8-2021

Predictive modeling of clinical outcomes for hospitalized COVID-19 patients utilizing CyTOF and clinical data.

Onajia Stubblefield

PREDICTIVE MODELING OF CLINICAL OUTCOMES FOR HOSPITALIZED COVID-19 PATIENTS UTILIZING CYTOF AND CLINICAL DATA

By

Onajia Stubblefield
B.S. The University of Louisville, 2019

A Thesis Submitted to the Faculty of the
School of Public Health and Information Sciences of the
University of Louisville
in Partial Fulfillment of the Requirements for the Degree of

Master of Science
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, KY

August 2021

PREDICTIVE MODELING OF CLINICAL OUTCOMES FOR HOSPITALIZED COVID-19 PATIENTS UTILIZING CYTOF AND CLINICAL DATA

By

Onajia Stubblefield
B.S. The University of Louisville, 2019

A Thesis Approved on

July 22, 2021

by the following Thesis Committee:

Dr. Maiying Kong, Committee Chair

Dr. Shesh Rai, Thesis Co-Chair

Dr. Jiapeng Huang

Dr. Riten Mitra

DEDICATION

This thesis is dedicated to friends and family.

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Dr. Maiying Kong, for her weekly meetings with me and the guidance she provided for this project. I would like to thank the other committee members as well: Dr. Shesh Rai and Dr. Riten Mitra for their continuous support and Dr. Jiapeng Huang for providing the clinical data in this analysis. I thank Dr. Anne Geller for her time and guidance with gating and providing knowledge on immunological aspects of the study. I thank Dr. Sarah Bush for helping us navigate the clinical data. I thank Triparna Poddar for assisting with R code and gating. Finally, I would also like to Dr. Jeremy Gaskins for the support throughout my time in the program.

ABSTRACT

PREDICTIVE MODELING OF CLINICAL OUTCOMES FOR HOSPITALIZED COVID-19 PATIENTS UTILIZING CYTOF AND CLINICAL DATA

Onajia Stubblefield

July 22, 2021

In December 2019, an outbreak of a novel coronavirus initiated a global pandemic. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a virus that causes the disease coronavirus disease 2019 (COVID-19). Symptoms of infection with COVID-19 vary widely between individuals. While some infected individuals are asymptomatic, others need more extensive care and require hospitalization. Indeed, the COVID-19 pandemic was characterized by a shortage of hospital beds which presented additional complications in providing adequate care for patients. In this study, we used a combination of T cell population data collected from mass cytometry analysis and clinical markers to form a predictive model of clinical outcomes for hospitalized COVID-19 patients. This paper details the steps and analysis towards the design of the final model including data acquirement and preprocessing, missing data handling via multiple imputation, and repeated imputations inferences.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER I
INTRODUCTION

1.1 Introduction

In December 2019, a new virus emerged that quickly spread around the globe in early 2020: severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). SARS-CoV-2 triggers a clinical disease named coronavirus disease 2019 or, succinctly, COVID-19 (Subbarao and Mahanty, 2020). COVID-19 causes a respiratory tract infection that can affect the upper respiratory tract including sinuses, noses, and throat and/or lower respiratory tract including the windpipe and lungs. Symptoms of the virus include fever, coughing, fatigue, and muscle and body aches (Adil et al., 2013). Interestingly, symptomatic expressions of the virus can vary significantly between individuals. Severity of symptoms can range from asymptomatic yet still potentially contagious to mildly or moderately symptomatic to severe. Severe cases, approximately 20% of all cases, require mechanical ventilation and could result in death (Adil et al., 2013).

SARS-CoV-2 viral infection leads to humoral and cellular responses from the immune system (Shah et al., 2020). B cells are the primary drivers of the humoral response in the human system, developing antibodies that bind to target antigens on viruses. T cells are the primary drivers of cellular response, also known as cell-mediated immunity. In cell-mediated immunity, whose chief purpose is elimination of virally infected cells, naïve T cells that encounter viral antigen proliferate and differentiate to produce memory T cells which rapidly initiate a secondary response upon subsequent infections (Shah et al., 2020). T cell counts have been shown to reflect the severity of COVID-19 (Zheng et al., 2020; Diao et al., 2020; Liu et al., 2020). Studies associating T cells with COVID-19 primarily look at the cell counts. Furthermore, the scope of these studies is commonly limited to the CD4+ and CD8+ populations without examination of their subtypes: memory T cells and naïve T cells.

In addition to T cell populations, previous research has also used laboratory clinical markers to monitor the status of hospitalized COVID-19 positive patients. Generally, severe patients exhibit increased levels of D-Dimers, C-reactive Protein (CRP), ferritin, and lactate dehydrogenase (LDH) over the course of their hospitalization (Kermali et al., 2020). Furthermore, certain comorbidities are associated with increased COVID-19 mortality including lung cancer (Passaro et al., 2021), chronic kidney disease (Sanyalu et al., 2020), and human immunodeficiency virus (HIV) (Ssentogo et al., 2021).

In this study, we collected information on T cell populations from hospitalized COVID-19 positive patients as well as a variety of laboratory values and comorbidity data from their electronic health records. Our goal was to utilize this data and create links to the clinical outcomes (i.e. discharged from hospital, ventilation, transferal to ICU, death, etc.) of these individuals. We hypothesize that measurements performed only at hospital admission narrowly characterize the clinical impact and course of COVID-19. Clinical status, initially measured laboratory values, and immune cell populations fluctuate throughout the duration of infection. Thus, longitudinal monitoring and analysis is essential for a holistic view of the immune system reactions and other physiological responses to SARS-CoV-2 infection. In this paper, we present our modeling of longitudinal COVID-19 clinical outcomes constructed using naïve and memory T-cell population proportions, laboratory values, and comorbidities. This study will hopefully help clinicians determine the most important markers and factors in monitoring and predicting the trajectory of SARS-CoV-2 infection. Ultimately, this could help evaluate the most efficient allocation of resources such as ICU beds and mechanical ventilators.

The structure of the paper will be as follows: in this Introduction, we will provide a brief overview of longitudinal studies and longitudinal data, discussing balanced versus unbalanced data and the advantages and disadvantages of longitudinal studies. We will also provide the definitions of missing data and multiple imputation which we implement as a resolution to missing data. These introductions will lay the foundation for our study.

In the Methods and Materials section, we provide the inclusion and exclusion criteria of the study's subjects, describe how we transformed the clinical outcomes to ordinal data, and detail the attainment of the lab values or clinical markers. In addition, we describe the processes that occur in a mass cytometer which ultimately produce Flow Cytometry Standard (FCS) files to analyze. We then share our gating strategy to collect information on the T cell populations from these FCS files. Lastly, we summarize the cytometry by time of flight (CyTOF) data and statistical analyses methods and packages used.

The CyTOF Data Analysis section of this paper will discuss additional, insightful visualization options of the CyTOF data such as t-SNE, principal component analysis (PCA), and heatmap plots. The visualizations were used cross-check the values obtained from the gating performed in the Methods and Materials section.

The Statistical Analysis body of this paper will cover the statistical theory and methods that form the bulk of our study. Among these include the preprocessing of variables using Pearson correlation, MICE (Multivariate Imputation via Chained Equations) imputation to impute missing clinical marker values, repeated measurements inferences, and creation of mixed effect models for longitudinal data.

The fifth section in the body of the paper will lay out the results of the study, including descriptions of the demographic data overall. We also included tables that summarize the data (Table 1), give the significantly correlated variables (Table 2), and provide the final output of the pooled cumulative link mixed model for ordinal outcomes (Table 3).

The fifth section will provide more in-depth discussion of the results and potential opportunities for future research.

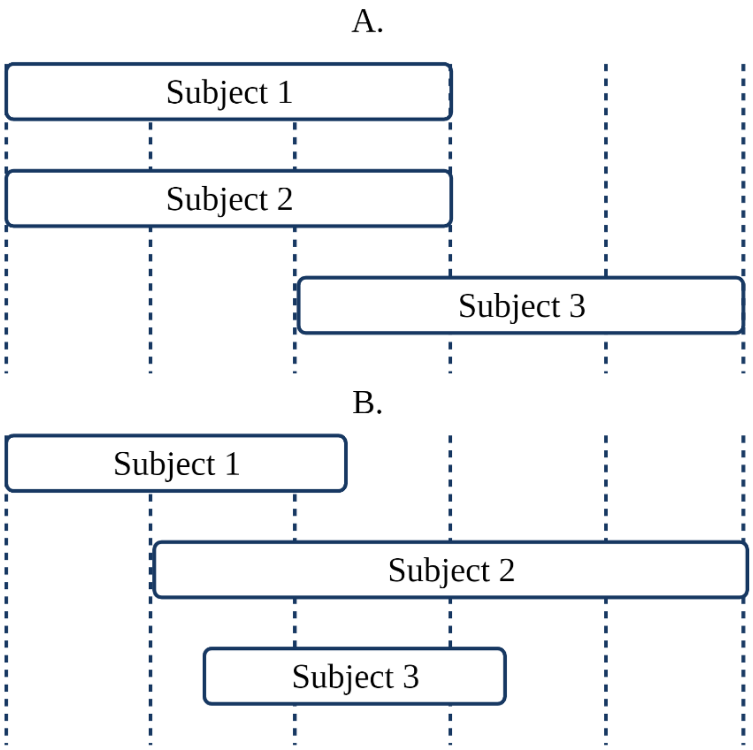
The Appendix of this paper provides additional tables, models, and a glossary related to the study that were deemed noncritical for inclusion in the body of the paper, but useful nevertheless for the reader who may want even more details regarding the analysis.

1.2 Introduction to longitudinal studies

Longitudinal studies are a dynamic approach to scientific investigation. In a longitudinal study, variables from a defined group of individuals are followed over an extended period of time (Coggon et al., 2009, Chapter 7). The purpose of a longitudinal study is to characterize changes in the response of interest over time in relation to the selected covariates. There are variety of types of longitudinal studies including cohort studies, panel studies, and record linkage studies - all of which can be prospective or retrospective (Caruana et al., 2015).

There are two types of design in longitudinal data analysis: balanced and unbalanced. In balanced design, repeated measurements are taken at the same intervals (Liu, 2016). In unbalanced design, the set of time points for the subjects are different (Liu, 2016). In addition to our data being unbalanced with different time intervals between samples, some patients had more observations than others.

Figure 1. Balanced and unbalanced longitudinal data. Panel A. Visualization of balanced longitudinal data; Panel B. Visualization of unbalanced longitudinal data.



Longitudinal studies offer numerous benefits. Among these include the ability to identify and relate events to specific exposures, establishing a sequence of events after exposure, eliminating recall bias in subjects, and monitoring change in particular individuals over the course of time (Caruana et al., 2015). Simultaneously, longitudinal studies come with obstacles as well. The longer a study is, the more likely it is for individuals to follow-up at subsequent times (Caruana et al., 2015). These types of studies are, in general, more costly in terms of financial demands and time requirements (Caruana et al., 2015). To counter this, investigators typically opt for a smaller group of subjects.

1.3 Longitudinal data analysis

A variety of statistical methods have been developed that can accommodate for different classes of response variables in longitudinal data. If the response variable is continuous, linear mixed effect models are commonly used whereas if the response variable is discrete, generalized linear mixed effect models are utilized. Discrete response variables include nominal responsible variables, which contain no quantitative value, and ordinal response variables, in which order matters but the differences between levels is unknown. Treating the clinical outcome as an ordinal response variable, the model of interest for this study was the cumulative link mixed model. We will delve more into this model in the Analysis section.

1.4 Missing data & multiple imputation

In statistics, missing data occur when any observation of interest has no stored data value for any variable. Though a longitudinal analysis can have a balanced design, if missing data is present, the data is unbalanced. Moreover, missing data is common in unbalanced longitudinal studies as well which, if neglected, can reduce the statistical power of study, induce bias in the estimation of parameters of interest, and reduce the representativeness of the samples (Kang 2013). If missing data are mishandled, incorrect inferences about the parameters can be drawn. To prevent these inappropriate inferences, there are different strategies and techniques for handling missing data. In this study, we investigate and utilize a multiple imputation strategy in which plausible data sets are imputed to replace the missing values. The imputed values in these data sets contain the natural variability and estimation uncertainty of the correct values which produces a valid statistical inference.

CHAPTER II

METHODS AND MATERIALS

2.1 Study participants

The Institutional Review Board at University of Louisville approved the present study and written informed consent was obtained from either subjects or their legal authorized representatives (IRB No. 20. 0321). Inclusion criteria were hospitalized adults, age 18 or older, at the University of Louisville Hospital with positive COVID-19 test results and provided consent to this study. Exclusion criteria included less than 18 years of age and/or refusal to participate. COVID-19 patients enrolled in this study were diagnosed with a 2019-CoV detection kit at the University of Louisville Hospital Laboratory using real-time reverse transcriptase–polymerase chain from nasal pharyngeal swab samples obtained from patients. All COVID-19 patients were followed by the research team daily and the clinical team was blinded to findings of the research analysis to avoid potential bias.

2.2 Clinical data & markers

The ordinal ranking of clinical outcomes from 1 to 5 is based on the clinical observation. A clinical outcome assigned the value of 1 meant the patient was discharged the same day. A value of 2 meant the patient was sent to the floor or extubated. If the patient was sent to the ICU that day, the clinical observation was given the value 3. If the patient was intubated, the clinical observation was assigned a value 4. Death on the day of draw was 5. This clinical ordinal outcome is our primary outcome of this study.

The demographic characteristics (age, sex, height, weight, Body Mass Index (BMI), clinical data (symptoms, comorbidities, laboratory findings, treatments, complications and outcomes) and results of cardiac examinations including biomarkers, ECG and echocardiography were collected prospectively. All data were independently reviewed and entered into a computer database. These data were then extracted to form a final data set along with the CyTOF data. The laboratory values can be found in the Discussion section in Table 1. For hospital laboratory CBC (complete blood count) tests, normal values are the following: white blood cell (4.1-10.8 x103 /µL); hemoglobin (13.7-17.5gram/dL); and platelet (140-370 x103 /µL). For hospital laboratory inflammatory and coagulation markers, normal values are the following: D-dimer (0.19-0.74 µgFEU/ml); ferritin (7-350 ng/ml); and lactate dehydrogenase (LDH) (100-242 Units/Liter). The selected comorbidities can be found in Table 1 in the Discussion section. A more thorough description of the attainment of the laboratory values can be found in Morrissey et al. (2020). The clinical outcomes (discharge, mortality, and length of stay) were monitored up to September 16, 2020.

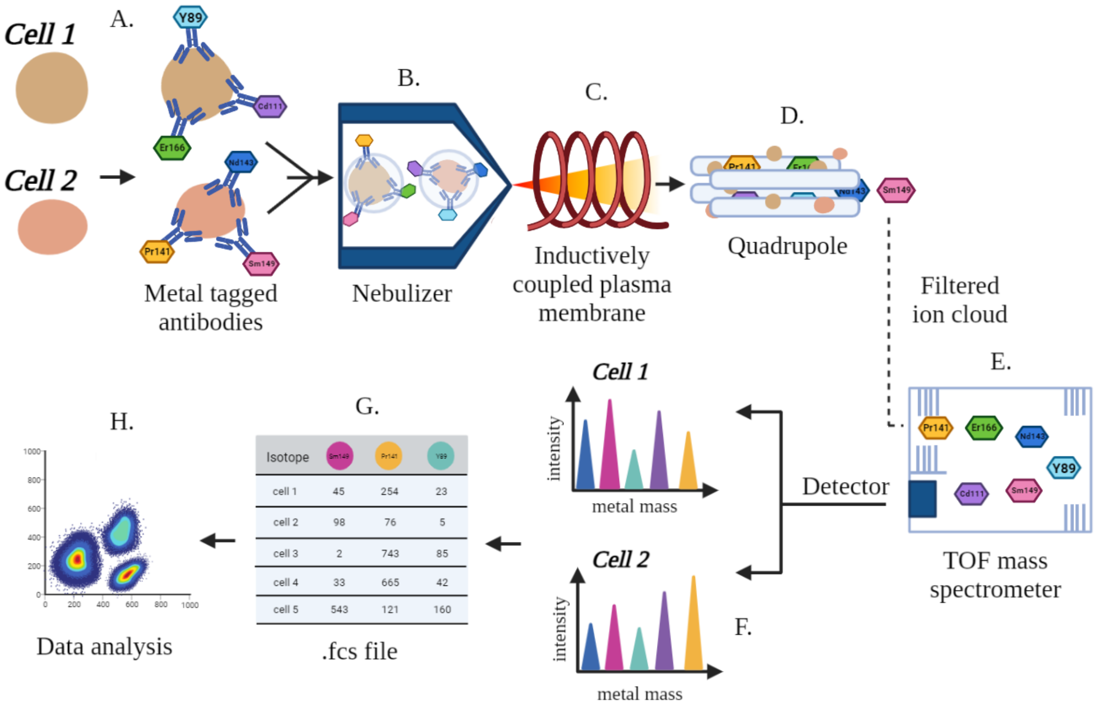
2.3 Mass cytometry overview

Mass cytometry is a next generation flow cytometry platform which uses elemental mass spectrometry to detect and quantify metal-conjugated antibodies that are bound extracellularly and/or intracellularly to components of interests on single cells. 44 markers were used for the mass cytometry analysis depicted in the Appendix section in

Table A1. The description of the CyTOF mass cytometry sample preparation and data acquisition for this study can be found in Morrissey et al. (2020). Here, we provide a brief overview of how mass cytometry works.

Mass cytometry is where we take single cell dispersions and flow them into a series of apparatuses that “burn” the cells and ionize the contents. Mass spectrometry is then used to analyze the cells. In the first step of mass cytometry, cells are labeled with a panel of metal-conjugated antibodies which can target cell surface markers, cellular proteins, or other epitopes under investigation. The cells are then loaded into the mass cytometer where they are injected through a nebulizer. The nebulizer orders the cells in a single line and encases each one in a liquid droplet (Spitzer and Nolan, 2016). These cell-containing droplets are sent to the inductively couple plasma (ICP) membrane where the cells are ionized and burned. This generates a particle cloud created from the atoms of each cell. Afterwards, the cells pass through a quadrupole mass filter where low atomic mass atoms are separated from the high mass ions. The selected high mass ions then enter the time of flight mass spectrometer. In the time of flight mass spectrometer is a detector which quantifies the abundance of each heavy metal present on a per-cell basis. The raw data collected by the detector are converted into an electrical signal which is analyzed by the mass cytometry instrument software to identify cell events (Spitzer and Nolan, 2016). Cell events are simply single cell readings. For each identified cell event, the signal intensity in each channel is quantified and a Flow Cytometry Standard (.fcs) file, which is used for analysis later, is generated from the data (Spitzer and Nolan, 2016).

Figure 2. Workflow of mass cytometry. *Key*: A. Antibodies conjugated to metal isotopes; B. Single-cell suspensions carried into nebulizer; C. Cells pass through a plasma torch; D. Quadrupoles filter and purify ion clouds from the vaporized cells; E. Times of flight (TOFs) of the ions are measured by a detector in the mass spectrometer; F. Individualized cells are profiled by atomic mass; G. FCS file is created to store the information; H. Analysis of the events using the .fcs file.



Isotope	Sm149	Pr141	Y89
cell 1	45	254	23
cell 2	98	76	5
cell 3	2	743	85
cell 4	33	665	42
cell 5	543	121	160

2.4 Gating strategy

Conventional flow cytometers and mass cytometers produce .fcs files that can be manually analyzed using programs such as FlowJo and Cytobank, or computationally using Bioconductor packages such as the flowCore package in R (Ellis et al., 2021). The .fcs file is a data matrix in which every column represents a distinct isotope measured and each row represents a single cell scan of the detector. Using FlowJo, we completed the bottleneck step of analyzing CyTOF data – gating. Gating is the sequential identification and refinement of a cellular population of interest using a panel of markers (Li et al., 2017). It is performed via heat scatter plots such as those depicted in Figures 3 and 4, located at the end of this chapter, which illustrate our strategy. Warm zones colored red and yellow represent a higher density of events, while cool zones colored blue and green represent a lower density of events. In this study, our cellular populations of interest were naïve and memory T cells.

Debris and dust can become trapped in the mass cytometer and result in pseudo-events. In addition, polystyrene beads are intentionally added to monitor the performance of the mass cytometer. To distinguish the cells from debris, dust, and beads, Iridium-191 is used to tag DNA while beads are tagged with Cerium-140. In Figures 3 and 4, events with high expression of the isotope are further right on the x-axis or upwards on the y-axis. We then gate on the cells.

From the cells, we select the live cells. Dead cells have compromised membranes which allows to the cell marker cisplatin or Platinum-195 to react with the dead cells' proteins. Thus, cells with high expression of cisplatin are dead. Unlike the other gating selections, we select the cells with low expression of the marker or lower on the y-axis.

At times, cells will cohere to each other in the mass cytometer and be read as a single event. Thus, we gate on true singlets or single cells. Singlet events are characterized by shorter event lengths and less DNA in FlowJo.

Immune cells are characterized by the CD45 antigen. After gating on the CD45+ singlets, we have finished the "preprocessing" for gating on T cells and proceed to gate for the naïve and memory T cells.

We distinguish T cells from other immune cells by selecting immune cells with high expression of the CD3 antigen. From there, we can see two distinct sets of populations: CD4+ and CD8+ T cells. Finally, we gated on the CD4RO+ (memory T cells) and CD45RA+ (naïve T cells) populations from CD4+ and CD8+ T cells.

2.5 CyTOF data analysis

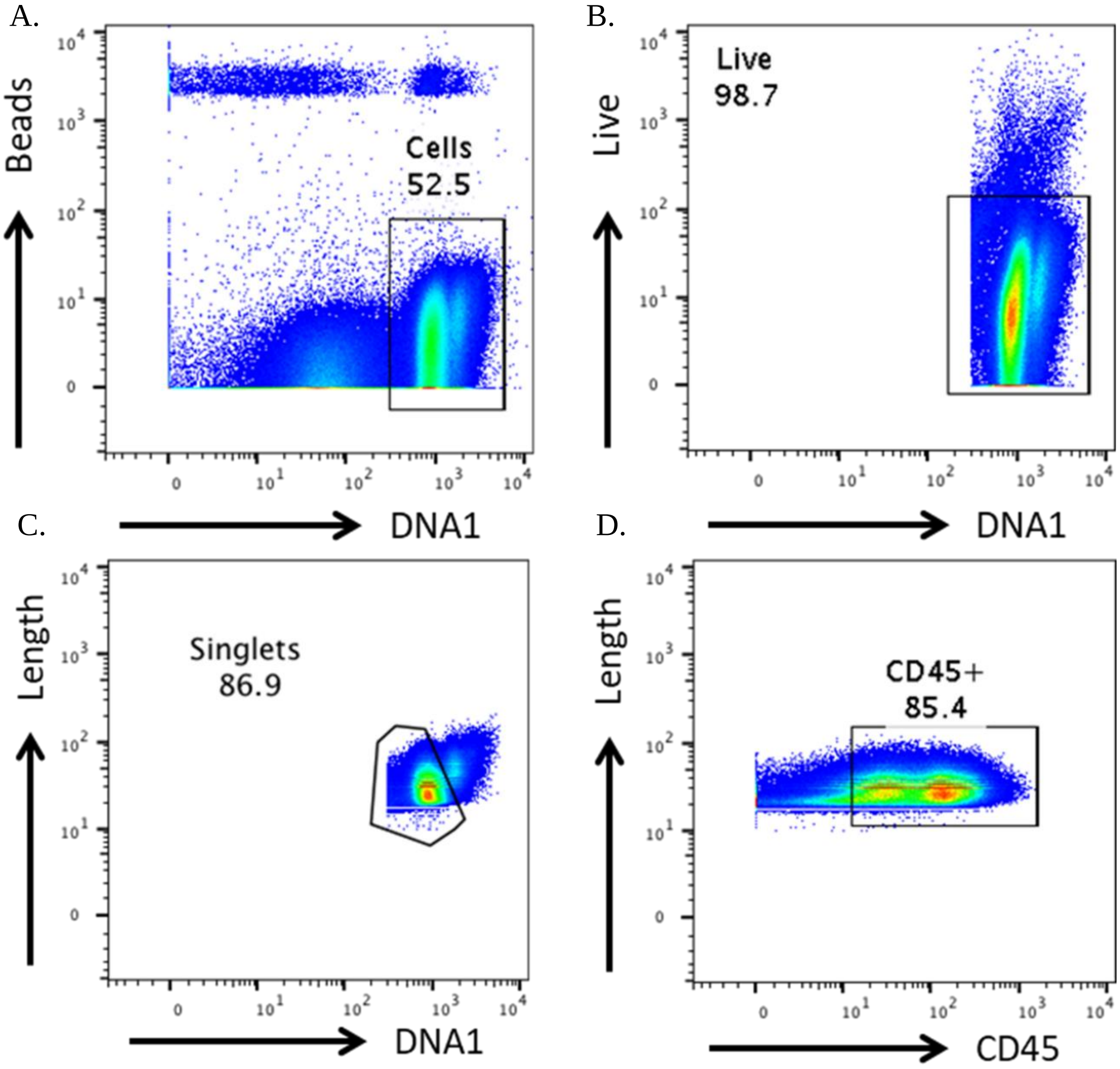
In addition to FlowJo, CyTOF data was also analyzed using an R-based pipeline from Bioconductor.org (Nowicka et al., 2017). Patients with observations in both moderate and severe status were selected for clustering. Moderate was defined as hospitalization without mechanical ventilation and severe was defined as hospitalization and required mechanical ventilation. FlowJo workspace files were imported into RStudio using the read.FCS function within flowCore (Ellis et al., 2021). An arcsinh transformation with a cofactor of 5 was applied to the data using the “apply” function within flowCore (Ellis et al., 2021). Diagnostic plots which included histograms of the marker expressions and a principal component analysis plot were created. Cell population clustering was conducted using FlowSOM (Gassen et al., 2015) and ConsensusClusterPlus (Wilkerson and Hayes, 2010) within the Bioconductor CyTOF workflow (Nowicka et al., 2017). A heatmap was then used to visualize the characteristics of the identified cell clusters using the median marker expression in each cluster. Last, t-SNE plots were used as a dimensionality reduction measure to examine the CD4+ and CD8+ T cell populations.

2.6 Statistical analysis & modeling

The statistical analyses were carried out in the statistical software R (https://www.r-project.org/). A statistical test was claimed significant if $p < 0.05$. First, the five number summary statistics were presented for each appropriate variable.

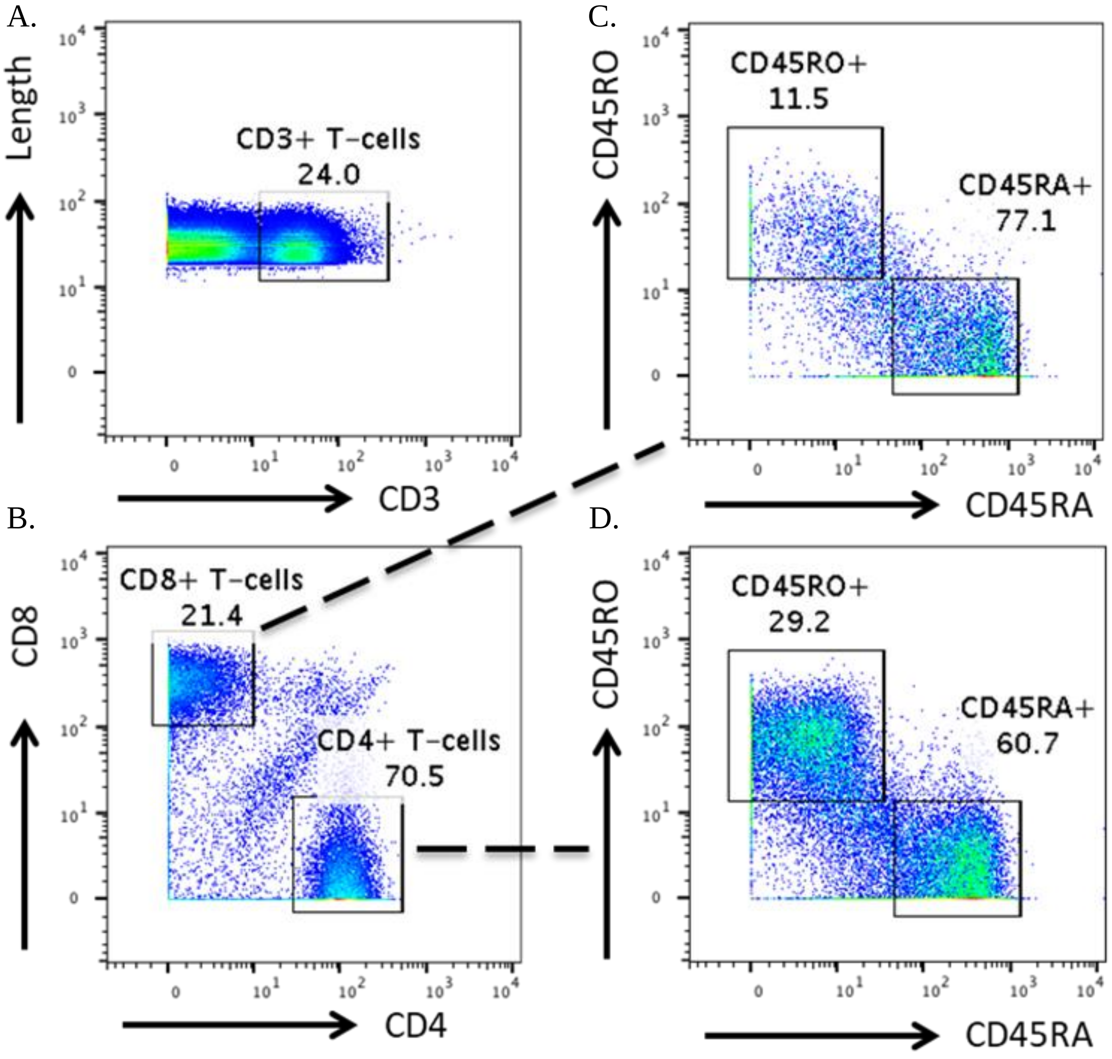
Percentages of patients were computed for comorbidities. This information is presented in Table 1. PCA plots were created to characterize the patients by their comorbidities. For feature selection, we examined the association among the variables using the marginal Pearson correlation coefficient and tested its significance. The marginal Pearson correlation coefficient captures the association between two variables at the population level. Multiple imputations for the missing data were carried out using the mice package (van Buuren S, 2011). Since we have varied number of observations for each patient and an ordinal response variable, we applied cumulative link mixed models to the imputed data sets. The clmm2 function within the ordinal package (Christensen, 2019) fit cumulative link mixed models with random effects, where patients were considered random effects. The results of each of the models were pooled using repeated imputation inferences.

Figure 3. Gating strategy for CD45+ immune cells. Panel A. Gate on cells using Beads (y-axis) and DNA1 (x-axis); Panel B. Gate on live cells using Live (y-axis) and DNA1 (x-axis); Panel C. Gate on singlets using Length (y-axis) and DNA1 (x-axis); Panel D. Gate on CD45+ cells (immune cells) using Length (y-axis) and CD45 (x-axis).



Notes: Warm zones, colored red and yellow, represent a higher density of events, while cool zones, colored blue and green, represent a lower density of events. Population proportions of the total events displayed are indicated under the gate names.

Figure 4. Gating strategy for CD45RO+ and CD45RA+ T cells. Panel A. Gate on T cells (CD3+) using Length (y-axis) and CD3 (x-axis); Panel B. Gate on CD8+ and CD4+ T cells using CD8 (y-axis) and CD4 (x-axis); Panel C. Gate on CD45RA+ and CD45RO+ populations of CD8 T cells; Panel D. Gate on CD45RA+ and CD45RO+ populations of CD4 T cells.



Notes: Warm zones, colored red and yellow, represent a higher density of events, while cool zones, colored blue and green, represent a lower density of events. Population proportions of the total events displayed are indicated under the gate names.

CHAPTER III

CYTOF DATA ANALYSIS

3.1 Inspection of CyTOF data samples

When analyzing sample data, it is important to verify that the samples are representative of the population. One way to check if this was true for the CyTOF data in our study was to compare the moderate samples to the severe samples. To do this, we took moderate and severe samples from patients who had both conditions at separate points in time. These patients were patients 3, 4, 5, 8, and 37. In addition, we would need to inspect if the marker expressions had any abnormalities, particularly inconsistent ranges or dissimilar distributions for a subset of samples (Nowicka et al., 2017). These could suggest issues with data collection or batch effects (Nowicka et al., 2017).

For the latter issue, histograms were created for the markers of the more generalized cell populations frequently investigated, colored by the condition of the patient during the draw. For the former, a principal component analysis (PCA) plot was created to show the relationships between samples based on marker expressions. PCA is an unsupervised dimensionality-reduction method that transforms a set of n-dimensional vector samples $X = \{x_1, x_2, \dots, x_m\}$ into another set $Y = \{y_1, y_2, \dots, y_m\}$ of the same dimensionality (Kantardzic, 2020, p. 80). However, Y contain most of the information in the first few dimensions - the principal components (Kantardzic, 2020, p. 80). Thus, we can reduce the dimensions of the data with a low loss of information (Kantardzic, 2020, p. 80). The technique can be summarized in 6 steps. First, we remove the labels from the data set (i.e. patient ID and draw number, condition). Next, we use the mean from every dimension of the new data set to compute the covariance matrix. The eigenvectors and corresponding eigenvalues of the covariance matrix are then computed. Afterward, we select j eigenvectors with the j largest eigenvalues to form a n×j dimensional matrix A, where n is the dimension of the original samples (Kantardzic, 2020, p. 80). Finally, A is used to transform X into Y. The resultant PCA plot is shown in Figure 6. The first notable observation is that there was less variation between moderate samples than the severe samples. The second notable observation is that, with the exception of patient 4 at draw 1, there is clear separation of the moderate vs. severe samples. This demonstrates what we expect globally: there is a difference in marker expressions and thus immune cell populations between moderate and severe cases of COVID-19.

Figure 5. Distributions of cell marker expression

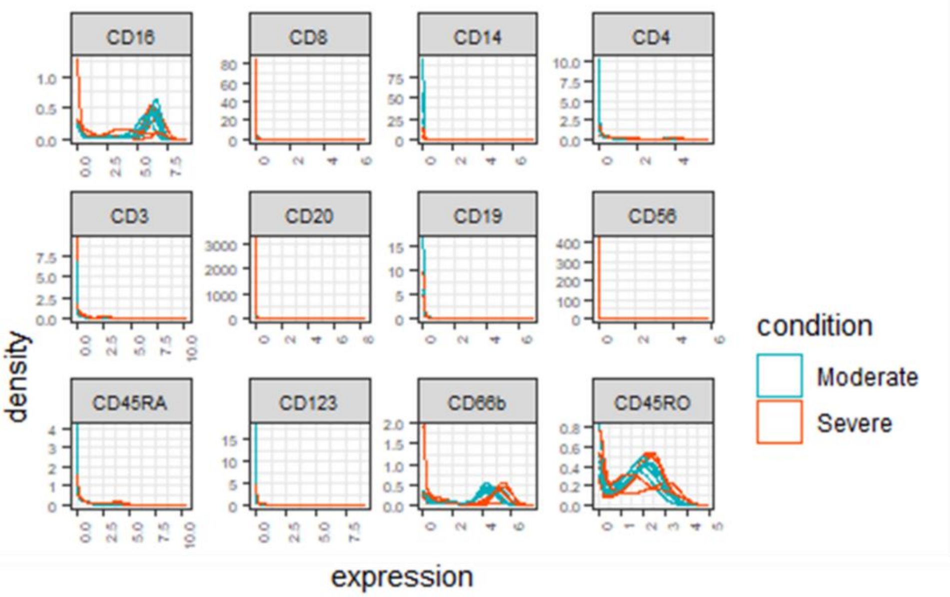
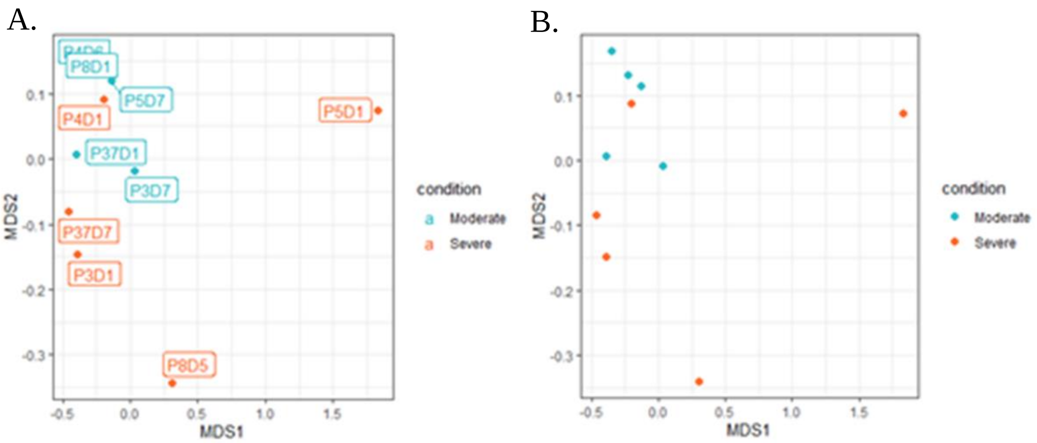


Figure 6. Principal component analysis plots of patient sample



Notes: The PCA plots are the same. Panel A includes the patient labels (P) and draw (D). Moderate observations are blue and severe observations are orange.

3.2 T cell population identification

Two disadvantages of manual gating are bias from the scientist and obscure separation of populations in the density plots. The FlowSOM and ConsensusClusterPlus packages from Bioconductor.org employ clustering techniques to help eliminate these concerns (Nowicka et al., 2017). Thus, we considered a cross-check necessary.

In the first step of FlowSOM, the data is read and preprocessed. The samples are combined into one data matrix with the markers serving as the columns and the events as the rows. The values for each column are scaled according to Z score normalization.

$$z_{ij} = \frac{c_{ij} - mean(c_{1j}, \dots, c_{nj})}{stdev(c_{1j}, \dots, c_{nj})}$$

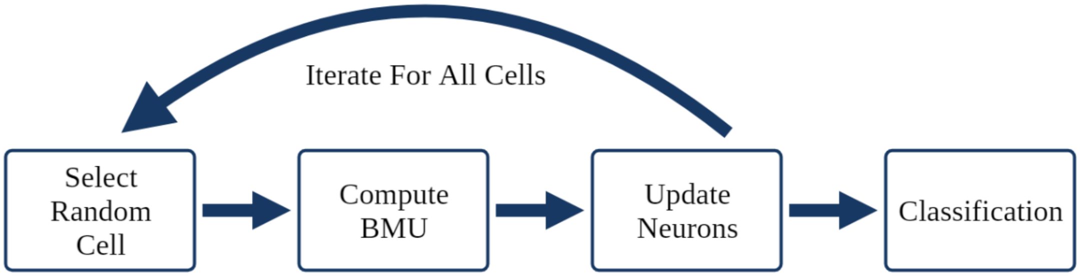
With this transformation, each column now has a mean of 0 and a standard deviation of 1 (Gassen et al., 2015). Now, each marker will hold the same weight in subsequent steps while the differences between ranges within the markers have still been preserved.

A self-organizing map (SOM) is then created using the function BuildSOM. Like PCA, a self-organizing map is an unsupervised clustering technique that performs dimensionality reduction. In this artificial neural network, cells are assigned according to their likeness to 100 grid points of the SOM using an algorithm. The algorithm is as follows: we have 100 d-dimensional nodes or “neurons”. We initialize the nodes with random cells of the dataset (Gassen et al., 2015). We use the Chebyshev distance to define a neighborhood function in the two-dimensional network (Gassen et al., 2015).

$$dist(A,B) = \max\left(|x_A - x_B|, |y_A - y_B|\right)$$

The SOM then learns in an iterative fashion. Data points are recursively selected and matched to the nearest node. The nearest node, termed the Best Matching Unit (BMU), and its neighborhood of nodes are moved closer to the data point. Neighbors that are farther away will shift less. The distance moved by the BMU is the learning rate, α, and the radius of the BMU is the size of the neighborhood, ε (Gassen et al., 2015). These are decreased as the algorithm iterates (Gassen et al., 2015). Upon conclusion of the algorithm, each cell is assigned to the node it most resembles which provides the final clustering (Gassen et al., 2015).

Figure 7. Self-organizing map algorithm



Lastly, meta-clustering of the SOM grid points is completed by the ConsensusClusterPlus function utilizing consensus hierarchical clustering. This method works by subsampling the points repeatedly and generating a hierarchical clustering for each subsampling (Gassen et al., 2015). The frequency with which the same points are clustered together is then used to determine the final clustering (Gassen et al., 2015).

There are two categories for hierarchical clustering algorithms: divisible algorithms and agglomerative algorithms. A divisible algorithm starts from the entire data set and divides the set into a partition of subsets. These subsets are divided into smaller sets, and so on. In an agglomerative approach, each data point or node center is its own initial cluster. The two closest clusters are identified and merged into one cluster. As the process repeats, the clusters are merged into broader divisions and continue to “agglomerate” until all objects are grouped into one cluster. Divisible algorithms could be thought of as top-down approaches while agglomerative algorithms could be considered bottom-up approaches.

Average linkage, a type of agglomerative approach, was used by ConsensusClusterPlus to create the hierarchical clusters. In average linkage, we utilize the mean inter-cluster dissimilarity by computing all pairwise distances between the observations, x, in cluster A and cluster B and recording the average.

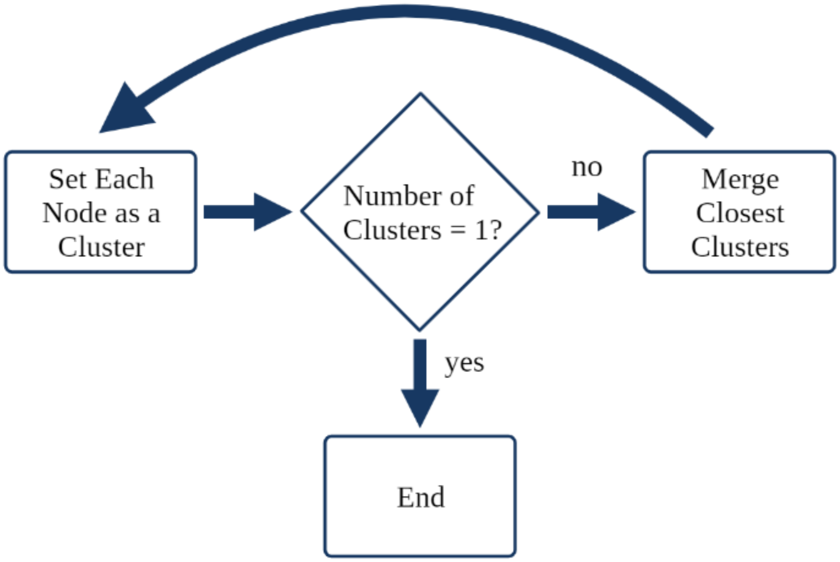
$$L(r,s) = \frac{1}{n_a n_b}\sum_{i=1}^{n_a}\sum_{j=1}^{n_b} D(x_{ai}, x_{bj})$$

Euclidean distance defined below served as the distance measure between points:

$$dist(A,B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

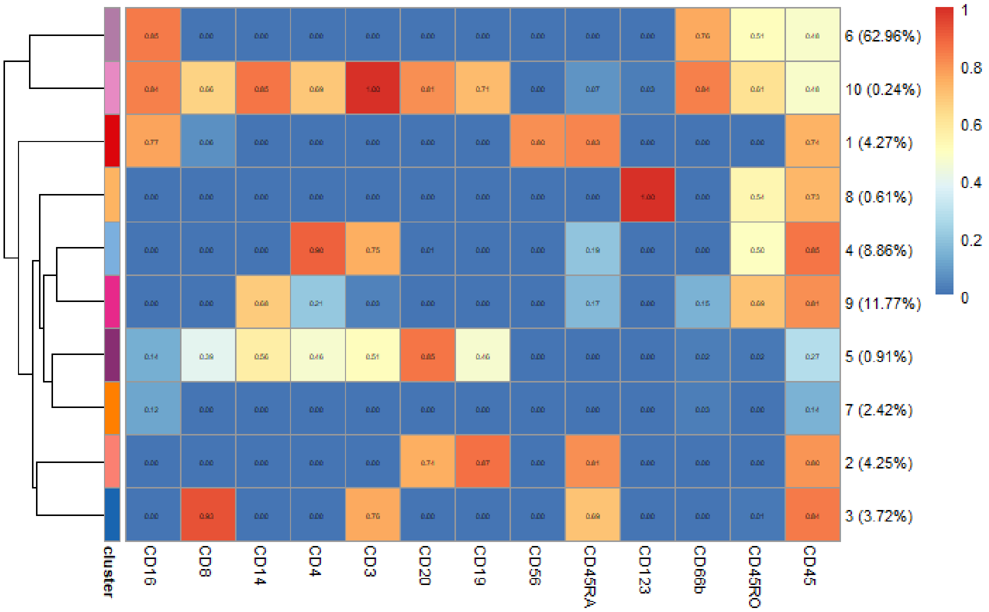
We can then specify the number of clusters and ConsensusClusterPlus function will cut the computed hierarchical clustering tree - the dendrogram - at the appropriate dissimilarity level, forming a partition. The selected number of clusters was 10. The hierarchical clustering map is depicted in Figure 9.

Figure 8. Hierarchical clustering algorithm



T-distributed stochastic neighbor embedding (t-SNE) is a powerful tool to visualize and explore single cell data. Cells in similar local neighborhoods of a high-dimensional space are clustered together in a low-dimensional space, usually two-dimensional. These clusters are then identified as specific cell types such as T cells, B cells, and neutrophils, among others. T-SNE is typically preferred over PCA for single cell analysis because, as Kobak and Berens (2019) state, it preserves the local structure better. We are more interested in clustering cells with high similarity, indicating a cell population, than the distance between cell populations. For the latter component, the global structure of the data, t-SNE performs worse than PCA (Kobak and Berens, 2019).

Figure 9. Heatmap of cell marker expression



Notes: the heatmap is colored according to the median of the arcsinh transformed values in the clusters. Higher medians are red, lower medians are blue.

In the first step of t-SNE, the algorithm computes the similarities between points in the high dimensional space. Similarity between two points of the data is the conditional probability that x_i would select x_j as its neighbor if neighbors were picked in proportion to their probability under a Gaussian centered at x_i (Van der Maaten and Hinton, 2008). These probabilities are then renormalized by dividing by the probabilities of x_i to all x_j.

$$p_{j|i} = \frac{\exp\left(-||x_i - x_j||^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-||x_i - x_j||^2/2\sigma_i^2\right)}, \qquad and$$

$$\sum\nolimits_j p_{j|i} = 1$$

Now, $p_{i|j} \neq p_{j|i}$ so we define

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

The second step is similar to the first step. However, instead we use a Student t-distribution with a single degree of freedom, also known as the Cauchy distribution. Van der Maaten and Hinton (2008) discuss that the Student t-distribution has greater kurtosis, or heavier tails, than the Gaussian distribution. This reduces crowding and allows for better modeling of highly separated points. The result is a second set of probabilities, Q_{ij}, in the low dimensional space:

$$Q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_k \sum_{l \neq k}(1 + ||y_i - y_j||^2)^{-1}}$$

In the final step of the algorithm, we minimize the Kullback-Leibler divergence between the conditional probabilities using gradient descent.

$$KL(P \,||\, Q) = \sum_{i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}}$$

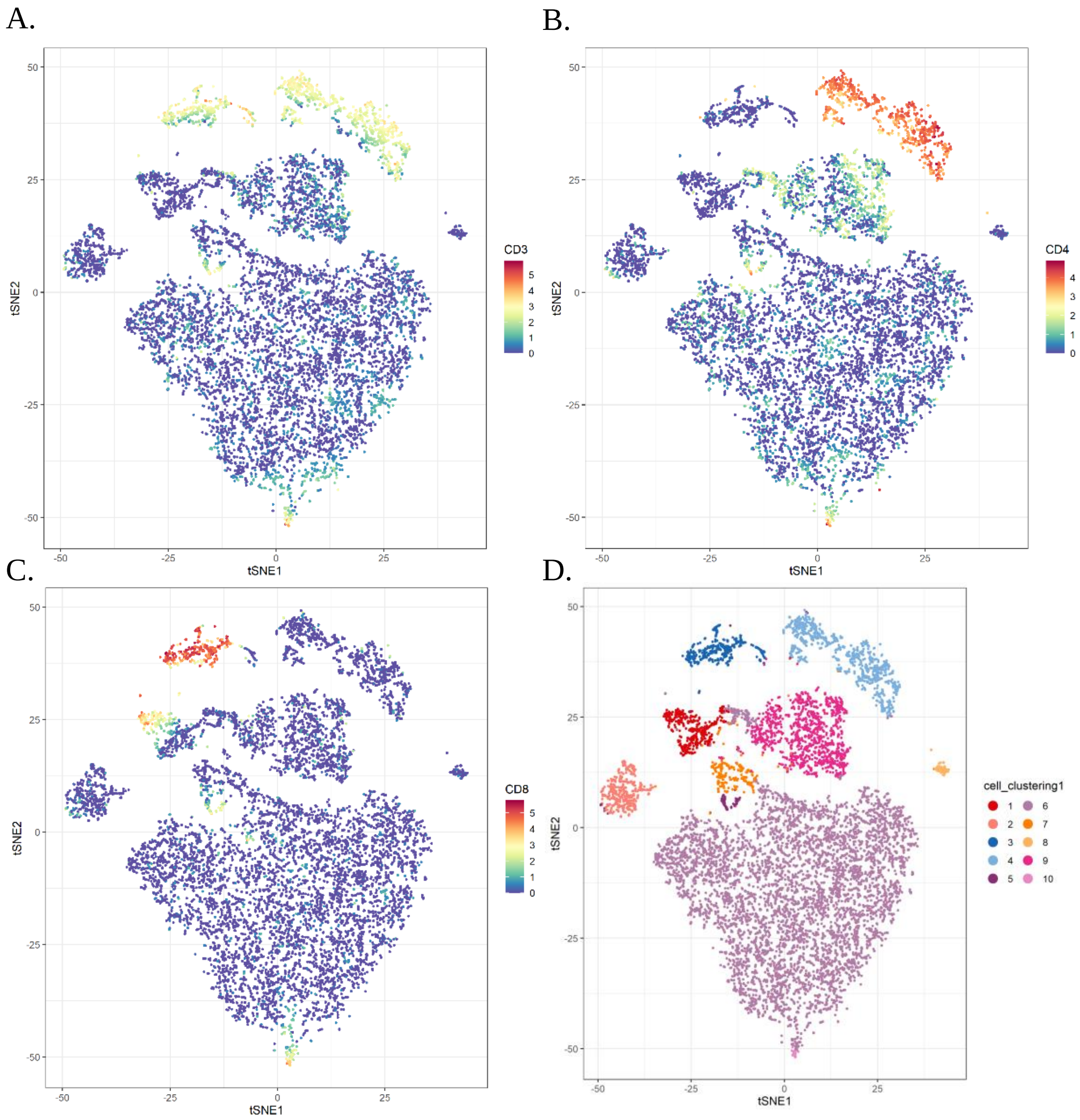
With this, as Kobak and Berens (2019) affirm, the set of probabilities from the low-dimensional space, Q_{ij}, are made to reflect those of the high-dimensional space, P_{ij}. From this optimization, the locations of the cells in the low dimension are used for visualization.

We can use a collection of markers to highlight where cell types of interest are located on the t-SNE map. The t-SNE maps in Figure 10 are colored according to the expression level of their respective marker. With consultation from an immunologist, clusters 3 and 4 and were identified to be CD8 and CD4 T cells, respectively. The heatmap shows that that these populations formed 3.72% and 8.86% of the immune cells.

These values were cross checked with the gates obtained from the analysis in FlowJo.

With some variation, our percentages were consistent with the clustering.

Figure 10. T-SNE plots of the sampled cell populations. Panel A. CD3+ cells are indicated at the upper portion of the t-SNE plot; Panel B. CD4+ cells (cluster 4) are in the upper right portion of the t-SNE plot; Panel C. CD8+ cells (cluster 3) are in the upper left portion of the t-SNE plot; Panel D. Clusters 3 and 4 are identified as the CD8+ and CD4+ T cells, respectively.



Notes: Figures10A – 10C are colored according to the arcsinh transformed values. Higher expressions are red/yellow, lower expressions are blue. Figure 10D colors the cells by the formed clusters.

CHAPTER IV

STATISTICAL ANALYSIS

4.1 Comorbidity selection

High-dimensional data can contain a substantial amount of irrelevant information. In statistics, we call this information “noise”. Noise can lower the quality of the analysis and contributes to the “curse of dimensionality”. It can also lead to overfitting of a model when it is taken as concept to be learned by the algorithm and, thus, the model will fail to generalize to the population or other data sets as well. Finally, with the inclusion of additional variables, this further decreases the comprehensibility of the model.

There were 262 comorbidity factors within the clinical data, where 1 indicates presence of a certain comorbidity and 0 indicates absence of the comorbidity. To reduce noisy variables, comorbidities with less than a count of two for all patients were automatically excluded from the data. This left a total of 88 comorbidity factors. We further reduced the number of comorbidities by excluding those present in only 5% or less of the patients. The comorbidities at the end of the selection were diabetes, hypertension, hyperlipidemia, coronary artery disease, COPD, asthma, and obstructive sleep apnea. We use the 88 comorbidities and the final selected comorbidities to illustrate the dimension reduction with Multiple Correspondence Analysis (MCA).

Although PCA could be used for dimension reduction, PCA is used on continuous variables. MCA is used for categorical or nominal variables. Greenacre and Nenadic (2015) explain the attainment of the principal coordinate using the Burt matrix. In MCA, each factor, K, has $\mathrm{J_k}$ levels where $\sum_{\mathrm{k}} \mathrm{J_k} = \mathrm{J}$. We denote the index of i as the i^{th} observation. So then, Z = $\{z_{ij}\}$ is the indicator matrix for the comorbidities. B is the Burt matrix and is acquired from B = $Z^T Z$ (Greenacre and Nenadic, 2015). To complete the MCA, there are 4 following steps. In the first step, we compute the correspondence matrix P by dividing B by its grand total $n = \sum_{i,j} b_{ij}$ (Greenacre and Nenadic, 2015):

$$P = \{p_{ij}\} = \left\{\frac{b_{ij}}{n}\right\}.$$

We also compute the row totals r_i.

Second, an eigenvalue-eigenvector decomposition is completed on standardized residuals, S (Greenacre and Nenadic, 2015):

$$S = \{s_{ij}\} = \frac{p_{ij} - r_i r_j}{\sqrt{r_i r_j}}.$$

The decomposition returns the eigenvectors, E = $\{e_{is}\}$, and eigenvalues λ_s from the solution of S = $V \Lambda V^T$ (Greenacre and Nenadic, 2015).

The ith row (or column) standard coordinate for the sth dimension is obtained as

$$a_{is} = \frac{v_{is}}{\sqrt{r_i}}$$

Finally, Greenacre and Nenadic (2015) state that principal coordinates from MCA can be obtained from

$$f_{is} = a_{is}\lambda_s$$

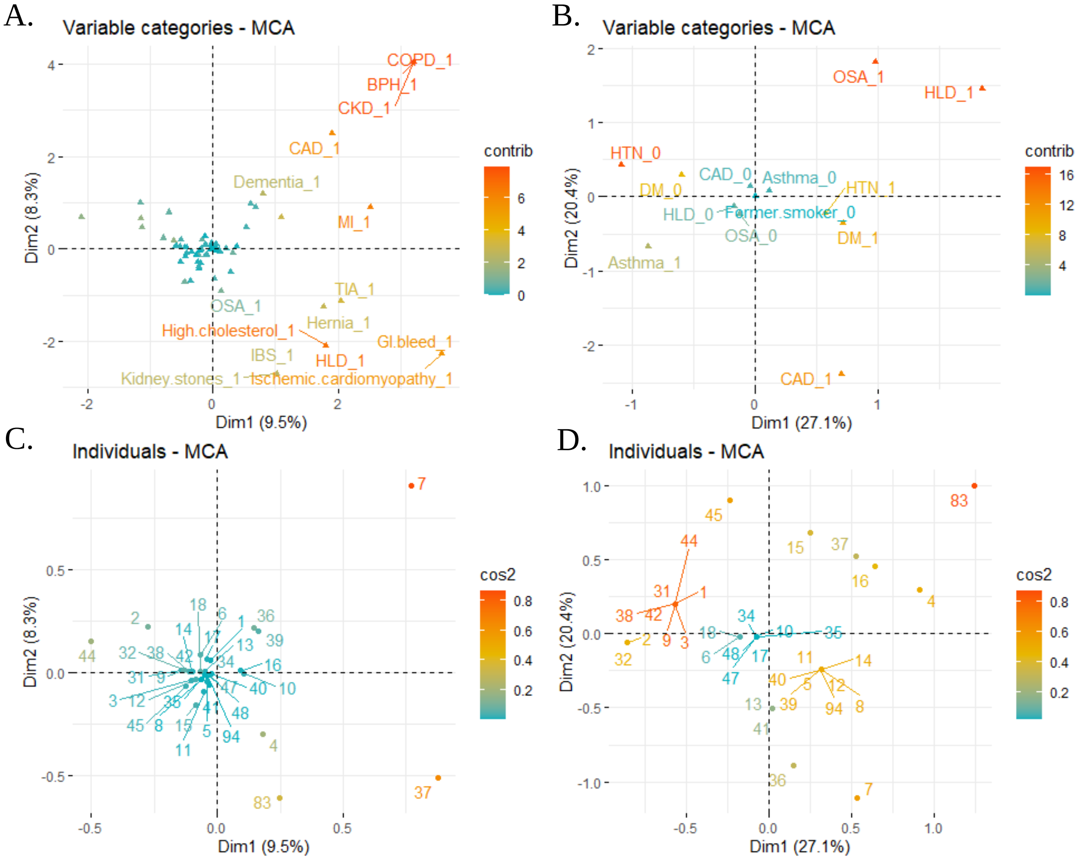
The quality of the representation is measured by the squared cosine (cos2). The squared cosine measures the degree of association between the factors or individuals in the MCA plot and the two dimensions. The squared cosine is calculated by dividing the absolute contribution to an axis by the sum of its absolute contribution to all axes in the analysis. The squared cosine for row and columns, respectively, are

$$\theta^2 = \frac{f_{il}^2}{\sum f_{il}^2} \text{ and } \theta^2 = \frac{g_{il}^2}{\sum g_{il}^2}$$

where we have row i, column j, and factor l (Salkind, 2010).

Comparing the variables squared cosine plots in Figure 10, many of the 88 comorbidities overlapped and did not provide much contribution (if any) to the dimensions. Using the 7 selected comorbidities, there were higher contributions on average from each of the variables. Both plots appear to be separated into two halves: absence of comorbidities on the left and the presence of comorbidities on the right. Examining the individuals MCA plots, we see three clusters of patients from the updated group of comorbidities that were not visible using the 88 comorbidities: those without hypertension or diabetes (red), those with hypertension and diabetes (orange), and those with a different combination of the two that separates them from the former two (blue).

Figure 11. Multiple correspondence analysis, contribution, and squared cosine. Panel A. PCA map of the 88 comorbidities; Panel B. PCA map of the 7 selected comorbidities; Panel C. PCA map of the patients using the 88 comorbidities; Panel D. PCA map of the patients using the 7 selected comorbidities.



Notes: Variables and individuals are colored according to the cos2 values. Red signifies a higher cos2 while blue signifies a lower cos2. A sum of cos2 close to one means a variable or patient is well represented by the two dimensions.

4.2 Correlations

Multicollinearity is the occurrence of significant dependency or association between two or more independent variables (Kim, 2019). A high intercorrelation between predictor variables is indicative of multicollinearity in the data. Multicollinearity can produce skewed results which result in erroneous interpretations of the data. Though some models can account for multicollinearity and will remove one of the correlated variables (Kim, 2019), computing the correlations between the independent variables is often one of the first steps in data analysis. To avoid multicollinearity influencing the model in our study, Pearson’s correlations were computed.

Pearson’s correlation measures the strength of the linear relationship between a pair of variables (Kirch, 2008). For Pearson's correlation coefficient, the null hypothesis is that the correlation between a pair of variables, ρ, is equal to 0 and the alternative hypothesis is that the correlation is not equal to 0.

$$\mathrm{H}_0: \mathrm{\rho} = 0 \text{ vs. } H_1: \mathrm{\rho} \neq 0$$

The correlation coefficient assumes a value between −1 and +1. If one variable trends towards decreasing as the other increases, the correlation coefficient is negative. Conversely, if the two variables tend to increase conjunctively the correlation coefficient is positive. The closer the correlation coefficient is to -1 or +1, the stronger the trend. For the variables x and y, the correlation statistic, r, is calculated by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{(\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where n is the number of observations with values of x_i and y_i for the ith individual (Mukaka, 2012).

The p-value of a statistical test is the probability of obtaining test results equally or more extreme than the observed results under the condition that the null hypothesis is correct (Greenland et al., 2016). In other words, it is the probability that an observed difference could have occurred just by random chance. The p-value for Pearson's correlation coefficient uses the t-distribution (Park, 2014).

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The p-value is obtained as 2 × P(T > t), where T follows a t-distribution with n – 2 degrees of freedom. If the p-value is less than the α level, the threshold value used to determine if a test statistic is statistically significant, we reject the null hypothesis and accept the alternative hypothesis. If the p-value is greater than α level, we fail to reject the null hypothesis.

The correlations within each of the subtypes of data were computed: laboratory values, comorbidities, and CyTOF data. Several significantly large correlations were found within the CyTOF data and a few within the laboratory values. Next, the interclass correlations were calculated, comparing the CyTOF data with the comorbidities first, the CyTOF data with the laboratory values second, and the comorbidities with the laboratory values last. Variables which were highly correlated were identified. Variables with clinical importance were retained while those with less importance or a larger proportion of missing values were removed from further analysis. From this step, the variables neutrophil count, IL-6, and c-reactive protein, a, d, f, and h were removed.

4.3 MICE imputation

As is common with clinical data, missing values were present for the lab values. Simply discarding samples with missing data could have potentially reduced power and biased model outputs (Kang 2013).

There are three types of missing data and the type of missing data the data falls under should be used to determine how the missing values are imputed. One of those missing types of data is missing completely at random (MCAR). Data are MCAR when the probability that values are missing is independent of the observed and unobserved data (Mack et al., 2018). In other words, the cause that the data is missing is *completely random*. An example would be unplanned exhaustion of laboratory instruments. Classifying missing data as MCAR is a strong assumption with a strong probability of being incorrect (Mack et al., 2018). Missing at random (MAR) is a weaker assumption that states the missingness can be explained by some of the observed data (Mack et al., 2018). For example, males might be less likely to participate in blood sampling to detect white blood cells, however, this is not related to their white blood cell count. Finally, missing data can be missing not at random (MNAR). MNAR data occur when the probability that a variable is missing is related to value of the variable (Nakagawa, 2015). For instance, participants with more severe depression are less likely to participate in a survey asking them to rate their depression.

We believed the missing data to be MAR or MCAR, thus, we required a technique that would be optimal for imputing the missing values. Multivariate Imputation via Chained Equations (MICE) is a multiple imputation method that assumes that the missing data are MAR. MICE takes a divide and conquer approach to imputing data in which Azur et al. (2011) summarize in six steps.

Step 1. The missing values in each variable are replaced with a temporary "place holder". This place holder is computed from a simple imputation such as imputing the mean.

Step 2. MICE removes the placeholder values from one variable.

Step 3. The observed values from the variable in step 2 are regressed on all or part of the variables of the imputation model. Linear regression is the default method for continuous missing variables; however, MICE can predict other types of variables as well. Logistic regression is used for binary categorical missing values, predictive mean matching (PMM) is used for numeric variables, Bayesian polytomous regression is used for categorical variables with at least 2 levels, and the proportional odds model is used for ordered data with at least 2 levels. Table A2 shows the imputation methods for the covariates in our models.

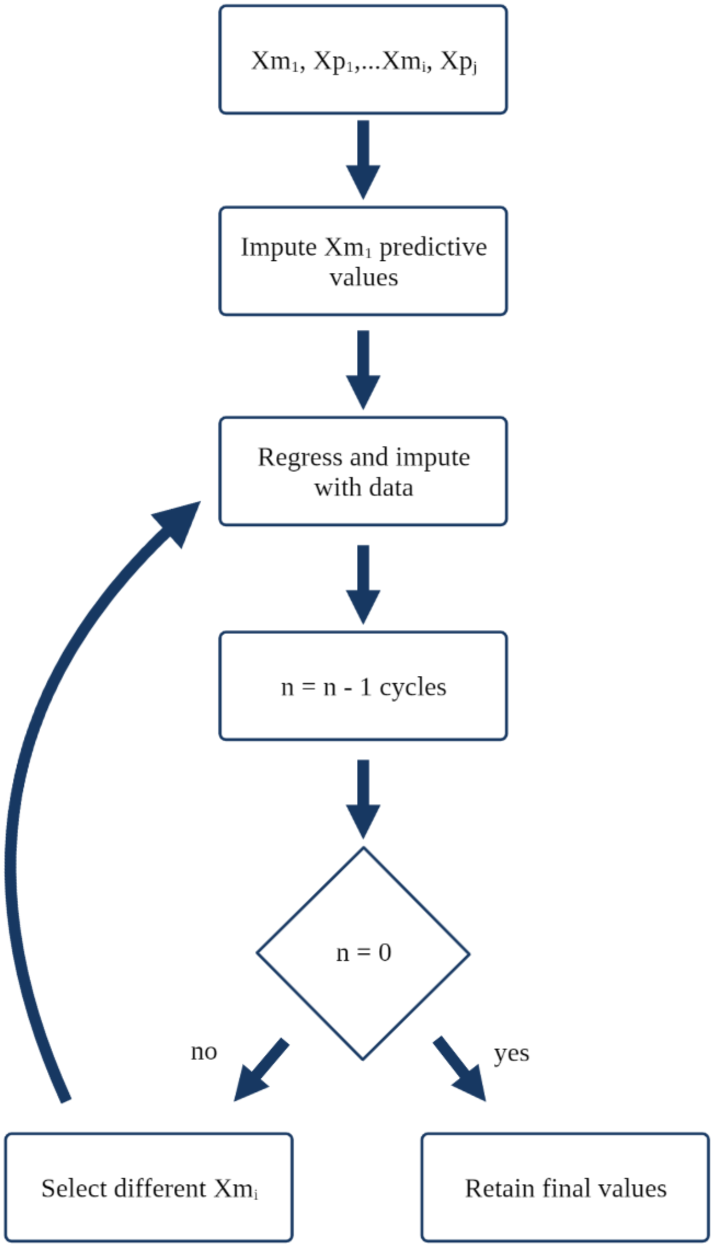
Step 4. The regression model created in Step 3 is used to predict or "impute" the missing values of the variable in Step 2. The newly completed variable will be used as an independent variable in subsequent regression models.

Step 5. MICE repeats Steps 2 – 4 for every variable with missing data. Each completion of Steps 1 – 5 is termed a "cycle". Upon the conclusion of a cycle, the missing values for the attended variable have been restored with predictions from regression models that indicate the associations observed in the data.

Step 6. Multiple cycles (we can specify the number, *n*) are run and at the end of each cycle, the imputed values are updated. The final values are retained from the final cycle.

After *n* cycles, one complete data set has been produced. The estimation of parameters and their variance can be obtained based on the complete data set. For our study, the final estimates of parameters and their variance will be based on repeated imputed inferences (Rubin, 1996), introduced below.

Figure 12. MICE algorithm


Xm_1, Xp_1,...Xm_i, Xp_j
Impute Xm_1 predictive values
Regress and impute with data
n = n - 1 cycles
n = 0
no
yes
Select different Xm_i
Retain final values

4.3 Repeated imputations inferences

If *m* complete data sets are produced, each completed data set can then be analyzed to yield *m* completed-data statistics, which are usually model estimates, say $\hat{Q}_l$ for l = 1, …, *m*, and *m* associated variance-covariance matrices, say $\widehat{U}_l$ for l = 1,…, *m*. Using the repeated-imputation inferences, we are able to pool the m models' estimates and determine the significance of the covariates. The repeated-imputation estimate is

$$\bar{Q}_m = \sum_{l=1}^{m} \hat{Q}_l / m$$

and the associated variance-covariance of $\bar{Q}_m$ is

$$T_m = \bar{U}_m + \frac{m+1}{m} B_m$$

Here, the within-imputation variability is given by

$$\bar{U}_m = \sum_{i=1}^{m} \widehat{U}_l / m \,,$$

and the between-imputation variability is given by

$$B_m = \sum_{i=1}^{m} \left(\hat{Q}_l - \bar{Q}_m\right)\left(\hat{Q}_l - \bar{Q}_m\right)' / (m-1)$$

The m repeated-imputation inference takes $(Q - \bar{Q}_m)$ to be a random variable with normal distribution and variance-covariance matrix T_m. Thus, letting m = ∞

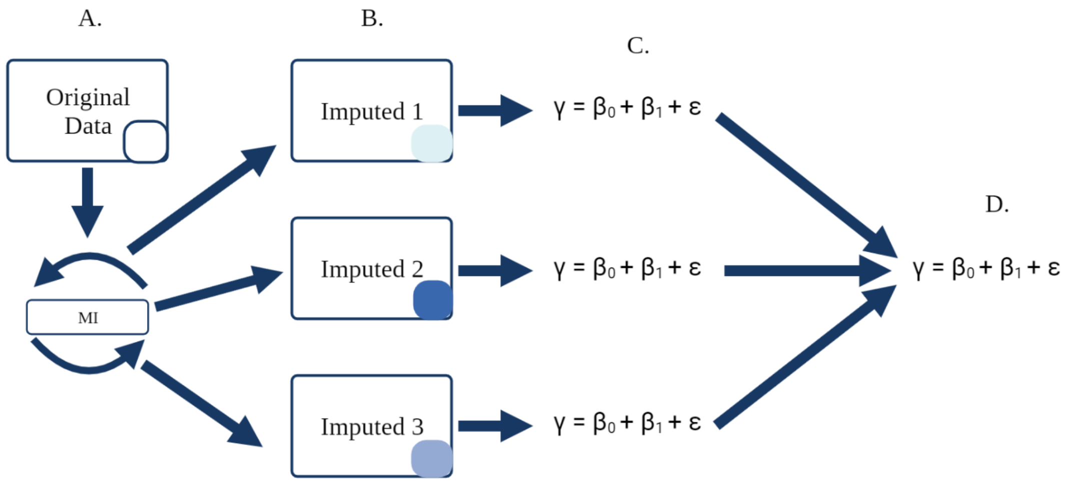
$$(Q - \bar{Q}_\infty) \sim N(O, T_\infty) \,,$$

where $T_\infty = \bar{U}_\infty + B_\infty$ (Rubin, 1996).

The Wald test statistic, *W*, can be calculated by

$$W = \frac{B_m}{\sqrt{T_m}}$$

Figure 13. Repeated imputations. *Key*: A. Original data set with missing data; B. Imputed data sets; C. Regression equations from each of the imputed data sets; D. Pooled regression model using repeated imputation inferences.


A.
B.
C.
D.
Original Data
MI
Imputed 1
Imputed 2
Imputed 3
$\gamma = \beta_0 + \beta_1 + \varepsilon$
$\gamma = \beta_0 + \beta_1 + \varepsilon$
$\gamma = \beta_0 + \beta_1 + \varepsilon$
$\gamma = \beta_0 + \beta_1 + \varepsilon$

4.4 Cumulative link mixed model

A cumulative link mixed model was created for ordinal regression with clinical outcome as the response variable.

Let $\pi_{ij} = \Pr\{Y_i = j \mid X = x_i\}$ the probability that the response of an individual *i* with characteristics x_i falls in the j^{th} category and let γ_{ij} denote the corresponding cumulative probability.

$$\gamma_{ij} = \Pr(Y_i \leq j \mid X = x_i)$$

Equivalently,

$$\gamma_{ij} = \pi_{i1} + \pi_{i2} + \ \dots \ + \pi_{ij}$$

Now, define $G^{-1}()$ as the logit function of a probability, p.

$$G^{-1} = logit(p) = log\ (\frac{p}{1-p})$$

Then, the general form of the cumulative link model is

$$G^{-1}(\gamma_{ij}) = \ \alpha_j - X\beta,$$

where α_j is a constant representing the threshold or intercept for level j and β is the vector of coefficients for each covariate.

The cumulative link regression model above assumes independence of the observations. When multiple measures are derived from the same individual or across time, as is the case with longitudinal analysis, this assumption is violated (Schmidt, 2012). In a cumulative link mixed model, a random effect is introduced to the cumulative link model to account for dependent observations (Schmidt, 2012). The general form of the cumulative link mixed model is

$$G^{-1}(\gamma_{ij} \mid X = x_i) = \ \alpha_j - (Z_i u_i + X_i\beta)$$

where Z_i and u_i are the random effects design matrix and random effects, respectively, for i^{th} subject, and $u_i \sim N(O, \sigma_u^2)$ (Schmidt, 2012). Together, $\beta, \alpha_j,$ and, u_i are the complete-data estimates. Here, we are primarily interested in the parameters β and u_i, which are the $\hat{Q}$ in multiple imputation. We use the MICE package in R to form *m* complete data sets, and for each complete data set we apply the cumulative link mixed models to obtain the parameter estimates ($\hat{Q}_l$) and their variance ($\widehat{U}_l$) for l = 1, …, *m*. Further, we apply Rubin’s repeated imputations inference to obtain the final estimate $\bar{Q}_m$

and its variance T_m for inference about the importance of the predictive variables on the clinical outcome.

CHAPTER V

RESULTS

Table 1 displays the clinical characteristics, comorbidities, CyTOF T cell population percentages, laboratory values at the time of observations as well as the distribution of the clinical outcome in severity in scale from 1 to 5. The median age was 64 years old with IQR as [50; 73], 57% female, with median BMI of 31 and IQR [24; 40] (Table 1). Comorbidities were present in most patients with diabetes (45.7%), hypertension (65.7%), hyperlipidemia (8.6%), coronary artery disease (5.7%), asthma (11.4%), and obstructive sleep apnea (11.4%) (Table 1). T cell population percentages were acquired from every observation. The median percentages and their IQR were as follows: CD4+CD45RA+ of T cells 21.3% [12.0%; 28.7%]; CD4+CD45RA+ of CD4+ T cells 35.1% [21.8%; 46.9%]; CD4+CD45RO+ of T cells 25.2% [19.6%; 33.8%]; CD4+CD45RO+ of CD4+ T cells 47.7% [35.6%; 63.1%]; CD8+CD45RA+ of T cells 14.2% [8.0%; 18.7%]; CD8+CD45RA+ of CD8+ T cells 64.9% [49.2%; 78.0%]; and CD8+CD45RO+ of CD8+ T cells 3.2% [2.1%; 7.1%]; CD8+CD45RO+ of T cells 17.8% [8.0%; 18.7%] (Table 1). The study’s patients showed a deviation from normal ranges within the available D-Dimer, Ferritin, C-reactive protein (CRP), Hgb, neutrophil percentage, lymphocyte percentage, lactate dehydrogenase, interleukin 6, and P/F ratio values. For these tests, the median was outside of the normal range (Table 1). Outcomes

Table 1. Demographic characteristics, comorbidities, T cell population percentages, clinical markers, outcome counts

Variable	**Value**	**Normal Range**
Number of observations	112	
Number of patients	35	
Patient Characteristics		
Age (y.o)	64 [50; 73]/(28|95)	
Gender (female) (%)	20 (57)	
BMI (kg/m^2)	31 [24; 40]/(18|54)	
Comorbidities (%)		
Diabetes (DM)	16 (45.7)	
Hypertension (HTN)	23 (65.7)	
Hyperlipidemia (HLD)	3 (8.6)	
Coronary Artery Disease (CAD)	2 (5.7)	
Asthma	4 (11.4)	
Obstructive Sleep Apnea (OSA)	4 (11.4)	
CyTOF T Cell Populations (%)		
CD4+CD45RA+ of T cells (a)	21.3 [12.0; 28.7]/(1.8|51.8)	
CD4+CD45RA+ of CD4+ T cells (b)	35.1 [21.8; 46.9]/(12.4|65.3)	
CD4+CD45RO+ of T cells (c)	25.2 [19.6; 33.8]/(5.6|49.9)	
CD4+CD45RO+ of CD4+ T cells (d)	47.7 [35.6; 63.1]/(18.9|75.7)	
CD8+CD45RA+ of T cells (e)	14.2 [8.0; 18.7]/(1.1|41.8)	
CD8+CD45RA+ of CD8+ T cells (f)	64.9 [49.2; 78.0]/(21.0|89.1)	
CD8+CD45RO+ of T cells (g)	3.2 [2.1; 7.1]/(.2|21.6)	
CD8+CD45RO+ of CD8+ T cells (h)	17.8 [8.0; 18.7]/(1.1|41.8)	
Laboratory Values		
D-Dimer (miss = 59) (µgFEU/mL)	2.3 [.87; 6.4]/(.21|32.0)	.19 - .74
Ferritin (miss = 60) (ng/mL)	385 [279.5; 729.2]/(17|6231.0)	7 - 350
CRP (miss = 61) (mg/L)	115.3 [51.1; 175.8]/(7.1|439.2)	0.0 - 10.0
WBC (miss = 4) (× 10^3/µL)	8.9 [5.9; 12.1]/(1.0|25.5)	4.1 - 10.8
Hgb (miss = 4) (gram/dL)	9.9 [8.3; 11.4]/(6.6|14.5)	13.7 - 17.5
Platelet (miss = 4) (× 10^3/µL)	228.5[158.0; 298.5]/(28.3|595.0)	140 - 370
Neutrophil % (miss = 35)	74 [64.8; 82.2]/(5.0|94.2)	40 - 60
Neutrophil Count (miss = 56) (× 10^3/µL)	5.3 [2.9; 7.2]/(1.5|18.1)	1.56 – 6.45
Lymphocyte % (miss = 36) (/µL)	13.3 [8.4; 21.6]/(3.8|36.6)	20 - 40
Lymphocyte Count (miss = 50) (× 10^3/µL)	.8 [.3; 1.2]/(0.0|2.6)	0.95 – 3.07
LDH (miss - 68) (units/L)	318.5 [225.0; 443.2]/(104.0|729.0)	100 - 242
IL-6 (miss = 107) (pg /mL)	62.3 [46; 77]/(24.6|126.2)	5 - 15
Creatinine (miss = 22) (mg/dL)	1.2 [.74; 1.9]/(.4|6.2)	.59 - 1.35
P/F Ratio (miss = 55)	180.0 [111.0; 226.7]/(52.0|310.0)	≥ 400
Outcomes		
Discharge (1)	5	
To floor/Extubation (2)	39	
To ICU (3)	14	
Intubation (4)	53	
Death (5)	1	

Notes: Data are expressed as medians [Interquartile IQ] or (min|max).

were evaluated the day of the blood samples for CyTOF data. For clinical outcomes, 5 observations were discharge, 39 observations were on the floor or extubated, 14 observations were in the ICU, and 1 observation occurred on the same day of death (Table 1).

Pearson correlations were computed using the pairwise complete observations. That is, for any two pair of variables, the observations with complete data for the variables of interest were used to calculate the correlation. Significant ($p <.05$) Pearson correlations with an absolute value of r greater than or equal to .7 were selected for during feature selection. The significant ($|r| > .7$, $p < .05$) results of the Pearson correlation computations are in Table 2.

We found significant correlations within two categories of the data. Within the laboratory values, an increase in white blood cell count (WBC) was highly associated with an increase in Neutrophil count ($r = .98$, $p < .05$), IL-6 had a strong positive correlation with LDH ($r = .99$, $p < .05$), and CRP was inversely correlated with P/F ratio ($r = -.78$, $p < .05$) (Table 2). The clinical marker with the highest percentage of complete data was selected for inclusion within the final model. Within the CyTOF data, increases in the T cell percentages of CD4+CD45RA+ T cells (a) were associated with increases in CD4+ T cell percentages of CD4+CD45RA+ T cells (b) ($r = 0.89$, $p < .05$) (Table 2). Increases in the CD8+ T cell percentages of CD8+CD45RO+ T cells (h) were associated with increases in the T cell percentages of CD8+CD45RO+ T cells (g) ($r = .74$, $p < .05$) (Table 2). The following were negatively correlated: T cell percentages of CD4+CD45RA+ T cells (a) with CD4+ T cell percentages of CD4+CD45RO+ T cells (d) ($r = -.86$, $p < .05$), CD4+ T cell percentages of CD4+CD45RA+ T cells (b) with T cell percentages of CD4+CD45RO+ T cells (c) (r = -.97, p < .05), CD8+ T cell percentages of CD8+CD45RA+ T cells (f) with T cell percentages of CD8+CD45RO+ T cells (g) (r = -.71, p < .05), and CD8+ T cell percentages of CD8+CD45RA+ T cells (f) with CD8+ T cell percentages of CD8+CD45RO+ T cells (h) (r = -.97, p < .05) (Table 2). Essentially, there were two groups of high correlation with 3 variables within each. The variable with the highest average absolute value of r within each group was selected for inclusion within the final model. These were variables b and g.

Table 2. Significantly correlated variables

Variable 1	Variable 2	Correlation	P-value
WBC	Neutrophil Count	0.98	0.00
IL-6	LDH	0.99	0.04
CRP	P/F Ratio	-0.71	0.00
a	b	0.89	0.00
a	d	-0.86	0.00
b	d	-0.97	0.00
f	g	-0.71	0.00
f	h	-0.97	0.00
g	h	0.74	0.00

The estimates, standard errors, Wald test statistics, and p-values from repeated imputation inferences utilizing the cumulative linked mixed models are depicted in Table 3. BMI, Hgb, and the CD4+CD45RA+ T cells percentage of CD4+ T cells were found to be significant predictors (p < .05) of clinical outcome with estimates of .27, -1.39, and .06, respectively. At the significance level of .10, this group also includes WBC, HLD, and the CD4+ T cells percentage of CD4+CD45RA+ T cells with estimates of .42, 9.05, and .06, respectively. From the model

$$G^{-1}(\gamma_{ij}) = \alpha_j - (Z_i u_i + X_i \beta),$$

positive coefficients lead to an increase in the probability of higher-level categories. Thus, increased levels of covariates with a positive coefficient were associated with worse clinical outcomes. From the covariates with significant p-values, an increase in BMI and the percentage of CD4+ T cells composed of CD4+CD45RA+ T cells signified a poorer clinical outcome. Conversely, an increase of covariates with negative coefficients lead to a decrease in the probability of higher-level categories. From the covariates with significant p-values, an increase in Hgb levels was associated with better clinical outcomes.

The thresholds (e.g., 1|2, 2|3, 3|4, 4|5) are intercepts for each of the ordinal levels in the cumulative link mixed model. Under the proportional odds assumption, these are assumed to be constant for all values of the remaining independent variables.

Random effect generates the correlation expected between observations from the same patient and allows inferences to be made to the population from which the groups were sampled. Negligence to consider the correlations of within patient observations could have resulted in biased estimates and invalid statistical inferences. Thus, we took the patient effects to be random and assumed that the patient effects were independent and identically distributed normal: $u(patient_i) \sim N(O, \sigma_u^2)$.

One method of determining if a variable should be treated as a random effect is the intraclass correlation (ICC) (Theobald, 2018). The ICC is the measure of the clustering in a variable and is given by

$$ICC = \frac{\sigma_u^2}{\sigma^2 + \sigma_u^2}$$

Here, σ^2 represents the residual variance, which is assumed to be one in cumulative link mixed models. σ_u^2 represents the variance of the random effect (Schmidt, 2012), 7.655 from the model. When ICC = 0, there is no clustering, i.e., observations within a given patient are just about independent. When ICC = 1, there is complete clustering (Theobold, 2018), i.e., observations within a given patient from the study were highly similar. The ICC was .88, thus, patient should be treated as a random effect as a patients' clinical outcomes from COVID-19 were highly correlated.

Table 3. Pooled model estimates which include patient as random effect

	Estimate	Std. Error	z value	Pr(>|z|)
1|2	-16.06	5.58	-2.88	0.00
2|3	-3.25	4.96	-0.66	0.51
3|4	-0.89	4.82	-0.18	0.85
4|5	15.40	6.75	2.28	0.02
BMI	0.27	0.13	2.08	0.04
WBC	0.42	0.23	1.83	0.06
Hgb	-1.39	0.596	-2.33	0.02
Platelet	0.00	0.01	0.00	0.74
Creatinine	0.63	1.24	0.51	0.61
HTN	0.67	2.16	0.31	0.76
DM	-3.34	2.36	-1.42	0.16
HLD	9.05	5.22	1.73	0.08
CAD	0.6	6.35	0.09	0.99
Asthma	5.02	4.18	1.20	0.23
OSA	0.62	4.40	0.14	0.89
b	0.06	0.03	2.00	0.05
c	-0.02	0.04	-0.50	0.71
e	-0.16	0.10	-1.60	0.11
g	-0.42	0.11	-3.82	0.00
random effect	7.655	2.96	2.59	0.01

Notes: Table 3 is produced from 10 sets of estimates reported in the appendix in Table A3. Each set of estimates was obtained from an imputed data set.

CHAPTER VI

DISCUSSION

6.1 Observations

The study’s patients showed a deviation from normal ranges within the available D-Dimer, Ferritin, C-reactive protein (CRP), Hgb, neutrophil percentage, lymphocyte percentage and count, lactate dehydrogenase, interleukin 6, and P/F ratio values. For these tests, the median was outside of the normal range (Table 1).

D-dimer is a protein fragment produced when a blood clot is dissolved in the body and, as a clinical marker, is used to monitor coagulation state. High levels of D-dimer in the blood indicate a major clot like deep vein thrombosis (DVT). Our increased levels of D-dimer compared to normal D-dimer ranges suggests a higher risk of thrombosis in COVID-19 patients. Similarly, Yao et al. (2020), concluded that D-dimer levels are commonly elevated in COVID-19.

Ferritin is a blood protein that stores iron that is used to monitor systemic inflammation. One explanation for the high levels for elevated Ferritin in COVID-19 patients is that the natural immune response might limit iron turnover during infections to prevent pathogens from using it (Bozkurt et al., 2021). Elevated ferritin levels have been shown to be a characteristic of severe COVID-19 patients in several studies (Banchini et al., 2021; Vargas-Vargas and Cortés-Rojo, 2020).

An Hgb test measure how much hemoglobin red blood cells contain. Hgb helps red blood cells transport oxygen from the lungs to the body. Opposite of many of the clinical markers, Hgb was lower than normal ranges for study's patients. This concurs with a study by Dinevari et al. (2021) study which showed a high prevalence of anemia in hospitalized patients with COVID-19.

Neutrophils and Lymphocytes are both types of white blood cell types. Lymphopenia (Jafarzadeh et al., 2021) and high neutrophil (Huang et al., 2020) counts have been characteristic of COVID-19 patients. In addition, a higher level of neutrophile-to-lymphocyte (i.e. a higher neutrophil percentage) has been linked to severe COVID-19 (Kong et al., 2020).

Increased levels of LDH in the blood, as seen in this study, signify early-stage myocardial infarction and hemolysis (Szarpak et. al, 2020). Szarpak et al. (2020) and Henry et al. (2020) found LDH levels to be markers of COVID-19 severity and predictors of survival.

Interleukin 6, or IL-6, is one of the overproduced inflammatory proteins, called cytokines, that are associated with a condition known as cytokine release syndrome or cytokine storm (Jose and Manuel, 2020). Cytokine storms, if not de-escalated, lead to increased chance of vascular hyperpermeability, multi-organ failure, and death (Jose and Manuel, 2020). As in our study, IL-6 levels have been shown to be elevated in the peripheral blood of COVID-19 patients (Chen et al., 2020).

P/F ratio is an oxygenation index used to identify hypoxia and respiratory distress. However, its use for this purpose is controversial (Tobin et al. 2021, Gu et al., 2021) so we will avoid further discussion on this subject.

For the correlations, neutrophils are a type of white blood cell, so it was reasonable that a strong, significant correlation was detected. Our study results align with Ede et al. (2013) in that there was a positive correlation between LDH and IL-6. IL-6 is a cytokine that mediates inflammation whereas LDH is the product of inflammatory injury. This correlation makes sense. C-reactive protein and P/F ratio were inversely correlated; however, once again, we will avoid any discussion on P/F ratio due to its controversial usage. Finally, there were several correlations between the CyTOF T cell population percentages. This was expected as an increase one population's percentage would guarantee a decrease in another's. We believed it important to show the percentages of complementary populations to gain a holistic view of the T cells and then select the percentages to be used for the model. Ultimately, the CD4+ T cell percentages of CD4+CD45RA+ T cells and the CD8+ T cell percentages of CD8+CD45RO+ T cells were used.

The cumulative link mixed model determined BMI, Hgb, and CD8+ T cell percentages of CD8+CD45RO+ T cells to be significant predictors of COVID-19 severity. The BMI and Hgb factors significance is in agreement with several other studies. Hendren et al. (2020) and Dinevari et al. (2021) are just two examples. Our model supports a hypothesis that an increase in the percentage of memory CD8+ T cells over T cells overall was associated with improved clinical outcomes.

6.2 Limitations

There were only 112 observations and 35 patients. In addition to this being low sample size of the dataset, with an average of 3.2 observation per patient, there were few observations from each patient. Cumulative link model estimates can be unstable if there are a small number of observations within clusters or if there are few clusters from which to estimate within group correlation (Schmidt, 2012). Furthermore, only 1 observation had the outcome of the death. In general, a small sample size reduces the confidence level of any study and reduces the statistical power, the ability to detect an effect when there is one to be detected. It could very well be that other covariates in our model were significant predictors, but due to small sample size and inappreciable effect, the model was not able to detect these associations.

Another limitation of this study was the missing data within the laboratory values. This caused us to exclude several potentially significant covariates from our model. Although we were able to perform multiple imputation, imputed data holds less value than observed data. The values imputed were regressed only with the covariates included in the model. Though we selected covariates commonly measured in COVID-19 studies, an incorrectly chosen conditioning set can make the imputations endogenous and lead to bias (Mittag, 2013). There were other variables that would have been better to impute values for. This is why in addition to utilization of the repeated imputation inferences, to counter any potential errors in imputation, we selected laboratory values with less than 20% of the data missing.

There were 262 potential comorbidities that could have been included in the data. However, over half of these were excluded due to occurrence in only 2 or fewer patients. These were potential explanators of laboratory values in multiple imputation or COVID severity in the cumulative link mixed models as well. However, to simplify the model and due a likely outcome of insignificance because of low sample size, we chose to exclude these values.

When individuals in a longitudinal study have the same number of repeated measurements collected in similar time intervals, the study is considered to be "balanced". Time between a draw and the previous one varied in the data.

Although cross-checked with manual and computational methods, it is unlikely that we obtained the exact percentages of the T cell population for every observation. Errors are almost bound to occur when manual gating. Divisions and selections of the populations is subjective to the scientist doing the gating. Nevertheless, we believe these values to be close to the true values after thorough review from multiple individuals and experts.

6.3 Future studies

Our longitudinal study found BMI, Hgb, and CD8+ T cell percentages of CD8+CD45RO+ T cells to be significant predictors of COVID-19 severity. However, the physiological pathways and cellular etiology of COVID-19 involves more than just T cells and hemoglobin. Potential future studies of our work could look at the interaction of of CD8+CD45RO+ T cells with other cell populations and how this affects COVID-19 severity. Furthermore, we propose more extensive longitudinal testing to characterize the effect of COVID-19 infection after discharge from the hospital. This would provide a more comprehensive perspective of infection over time.

REFERENCES

Adil, M. T., Rahman, R., Whitelaw, D., Jain, V., Al-Taan, O., Rashid, F., Munasinghe, A., & Jambulingam, P. (2021). “SARS-CoV-2 and the pandemic of COVID-19.” Postgraduate Medical Journal **97**(1144): 110–116. https://doi.org/10.1136/postgradmedj-2020-138386INtro

Akoglu H. (2018). “User's guide to correlation coefficients.” Turkish Journal of Emergency Medicine **18**(3): 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Azur, Melissa & Stuart, Elizabeth & Frangakis, Constantine & Leaf, Philip. (2011). “Multiple Imputation by Chained Equations: What is it and how does it work?” International Journal of Methods in Psychiatric Research **20**(1): 40-9.

Banchini, F., Cattaneo, G. M., & Capelli, P. (2021). “Serum ferritin levels in inflammation: a retrospective comparative analysis between COVID-19 and emergency surgical non-COVID-19 patients.” World Journal of Emergency Surgery **16**(1): 9. https://doi.org/10.1186/s13017-021-00354-3

Bozkurt, F. T., Tercan, M., Patmano, G., Bingol Tanrıverdi, T., Demir, H. A., & Yurekli, U. F. (2021). “Can Ferritin Levels Predict the Severity of Illness in Patients With COVID-19?” Cureus **13**(1), e12832. https://doi.org/10.7759/cureus.12832

Caruana, E. J., Roman, M., Hernández-Sánchez, J., & Solli, P. (2015). “Longitudinal studies.” Journal of Thoracic Disease **7**(11): 537–540. https://doi.org/10.3978/j.issn.2072-1439.2015.10.63

Chen, G., Wu, D., Guo, W., Cao, Y., Huang, D., Wang, H., Wang, T., Zhang, X., Chen, H., Yu, H., Zhang, X., Zhang, M., Wu, S., Song, J., Chen, T., Han, M., Li, S., Luo, X., Zhao, J., & Ning, Q. (2020). “Clinical and immunological features of severe and moderate coronavirus disease 2019.” The Journal of Clinical Investigation **130**(5): 2620–2629. https://doi.org/10.1172/JCI137244

Christensen RHB (2019). “ordinal—Regression Models for Ordinal Data.” R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Coggon, D., Barker, D., Rose, G. (2009). Epidemiology for the Uninitiated. Hoboken, NJ: BMJ Books.

Diao, B., Wang, C., Tan, Y., Chen, X., Liu, Y., Ning, L., Chen, L., Li, M., Liu, Y., Wang, G., Yuan, Z., Feng, Z., Zhang, Y., Wu, Y., & Chen, Y. (2020). “Reduction and Functional Exhaustion of T Cells in Patients With Coronavirus Disease 2019 (COVID-19).” Frontiers in Immunology **11**(827). https://doi.org/10.3389/fimmu.2020.00827

Dinevari, Masood & Somi, Mohammad & Majd, Elham & Abbasalizad Farhangi, Mahdieh & Nikniaz, Zeinab. (2021). “Anemia predicts poor outcomes of COVID-19 in hospitalized patients: a prospective study in Iran.” BMC Infectious Diseases **21**(170)

Ede, L. C., O'Brien, J., Chonmaitree, T., Han, Y., Patel, J. A. (2013). “Lactate dehydrogenase as a marker of nasopharyngeal inflammatory injury during viral upper respiratory infection: implications for acute otitis media.” Pediatric Research **73**(3): 349–354. https://doi.org/10.1038/pr.2012.179

Ellis B., Haaland P., Hahne F., Le Meur N., Gopalakrishnan N., Spidlen J., Jiang M., Finak G. (2021). “flowCore: Basic structures for flow cytometry data.” R package version 2.4.0.

Greenacre, Michael & Nenadic, Oleg. (2005). “Computation of Multiple Correspondence Analysis, with Code in R.” Department of Economics and Business, Universitat Pompeu Fabra, Economics Working Papers.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). “Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations.” European Journal of Epidemiology **31**(4): 337–350. https://doi.org/10.1007/s10654-016-0149-3

Gu, Y., Wang, D., Chen, C., Lu, W., Liu, H., Lv, T., Song, Y., & Zhang, F. (2021). “PaO2/FiO2 and IL-6 are risk factors of mortality for intensive care COVID-19 patients.” Scientific Reports **11**(1): 7334. https://doi.org/10.1038/s41598-021-86676-3

Hendren, Nicholas & Lemos, James & Ayers, Colby & Das, Sandeep & Rao, Anjali & Carter, Spencer & Rosenblatt, Anna & Walchok, Jason & Omar, Wally & Khera, Rohan & Hegde, Anita & Drazner, Mark & Neeland, Ian & Grodin, Justin. (2020). “Association of Body Mass Index and Age With Morbidity and Mortality in Patients Hospitalized With COVID-19: Results From the American Heart Association COVID-19” Cardiovascular Disease Registry.” Circulation **143**(2): 135 – 144.

Henry, B. M., Aggarwal, G., Wong, J., Benoit, S., Vikse, J., Plebani, M., & Lippi, G. (2020). “Lactate dehydrogenase levels predict coronavirus disease 2019 (COVID-19) severity and mortality: A pooled analysis.” The American Journal of Emergency Medicine **38**(9): 1722–1726. https://doi.org/10.1016/j.ajem.2020.05.073

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., … Cao, B. (2020). "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China." Lancet (London, England) **395**(10223): 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Jafarzadeh, A., Jafarzadeh, S., Nozari, P., Mokhtari, P., & Nemati, M. (2021). "Lymphopenia an important immunological abnormality in patients with COVID-19: Possible mechanisms." Scandinavian Journal of Immunology **93**(2): 12967. https://doi.org/10.1111/sji.12967

Jose, R. J., & Manuel, A. (2020). "COVID-19 cytokine storm: the interplay between inflammation and coagulation." The Lancet. Respiratory Medicine **8**(6): 46–47. https://doi.org/10.1016/S2213-2600(20)30216-2

Kang H. (2013). "The prevention and handling of the missing data." Korean Journal of Anesthesiology **64**(5): 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Kantardzic, M. (2020). Data mining: Concepts, models, methods, and algorithms. Piscataway, NJ: IEEE Press

Kermali, M., Khalsa, R. K., Pillai, K., Ismail, Z., & Harky, A. (2020). "The role of biomarkers in diagnosis of COVID-19 - A systematic review." Life sciences **254**: 117788. https://doi.org/10.1016/j.lfs.2020.117788

Kim J. H. (2019). "Multicollinearity and misleading statistical results." Korean Journal of Anesthesiology **72**(6): 558–569. https://doi.org/10.4097/kja.19087

Kirch, W. (ed). (2008). "Pearson's Correlation Coefficient." In: Encyclopedia of Public Health (Dordrecht: Springer Netherlands), 1090–1. Available online at: http://link.springer.com/10.1007/978-1-4020-5614-7_2569

Kobak, D., & Berens, P. (2019). "The art of using t-SNE for single-cell transcriptomics." Nature Communications **10**(1): 5416. https://doi.org/10.1038/s41467-019-13056-x

Kong, M., Zhang, H., Cao, X., Mao, X., & Lu, Z. (2020). "Higher level of neutrophil-to-lymphocyte is associated with severe COVID-19." Epidemiology and Infection **148**: 139. https://doi.org/10.1017/S0950268820001557

Li, H., Shaham, U., Stanton, K. P., Yao, Y., Montgomery, R. R., & Kluger, Y. (2017). "Gating mass cytometry data by deep learning." Bioinformatics (Oxford, England) **33**(21): 3423–3430. https://doi.org/10.1093/bioinformatics/btx448

Liu, X. (2016). In Methods and applications of longitudinal data analysis (pp. 1–18). essay, Academic Press.

Mack, C., Su, Z., & Westreich, D. (2018). “Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User’s Guide, Third Edition.” Agency for Healthcare Research and Quality (US).

Mahdiyasa, Adilan & Pasaribu, U. (2019). “Multiple Correspondence Analysis Using Burt Matrix: A Study of Bandung Institute of Technology student Characteristics.” IOP Conference Series: Materials Science and Engineering **598**. 012012. 10.1088/1757-899X/598/1/012012.

Mittag, N. (2013). Imputations: Benefits, Risks and a Method for Missing Data.

Morrissey, Samantha & Geller, Anne & Hu, Xiaoling & Tieri, David & Cooke, Elizabeth & Ding, Chuanlin & Woeste, Matthew & Zhange, Huang-ge & Cavallazi, Rodrigo & Clifford, Sean & Chen, James & Kong, Maiying & Watson, Corey & Huang, Jiapeng & Yan, Jun. (2020). “Emergence of Low-density Inflammatory Neutrophils Correlates with Hypercoagulable State and Disease Severity in COVID-19 Patients.” 10.1101/2020.05.22.20106724.

Mukaka M. M. (2012). “Statistics corner: A guide to appropriate use of correlation coefficient in medical research.” Malawi Medical Journal: The journal of Medical Association of Malawi **24**(3): 69–71.

Nakagawa, S. (2015). Chapter 4 Missing data: mechanisms, methods, and messages.

Nowicka, M., Krieg, C., Crowell, H. L., Weber, L. M., Hartmann, F. J., Guglietta, S., Becher, B., Levesque, M. P., & Robinson, M. D. (2017). “CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets.” F1000Research **6**: 748. https://doi.org/10.12688/f1000research.11622.3

Park Y. G. (2014). Comments on statistical issues in September 2014. Korean Journal of Family Medicine, **35**(5): 257–258. https://doi.org/10.4082/kjfm.2014.35.5.257

Passaro, Antonio & Bestvina, Christine & Velez, Maria & Garassino, Marina & Garon, Edward & Peters, Solange. (2021). “Severity of COVID-19 in patients with lung cancer: Evidence and challenges.” Journal for ImmunoTherapy of Cancer **9**: 002266. 10.1136/jitc-2020-002266.

Rubin, D. (1996). “Multiple Imputation After 18 Years.” Journal of the American Statistical Association **91**(434): 473-489. doi:10.2307/2291635

Salkind, N. J. (2010). Encyclopedia of research design (Vols. 1-0). Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412961288

Sanyaolu, A., Okorie, C., Marinkovic, A., Patidar, R., Younis, K., Desai, P., Hosein, Z., Padda, I., Mangat, J., & Altaf, M. (2020). “Comorbidity and its Impact on Patients with COVID-19.” SN Comprehensive Clinical Medicine, 1–8. Advance online publication. https://doi.org/10.1007/s42399-020-00363-4

Schmidt, J. (2012). Ordinal Response Mixed Models: A Case Study.

Shah, V. K., Firmal, P., Alam, A., Ganguly, D., & Chattopadhyay, S. (2020). “Overview of Immune Response During SARS-CoV-2 Infection: Lessons From the Past.” Frontiers in Immunology **11**: 1949. https://doi.org/10.3389/fimmu.2020.01949

Spitzer, M. H., & Nolan, G. P. (2016). “Mass Cytometry: Single Cells, Many Features.” Cell **165**(4), 780–791. https://doi.org/10.1016/j.cell.2016.04.019

Ssentongo, P., Heilbrunn, E. S., Ssentongo, A. E., Advani, S., Chinchilli, V. M., Nunez, J. J., & Du, P. (2021). “Epidemiology and outcomes of COVID-19 in HIV-infected individuals: a systematic review and meta-analysis.” Scientific Reports **11**(1): 6283. https://doi.org/10.1038/s41598-021-85359-3

Subbarao, K., & Mahanty, S. (2020). “Respiratory Virus Infections: Understanding COVID-19.” Immunity **52**(6): 905–909. https://doi.org/10.1016/j.immuni.2020.05.004

Szarpak, L., Ruetzler, K., Safiejko, K., Hampel, M., Pruc, M., Kanczuga-Koda, L., Filipiak, K. J., & Jaguszewski, M. J. (2020). “Lactate dehydrogenase level as a COVID-19 severity marker.” The American Journal of Emergency Medicine, S0735-6757(20)31034-2. Advance online publication. https://doi.org/10.1016/j.ajem.2020.11.025

Theobald E. (2018). “Students Are Rarely Independent: When, Why, and How to Use Random Effects in Discipline-Based Education Research.” CBE Life Sciences Education **17**(3): rm2. https://doi.org/10.1187/cbe.17-12-0280

Tobin, M. J., Jubran, A., & Laghi, F. (2021). “PaO2 /FIO2 ratio: the mismeasure of oxygenation in COVID-19.” The European Respiratory Journal **57**(3): 2100274. https://doi.org/10.1183/13993003.00274-2021

van Buuren S, Groothuis-Oudshoorn K (2011). “mice: Multivariate Imputation by Chained Equations in R.” Journal of Statistical Software **45**(3): 1-67. https://www.jstatsoft.org/v45/i03/.

van der Maaten, Laurens & Hinton, Geoffrey. (2008). “Visualizing data using t-SNE.” Journal of Machine Learning Research. **9**: 2579-2605.

Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., & Saeys, Y. (2015). “FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data.” Cytometry. Part A: The Journal of the International Society for Analytical Cytology **87**(7): 636–645. https://doi.org/10.1002/cyto.a.22625

Vargas-Vargas, M., & Cortés-Rojo, C. (2020). “Ferritin levels and COVID-19.” Revista Panamericana de Salud Publica = Pan American Journal of Public Health, **44**: 72. https://doi.org/10.26633/RPSP.2020.72

Wilkerson, M. D., & Hayes, D. N. (2010). "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking." Bioinformatics (Oxford, England) **26**(12): 1572–1573. https://doi.org/10.1093/bioinformatics/btq170

Yao, Y., Cao, J., Wang, Q., Shi, Q., Liu, K., Luo, Z., Chen, X., Chen, S., Yu, K., Huang, Z., & Hu, B. (2020). "D-dimer as a biomarker for disease severity and mortality in COVID-19 patients: a case control study." Journal of Intensive Care **8**.

Zheng, H. Y., Zhang, M., Yang, C. X., Zhang, N., Wang, X. C., Yang, X. P., Dong, X. Q., & Zheng, Y. T. (2020). "Elevated exhaustion levels and reduced functional diversity of T cells in peripheral blood may predict severe progression in COVID-19 patients." Cellular & molecular immunology **17**(5): 541–543. https://doi.org/10.1038/s41423-020-0401-3

APPENDIX

7.1 Supplementary Tables

Table A1. Mass cytometry antibody panel

Antigen	Symbol & Mass	Antibody Clone	Source
CD16	Bi209	3G8	Fluidigm
CD8	Cd106	RPA-T8	Biolegend
CD14	Cd110	M5E2	Biolegend
CD4	Cd111	RPA-T4	Biolegend
CD11b	Cd112	IRCF44	Biolegend
CD3	Cd113	UCHT1	Biolegend
CD20	Cd114	2H7	Biolegend
CD19	Cd116	HIB19	Biolegend
CD80	Dy161	2D10.4	Fluidigm
CD79b	Dy162	CB3-1	Fluidigm
CXCR3	Dy163	G025H7	Fluidigm
CXCR5	Dy164	RF8B2	Fluidigm
CD44	Er166	BJ18	Fluidigm
CD27	Er167	L128	Fluidigm
CD40L	Er168	24-31	Fluidigm
CTLA-4	Er170	14D3	Fluidigm
LAMP1	Eu151	H4A3	Fluidigm
gdTCR	Eu153	B1	Biolegend
CD56	Gd155	HCD56	Fluidigm
CD86	Gd156	IT2.2	Fluidigm
TLR4	Gd158	HTA125	Fluidigm
CD28	Gd160	CD28.2	Fluidigm
CD45RA	Ho165	HI100	Biolegend
CD274	Lu175	29E.2A3	Fluidigm
CD40	Nd142	5C3	Fluidigm
CD123	Nd143	6H6	Fluidigm
CD69	Nd144	FN50	Fluidigm
CD163	Nd145	GHI/61	Fluidigm
IgD	Nd146	IA6-2	Fluidigm
CD66b	Nd148	G10F5	Biolegend
LAG-3	Nd150	11C3C65	Fluidigm
CD196	Pr141	G034E4	Fluidigm
CD11c	Sm147	Bu15	Fluidigm
CD45RO	Sm149	UCHL1	Fluidigm
CD21	Sm152	BL13	Fluidigm
TIM-3	Sm154	F38-2E2	Fluidigm
CD197	Tb159	G043H7	Fluidigm
CD25	Tm169	2A3	Fluidigm
CD45	Y89	HI30	Fluidigm
CD68	Yb171	Y1/82A	Fluidigm
CD38	Yb172	HIT2	Fluidigm
HLA-Dr	Yb173	L243	Fluidigm
CD279	Yb174	EH12.2H7	Fluidigm
CD127	Yb176	A019D5	Fluidigm

Table A2. Method of imputation

Method

BMI	WBC	Hgb	Platelet	Creatinine	HTN	DM
""	"pmm"	"pmm"	"pmm"	"pmm"	""	""
HLD	CAD	Asthma	OSA	b	c	e
""	""	""	""	""	""	""
g						
""						

Table A3. Coefficients of Cumulative Link Mixed Models

	Model									
	1	2	3	4	5	6	7	8	9	10
1|2	-15.71053	-18.37015	-19.32589	-17.29712	-12.51050	-15.49182	-16.36435	-15.64600	-17.43491	-19.22875
2|3	-2.99596	-4.61120	-6.59951	-3.16926	-0.28828	-2.79137	-4.07530	-3.19572	-4.46790	-4.14137
3|4	-0.67670	-2.12144	-4.18905	-0.71481	1.94791	-0.43564	-1.65859	-0.87314	-2.06066	-1.57264
4|5	15.97543	15.45462	11.86630	16.89943	17.21479	16.01301	14.77046	15.44650	15.72848	17.90211
BMI	0.28325	0.30277	0.25625	0.31279	0.25518	0.26417	0.23802	0.26055	0.26173	0.38156
WBC	0.40833	0.52933	0.45258	0.47080	0.40157	0.46022	0.49360	0.42258	0.57920	0.63911
Hgb	-1.33331	-1.48136	-1.64508	-1.54991	-1.23218	-1.50129	-1.65449	-1.57496	-1.67238	-1.74574
Platelet	-0.00335	-0.00666	-0.00268	-0.00131	0.00032	-0.00118	-0.00244	-0.00050	-0.00399	-0.00485
Creatinine	0.15261	0.41165	0.66741	0.80092	0.47603	0.68993	0.86358	0.67598	0.29652	-0.33768
HTN	0.63733	0.15853	0.97239	0.31534	0.91250	0.93713	1.21346	1.20583	1.27797	0.10076
DM	-2.95310	-3.29598	-3.87782	-3.99140	-3.23291	-3.84055	-3.96673	-3.86795	-3.13702	-3.26825
HLD	8.90489	9.90209	8.45773	10.58712	9.41208	9.01481	7.50451	8.30791	7.32435	10.16687
CAD	-0.12008	-0.76278	0.24382	0.33302	0.91756	0.56043	0.08025	0.54829	0.15501	-0.69353
Asthma	4.76101	5.72661	5.36584	6.46767	5.00653	5.41895	5.08342	5.14513	4.53131	5.79405
OSA	0.45178	0.48066	0.99883	1.01678	-0.28447	0.80912	1.79139	1.16601	1.60342	-0.09464
b	0.07093	0.03310	0.03998	0.05901	0.08771	0.07427	0.07085	0.07733	0.04841	0.04390
c	-0.01629	-0.00550	-0.01071	-0.02100	-0.01207	-0.01170	0.00751	-0.00213	0.00578	-0.00212
e	-0.16683	-0.16362	-0.14494	-0.18290	-0.16480	-0.13570	-0.09584	-0.12102	-0.09770	-0.16108
g	-0.41529	-0.46038	-0.43821	-0.42611	-0.38693	-0.40085	-0.38272	-0.41366	-0.43272	-0.45423
random effect	7.71630	8.88920	7.00886	8.95262	7.20870	7.37108	6.52085	6.92264	7.57590	9.86865

7. 2 Acronyms

BMI – body mass index

BMU – best matching unit

CAD – coronary artery disease

COPD – chronic obstructive pulmonary disease

COVID-19 – coronavirus disease of 2019

CRP – C-reactive protein

DM – diabetes mellitus

ECG - electrocardiogram

FCS – flow cytometry standard

Hgb - hemoglobin

ICC – intraclass correlation

ICP – inductively coupled plasma (membrane)

LDH – lactate dehydrogenase

MAR – missing at random

MCA – multiple correspondence analysis

MCAR – missing completely at random

MICE – multivariate imputation by chained equations

MNAR – missing not at random

OSA – obstructive sleep apnea

PCA – principal component analysis

PMM – predictive mean matching

SOM – self-organizing map

TOF – time of flight

CURRICULUM VITA

NAME:	Onajia Josiah Stubblefield
ADDRESS:	University of Louisville School of Public Health and Information Sciences 485 E. Gray St Louisville, KY 40202
DOB:	Dyersburg, TN – April 6, 1997
EDUCATION:	B.S., Mathematics The University of Louisville 2015 – 2019 Honors: *Cum laude*, Honors program
AWARDS:	Woodford R. Porter Scholar 2015
	Gates-Millennium Scholar 2015
	National Achievement Scholar 2015
	Kentucky Governor's Scholar 2014