A phenome-wide association study (PheWAS) of COVID-19 outcomes by race using the electronic health records data in Michigan Medicine

Maxwell Salvatore, MPH[†1]; Tian Gu, MS[†1]; Jasmine A. Mack, MPH[1]; Swaraaj Prabhu Sankar, MS[2,3]; Snehal Patil, MS[1,4]; Thomas S. Valley, MD[5,6]; Karandeep Singh, MD[6,7]; Brahmajee K. Nallamothu, MD[8]; Sachin Kheterpal, MD[6,9]; Lynda Lisabeth, PhD[10]; Lars G. Fritsche, PhD[1,2,11]; Bhramar Mukherjee, PhD[1,2,10*]

[†]The first two authors contributed equally

Author Affiliations

[1] Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA

[2] Rogel Cancer Center, University of Michigan Medicine, Ann Arbor, MI 48109, United States

[3] Data Office for Clinical and Translational Research, University of Michigan, Ann Arbor, MI 48109, USA

[4] Precision Health, University of Michigan, Ann Arbor, MI 48109, USA

[5] Division of Pulmonary and Critical Care Medicine and Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, USA

[6] Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, MI 48109, USA

[7] Department of Learning Health Sciences, University of Michigan, Ann Arbor, MI 48109, USA

[8] Division of Cardiovascular Medicine and Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI 48109, USA

[9] Department of Anesthesiology, University of Michigan Medical School, Ann Arbor, MI 48109, USA

[10] Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA

[11] Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA

*** Corresponding author:**

Bhramar Mukherjee
Department of Biostatistics
University of Michigan School of Public Health
1415 Washington Heights
Ann Arbor, MI 48109
United States
bhramar@umich.edu
+1 (734)-764-6544

Abstract

Blacks/African Americans are overrepresented in the number of hospitalizations and deaths from COVID-19 in the United States, which could be explained through differences in prevalence of existing comorbidities. We performed a disease-disease phenome-wide association study (PheWAS) using data representing 5,698 COVID-19 patients from a large academic medical center, stratified by race. We explore the association of 1,043 pre-occurring conditions with several COVID-19 outcomes: testing positive, hospitalization, ICU admission, and mortality. *Obesity*, *iron deficiency anemia* and *type II diabetes* were associated with susceptibility in the full cohort, while *ill-defined descriptions/complications of heart disease* and *stage III chronic kidney disease* were associated among non-Hispanic White (NHW) and non-Hispanic Black/African American (NHAA) patients, respectively. The top phenotype hits in the full, NHW, and NHAA cohorts for hospitalization were *acute renal failure*, *hypertension*, and *insufficiency/arrest respiratory failure*, respectively. Suggestive relationships between respiratory issues and COVID-19-related ICU admission and mortality were observed, while circulatory system diseases showed stronger association in NHAA patients. We were able to replicate some known comorbidities related to COVID-19 outcomes while discovering potentially unknown associations, such as endocrine/metabolic conditions related to hospitalization and mental disorders related to mortality, for future validation. We provide interactive PheWAS visualization for broader exploration.

Introduction

The emergence of electronic health records (EHR) and rise of EHR-linked biobanks has made it possible for researchers to explore -omics-based relationships agnostically on a large scale instead of targeted hypothesis testing. Introduced by Denny et al. in 2010, a phenome-wide association study (PheWAS) is an omnibus scan to identify gene-disease associations across the medical phenome.[1] Due to computational advances and development of widely available analytic frameworks,[2–6] PheWAS are now relatively easy to implement. The main goal of a PheWAS is to replicate known gene-disease relationships and to search for hidden and unanticipated associations.

In December 2019, a patient was first diagnosed with COVID-19, the disease caused by a novel coronavirus, SARS-CoV-2.[7] It quickly spread across the globe, earning both the designation of pandemic by the World Health Organization on March 11[8] and a dedicated ICD-10 code. In the US, the first case was confirmed in a traveler returning from Wuhan, China in Washington state on January 21.[9] As of June 29, there are 2,593,169 confirmed cases in the US,[10] representing approximately 25% of all global cases. Because COVID-19 is a respiratory disease and produces flu-like symptoms, testing strategies in the US initially focused on those with symptoms, the elderly, and those with pre-existing conditions[11] - populations who are at risk of severe disease and complications. However, because COVID-19 is a novel disease, only a handful of pre-existing phenotypes are known to be associated with developing symptoms or experiencing adverse outcomes. These include liver, kidney, heart, and respiratory disease.

There has been a remarkable surge within the academic and medical communities to research COVID-19.[12] However, only recently have there been studies examining disparities in broad COVID-19 associated conditions and outcomes in US patient cohorts.[13–16] Instead of a hypothesis driven approach based on the literature, this study applies an agnostic *disease-disease* PheWAS framework to COVID-19 outcomes (to our knowledge, the first of its kind) in a cohort of 5,698 patients who were tested or treated at a large academic medical center. We look at susceptibility and prognosis

among all COVID-19 patients as well as separately among non-Hispanic White (NHW) and non-Hispanic Black/African American (NHAA) patients. The primary objective of this study is to agnostically identify conditions present in an individual's medical record that may be associated with developing COVID-19 symptoms and with hospitalization, ICU admission, and mortality via a large-scale scan. The secondary objective is to compare and contrast the phenome-wide association analyses across races.

Subjects and Methods

COVID-19 cohort: We extracted the EHR for patients tested for COVID-19 at the University of Michigan Health System, also known as Michigan Medicine (MM), from March 10, 2020 to April 22, 2020. A total of 5,500 patients (96.5%) who were tested at MM and 198 patients (3.5%) who were treated for COVID-19 in MM, but tested elsewhere, constituted our initial study cohort of 5,698 patients, of whom 1,119 tested positive. Since the testing protocol in MM[17] focused on prioritized testing (e.g., testing symptomatic patients, those at the highest risk of exposure and those likely to experience fatal outcomes due to existing comorbidities), this is a non-random sample of the population.

Control selection: Controls were used in the susceptibility models and not in the prognosis models. We created two sets of alive control samples from the MM patient database to compare and contrast the testing positive, one "unmatched control sample" consists of 7,211 randomly drawn patients and another 13,351 "matched control sample" using 1:3 frequency-matching on race (NHW/NHAA), sex and age (above/below 50). We used unmatched controls in the analysis with the full cohort (including all races) and matched controls in the race-stratified analysis. Study protocols were reviewed and approved by the University of Michigan Medical School Institutional Review Board (IRB ID HUM00180294 and HUM00155849).

Classifying patients who were still in hospital and ICU: We categorized patients into non-hospitalized, hospitalized (includes ICU stays), and hospitalized with ICU stay

based on the admission and discharge data. 166 patients were still admitted in the hospital of which 113 had at least one ICU state and 53 had no ICU stay at the time of the data extraction. We performed a sensitivity analysis by excluding these patients whose final prognostic outcome was unknown at the time of data abstraction.

Generation of the medical phenome: We constructed the medical phenome by extracting available International Classification of Diseases (ICD; ninth and tenth editions) code from EHR and forming them up to 1,781 traits using the PheWAS R package (as described in detail elsewhere).[1] Each of these traits was coded as a binary risk factor (present/absent) and used as a predictor in the association models with COVID-19 outcomes. The analyses in this study were restricted to traits that ever appeared in the EHR of at least 10 COVID-19 positive patients. To differentiate *pre-existing* conditions from phenotypes related to COVID-19 testing/treatment, we applied a 14-day-prior restriction on the tested cohort by removing diagnoses that first appeared within the 14 days before the first test or diagnosis date, whichever was earlier. We use the term "pre-existing" liberally to include not only chronic conditions but also acute health events that were diagnosed at any point in the patient's EHR prior to COVID diagnosis. Further, we realize that the aggregation of ICD codes into phecodes may result in clinically unusable or unclear phenotypes. While the PheWAS is performed on phecodes, one can view the mapping of ICD-to-phecode relationships on this website: http://shiny.sph.umich.edu/ICD_Coding/ (Michigan Genomics Initiative [MGI] mapping applies to this manuscript).

Description of variables: A summary data dictionary is available with the source and definition of each variable used in our analysis (Supplementary eTable 2A).

Statistical analysis: We performed two types of comparisons in this study (detailed definition in Supplementary eTable 2B):

(a) Predictors of COVID-19 susceptibility: comparing those who were diagnosed with COVID-19 with those who were not (unmatched controls)

(b) Predictors of three COVID-19 prognostic outcomes: among those who were diagnosed, (i) comparing those who were hospitalized with those who were not, (ii) those who were admitted to ICU with those who were not, and (iii) those who died with those who did not (no untested controls were used, only considers tested positive cohort).

All COVID-19 outcomes of interest are binary; thus, logistic regression was our primary tool. All logistic regression models were of the following form:

$$\text{logit}\, P(Y_{COVID} = 1 | Covariates, PheCode_k) = \beta_0 + \beta_{Cov}^{\top} Covariates + \beta_k I[Phecode_k = 1]$$

$k = 1, \cdots 1043$. Here Y_{COVID} is various COVID-19 related outcomes under consideration (e.g., COVID-19 positive, hospitalization and so on). The Firth correction was used to address potential separation issues in logistic regression models.[18–20] Full models were adjusted for age, sex, race, and four census tract-level socioeconomic indicators: proportion with less than high school education, proportion unemployed, proportion with annual income below the federal poverty level, and population density (persons per mile^2). The socioeconomic characteristics are defined by US census tract (corresponding to the residential address available in each patient's EHR) for the year 2010 and are from the National Neighborhood Data Archive (NaNDA).[21] PheWAS adjusting for an additional comorbidity score covariate (indicating whether the patient was diagnosed with conditions across seven disease categories associated with COVID-19 susceptibility and adverse outcomes: respiratory, circulatory, any cancer, type II diabetes, kidney, liver, and autoimmune; ranges from 0 to 7) is included on our accompanying website: http://prsweb-dev.sph.umich.edu:8080/covidphewas/.

For all models, we report the Firth corrected estimate of the odds ratio, 95% Wald-type confidence interval and P-value. A conservative Bonferroni multiple testing correction was implemented to conclude statistically significant results of susceptibility (P=0.05/1043), and P < .05 was used as a threshold for suggestive traits in the prognostic results where the sample size was limited. In the PheWAS plots in Figures 1 and 2, these thresholds are represented by the horizontal, dashed red and orange lines, respectively. The x-axis are individual disease codes, color-coded by their

corresponding disease category as described in the figure legend. The y-axis represents the $-\log_{10}$ transformed p-value of the association. Each point is represented by either an upward triangle indicating a positive association or a downward triangle indicating a negative association.

Stratified analysis for NHW and NHAA: Since the susceptibility and prognostic factors are potentially different across races, we carried out the entire analysis stratified by race. We restricted our attention to NHW and NHAA due to limitations of sample size for other racial groups. Supplementary eTable 1 contains descriptive statistics stratified by race. Matched controls were used in the model for COVID-19 susceptibility, as the proportion of NHW and NHAA in unmatched controls are not comparable to the stratified population under study.

Results

There were 5,698 patients who were either tested for or diagnosed with COVID-19 (n_{tested}; specifically, 5,500 patients were tested and 198 patients were diagnosed with COVID-19 and transferred into Michigan Medicine [MM]), 7,211 unmatched controls ($n_{unmatched}$), and 13,351 matched controls ($n_{matched}$) eligible for inclusion in this study. Of the 26,260 individuals eligible for inclusion, our study population comprised 23,769 individuals (n_{tested}=5,225 [$n_{positive}$=1068]; $n_{unmatched}$=6,811; $n_{matched}$=11,733) who had available International Classification of Disease (ICD; ninth and tenth editions) code data before applying the 14-day-prior to testing restriction to the EHR. Among the 5,225 tested individuals, 4,622 had pre-existing diagnoses data 14 days prior to diagnosis/first test, which yield the final sample size of 23,166 (n_{tested}=4,662 [$n_{positive}$=778]; $n_{unmatched}$=6,811; $n_{matched}$=11,733). Furthermore, a total of 1,781 qualified ICD-code-based phenotypes, referred to as PheWAS traits, were initially screened and 1,043 had at least 10 occurrences in our COVID-19 positive cohort. Thus, we analyzed 1,043 unique phecodes from 17 different disease categories.

Of those 4,622 who were tested for COVID-19, 36.3% (1,676/4,615) were males and the median age was 47 years. The majority were NHW (66% [3,051]) while 17% were

NHAA (785). Out of the study cohort, 16.8% (778) were tested positive (Table 1). Among the 778 positive patients, 49.4% (384) were NHW, 34.8% (271) were NHAA, 35.0% (272) were hospitalized, 13.8% (107) were admitted to ICU and 2.3% (18) died.

Table 1. Descriptive Characteristics of the COVID-19 Tested/Diagnosed cohort

	COVID-19 Tested					
			COVID-19 Positive			
	Overall (N=4622)	Negative (N=3844)	Overall (N=778)	Hospitalized (n=272)	ICU (n=107)	Deceased (n=18)
Variable	**No./No. (%)**					
Age						
mean (SD)	47.4 (20.6)	46.4 (20.9)	52.3 (18.0)	62.8 (15.8)	61.8 (13.7)	69.7 (13.5)
median (IQR)	47 (32, 63)	46 (31, 62)	53 (37.3, 65)	64.0 [54, 75]	62.0 [56, 70]	70 (62, 80.8)
Male Sex	1676/4615 (36.3)	1333/3838 (34.7)	343/777 (44.1)	158 (58.1)	70 (65.4)	13 (72.2)
BMI, mean (SD); No.	29.8 (7.6); 4140	29.4 (7.5); 3399	31.7 (7.8); 741	32.7 (8.0); 266	33.4 (7.5); 106	31.7 (7.9)
Race/Ethnicity, No. (%)						
NHW	3051 (66.0)	2667 (69.4)	384 (49.4)	125 (46.0%)	45 (42.1%)	9 (50.0)
NHAA	785 (17.0)	514 (13.4)	271 (34.8)	111 (40.8%)	48 (44.9%)	7 (38.9)
Other Ethnicity	419 (9.1)	345 (9.0)	74 (9.5)	24 (8.8%)	7 (6.5%)	1 (5.6)
Unknown Ethnicity	367 (7.9)	318 (8.3)	49 (6.3)	12 (4.4%)	7 (6.5%)	1 (5.6)
SES, mean (SD); No.						
% < HS Education	0.08 (0.07); 4132	0.08 (0.06); 3411	0.09 (0.08); 721	0.10 (0.09); 248	0.12 (0.10); 96	0.14 (0.11); 17
% Unemployed	0.07 (0.04); 4132	0.07 (0.03); 3411	0.08 (0.04); 721	0.08 (0.05); 248	0.09 (0.05); 96	0.09 (0.03); 17
% Annual Income < FPL	0.13 (0.12); 4132	0.12 (0.11); 3411	0.14 (0.12); 721	0.15 (0.13); 248	0.17 (0.14); 96	0.20 (0.14); 17
Persons per square mile^2	2630 (2320); 4132	2530 (2300); 3411	3140 (2320); 721	3580 (2580); 248	3810 (2570); 96	4550 (3310); 17

Abbreviations: BMI, body mass index; NHW, non-Hispanic Whites; NHAA, non-Hispanic Blacks/African Americans; SES, census tract-level socioeconomic status; HS, high school; FPL, federal poverty level.

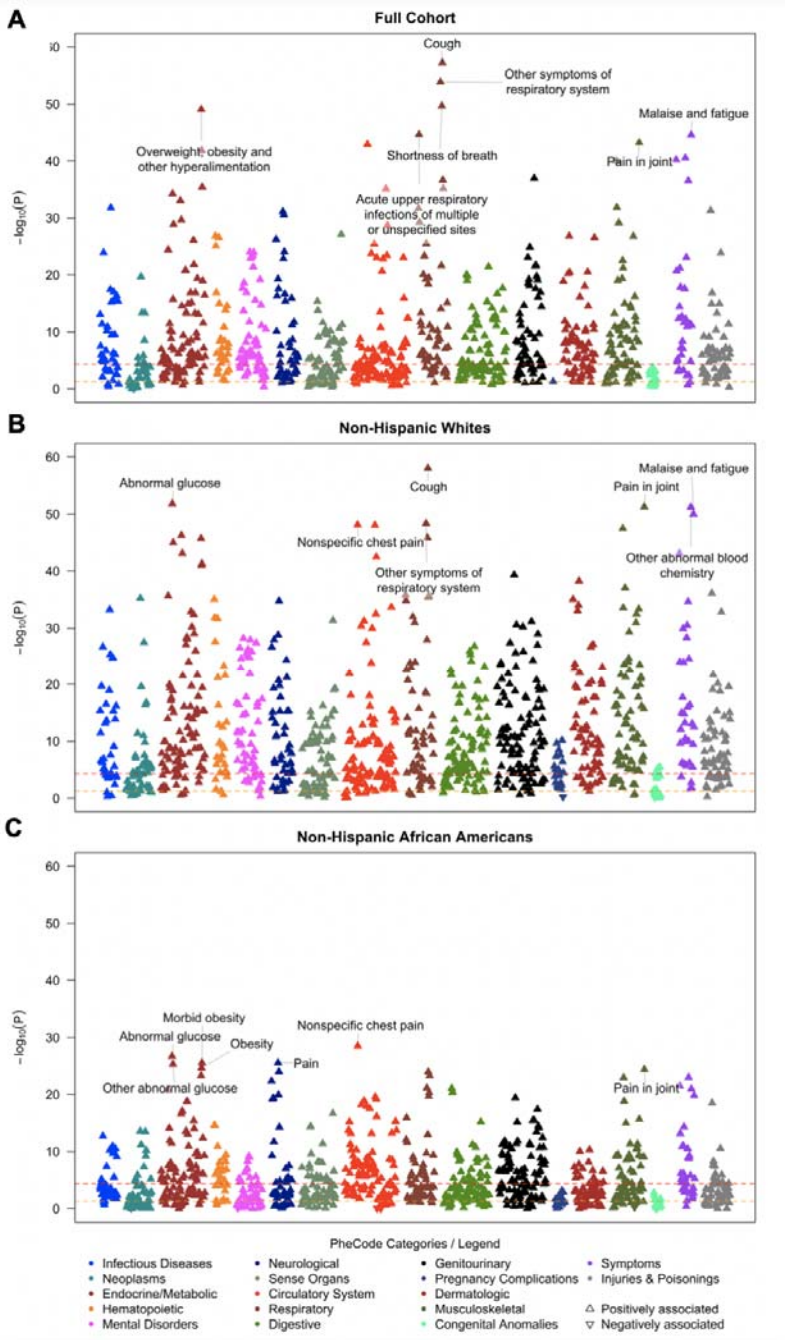
Phenome-wide comorbidity association analysis

The top 50 traits from the comorbidity PheWAS can be found in Supplementary eTable S3, S3A and S3B for the full cohort and for NHW and NHAA, respectively. Interactive versions of the PheWAS plots are online at http://prsweb-dev.sph.umich.edu:8080/covidphewas/. This resource also provides tables with the adjusted odds ratios, 95% confidence intervals, p-values, and counts of occurrence in cases and controls for all traits included in the PheWAS performed.

Full cohort susceptibility: For susceptibility, when comparing the positives and the unmatched controls, 538 traits were identified after applying Bonferroni correction. This demonstrates that patients who were tested and tested positive were sicker than the general population. As illustrated in Figure 1A, we found strong and positive associations with various comorbidities and COVID-19 positive diagnosis (e.g., pain in joint [$P=5.97x10^{-44}$]; respiratory abnormalities [$P=7.3x10^{-36}$]; complications of heart disease [$P=7.82x10^{-36}$]). Overall, the findings were consistent with previously identified COVID-19 risk factors (e.g., obesity [$P=8.92x10^{-50}$] and diabetes mellitus [$P=3.8x10^{-25}$] were associated with higher risk of being test positive).[22] In contrast, the comparison between those who tested positive for COVID-19 and those who tested negative leads to counterintuitive findings (361 traits showed protective effect out of 369 significant traits under Bonferroni correction, such as non-hypertensive congestive heart failure [odds ratio [OR]=0.39, $P=5.8x10^{-10}$] and acute renal failure [OR=0.47, $P=7.7x10^{-8}$]) contradicting findings in other COVID-19 studies[23,24] (Supplementary eTable S3). This amplifies the need for choosing an appropriate control group.

Race-stratified susceptibility: As shown in Figure 1B, we identified 734 traits in NHW, including 84 genitourinary, 79 endocrine/metabolic and 66 circulatory system diseases, such as hematuria ($P=4.16x10^{-28}$), abnormal glucose ($P=1.57x10^{-52}$) and Ill-defined descriptions and complications of heart disease ($P=9x10^{-49}$). In addition, as shown in Figure 1C, we observed 406 traits in NHAA, including 61 genitourinary, 59 circulatory system, and 52 endocrine/metabolic diseases, where some of the top traits includes cardiac conduction disorders ($P=5.73x10^{-19}$) and stage III chronic kidney disease ($P=1.32x10^{-12}$).

Figure 1. Manhattan plot showing the phenome-wide association between disease codes and testing positive for COVID-19. Models are adjusted for age, sex, race (full cohort only), and four census tract-level socioeconomic indicators: proportion with less than high school education, proportion unemployed, proportion with annual income below the federal poverty level, and population density (persons per mile^2). The x-axis are individual disease codes, color-coded by their corresponding disease category as described in the legend. The y-axis represents the -$\log_{10}$ transformed p-value of the association. The dashed, horizontal lines represent the p = 0.05 (in orange) and the Bonferroni corrected p-value (0.05 / 1,043; in red). Each point is represented by either an upward triangle indicating a positive association or a downward triangle indicating a negative association.



Full cohort prognostic associations: As the disease outcome progresses (from hospitalized to ICU, and to deceased), stronger associations with respiratory diseases, circulatory system diseases, kidney diseases and type II diabetes were observed compared with other comorbidities. Four traits were phenome-wide significantly associated with hospitalization—renal failure ($P=3.03x10^{-5}$), acute renal failure ($P=3.40x10^{-5}$), acid-base balance disorder ($P=6.57x10^{-5}$), and hypertensive heart and/or renal disease ($P=8.24x10^{-5}$). In addition, 127 traits were identified to be associated with hospitalization (Figure 2A), 56 associated with ICU admission (Figure 2D), and 239 associated with mortality (Figure 2G), under threshold $P<0.05$. For example, patients with pulmonary heart diseases ($P=1.51x10^{-4}$) or diabetic complications such as chronic ulcer of leg/foot ($P=8.57x10^{-4}$) showed an association with hospitalization; respiratory failures such as chronic airway obstruction ($P=4.63x10^{-4}$) and bronchiectasis ($P=6.38x10^{-4}$) were identified as the top threats of admission to ICU; and previous history of pleurisy ($P=2.4x10^{-5}$) was phenome-wide significantly associated with COVID-19 mortality (Figure 2G).

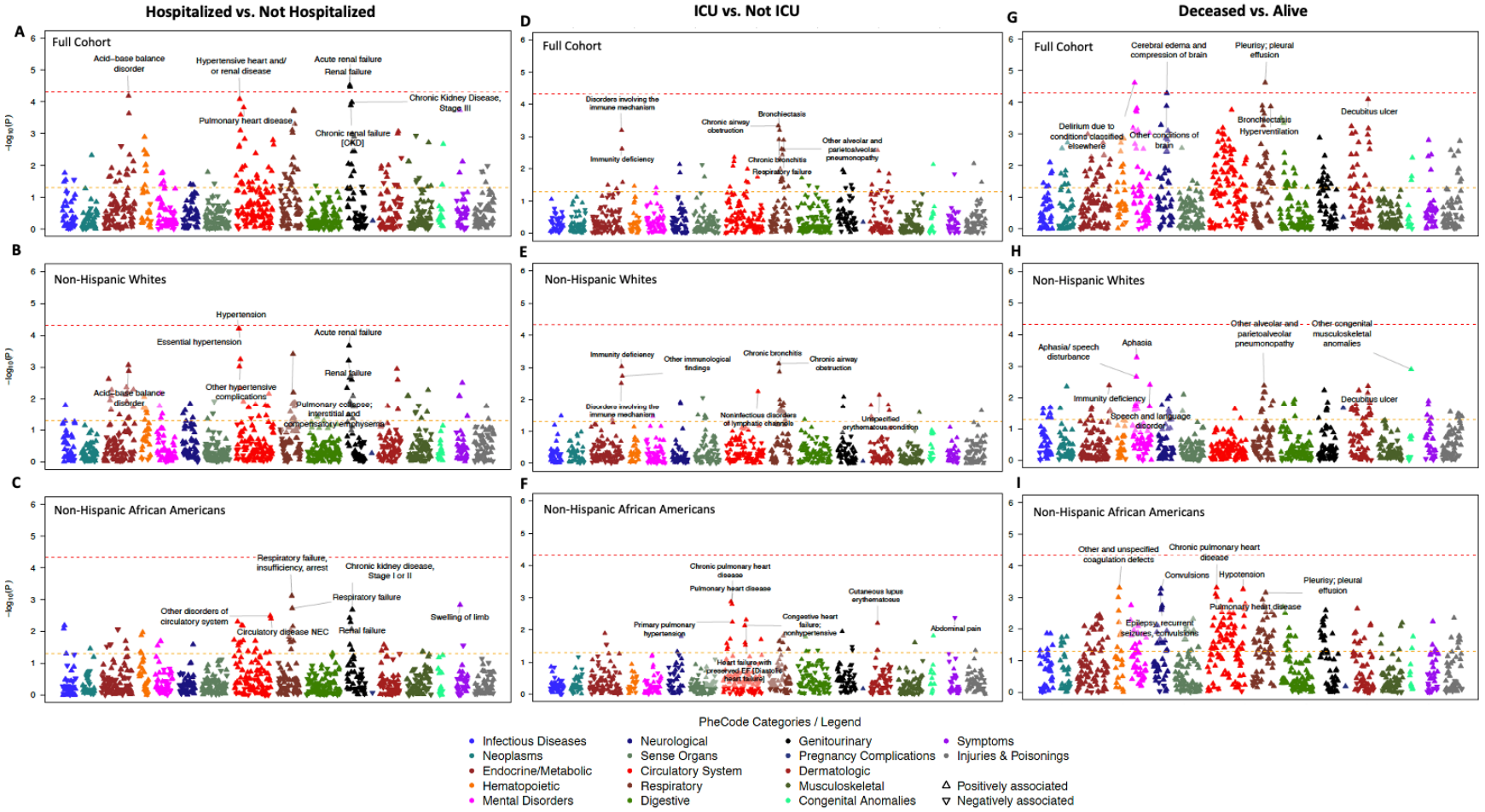
Race-stratified prognostic associations:

In NHW, we identified no phenome-wide significantly associated trait, but 93 traits were nominally associated with hospitalization, 40 with ICU admission, and 88 associated with COVID-19 mortality. Specifically, hypertension was identified as the top trait for hospitalization ($P=6.22x10^{-5}$; Figure 2B); chronic airway obstruction ($P=7.5x10^{-4}$) and chronic bronchitis ($P=0.001$) were associated with high risk of ICU admission (Figure 2E); Unexpected associations included disorders such as aphasia ($P=5.31x10^{-4}$) and benign neoplasm of lip/oral cavity/pharynx ($P=0.004$) showed strong signs of COVID-19 mortality (Figure 2H).

In NHAA, no phenome-wide significantly associated trait was detected but we identified a total of 59 traits nominally associated with hospitalization (Figure 2C), 38 with ICU admission, and 229 associated with COVID-19 mortality. Different from NHW, various circulatory heart diseases were observed as top traits associated with ICU admission, of which pulmonary heart disease ($P=0.001$), chronic pulmonary heart disease ($P=0.002$)

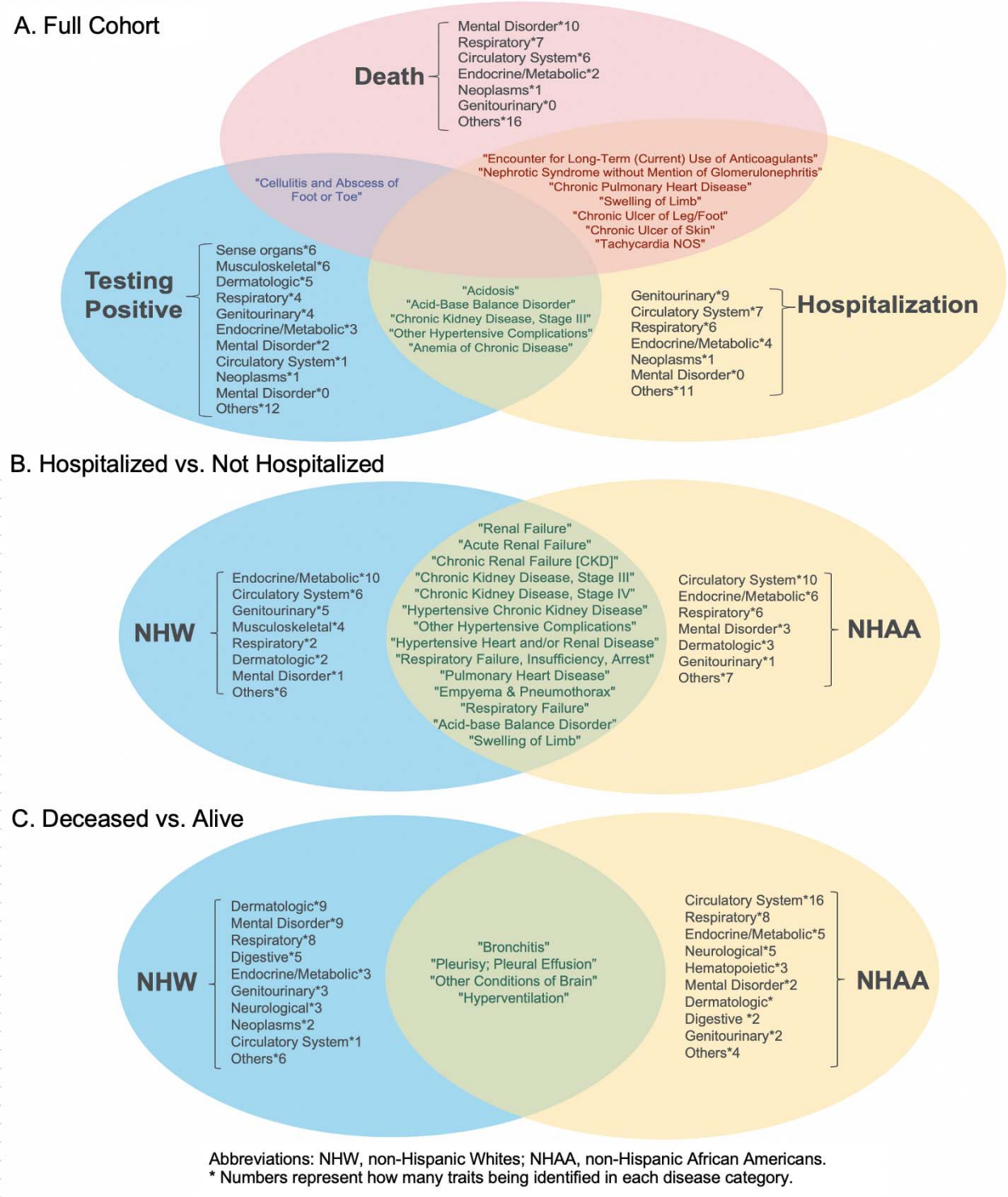
and diastolic heart failure (P=0.005) were among the top five (Figure 2F). As shown in Figure 2H and Figure 2I, both of the number and strength of association between circulatory system disorders and COVID-19 mortality was higher in NHAA patients compared with NHW, with a total of 62 traits identified in NHAA while only 2 in NHW. Similarly, we observe a higher prevalence of genitourinary diseases in NHAA associated with COVID-19 mortality such as acute renal failure (P=0.005) and stage I/II chronic kidney disease (P=0.004) compared with NHW. Moreover, we also observe an association between coagulation defects and COVID-19 mortality (P= 0.0005) that was not observed in NHW.

Figure 2. Manhattan plot showing the phenome-wide association between disease conditions and prognostic outcomes for COVID-19. Models are adjusted for age, sex, race (full cohort only), and three census tract-level socioeconomic indicators: proportion with less than high school education, proportion unemployed, and proportion with annual income below the federal poverty level. The x-axis are individual disease codes, color-coded by their corresponding disease category as described in the shared legend. The y-axis represents the -$\log_{10}$ transformed p-value of the association. The dashed, horizontal lines represent the p = 0.05 (in orange) and the Bonferroni corrected p-value (0.05 / 1,043; in red). Each point is represented by either an upward triangle indicating a positive association or a downward triangle indicating a negative association.



Summary Takeaways: In summary, (i) in all cohorts, as the disease progressed to increasingly severe prognosis, the associated phenotypes concentrated in kidney, respiratory and circulatory system diseases (Figure 3A); pre-existing chronic diseases such as *stage III chronic kidney disease*, *anemia of chronic disease* and *chronic pulmonary heart disease* appeared to be associated with poor prognosis, while mental disorders distinctly showed association to COVID-19 mortality; (ii) When comparing NHW and NHAA, kidney diseases showed an association with hospitalization in both races whereas endocrine/metabolic problems were the largest number of hits in NHW and circulatory system diseases were strongest hits in NHAA (Figure 3B); (iii) Circulatory system diseases including various heart diseases stood out as the top threats associated with ICU admission distinctively in NHAA; (iv) Towards mortality, associations with respiratory problems were observed in NHW and NHAA, while associations with dermatologic and mental issues were often seen in NHW and associations with circulatory system diseases were especially prevalent in NHAA (Figure 3C).

Figure 3. Venn diagrams of the top 50 traits. Each circle represents the top 50 hits from the full cohort PheWAS (panel A) and the racial PheWAS (panels B and C), respectively. Traits shared across PheWAS are stated, while the corresponding number of traits within a given disease category that are unique to that PheWAS are also provided.



Abbreviations: NHW, non-Hispanic Whites; NHAA, non-Hispanic African Americans.
* Numbers represent how many traits being identified in each disease category.

Discussion

Using data from a cohort of tested/diagnosed COVID-19 patients at MM, we performed what we believe is the first PheWAS looking at COVID-19, stratified by race. This technique allows us to explore and identify potentially associated conditions across the medical phenome that are associated with susceptibility, hospitalization, ICU admission or mortality. Our results yield many previously known or plausibly associated phenotypes for increasingly severe prognosis, namely pulmonary diseases, such as pulmonary heart disease, respiratory failure and bronchitis. Our stratified analysis showed that respiratory conditions appear to be associated with more severe outcomes among NHW while coagulation renal disease and heart disease are more strongly associated with severe outcomes among NHAA. Our results can inform targeted prevention across racial groups, which includes increased testing and encouraging self-isolation from household members with specific disease profiles along with education of enhanced public health prevention guidelines.

There are several limitations to this analysis. First, there is the agnostic nature of PheWAS, which can identify potentially spurious associations. While we feel that many of the top traits have been highlighted elsewhere and are biologically plausible, there is currently no process in place for rapidly discerning potentially novel from spurious associations[25] beyond extensive manual review and follow-up research, particularly for a novel disease. Second, many of the issues with utilizing EHR data for research purposes also applies here including inaccurate data from billing codes[26] and failure of physicians to report/record problems.[27] However, Wei et al. (2017)[28] showed that manually curated phecodes, as used in this study, were better at identifying phenotypes than other phenotype classification coding systems, including raw ICD codes. Third, the sample size for a PheWAS is still rather small to be able to identify statistically significant associations. Moreover, we did not distinguish between transfer patients (i.e., those who were diagnosed elsewhere and transferred to MM for treatment), who may have been sicker patients than the cohort diagnosed at MM. However, given that this is an emerging and novel disease, we feel it is important to identify suggestive associations so that future research and clinicians can potentially consider other

conditions outside those that have been previously identified – namely, pulmonary and cardiovascular conditions. Another limitation of our analysis is scanning through each phenotype one at a time though they occur in a correlated and interactive manner. A richer multivariate model needs to be constructed with more complex features. Finally, we focus on association analysis and refrain from risk prediction and risk stratification which are the obvious logical next steps.

This work contributes to a new area of COVID-19 research that rigorously examines racial differences in disease susceptibility and prognosis. Moreover, we incorporated census tract-level SES covariates, which are important to consider when comparing races. We found several potentially novel diseases unexpectedly associated with different outcomes in the course of COVID-19 progression and that some disease profiles differ by race. We hope this exploratory effort will inspire hypothesis generation for future research that might result in targeted prevention and care as we are still combatting this pandemic. In this spirit, we have made all PheWAS results available for exploration here: http://prsweb-dev.sph.umich.edu:8080/covidphewas/. Future work include: (i) constructing a COVID-19 comorbidity index to identify individuals who are at particularly high risk of being diagnosed with and developing severe COVID-19 outcomes; (ii) assess multivariate prediction using a complex non-linear phenome-space to account for interactions using modern machine learning tools and provide individual level predictions for absolute risk (iii) follow the EHR of COVID patients released from the hospital prospectively, to track enrichment of specific diseases.

Author contributions

MS, TG, BM and JAM wrote initial drafts and revisions. SPS and LGF procured and prepared data. TG and LGF performed data analysis and prepared figures and tables. SP led the development of the accompanying web application, with contribution from LGF. TSV, KS, BKN, SK, and LL edited, reviewed and provided in-depth review guidance of each draft. BM provided leadership and revised each draft. All authors contributed expertise and provided meaningful feedback throughout draft and revision process.

Competing interests

The authors have no competing interests to declare.

Data availability

Data cannot be shared publicly due to patient confidentiality. The data underlying the results presented in the study are available from University of Michigan Data Office for Clinical & Translational Research for researchers who meet the criteria for access to confidential data.

Materials & Correspondence

Materials requests and correspondence should be directed to Bhramar Mukherjee via email at bhramar@umich.edu.

References

1 Denny JC, Ritchie MD, Basford MA, *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010; **26**: 1205–10.

2 Gagliano Taliun SA, VandeHaar P, Boughton AP, *et al.* Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat Genet* 2020; **52**: 550–2.

3 Verma A, Lucas A, Verma SS, *et al.* PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. *Am J Hum Genet* 2018; **102**: 592–608.

4 Zhao X, Geng X, Srinivasasainagendra V, *et al.* A PheWAS study of a large observational epidemiological cohort of African Americans from the REGARDS study. *BMC Med Genomics* 2019; **12**: 26.

5 Cai T, Zhang Y, Ho Y-L, *et al.* Association of Interleukin 6 Receptor Variant With Cardiovascular Disease Effects of Interleukin 6 Receptor Blocking Therapy. *JAMA Cardiol* 2018; **3**: 849.

6 Leppert B, Millard LAC, Riglin L, *et al.* A cross-disorder PRS-pheWAS of 5 major psychiatric disorders in UK Biobank. *PLOS Genet* 2020; **16**: e1008185.

7 Bryner J. 1st known case of coronavirus traced back to November in China. Live Sci. 2020; published online March 14. https://www.livescience.com/first-case-coronavirus-found.html (accessed March 19, 2020).

8 WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. World Heal. Organ. 2020; published online March 11. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed March 19, 2020).

9 Centers for Disease Control and Prevention. First travel-related case of 2019 novel coronavirus detected in United States. 2020; published online Jan 21. https://www.cdc.gov/media/releases/2020/p0121-novel-coronavirus-travel-case.html.

10 Microsoft bing COVID-19 Tracker. Microsoft Corp. https://www.bing.com/covid (accessed June 29, 2020).

11 Centers for Disease Control and Prevention. Overview of testing for SARS-CoV-2. 2020. https://www.cdc.gov/coronavirus/2019-ncov/hcp/testing-overview.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fhcp%2Fclinical-criteria.html (accessed June 15, 2020).

12 Brainard J. Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? Sci. Mag. 2020; published online May 13. https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat.

13 Price-Haywood EG, Burton J, Fort D, Seoane L. Hospitalization and Mortality among Black Patients and White Patients with Covid-19. *N Engl J Med* 2020; : NEJMsa2011686.

14 Brandt EB, Beck AF, Mersha TB. Air pollution, racial disparities, and COVID-19 mortality. *J Allergy Clin Immunol* 2020; published online May. DOI:10.1016/j.jaci.2020.04.035.

15 Laurencin CT, McClinton A. The COVID-19 Pandemic: a Call to Action to Identify and Address Racial and Ethnic Disparities. *J Racial Ethn Heal Disparities* 2020; **7**: 398–402.

16 Gu T, Mack JA, Salvatore M, *et al.* COVID-19 outcomes, risk factors and associations by race: a comprehensive analysis using electronic health records data in Michigan Medicine. *medRxiv* 2020. DOI:10.1101/2020.06.16.20133140.

17 Michigan Medicine. Indications for COVID-19 diagnostics testing for adult patients in all clinical settings. 2020. http://www.med.umich.edu/asp/pdf/adult_guidelines/COVID-19-testing.pdf (accessed June 9, 2020).

18 FIRTH D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**: 27–38.

19 Kosmidis I, Pagui ECK, Sartori N. Mean and median bias reduction in generalized linear models. 2018; published online April 11. http://arxiv.org/abs/1804.04085.

20 Kosmidis I. brglm2: Bias Reduction in Generalized Linear Models. 2020. https://cran.r-project.org/package=brglm2.

21 NaNDA | Social Environment and Health Program.

https://seh.isr.umich.edu/signature-projects/nanda/ (accessed June 10, 2020).

22 Coronavirus Disease 2019 (COVID-19): Are You at Higher Risk for Severe Illness? Centers Dis. Control Prev. 2020. https://www.cdc.gov/coronavirus/2019-ncov/specific-groups/high-risk-complications.html (accessed March 19, 2020).

23 de Lusignan S, Dorward J, Correa A, *et al.* Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance Centre primary care network: a cross-sectional study. *Lancet Infect Dis* 2020; published online May. DOI:10.1016/S1473-3099(20)30371-6.

24 Miyara M, Tubach F, POURCHER V, *et al.* Low rate of daily active tobacco smoking in patients with symptomatic COVID-19. *Qeios* 2020; published online May 9. DOI:10.32388/WPP19W.4.

25 Hanauer DA, Saeed M, Zheng K, *et al.* Applying MetaMap to Medline for identifying novel associations in a large clinical dataset: a feasibility analysis. *J Am Med Informatics Assoc* 2014; **21**: 925–37.

26 Rhodes ET, Laffel LMB, Gonzalez T V., Ludwig DS. Accuracy of Administrative Coding for Type 2 Diabetes in Children, Adolescents, and Young Adults. *Diabetes Care* 2007; **30**: 141–3.

27 Williams C, Mosley-Williams A, C M. Accuracy of provider generated computerized problem lists in the Veterans Administration. *AMIA Annu Symp Proc* 2007; : 1155.

28 Wei W-Q, Bastarache LA, Carroll RJ, *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* 2017; **12**: e0175508.