Abridged version

A Comprehensive Analysis of Health Equity in the Veterans Affairs Healthcare System Using COVID-19 Patient Data

13 August 2021

Bitscopic Inc., 715 Colorado Avenue, Suite B, Palo Alto, CA 94303

Table of Contents

Introduction

Bitscopic has been working with the VA healthcare system on a range of medical data-related projects for over a decade. Over the past 18 months, it has worked on multiple COVID-19-related projects in the areas of Machine Learning-based diagnosis, therapies, clustering, and genetic sequence analysis (Bayat et al. 2020). This project is a collaboration between Bitscopic, the Office of Healthcare Innovation and Learning, and the Office of the Chief Technology Officer with the goal of utilizing Bitscopic's expertise in analyzing COVID-19 positive patient data to provide insights into demographics, health equity, comorbidity-associated risks, outcomes, therapeutics, care paths, and best practices. These are not only relevant to improving short-term mortality and morbidity but also in decreasing risk and improving treatment for patients at high risk of long-term COVID-19-related complications such as heart attacks, strokes, and Post-COVID-19 Syndrome, better known as "long-haul COVID" or Post-Acute Sequelae of SARS-CoV-2 (PASC).

Bitscopic's Praedico® platform normalizes and standardizes medical data across the VA healthcare system across numerous types of inpatient and outpatient lab data, vitals, pharmacy and comorbidity data, among others (Holodniy et al. 2015). With this standardized data, we analyzed a dataset of 148,831 U.S. Veteran patients who had a positive SARS-CoV-2 qualitative polymerase chain-reaction (PCR) or antigen assay result between March 2nd, 2020 and April 7th, 2021. Out of these patients, 25,655 were admitted to the hospital within three days of their first COVID-19 positive test result, and it is this initial dataset that Bitscopic is utilizing for its machine learning efforts.

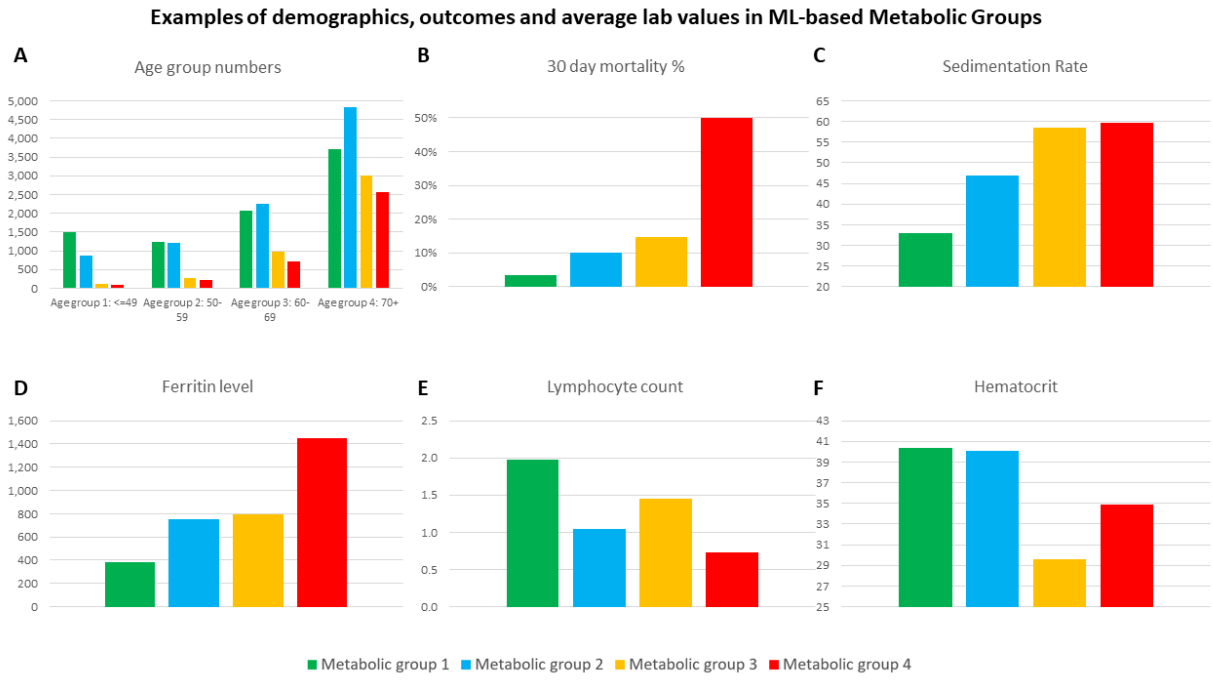
Using this rich normalized dataset of 25,655 patients, we have utilized Principal Component Analysis and K-means analysis to separate patients into two types of clusters, the first based largely on the patient's lab results, which we call a Metabolic Cluster, and the second based on their comorbidities, e.g., hypertension, diabetes, and cancer history. The utility of this approach has been validated by other groups that have reached similar conclusions to Bitscopic's team, an example being Benito-Léon et al's work on the use of unsupervised machine learning to identify age and gender-independent COVID19 patient groups published last month (Benito-Leon et al. 2021). Another example of a Machine Learning clinical support tool was published last month for the diagnosis of Inflammatory Bowel Disease types in children, in whom diagnoses are more uncertain. They used Machine Learning to take inflammation-related hospital labs such as C-Reactive Protein and Sedimentation Rate (both of which are labs we include in our input) along with inflamed bowel location data to predict the type of Inflammatory Bowel Disease the patient had, and achieved almost 91% accuracy relative to the final clinical diagnoses given to the children (Schneider et al. 2021).

Every patient therefore belongs to a Metabolic cluster and a Comorbidity cluster, with four possibilities for each. Patients who belong to the same cluster, sometimes referred to as a phenotypic cluster, are therefore relatively similar to each other. It should be noted that due to the clusters having been generated in an unsupervised manner, albeit of carefully normalized and curated data with low missingness, their relative sizes are also different. They also have very different outcomes, with patients in metabolic cluster 4 having a 50% 30-day all-cause mortality rate as opposed to an 3.3% 30-day all-cause mortality rate for cluster 1. Graphs demonstrating examples of the differences between the metabolic clusters are shown below.

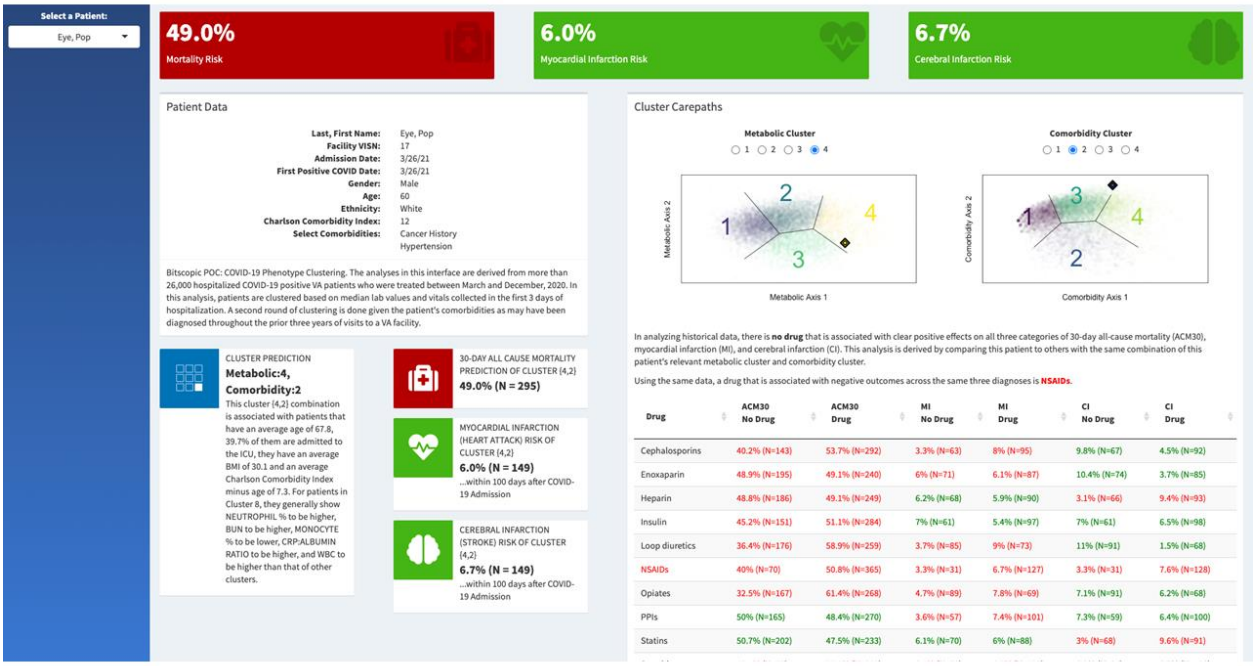
We next analyzed the outcomes of patients in each of these clusters. One interesting way to do this is to examine associations between the therapies and therapy classes, e.g., NSAIDS, that patients received during their first five days of hospitalization and their short-, medium- and long-term outcomes, including all-cause mortality, rate of developing strokes or heart attacks, and development of Long Haul COVID (where that information is available, either through clinical records or careful analysis of ICD-10 codes despite Long-COVID/PASC ICD-10 codes having still yet to be defined). An example of similar work we had performed prior to the beginning of this project was on the drug Ondansetron, which we found to be associated with reduced mortality, especially in patients with high comorbidities and which was published in Open Forum Infectious Diseases (Bayat et al. 2021).

The final two outcomes of the project are to prepare a written report that summarizes and explores the findings of the project and to produce a prototype dashboard tool which will allow providers and researchers with a means of analyzing a patient in the context of the clusters that they have been assigned to in terms of their predicted outcomes as well as the outcome associations of patients in that cluster for various therapies, highlighting those therapies that have better or worse outcomes. It should be borne in mind that the display of therapies associated with

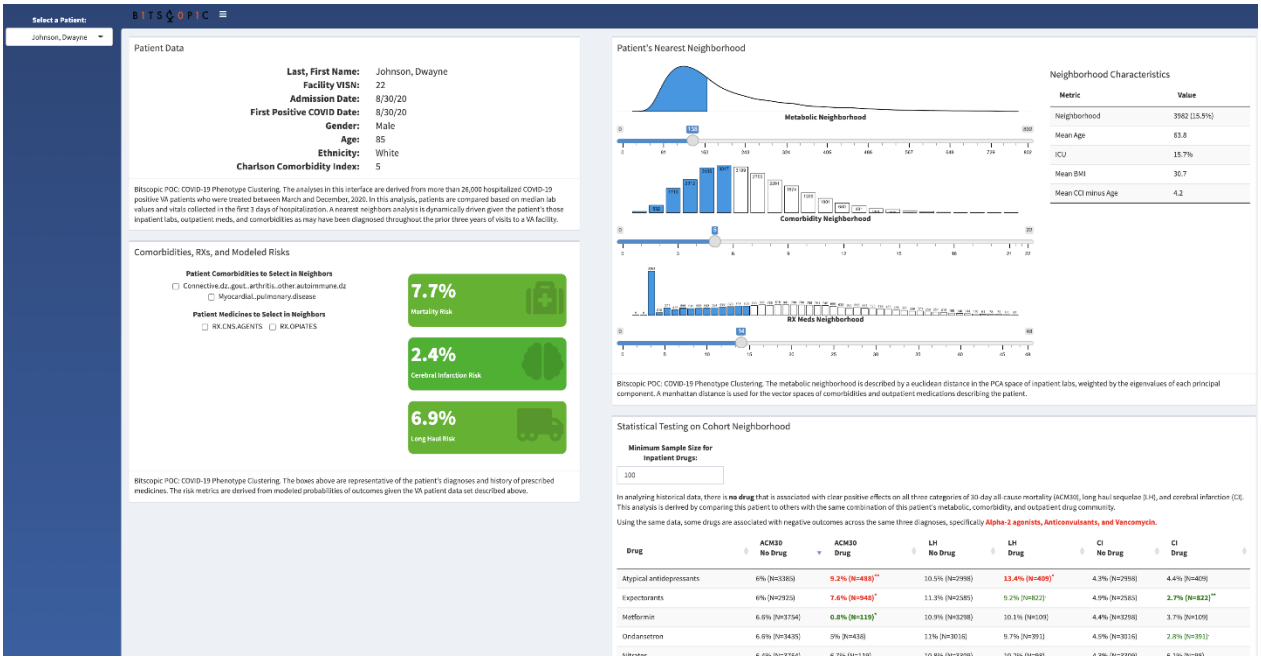
better or worse outcomes for a patient's cluster does not implicate a causational relationship but in the case of a therapy with a positive outcome for patients in that cluster and which the provider can see a potential benefit, such as a potassium supplement in a patient who is low in potassium and in a specific cluster, the provider may choose to prescribe it. In this way, the tool could be utilized for clinical support or care optimization applications, and potentially lead to randomized clinical trials in patients belonging to specific clusters and better outcomes. This prototype is shown below. Once this project has completed its objectives, Bitscopic has agreed with OHIL and the Office of the Chief Technology Officer to follow up with the development of a next generation of the prototype that will serve as a clinical support tool for improved care plans.



Decisioning GUI Iteration 1



Decisioning GUI Iteration 2



Methods

Sources of Data

The U.S. Department of Veterans Affairs (VA) healthcare system serves over 9 million veterans at over 1,200 Veterans Health Administration sites of care throughout the United States and U.S. territories (Health 2015). All VA facilities were included in this analysis. The data analyzed here is from a dataset of 148,831 U.S. veteran patients who had a positive SARS-CoV-2 qualitative polymerase chain-reaction (PCR) or antigen assay result entered officially into the VA's Lab Chemistry system between March 2nd, 2020 and April 7th, 2021. Out of these patients, 25,655 were admitted to the hospital within 3.00 days of their first COVID-19 positive test result. Inpatient barcode medication administration (BCMA) records and SARS-CoV-2 test sample times permitted us to calculate the time interval between the first positive SARS-CoV-2 test and the administration of specific drug doses. SARS-CoV-2 test data, BCMA, inpatient and outpatient medication, laboratory data for the hospital stay of interest, intensive care unit (ICU) admission status, as well as comorbidity, demographic, and self-reported ethnicity data for the prior three years of outpatient and inpatient visits were included in our analysis. Relevant data sources from VA sites were maintained, integrated, and normalized using the Bitscopic Praedico® platform (Holodniy et al. 2015).

Outline of Results Section

In the following sections, we perform critical analysis on hospitalized COVID-19 patients and their outcomes in **Part 1** and subphenotyping and model prototyping in **Part 2**.

For Part 1, we are interested in assessing how patients from different demographics (ethnic groups/race, gender, urban vs. rural) differ in terms of rate of COVID-19 testing, ICU admission, days of hospitalization, mortality, and vaccinations for COVID-19. We also investigate what rates COVID-19 positive VA patients develop new onset myocardial infarctions, strokes and Long Haul COVID-associated sequelae.

In Part 2, we examine whether COVID-19 positive VA patients can be subphenotyped in an unsupervised manner (e.g., using PCA analysis) based on their lab results, vitals, comorbidities, prescription drugs and demographic backgrounds, and outcomes, and if so, whether demographics plays a role.

Figures are numbered based on which part they belong to.

Part 1: COVID-19 Hospitalization Data, Analysis, and Issues Surrounding Health Equity

Selected Findings

Figure 1: Hospitalized COVID19 positive VA patient 30 day mortality by ethnicity

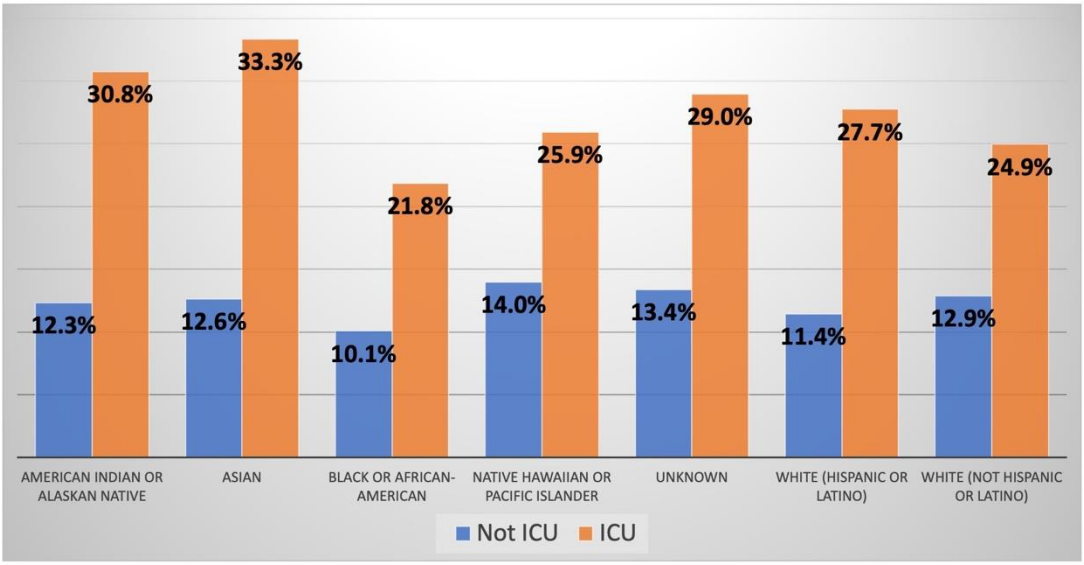
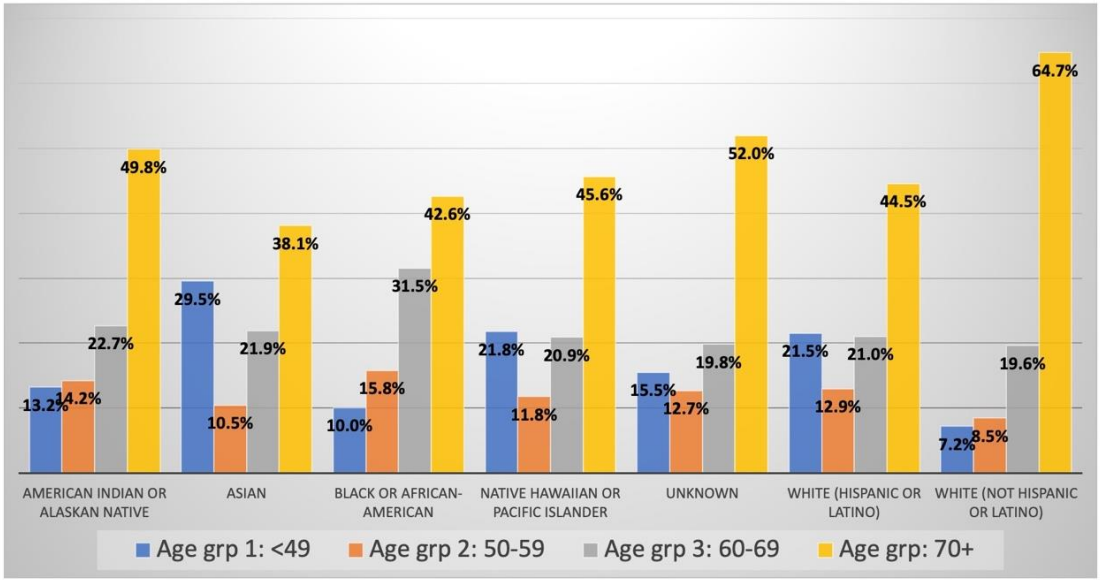


Figure 1 depicts that VA patients across all ethnicities who were admitted to the ICU experienced significantly higher 30-day mortality than patients who were not admitted. Asian patients admitted to the ICU experienced the highest 30-day mortality at 33.3% whereas Black or African-American patients admitted to the ICU experienced the lowest mortality at 21.8%. For non-ICU stays, Native Hawaiian or Pacific Islander patients experienced the highest 30-day mortality at 14.0% while Black or African-American patients experienced the lowest mortality at 10.1%. This data is interesting because most studies find that Black or African-American patients have significantly higher mortality than their White counterparts. However, it has also been noted in the literature that despite greater hospitalization rates for African-Americans, they do not suffer worse outcomes if provided similar care, which appears to be the case in the VA system (Krishnamoorthy et al. 2021).

Figure 2: Hospitalized COVID19 positive VA patient age group percentages by ethnicity

Figure 2 indicates that across all ethnicities, VA patients who are 70+ years old were the most hospitalized age group (Group 4). The second highest hospitalized age group are patients



aged 60-69 (Group 3). Moreover, White (not Hispanic or Latino) patients had the highest percent hospitalization of the 70+ age group, 64.7%, and the lowest percent hospitalization of those <49 years old (Group 1), 7.2%, compared to all other ethnicities. Interestingly, Asians had the highest percent hospitalization of patients <49, 29.5%, and the lowest percent hospitalization of those 70+, 38.1%. This data is compatible with other studies such as that by Gold et al, which found that hospitalization and outcomes were directly proportional to age (Gold et al. 2020).

Figure 3: Hospitalized COVID19 positive VA patient percentage developing stroke or MI within 100 days of first COVID19 diagnosis by ethnicity

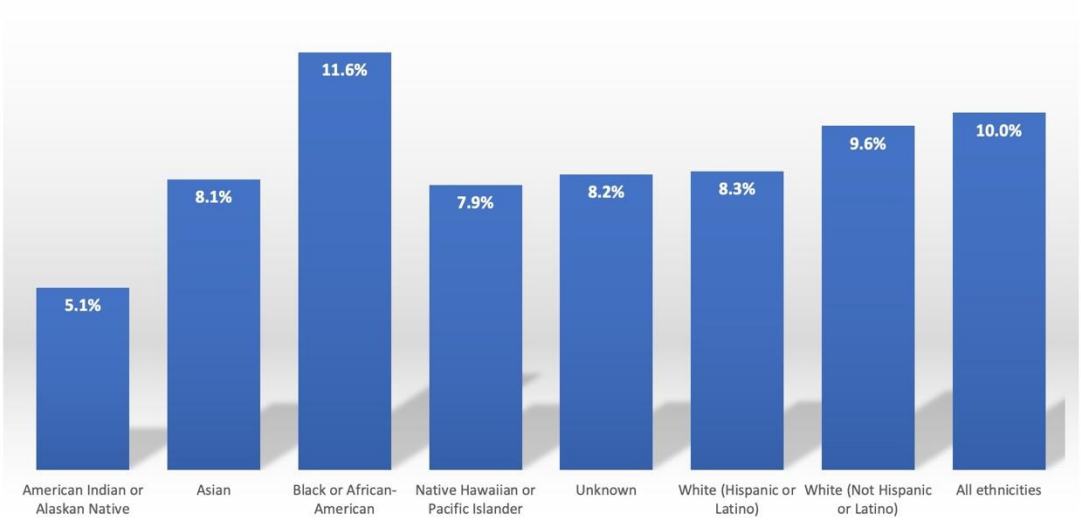
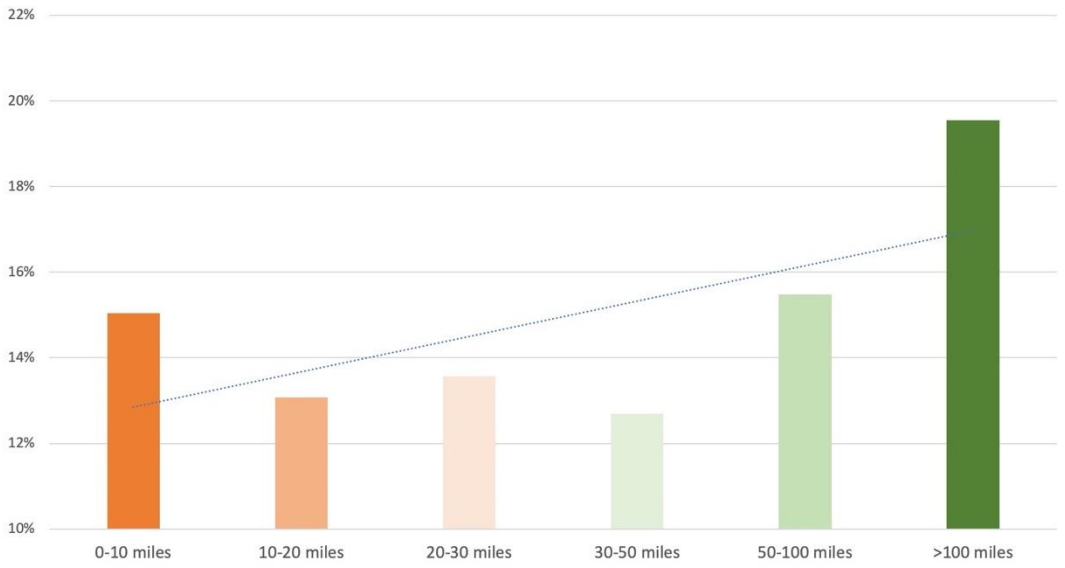


Figure 3 demonstrates that a higher percentage of Black or African-American VA patients develop a stroke or MI within 100 days of the first COVID-19 diagnosis when compared to all other ethnicities. Specifically, 11.6% of Black or African American patients develop a stroke or MI within 100 days of the first COVID-19 diagnosis compared to 10% of patients from all ethnicities. In contrast, only 5.1% of American Indian or Alaskan Native patients develop a stroke or MI within 100 days of the first COVID-19 diagnosis—the lowest percentage when compared to all other ethnicities and less than half the rate for Black and African-American patients. Furthermore, Whites (not Hispanic or Latino) have the second highest percentage of patients—9.6%—who develop a stroke or MI. This data is consistent with other reports, such as that by Lekoubou et al, showing that the prevalence of ischemic stroke in Blacks, non-Hispanic Whites and Hispanics was 1.26% (95% CI: 0.86% to 1.83%), 0.84% (95% CI: 0.51% to 1.37%) and 0.49% (95% CI: 0.26% to 0.88%) respectively. After adjusting for age, sex, hypertension, diabetes, obesity, drinking and smoking, they found that the likelihood of stroke was higher in Black than non-Black patients (adjusted odds ratio, 2.76; 95% CI, 1.13 to 7.15, p=0.03) (Gold et al. 2020).

Figure 4: 30 day mortality of hospitalized COVID19 patients by distance from VA hospital



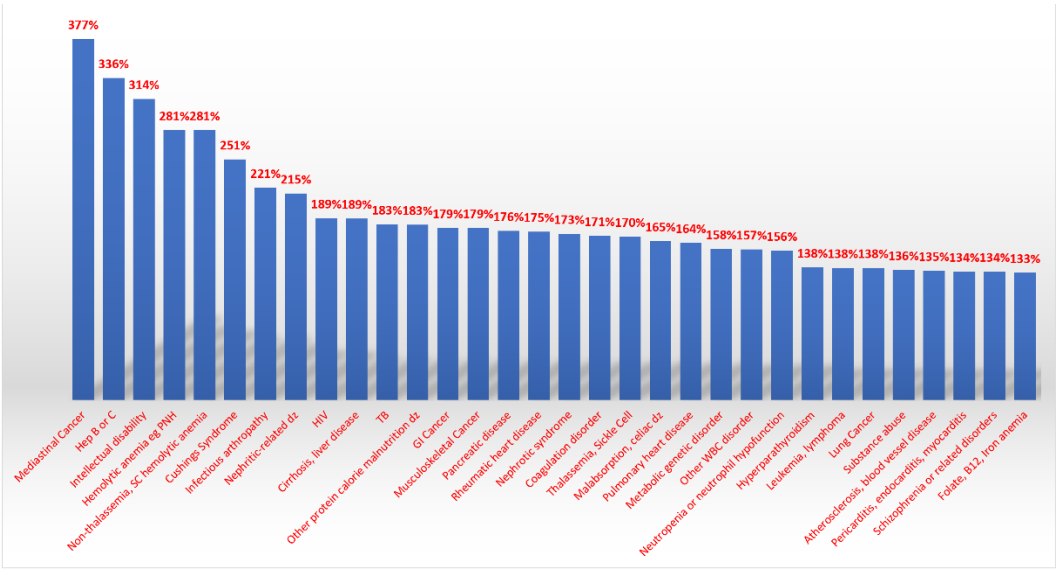
Patient zip codes and VA hospitals (not including VA-affiliated hospitals, clinics, mobile clinics, etc.) were mapped to their latitude and longitude coordinates, and the minimum linear distance was calculated for each patient. Figure 4 depicts that the 30-day mortality rate was highest for patients who were >100 miles away from the nearest VA hospital and lowest for patients 30-50 miles away from the nearest VA hospital.

Figure 5: Breakthrough COVID19 infections >14 days after 1 or 2 vaccine doses through 27 April 2021

	Janssen/J&J	Moderna 1 dose	Pfizer 1 dose	Moderna 2 doses	Pfizer 2 doses
American Indian or Alaskan Native	0/1,138 (0.000%)	2/2,137 (0.094%)	1/1,623 (0.062%)	8/10,099 (0.079%)	12/8,646 (0.139%)
Asian	0/1,434 (0.000%)	3/3,236 (0.093%)	0/3,931 (0.000%)	10/13,699 (0.073%)	9/13,839 (0.065%)
Black or African-American	12/18,222 (0.066%)	61/38,296 (0.159%)	85/33,836 (0.251%)	224/174,777 (0.128%)	371/215,163 (0.172%)
Native Hawaiian or Pacific Islander	1/1,227 (0.081%)	1/2,456 (0.041%)	4/1,948 (0.205%)	10/12,838 (0.078%)	17/11,096 (0.153%)
White (Unknown)	0/1572 (0.000%)	4/3,007 (0.133%)	3/1,923 (0.156%)	10/13,629 (0.073%)	12/11,082 (0.108%)
White (Hispanic or Latino)	5/4,879 (0.102%)	14/13,524 (0.106%)	20/7,858 (0.255%)	73/57,352 (0.127%)	74/44,853 (0.165%)
White (Not Hispanic or Latino)	23/68,723 (0.033%)	177/124,287 (0.142%)	134/73,059 (0.183%)	780/794,581 (0.098%)	901/610,175 (0.148%)
Unknown	5/18,611 (0.027%)	11/29,724 (0.037%)	13/31,484 (0.041%)	66/95,749 (0.069%)	68/107,783 (0.063%)
	46/115,806 (0.040%)	273/216,397 (0.126%)	260/155,652 (0.167%)	1,181/1,172,724 (0.101%)	1,464/1,022,637 (0.143%)

As shown in Figure 5, the data indicates that the three vaccines are all effective at producing very low rates of infections. Among Hispanics and African-Americans, there is a tendency for a greater number of breakthrough infections for all three vaccines, particularly when patients do not receive their second Pfizer dose. Across all demographics, the likelihood of breakthrough infections decreases when the second dose of Moderna or Pfizer is received.

Figure 6: Analysis of Long Haul COVID19 patients



	Mediastinal Cancer	Hep B or C	Intellectual disability	Hemolytic anemia, e.g. PNH	Non-thalassemia, SC hemolytic anemia	Cushings Syndrome	Infectious arthropathy	Nephritic-related dz	HIV	Cirrhosis, liver disease	TB
% in non-Long Haul	0.03%	3.38%	0.02%	0.16%	0.16%	0.03%	0.75%	0.19%	0.92%	7.28%	0.13%
% in Long Haul	0.10%	11.37%	0.07%	0.44%	0.44%	0.07%	1.67%	0.41%	1.74%	13.78%	0.24%

	Other protein calorie malnutrition dz	GI Cancer	Musculoskeletal Cancer	Pancreatic disease	Rheumatic heart disease	Nephrotic syndrome	Coagulation disorder	Thalassemia, Sickle Cell	Malabsorption, celiac dz	Pulmonary heart disease
% in non-Long Haul	1.84%	1.82%	0.08%	2.26%	0.87%	0.31%	3.88%	0.38%	0.49%	4.90%
% in Long Haul	3.37%	3.27%	0.14%	3.98%	1.53%	0.54%	6.64%	0.65%	0.82%	8.03%

	Metabolic genetic disorder	Other WBC disorder	Neutropenia or neutrophil hypofunction	Hyperparathyroidism	Leukemia, lymphoma	Lung Cancer	Substance abuse	Atherosclerosis, blood vessel disease	Pericarditis, endocarditis, myocarditis	Schizophrenia or related disorders	Folate, B12, Iron anemia
% in non-Long Haul	6.79%	4.25%	0.70%	1.43%	3.12%	1.36%	22.80%	23.00%	6.02%	5.26%	11.05%
% in Long Haul	10.69%	6.67%	1.09%	1.97%	4.29%	1.87%	30.91%	31.01%	8.07%	7.05%	14.67%

We used the ICD10 codes G93.3, Z86.19, and B94.8 to identify those of the 21,374 hospitalized patients in our dataset who survived at least 60 days that had been diagnosed with Long Haul COVID. 2,938 out of the 21,374 patients, or 15.94% developed Long Haul COVID. Figure 6 displays the comorbidities that were [3]1.3x more likely to be found in the Long Haul COVID patients vs. the non-Long Haul COVID patients. What is noticeable is that half of the top hits are comorbidities associated with immune system dysfunction, particularly enriched for

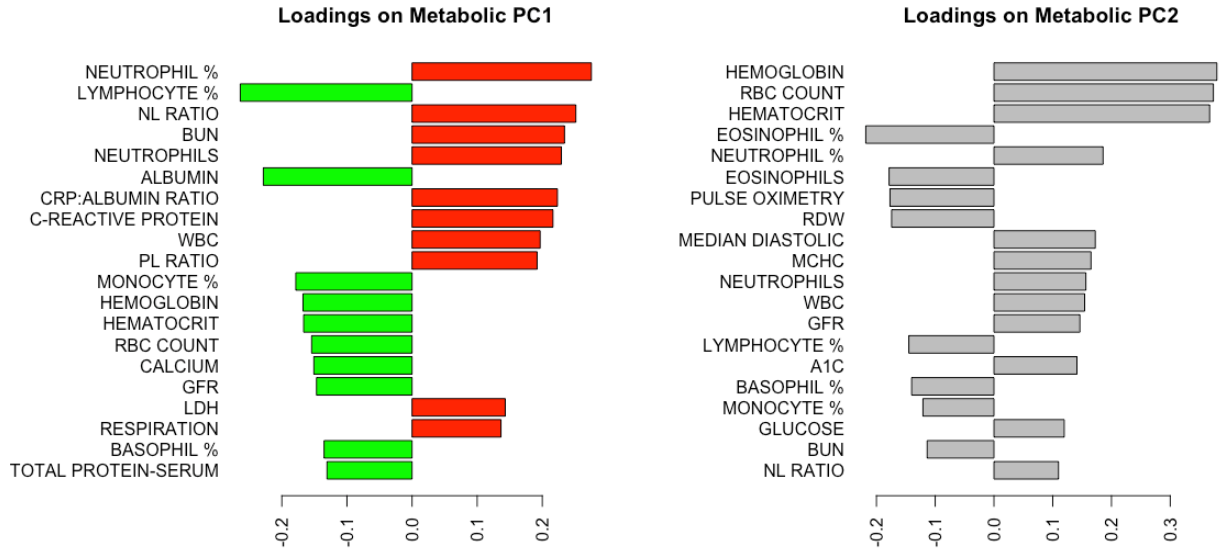
different types of anemia, White Blood Cell dysfunction, coagulation disorders, cirrhosis, Hepatitis B or C, leukemia/lymphoma, mediastinal cancer (which typically involves T cells), vitamin deficiencies that relate to anemia, etc. Note the prominence of Hepatitis B or C and cirrhosis, with 11.37% of Long Haul patients having a chronic Hepatitis B or C infection vs. only 3.38% of non-Long Haul patients. It is therefore clear that patients who have any kind of immunocompromised state should be monitored closely for the development of Long Haul COVID.

Part 2: Patient subphenotyping using PCA Analysis and Prototype Development

Selected Findings

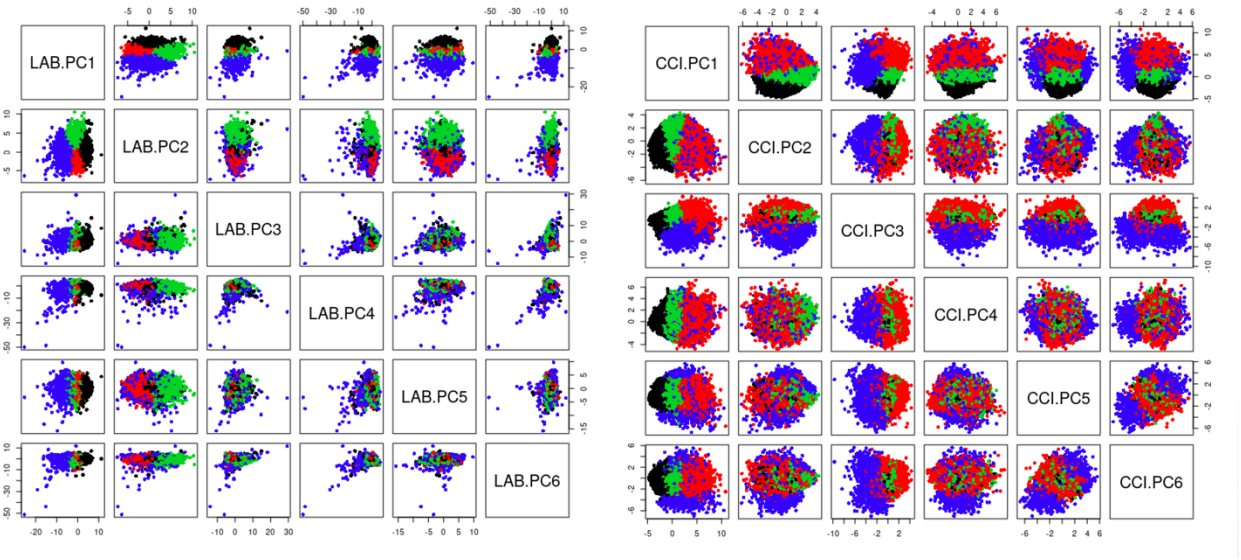
2A. Unsupervised Modeling

Figure 7: Metabolic Factors (Inpatient Labs' Principal Components)



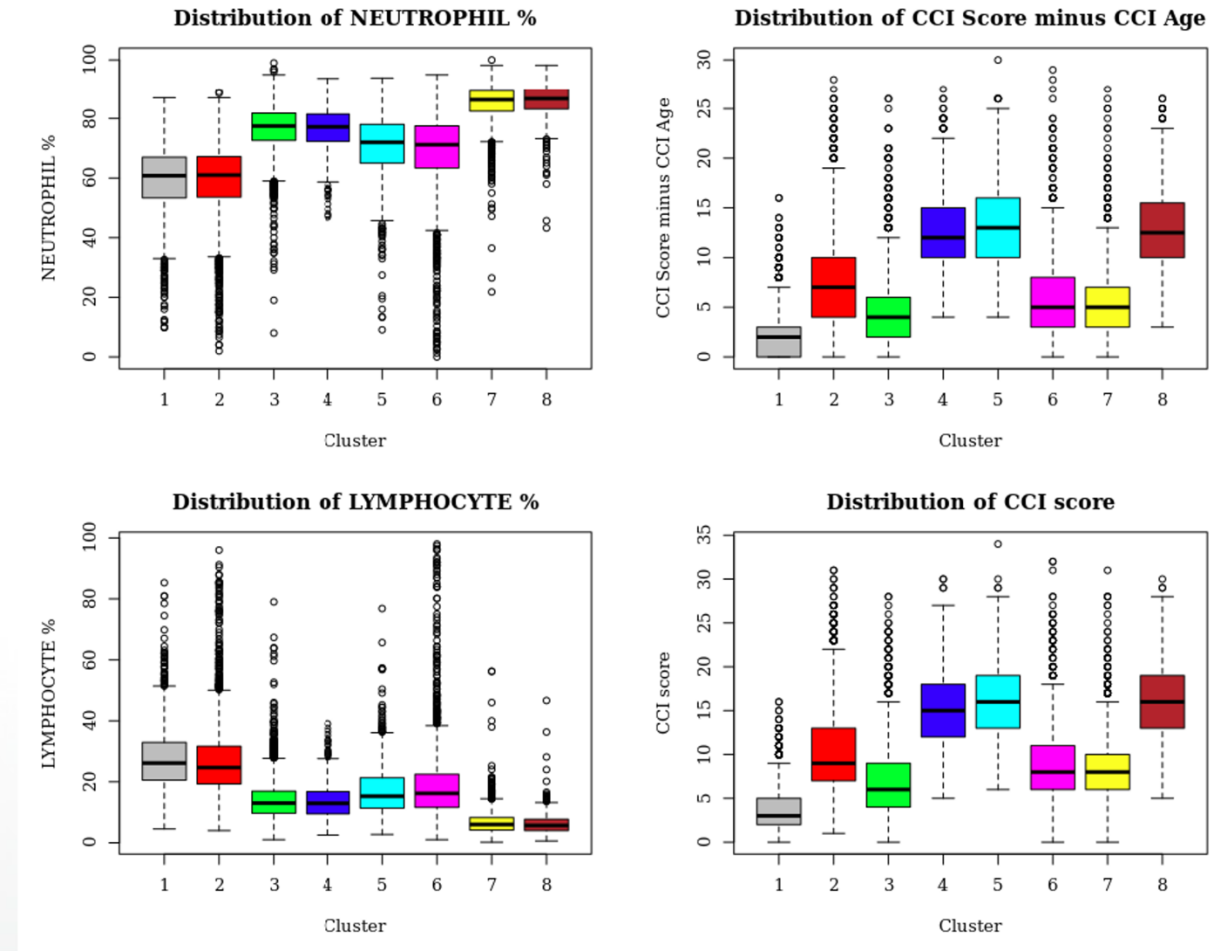
PC1 deserves further discussion. It is well accepted that a patient's age is a strong predictor of all-cause mortality from SARS-Cov-2 infection. Considering the size of the cohort of VA patients, it is worth noting that a model fit on PC1 outperforms predictions using age. Two logistic regressions were performed on PC1 and age, respectively. In Figure 7, the predictions from those regressions are run on the same 20% holdout sample. The predictions are quantiled into 20 ordinal buckets, with the lowest propensity for all-cause mortality in bucket one and the highest propensity in bucket 20. The model on age shows an approximately 4-fold increase in relative risk among the oldest 5% of the holdout sample. The model shows that the 5% with the highest scores on the first metabolic principal component PC1 have a 9-fold increase in relative risk. This analysis demonstrates that unsupervised methods can provide very significant signals that can become powerful predictors when used in supervised methods, for example the criticality of the COVID-positive patient's condition.

Figure 8: Visual of 4 Inpatient Lab Clusters and 4 Comorbidity Clusters of Hospitalized COVID-19 Patients using PCA and K-means



Four clusters showed the most robust outputs for both groups of data elements. The numbers show up with lab clusters of population N_{Lab} = {7908, 9021, 5298, 3412} and N_{Comorb} = {2396, 3259, 11299, 8685}. When overlaying these cluster definitions on the entire set of patients, there are 16 possible combinations of clusters. In order to assess for phenotype differences among these 16 clusters, one-way distributions were compared for all 120 combinations of cluster pairs, e.g., lab cluster 2 & comorbidity cluster 1 compared with lab cluster 3 & comorbidity cluster 3. One example of this type of distributional comparison can be seen in Figure 8. This shows how dramatically different the percent of neutrophils are between the two cluster combinations just mentioned. The red bars represent lab and comorbidity cluster {2,1}, and the gray bars represent lab and comorbidity cluster {3,3}.

Figure 9: Examples of Four Top Features Distinguishing Eight Different Subphenotyped Clusters



2B. Supervised Modeling

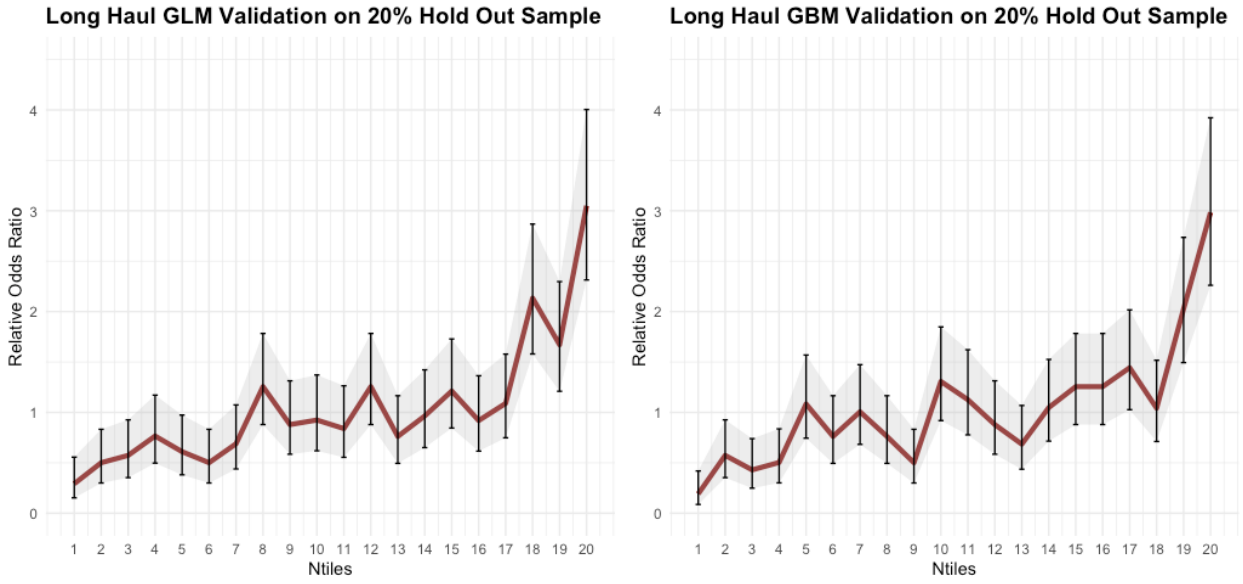
Table 1: Coefficient Estimates for Predictors in the Long Haul GLM

Predictor	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-11.5455	2.0437	-5.6494	<0.0001
Age	0.0059	0.0017	3.3936	0.0007
Atherosclerosis, blood vessel disease	0.1577	0.0523	3.0134	0.0026
CCI.PC1[2]	22.3044	3.0365	7.3454	<0.0001
CCI.PC1[3]	-11.032	2.9697	-3.7149	0.0002
Cirrhosis, liver disease	0.3017	0.0755	3.9934	0.0001
log(1+EOSINOPHIL %)	7.0827	3.3696	2.1019	0.0356
Hep B or C	0.9937	0.0873	11.3806	<0.0001
LAB.PC2[3]	13.0068	4.0699	3.1958	0.0014

[*log*(1+NEUTROPHILS)]^3	-11.0997	4.2347	-2.6211	0.0088
ÖNRBC	0.6312	0.2116	2.9828	0.0029
PULSE.OXIMETRY	0.0343	0.0112	3.0636	0.0022
log(1+RDW)	4.4414	1.7641	2.5176	0.0118
ÖRDW	-1.7053	0.8062	-2.1151	0.0344
RX.ANTIBACTERIALS	0.1433	0.0501	2.8572	0.0043
RX.CNS.AGENTS	0.1255	0.047	2.6685	0.0076
RX.NICOTINE	0.3768	0.1092	3.4502	0.0006
Substance.abuse	0.2291	0.0542	4.2282	<0.0001

The odds ratio plots in Figure 10 suggest that the GLM and GBM perform very similarly with the lowest three probability buckets having an approximately 6.4% chance of long haul, and the highest probability bucket a 31% chance. The fact that the data is not as predictive as the mortality and stroke models suggests that there are more signals for a causal mechanism that are not adequately captured with the current data set.

Figure 10: GLM and GBM Predicted Relative Odds Ratios for Stroke Risk



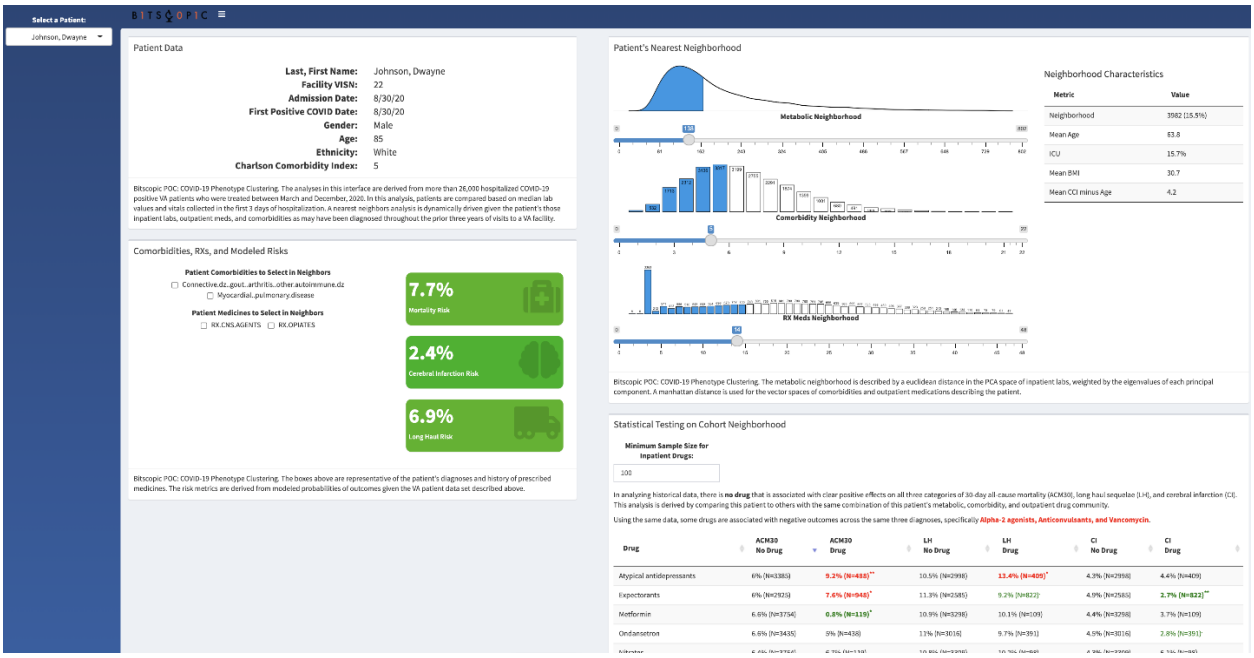
A variable excluded from the model but was found to be statistically significant is the distance to a VA hospital; this value is negatively correlated with the diagnosis of long haul which may imply that patients who are further from their respective facility are less likely to come in for long haul symptoms. In Figure 10, the relative odds ratio is shown as distance from a VA hospital increases. The 10% of patients who live furthest from their closest VA are exactly half as likely to develop a long haul case than the 10% of patients living closest to their nearest VA.

2C. Patient-centered Clustering

In Figure 11, a screenshot of the software prototype to advance the VA's clinical decisioning for COVID-19 inpatient treatments is shown. The use case for this tool is to actualize the VA's epidemiological data to provide statistically driven decisions that support a clinician's pharmaceutical interventions. Using this tool, a clinician can interactively choose a cohort that showed similar vitals, similar comorbidities, and may be taking similar outpatient medications. The screenshot shows four main components:

1. Patient Data – This box has straightforward descriptive elements concerning the patient's demographics.
2. Comorbidities, RXs, and Modeled Risks
    a. The ICD10 codes corresponding to patient diagnoses are normalized and listed in this box with the ability to "force" the requirement that the comparable cohort of epidemiological data shares a comorbidity.
    b. Outpatient medications (where available) are listed in a similar fashion as the comorbidities.
    c. The three colored boxes indicate the predictions of the three models described in Section 2B.
3. Patient's Nearest Neighborhood – This section allows the clinician to interactively weight their preference for the "distance" from the patient in question. It includes four elements:
    a. Metabolic Neighborhood – This slider shows the distribution of Euclidean distances from the selected patient to the rest of the patients in the epidemiological data.
    b. Comorbidity Neighborhood – This slider shows the distribution of Hamming distances from the selected patient to the rest of the patients in the epidemiological data.
    c. RX Meds Neighborhood – This slider shows the distribution of Hamming distances from the selected patient's prescription medications on record to the rest of the patients' indicators for medications in the epidemiological data
    d. Neighborhood Characteristics – This table gives simple counts and averages of the cohort that survives the filters applied using the sliders described.
4. Statistical Testing on Cohort / Neighborhood – This component contains a table of inpatient rates of the targeted outcomes, as well as auto-generated text describing conclusions; these conclusions may be drawn given statistically significant beneficial or poor outcomes across all three targets given the delivery of an inpatient medication during hospitalization associated with a positive COVID-19 diagnosis.

Figure 11: The Decision Support Tool



Executive Summary of Actionable Items and Discussion

In this wide-ranging study of COVID-19 patients in the VA system, the following noteworthy items were recognized in Part 1.

1. Black or African-American VA patients who are COVID-19 positive are more likely to be hospitalized and admitted to the ICU when compared to other ethnicities, but they have the lowest 30-day and 60-day mortality out of all ethnicities. In contrast, Asian VA patients had the highest 30-day and 60-day ICU mortalities of all ethnicities.
2. Across all ethnicities, veterans who are greater than 70 years old are the most likely to be hospitalized. The likelihood of hospitalization increases significantly in patients over 70. The general relationship is that hospitalization is directly proportional to the age of the patient. Of the hospitalized patients, those belonging to minority groups tended to be significantly younger than non-Hispanic White patients. White patients had the lowest number of hospitalized patients under 49 and the highest number of patients over 70 hospitalized out of all ethnicities.
3. Although Black or African-American patients have the lowest mortality among all ethnicities, they have the highest chance of developing a stroke or MI within 100 days of their first COVID-19 diagnosis. Further, they are also more likely than any other ethnicity to have a cancer history and obesity.
4. VA patients from rural areas tended to be older and had a higher 30-day mortality than patients from urban areas. Further, a general trend was that as the distance

from the nearest VA hospital increases, the 30-day mortality rate also increased, potentially signifying the importance of being close to a VA facility. Thus, to make access to healthcare easier for veterans, distance from the nearest VA facility is an important measure to consider when planning future facilities, collaborations with private hospitals, and other actions.

5. All three vaccines—Pfizer, Moderna, and Janssen—are effective at producing very low rates of breakthrough infections. Among Hispanics and African-Americans, there is a tendency for a greater number of breakthrough infections for all three vaccines, particularly when patients did not receive their second Pfizer dose. Across all demographics, the likelihood of breakthrough infections decreases two weeks after the second dose of Moderna or Pfizer is received.
6. We noted that patients with certain comorbidities associated with immune system dysfunction, particularly different types of anemia, White Blood Cell dysfunction, coagulation disorders, cirrhosis, Hepatitis B or C, leukemia/lymphoma, mediastinal cancer (which typically involves T cells), and vitamin deficiencies that relate to anemia, etc. were more likely to develop Long Haul COVID and may require additional observation.

Part 2

7. With a dataset as large as the VA's, simple dimension reductions such as the PCAs shown in Section 2A can reveal structural signals that are highly predictive of the criticality of a patient's infection. These signals, such as the first metabolic principal component described in Section 2A, may also be generalizable to other diagnoses and enable predictive scores highlighting patients' immediate needs. Bitscopic sees opportunities to leverage this technique for many more use cases.
8. The predictive model for all cause mortality was remarkably predictive with more than 80% of the patients in the highest 5% of scores passing within 60 days of admission. This model could be leveraged using Praedico or PraediAlert to identify highly critical patients. While this is a prototype, there is a considerable effort required to ensure usability of the software. Any analytic deliverable is only as useful as the decisions it modifies. Further, any modification of a clinician's decisioning is going to affect, or bias, the forward-looking outcomes that will feedback to the data tool.
9. Examining in more detail the most important features used by the Long Haul Generalized Linear Model *GLM) Predictor (see Table 1), they included a history of atherosclerosis, a history of cirrhosis or liver disease, Hepatitis B or C infection, a high eosinophil %, a low neutrophil count, a high nucleated Red Blood Cell count, and a history of nicotine or substance abuse among others. Together, these indicate that patients who have liver dysfunction, atherosclerosis, and addiction to tobacco or other substances are associated with the development of Long Haul COVID. This implies that patients who have cirrhosis or smokers,

among other things, should be monitored more closely for the development of Long Haul. This is also in concordance with our analysis finding that patients with cirrhosis/hepatitis B/C were more likely to develop Long Covid, for instance.

Limitations

A limitation of the tool as currently developed is that it is essentially observational and therefore associations between certain medications, comorbidities, or lab values are often not going to be causational but correlational. This is part and parcel of this being a tool based on vast amounts of observational data. While the ability to force certain variables allows one to control for datasets containing patients with only those variables is useful, it still does not eliminate confounders. Further data and analysis are necessary to reduce the size of these limitations.

Author contributions

Vafa Bayat, Eugene Geis, Tami Utz, Farshid Sedghi, Renle Chu, Amanda Purnell, Marian Adly, and members of the Long Covid Steering Committee contributed to study conception, design, data analysis, and/or the writing of this report.

Funding

This work was supported by a contract from the VA OHIL.

Declaration of interests and acknowledgements

The authors, with the exception of Amanda Purnell and Marian Adly, are all employees of Bitscopic, Inc. Otherwise they declare no competing interests. The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

References

Bayat, V., S. Phelps, R. Ryono, C. Lee, H. Parekh, J. Mewton, F. Sedghi, P. Etminani, and M. Holodniy. 2020. 'A SARS-CoV-2 Prediction Model from Standard Laboratory Tests', *Clin Infect Dis*.

Bayat, V., R. Ryono, S. Phelps, E. Geis, F. Sedghi, P. Etminani, and M. Holodniy. 2021. 'Reduced Mortality with Ondansetron Use in SARS-CoV-2 Infected Inpatients', *Open Forum Infect Dis*, (accepted).

Benito-Leon, J., M. D. Del Castillo, A. Estirado, R. Ghosh, S. Dubey, and J. I. Serrano. 2021. 'Using Unsupervised Machine Learning to Identify Age- and Sex-Independent Severity Subgroups Among Patients with COVID-19: Observational Longitudinal Study', *J Med Internet Res*, 23: e25988.

Gold, J. A. W., K. K. Wong, C. M. Szablewski, P. R. Patel, J. Rossow, J. da Silva, P. Natarajan, S. B. Morris, R. N. Fanfair, J. Rogers-Brown, B. B. Bruce, S. D. Browning, A. C.

Hernandez-Romieu, N. W. Furukawa, M. Kang, M. E. Evans, N. Oosmanally, M. Tobin-D'Angelo, C. Drenzek, D. J. Murphy, J. Hollberg, J. M. Blum, R. Jansen, D. W. Wright, W. M. Sewell, 3rd, J. D. Owens, B. Lefkove, F. W. Brown, D. C. Burton, T. M. Uyeki, S. R. Bialek, and B. R. Jackson. 2020. 'Characteristics and Clinical Outcomes of Adult Patients Hospitalized with COVID-19 - Georgia, March 2020', *MMWR Morb Mortal Wkly Rep*, 69: 545-50.

Health, RAND. 2015. "Resources and Capabilities of the Department of Veterans Affairs to Provide Timely and Accessible Care to Veterans." In.

Holodniy, M, C Winston, CA Lucero-Obusan, G Oda, A Mostaghimi, JA Pavlin, P Etminani, C Lee, and F Sedghi. 2015. 'Evaluation of Praedico™, A Next Generation Big Data Biosurveillance Application', *Online J Public Health Inform*, 7: e133.

Krishnamoorthy, G., C. Arsene, N. Jena, S. M. Mogulla, R. Coakley, J. Khine, N. Khosrodad, A. Klein, and A. A. Sule. 2021. 'Racial disparities in COVID-19 hospitalizations do not lead to disparities in outcomes', *Public Health*, 190: 93-98.

Schneider, N., K. Sohrabi, H. Schneider, K. P. Zimmer, P. Fischer, J. de Laffolie, and Cedata-Gpge Study Group. 2021. 'Machine Learning Classification of Inflammatory Bowel Disease in Children Based on a Large Real-World Pediatric Cohort CEDATA-GPGE(R) Registry', *Front Med (Lausanne)*, 8: 666190.