficients were used to compute individual predicted probabilities of hospitalization in the event of SARS-CoV-2 infection for the entire Salus Medicare cohort of 16 million beneficiaries. The regional percentage of the cohort population over a predicted probability of 0.55, is displayed on a digital risk map at the county and zip code levels. Additional mapping data was also produced to display Covid-19 confirmed cases and their outcomes at the national and local level, with recent, cumulative and times series displays.

RESULTS:

Study population characteristics

Socio-demographic and clinical characteristics of the 16 million study population, as of November 6, 2020, and its subsets of Covid-19 cases (n=910,360) who were either hospitalized with severe disease (n=317,189) or who only required outpatient Covid-19 care (n=388,424), or who died (n=122,613) are summarized in Table 1.

Table 1 displays comparisons between non-SARS-CoV-2 infected FFS Medicare beneficiaries and Covid-19 cases, and among them, hospitalized vs. non-hospitalized and deceased cases. Among Covid-19 cases, the severe disease groups of hospitalized and deceased patients show higher frequencies of diabetes, COPD, ESRD, hypertension, ischemic heart disease, cerebrovascular disease, chronic kidney disease, chronic lung disease, chronic liver disease, and congestive heart failure although the effect sizes for these differences are small (see Supplemental Appendix).

Predictors

We used logistic regression analysis to determine the relative importance of significant predictors of severe Covid-19 at the individual Medicare beneficiary level. The predictor variables for hospitalization are shown on Figure 1 (hospitalization model validation set AUROC = 0.65, balanced accuracy 0.62 using threshold 0.50) and the predictor variables for the death model are shown in Figure 2 (death model validation set AUROC = 0.71, balanced accuracy 0.65 using threshold 0.50). Variables excluded from the models based on the specified selection criteria to remove insignificant variables are listed in the legends of Figures 1 and 2.

Demographic factors are among the strongest predictors of hospitalization and death: North American Native ethnicity (hospitalization OR 2.01; 95% CI 1.85 - 2.17) and Black ethnicity (hospitalization OR 1.54; 95% CI 1.51 - 1.57), and the 85+ age group (death OR 5.73; 95% CI 5.44 - 6.05). Among socio-economic factors extracted from SVI data, the strongest predictor was living in a zip code with the lowest quartile of income which is associated with an increased probability of severe Covid-19 (hospitalization OR 1.20; 95% CI 1.18 - 1.22).

On the clinical side, End Stage Renal Disease (OR 2.45; 95% CI 2.35 - 2.56), pulmonary fibrosis and pulmonary hypertension (OR 1.53; 95% CI 1.49 - 1.57) and morbid obesity (OR 1.57; 95% CI 1.53 - 1.60) are leading risk factors for hospitalization. COPD (hospitalization OR 1.17; 95% CI 1.15 - 1.19) and beta-2 agonist bronchodilators (OR 1.1795%1.14 - 1.20) are modest and independent risk factors for severe Covid-19. On the other hand, diabetes is not a hospitalization predictor (OR 1.02; 95% CI 1.01 - 1.04) but diabetic Covid-19 patients are at a slightly higher risk of dying from the disease (OR 1.09; 95% CI 1.06 - 1.11). Hypertension and asthma are not associated with higher odds of Covid-19 hospitalizations or death. ACE inhibitors, angiotensin II blockers, and NSAIDs have modest associations with lower rates of Covid-19-related hospitalization or death. More than one prior hospitalization since October 2019 is strongly associated to both Covid-19 related hospitalization (OR 1.74; 95% CI 1.72 - 1.77), and death (OR 1.57; 95% CI 1.53 - 1.61).

The regression analyses indicate an association of prior influenza vaccination with lower Odds Ratio of hospitalization (OR 0.81; 95% CI 0.80 - 0.82), and death (OR 0.72; 95% CI 0.70 - 0.73). Pneumococcal vaccination is associated with reduced Odds Ratio of death (OR 0.84; 95% CI 0.81 - 0.87), as recently proposed,[23] but not hospitalization.

To test the hypothesis that these vaccinations simply represent preventive behaviors that are also associated with less hospitalization and death, we created a preventive behavior variable based on the use of preventive screenings (colonoscopy and mammography) as well as vaccinations. Indeed, these preventive behaviors were associated with lowest Odds Ratios of developing severe Covid-19 (hospitalization OR 0.74; 95% CI 0.72 - 0.75, death OR 0.64; 95% CI 0.62 - 0.66). However, the use of influenza or pneumococcal vaccines were poorly correlated with the other preventive behavior of prior use of preventive screenings, and influenza vaccination retained a significant reduction in Odds Ratio, even when prior preventive screenings were accounted for in the models.

We also built random forest models on the same Covid-19 cases in our study population for both hospitalization versus outpatient care and death versus survival outcomes of Covid-19. In the validation set, the hospitalization model achieved an AUROC of 0.68, and the death model achieved an AUROC of 0.70. We calculated the Gini importance of each feature in the models to determine which variables were the most important for determining severe disease outcomes in our sample (note that the Feature Importance values sum to one).

The Feature Importance (FI) values for the random forest hospitalization and death models are shown in Figure 3. A history of prior hospitalizations before the diagnosis of Covid-19 was the most important variable in the hospitalization model (FI 0.153). From a clinical perspective, chronic kidney disease (FI 0.072), End Stage Renal Disease (FI 0.051), and Congestive Heart Failure (FI 0.042) were among the most important pre-existing comorbidities for predicting Covid-19 hospitalization. Of the demographic variables, male sex (FI 0.045) and black race (FI 0.038) were the most important features.

In the random forest death model, the most important variable in determining Covid-19 survival was age. The Feature Importance summed across the three age groups (65-74, 75-84, and 85+) was 0.243 with age greater than 85 being the most important feature in the model (FI 0.135). The most important comorbidity in the model included chronic kidney disease (FI 0.042) and Congestive Heart Failure (FI 0.045). In addition, recent prior hospitalization was an important predictor of patient death (FI 0.061).

The random forest models reported that preventive health behavior with use of preventive screenings and immunizations among the most important variables in predicting patient outcomes. Of note, a history of preventive screenings was the third most important feature in the hospitalization model (FI 0.055) and the second most important feature in the death model (FI 0.077). Moreover, a flu vaccine was the eighth most important variable in predicting whether a patient will die from Covid-19.

Mapping

The hospitalization model logistic regression coefficients were used to calculate the predicted probability for hospitalization in the event of SARS-CoV-2 infection, at the individual level, for the entire 16 million cohort. We mapped areas in which the percentage of this cohort with a predicted probability of hospitalization was greater than 0.55 for every residential zip code in the U.S. Figure 4 shows an analysis of a portion of the Los Angeles metropolitan area with zip codes displaying a wide range of population percentages over this threshold which were positively correlated with the cumulative Covid-19 case hospitalization rates in these zip codes (Pearson correlation coefficient 0.65 $p < 0.001$; R^2 value for linear regression (figure 3, panel C) 0.43 with p value < 0.001 for intercept and coefficient). We further conducted the same correlation and linear regression analyses in a set of zip codes drawn from other metro areas shown in Figure S1 in the Appendix. All correlation coefficients in these regions (Washington D.C., 0.61; Houston, 0.66; Miami, 0.72; Phoenix, 0.72; and New York City, 0.52) and linear

regression coefficients and intercepts had p values < 0.001. R^2 values for the other metropolitan area linear regressions were Houston (0.44), Phoenix (0.52), New York City (0.27), Miami-Dade- Palm Beach (0.51), Washington DC (0.37). This analysis has been extended to include the 50 largest metropolitan areas in the U.S. In only five of these regions for reasons which have yet to be determined, correlation and linear regression analyses showed no correlation between the zip code level Covid-19 case hospitalization rate and the population risk level.

DISCUSSION:

The severe Covid-19 risk models and the mapping of their outcomes are based, to our knowledge, on the largest Covid-19 dataset assembled to date for this purpose. All members of the study cohort are active users of healthcare services, presenting on average a frailer clinical profile than the general Medicare population,[24] as shown by their clinical characteristics in Table 1. While our observed 910,360 Medicare FFS cases represent a little less than 10% of the total number of Covid-19 cases in the U.S., they contribute close to 50% of all CDC estimated Covid-19-related hospitalizations and a majority of Medicare hospitalizations.[25] The scale of this dataset has allowed us to quantitively and qualitatively validate our hospitalization model predictions with actual Covid-19 case hospitalization rates in multiple geographic areas.

The countrywide distribution, and local representation of the Salus Medicare cohort also enabled Project Salus to produce the first county and zip code level mapping for this population at higher risk for severe Covid-19 and related hospitalizations. Beyond its use for epidemic mapping, the Salus Medicare Covid-19 platform provides a disease surveillance tool for Covid-19, which can be extended to influenza to evaluate the potential compounding effect of these two respiratory diseases on hospitalizations, as influenza undergoes its seasonal progress in the coming months.

Our combined use of logistic regression and random forest analyses, as applied to this very large Medicare population, allowed us to test the current CDC listing of clinical and socio-demographic risk factors for severe Covid-19. Our different analytic approaches provide two different perspectives: the Odds Ratios from the logistic regression help identify individual risk while the Feature Importances in the random forest analyses identify the most important variables for predicting severe Covid-19 outcomes for the entire cohort. For example, in the hospitalization model, the second highest Odds Ratio was associated with North American Native ethnicity. Unlike Odds Ratios which are not influenced by the prevalence of the feature in the population, Feature Importances are more influenced by how common a feature is within the sample. Thus, while North American Native ethnicity has a high Odds Ratio in the hospitalization logistic regression model, its Feature Importance in the random forest model is low (ranked 40th out of 48 variables) due to the low frequency of this feature in the sample (approximately 1% of the sample).

Both logistic regression and random forest analyses affirm the critical risk factors of ethnicity and older age, as most recently identified by the CDC,[26] and morbid obesity as previously reported.[27] However, contrary to prior descriptive analyses performed on smaller population sizes of hospitalized patients only, our analyses based on a very large and nationwide dataset including both outpatient and inpatient Covid-19 cases, reveal the lack of or modest effect of hypertension, diabetes,[28] COPD,[29] and asthma,[30] in our mostly older Medicare beneficiary population. Both analytical approaches also show that prior hospitalizations, a known marker of frailty in aged Medicare beneficiaries,[16] is one of the most significant individual characteristics associated with severe Covid-19 outcomes. Our findings of the significant and independent associations of less severe Covid-19 with prior influenza or pneumococcal immunizations, and use of preventive screenings call for further investigation to address possible

selection bias in this non-randomized observational study which may exist even with the very large size of this cohort.[31,32]

While we made our best efforts in developing these regression and random forest models using rigorous variable selection methods to address potential issues of causal modeling,[33] more iterations of our models will be our next steps before ascertaining whether the associations we found in our models might not be causative (e.g., that prior vaccinations being associated with less severe Covid-19, do not equate to a protective effect of these vaccinations against severe Covid-19).

The independent association of prior use of preventive screenings, a possible marker for a higher rate of Covid-19 preventive behavior (social distancing, hand washing, mask wearing), with less severe Covid-19, is a reminder of the importance of such behaviors.

Regardless of the above limitations, the models we have developed from two different analytical approaches which produced different measures of variable assessment (Odds Ratios at the individual beneficiary level versus Feature Importance at the population group level), provide a comprehensive analysis for clinicians and policy makers to consider. Specifically, and in preparation for the Covid-19 immunization campaign, our models integrating both socio-economic factors and individual clinical data, respond to the recommendations of the National Academies for prioritization and allocation of Covid-19 vaccines when they will become available. They could be used by the Medicare program, in collaboration with state and local health officials to affirmatively invite or encourage highest risk beneficiaries to seek early vaccination and can be used as a tool for vaccine allocation. In Figure 4, we show on a histogram the distribution of the predicted probabilities of hospitalization for SARS-CoV-2 infected patients in the Salus Medicare population which can identify priority groupings for Covid-19 vaccination, according to the quantities of vaccine doses available. Further, once receipt of vaccination is linked to Medicare claims, the system could be used to support post-licensure pharmacovigilance and effectiveness studies. These are of paramount importance, especially in the early phases of a vaccination campaign.

To conclude, we believe that this quantitative analysis of risk for severe Covid-19 provides important new insights which should be considered before finalization of ACIP Covid-19 vaccine prioritization recommendations. State and local governments could consider asking the DoD, through the Mission Assignment process, to provide the Salus Medicare risk mapping to their jurisdictions. The data rich Salus Medicare dataset, which underscores the value of Medicare claim data for epidemiologic surveillance, with its size and nationwide representation, can also augment both the ILINet and COVID-NET disease surveillance systems and complement the existing vaccine monitoring systems for tracking both the safety and efficacy of Covid-19 vaccination in the high risk Medicare population.

We thank Robert Lyden, Travis Yatsko, Adrien Cirou, and Franz Krachtus of Humetrix for their assistance in project management and in the production of the manuscript figures, tables and video.

Supported by the Johns Hopkins University Applied Physics Laboratory (JHU-APL) under a prime contract with the Department of Defense Joint Artificial Intelligence Center (JAIC) for Project Salus.

Department of Defense Disclaimer:

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of Defense, nor the U.S. Government. Authors are military service

---

REFERENCES

1. Centers for Disease Control and Prevention. CDC COVID Data Tracker (https://covid.cdc.gov/covid-data-tracker/#cases_casesinlast7days).
2. Centers for Disease Control and Prevention. Older Adults. September 11, 2020 (https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html).
3. Centers for Disease Control and Prevention. People at Increased Risk. September 11, 2020 (https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/index.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fneed-extra-precautions%2Fpeople-at-increased-risk.html).
4. Sara Oliver M.D. Centers for Disease Control and Prevention. ACIP COVID-19 Vaccines Work Group. September 22, 2020 (https://www.cdc.gov/vaccines/acip/meetings/downloads/slides-2020-09/COVID-06-Oliver.pdf).
5. The National Academies of Sciences, Engineering, and Medicine. National Academies Release Framework for Equitable Allocation of a COVID-19 Vaccine for Adoption by HHS, State, Tribal, Local, and Territorial Authorities. October 2, 2020 (https://www.nationalacademies.org/news/2020/10/national-academies-release-framework-for-equitable-allocation-of-a-covid-19-vaccine-for-adoption-by-hhs-state-tribal-local-and-territorial-authorities).
6. Centers for Disease Control and Prevention. CDC SVI 2018 Documentation. January 31, 2020 (https://svi.cdc.gov/Documents/Data/2018_SVI_Data/SVI2018Documentation.pdf).
7. Surgo Foundation. The COVID-19 Community Vulnerability Index (CCVI) (https://precisionforcovid.org/ccvi).
8. Kathleen Dooling, M.D., M.P.H. Centers for Disease Control and Prevention. ACIP COVID-19 Vaccines Work Group. September 22, 2020 (https://www.cdc.gov/vaccines/acip/meetings/downloads/slides-2020-09/COVID-07-Dooling.pdf).
9. Centers for Disease Control and Prevention. Percent of U.S. Adults 55 and Over with Chronic Conditions (https://www.cdc.gov/nchs/health_policy/adult_chronic_conditions.htm).
10. Centers for Disease Control and Prevention. People with Certain Medical Conditions. September 11, 2020 (https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html).
11. CDC COVID-19 Response Team. Preliminary Estimates of the Prevalence of Selected Underlying Health Conditions Among Patients with Coronavirus Disease 2019 — United States, February 12 -March 28, 2020. Centers for Disease Control and Prevention 2020;69(13):382-386.
12. Lithander F, Neumann S, Tenison E, et al. COVID-19 in older people: a rapid clinical review. Oxford Academic 2020;49(4):501-515.

13. Garg S, Kim L, Whitaker M, et al. Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 — COVID-NET, 14 States, March 1 -30, 2020. Centers for Disease Control and Prevention 2020;69(15):458-464.
14. Centers for Medicare & Medicaid Services. Preliminary Medicare COVID-19 Data Snapshot (https://www.cms.gov/research-statistics-data-systems/preliminary-medicare-covid-19-data-snapshot).
15. Marvel S, House J, Wheeler M, et al. The COVID-19 Pandemic Vulnerability Index (PVI) Dashboard: Monitoring county-level vulnerability using visualization, statistical modeling, and machine learning. medRxiv 2020. DOI: 10.1101/2020.08.10.20169649.
16. Experton B, Li Z, Branch L, Ozminkowski R, Mellon-Lacey D. The impact of payor/provider type on health care use and expenditures among the frail elderly. American Journal of Public Health 1997;87(2):210-216.
17. Melinda Beck. The Wall Street Journal. Next in Tech: App Helps Patients Track Care. December 16, 2013 (https://www.wsj.com/articles/SB10001424052702303330204579248420368822400).
18. AI in Defense DoD's Artificial Intelligence Blog. The JAIC Forges Ahead. May 20, 2020 (https://www.ai.mil/blog_05_20_20-the_jaic_forges_ahead.html).
19. R Core Team (2019). R: A language and environment for statistical computing and graphics. R Foundation for Statistical Computing, Vienna, Austria. (https://www.R-project.org/).
20. Frank E Harrell Jr. rms: Regression Modeling Strategies. July 18, 2020 (https://CRAN.R-project.org/package=rms).
21. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77. DOI: 10.1186/1471-2105-12-77.
22. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 2010;33(1):1-22. DOI: 10.18637/jss.v033.i01
23. Root-Bernstein R. Possible Cross-Reactivity between SARS-CoV-2 Proteins, CRM197 and Proteins in Pneumococcal Vaccines May Protect Against Symptomatic SARS-CoV-2 Disease and Death. Vaccines 2020;8(4):559-579; DOI:10.3390/vaccines8040559.
24. Centers for Medicare & Medicaid Services. Prevalence of Chronic Conditions among Fee-for-Service Beneficiaries: 2017 (https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/Chartbook_Charts).
25. COVID-NET. COVID-19 Laboratory Confirmed Hospitalizations. September 26, 2020. (https://gis.cdc.gov/grasp/covidnet/COVID19_5.html).
26. Centers for Disease Control and Prevention. COVID-19 Hospitalization and Death by Age. Last updated August 18, 2020. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html
27. Palaiodimos L, Kokkinidis D, Li W, et al. Severe Obesity, Increasing Age and Male Sex Are Independently Associated with Worse In-Hospital Outcomes, and Higher In-Hospital Mortality, in a Cohort of Patients with COVID-19 in the Bronx, New York. Metabolism 2020;108(154262). DOI: 10.1016/j.metabol.2020.154262

28. Richardson S, Hirsch J, Narasimhan M, et al. Presenting Characteristics, Comorbidities, and Outcomes among 5700 Patients Hospitalized with COVID-19 in the New York City Area. JAMA 2020;323(20):2052-2059. DOI: 10.1001/jama.2020.6775
29. Lippi G, Henry B. Chronic Obtrusive Pulmonary Disease Is Associated with Severe Coronavirus Disease (COVID-19). Respiratory Medicine 2020;167(105941). DOI: 10.1016/j.rmed.2020.105941
30. Song J, Zeng M, Wang H, et al. Distinct Effects of Asthma and COPD Comorbidity on Disease Expression and Outcome in Patients with COVID-19. European Journal of Allergy and Clinical Immunology 2020. DOI: 10.1111/all.14517
31. Heckman J.J. (1990) Selection Bias and Self-selection. In: Eatwell J., Milgate M., Newman P. (eds) Econometrics. The New Palgrave. Palgrave Macmillan, London. DOI: 10.1007/978-1-349-20570-7_29
32. Terzal J. Mata implementation of Gauss-Legendre quadrature in the M-estimation context: Correcting for sample-selection bias in a generic nonlinear setting. EconPapers 2019 (https://econpapers.repec.org/paper/bocscon19/31.htm).
33. Judea Pearl. The Causal Foundations of Structural Equation Modeling. Handbook of Structural Equation Modeling. New York: Guilford Press 2012 (https://apps.dtic.mil/sti/pdfs/ADA557445.pdf).

Table 1. **Demographic, Clinical and Socioeconomic Characteristics of the Project Salus Medicare Cohort**

Variable	Non-Covid-19 Cases†	Covid-19 Outpatients‡	Covid-19 Hospitalized¶	Covid-19 Deaths#
Total	14,424,612	388,424	317,189	122,613
Age	Median (IQR)	Median (IQR)	Median (IQR)	Median (IQR)
Age	73 (67 - 80) ✦✦✦	73 (68 - 82)	76 (68 - 84)✦✦✦	82 (73 - 89)✦✦✦
	%	% (no.)	% (no.)	% (no.)
Under 65	15.3% (2,203,457)***	14.5% (56,513)	14.4% (45,810)	7.5% (9,244)***
From 65 to 74	41.4% (5,977,645)***	40.0% (155,300)	32.0% (101,543)***	21.6% (26,511)***
From 75 to 84	28.9% (4,162,510)***	27.0% (105,004)	30.5% (96,899)***	31.1% (38,113)***
Over 85	14.4% (2,081,000)***	18.4% (71,607)	23.0% (72,937)***	39.8% (48,745)***
Sex	%	% (no.)	% (no.)	% (no.)
Male	43.4% (6,253,297)	40.4% (157,027)	48.4% (153,653)***	48.5% (59,471)***
Female	56.6% (8,171,313)	59.6% (231,397)	51.6% (163,536)***	51.5% (63,142)***
Race	%	% (no.)	% (no.)	% (no.)
North Americ. Native	0.6% (90,102)***	0.5% (2,028)	1.0% (3,089)***	0.9% (1,052)***
Black	9.6% (1,379,610)***	12.7% (49,405)	18.7% (59,443)***	17.1% (20,944)***
Hispanic	2.0% (290,587)***	3.9% (15,000)	4.8% (15,313)***	4.4% (5,385)***
Asian	1.9% (272,480)***	2.1% (8,167)	2.3% (7,423)***	2.5% (3,054)***
White	82.3% (11,865,528)***	76.9% (298,554)	69.8% (221,542)***	72.4% (88,805)***
Other	1.6% (223,904)***	1.7% (6,526)	1.7% (5,510)	1.7% (2,073)*
Unknown	2.1% (302,401)***	2.3% (8,744)	1.5% (4,869)***	1.1% (1,300)***
Income (SVI EPL_PCI)	%	% (no.)	% (no.)	% (no.)
Upper quartiles	88.8% (12,807,172)***	88.79% (335,336)	82.1% (260,540)***	83.3% (102,081)***
Poorest quartile	11.2% (1,617,440)***	11.21% (53,088)	17.9% (56,649)***	16.7% (20,532)***
Poverty (SVI EPL_POV)	%	% (no.)	% (no.)	% (no.)
Upper quartiles	89.2% (12,870,718)***	89.23% (335,974)	82.9% (263,081)***	83.7% (102,682)***
Poorest quartile	10.8% (1,553,894)***	10.77% (52,450)	17.1% (54,108)***	16.3% (19,931)***
Housing	Median (IQR)	Median (IQR)	Median (IQR)	Median (IQR)
Crowded Quarters (SVI EPL_CROWD)	0.44 (0.29 - 0.59)✦✦✦	0.46 (0.30 - 0.66)	0.48 (0.32 - 0.67) ✦✦✦	0.47 (0.31 - 0.66) ✦✦✦
Multi-Unit (SVI EPL_MUNIT)	0.49 (0.33 - 0.64)✦✦✦	0.53 (0.35 - 0.70)	0.51 (0.35 - 0.68) ✦✦✦	0.52 (0.35 - 0.69) ✦✦✦
	%	% (no.)	% (no.)	% (no.)
Disabled Status	24.9% (3,590,681)***	26.1% (101,185)	27.7% (87,718)***	21.9% (26,867)***
Dual Status	21.8% (3,148,657)***	38.0% (147602)	40.8% (129,282)***	48.3% (59,274)***
Prior Hospitalization	%	% (no.)	% (no.)	% (no.)
0	100.0% (14,424,612)***	79.2% (307,588)	64.5% (204,628)***	62.6% (76,777)***
1 or more	0.0% NA	20.8% (80,836)	35.5% (112,561)***	37.4% (45,836)***
Clinical Variables	%	% (no.)	% (no.)	% (no.)
ESRD	1.8% (258,842)***	2.1% (8,225)	6.4% (20,287)***	5.2% (6,319)***
Chronic Kidney Disease	36.9% (5,326,575)***	40.0% (155,291)	53.4% (169,412)***	58.0% (71,153)***
Pulmonary Fibrosis & HTN	6.5% (933,246)***	5.0% (19,330)	9.6% (30,524)***	10.0% (12,301)***
Chronic Liver Disease	2.6% (371,964)***	2.5% (9,700)	3.8% (11,959)***	3.6% (4,440)***
COPD	26.0% (3,745,984)***	28.0% (108,697)	36.3% (115,223)***	39.8% (48,804)***
CHF	24.5% (3,540,526)***	30.3% (117,573)	41.2% (130,783)***	48.7% (59,731)***
Stroke/TIA	13.6% (1,959,028)***	18.3% (71,078)	22.7% (72,074)***	28.6% (35,044)***
Diabetes	37.5% (5,406,693)***	45.1% (175,109)	53.2% (168,839)***	56.0% (68,647)***
Hypertension	76.3% (11,011,439)***	79.5% (308,975)	85.0% (269,612)***	88.9% (108,994)***
Acute MI	0.9% (130,315)***	0.8% (3,099)	1.5% (4,761)***	1.6% (2,002)***
Ischemic Heart Disease	44.3% (6,386,752)***	49.5% (192,160)	57.9% (183,770)***	64.8% (79,444)***
Asthma	16.3% (2,350,070)***	17.6% (68,527)	19.8% (62,737)***	18.8% (23,111)***
Chemotherapy	9.3% (1,341,872)***	11.2% (43,378)	14.1% (44,639)***	13.4% (16,369)***
Obesity	14.9% (2,153,115)***	16.0% (62,252)	16.4% (51,907)***	11.6% (14,234)***
Morbid Obesity	8.6% (1,236,500)***	9.5% (37,070)	13.7% (43,298)***	9.5% (11,624)***

† Asterisks shown in this column indicate p values for the differences between individuals who have not been diagnosed with Covid-19 and confirmed Covid-19 cases in the Salus Medicare cohort.
‡ Covid-19 outpatients are defined are individuals who did not require hospitalization for the disease and remained alive at least 30 days after diagnosis. Covid-19 cases who either did not die or were not hospitalized for the disease but who had no claims more than 30 days after their Covid-19 diagnosis are not shown in this table.
¶ Covid-19 hospitalized cases are those requiring inpatient admission for management of their disease. Asterisks shown in this column indicate the p values for differences between this group and the Covid-19 outpatient group.
# Covid-19 deaths are deaths attributed to Covid-19 based on the timing of death in relation to the date of diagnosis. Asterisks shown in this column indicate the p values for differences between Covid-19 cases who died from the disease and those who survived.
* p < 0.05; ** p < 0.01; *** p <0.001 by Chi Square test
✦ p < 0.05; ✦✦ p <0.01; ✦✦✦ p <0.001 by Mann-Whitney test

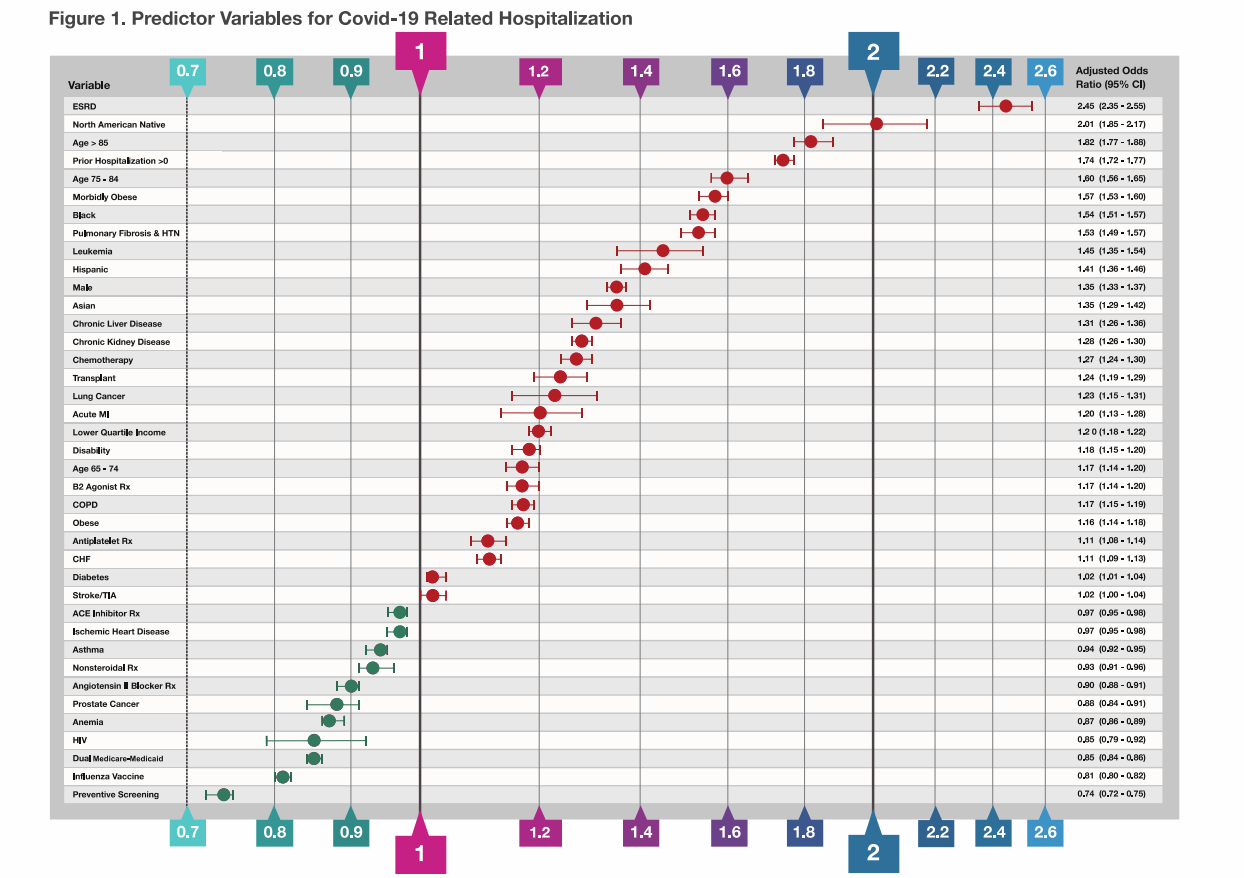

Figure 1: Predictor Variables for Covid-19 Related Hospitalization. The independent variable Odds Ratios were determined by binary logistic regression analysis of confirmed Covid-19 cases that required hospitalization for the disease and for those that were managed with outpatient care only. In addition to the thirty-nine variables shown in the figure, the following variables were included in the model based on the variable selection criteria described in Methods but are not shown: colorectal cancer (OR 1.07; 95% CI 1.01- 1.14), endometrial cancer (OR 1.12; 95% CI 1.00 - 1.25) in the second half of 2019, unknown ethnicity (OR 0.96; 95% CI 0.91 - 1.00), prescriptions overlapping the Covid-19 diagnosis date of Azithromycin (OR 1.15; 95% CI 1.11 - 1.18), chloroquine and hydroxychloroquine drugs (OR 0.96; 95% CI 0.91 - 1.01), anticoagulant drugs (OR 1.06; 95% CI 1.04 - 1.08), opioid drugs (OR 1.03; 95% CI 1.01 - 1.05) and H2 blocker drugs (OR 1.03; 95% CI 0.99 - 1.06); Variables excluded from the model based on the variable selection criteria included: “other” ethnicity, a history breast cancer in the second half of 2019, prescriptions for immunosuppressive and inhaled corticosteroid drugs overlapping the Covid-19 diagnosis date, hypertension and pneumococcal vaccinations.

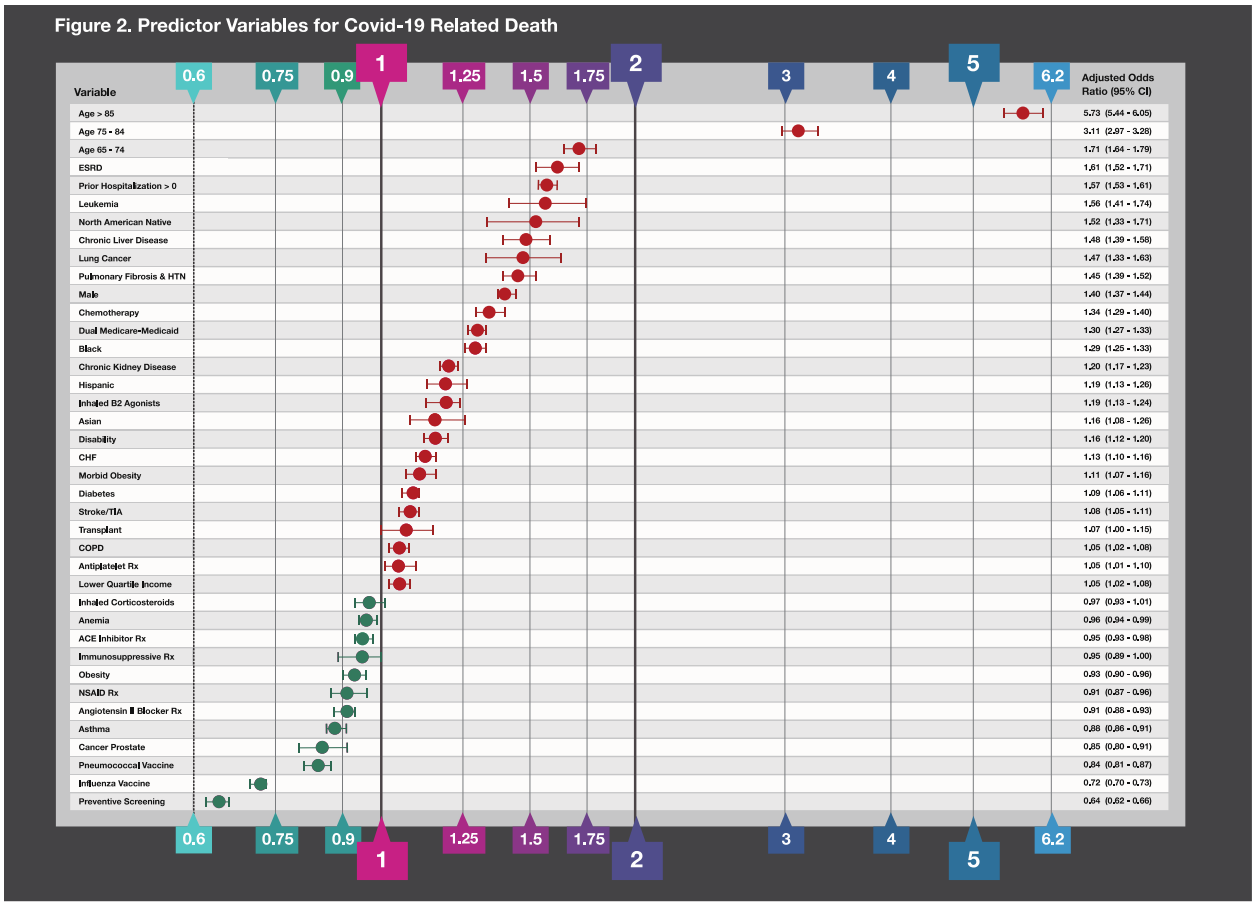

Figure 2: Predictor Variables for Covid-19 Related Death. The independent variable Odds Ratios were determined by binary logistic regression analysis of confirmed Covid-19 cases that survived and those that died within 60 days of Covid-19 diagnosis. In addition to the thirty nine variables shown in the figure, the following variables were included in the model based on the variable selection criteria described in Methods but are not shown: endometrial cancer (OR 1.29; 95% CI 1.06 - 1.54) in the second half of 2019, prescriptions overlapping the Covid-19 diagnosis date for H2 Blocker drugs (OR 1.13; 95% CI 1.07 - 1.19), chloroquine and hydroxychloroquine drugs (OR 1.19; 95% CI 1.10 - 1.29) and Azithromycin (OR 1.18; 95% CI 1.10 - 1.29); the “other” ethnicity (OR 1.08; 95% CI 0.99 - 1.18) and “unknown” ethnicity (OR 0.87; 95% CI 0.79 - 0.95). Variables excluded from the model based on the variable selection criteria include a history of colorectal cancer and breast cancer, or acute MI in the second half of 2019 and the following chronic conditions: hypertension, ischemic heart disease, HIV, and prescriptions for opioid and anticoagulant drugs overlapping the Covid-19 diagnosis dates.

Figure 3. Random Forest Hospitalization & Death Model Feature Importance

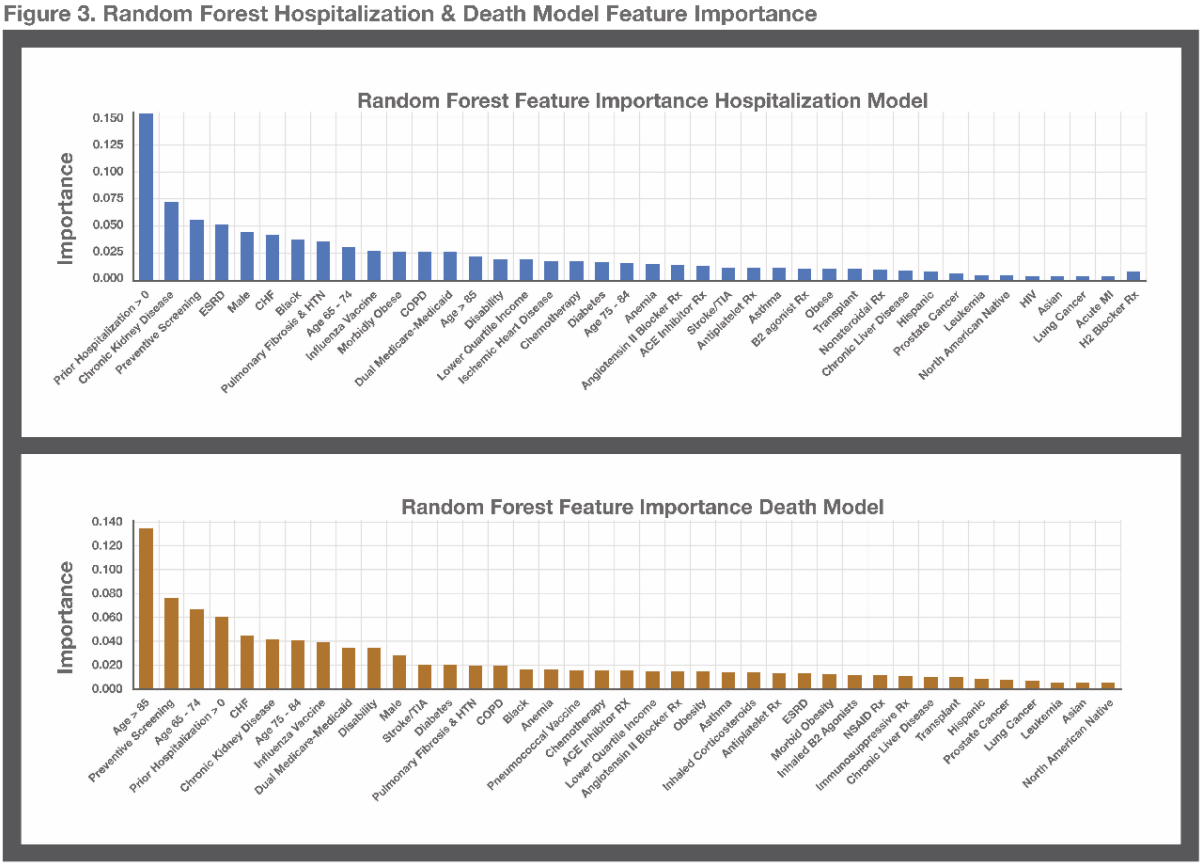

Figure 3: Random Forest Hospitalization and Death Model Feature Importance. Variables that were selected for inclusion in the Hospitalization and Death logistic regression models were used to build these two random forest models. The Feature Importance values for the variables not shown in the Hospitalization model graph are: prescriptions overlapping the Covid-19 diagnosis date for anticoagulants (0.01420, opioids drugs (0.0131), Azithromycin (0.0103), chloroquine and hydroxychloroquine drugs (0.0058); colorectal cancer (0.0043) and endometrial cancer (0.002) in second half of 2019; the Medicare defined “other” (0.0038) and “unknown” (0.0038) ethnicities. The Feature Importance values for the variables not shown in the Death model graph include: prescriptions overlapping the Covid-19 diagnosis date for Azithromycin (0.0103), chloroquine and hydroxychloroquine drugs (0.006), a history of endometrial cancer (0.002) in the second half of 2019, and “other” (0.0043) and “unknown” (0.0037) ethnicities.

Figure 4. Los Angeles Covid-19 Hospitalization Risk Map

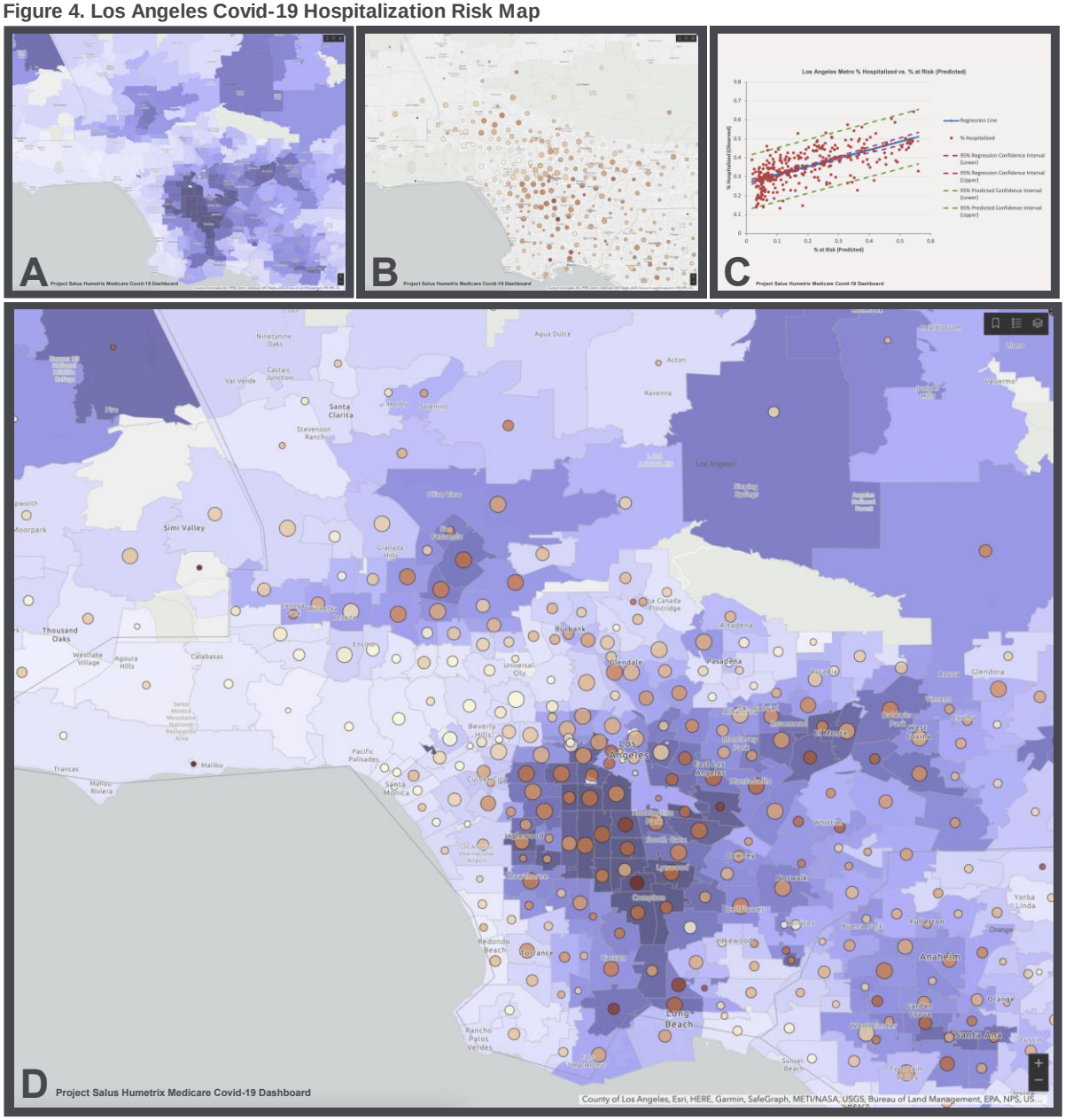

Figure 4: Los Angeles Covid-19 Hospitalization Risk Map. Panel A shows the percentage of the Salus cohort with a predicted probability of hospitalization when diagnosed with Covid-19 of over 0.55 on a light blue to dark lavender color scale. Panel B shows the cumulative number of hospitalizations per zip code (increasing size of circles denotes a higher hospitalization count) with the percentage of cases requiring hospitalization shown on a beige to dark orange scale. Panel C shows a linear regression analysis of the case hospitalization rate (Y axis) as a function of the risk level in each zip code (regression $R^2 = 0.43$); Panel D is an overlay of panel B on Panel A and demonstrates that zip codes with the highest predicted probabilities of hospitalization with Covid-19 tend to have higher observed percentage of cases requiring hospitalization and vice versa.

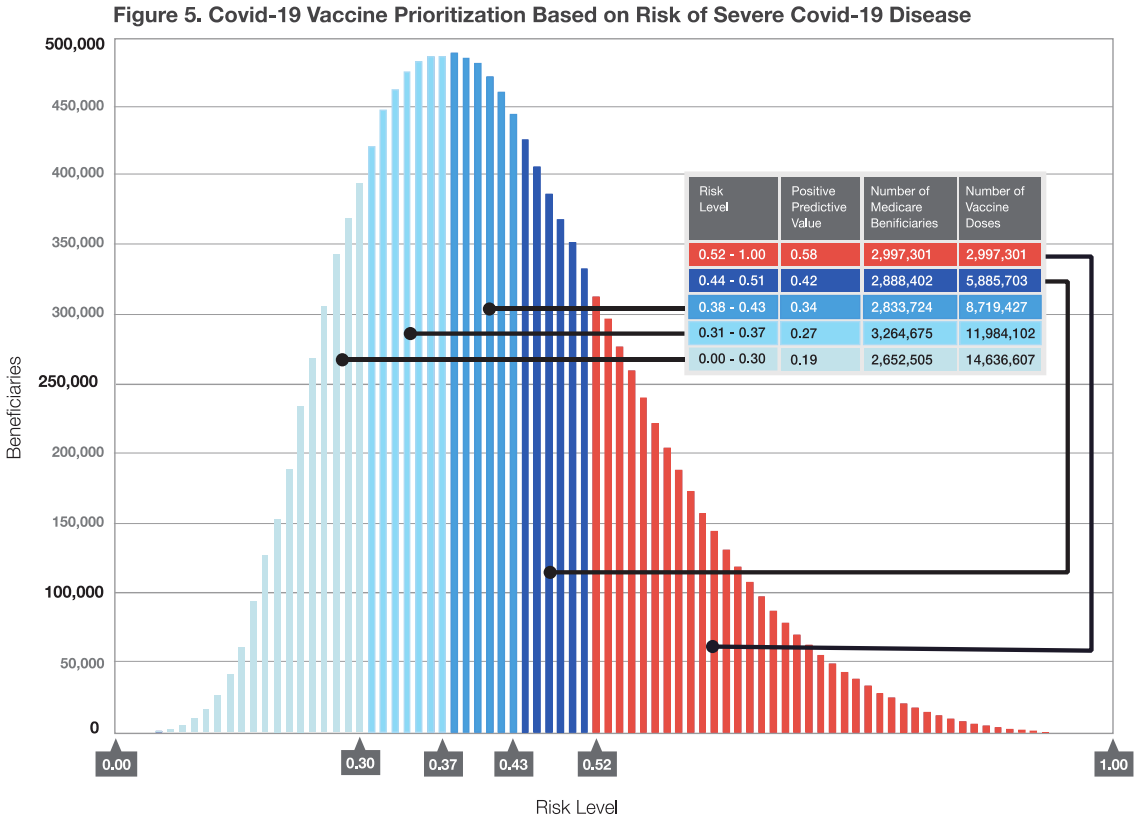

Figure 5 - Covid-19 Vaccine Prioritization Based on Risk of Severe Covid-19 Disease. The logistic regression model coefficients for the independent variables shown in figure 1 were used to calculate the predicted probabilities of hospitalization in the Salus cohort. The distribution of predicted probabilities was split into 5 groups shown in the table of approximately 3M beneficiaries each to enable stratification of the cohort by risk of severe disease in order to prioritize individuals for Covid-19 vaccination.

Supplemental Appendix

Supplemental Methods:

Medicare claim data processing; The Humetrix SaaS platform installed in the SUNet secure classified government network provides automated pre-processing of weekly updates downloaded to the Humetrix SUNet enclave of Medicare Part A inpatient and outpatient, Hospice and SNF, Part B Carrier claims and Part D (PDE) claims data to generate output files containing derived variables used to train logistic regression models as described below. The platform's medical terminology service includes a complete collection of AMA CPT-4, FDA NDC, NNPES NPI, ICD-10-CM, CMS Level II HCPCS, and NLM RxNorm codes and provides automated identification grouping of ICD-10-CM codes to identify chronic condition categories and NDC drug code to RxNorm ingredient code mappings to identify pharmaceutical classes of active pharmaceutical ingredients from Medicare claim data.

Demographic and Coverage Variables: weekly updates of CMS "Master Beneficiary Summary 2020 File" (MBSF_2020) data were processed, extracting variables: ORIG_REASON_FOR_ENTITLEMENT (Disability), ZIP_CD (residential zip code), YOB, SEX_CODE, RACE_CODE, and DUAL_STUS_CD 01-12 (Dual Medicare-Medicaid insurance).

Social Vulnerability Index (SVI) variables associated with zip codes: these were derived from CDC data, which are categorized by census tract. The variables analyzed using binary logistic regression to determine significant predictor variables for hospitalization or death due to Covid-19 included the following SVI variables: EPL_PCI, EPL_POV, EPL_NOHSDP, RPL_THEME1, EPL_CROWD, EPL_GROUPQ, EPL_MUNIT, EPL_THEME4, RPL_THEMES. Among these SVI variables, living in a zip code in the lowest quartile of income computed from the EPL_PCI was the only variable selected for the hospitalization and death logistic regression modelling described in this paper. The other SVI variables were excluded either because of their high correlation with EPL_PCI or because they were consistently excluded by the variable selection criteria in prior logistic regressions.

Chronic Condition Summary Variables: Chronic conditions in the MBSF_2019 chronic condition segment file was extracted, including variables:

1. acute myocardial infarction July-December 2019 (AMI)
2. chronic kidney disease (CHRONIC_KIDNEY_EVER), COPD (COPD_EVER)
3. congestive heart failure (CHF_EVER)
4. diabetes mellitus (DIABETES_EVER)
5. ischemic heart disease (ISCHEMICHEART_EVER)
6. stroke/transient ischemic attack (STROKE_TIA_EVER)
7. recent breast cancer July-December 2019 (CANCER_BREAST)
8. recent colorectal cancer July-December 2019 (CANCER_COLORECTAL)
9. recent prostate cancer July-December 2019 (CANCER_PROSTATE)
10. recent lung cancer July-December 2019 (CANCER_LUNG)
11. recent endometrial cancer July-December 2019 (CANCER_ENDOMETRIAL)
12. anemia July-December 2019 (ANEMIA)
13. asthma (ASTHMA_EVER)
14. hypertension (HYPERT_EVER)

End Stage Renal Disease (ESRD) was identified by finding a "Y" code in the ESRD_INDICATOR field of the MBSF_2020 file.

Humetrix SaaS platform identification of additional chronic conditions present prior to Covid-19 diagnosis date: The following chronic condition variables were identified by analyzing ICD10 codes in these CMS claim files: (Beneficiary Part A institutional Inpatient and Outpatient claims, Part B Carrier claims, SNF and Hospice claims) from October 1, 2019 through September 30, 2020:

1. leukemia (includes acute and chronic myeloid and lymphocytic leukemias as well as less common leukemias),
2. Pulmonary fibrosis or pulmonary hypertension (include idiopathic pulmonary fibrosis, interstitial lung disease, pneumoconioses, pulmonary sarcoidosis, pulmonary hypertension),
3. chronic liver disease (includes alcoholic cirrhosis, primary biliary cirrhosis, chronic viral hepatitis due to hepatitis B and C, alcoholic fatty liver, primary sclerosing cholangitis, Wilson's disease)
4. HIV/AIDS
5. organ transplant (includes following transplants: lung, bone, heart, liver, pancreas, intestine, kidney, bone marrow),
6. morbid obesity (BMI over 40),
7. obesity (BMI 30 - 40).

Medication variables: PDE and Claims Data Files were analyzed by the Humetrix SaaS system to derive the factors listed below. Note coding of non-chemotherapy drugs as "True" signifies at least one applicable prescription fill was identified in 2020. These variables were all identified by mapping NDC drug product codes to RxNorm ingredient codes for these pharmaceutical classes of drugs:

1. chemotherapy: signifies that a beneficiary at any time in 2020 either had an ICD10 code for chemotherapy in part A Institutional or Part B Carrier claims, had a CPT-4 code indicating administration of parenteral chemotherapy in a Part B claim, or had a pharmacy (PDE) claim with an NDC code which mapped to an RxNorm ingredient code for an active pharmaceutical ingredient belonging to multiple classes of chemotherapeutic agents.
2. anticoagulant drugs
3. antiplatelet drugs
4. inhaled Beta-2 agonists
5. inhaled corticosteroid drugs
6. opioid drugs
7. histamine type-2 receptor blockers
8. angiotensin converting enzyme inhibitors (ACE inhibitors)
9. angiotensin II receptor blockers
10. non-steroidal anti-inflammatory (NSAID) drugs
11. immunosuppressive drugs of diverse pharmaceutical classes
12. Azithromycin and Chloroquine drugs (includes both Chloroquine and Hydroxychloroquine). A value of "True" was assigned if a sufficient quantity of these drugs was filled to extend up to, or 10 days beyond, the Covid-19 diagnosis date.

Immunization Variables: Influenza and pneumococcal vaccinations were identified using their respective CPT-4 codes found in Part B claims from October 1, 2019 until September 30, 2020. Shingrix® herpes zoster vaccinations were identified by NDC codes found in Part D (PDE) claims from January 1, 2020 through September 30, 2020.

Covid-19 outcome variables:

1. Covid-19 hospitalizations were identified either by Part B Carrier claims with place of service code and CPT codes indicating inpatient care with a date of service no more than 14 days after or 10 days before the Covid-19 diagnosis date, or by finding Part A Inpatient claims where the data of admission no more than 14 days after or 10 days before the Covid-19 diagnosis date.
2. Covid-19 cases managed by outpatient care only were defined as cases who were not hospitalized in the 30 days after the Covid-19 diagnosis date and who did not die within 60 days of the diagnosis date.
3. Deaths due to Covid-19 were defined as either deaths which occurred during a Covid-19 hospitalization using the inpatient PTNT_DSCHRG_STUS_CD code, or a death which occurred within 60 days of the Covid-19 diagnosis as reported in the DEATH_DT field of the MBSF_2020 file.

Supplemental Results:

Effect size calculation (Table 1): Due to the very large sizes of the compared populations, the p values for the statistical tests presented in Table 1 are almost all very significant. We used Cramér's V to calculate effect sizes values for the comparisons between groups for binary and categorical variables (age, sex, race, income, poverty, dual Medicare-Medicaid dual status, disabled status, prior hospitalization, and clinical variables) all of which had a computed Cramér's V values under 0.17 indicating a small effect size. We used Vargha and Delanay's A (VDA) to calculate effect sizes for comparisons between groups for the quantitative variables (age in years, SVI EPL_CROWD and SVI EPL_MUNIT) and found Vargha and Delanay's A values of 0.42 and 0.66 also indicative of small effect size.

Analysis of Hospitalization model predictions of Covid-19 case hospitalization rates in six large metropolitan areas (see following pages).

Figure S1

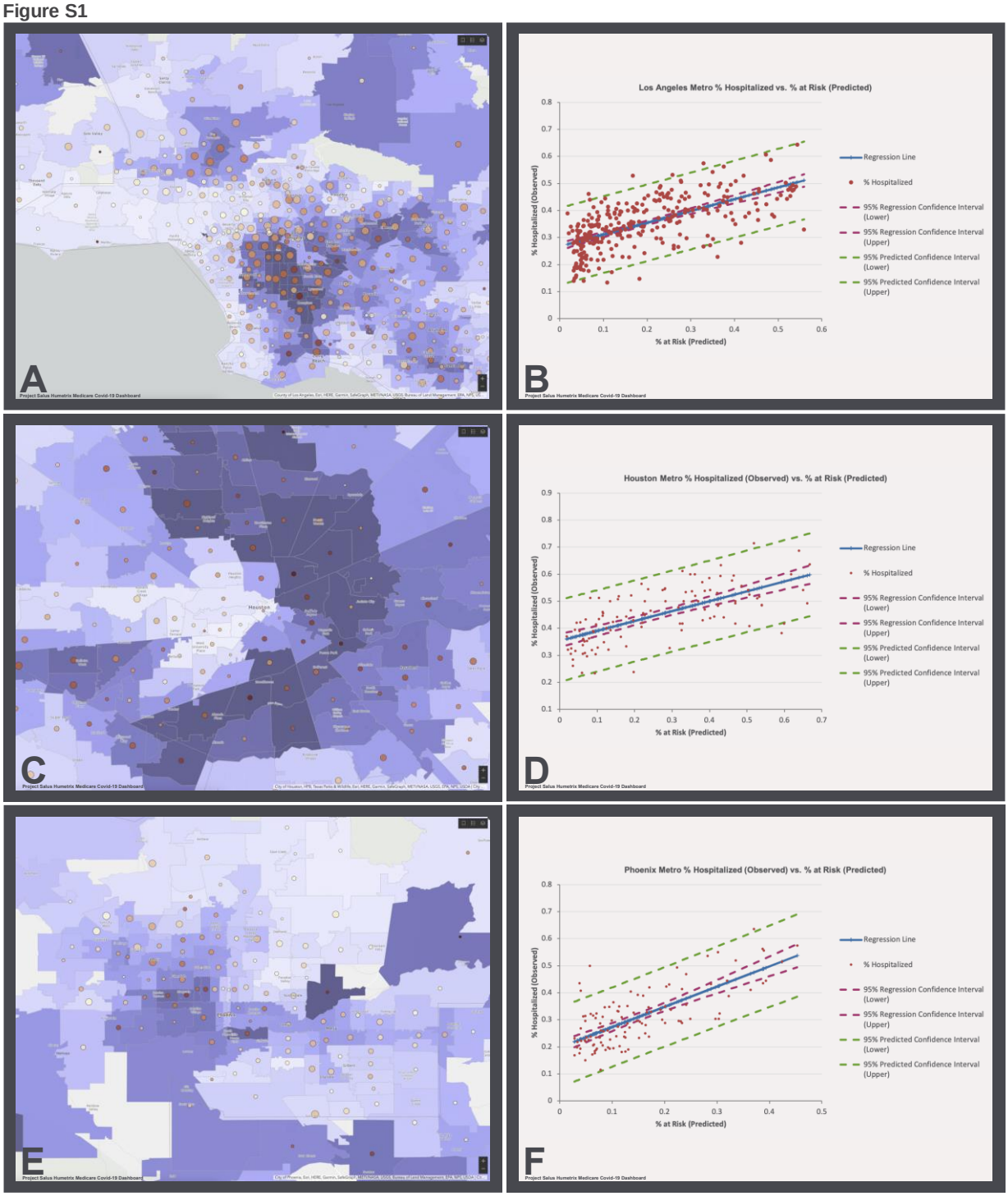

Figure S1 Continued

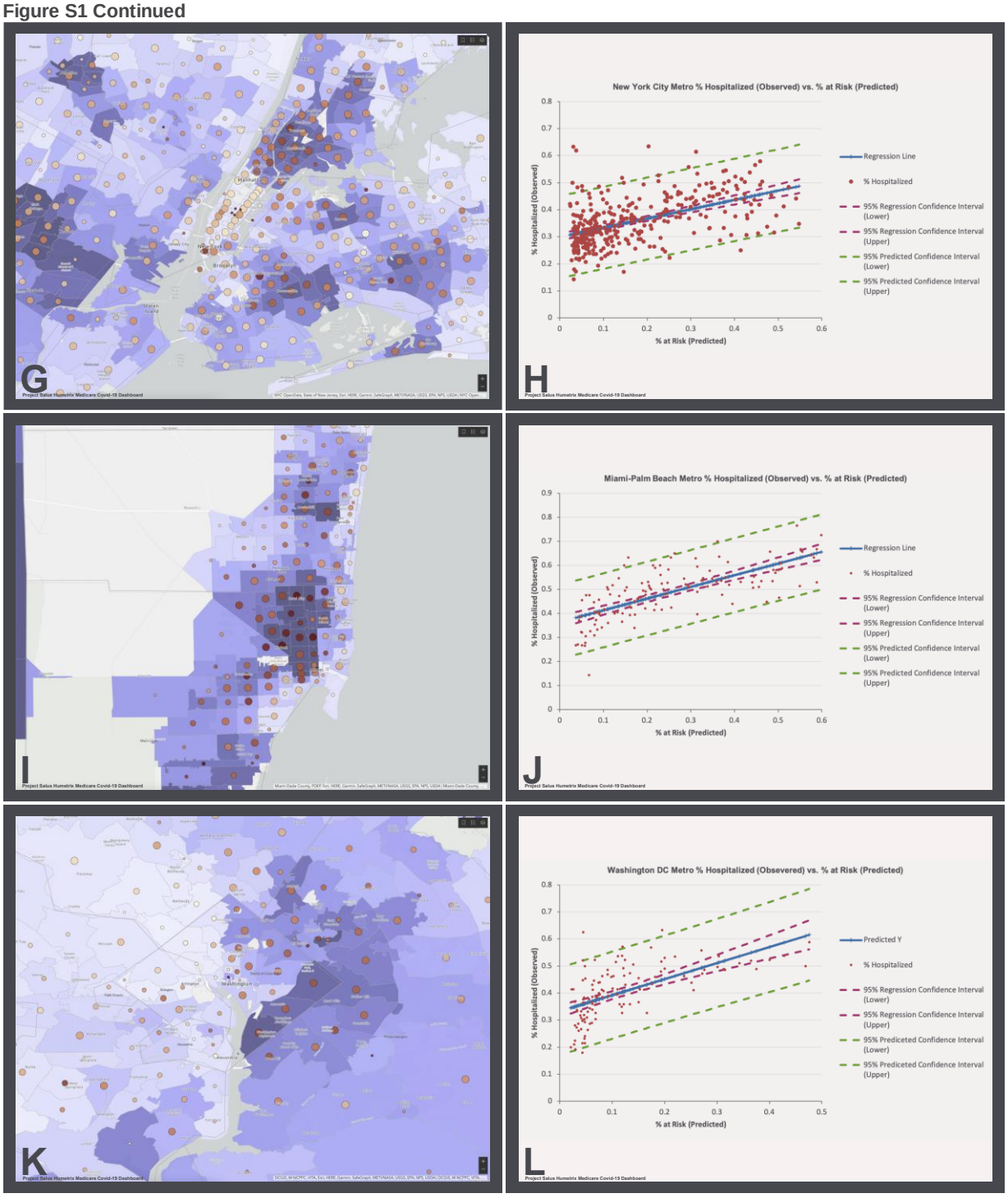

Figure S1: Metropolitan region risk maps for severe Covid-19 (A) Los Angeles County, (C) Houston, (E) Phoenix, (G) New York City, (I) Miami-Dade - Palm Beach, (K) Washington DC. In each metropolitan region, zip codes with the higher predicted probabilities of hospitalization with Covid-19 based on the logistic regression hospitalization model are shown in darker shades of lavender. Circles denote observed cumulative hospitalizations due to Covid-19 extracted from claims data with the size of the circles scaled to the number of hospitalizations. The percentage of cases requiring hospitalization is

(Figure S1 Legend Continued)

displayed by the color of the circles on a beige to dark orange-red color scale with the darker circles indicating higher zip codes with higher percentages of cases requiring hospitalization for Covid-19. Panels B, D, F, H, J, L shows corresponding linear regression analyses of the case hospitalization rates (Y axis) as a function of the risk level in each zip code in each metropolitan region. The following R^2 values for the regressions were: Los Angeles (0.43), Houston (0.44), Phoenix (0.52), New York City (0.27), Miami-Dade-Palm Beach (0.51), Washington DC (0.37).