

Social and Clinical Determinants of COVID-19 Outcomes: Modeling Real-World Data from a Pandemic

Epicenter

Jyothi Manohar MPH^{1,5}, Sajjad Abedian MS², Rachel Martini BS³, Scott Kulm BE^{1,4}, Mirella Salvatore MD^{5,6}, Kaylee Ho MS⁶, Paul Christos DrPH⁶, Thomas Campion PhD^{6,7}, Julianne Imperato-McGinley MD⁷, Said Ibrahim MD/MPH/MBA^{6,7}, Teresa H. Evering⁵ MD/MS, Erica Phillips^{5,10} MD, Rulla Tamimi^{6,11} ScD, Vivian Bea³ MD, Onyinye D. Balogun¹² MD, Andrea Sboner^{1,11,13}, PhD, Olivier Elemento^{1,4,7,8,11,13} PhD and Melissa Boneta Davis PhD^{1,3,9*}

¹Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medicine, New York City, NY

²Information Technology and Services Department, Weill Cornell Medicine, New York City, NY

³Department of Surgery, Weill Cornell Medicine, New York City, NY

⁴Department of Physiology, Weill Cornell Medicine, New York City, NY

⁵Department of Medicine, Weill Cornell Medicine, New York City, NY

⁶Department of Population Health Sciences, Weill Cornell Medicine, New York City, NY

⁷Clinical Translational Science Center, Weill Cornell Medicine, New York City, NY

⁸WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York City, NY

⁹Institute for the Study of Breast Cancer Subtypes, Weill Cornell Medicine, New York City, NY

¹⁰Department of Integrative Medicine, Weill Cornell Medicine, New York City, NY

¹¹Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York City, NY

¹²Department of Radiation Oncology, Weill Cornell Medicine, New York City, NY

¹³Institute for Computational Biomedicine, Weill Cornell Medicine, New York City, NY

*Corresponding Author:

Melissa Boneta Davis, PhD

Department of Surgery, Weill Cornell Medicine

420 E 70th Street

New York City, NY 10021

Phone: (646) 962-2855

Fax: (646) 962-0468

Email: mbd4001@med.cornell.edu

Manuscript word count: 3660

32 **KEY POINTS**

33 **QUESTION:** What is the impact of patient self-reported race, ethnicity, socioeconomic status, and clinical profile
34 on COVID-19 hospitalizations, severity, and mortality?

35 **FINDINGS:** In patients diagnosed with COVID-19, being over 50 years of age, having type 2 diabetes and
36 hypertension were the most important risk factors for hospitalization and severe outcomes regardless of patient
37 race or socioeconomic status.

38 **MEANING:** In this large sample of patients diagnosed with COVID-19 in New York City, we found that clinical
39 comorbidity, more so than social determinants of health, was associated with important patient outcomes.

40

41 **ABSTRACT**

42 **IMPORTANCE:** As the United States continues to accumulate COVID-19 cases and deaths, and disparities persist,
43 defining the impact of risk factors for poor outcomes across patient groups is imperative.

44 **OBJECTIVE:** Our objective is to use real-world healthcare data to quantify the impact of demographic, clinical,
45 and social determinants associated with adverse COVID-19 outcomes, to identify high-risk scenarios and
46 dynamics of risk among racial and ethnic groups.

47 **DESIGN:** A retrospective cohort of COVID-19 patients diagnosed between March 1 and August 20, 2020. Fully
48 adjusted logistical regression models for hospitalization, severe disease and mortality outcomes across 1-the
49 entire cohort and 2- within self-reported race/ethnicity groups.

50 **SETTING:** Three sites of the NewYork-Presbyterian health care system serving all boroughs of New York City.
51 Data was obtained through automated data abstraction from electronic medical records.

52 **PARTICIPANTS:** During the study timeframe, 110,498 individuals were tested for SARS-CoV-2 in the NewYork-
53 Presbyterian health care system; 11,930 patients were confirmed for COVID-19 by RT-PCR or covid-19 clinical
54 diagnosis.

55 **MAIN OUTCOMES AND MEASURES:** The predictors of interest were patient race/ethnicity, and covariates
56 included demographics, comorbidities, and census tract neighborhood socio-economic status. The outcomes of
57 interest were COVID-19 hospitalization, severe disease, and death.

58 **RESULTS:** Of confirmed COVID-19 patients, 4,895 were hospitalized, 1,070 developed severe disease and 1,654
59 suffered COVID-19 related death. Clinical factors had stronger impacts than social determinants and several
60 showed race-group specificities, which varied among outcomes. The most significant factors in our all-patients
61 models included: age over 80 (OR=5.78, $p= 2.29 \times 10^{-24}$) and hypertension (OR=1.89, $p=1.26 \times 10^{-10}$) having the
62 highest impact on hospitalization, while Type 2 Diabetes was associated with all three outcomes (hospitalization:

63 OR=1.48, $p=1.39 \times 10^{-04}$; severe disease: OR=1.46, $p=4.47 \times 10^{-09}$; mortality: OR=1.27, $p=0.001$). In race-specific
64 models, COPD increased risk of hospitalization only in Non-Hispanics (NH)-Whites (OR=2.70, $p=0.009$). Obesity
65 (BMI 30+) showed race-specific risk with severe disease NH-Whites (OR=1.48, $p=0.038$) and NH-Blacks (OR=1.77,
66 $p=0.025$). For mortality, Cancer was the only risk factor in Hispanics (OR=1.97, $p=0.043$), and heart failure was
67 only a risk in NH-Asians (OR=2.62, $p=0.001$).

68 **CONCLUSIONS AND RELEVANCE.** Comorbidities were more influential on COVID-19 outcomes than social
69 determinants, suggesting clinical factors are more predictive of adverse trajectory than social factors.

70

71 **INTRODUCTION**

72 As of March 1, 2021, over 25 million coronavirus disease 2019 (COVID-19) cases have resulted in more than
73 500,000 deaths in the US. New York City (NYC) bore the brunt of the initial wave of infections with over 60,000
74 hospitalized cases and 10,000 deaths. Early in the crisis, it was clear that the incidence and outcome of COVID-19
75 infection differed vastly across patient populations. For example, older, male individuals with clinical co-
76 morbidities were shown to have worst outcomes than other groups¹. Ethnic and racial disparities in COVID-19
77 incidence and mortality also emerged in the US, where African American and Hispanic groups suffer
78 disproportionate incidence¹⁻⁵ and related death rates^{3,5}. These disparities may partially be explained by the
79 prevalence of comorbidities^{5,6}. However, recent investigations also suggest factors related to socioeconomic
80 status, such as employment and neighborhood density may play a role. Specifically, individuals living in higher
81 poverty and/or higher density areas were found to have higher COVID-19 infection rates, and residents in
82 COVID-19 hotspot areas were more likely to be younger, non-White, and have multiple co-morbidities,
83 compared to those who live in areas with lower risk of COVID-19 infection^{2,3,5}. The influence of poverty levels on
84 death outcomes are less clear, as some groups have reported higher death rates among individuals living in
85 higher poverty areas³, and others have reported a protective survival benefit among those in high poverty⁵ areas
86 or with non-White race/ethnicity⁶.

87 Despite many studies, a full understanding of specific factors of differential risk of disease trajectory across
88 diverse populations is still unclear. In particular, the role of social determinants of health (SDOH) when
89 controlled for clinical factors remains unclear. Most importantly, risk analyses have only been performed in the
90 general population and variance of risk factors among racial/ethnic group is unknown. Finally, whether risk
91 factor status can be developed into predictive outcome models to be used in high-risk patient triage to prevent
92 severe outcomes is equally unknown.

93 To address these issues, we used a real-world dataset (RWD) from NewYork-Presbyterian (NYP) Hospitals, a
94 healthcare system that provides primary and specialty clinical services for a diverse population of patients from

95 all five NYC boroughs, and was a major contributor to disease management during the initial COVID-19 spike in
96 NYC. Specifically, we sought to quantify the relative impact of risk factors related to patient demographics, social
97 determinants of health and clinical comorbidities on COVID-19 disease outcomes across patient groups. We
98 hypothesized that multivariate modeling in a large and diverse cohort would reveal the contribution of each
99 factor and help identify specific reasons why certain population groups have had worse outcomes, compared to
100 others. We anticipate these findings will facilitate better triage and prioritization of patients with high-risk of
101 disease severity and/or mortality.

102

103 **METHODS**

104 **Setting.** Weill Cornell Medicine (WCM), located in New York City, has over 20 outpatient sites across the city.
105 WCM physicians also hold admitting privileges at NewYork-Presbyterian (NYP) Hospital. WCM is an academic
106 medical center, approximately 1,000 attending physicians and over 250,000 patient visits per year. This study
107 was approved by the WCM Institutional Review Board (IRB).

108 **Data Sources - *Integrated Data Registry/Repository (IDR)*.** Our system clinicians used the EpicCare® Ambulatory
109 Electronic Health Record (EHR) system in conjunction with AllScripts Acute EHR system to document clinical care
110 in the outpatient and inpatient settings, respectively. We aggregated raw data from all these disparate EHR
111 source systems which then were transformed and modeled in WCM's COVID centralized data repository, COVID
112 Institutional Data Repository (IDR)⁷. The NYP hospital system catchment includes patients from all NYC
113 boroughs. Of the 100,000+ patients in our COVID IDR, which represented three NYP sites – designated as Sites 1,
114 2 and 3, we included patients with confirmed COVID-19.

115 **Participants.** COVID-19 confirmed cases were defined as patients that had nasopharyngeal swab PCR testing
116 performed with “Detected” results or those who received a COVID-19 ICD-10 diagnosis (excluding those that
117 were also confirmed as “Not Detected” by PCR assay). We utilized self-reported race and ethnicity subcategories
118 that were recorded as separate fields in the EHR. Patients who identified as “Hispanic or Latino or Spanish

119 Origin” ethnicity were designated as “Hispanic or Latino or Spanish Origin”, and all others were classified as the
120 Non-Hispanic race categories they identified. Thus, we designated patients to mutually exclusive self-identified
121 categories: Hispanic (HISP); Non-Hispanic Black (NH-Black); Non-Hispanic White (NH-White); and Non-Hispanic
122 Asian (NH-Asian) (Table 1.)

123 For the race-specific regression analyses, age was binned as either less than 50 years or 50+ years. We extracted
124 co-morbidity data using ICD-10 Codes as documented in the EHR (ICD10 codes: type 2 diabetes (T2D), E11; type
125 1 diabetes (T1D), E10; hypertension (HTN), I10; heart failure (HF), I50; cardiovascular disease (CVD), I25; cancer,
126 C80; chronic obstructive pulmonary disease (COPD), J44; asthma, J45; depression, F32). Body Mass Index (BMI)
127 within three months of COVID-19 diagnosis was extracted from the EHR. This variable was then categorized as
128 “<30 (non-obese)” or “30+ (obese)”. Hospital Site was defined as the location at which the patient was either
129 tested for COVID-19 or first received a COVID-19 ICD-10 diagnosis. Social Determinants of Health (SDOH) were
130 quantified using the Neighborhood Deprivation Index (NDI), calculated as previously described in Messer et al⁸
131 using census tract data summarizing five socio-demographic domains associated with health outcomes,
132 including income/poverty, education, employment, housing, and occupation, using principal components
133 analysis. NDI was assigned to each patient using the longitude and latitude coordinates of street addresses, and
134 this value was categorized into three ranges: [-2.07, -0.696] (low), (-0.696, 0.544] (medium), and (0.544, 3.09]
135 (high). We classified available insurance data for patients as either Commercial, Medicare, Medicaid or Hybrid
136 (i.e., different primary and secondary insurances). The severe disease category included hospitalized patients
137 who were intubated, were on vasopressors, had a diagnosis of acute respiratory distress syndrome (ARDS),
138 and/or were on kidney dialysis (ICD10 codes: ARDS, J80; renal dialysis, Z99.2 or Z49). These criteria were based
139 on criteria from the WHO Ordinal Scale for Clinical Improvement, as well as clinical manifestation of severe
140 disease in COVID-19 patients⁹.

141 **Statistical Analysis.** Demographic, clinical, and socioeconomic characteristics of our cohort were summarized
142 using descriptive statistics for each model group; including COVID-19 positive, hospitalized, severe disease, and

143 deceased subsets (**Table 1**). Multivariate logistic regression models were fit from the covariates against each of
144 the outcome variables (hospitalization, severe disease, and death). The coefficients and standard error
145 generated from this regression were converted to odds ratios, and p-values were generated using a Wald Test.
146 The predictive capacity of each model was determined by re-fitting the model on 80% of the data, and then
147 assessing its performance on the remaining 20%. The data was split such that the case to control ratio in each
148 subset was equal. The true positive and false positive rates generated when comparing the predictions to true
149 outcomes at a series of different prediction cut-offs were organized into a Receiver Operator Curve (ROC). The
150 area under the ROC curve (AUC) was extracted to assess predictive ability. For all models, patients with
151 unknown or missing data were included in the overall cohort to avoid bias, but the “Unknown” categories for
152 each covariate was excluded from data visualization. R (version 4.0.1) was utilized for all computations.

153

154 **RESULTS**

155 **Baseline Sample Clinical and Demographic Characteristics:**

156 Our patient cohort consists of 11,930 COVID-19 patients between March and August of 2020, which was 12% of
157 patients tested in that period (**Supplemental Figure 1**). Nearly 50% of COVID-19 positive patients (n=4,895) were
158 hospitalized in the NYP system in this timeframe. Of these hospitalized patients, 22% (n=1,070) developed
159 severe disease, defined by intubation with vasopressor use, ARDS diagnosis, and/or kidney dialysis
160 (**Supplemental Figure 1**). Deceased patients (n=1,654) were defined by the vital status in the EHR. **Supplemental**
161 **Table 1** contains descriptive statistics of all sixteen variables (described in Methods) across the three outcome
162 groups.

163 We utilized patients’ self-reported race and ethnicity identities, as described in methods. Our diverse multi-
164 ethnic cohort resides across all five NYC boroughs (**Figure 1A**) and received COVID-19 tests at one of three sites
165 in our network (**Supplemental Figure 3**). There is similar age distribution among race/ethnicity groups (**Figure**
166 **1B**), comparable to the general population (overall mean age = 57) (**Supplemental Figure 2**). More than 80% of

167 racial/ethnic minority patients have moderate to high NDI (**Figure 1C and Supplemental Figure 4**), indicating
168 they live in more deprived neighborhoods. Most patients reside in neighborhoods that are enriched for their
169 same race (**Figure 1A**).

170 **Factors that impact risk of hospitalization among COVID-19 patients**

171 We first determined factors associated with hospitalization of COVID-19 patients. Specifically, we quantified
172 the risk of eighteen factors using a fully adjusted multivariate model (**Figure 2A**). Similar to other studies¹⁰, we
173 found risk of hospitalization increased with age, starting at age 30 (OR=1.79, $p=7.91 \times 10^{-06}$) up to 80+ (OR=5.78,
174 $p=2.29 \times 10^{-24}$) and male sex (OR=1.50, $p=5.92 \times 10^{-08}$). Interestingly, among race/ethnicity variables, NH-Black and
175 NH-Other patients had the strongest association in the hospitalization model, but for being less likely to be
176 hospitalized (NH-B: OR=0.60, $p=1.15 \times 10^{-04}$; NH-O: OR=0.73, $p=0.027$). For clinical factors within the entire cohort,
177 all co-variates were associated with increased odds of hospitalization except T1D, asthma, and obesity (HTN:
178 OR=1.89, $p=1.26 \times 10^{-10}$; Depression: OR=1.76, $p=3.07 \times 10^{-05}$; T2D: OR=1.48, $p=1.39 \times 10^{-04}$; HF: OR=1.42, $p=0.021$;
179 Cancer: OR=0.55, $p=0.028$; COPD: OR=1.49, $p=0.045$; CVD: OR=1.32, $p=0.049$). Lastly, there was an intriguingly
180 high association with specific hospital testing sites, with Site 3 having the highest odds of hospitalization
181 (OR=3.32, $p=4.72 \times 10^{-22}$). Insurance type was also shown to be significantly associated with hospitalization
182 (Medicaid: OR=1.89, $p=0.016$; Medicare: OR=1.61, $p=0.033$). Overall, the clinical factors had the highest
183 magnitude of risk, compared to other determinant types.

184 **Factors that impact risk of disease severity among COVID-19 patients:**

185 We then sought to determine the magnitude of risk for factors associated with severe disease among COVID-19
186 hospitalized patients, using a similar model approach as our hospitalization risk model (**Figure 2B**). Of the
187 demographic factors, males were 1.58 times more likely to develop severe disease than females ($p=4.47 \times 10^{-09}$).
188 Risk of severe disease increased with age, starting at ages 40-49 (OR=2.30, $p=0.002$) up to ages 70-79 (OR: 3.31,
189 $p=4.48 \times 10^{-06}$). NH-Asian and Hispanic patients were more likely to develop severe disease (NH-A: OR= 1.79, $p=$
190 4.47×10^{-09} ; HISP: OR=1.33, $p=0.013$), compared to NH-White patients. The most significant clinical factors

191 associated with severe COVID-19 were type 2 diabetes (OR=1.46, $p=4.47 \times 10^{-09}$), heart failure (OR=1.57,
192 $p=4.47 \times 10^{-09}$), and obesity (OR=1.19, $p=0.042$). Social determinants, such as NDI and insurance status had no
193 significant impact on severe disease outcome overall. A significant association with specific hospital testing site
194 was observed (Site 2: OR=0.49, $p=4.47 \times 10^{-09}$; Site 3: OR: 0.47, $p=4.47 \times 10^{-09}$). Thus, in this cohort, the sites
195 associated with higher likelihood of hospitalization have a lower likelihood of developing severe disease after
196 hospitalization.

197 **Factors that increase risk of mortality among COVID-19 patients:**

198 Lastly, we measured the impact of risk factors leading to death among confirmed COVID-19 patients, as opposed
199 to recovering from infection (**Figure 2C**). Our third model includes 1,654 deceased patients whose death
200 occurred following a COVID-19 diagnosis, without regard to hospitalization. Upon model fitting, we found that
201 age was the most significant demographic factor associated with COVID-19 mortality, in a “dose-dependent”
202 fashion, starting at ages 40-49 (OR=4.33, $p=5.25 \times 10^{-05}$) ranging to age 80+ (OR=66.5, $p=1.70 \times 10^{-34}$). Male sex
203 (OR=1.53, $p=3.82 \times 10^{-12}$), Hispanic ethnicity (OR=1.26, $p=0.007$), and NH-Asian race (OR=1.27, $p=0.03$) were also
204 associated with mortality among COVID-19 patients. Type 2 diabetes (OR=1.27, $p=0.001$), type 1 diabetes
205 (OR=1.78, $p=0.02$), hypertension (OR=0.86, $p=0.033$), heart failure (OR=1.43, $p=9.06 \times 10^{-05}$), cancer (OR=1.51,
206 $p=0.033$), and obesity (OR=1.20, $p=0.025$) were the all associated with odds of mortality. NDI, smoking status,
207 and insurance type were not significant risk factors, after adjusting for the other variables.

208 **Differences in COVID-19 outcome risk factors across Patient Race/Ethnicity**

209 Several reports have indicated the disproportionately higher rates of COVID-19 related hospitalization and
210 mortality in the US Hispanic and NH-Black populations, compared to NH-White¹⁻⁶. Our full cohort analyses have
211 identified risk factors that are independent of race groups. To address whether some risk factors are more
212 important in certain race groups, we fit new models within specific race groups, including NH-Black, NH-Asian,
213 NH-White and Hispanic categories (**Figure 3**). Surprisingly, we found few risk factors that were consistent across
214 race groups, with hypertension being the sole risk factor for hospitalization across all race groups (**Figure 3A**).

215 **NH-W: OR=2.17, p=1.31x10⁻⁰⁴; NH-B: OR=1.68, p=0.031; NH-A: OR=2.68, p=0.004; HISP: OR=2.09, p=3.44x10⁻⁰⁴.**

216 Age (NH-W: OR=3.17, p=4.87x10⁻⁰⁹; NH-A: OR=3.42, p=7.97x10⁻⁰⁶; HISP: OR=1.61, p=0.04) and depression (NH-

217 W: OR=1.85, p=0.01; NH-A: OR=12, p=0.02; HISP: OR=1.93, p=0.018) were significant risk factors in all but the

218 NH-Black population (**Figure 3A**). Male sex was associated with higher odds of hospitalization among NH-Asian

219 patients (NH-A: OR=2.62, p=1.12x10⁻⁰⁴) and Hispanic patients (HISP: OR=1.56, p=0.002). We also found several

220 clinical risk factors for hospitalization were only significant in specific race groups. For example, type 2 diabetes

221 was only a significant risk factor in NH-Black patients (NH-B: OR=1.81, p=0.014), heart failure was a significant

222 risk factor in NH-Asian patients (NH-A: OR=4.31, p=0.037), COPD was a significant risk factor in NH-White

223 patients (NH-W: OR=2.70, p=0.009). Interestingly, several variables that were not significantly associated with

224 odds of hospitalization in the overall cohort model were identified as race-specific risk factors. These included

225 cancer diagnoses in NH-Black and Hispanic patients (NH-B: OR=0.20, p= 0.005; HISP: OR=0.25, p= 0.022),

226 smoking status (NH-A: OR_{Quit}=0.23, p_{Quit}=0.008; NH-A: OR_{Yes (current or passive)}=0.02, p_{Yes (current or passive)}=0.006; HISP:

227 OR_{Yes (current or passive)}=0.32, p_{Yes (current or passive)}=0.046), and insurance type (NH-W: OR_{Hybrid}=0.41, p_{Hybrid}=0.034; NH-B:

228 OR_{Hybrid}=3.11, p_{Hybrid}=0.049, OR_{Medicaid}=4.56, p_{Medicaid}=0.028; HISP: OR_{Medicare}=3.46, p_{Medicare}=0.009) (**Figure 3B**).

229 Lastly, NDI had differential impact across race groups (**Supplemental Figure 5**) where moderate NDI was

230 associated with hospitalization in NH-Asian patients (NH-A: OR: 4.00, p=0.006) and high NDI scores were

231 associated with hospitalization in NH-White patients (NH-W: OR=2.10, p=0.032).

232 For demographic variables in the race-group severe COVID-19 disease model (**Figure 3B**), surprisingly, age was

233 only a significant severity risk factor among NH-Asian and Hispanic patients (NH-A: OR=2.66, p= 0.002; HISP:

234 OR=1.92, p=9.77x10⁻⁰⁵). Similarly, male sex was a significant severe disease risk factor for all but NH-Black

235 patients, with the highest impact in Hispanic patients (NH-W: OR=1.49, p=0.02; NH-A: OR=1.56, p= 0.022; HISP:

236 OR=2.14, p= 1.50x10⁻⁰⁷). Across clinical factors, type 2 diabetes was a significant factor only in White patients

237 (W: OR=1.84, p=0.001), cardiovascular disease was a race-specific risk factor of severe disease among NH-White

238 patients (NH-W: OR=1.52, p=0.043), and obesity was a significant clinical risk factor among NH-White and NH-

239 Black patients (NH-W: OR: 1.48, $p=0.038$; NH-B: OR=1.77, $p=0.025$). Odds of severe disease was lower among
240 patients tested at hospital site 2 (NH-W: OR=0.34, $p=0.011$; NH-A: 0.42, $p=0.006$; HISP: OR=0.42, $p=0.006$) and
241 site 3 (NH-W: OR=0.29, $p=2.06 \times 10^{-06}$; HISP: OR=0.32, $p=5.38 \times 10^{-07}$).

242 Finally, the only significant risk factors for increased odds of mortality among all race groups was age (NH-W:
243 OR=9.44, $p=5.45 \times 10^{-17}$; NH-B: OR=6.83, $p=1.60 \times 10^{-10}$; NH-A: OR=11.40, $p=2.00 \times 10^{-09}$; HISP: OR=6.09, $p=9.52 \times 10^{-27}$) (Figure 3C). Male sex was associated with increased odds of mortality only among Hispanic patients (HISP:
245 OR=1.42, $p=0.01$). Among clinical characteristics, only heart failure (NH-W: OR=1.99, $p=5.76 \times 10^{-05}$; NH-A:
246 OR=2.62, $p=0.001$) and cancer (NH-W: OR=2.42, $p=0.004$; HISP: OR=1.97, $p=0.043$) were shown to be statistically
247 significant risk factors of mortality within race groups. Among the Hispanic population, smoking status (HISP:
248 OR_{Quit}=1.49, $p=0.034$) and insurance type (HISP: OR_{Hybrid}=2.80, $p=0.012$) were shown to be associated with
249 increased odds of mortality. Surprisingly, NDI had relatively small impact in any race-group's mortality risk,
250 which suggests social determinants were not the primary drivers of mortality disparities, when considered in
251 context with the risk associated with clinical factors.

252 **Predictive efficacy of the multivariate models:**

253 We hypothesized that the multivariate logistic regression models could be developed into predictive models.
254 The predictive capacity of each model was assessed using the area under the receiver operator curve (AUC) as
255 discussed in Methods. We found that all models had some predictive power; however, accuracy varied among
256 outcomes (Figure 4). Overall, the hospitalization model was highly predictive in the overall cohort (AUC=0.969)
257 (Figure 4Ai), while the severe disease and death models performed with lower accuracy (AUC=0.698 and
258 AUC=0.812, respectively) (Figure 4A-vi and xi).

259 Using models fitted with only specific race groups, we found that accuracy of the hospitalization models remains
260 high across all 4 racial/ethnic groups, with NH-White group having the highest accuracy (NH-W: AUC=0.970,
261 Figure 4A-v) and the NH-Asian group having the lowest accuracy (NH-A: AUC=0.942, Figure 4A-iii). Severe

262 disease risk models were least accurate overall, ranging from the lowest AUC in the Hispanic group (HISP:
263 AUC=0.659, **Figure 4A-ix**) to the highest AUC among the NH-Asian group (NH-A: AUC=0.691, **Figure 4A-viii**). The
264 death risk models performed with high accuracy, but to varying degrees across race groups. The NH-Black group
265 had the lowest accuracy with our model (AUC=0.719, **Figure 4A-xii**), while NH-White patients had the highest
266 accuracy (AUC=0.770, **Figure 4A-xv**). The attenuated accuracy of our models in specific race groups likely reflects
267 the differential influence of each factor toward disease outcomes among race groups.

268

269 **DISCUSSION**

270 Overall, our findings in this large New York City cohort (n=11,930) of COVID-19 patients, diagnosed between
271 March and August of 2020, present a unique perspective of relative risks impacting disparities. While we confirm
272 the vulnerabilities of certain race/ethnicity groups, we also uncovered that typical demographic and social
273 determinants that were implicated in COVID-19 disparities were not as impactful as the clinical factors.
274 Interestingly, we also observed differences in which factors associated with severe disease risk of hospitalized
275 patients among race-groups.

276 Several of our results agree with recently published reports from other academic health centers¹⁰ and the
277 Veteran Affairs Hospitals^{6,11}, including the impact of age and sex on hospitalization and severe outcomes. One
278 consistently surprising result among these studies was that NH-Black patients were less likely to be hospitalized
279 after a COVID-19 diagnosis, compared to other race groups. This is despite public health data reports indicating
280 disproportionately high rates of NH-Black hospitalization, nation-wide. There may be several reasons for this
281 difference. First, NH-Black patients in our cohort, like other large academic institutions, may represent a specific
282 subset of the greater NYC NH-Black population, which illustrates the jeopardy in generalizations of race-group
283 risk when these populations are distributed across broad social strata. Secondly, based on their residential
284 locations, NH-Black patients in our cohort were likely to travel beyond their boroughs for access to our hospital

285 testing sites (**Supplemental Figure 3**). At the peak of COVID-19, this was likely necessary because of the limited
286 access to testing in certain areas. This also presents a possibility that, if their COVID-19 condition escalated to
287 require urgent care or hospitalization, they may have sought care in alternative healthcare systems in closer
288 proximity to their residence, which we would not be able to track in our cohort data sources.

289 Our study describes race-group specific risk factors, compared to other publications that generalize risk factors
290 to be similar across the entire population at-large. For instance, in our hospitalized severe disease model, we
291 found that age was only significant in the NH-Asian and Hispanic groups, while obesity was the only significant
292 factor for severe disease in NH-Black patients. This suggests that generalized use of risk factors for prioritization
293 models across the population at-large would inherently benefit certain race/ethnicity groups more than others,
294 de-prioritizing the high-risk status of specific race groups, leading to greater disparities across the continuum of
295 disease management and prevention.

296 The results of this study can have immediate transformative clinical impact. Our use of fully adjusted
297 multivariate models assesses probability of severe outcomes in the mixed context of social, demographic, and
298 clinical factors, providing a comprehensive approach to reduce racial disparities through prioritization of risk.
299 Ranked severe disease risk factors, by effect size across all models (**Figure 4B**), gives insight to differential
300 disease course as well as a potential plan of prioritization. First, the impact of Type 2 Diabetes on severe disease
301 was significant in NH-White patients, but not in NH-Black patients after adjusting for BMI. This suggests that
302 while this disorder is typically more prevalent in the NH-Black population, obesity can influence its role in
303 COVID-19 outcomes. The differential impact of clinical risk factors across race groups may indicate population-
304 level differences in chronic disease management/status may influence physiology and biological determinants of
305 COVID-19 disease trajectory. Next, we found that hospitalization is a benefit to outcome. Specifically, in the
306 hospital sites where hospitalization was more likely (**Figure 2A**), severe disease was less likely (**Figure 2B**). This
307 may suggest that severe disease can be avoided, for patients treated at certain sites, if hospitalization occurs.

308 This finding was consistent for most race groups, excluding NH-Black patients (**Figures 3B and 4B**) who were also
309 found to be less likely to be hospitalized.

310 There are important limitations to consider in interpreting our results. First, our study has the typical limitations
311 RWD data collection from EHR systems. However, use of RWD is advantageous because in the acute and fast-
312 paced context of a deadly pandemic, urgency of saving lives precedes the priority of appropriate study designs
313 and specialized data collection. However, completeness and accuracy of some variables that are typically part of
314 self-reported intake surveys, including smoking status, medical history and loss of follow-up is debatable. The
315 NDI application of SDOH is a powerful indicator of patients' 'placemats' or integrated neighborhood-level
316 determinants, which is often more predictive of health outcomes than direct measures of socioeconomic status.
317 However, the NDI does not directly measure some key factors, such as systemic racism or social capital¹² though
318 the measures included in calculating the NDI are certainly influenced by these factors. (**Supplemental Figure 3**).

319 **CONCLUSIONS**

320 In conclusion, our large cohort study of over 11,000 COVID-19 patients reveals a slightly different perspective
321 than what has previously been reported nationwide. We find that clinical comorbidity factors, such as Type 2
322 Diabetes, are more significant to COVID-19 outcome than social determinants. We also found risk factors are not
323 generalizable across all race groups. Finally, we demonstrate that poor outcomes such as hospitalization risk can
324 be predicted reliably using pre-infection data. This indicates that individuals at risk can in theory be reliably
325 identified and triaged for highly effective care including vaccination.

326

327 **Acknowledgements:** This work was funded by several awards to the co-authors; including a COVID-19 Research
328 Grant from the Weill Cornell Office of Research (MBD), a NIH (National Institutes of Health) COVID-19 Disparities
329 Supplement for the WCM CTSC (Clinical and Translational Science Center) NIH UL1TR002384 grant (JIM, OE, SI,
330 MBD, JM).

331 **TABLES**

332 **Supplemental Table 1. Descriptive statistics of demographic, clinical, and socioeconomic factors of confirmed**
333 **COVID-19 patients.** Analysis of electronic health record data from New York-Presbyterian from March 1, 2020 to
334 August 20, 2020.

335 **FIGURE TITLES AND LEGENDS:**

336 **Figure 1. Residential area, NDI and race distribution of NYP COVID-19 patients are representative of general**
337 **population.** (A) Geographic distribution of COVID-19 patients by latitude-longitude coordinates, color-coded by
338 self-reported race/ethnicity groups. Inset indicates the prevalence of specific race groups across the patient's
339 neighborhoods, based on census-track data. (B) Age distribution of cohort, by race. Bars indicate the mean age
340 for each race group. Color coding of points indicate self-reported sex (Blue=Female, Red=Male). (C) Distribution
341 of Neighborhood Deprivation Index (NDI) scores, by race. Column groups are based on our severe disease status
342 model and mortality outcome models.

343 **Figure 2. Fully adjusted multivariate models of hospitalization, severe disease, and mortality outcomes among**
344 **confirmed COVID-19 patients.** (A) Analysis of odds of hospitalization among all confirmed COVID-19 patients. (B)
345 Analysis of odds of severe disease among hospitalization confirmed COVID-19 patients. (C) Analysis of odds of
346 mortality among all confirmed COVID-19 patients. Analyses conducted using electronic health record data from
347 NewYork-Presbyterian from March 1, 2020 to August 20, 2020. Odds Ratios shown with 95% confidence
348 intervals in parentheses. NH = Non-Hispanic; T2D = Type 2 Diabetes; T1D = Type 1 Diabetes; HTN =
349 Hypertension; HF = Heart Failure; CVD = Cardiovascular Disease; COPD = Chronic Obstructive Pulmonary Disease.

350 **Figure 3. Fully adjusted multivariate models of hospitalization, severe disease, and mortality outcomes among**
351 **confirmed COVID-19 patients, by race/ethnicity.** (A) Odds of hospitalization among confirmed COVID-19, by
352 race/ethnicity group. (B) Odds of severe disease among hospitalization confirmed COVID-19 patients, by
353 race/ethnicity group. (C) Odds of mortality among all confirmed COVID-19 patients, by race/ethnicity group.

354 Analyses conducted using electronic health record data from NewYork-Presbyterian from March 1, 2020 to
355 August 20, 2020. Odds Ratios shown with 95% confidence intervals in parentheses. NH = Non-Hispanic; T2D =
356 Type 2 Diabetes; T1D = Type 1 Diabetes; HTN = Hypertension; HF = Heart Failure; CVD = Cardiovascular Disease;
357 COPD = Chronic Obstructive Pulmonary Disease.

358 **Figure 4A. Performance of Risk Models and Ranked Risk Factors. A.** Receiver operating characteristic (ROC)
359 curves and areas under the curves (AUC) for each fully adjusted multivariate regression model: (i.-v) odds of
360 hospitalization among confirmed COVID-19 patients overall and by race groups; (vi-x) odds of severe disease
361 among hospitalized confirmed COVID-19 patients overall and by race groups; (xi-xv) odds of mortality among
362 confirmed COVID-19 patients overall and by race groups. NH = Non-Hispanic.

363 **Figure 4B. Ranked risk factors in severe disease models.** Combined significant severe disease risk factors
364 identified in the all-inclusive models and stratified within race groups. Bar graphs represent estimated effect
365 measurements for indicated subset and factors, with Standard Error bars. P-values are color coded with
366 increasing significance, red=most significant, blue=least significant.

367 **Supplemental Figure 1. Study Flow Chart.** Patient population in New York Presbyterian COVID-19 Data
368 repository. NYP = NewYork-Presbyterian; IDR = Institutional Data Repository; EHR = Electronic Health Record.

369 **Supplemental Table 1. Descriptive statistics of demographic, clinical, and socioeconomic factors of confirmed**
370 **COVID-19 patients.** Analysis of electronic health record data from NewYork-Presbyterian from March 1, 2020 to
371 August 20, 2020.

372 Supplemental Figure 2. Age distribution across the full dataset and individual race groups are comparable to the
373 NYC demographics. (A) Histogram of the age groupings in the COVID IDR indicate that ~60% of our population
374 falls within the age range of 20-65. (B) This corresponds to the 2019 NYC census where 58% of the population is
375 within the age range of 18-65 (panel B). (C) Age distribution across each race group shows correlated average

376 age around 55-57, except for NH-Asians having a slightly higher average age. (D) COVID IDR age groups by
377 decade show the generally even spread of age, indicating little to no bias in our dataset.

378 **Supplemental Figure 3. Testing sites compared to residential locations of patients.** Residential map of patients,
379 color coded for the testing site where they received a COVID-19 test result. Most of the patients were tested in
380 the vicinity of their residents, though a significant proportion tested in different boroughs than where they
381 resided. The majority of Brooklyn residents tested at Site 1.

382 **Supplemental Figure 4. Box plots of NDI distribution by race/ethnicity for primary outcomes of interest:** (A)
383 Confirmed COVID-19 patients; (B) Hospitalized confirmed COVID-19 patients; (C) Severe disease among
384 hospitalized confirmed COVID-19 patients; (D) Deceased confirmed COVID-19 patients. NH = Non-Hispanic.

385 **Supplemental Figure 5. Distribution of NDI scores for patient race groups by hospital testing sites.** Boxplots
386 indicate the NDI range and average for each race group that utilized the indicated hospital sites. Floating blue
387 bar indicates hospital group NDI score average. This indicates that each hospital has a range of patients from
388 areas with varying NDI. Overall, Site 2 had the highest NDI and Site 1 had the lowest average NDI. The contrast
389 of white patients at Site 2 indicates that NDI can vary greatly between race groups in the same region of a
390 borough, a characteristic that reflects red-lining and systemic racial bias.

	Confirmed COVID-19 patients	Confirmed COVID-19 patients admitted to hospital	Severe disease (i.e., Vasopressor, intubation, ARDS, or renal dialysis) among hospitalized confirmed COVID-19 patients	Deaths among confirmed COVID-19 patients
	n=11,930	n=4,895	n=1,070	n=1,654
	n (%)	n (%)	n (%)	n (%)
Age mean, years	57.26	62.48	64.03	74.58
Age Group, years				
<30	1097 (9.20)	236 (4.82)	20 (1.87)	9 (0.54)
30-39	1630 (13.66)	473 (9.66)	57 (5.33)	27 (1.63)
40-49	1549 (12.98)	511 (10.44)	99 (9.25)	57 (3.45)
50-59	1973 (16.54)	816 (16.67)	201 (18.79)	155 (9.37)
60-69	2219 (18.60)	1026 (20.96)	296 (27.66)	302 (18.26)
70-79	1769 (14.83)	891 (18.20)	255 (23.83)	460 (27.81)
80+	1663 (13.94)	933 (19.06)	140 (13.08)	643 (38.88)
Unknown (missing)	30 (0.25)	9 (0.18)	2 (0.19)	1 (0.06)
Sex				
Female	6051 (50.72)	2135 (43.62)	359 (33.55)	686 (41.48)
Male	5845 (48.99)	2750 (56.18)	708 (66.17)	967 (58.46)
Unknown (missing)	34 (0.28)	10 (0.20)	3 (0.28)	1 (0.06)
Race/Ethnicity				
Non-Hispanic White	3004 (25.18)	1099 (22.45)	214 (20.00)	408 (24.67)
Hispanic or Latino or Spanish Origin	3459 (28.99)	1514 (30.93)	315 (29.44)	496 (29.99)
Non-Hispanic American Indian/Alaskan Native/Native Hawaiian/Other Pacific Islander	33 (0.28)	22 (0.45)	4 (0.37)	4 (0.24)
Non-Hispanic Asian	1143 (9.58)	748 (15.28)	192 (17.94)	221 (13.36)
Non-Hispanic Black	1788 (14.99)	520 (10.62)	113 (10.56)	229 (13.85)
Non-Hispanic Other	1089 (9.13)	507 (10.36)	103 (9.63)	144 (8.71)
Unknown (or missing)	1414 (11.85)	485 (9.91)	129 (12.06)	152 (9.19)
Comorbidities				
Type 2 Diabetes (ICD10: E11)	2662 (22.31)	1467 (29.97)	464 (43.36)	563 (34.04)
Type 1 Diabetes (ICD10: E10)	142 (1.19)	78 (1.59)	30 (2.80)	26 (1.57)
Hypertension (ICD10: I10)	4492 (37.65)	2308 (47.15)	648 (60.56)	838 (50.67)
Heart Failure (ICD10: I50)	994 (8.33)	599 (12.24)	202 (18.88)	276 (16.69)
Cardiovascular Disease (ICD10: I25)	1321 (11.07)	788 (16.10)	259 (24.21)	320 (19.35)
Cancer (ICD10: C80)	211 (1.77)	71 (1.45)	22 (2.06)	44 (2.66)
Chronic Obstructive Pulmonary Disease (ICD10: J44)	536 (4.49)	324 (6.62)	90 (8.41)	131 (7.92)
Asthma (ICD10: J45)	1130 (9.47)	439 (8.97)	120 (11.21)	111 (6.71)
Depression (ICD10: F32)	1107 (9.28)	480 (9.81)	117 (10.93)	131 (7.92)
Body Mass Index (BMI)				
<30 (Non-obese)	4918 (41.22)	2806 (57.33)	605 (56.54)	828 (50.06)
30+ (Obese)	2403 (20.14)	1333 (27.23)	312 (29.16)	330 (19.95)
Unknown (missing)	4609 (38.63)	756 (15.44)	153 (14.30)	496 (29.99)
Smoking Status				
Never	1407 (33.67)	873 (17.83)	226 (21.12)	303 (18.32)
Unknown (missing)	6277 (52.62)	3530 (72.11)	687 (64.21)	1140 (68.92)
Passive	15 (0.13)	7 (0.14)	1 (0.09)	3 (0.18)
Quit	1382 (11.58)	431 (8.80)	143 (13.36)	187 (11.31)
Yes (current)	239 (2.00)	54 (1.10)	13 (1.21)	21 (1.27)
Hospital Site				
Site 1	2432 (20.39)	1706 (34.85)	516 (48.50)	266 (16.08)
Site 2	549 (4.60)	443 (9.05)	81 (7.57)	99 (5.99)
Site 3	3164 (26.52)	2614 (53.4)	424 (39.63)	668 (40.39)
Unknown (missing)	5785 (48.49)	132 (2.70)	46 (4.30)	621 (37.55)
Insurance Type				
Commercial	2871 (24.07)	1230 (25.13)	324 (30.28)	217 (13.12)
Hybrid	247 (2.07)	131 (2.68)	37 (3.46)	39 (2.36)
Medicaid	220 (1.84)	156 (3.19)	43 (4.02)	21 (1.27)
Medicare	404 (3.39)	314 (6.41)	102 (9.53)	92 (5.56)
Unknown (missing)	8188 (68.63)	3064 (62.59)	564 (52.71)	1285 (77.69)
Neighborhood Deprivation Index (NDI)				
[-2.07,-0.696]	1788 (14.99)	624 (15.73)	145 (13.55)	217 (13.12)
[-0.696,0.544]	1791 (15.01)	770 (14.65)	188 (17.57)	278 (16.81)
[0.544,3.09]	1783 (14.95)	717 (12.75)	191 (17.85)	260 (15.72)
Unknown (missing)	6568 (55.05)	2784 (56.87)	546 (51.03)	899 (54.35)

391

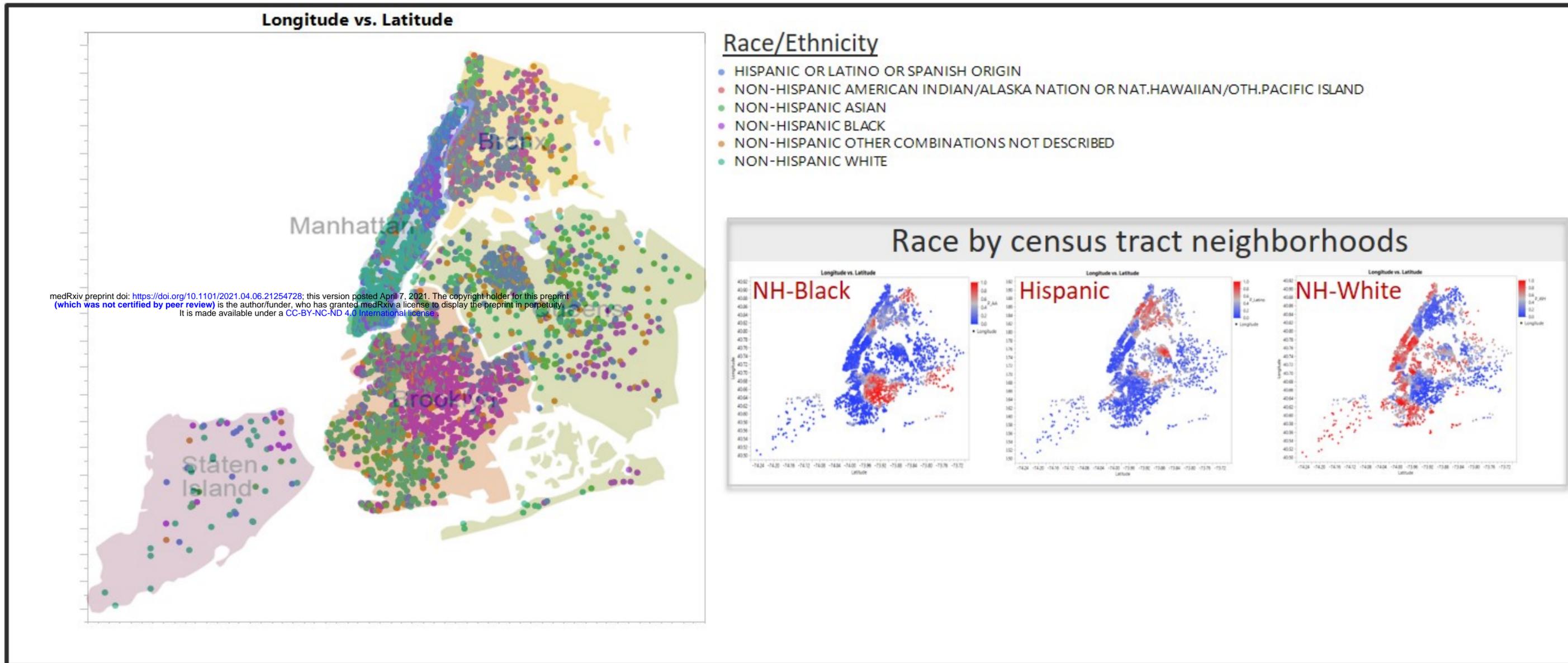
392

393 **References:**

- 394 1. Argenziano MG, Bruce SL, Slater CL, et al. Characterization and clinical course of 1000 patients with
395 coronavirus disease 2019 in New York: retrospective case series. *BMJ*. 2020;369:m1996.
- 396 2. Hanson AE, Hains DS, Schwaderer AL, Starr MC. Variation in COVID-19 Diagnosis by Zip Code and Race
397 and Ethnicity in Indiana. *Front Public Health*. 2020;8:593861.
- 398 3. Chen JT, Krieger N. Revealing the Unequal Burden of COVID-19 by Income, Race/Ethnicity, and
399 Household Crowding: US County Versus Zip Code Analyses. *J Public Health Manag Pract*. 2021;27 Suppl
400 1, COVID-19 and Public Health: Looking Back, Moving Forward:S43-S56.
- 401 4. Munoz-Price LS, Nattinger AB, Rivera F, et al. Racial Disparities in Incidence and Outcomes Among
402 Patients With COVID-19. *JAMA Netw Open*. 2020;3(9):e2021892.
- 403 5. Little C, Alsen M, Barlow J, et al. The Impact of Socioeconomic Status on the Clinical Outcomes of COVID-
404 19; a Retrospective Cohort Study. *J Community Health*. 2021.
- 405 6. Kabarriti R, Brodin NP, Maron MI, et al. Association of Race and Ethnicity With Comorbidities and
406 Survival Among Patients With COVID-19 at an Urban Medical Center in New York. *JAMA Netw Open*.
407 2020;3(9):e2019795.
- 408 7. Sholle ET, Kabariti J, Johnson SB, et al. Secondary Use of Patients' Electronic Records (SUPER): An
409 Approach for Meeting Specific Data Needs of Clinical and Translational Researchers. *AMIA Annu Symp*
410 *Proc*. 2017;2017:1581-1588.
- 411 8. Messer LC, Laraia BA, Kaufman JS, et al. The development of a standardized neighborhood deprivation
412 index. *J Urban Health*. 2006;83(6):1041-1062.
- 413 9. Blueprint WHORD. *Novel Coronavirus: COVID-19 Therapeutic Trial Synopsis*. World Health Organization
414 February 2020 2020.
- 415 10. Ogedegbe G, Ravenell J, Adhikari S, et al. Assessment of Racial/Ethnic Disparities in Hospitalization and
416 Mortality in Patients With COVID-19 in New York City. *JAMA Netw Open*. 2020;3(12):e2026881.
417

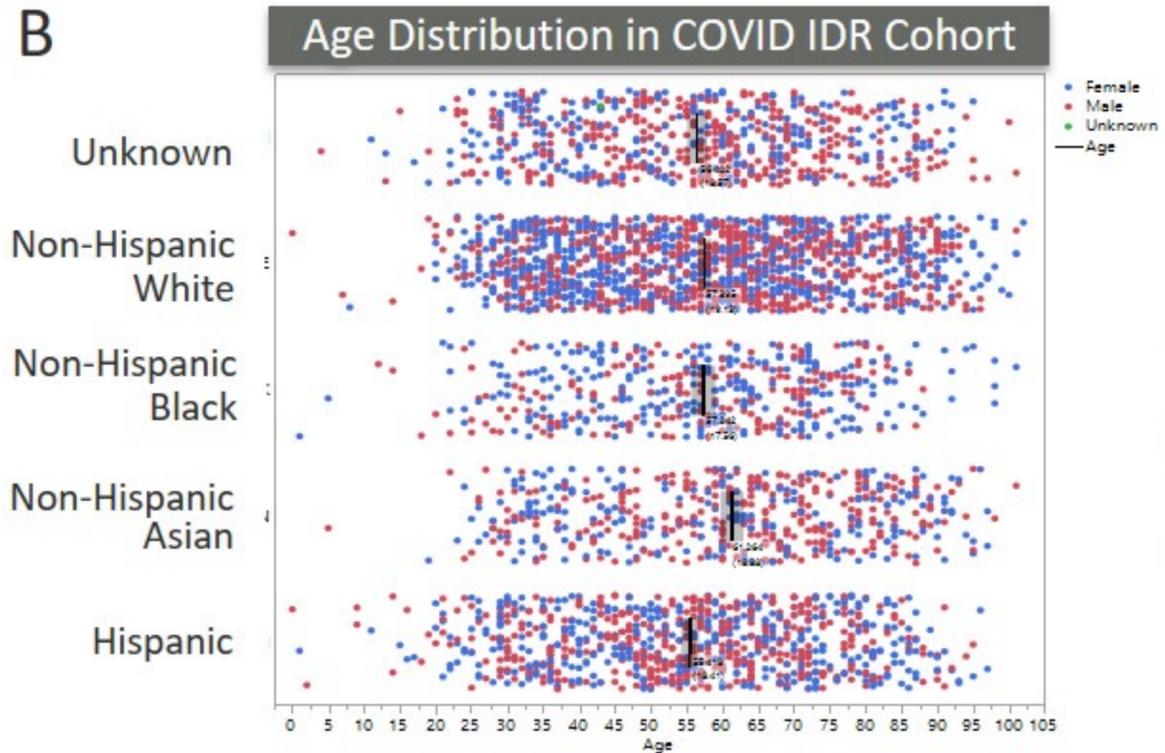
Figure 1. Residential area, NDI and race distribution of NYP COVID patients are representative of general population

A



medRxiv preprint doi: <https://doi.org/10.1101/2021.04.06.21254728>; this version posted April 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

B



C

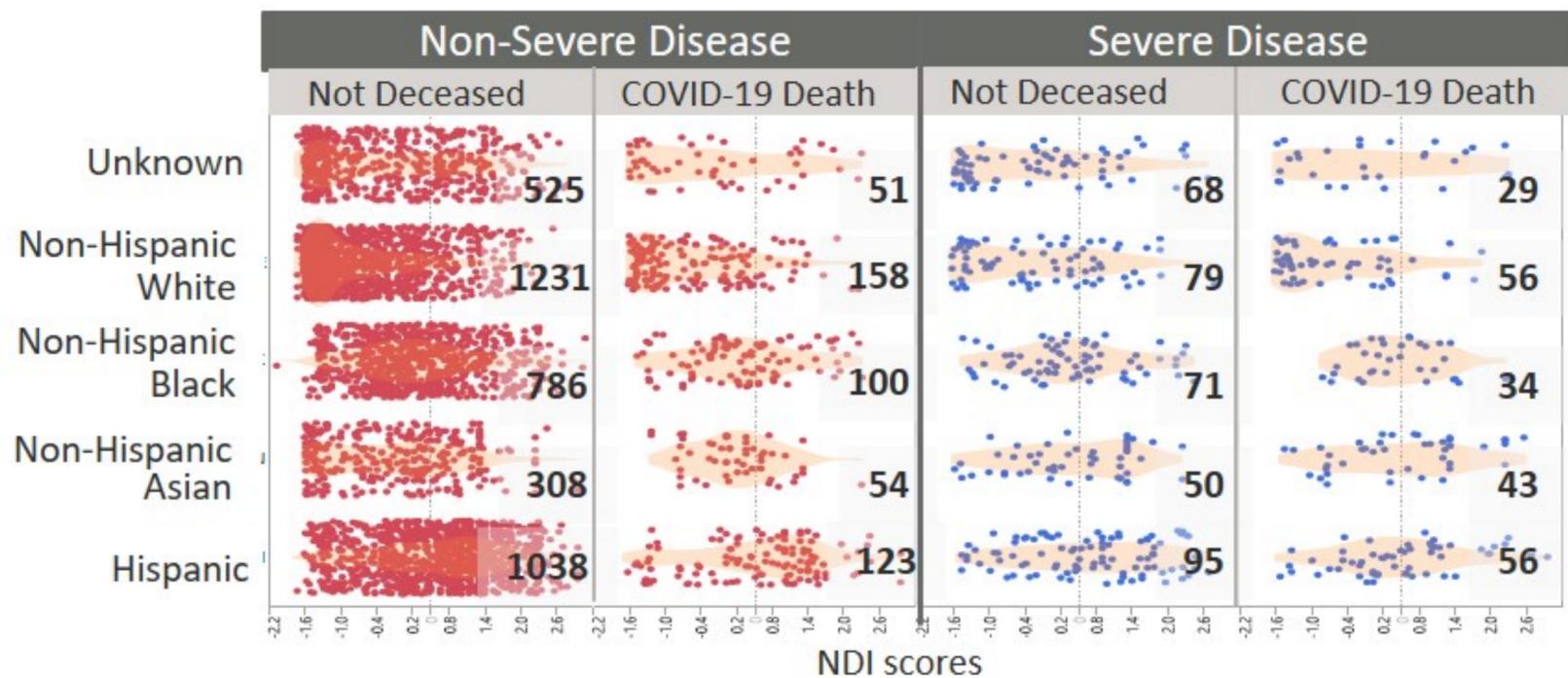
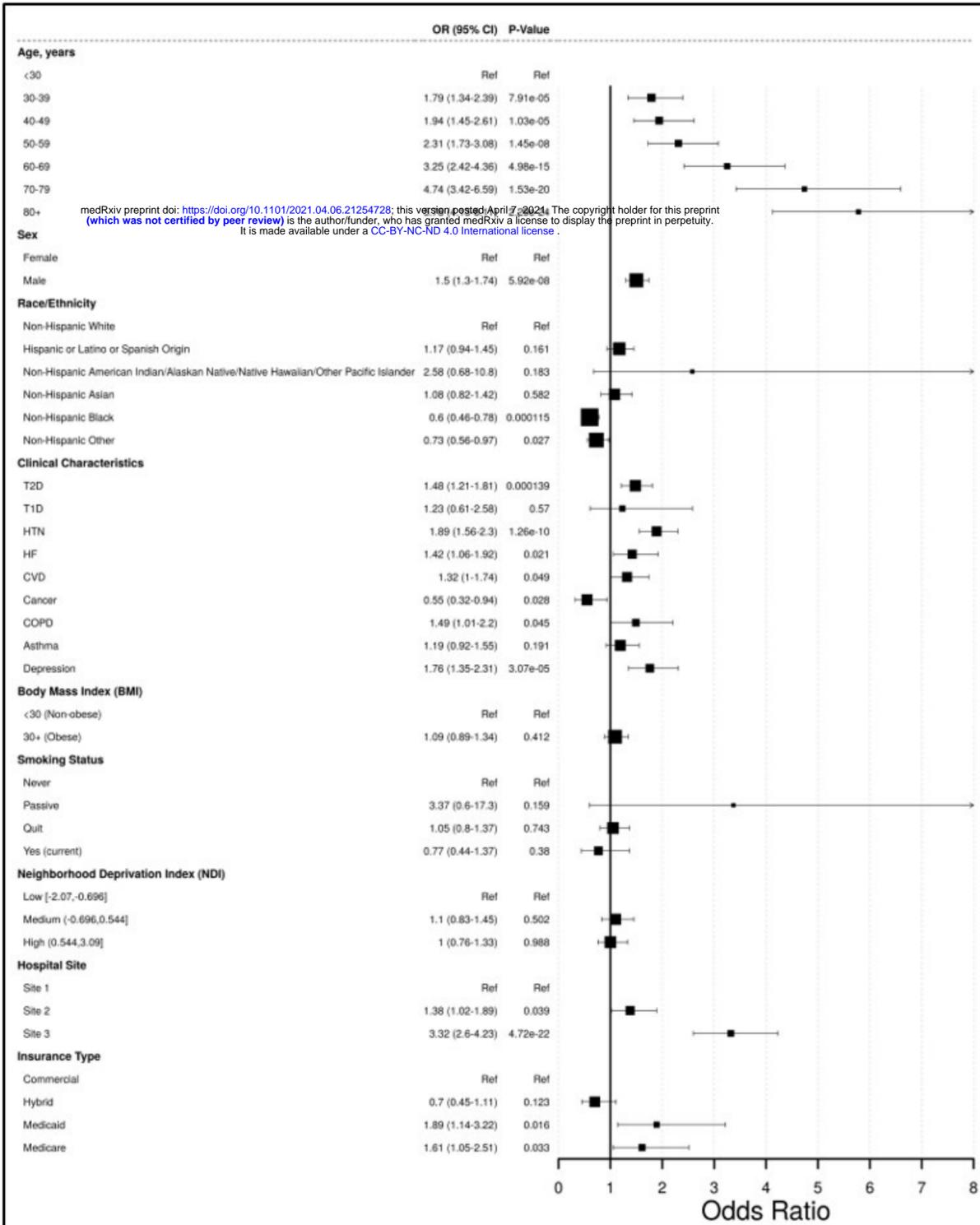
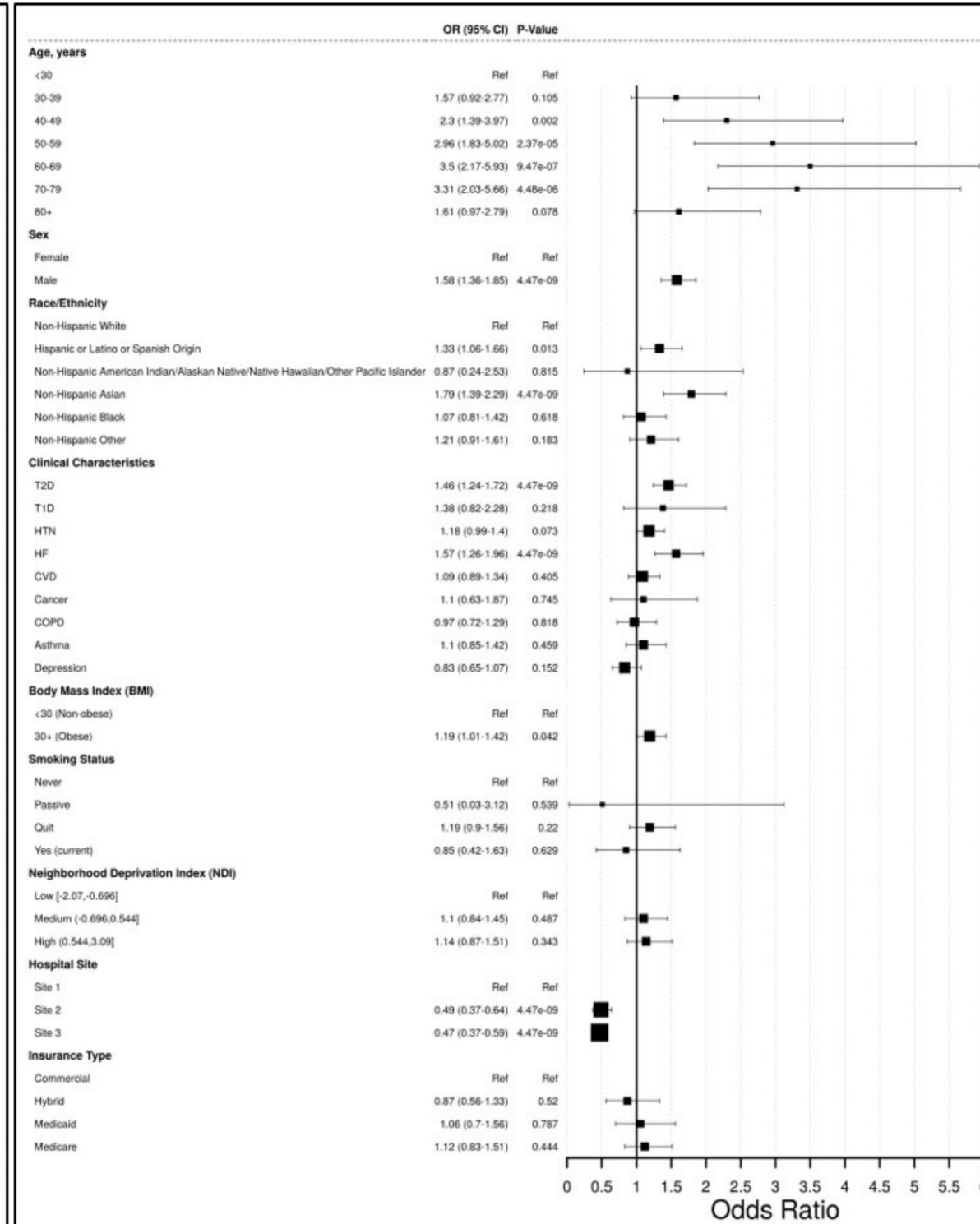


Figure 2. Fully adjusted multivariate models of hospitalization, severe disease, and mortality outcomes among confirmed COVID-19 patients.

A



B



C

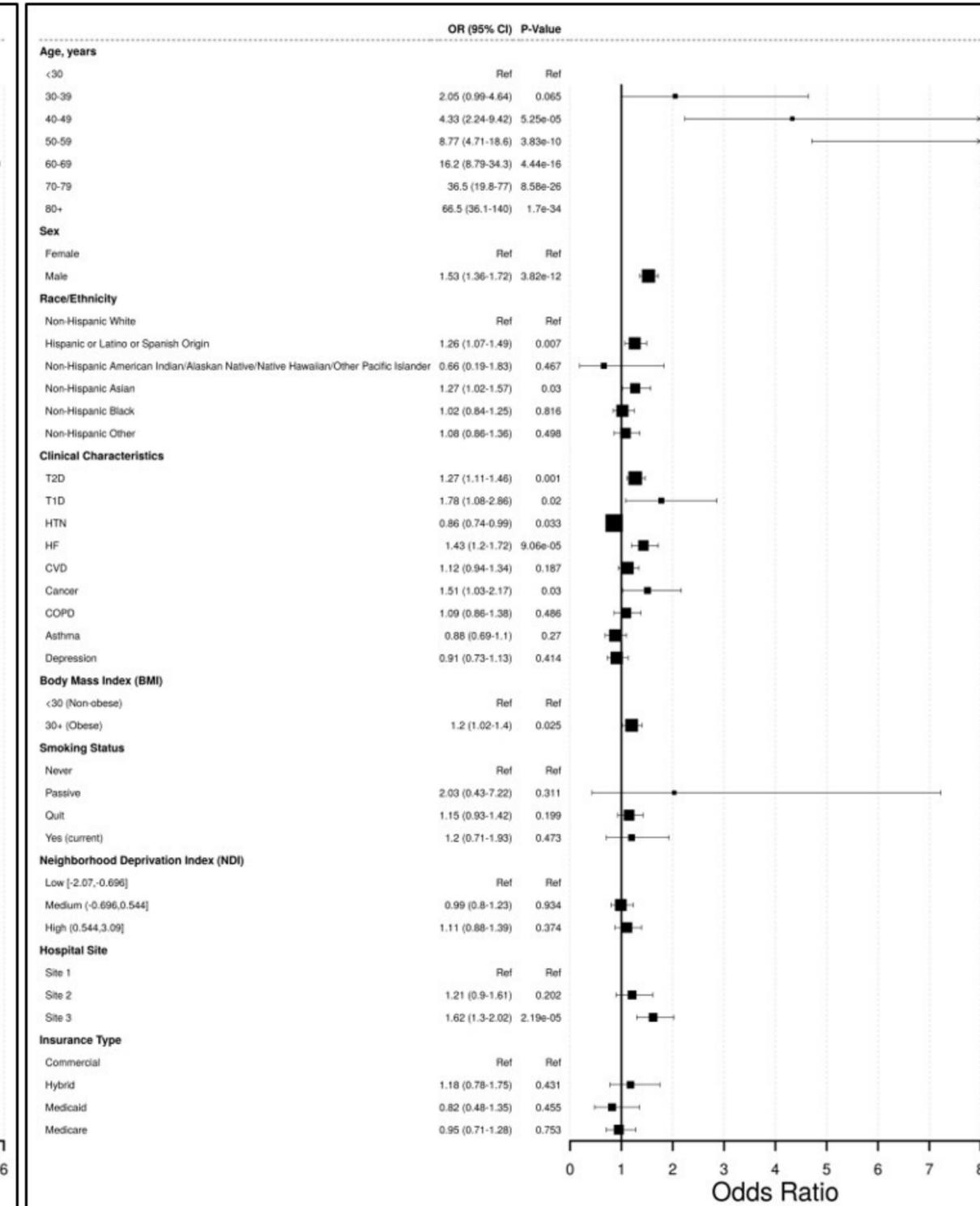
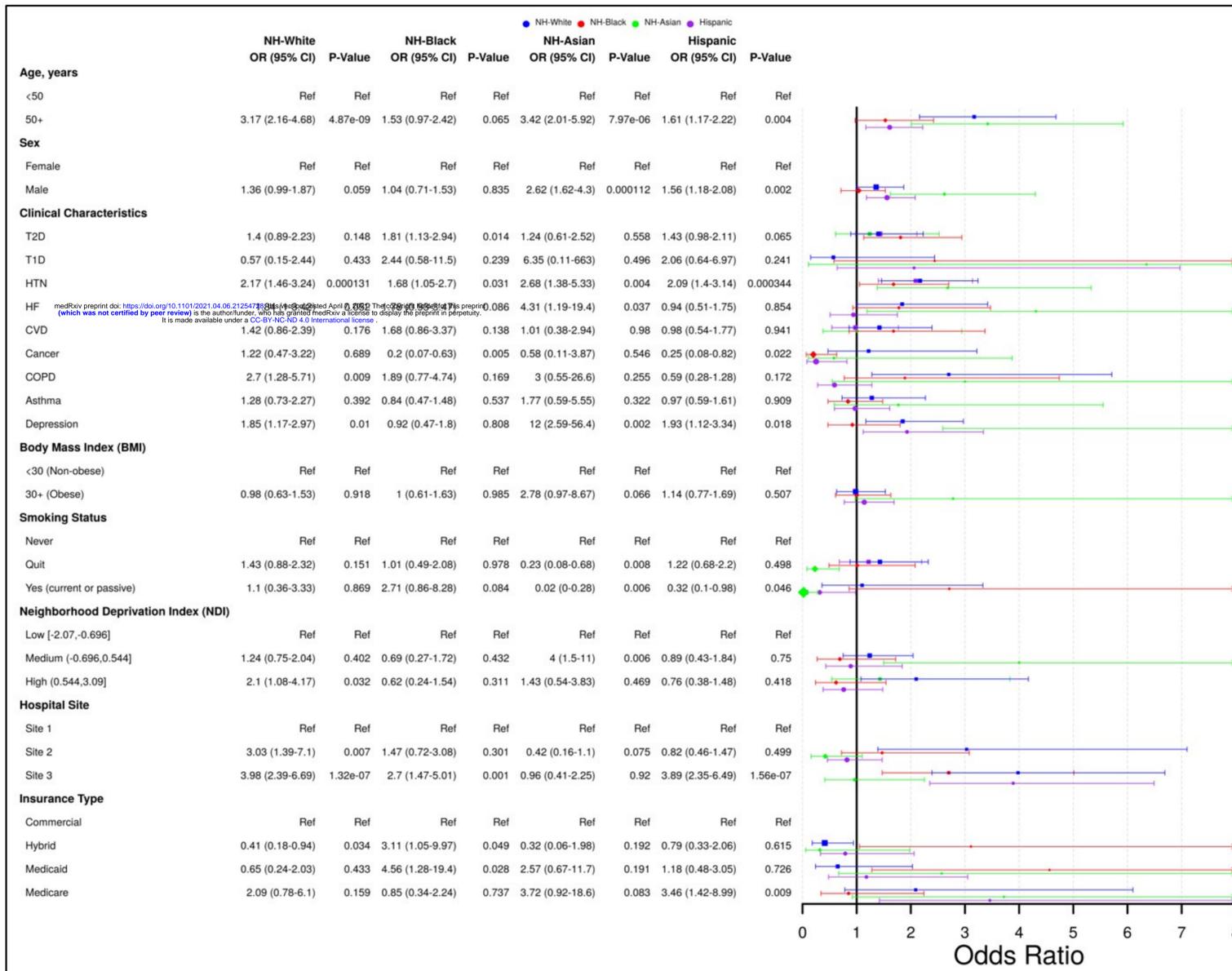
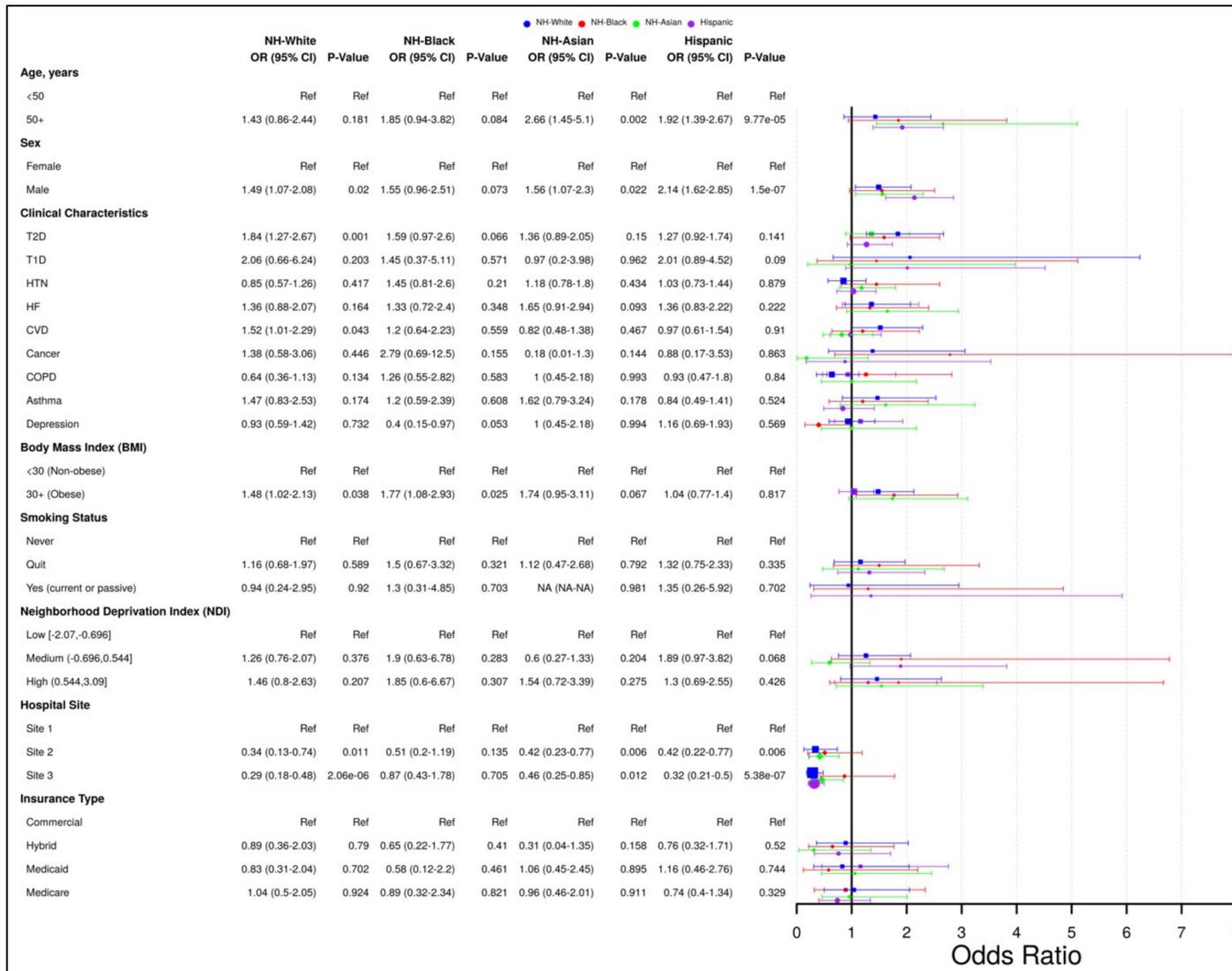


Figure 3. Fully adjusted multivariate models of hospitalization, severe disease, and mortality outcomes among confirmed COVID-19 patients, by race/ethnicity.

A



B



C

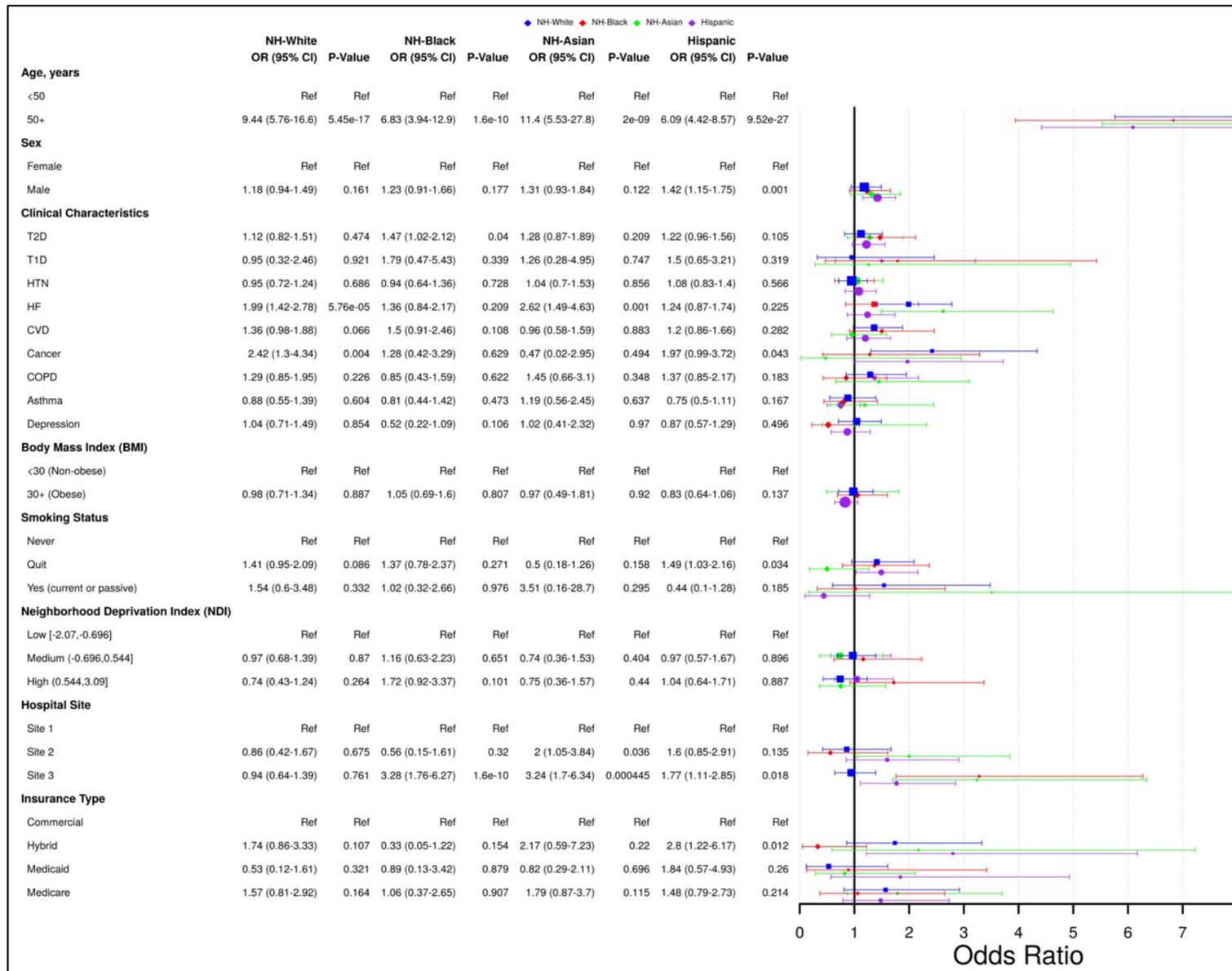
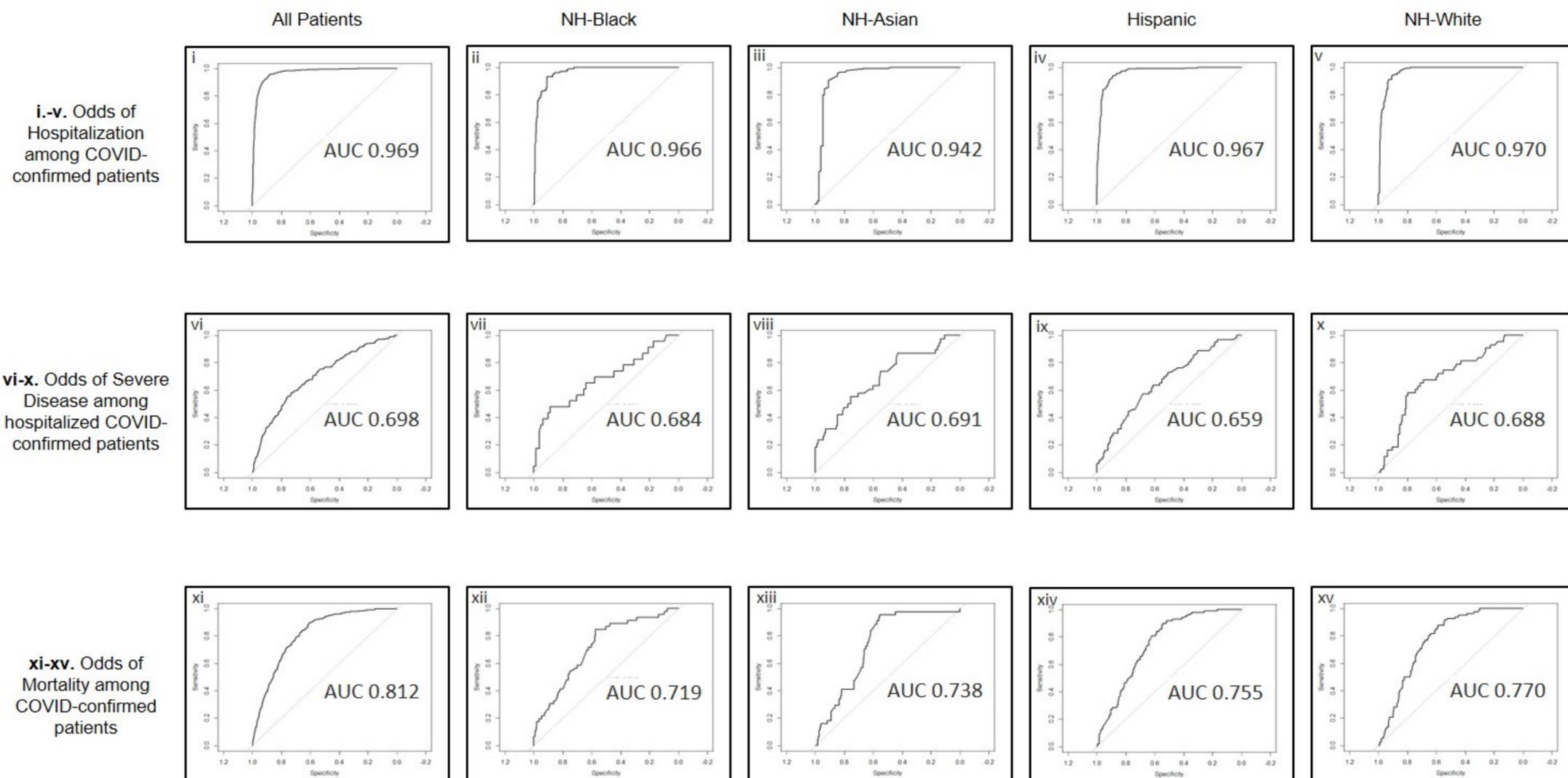


Figure 4. Performance of Risk Models and Ranked Risk Factors.

(A) Receiver operating characteristic (ROC) curves and areas under the curves (AUC) for each fully adjusted multivariate regression model



(B) Ranked risk factors in severe disease models

