Using electronic health records to understand COVID-19 risks

Vijendra Ramlall

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2023

Abstract

Using electronic health records to understand COVID-19 risks

Vijendra Ramlall

On December 31, 2019, a new disease, which would in due time would come to be identified as COVID-19, was reported to the World Health Organization. During the two and a half years since the emergence of COVID-19 and the more than two years since the start of the COVID-19 pandemic, which is caused by infection of SARS-CoV-2, more than 500 million cases have been reported around the world with more than six million deaths attributed it with than 85 million cases and more than one million deaths from the United States of America. This novel disease has had profound economic, political, public health and social impact in the United States and around the world. Subsequent research, both concurrent and ongoing, throughout the pandemic has been necessary to identify population at risk of SARS-CoV-2 infection, severe disease, beneficial treatments, death and long-term complications. Clinical data, sourced from electronic health records, had been paramount to identifying these risks.

The novelty of SARS-CoV-2 and COVID-19 brought uncertainty as to who was at risk of infection, who was at risk for death, how should patients be treated and what are the long-term effects. At the start of the pandemic, there was a focus on public health measures, such as proper hygiene, quarantining when sick and reducing close contacts. As the number of cases continued

to rise and hospitals became inundated with patients, researchers set out to identify patients at risk for severe disease and death and to identify existing treatment options that may benefit patients who were hospitalized and suffering from severe disease. Clinical trials and on-going retrospective analysis of patients helped to identify beneficial treatments for patients as well as rule out treatments that were not beneficial or associated with negative outcomes. In one of our studies were identified patients who had a history of macular degeneration and coagulation disorders were at increased risk for severe disease and death as a result of COIVD-19 and identified variants in gene underpinning the inflammatory response as associated with altered risk. In another study using retrospective analysis, we utilized clinical data to identify patients who were intubation and investigated the effect of steroid hormone exposure on the survival of these patients. Our analysis indicated that exposure to melatonin between intubation and extubation was significantly associated with survival in COVID-19 patients and in mechanically ventilated COVID-19 patients. This association was observed when accounting for patient demographics and previous clinical history.

As multiple vaccines have been developed and distributed and therapeutics have become widely available, surges in case counts have not been associated with a proportional rise in hospitalizations and death. Research has shifted to trying to understand the long-term impact of COVID-19 on the health of patients. While viral infections are not uncommon, some can have lasting impacts on patients. With more than 500 million cases reported worldwide long-term analysis of COVID-19 patients and their health after COVID-19 will remain important. Additionally, the incomplete success of vaccination campaigns also highlights the need to monitor any future endemic spikes. While clinical data has been important for conducting studies, they are incomplete and lead to challenges as we transition to an endemic state. To that

end, we trained a random forest classifier to assign a probability of a patient having had COVID-19 during each of their visits and utilized these probabilities to identify clinical phenotypes that are associated with patients who had COVID-19. Within one year, our analysis identified myocardial infarction, urinary tract infection, type 2 diabetes and acute renal failure as being associated with higher probabilities of COVID-19.

The projects presented here demonstrate how to use electronic health records to identify patients at risk for severe disease and death, monitor drug exposure and evaluate its effect on survival of patients with severe COVID-19, how to use machine learning to circumvent the limitations of using clinical data and sets a foundation for further work in identifying the effects of COVID-19. Moreover, these projects also show methods that can be applied to any future emerging disease.

Table of Contents

List of Tables

List of Figures

Acknowledgments

The past four years have been some of the most incredible and impactful of my life both in and out of school. While the COVID-19 pandemic disrupted the traditional progression of graduate school, it gave me the opportunity to be among the thousands of researchers working to understand this new disease in more ways than I could have previously foreseen. In responding to the situation that we have been in for the last two and a half years, I was able to utilized my background, to learn new methods and to develop a new pipeline all while working with live data.

First and foremost, I would like to thank my advisor, Nicholas Tatonetti, for being an amazing mentor. Nick gave me the chance to explore and learn about new areas of research and to make mistakes that would go on to shape my abilites. In joining the Tatonetti lab, I got an opportunity to learn to work with clinical, data, something I had no experience with prior to my rotation, and apply the work that I had done prior to coming to Columbia. I am thankful to all the members of the Tatonetti Lab I have met during my tenure (Anna Basile, Theresa Koleck, Rami Vanguri, Joe Romano, Alexander Yahi, Benjamin May, Phyllis M. Thangaraj, Nicholas

Giangreco, Jenna Kefeli, Katie LaRow Brown, Michael Zietz, Undina Gisladottir, Pietro Belloni, Payal Chandak, Yutaro Tanaka and Kai Chen) for their advice and friendship over the years. I am grateful to Anna, who worked with me during my rotation, for helping me to get my start in the lab. Quite importantly, I am indebted to Ben for his help keeping our live clinical dataset up to date, walking me through the Epic platform and the data we had available and helping me determine, which of my project ideas were possible. During the height of the pandemic, Ben was downloading and transferring data to me daily (including on Saturdays and Sundays).

In deciding to apply to graduate school and identifying areas that were of interest to me, my previous research experience proved invaluable. I am thankful to Trevor Siggers, who welcomed a first semester freshman into his lab, for allowing me to learn molecular biology and biochemistry lab work and computational work. Without all the time that Brian Barron, Ashely Penvose and Nima Mohaghegh spent teaching, I would have never been able to lead my own projects in the Siggers Lab or elsewhere. I am thankful to Karina Yazdanbakhsh for giving me my introduction to translational work, which has become the driving force for me.

I am also thankful to the Department of Physiology and Cellular Biophysics and the program in Cellular and Molecular Physiology & Biophysics for welcoming me to Columbia University and inviting me to be a part of a diverse program where I was able to learn about numerous areas outside of my scope. I am particularly grateful to Panagiota Apostolou, Pedro del Rivero Morfin and Shavonne Teng, with whom I started graduated school, for the great friendships since our interview day.

Thank you to Henry Colecraft, Kam Leong and Andrew Marks for their mentorship completing my qualifying examination and developing a great initial project. Thank you to Benjamin Glickberg, Christopher Mason, Sagi Shapira and Anne-Catrin Uhlemann for advice

and willingness to collaborate on projects. Thank you to X. Shawn Liu, Anne-Catrin Uhlemann and Lawrence Shapiro for their mentorship and guidance during my committee meetings.

Dedication

To my mom (Yvonne), brother (Vick), and sister (Hanna), my extended family and my friends. You have given me every opportunity to explore any and everything that has even remotely sparked my curiosity (and there were a lot). Moreover, you have listened and continued to listen to me talk about whatever sparks my interest, sometimes *ad nauseam*. Thank you for the emotional and physical support you have provided me over the last four years.

Chapter 1: Introduction and Data Processing

1.1 Origins of COVID-19

The disease, which would become known as COVID-19, was first reported to the World Health Organization (WHO) on December 31st, 2019 by public health officials from the People's Republic of China as "cases of pneumonia of unknown etiology" [1]. SARS-CoV-2 was first identified as the cause of COVID-19 in January 2020 [2] and the first infection was retrospectively estimated to have occurred as early as October 2019 in Wuhan, Hubei Provence in China [3]. It was originally reported that the first cases of SARS-CoV-2 were connected to the Huanan Seafood market [4,5], though there has been controversy surrounding the source of SARS-CoV-2 [6]. A study out of the University of Barcelona, which screened wastewater samples for SARS-CoV-2 using realtime quantitative polymerase chain reaction (RT-qPCR) test suggested that SARS-CoV-2 was circulating in Barcelona in early 2019 [7]. However, the study used an incomplete RT-qPCR test and the results could not be reproduced due to a lack of sample. Additionally, political discourse has suggested that SARS-CoV-2 was developed in a laboratory [8]. Recent studies have indicated, with high confidence, that SARS-CoV-2 is a zoonotic infection, which originated in bats, and either jumped to humans either directly [9,10] or via *Nyctereutes procyonoides* (racoon dogs) [11].

1.2 Molecular Mechanism of SARS-CoV-2 Infection

SARS-CoV-2 infection of human cells has been identified to occur via two routes, both of which require interaction with angiotensin-converting enzyme 2 (ACE2) [12]. In the cell surface mechanism of entry, the virus binds to the cell via an interaction between ACE2 and the spike protein on the virial capsid. This is followed by a cleavage of the spike protein by transmembrane serine protease 2 (TMPRSS2), which leads to fusion between the viral capsid and the cellular membrane. This leads to the entry of the viral RNA into the cytoplasm

In the endosomal mechanism of entry, the virus binds to the cell via an interaction between ACE2 and the viral capsid is internalized in an endosome. This is followed by endosomal acidification and cleavage of the spike protein by cathepsin L, which leads to the fusion between the viral capsid and the endosomal membrane. This leads to the entry of the viral RNA into the cytoplasm.

Both routes lead to the entry of viral RNA into the cytoplasm, where host ribosomes translate the replicase polyproteins. These replicase polyproteins protein are processed by viral proteins, which in turn form replication and transcription complexes. These complexes replicate the genome of SARS-CoV-2 and produce accessory proteins and then co-opt the endogenous endoplasmic reticulum to package new viral particles for exocytosis.

1.3 Spread of SARS-CoV-2 and COVID-19

Initial identification of cases was principally done based on clinical symptoms in the absence of other causes [13]. With the release of the sequence of SARS-CoV-2's viral genome, RT-qPCR tests would be developed to identify infected individuals [2, 14]. The first lab confirmed case of COVID-19 outside of the People's Republic of China was identified in Thailand in mid-January 2020 [2]. By the end of January 2020, the first cases were being

reported in Europe and in the United States [2, 15]. Initial tracking of the spread of SARS-COV-2 was hindered by last of tests and test were initially prioritized for patients seeking treatment in hospitals [16].

At the end of January, there were almost 9,927 cases of confirmed COVID-19 around the world with the most being identified in the China [17]. Within the next four weeks, the number of cases has ballooned to 86,023 with China, Italy and Iran leading the case counts [17]. With cases being reported in 114 counties around the world and the evident extensive community spread of COVID-19, the WHO official declared a pandemic on March 11th, 2020 [18].

In the 28 months since the start of the pandemic and the 33 months since the first purported infection (through June 2022), more than 500 million cases have been reported around the world and more than six million people have died [19].

1.4 COVID-19 in the United States

Following the increase in cases in Asia as well as across Europe, the United States began screening passengers entering at major hubs across from Asian countries [2]. By mid-January 2020, the United States began screening passengers coming from Asian countries at ports of entry (New York, Los Angeles, San Francisco) for symptoms [2]. At the time in the United States, nasopharyngeal RT-qPCR testing not scaled up enough to test all passengers arriving from Asian countries at the these ports of entries, nor passengers arriving from Asian counties at other port of entries (e.g. Atlanta, Chicago), not passengers arriving from Europe or other parts of the world [20]. Moreover, as there was still a limited understanding of SARS-CoV-2 and COVID-19 and there was a reluctance from the People's Republic of China to acknowledge that a large part of their cases was due to transmission between individuals without a connection to

the Huanan Seafood Market, testing had not been expanded to the general public who may have interacted with symptoms passengers [21].

The United Sates reported its first confirmed case of COVID-19 in a 56 year old male Washington state resident who had recently returned from Wuhan, PRC [2, 22]. In the weeks that followed, the executive branch of the federal government set up a taskforce to combat COVID-19 with the goal of increasing testing and preparing the public health response [2, 23]. There were 1,147 cases of COVID-19 and 33 deaths attributed reported by March 11th, 2020 when the WHO declare the situation to be a pandemic [2, 17]. There were 2,219 case of COVID-19 United States and 51 deaths reported by March 13th, 2020 when the United States declared a national emergency [2, 17, 24].

Under the guidance of the federal government, air travel was severely curtailed and water travel via cruise ships were halted [2]. Flights originating in Asian countries were severely limited and priorities were given to national returning back to the United States [25]. Passengers arriving were required to quarantine [26]. The United States quarantine requirements were enforced based on an honor policy, unlike other counties, such as the PRC and Australia, where quarantine was done at specific sites, often hotels, with testing through government labs [27].

In the 28 months since the start of the pandemic and the 33 months since the first purported infection, more than 87 million cases have been reported in the United States and more than one million people have died [19].

1.5 COVID-19 in New York State

The first confirmed case of COVID-19 in New York City was reported on March 1st, 2020 [28], though a study from Mount Sinai suggested that SARS-CoV-2 as circulating as early as January 2020 [29]. In the three months that followed, New York City was the epicenter of the

pandemic [30], though by the summer of 2020 had among the lowest rates in the United States [17, 32]. Following weeks of public health guidance for sick individuals to quarantine and for individuals to socially distance, public school and private business were ordered closed on March 16th, 2020 and March 22nd, 2020, respectively [33, 34]. A Pause Order went into effect on March 22nd, 2020 for New York State [35]. In the 28 months since the start of the pandemic and the 33 months since the first purported infection, more than 2.5 million cases have been reported in the New York City and more than 40,000 people have died [19].

1.6 Electronic Health Records

Electronic health records (EHR) capture patient-specific information covering, but not limited to, demographics, diagnoses, medications given, procedures orders, etc. Utilizing the date- and timestamp that accompanies the non-demographic data allows for a live, dynamic and evolving view of how any specific patient is being cared for. With COVID-19 EHR presented data that also mirrored the public health response, e.g. testing, treatment and hospitalization, and allowed for the public health response to be altered as necessary and for officials to respond to new data and studies.

Using EHR data in live and retrospective research, however, is not without its limitations. Borne out of its intended, EHR data is coded for billing purposes often for recoups cost from insurance companies or, as was the case with COVID-19, the federal government of the United States. While diagnoses are able to capture symptoms and conditions that patients are experiencing, they can be limiting when miscellaneous and “not otherwise described" diagnosis codes are utilized. Similarly, information on the procedures ordered can be informative of the diagnostic path or treatment course being utilized by the physician, however the results of those data cannot be readily incorporated into algorithm due to the form of the results, for example

images from magnetic resonance imaging or computed tomography scans, and interpretation of these images from radiologist require natural language processing of notes. Finally, as EHR data is largely entered by a healthcare professional, there is potential human error in the system. For example, patients maybe temporarily duplicated in the system or incorrect information may be enter and fixed at a later date.

That said, laboratory measurements are able to provide in-depth quantitative data, for example complete blood cell counts or blood lipid concentrations. While it is difficult when multiple laboratories are used, internal and external ontologies allow for data to be readily utilized in algorithms. Additionally, the date- and timestamp accompanying procedure data allows, for oxygen treatment periods of patients to be determined. Combined with flowsheet data, it is further possible to identify the method of oxygen treatment, e.g. nasal cannula or endotracheal tube.

The evolving nature of COVID-19 itself was mirrored in the EHR data and the limitations of it varied. Early on in the United States, testing was limited and the public health infrastructure was unable to readily respond to the increasing need for testing. As such, many early studies, including the work presented in Chapter 2, relied on patients being diagnosed with COVID-19 without having an accompanying nasopharyngeal PCR test. Additionally, without a diagnosis code for COVID-19 early on, COVID-19 patients were being identified based on symptoms identified and included in diagnosis date. As the pandemic proceeded and testing became for widely available, having been set up at individual institutions and larger laboratories, data become more reliable and analysis could utilized nasopharyngeal PCR test results to identify patient infected with SARS-CoV-2 (as was done in Chapter 3).

In utilizing data from EHR in the United States, we are presented with the additional limitations of having institution specific information. Patients are able to seek treatment at non-affiliated primary care groups, outside hospitals or urgent care facilities and due to the lack of a universal healthcare recording system, healthcare professionals and researcher using EHR data will only be aware of those encounters if the patient share the details. Sharing of data is becoming easier with the adoption of EPIC by most New York City hospitals, however coding practices can differ between institutions and the sharing procedure is patient initiated.

1.7 Ethics Statement

The work presented in the following three chapters was done using data from patents who sought treatment at New York-Presbyterian (NYP) who had at least one interaction with Columbia University Irving Medical Center (CUIMC) since February 1st, 2020 and any previous data that in our clinical data warehouse (CDW) at NYP/CUIMC. However, these studies were conducted at different points during the pandemic and are censored at April 2020 (Chapter 2), December 2020 (Chapter 3) and March 2022 (Chapter 4). These studies utilizing data in our CDW were approved by the CUIMC Institutional Review Board (IRB# AAAL0601) and the requirement for an informed consent was waived. The data for patients seeking treatment since February 1st, 2020 was made available by a data request associated with AAAL0601, which was submitted to and approved by the Tri-Institutional Request Assessment Committee of New York-Presbyterian, Columbia and Cornell.

Additionally. the work presented in Chapter 2 utilized data from volunteers whose clinical and genomic data are a part of the UK Biobank. The work presented in Chapter 2 also utilized results from SARS-CoV-2 nasopharyngeal RT-qPCR tests for the volunteers in the UK

Biobank. The work done using data from the UK Biobank was conducted under application number 41039.

1.8 Data Sources

The data within the CDW at NYP/CUIMC, which was clinical data available prior to February 1st, 2020, is termed the historical dataset, and the data concurrently available throughout the pandemic since February 1st, 2020 is termed the live dataset. While the historical data did not change, it was utilized differently in different studies. Additionally, when the same information was noted in both historical and live datasets, preference was given to data in the live dataset. For example, if a patient declined to identify their race in the historical data set, but identified their race as Asian in the live dataset, the patient's race indicated as Asian for the analyses below.

The next three subsections will provide a general overview of the data. Any deviations in how the data is utilized will be noted in the methods sections of the subsequent chapters.

1.8.1 Historical Data

From the historical dataset, the demographics data, conditions data and measurements data were utilized (Table 1.1). The demographics data identified gender, date of birth, date of death, if the person had died, race and ethnicity. The conditions data identified conditions diagnosed and the date of diagnosis. The measurements data identified the item being analyzed (e.g. cholesterol, hemoglobin A1C), the date of the measurement, the results of the analysis and the units. The individual was identified using a patient specific unique integer.

Table 4.1: Description of fields in historical dataset

Data Subset	Data Field	Description
Demographics	Person ID	Unique identifier to each patient
	Gender	Female, Male, Other, Ambiguous, Unknown, No matching term
	Date of Birth	Date
	Date of Death	Only if the patient has died
	Race	Asian, Black or African American, White, Other, Unknown, Native American, Hawaiian or Other Pacific Islander, American Indian or Alaskan Native, No matching term
	Ethnicity	Hispanic or Latino or of Spanish origin, Not Hispanic or Latino or of Spanish origin, No matching term
Conditions	Person ID	Unique identifier to each patient
	Condition code	Using Systematized Nomenclature of Medicine - Clinical Terminology (SNOMED-CT) ontology
	Date of Diagnosis	Date
Measurements	Person ID	Unique identifier to each patient
	Measurement code	Using Logical Observation Identifiers Names and Codes (LOINC) ontology
	Date of measurement	Date
	Value of measurement	Integer representing result of the measurement
	Source value of measurement	Text representing result of the measurement
	Units of measurement	Text indicating the units of the value of the measurement

1.8.2 Live Data

From the live dataset, the demographics data, admissions, admissions/discharge/transfer, diagnosis, measurements, vitals, medication, medication administration record, orders and smoking data were utilized (Table 1.2). The demographics data identified gender, date of birth, date of death, if the person had died, race and ethnicity. The admissions data outlined the start date of each visit, the status of the visit (at the time of data retrieval) and the discharge date, if the visit had finished, and encounter identification number for the visit. The admission/discharge/transfer data contained each patient interaction with the visit. The diagnosis data identified conditions diagnosed and the date of diagnosis. The measurements data identified the item being analyzed (e.g. cholesterol, hemoglobin A1C), the date of the measurement, the results of the analysis, the units and a unique order number. The vitals data identified the patient's pulse, respiratory rate, blood pressure, body temperature and blood oxygen saturation. The medication data identified the medication being administered, the method of delivery the quantity, dose, the start date for the medication, the end date for the medication (if the medication was given over a period of time) and a unique order number. The medication administration record data identified the medication being administered, the date the medication was given and a unique order number. The orders data identified the procedure being ordered, the status of the order, the reason for the cancellation of the order (if the order was cancelled), the order type, the order date and a unique order number. The smoking data identified whether or not the patient used tobacco, cigarettes, pipes, cigars, snuff, chewing tobacco or smokeless tobacco products, the start and end dates for use of non-smokeless products, if applicable, and the start and end dates for smokeless products, if applicable.

Medication names were mapped to RXNorm identification number to facilitate analysis. Historical data and live data were able to be used in concert with the other using mappings between the person identification number used in the historical data and the patient's medical reference number used in the live data.

Table 5.2: Description of fields in Live Dataset

Data Subset	Data Field	Description
Demographics	MRN	Medical Record Number unique to each patient
	Gender	Female, Male, Unknown, Non-binary, X, NULL (not indicated)
	Date of Birth	Date
	Date of Death	Only if the patient has died
	Race	Asian, Black or African American, White, Other, Unknown, Ashkenazi Jewish, Hawaiian or Other Pacific Islander, American Indian or Alaskan Native, Sephardic Jewish, Declined, NULL (not indicated)
	Ethnicity	Hispanic or Latino or of Spanish origin, Not Hispanic or Latino or of Spanish origin, Declined, NULL (not indicated)
Admissions	MRN	Medical Record Number unique to each patient
	Admission date	Date of the start of encounter
	Visit Status	Current visit stage
	Discharge date	Date of the end of encounter, if finished
	Encounter ID	Identifier unique to each admission
Admissions/ Discharge/Transfer	MRN	Medical Record Number unique to each patient
	Encounter ID	Identifier unique to each admission
	Series number	Cardinal number indicating each interaction within the encounter
	Event Time	Start date and time of interaction
	Effective Time	End data and time of interaction
Diagnosis	MRN	Medical Record Number unique to each patient
	Diagnosis code	Using International Classification of Diseases version 10 (ICD-10) ontology
	Date of Diagnosis	Date

Table 6.2: Description of fields in Live Dataset (cont.)

Data Subset	Data Field	Description
Measurements	MRN	Medical Record Number unique to each patient
	Order number	Unique identifier for each test ordered
	Procedure name	Name of test ordered
	Component name	Name of each metric analyzed in the test
	Value	Integer or text representing the result of each component analyzed
	Units	Text indicating the units of the value of the measurement
	Order date	Date test was order
	Result date	Date results were obtained
	Measurement code	Using Logical Observation Identifiers Names and Codes (LOINC) ontology
Vitals	MRN	Medical Record Number unique to each patient
	Date	Date field
	Pulse	Number of heart beats per minute
	Respiratory rate	Number of respirations per minute
	Blood pressure	Systolic/Diastolic measurements
	Temperature	Body temperature
	SpO2	Blood oxygen saturation
	Body mass index	Ratio of weight divide by height
Medication	MRN	Medical Record Number unique to each patient
	Order number	Unique identifier for each medication order
	Description of medication	Name of medication
	Delivery	Method of administration
	Quantity	Amount of medication
	Dosage	Amount of medication administered per interaction
	Start date	Date medication first given
	End date	Date medication stopped

Table 7.2: Description of fields in Live Dataset (cont.)

Data Subset	Data Field	Description
Medication Administration Record	MRN	Medical Record Number unique to each patient
	Order number	Unique identifier for each medication order
	Description of medication	Name of medication
	Date of interaction	Date medication given
Orders	MRN	Medical Record Number unique to each patient
	Order number	Unique identifier for each procedure order
	Procedure description	Description of the order
	Order status	The current progress of the order
	Reason for cancellation	If cancelled, why order was cancelled
	Order type	Order category
	Order date	Date order was entered
Smoking	MRN	Medical Record Number unique to each patient
	Tobacco	Whether or not the patient use tobacco products
	Cigarettes	Whether or not the patient use cigarettes
	Pipes	Whether or not the patient used a smoke pipe
	Cigars	Whether or not the patient used cigars
	Snuff	Whether or not the patient used snuff
	Chewing tobacco	Whether or not the patient used chewing tobacco
	Smokeless tobacco products	Whether or not the patient used smokeless tobacco products
	Smoking start date	When the patient started using smoking products, if applicable
	Smoking end date	When the patients stopped using smoking products, if applicable
	Smokeless start date	When the patient started using smokeless products, if applicable
	Smokeless end date	When the patients stopped using smokeless products, if applicable

1.8.3 UK Biobank Data

From the UK Biobank data, the demographics, SARS-CoV-2 testing data and the clinical data were utilized (Table 1.3). The demographics data identified the date of birth and race of participants. For privacy purposes, day of birth was not included in the UK Biobank dataset, so age was calculated from the 1st day of the month in which the participant was born. Testing data identified the date the specimen was collected, the type of specimen collected, the National Health Service laboratory that analyzed the specimen, the facility type where the specimen was collected (e.g. hospital, general practitioner clinic) and the result. Two testing results data were used – the first censored April 18th, 2020 and the second censored May 7th, 2020. The diagnosis data identified clinical diagnose identified by ICD-10 code and date of diagnosis. Participants were identified using a unique reference number (EID).

Additionally, the genotyping data for 337, 147 participants of White British decent was used for the genetic analysis. From the full UK Biobank data set, approximately 50,000 subjects were genotyped on the UK BiLEVE Array by Affymetrix and the remainder were genotyped using the Applied Biosystems UK Biobank Axiom Array. The genotype data covers move than 800,000 variants identified by their GRCh37 (hg19) position. Variants with a minor allele frequency greater than 0.005, a R-squared quality shore greater than 0.03 and a Hardy–Weinberg equilibrium test mid-*P* value less than 10^{-10}.

Table 8.3: Description of fields in UK Biobank Data

Data Subset	Data Field	Description
Demog-raphics	EID	Reference number unique to each participant
	Birth date	Date of birth (year and month only)
	Race	Participant identified race
Testing Results (censored April 18th, 2020 and May 7th, 2020)	EID	Reference number unique to each participant
	Specimen date	Date sample was obtained
	Specimen type	Type of specimen collected
	Laboratory	National Health Service (NHS) laboratory that tested the specimen
	Sample Origin	The type of NHS facility where the sample was collected
	Result	The SARS-CoV-2 test result for the specimen
Diagnoses	EID	Reference number unique to each participant
	Diagnosis Date	Date diagnosis was entered
	Diagnosis	Using International Classification of Diseases version 10 (ICD-10) ontology

1.9 Thesis Overview

Computational analyses have been at the forefront of the response to the COVID-19 pandemic. At the beginning of the pandemic, the focus was on identifying patients at risk of infection, severe disease and death. To that end, I present our work which identified coagulation and complement disorders as affecting risk of severe disease and death and identified genetic variants associated with these effects (Chapter 2). As the pandemic continued, researchers were focused on identifying methods of treating hospitalized patients and developing therapeutics and prophylactics. To that end, I present our work which identified melatonin as being significantly associated with increase survival of intubated COVID-19 patients (Chapter 3). As the situation approaches the endemic state, where spike in disease case counts can be predicted and handled without undue burden on the public health system, the focus of research has shifted to understand

the long-term effects of SARS-CoV-2 infection and COVID-19. To this end, I present our work which developed a novel method to assign a probability of COVID-19 to each patient at each visit, which we in turn used to interrogate more than 1,000 phenotypes for associations with patients who had COVID-19 (Chapter 4).

Chapter 2: Identifying clinical and genetic factors affecting COVID-19 susceptibility, severity and mortality

The work in this chapter is adapted in part from the following publication:

V. Ramlall, P. M. Thangaraj, C. Meydan, J. Foox, D. Butler, J. Kim, B. May, J K. De Freitas, B. S. Glickberg, C. E. Mason, N. P. Tatonetti and S. D. Shapira "Immune complement and coagulation dysfunction in adverse outcomes of SARS-CoV-2 infection". Nature Medicine, vol. 26, pp.1609-1615, October 2020.

DOI: 10.1038/s41591-020-1021-2

2.1 Introduction

In the six months since the start of the COVID-19 pandemic, there have been profound economic, social and public health effects across the world with over 11 million confirmed cases worldwide and over 530,000 deaths [19]. As researchers have been trying to identify patients who are most susceptible to infection as disease, age had been shown to be associated with disease severity and increased mortality driven in part by viral replication and comorbidities, which may influence immune pathology [36,37].

Virial infections exude their effects on their host directly through the initial infection and indirectly through downstream effects caused by interactions between the virus and the host, which can affect the regulatory programs controlling the host's immune pathology [38].

Identifying these interactions have the potential to influence public health measure by identifying groups at higher risk, calling attention to specific manifestations that would influence clinical care and treatment and suggest targets for therapeutic developments. In a study by the Honig and Shapira labs, researchers elucidated that coronavirus proteins structurally mimicked over 140 cellular proteins – notably, all strains of coronavirus queried mimicked proteins in the complement and coagulation pathways [39].

Among its other functions, the complement system mediates the immune response to pathogens, such as bacteria, parasites and viruses [40]. Dysregulation due to genetics, environmental stressors or clinical manifestations can alter the ability of the complement system to correctly direct the immune response and contribute to downstream pathologies due to inflammation [40, 41, 42]. Additionally, the complement system regulates coagulation pathways that are triggered by inflammation in a feedback mechanism that is important for controlling infection-induced pathogenesis. Based on the study from Honig and Shapira labs, coronavirus encodes proteins that mimic complement and coagulation factors may disrupt the endogenous response in humans and in turn allow for the development of induced pathologies. For example, complement dysfunction, as is present in early-onset and age-related macular degeneration (AMD) [41,42,43,44,45], or coagulation dysfunction, as is present in thrombocytopenia, thrombosis and hemorrhage) may impact clinical outcome of SARS-CoV-2 infection.

Based on the mimicry of complement and coagulation proteins by coronaviruses and the clinical observations of hypercoagulation in individuals infected with coronavirus [46,47], we aim to understand the role of complement and coagulatory function in SARS-CoV-2 infection and the effect on clinical outcome. Additionally, we aim to identify genetic variants affecting

complement and coagulation genes that may be associated with COVID-19 susceptibility and disease severity.

2.2 Methods

2.2.1 Software

We used PLINK v.2.00a2LM 64-bit Intel (26 August 2019) to run the genetic association analysis [48]. We used PLINK v.1.90b6.10 64-bit (17 June 2019) to identify haplotype blocks based on linkage disequilibrium. We used Jupyter Notebooks (jupyter-client v.5.3.4 and jupyter-core v.4.6.1) running Python 3.7, numpy 1.18.1 and scipy 1.4.1 for the permutation analyses.

2.2.2 Cohort identification

Between February 1st, 2020 and April 25th, 2020, 11,116 patients were treated at New York-Presbyterian/Columbia University Irving Medical Center (NYP/CUIMC). Of those patients, 6,393 patients either tested positive for SARS-CoV-2 infection or were clinically diagnosed with COVID-19. From the full set of patients, we identified historical data for 6,927 patients who had historical data available before September 24, 2019 in our clinical data warehouse at NYP/CUIMC. Patients' sex, age at first encounter on or after February 1st, 2020 (calculated from date of birth) and smoking status were identified from the live dataset. Patients' race and ethnicity were identified from the historical dataset. Patients who identified sex other than male or female were excluded from the analysis

2.2.3 Defining patient outcome

Using the live data set, we identified patients with severe disease as those requiring intubation during their encounter with NYP/CUIMC. Additionally, for our mortality analysis, we identified patients who died within 28 days of the start of their encounter with NYP/CUIMC using the live dataset.

2.2.4 Identifying patient comorbidities

Using the historical dataset and live dataset, we identified patients with macular degeneration, coagulation disorders, complement deficiency, hypertension, type 2 diabetes, coronary artery disease and obesity. Thrombocytopenia, thrombosis and hemorrhage were used as proxies for coagulation disorders.

2.2.5 Statistical modeling

We used MySQL and Python libraries (pymysql, pandas) to extract and prepare data for modeling. We performed survival analysis on the intubation orders and death using a Cox proportional-hazards model and visualized the risk using Kaplan–Meier curves using the lifelines Python package (v.0.24.4). Error estimates on the Kaplan–Meier curves were estimated using Greenwood's exponential formula. For univariate analysis of age, patients who were at least 65 years old were coded as 1, while those younger than 65 were coded as 0. For univariate analysis of sex, patients who identified as male were coded as 1, while those who identified a female were coded as 0. For univariate analysis of diseases history, patients who had a history of a disease were coded as 1, while those without were coded as 0. For survival analysis, time to event was calculated as time since first encounter to either intubation or death (depending on the analysis) and patients for whom the event occurred were coded as1, while patients for whom it did not were coded as 0. Analysis was censored at 28 days following the first encounter or the last encounter with NYP/CUIMC, whichever occurred first.

2.2.6 Conducting association studies

The UK Biobank contains genotype data for 502,682 participants that profiles approximately 805,426 variants. Of these participants, the genotyping data for 337,147 who were identified to be of White British ancestry were used in our analysis. Association studies were

conducted using PLINK two with filters for minor allele frequency greater than 0.005, a R-squared quality score greater than 0.03 and a Hardy–Weinberg equilibrium test mid-*P* value less than 10^{-10}. Additionally, analyses were performed using a logistic regression model with additive gene dosage and covariates including age at 2018, sex, first ten principal components (UK Biobank) and the genotyping array that the sample was carried out on. The α threshold for study-wide significance using an empirical permutation analysis. Association studies compared subjects that who tested positive for a SARS-CoV-2 infection and required hospitalization to the entire population of 337,147 subjects

2.2.7 Identifying haplotype blocks

Using the genotype data of the 337,147 participants of White British ancestry, we identified haplotype black based on linkage disequilibrium using PLINK1.9 where the lower 90% confidence interval is greater than 0.70 and the upper 90% confidence interval is at least 0.98. Haplotype blocks containing any part of the genes of interest were first identified and subsequently variants outside of the genes of interest, which were a part of the blocks. Of the full dataset of 805,426 variant profiled in the UK Biobank genotype data, 7,281 variants were within the genes of interest. After applying additional quality control filters using PLINK2, 936 variants remained for analysis.

2.3 Results

2.3.1 Identifying patient cohort

From the live dataset, we identified 11,116 patients who sought treatment at NYP/CUIMC between February 1st, 2020 and April 25th, 2020 (Table 2.1). Among those patients, 6,393 patients tested positive for SARS-CoV-2 infection by a RT-qPCR test or were clinically diagnosed with COVID-19. The average age of all patients was 52.0, while the average

age of COVID-19 patients was 57.1 years. Similar proportions of all patients and COVID-19 patients identified as Asian, Black or African American, White or Other; a similar proportion of both groups declined to identify their race (Table 2.1_. Similar proportions of all patients and COVID-19 patients identified as Hispanic or Latino or of Spanish origin and not Hispanic or Latino or of Spanish origin; a similar proportion of both groups identified ethnicity as other or declined to identify their ethnicity (Table 2.1). A similar proportional of patients in both groups were past or current smokers (Table 2.1). There was a higher proportion of all patients who required mechanical ventilation than COVID-19 patients (9.2% and 7.6%, respectively) (Table 2.1) and a similar proportion of patients in both groups died within 28 days (10.2% and 9.7%, respectively). A similar proportion of patients with a history of hypertension, type 2 diabetes, obesity and coronary artery disease (Table 2.2).

Table 2.1: Demographics and outcome frequencies of all patients and COVID-19 patients

	All Patients	COVID-19+
N	11,116	6,393
Average Age (IQR)	52.0 (34.7–69.5)	57.1 (41.5–72.0)
Male	4,980 44.8%	3,177 49.7%
Hispanic or Latin or Spanish origin	3,535 31.8%	2,186 34.2%
Not Hispanic or Latin or Spanish origin	4,391 39.5%	2,365 37.0%
Asian	300 2.7%	153 2.4%
Black or African American	2,357 21.2%	1,419 22.2%
White	3,479 31.3%	1,816 28.4%
Other	2,957 26.6%	1,784 27.9%
Past or Current Smoker	2,979 26.8%	1,643 25.7%
Mechanical ventilation required	1,023 9.2%	484 7.6%
Death within 28 days	1,134 10.2%	618 9.7%

From the historical and live dataset, we identified 88 patients with COVID-19 who had a history of macular degeneration, 4 who had a history of complement deficiency, 1,239 who had a history of coagulation disorders (Table 2.2). The average age of the COVID-19 patients with macular degeneration, complement deficiency and coagulation disorders were statistically equal to the full COVID-19 data set, though the patients were younger than the other two groups (Table 2.3). The proportion of COVID-19 patients with macular degeneration and coagulation disorders who identify as male is lower than in the full COVID-19 dataset. The proportion of COVID-19 patients with macular degeneration and coagulation disorders who identify as Hispanic or Latino or of Spanish origin is lower than in the full data set; the proportion who identify as not Hispanic or Latino or of Spanish origin is higher (Table 2.3). The proportion of COVID-19 patients with macular degeneration and coagulation disorders who identify as Black or African American is lower than in the full dataset; the proportion identifying as White is higher (Table 2.3). The proportion of COVID-19 patients with macular degeneration and coagulation disorders who required mechanical ventilation is greater than in the full dataset (Table 2.3). The mortality rate among COVID-19 patients with macular degeneration and coagulation disorders is higher than the mortality rate among the full dataset (Table 2.3).

Table 2.2: Past clinical history frequencies of all patients and COVID-19 patients

	All Patients	COVID-19+
N	11,116	6,393
History of hypertension	3,135 28.2%	1,988 31.1%
History of type 2 diabetes	1,401 12.6%	911 14.2%
History of obesity	1,334 12.0%	831 13.0%
History of coronary artery disease	2,979 26.8%	1,698 26.6%

Table 2.3: Demographics and outcome frequencies of COVID-19 subsets patients

	Macular degeneration	Complement deficiency	Coagulation disorders	Cough
N	88	4	1,239	725
Average Age (IQR)	74.1 (67.2–84.6)	57.9 (49.1–70.9)	61.8 (48.2–77.0)	59.2 (46.6–72.0)
Male	37 42.0%	2 50%	522 42.1%	387 53.4%
Hispanic or Latin or Spanish origin	22 25%	1 25%	383 30.9%	370 51.0%
Not Hispanic or Latin or Spanish origin	52 59.1%	2 50%	607 49.0%	183 25.2%
Asian	0 0%	0 0%	24 1.9%	12 1.7%
Black or African American	15 17.0%	1 25%	250 20.2%	132 18.2%
White	32 36.4%	0 0%	422 34.1%	204 28.1%
Other	25 28.4%	2 50%	302 24.4%	228 31.4%
Past or Current Smoker	26 29.5%	2 50%	331 26.7%	185 25.5%
Mechanical ventilation required	14 15.9%	0 0%	126 10.2%	80 11.1%
Death within 28 days	22 25%	0 0%	212 17.1%	110 15.2%

2.3.2 Cox proportional hazards analysis leading to intubation

We conducted a Cox proportional hazards analysis of COVID-19 patients to identify whether or not a history of macular degermation and coagulation disorder was associated with a specific outcome. Additionally we conducted similar analysis for COVID-19 patients to identify whether or not hypertension, type 2 diabetes, obesity, coronary artery disease, being over the age of 65 and being a current or past smoker was also associated with a specific outcome. Due to the small number of patients with a history of complement deficiency, we were unable to investigate the outcome associated with those patients. Patients with a cough were used as a reference int eh analysis.

In univariate analysis, history of macular degeneration (Hazards ratio = 2.16, 95% CI: 1.30–3.67, p = 4.63E-3 and coagulation disorders (Hazards ratio = 1.50, 95% CI:1.23–1.83, p = 9.64E-5) were associated with patients being intubated; similar hazards ratios were noted when controlling for age and sex (Hazards ratio = 1.83 , 95% CI: 1.07–3.13, p = 2.73E-2 and Hazards ratio = 1.47, 95% CI: 1.20–1.82), p = 2.43E-4, respectively). A history of hypertension, type 2 diabetes, obesity and coronary artery disease were associated with patient needing to be intubated in a univariate analysis and when controlling for age and sex (Table 2.4, Figure 2.1). Being a past or current smoker was not associated with patients being intubated (Hazards ratio = 1.13, 95% CI: 0.873–1.46, p = 0.353) in a univariate analysis and when controlling for age and sex (Hazards ratio = 0.962, 95% CI: 0.740–1.25, p = 0.775).

Table 2.4: Hazards ratios from univariate Cox proportional hazards analysis for intubation

Covariate	Univariate Hazards Ratio	Age and sex controlled Hazards ratios
History of macular degeneration	2.16 (1.30–3.67) p = 4.63E-3	1.83 (1.07–3.13) p = 2.73E-2
History of coagulation disorder	1.50 (1.23–1.83) p = 9.64E-5	1.47 (1.20–1.82) p = 2.43E-4
History of hypertension	1.74 (1.46–2.09) p = 1.29E-9	1.56 (1.29–1.88) p = 4.24E-6
History of type 2 diabetes	1.85 (1.50–2.29) p = 1.23E-8	1.63 (1.30–2.02) p = 2.32E-5
History of obesity	1.30 (1.02–1.65) p = 3.74E-2	1.46 (1.14–1.86) p = 2.83E-3
History of coronary artery disease	1.99 (1.66–2.39) p = 8.11E-14	1.80 (1.49–2.17) p =1.19E-3
Age ≥ 65	1.68 (1.41–2.01) p = 1.08E-8	1.30 (0.961–1.77) p = 8.85E-2
Past or current smoker	1.13 (0.873–1.46) p = 0.353	0.962 (0.740–1.25) p = 0.775
Cough	1.46 (1.14–1.86) p = 2.54E-3	1.41 (1.10–1.80) p = 6.47E-3

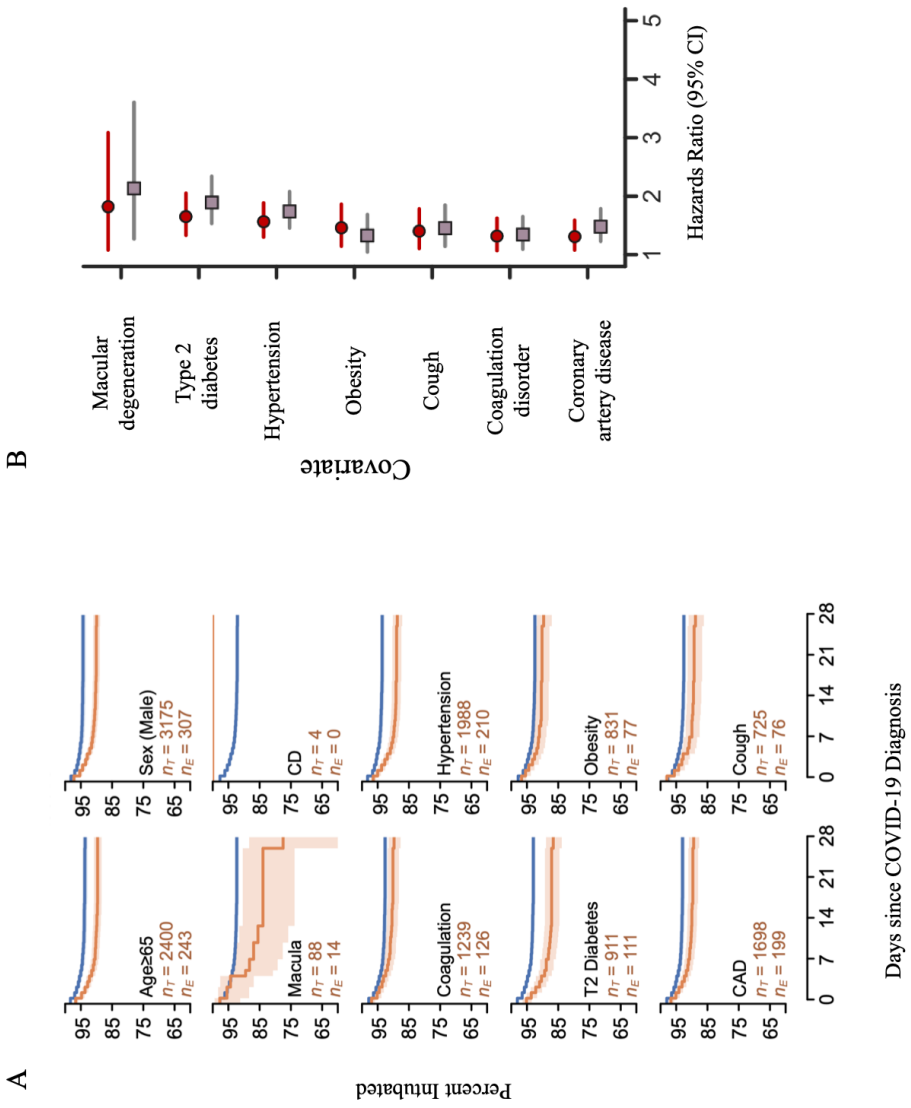


Figure 2.1: (A) Univariate Kaplan-Meier curves (B) Comparison of hazards ratios from Cox proportional hazards analysis for intubation

2.3.4 Cox proportional hazards analysis leading to death

Similar to the analysis of COVID-19 patients to identify the risk associated with needing intubation, we investigated the risk associated with death. In univariate analysis, history of macular degeneration (Hazards ratio = 2.99, 95% CI: 1.96–4.58, p = 4.39E-7) and coagulation disorders (Hazards ratio = 2.33, 95% CI:1.98–2.76, p = 1.85E-23) were associated with increased patient mortality. Similar hazards ratios were noted when controlling for age and sex in analysis of history of macular degeneration (Hazards ratio = 1.53, 95% CI: 0.998–2.35, p = 5.09E-2) and history of coagulation disorder (Hazards ratio = 1.81, 95% CI: 1.53–2.14), p = 3.43E-12) - albeit the association was not as significant for history of macular degeneration. A history of hypertension, type 2 diabetes, obesity and coronary artery disease were associated with increased patient mortality in a univariate analysis and when controlling for age and sex (Table 2.5, Figure 2.2). Being a past or current smoker was associated with increased patient mortality (Hazards ratio = 1.53, 95% CI: 1.21–1.92, p = 3.14E-4) in a univariate analysis and though there was no significant risk identified when controlling for age and sex (Hazards ratio = 1.08, 95% CI: 0.857–1.36, p = 0.512).

Table 2.5: Hazards ratios from univariate Cox proportional hazards survival analysis for death

Covariate	Univariate Hazards Ratio	Age and sex controlled Hazards ratios
History of macular degeneration	2.99 (1.96–4.58) p = 4.39E-7	1.53 (0.998–2.35) p = 5.09E-2
History of coagulation disorder	2.33 (1.98–2.76) p = 1.85E-23	1.81 (1.53–2.14) p = 3.43E-12
History of hypertension	3.75 (3.19–4/41) p = 3.90E-58	2.30 (1.96–2.71) p = 1.02E-23
History of type 2 diabetes	2.93 (2.47–3.48) p = 4.59E-35	1.98 (1.67–2.35) p = 5.32E-15
History of obesity	1.61 (1.32–1.98) p = 3.82E-6	1.92 (1.56–1.36) p = 7.13E-10
History of coronary artery disease	3.69 (3.15–4.33) p = 1.46E-58	2.23 (1.89–2.62) p =3.71E-22
Age ≥ 65	8.80 (7.14.–10.9) p = 1.51E-91	1.68 (1.23–2.28) p = 9.50E-4
Past or current smoker	1.53 (1.21–1.92) p = 3.14E-4	1.08 (0.857–1.36) p = 0.512
Cough	1.32 (1.06–1.65) p = 1.35E-2	1.32 (1.06–1.65) p = 1.49E-2

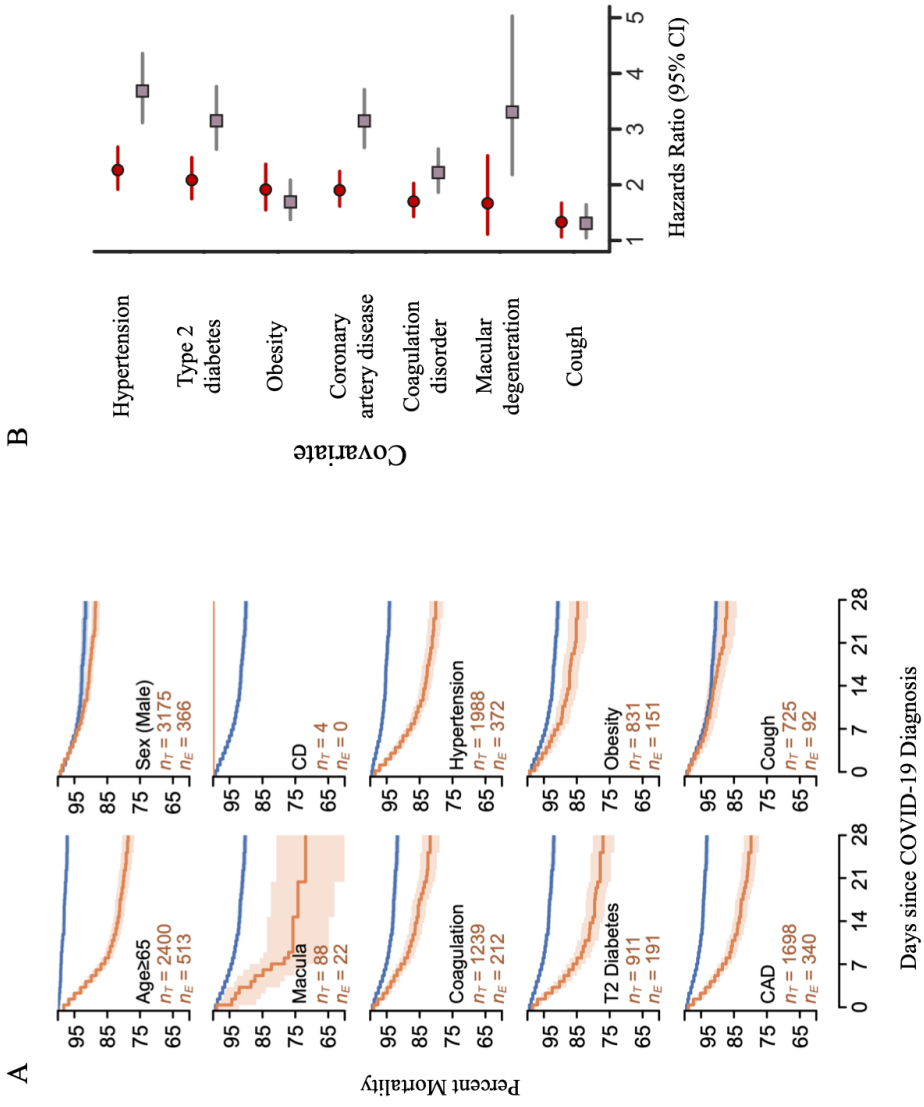


Figure 2.2: (A) Univariate Kaplan-Meier curves (B) Comparison of hazards ratios from Cox proportional hazards analysis for death

2.3.3 Identify genetic variants

With the noted associations between complement and coagulation dysfunction in the dataset at NYP/CUIMC, we sought to determine if specific genetic variation in the complement and coagulation pathways were associated with the adverse outcomes. From the Kyoto Encyclopedia of Genes and Genomes we identified 102 genes associated with the regulation of the complement and coagulation cascade. From the 805,426 variants profiled in the UK Biobank, we identified 2,888 that were within the sequence of those 102 genes or within 60 kbp upstream or downstream of the gene.

From the UK Biobank release of SARS-CoV-2 RT-qPCR test results in April 2020, we identified 388 individuals who tested positive for infection and 332 individuals who tested positive for infection and were hospitalized. A targeted association study using these individuals identified 11 variants across seven genes (F3, CFH, C4BPB, CR2, F13A1, SERPING1, SERPINF2 and CD3) with a significance value less than 0.001 (Table 2.6, Figure 2.3). The eight variants across genes F3, CFH, C4BPB, CR2, F13A1, SERPING1 and SERPINF2 have odds ratios suggesting the variants are associated with an adverse outcome (Odds ratio > 1), while the three variants in gene C3 are associated with less adverse outcomes in SARS-CoV-2 infected individuals (Odds ratio <1).

Table 2.6: Significant variants from April 2020 association study

Gene	Variant	Position	Odds Ratio	Significance
F3	rs72729504	1:94940206	1.93	4.24E-04
CFH	rs12064775	1:196600605	2.13	3.71E-04
C4BPA	rs45574833	1:207300070	2.65	1.20E-05
C4BPA	rs61821041	1:207352581	2.34	2.74E-04
CR2	rs61821114	1:207610967	2.40	3.94E-05
F13A1	rs3024329	6:6316135	1.43	9.88E-04
SERPING1	rs117284601	11:57425228	1.80	1.06E-04
SERPINF2	rs9913923	17:1703982	1.48	5.59E-04
C3	rs2230203	19:6710782	0.660	2.57E-04
C3	rs1047286	19:6713262	0.657	1.02E-04
C3	rs2230199	19:6718387	0.684	3.92E-04

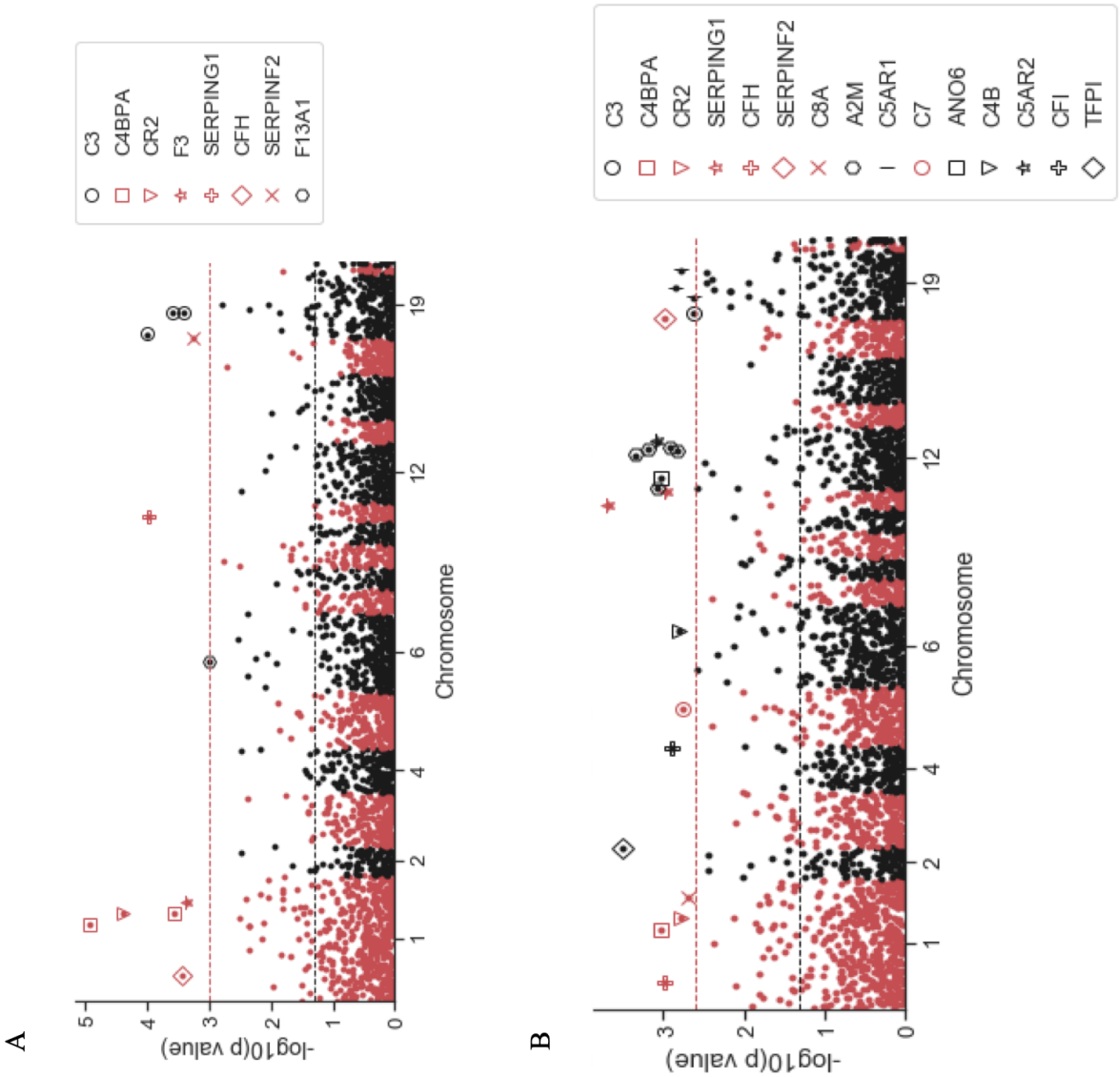


Figure 2.3: Manhattan plots of variants from individuals from (A) April 2020 and (B) May 2020.

From the UK Biobank release of SARS-CoV-2 RT-qPCR test results in May 2020, we identified 651 individuals who test positive for infection and 500 intubations who tested positive for infection and were hospitalized. A similar targeted association study focused on the data fro May 2020 identified 22 variants across 15 genes (C8A, CFH, C4BPA, CR2, TFPI, CFI, C7, C4B, SERPING1, A2M, ANO6, C5AR2, SERPINF2, C3, C5AR1) with a significance value less than 0.0025 (Table 2.7, Figure 2.3). The variants identified in C8A, CFH, C4BPA, CR2, TFPI, CFI, C7, C4B, ANO6, SERPINF2 and C5AR1 have odds ratios suggesting the variants are associated with adverse outcomes in SARS-CoV-2 infected individuals (Odds ratio > 1). The variant in C5AR2 is associated with less adverse outcomes in SARS-CoV-2 (Odds ratio < 1). Variants in SERPING1, A2M and C3 are associated with both adverse and less adverse outcomes.

Table 2.7: Significant variants from May 2020 association study

Gene	Variant	Position	Odds Ratio	Significance
C8A	rs7556361	1:57276195	1.31	2.06E-03
CFH	rs12064775	1:196600605	1.84	1.03E-03
C4BPA	rs45574833	1:207300070	1.99	9.25E-04
CR2	rs61821114	1:207610967	1.85	1.64E-03
TFPI	rs8176612	2:188349145	1.51	3.10E-04
CFI	rs78730002	4:110677014	1.77	1.28E-03
C7	rs55945585	5:40903166	1.62	1.72E-03
C4B	rs6447	6:32008924	1.83	1.59E-03
SERPING1	rs78958998	11:57317971	0.655	1.08E-03
SERPING1	rs117284601	11:57425228	1.62	2.10E-04
A2M	rs669	11:9232268	0.790	6.54E-04
A2M	rs10842898	12:9262289	0.795	8.50E-04
A2M	rs7297589	12:9273449	0.809	1.18E-03
A2M	rs61916033	12:9285480	1.36	1.46E-03
A2M	rs4883215	12:9314207	0.780	4.54E-04
ANO6	rs117316516	12:45797209	1.78	9.33E-04
C5AR2	rs74504130	12:47284908	0.549	8.60E-04
SERPINF2	rs9913923	17:1703982	1.37	1.03E-03
C3	rs1047286	19:6713262	0.776	2.32E-03
C5AR1	rs140754743	19:47803469	1.90	2.44E-03
C5AR1	rs4804049	19:47823484	1.91	1.66E-03
C5AR1	rs2910425	19:47847760	1.91	1.39E-03

Additionally, association studies were conducted for the 936 identified in the haplotype blocks. Using the data form the patients who tested positive for SARS-CoV-2 infection in April 2020, we identified 16 variants across eight genes (F5, CF, C4BPA, COLEC11, CF1, F13A1, ANO6 and C3) with a significance value less than 0.01 (Table 2.8). The variants in CFH, C4BPA, COLEC11, CF1, F13A1 and ANO6 have an odds ratios suggesting an association with adverse outcome (Odds ratio > 1). The variants in F5 and C3 have odds ratios suggesting an association with less adverse outcomes (Odds ratio <1). Using the data form the patients who tested positive for SARS-CoV-2 infection in May 2020, we identified 14 variants across ten genes (C4BPA, COLEC11, GGCX, TFPI, CF1, F13A1, A2M, ANO6, C3 and C5AR1) with a significance value less than 0.0075 (Table 2.9). The variants in C4BPA, COLEC11, TFPI, CF1, F13A1, ANO6 and C5AR1 have an odds ratios suggesting an association with adverse outcome (Odds ratio > 1). The variants in GGCX, A2M and C3 have odds ratios suggesting an association with less adverse outcomes (Odds ratio <1).

Table 2.8: Significant variants from April 2020 association study using haplotype blocks

Gene	Variant	Position	Odds Ratio	Significance
F5	rs2213866	1:169489512	0.767	4.52E-03
F5	rs6032	1:169511555	0.758	3.06E-03
F5	rs4525	1:169511734	0.767	4.50E-03
F5	rs4524	1:169511755	0.767	4.39E-03
CFH	rs35634602	1:196696857	1.53	7.20E-03
C4BPA	rs45574833	1:207300070	2.65	1.20E-05
C4BPA	rs75202466	1:207303477	1.83	9.12E-03
COLEC11	rs731034	2:3677022	1.33	3.37E-03
CFI	rs78730002	4:110677014	1.87	3.38E-03
CFI	rs79891491	4:110679456	1.61	6.83E-03
F13A1	rs3024329	6:6316135	1.43	9.88E-04
ANO6	rs117316516	12:45797209	1.85	3.19E-03
C3	rs344548	19:6685817	0.740	8.93E-03
C3	rs2230203	19:6710782	0.660	2.57E-04
C3	rs1047286	19:6713262	0.657	1.02E-04
C3	rs2230199	19:6718387	0.684	3.92E-04

Table 2.9: Significant variants from May 2020 association study using haplotype blocks

Gene	Variant	Position	Odds Ratio	Significance
C4BPA	rs45574833	1:207300070	1.99	9.25E-04
COLEC11	rs731034	2:3677022	1.27	3.55E-03
GGCX	rs12714145	2:85787341	0.826	3.63E-03
TFPI	rs8176612	2:188349145	1.51	3.10E-04
CFI	rs78730002	4:110677014	1.77	1.28E-03
F13A1	rs3024329	6:6316135	1.30	4.77E-03
A2M	rs669	12:9232268	0.790	6.54E-04
A2M	rs10842898	12:9262289	0.795	8.50E-04
ANO6	rs117316516	12:45797209	1.78	9.33E-04
C3	rs2230203	19:6710782	0.790	6.59E-03
C3	rs1047286	19:6713262	0.776	2.32E-03
C3	rs2230199	19:6718387	0.799	6.84E-03
C5AR1	rs4467185	19:47823038	1.85	3.46E-03
C5AR1	rs4804049	19:47823484	1.91	1.66E-03

2.4 Discussion

Analysis of viral genomes have identified viral proteins that can mimic the structure of endogenous human proteins. In a concurrent study, the Shapira lab at Columbia University identified that coronavirus structurally mimicked 140 human proteins a subset of which are important in the complement and coagulation pathways [49]. With the role of these proteins identified, we wanted to investigate whether or not patients with diseases resultant from dysfunction of these pathways were more likely to experience adverse outcomes (severe disease or mortality

Using the data for 11,116 patients who sought treatment at NYP/CUIMC, we identified that patients with a history of macular degeneration and coagulation disorders were at increased risk of severe disease and death when analyzing the covariate alone and when accounting for the age and se of the patient. These results were consistent with our hypothesis that dysfunction of pathways involved in the body's defense from pathogen or coagulation induced by inflation were associated with higher risk of adverse outcome. Additionally, the results identifying that history of hypertension, type 2 diabetes, obesity and coronary artery disease were all associated with increased risk of severe disease and death when analyzing the covariate alone and when accounting for the age and sex of the patient were consistent with other studies. Being over the age of 65 and identifying as male were also associated with an increased risk of severe disease or death.

Counterintuitively, this study identified that being a current or past smoker was not associated with an altered risk to disease severity even when accounting for the age and sex of the patients. Being a past or current smoker, was associated with an increased risk of death in COVID-19 patients, however the analysis when accounting for age and sex indicated that there was no altered risk of death. It was expected that patients who had pulmonary distress as a result of smoking would be at increased risk of severe disease or death, however our results suggest that is not the case. While researchers are still trying to understand this, it is thought that scarring in the interstitial space of the lung is preventing signal transduction to active these pathways.

In a partner study at New York-Presbyterian/Weill-Cornell Medical Center, researchers used RNA sequencing to transcriptional profile 650 nasopharyngeal swabs [50]. The results suggested that SARS-CoV-2 infection induced genes in pathways that modulate immune

function. The data showed that the regulation of the complement and coagulation pathways were influenced by SARS-CoV-2 infection. Their results concurred with previous results that showed poor clinical outcome in the upregulation of type I interferons and the dysregulation of the interleukin-6 inflammatory response [51].

In a final part to this study, we used SARS-CoV-2 test results and genotyping data from the UK Biobank to investigated whether genetic variants in genes associated with the complement and coagulation cascade were enriched among individuals who tested positive for SARS-CoV-2 and were hospitalized. The analysis using data available in April 2020 identified rs72729504 a variant in coagulation factor III, a protein which is associated with fibrin fragment D-dimer levels and used as a clinical marker for activated blood coagulation, as being associated with increased risk of adverse clinical outcome [52]. Variants (rs1047286, rs2230203 and rs2230199 in complement factor 3 are associated with a decreased risk of an adverse outcome suggesting a protective effect. Variants rs61821114 and rs61821041, which are known to decrease the expression of complement decay-accelerating factor 55 that functions to disrupt the inflammatory cascade and preventing immune-mediated damage [53].

In analysis of data available in May 2020, the association study identified variants rs10842898, rs669 and rs4883215 that are associated with decreased expression of α-2-macroglobulin, which regulates fibrin clot formation and inflammatory cascades [54]. Additionally, variants rs10842898 and rs669 that affect splice variants in mannose-6-phosphate receptor, which is a P-type lectin that regulates lysosomal cargo loading and participates in cellular responses to wound healing, cell growth and viral infection [55].

Finally, in association studies of variants in haplotype blocks, variant rs45574833 that causes a missense variant in complement component 4 binding protein alpha, which negatively

regulates the classical complement pathway [56]. Finally, variant rs731034 decreases expression of collection subfamily member 11, which binds carbohydrate antigens on microorganism to facilitate their recognition and phagocytosis.

Chapter 3: Investigating steroid hormone exposure on outcome in intubated and mechanically ventilated COVID-19 patients

The work in this chapter is adapted in part from the following publication:

V. Ramlall, J. Zucker and N. P. Tatonetti. "Melatonin is significantly associated with survival of intubated COVID-19 patients". medRχiv, October 2020.

DOI: 10.1101/2020.10.15.20213546

3.1 Introduction

The coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection has impacted every country in the world with more than 35 million confirmed cases and more than 1 million deaths globally - the United States accounts for more than 20% of total cases and deaths [19]. In the ten months since the first infections were reported outside of the original epicenter, clinical research remains focused on identifying treatments [57,58,59] and preventive measures [60, 61, 62] for SARS-CoV-2 infection.

Analyses of healthcare data from infected patients have identified the most frequent symptoms, e.g. fever, cough, fatigue, shortness of breath, loss of taste or smell [63], however less frequent symptoms influenced by comorbidities are also observed [49,64,65]. Respiratory distress remains the most significant and serious symptom of COVID-19 [66], which, in the most

severe of cases, can require endotracheal intubation and mechanical ventilation, and even lung transplants [67]. At the core of the public health emergency that has ravaged the world, is the limited amount of supplies and number of intensive care unit beds [68]. Furthermore, patients requiring respiratory support, intubation, oxygen supplementation or invasive mechanical ventilation are bearing the brunt of the limited availability of resources [69].

Among the possible therapies for SARS-CoV-2 infection being researched, hormone drugs, such as dexamethasone [69] and methylprednisolone [70], have proved promising. The dexamethasone study from the RECOVERY Collaborative Group in the UK found that patients overall (22.9%), patients requiring invasive mechanical ventilation (29.3%) and patients receiving oxygen without mechanical ventilation (23.3%) treated with dexamethasone had lower death rates at 28 days compared to those who were treated with usual care (25.7%, 41.4% and 26.2%, respectively) [69].

While steroid hormone drugs all share a common basic ring structure, they are used for specific circumstance dictated by ailment, potency, half-life and side effects. Based on the results from the RECOVERY trial in the UK, we were interested in understanding whether or not exposure to other hormone steroids led to similar effects, which would allow for the identification of other steroid hormones that may have similar therapeutic benefit.

3.2 Methods

3.2.1 Statistical modeling and software

We used Jupyter Notebooks (jupyter-client version 5.3.4 and jupyter-core version 4.6.1) running Python 3.7 and all fit models using the python lifelines package (version 0.24.4) [48]. We used MySQL and python libraries (pymysql, numpy, pandas and pickle) to extract and prepare the data for modeling.

3.2.2 Cohort identification

The data electronic health records for 189,987 patients who sought care at NYP/CUIMC between February 1st, 2020 and August 1st, 2020 were collected. From those patients, we identified a cohort of 13,394 patients who tested positive for SARS-CoV-2 infection using a nasopharyngeal RT-qPCR test or were clinically diagnosed with COVID-19. From those patients we identified 948 oxygen therapy periods for 791 patients who required oxygen therapy (both with and without mechanical ventilation). We identified 3,497 periods for 2,981 patients who required oxygen therapy (both with and without mechanical ventilation) , but were not diagnosed with COVID-19. We also identified 747 oxygen therpay periods for 637 patients, who sought care at NYP/CUIMC between February 1st, 2018 and August 1st, 2018, requiring oxygen therapy (both with and without mechanical ventilation) from our clinical data warehouse.

3.2.3 Identifying oxygen therapy periods and ventilator use

To begin, we used hospital admission and discharge data to identify visits lasting more than one day, which eliminated any patient who was admitted for an outpatient procedure. We identified the start of oxygen therapy using the display name and description of orders that had been completed - we used a similar method to identify the end of oxygen therapy - and filtered for visits where the start date and discharge date were not the same to remove any patients who were intubated during a surgical procedure. We used the end dates of oxygen therapy to define the end of each oxygen therapy period for each patient and identified the start of each oxygen therapy period by the first order to start oxygen therapy following any previous extubation order. Any periods beginning before February 1st, 2020 (i.e. those with an order to end oxygen therapy, but missing an order to begin oxygen therapy), were excluded from the analysis. Periods without

an accompanying order to end oxygen therapy censored up to the discharge date or the final date of analysis, August 1st, 2020. For patients who died within seven days following an order to end oxygen therapy, the periods were censored up to the death date.

For the 2018 oxygen therapy periods, patients, who had been started oxygen therapy, were identified using procedures identified as ‘Intubation, endotracheal, emergency procedure’, ‘Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening’, or ‘Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening Endoscopic’. We identified days on which patients began oxygen therapy using procedures identified as ‘'Unlisted procedure, larynx' or ‘Subsequent hospital care, per day, for the evaluation and management’. We used the dates of procedures to define the start of each oxygen therapy periods and the maximum date of care related to oxygen therapy following the start of oxygen therapy procedure as the end of each period.

For the 2020 oxygen therapy periods, patients requiring the use of a ventilator were identified using the display name and description of order that had not been cancelled such that the order date occurred during the oxygen therapy period. Oxygen therapy periods for each patient with mechanical ventilation order were coded as 1, while those without were coded as 0. For the 2018 oxygen therapy periods, patients requiring the use of a ventilator were identified as “Respiratory Ventilation, Less than 24 Consecutive Hours”, “Respiratory Ventilation, 24-96 Consecutive Hours”, and “Respiratory Ventilation, Greater than 96 Consecutive Hours”. Similarly, oxygen therapy periods for each patient with a ventilation order were coded as 1, while those without were coded as 0.

3.2.4 Identifying demographic information

For each patient within our cohorts of interest, we identified the patient's reported birth date, death date, if the patient had died, sex, race(s) and ethnicity. We calculated the age of each patient on their first admission to the hospital within the observation period. For sex, patients identifying their sex as male were identified as 1, while those who did not reported their sex as male (e.g. female or unknown) were coded as 0. For ethnicity, patients who reported their ethnicity as Hispanic or Latino or Spanish origin were coded as 1, while those who reported their ethnicity as Not Hispanic or Latino or Spanish origin or who did not report their ethnicity were coded as 0. For each possible racial group: American Indian or Alaskan Native, Native Hawaiian or Other Pacific Islander, Ashkenazi Jewish, Black or African American, Asian, and White, patients were coded as 1 if they reported identifying as a member of that racial group.

3.2.5 Identifying patient comorbidities

For patients in each cohort, we used the data available in the EHR at NYP/CUIMC and in the CDW at NYP/CUIMC to identify whether or not a patient had a history of asthma, cardiovascular disease, chronic kidney disease, chronic obstructive lung disease, coronary artery disease, delirium, diabetes mellitus, diabetic nephropathy, diabetic neuropathy, diabetic retinopathy, diabetic vasculopathy, heart failure, hypertension, insomnia, myocardial infarction, obesity, and respiratory disorder using ICD-10 diagnosis codes in the EHR and SNOMED-CT codes and relationships in data from the CDW. For each disease in our survival analysis, patients with a history were coded as 1, while those without a history were coded as 0.

3.2.6 Identifying patient drug treatments

For the 2020 oxygen therapy periods, we identified the drug names, the associated National Library of Medication RXNorm identification code and the time of order from orders

that had been completed or time of action from medication administration record. We then mapped RXNorm codes to DrugBank codes and utilized the associated DrugBank categories to identified drugs classified as hormones. Patients were considered as being treated with a drug before oxygen therapy if they had at least one completed order or administration between February 1st, 2020 and the start of an oxygen therapy period outside of any other oxygen therapy period. Similarly, patients were considered as being treated with a drug during oxygen therapy if they had at least one completed order or administration on or after the start of the oxygen therapy period through the end of the period.

For the oxygen therapy periods from 2018, we identified drug names, the associated RX Norm identification code, and the start date and end date of their drug regiment censored between February 1st, 2018 and August 1st, 2018. Patients were considered as being treated with a drug before oxygen therapy if any part of the treatment period occurred between the visit start day and the oxygen therapy start date outside of any other oxygen therapy period. Similarly patients were considered as being treated with a drug during oxygen therapy if any part of the treatment period occurred between the start of oxygen therapy and the end date or the censoring date.

3.2.7 Identify patient outcomes

For patients with a single oxygen therapy period, we identified patients for whom oxygen therapy was not beneficial as those who died within the period or within seven days following the end of that period. For patients with multiple oxygen therapy periods, we identified patients for whom oxygen therapy was not beneficial as those who died within the final oxygen therapy period or within seven days following the end of that period. For our survival analysis, oxygen therapy periods where the patient did not die within seven days of the end date were coded as 0,

while those who died within seven days were coded as 1. For oxygen therapy periods that did not result in death within seven days, time to event was equal to the length of the oxygen therapy period. For oxygen therapy periods that did not result in death within seven days, time to event was equal to the length of time from the start date of oxygen therapy to the death date of the patient.

3.3 Results

3.3.1 Identify patient outcomes

We conducted a retrospective observational study of 189,987 patients who sought care at NYP/ CUIMC between February 1st, 2020 and August 1st, 2020. We identified 13,394 patients who were diagnosed with COVID-19 or infected with SARS-CoV-2 and 948 oxygen therapy periods among the 791 patients who received oxygen therapy. Additionally, we identified 3,497 oxygen therapy periods among the 2,981 patients who required oxygen therapy and were not diagnosed with COVID-19 nor infected with SARS-CoV-2, which served as the controls in this study (Table 3.1). From our clinical data warehouse (CDW), we identified 747 oxygen therapy periods among the 637 patients who required oxygen therapy between February 1st, 2018 and August 1st, 2018 (Table 3.1). Of the oxygen therapy periods for COVID-19 patients, there were 315 periods where the patient required mechanical ventilation and 276 where the patient died within seven days of the end of oxygen therapy (i.e. negative outcome), 242 non-COVID-19 oxygen therapy periods where the patient required mechanical ventilation and 143 where the patient died within seven days of the end of oxygen therapy, 637 oxygen therapy periods from 2018 where the patient required mechanical ventilation and 174 where the patient died within seven days of the end of oxygen therapy (Table 3.1). The median (interquartile range) age of the COVID-19, non-COVID-19, and 2018 patients who required oxygen therapy was 56.52 (0 -

95.14) years, 46.44 (0-97.72) years and 52.62 (0 - 118.00), respectively, and 60.86%, 50.73% and 58.90%, respectively, self-identified as male (Table 3.1).

Additionally, 1.37% (N=13), 20.57% (N=195) and 28.38% (N=269) of the COVID-19 oxygen therapy periods, 2.17% (N=76), 15.67% (N=548) and 44.35% (N=1,151) of non-COVID-19 oxygen therapy periods and 1.47% (N=11), 14.99% (N=112), 35.74% (N=267) of oxygen therapy periods in 2018 were for patients who self-identified as Asian, Black or African American and White, respectively, and 46.20% (N=438) of the COVID-19 patients for COVID-19 patients who identified as of Hispanic or Latin or Spanish origin compared to 25.68% (N=898) of the non-COVID-19 oxygen therapy periods and 20.21% (N=151) of oxygen therapy periods from 2018 (Table 3.1).

Table 3.1: Demographics and outcome frequencies of COVID-19, non-COVID-19 and 2018 oxygen therapy periods' patients

	COVID-19 +	COVID-19 -	2018
N	948	3497	747
Average Age (IQR)	56.52 (0.0, 95.14)	46.44 (0.0, 97.72)	52.62 (0.0, 118.00)
Age ≥ 65	409 43.14%	1125 32.17%	315 42.17%
Male	577 60.86%	1774 50.73%	440 58.90%
Hispanic or Latin or Spanish origin	438 46.20%	898 25.68%	151 20.21%
American Indian or Alaskan Native	≤10 ≤1.05%	11 0.31%	≤10 ≤1.34%
Asian	13 1.37%	76 2.17%	11 1.47%
Black or African American	195 20.57%	548 15.67%	112 14.99%
Native Hawaiian or Other Pacific Islander	≤10 ≤1.05%	≤10 ≤0.27%	≤10 ≤1.34%
White	269 28.38%	1551 44.35%	267 35.74%
Mechanical ventilation required	315 33.23%	242 6.92%	637 85.27%
Death within 7 days of end of oxygen therapy	276 29.11%	143 4.09%	174 23.29%

3.3.2 Identify patient comorbidities

More than 50% of the COVID-19, non- COVID-19 and 2018 oxygen therapy periods were for patients who had a history of cardiovascular disease and respiratory disease (Table 3.2). Additionally, 70.78%, 43.04%, 20.89%, 43.78% and 47.47% of the COVID-19 oxygen therapy periods were for patients who had a history of chronic kidney disease, diabetes mellitus, heart failure, hypertension and obesity, respectively, compared to 37.36%, 20.65%, 14.81%, 33.77% and 35.05% of the non-COVID-19 oxygen therapy periods and 34.67%, 37.88%, 51.81%, 63.05% and 25.84% of oxygen therapy periods from 2018 (Table 3.2).

Table 3.2: Frequency of diseases of COVID-19 +, COVID-19 - and 2018 oxygen therapy periods' patients.

	COVID-19 +	COVID-19 -	2018
N	948	3497	747
Asthma	135 14.24%	336 9.61%	148 19.81%
Cardiovascular disease	662 69.83%	2149 61.45%	723 96.79%
Chronic kidney disease	671 70.78%	1303 37.26%	259 34.67%
Chronic obstructive lung disease	85 8.97%	244 6.98%	187 25.03%
Coronary artery disease	145 15.30%	539 15.41%	265 35.48%
Diabetes mellitus	408 43.04%	722 20.65%	283 37.88%
Diabetic nephropathy	56 5.91%	107 3.06%	64 8.57%
Diabetic neuropathy	52 5.49%	104 2.97%	57 7.63%
Diabetic retinopathy	28 2.95%	48 1.37%	23 3.08%
Diabetic vasculopathy	26 2.74%	58 1.66%	38 5.09%
Heart failure	198 20.89%	518 14.81%	387 51.81%
Hypertension	415 43.78%	1181 33.77%	471 63.05%

Table 3.2: Frequency of diseases of COVID-19 +, COVID-19 - and 2018 oxygen therapy periods' patients. (cont.)

Disease	COVID-19 +	COVID-19 -	2018
Myocardial infarction	80 8.44%	216 6.18%	160 21.42%
Obesity	450 47.47%	1225 35.03%	193 25.84%
Respiratory disease	706 74.47%	1819 52.02%	746 99.87%

3.3.3 Univariate analysis of demographics on outcome following oxygen therapy

Among oxygen therapy periods for COVID-19 patients, increasing age, as a continuous variable (HR: 1.05, 95% CI: 1.04 -1.06, p-value = 4.42E-24) and as a binary variable of age greater than or equal to 65 years (HR: 3.25, 95% CI: 2.52 - 4.9, p-value = 1.33E-19), and self-identifying as (HR: 4.36, 95% CI: 1.08 - 17.6, p-value = 3.80E-02) was significantly associated with a negative outcome (Table 3.3 and Figure 3.1). Among the subset of oxygen therapy periods where mechanical ventilation was required for COVID-19 patients increasing age, as a continuous variable (HR: 1.04, 95% CI: 1.02 -1.06, p-value = 1.39E-5) and as a binary variable of age greater than or equal to 65 (HR: 2.85, 95% CI: 1.84 - 4.40, p-value = 2.42E-06) was significantly associated with a negative outcome (Table 3.3 and Figure 3.2). Conversely, self-identifying race as Black or African American (HR: 0.347, 95% CI: 0.175 - 0.689, p-value = 2.49E-3) was significantly associated with a positive outcome (Table 3.3 and Figure 3.2).

Table 3.3: Demographic and disease univariate Cox proportional hazards ratios for COVID-19 + oxygen therapy periods. (± Not Determined)

Covariates	Intubated Patients		Mechanically Ventilated Patients	
	N	Hazards Ratio (95 % CI) P-value	N	Hazards Ratio (95 % CI) P-value
Age (continuous)	948	1.05 (1.04-1.06) 4.42E-24	315	1.04 (1.02-1.06) 1.39E-05
Age >= 65	409	3.25 (2.52-4.19) 1.33E-19	144	2.85 (1.84-4.40) 2.42E-06
Male	577	1.09 (0.848-1.39) 0.513	196	0.980 (0.639-1.50) 0.927
Hispanic or Latin or Spanish origin	438	1.00 (0.793-1.27) 0.969	166	1.09 (0.719-1.64) 0.697
American Indian or Alaskan Native	3	4.36 (1.08-17.6) 3.80E-02		±
Asian	13	1.00 (0.321-3.13) 0.995		±
Black or African American	195	0.923 (0.687-1.24) 0.595	62	0.344 (0.166-0.710) 3.92E-03
Native Hawaiian or Other Pacific Islander	3	0.602 (8.44E-02-4.29) 0.612		±
White	269	0.842 (0.634-1.12) 0.232	77	1.30 (0.813-2.07) 0.275

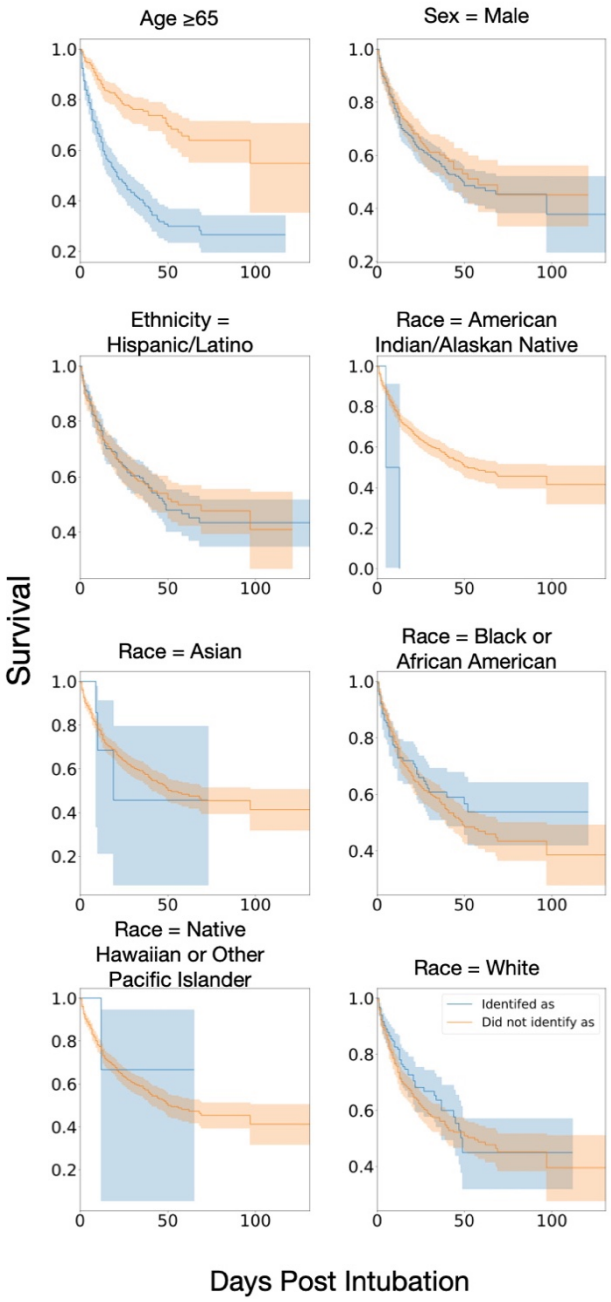


Figure 3.1: Kaplan-Meier curves for demographic covariates for COVID-19 oxygen therapy periods.

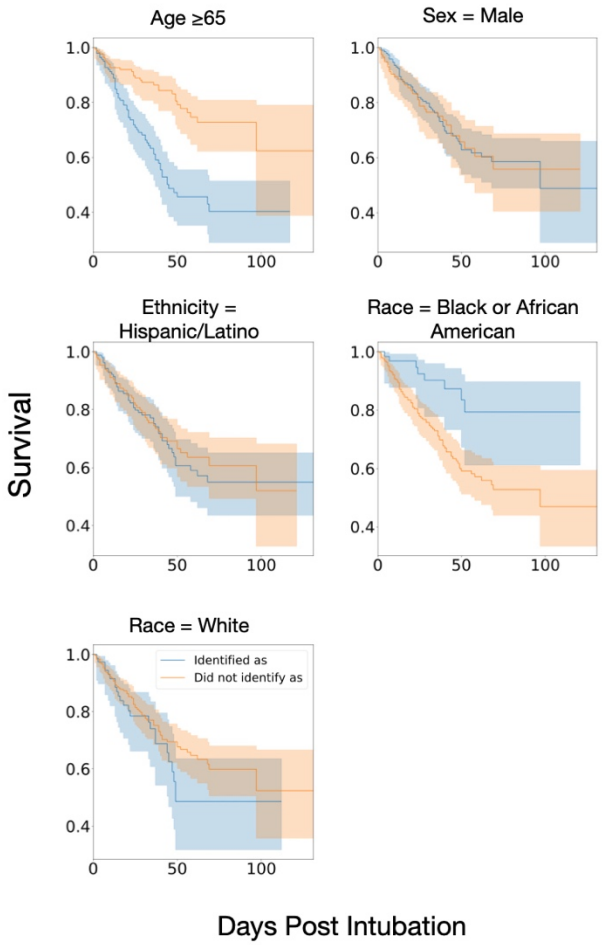


Figure 3.2: Kaplan-Meier curves for demographic covariates for COVID-19 oxygen therapy periods requiring mechanical ventilation.

3.3.4 Univariate analysis of comorbidities on outcome following oxygen therapy

Among oxygen therapy periods for COVID-19, having a history of chronic kidney disease (HR: 5.90, 95% CI: 3.40 - 10.4, p-value = 3.73E-10), chronic obstructive lung disease (HR: 1.82, 95% CI: 1.29 - 2.58, p-value = 7.37E-4), coronary artery disease (HR: 1.65, 95% CI: 1.23 - 2.21, p-value = 8.41E-04), diabetes mellitus (HR: 1.61, 95% CI: 1.27 - 2.04 , p-value =8.83E-05), hypertension (HR: 1.62, 95% CI: 1.20 - 1.93 , p-value = 5.20E-04) and myocardial infarction (HR: 1.56, 95% CI: 1.07 - 2.27, p-value = 1.96E-02) were associated with a negative

outcome. Having a history of respiratory disease (HR: 0.630, 95% CI: 0.480 - 0.830, p-value = 1.06E-30) was significantly associated with a positive outcome following oxygen therapy (Table 3.4 and Figure 3.3).

Among the subset of oxygen therapy periods where mechanical ventilation was required for COVID-19, having a history of chronic kidney disease (HR: 3.63, 95% CI: 1.32 - 9.89, p-value = 1.17E-02) and chronic obstructive lung disease (HR: 2.06, 95% CI: 1.07 - 3.98, p-value = 3.11E-02) were significantly associated with a negative outcome (Table 3.4 and Figure 3.4). Conversely, having a history of asthma (HR: 0.299, 95% CI: 9.50E-02 - 0.950, p-value = 4.00E-2) and respiratory disease (HR: 0.457, 95% CI: 0.273 - 0.766, p-value = 2.98E-3) were significantly associated with a positive outcome (Table 3.4 and Figure 3.4).

Table 3.4: Disease univariate Cox proportional hazards ratios for COVID-19 oxygen therapy periods. ± Not determined. (± Not Determined)

Covariates	Intubated Patients		Mechanically Ventilated Patients	
	N	Hazards Ratio (95 % CI) P-value	N	Hazards Ratio (95 % CI) P-value
Asthma	135	0.984 (0.696-1.39) 0.929	32	0.299 (9.46E-02-0.945) 3.98E-02
Cardiovascular disease	662	0.981 (0.752-1.28) 0.891	240	0.858 (0.538-1.37) 0.521
Chronic Kidney disease	671	5.94 (3.40-10.4) 3.73E-10	270	3.63 (1.33-9.89) 1.17E-02
Chronic obstructive lung disease	85	1.82 (1.29-2.58) 7.37E-04	20	2.06 (1.07-3.98) 3.11E-02
Coronary artery disease	145	1.65 (1.23-2.21) 8.41E-04	32	1.31 (0.695-2.46) 0.407
Diabetes mellitus	408	1.61 (1.27-2.04) 8.83E-05	138	1.33 (0.884-2.01) 0.171
Diabetic nephropathy	56	0.723 (0.414-1.26) 0.254	21	0.710 (0.260-1.94) 0.503
Diabetic neuropathy	52	1.16 (0.716-1.87) 0.552	15	0.719 (0.227-2.28) 0.574

Table 3.4: Disease univariate Cox proportional hazards ratios for COVID-19 oxygen therapy periods. ± Not determined. (± Not Determined) (cont.)

Covariates	Intubated Patients		Mechanically Ventilated Patients	
	N	Hazards Ratio (95 % CI) P-value	N	Hazards Ratio (95 % CI) P-value
Diabetic retinopathy	28	0.966 (0.478-1.95) 0.923	7	0.492 (6.85E-02-3.54) 0.481
Diabetic vasculopathy	26	0.958 (0.451-2.03) 0.910		±
Heart failure	198	1.30 (0.989-1.70) 5.98E-02	60	1.05 (0.635-1.74) 0.843
Hypertension	415	1.52 (1.20-1.93) 5.20E-04	138	1.04 (0.684-1.57) 0.864
Myocardial infarction	80	1.56 (1.07-2.27) 1.96E-02	13	0.742 (0.235-2.35) 0.612
Obesity	450	0.810 (0.639-1.03) 8.10E-02	193	0.691 (0.457-1.04) 7.87E-02

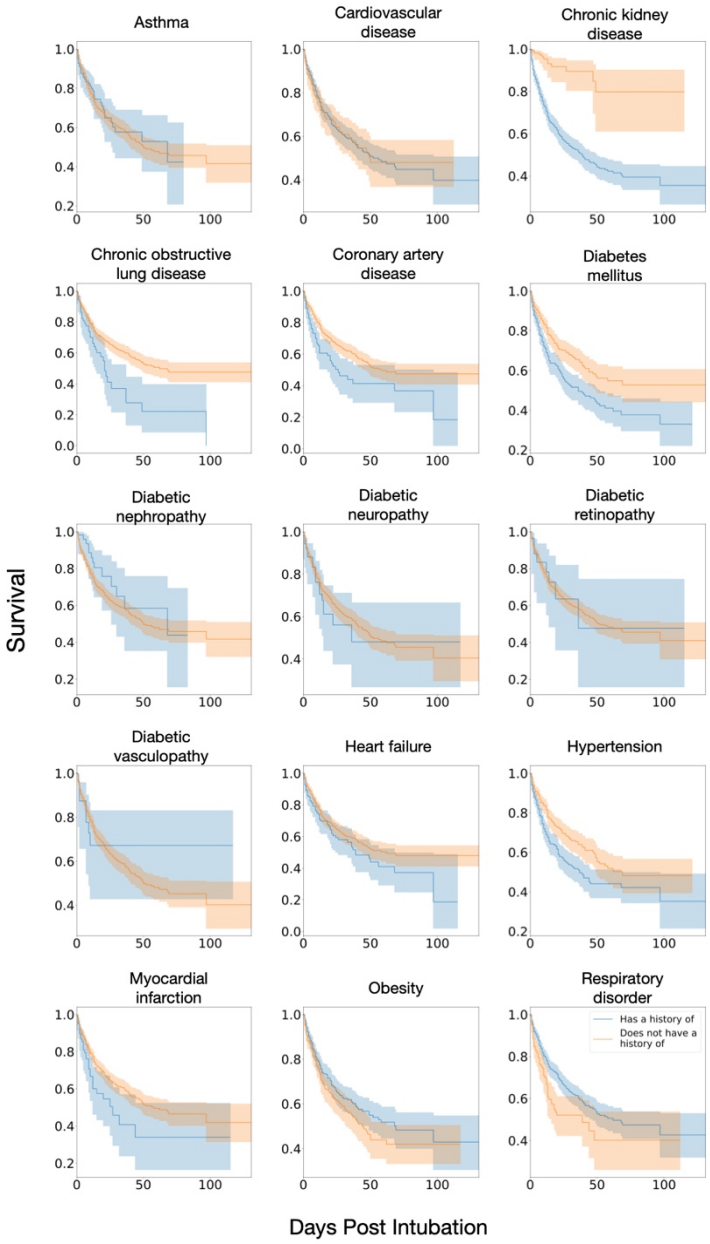


Figure 3.3: Kaplan-Meier curves for disease covariates for COVID-19 oxygen therapy periods.

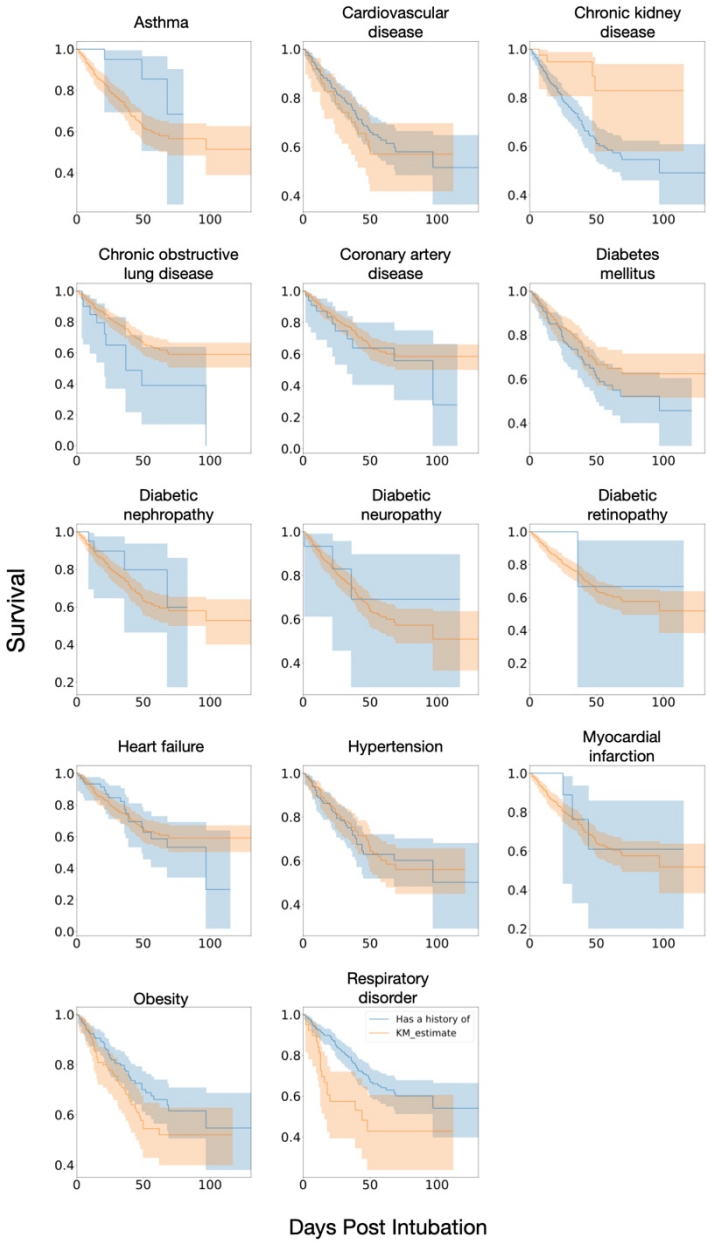


Figure 3.4: Kaplan-Meier curves for disease covariates for COVID-19 oxygen therapy periods requiring mechanical ventilation.

3.3.5 Identify drug exposure on outcome following oxygen therapy

Among the COVID-19 +, COVID-19 – and 2018 oxygen therapy periods' patients, the largest faction of patients were exposed to benzodiazepines prior to oxygen therapy (18.35%, 15.56%, 69.34%, respectively) (Table 3.5). Among the COVID-19 + oxygen therapy periods' patients, 9.07% of them were exposed to melatonin; comparatively, approximately half the fraction of COVID-19 – oxygen therapy periods' patients were exposed (4.92%), while approximately a third larger fraction of 2018 patients were exposed (Table 3.5).

Among the COVID-19 +, COVID-19 – and 2018 oxygen therapy periods' patients, the largest faction of patients were exposed to benzodiazepines after starting oxygen therapy (76.90%, 71.03%, 81.79%), respectively) (Table 3.6). Among the COVID-19 + and 2018 oxygen therapy periods' patients, at least one-fifth were exposed to melatonin after starting oxygen therapy (20.68%, 25.03%, respectively); comparatively, approximately half the fraction of COVID-19 – oxygen therapy periods' patients were exposed to melatonin after starting oxygen therapy (Table 3.6).

Table 3.5: Fraction of oxygen therapy periods where the patient was exposed to the drug before oxygen therapy. (± Not Determined)

Drug	COVID-19 +	COVID-19 -	2018
Quetiapine	52 5.49%	53 1.52%	77 10.31%
Trazodone	13 1.37%	28 0.80%	27 3.61%
Benzodiazepines	174 18.35%	544 15.56%	518 69.34%
Insulin glargine	138 14.56%	129 3.69%	±
Insulin Human	150 15.82%	162 4.63%	±
Dronabinol	≤10 ≤1.05%	≤10 ≤0.27%	±
Hydrocortisone	34 3.59%	74 2.12%	±
Triamcinolone	≤10 ≤1.05%	≤10 ≤0.27%	±
Budesonide	18 1.90%	67 1.92%	±
Melatonin	86 9.07%	172 4.92%	93 12.45%
Dexamethasone	≤10 ≤1.05%	67 1.92%	78 10.44%
Vasopressin	37 3.90%	73 2.09%	±

Table 3.5: Fraction of oxygen therapy periods where the patient was exposed to the drug before oxygen therapy. (± Not Determined) (cont.)

Drug	COVID-19 +	COVID-19 -	2018
Prednisone	41 4.32%	98 2.80%	±
Methylprednisolone	146 15.40%	104 2.97%	±
Levothyroxine	36 3.80%	78 2.23%	±
Fludrocortisone	≤10 ≤1.05%	≤10 ≤0.27%	±
Insulin Lispro	172 18.14%	237 6.78%	±

Table 3.6: Fraction of oxygen therapy periods where the patient was exposed to the drug during oxygen therapy periods. (± Not Determined)

Drug	COVID-19 +	COVID-19 -	2018
Quetiapine	252 26.58%	119 3.40%	217 29.05%
Trazodone	20 2.11%	55 1.57%	49 6.56%
Benzodiazepines	729 76.90%	2484 71.03%	611 81.79%
Insulin glargine	341 35.97%	273 7.81%	±
Insulin Human	486 51.27%	432 12.35%	±
Dronabinol	≤10 ≤1.05%	12 0.34%	±
Hydrocortisone	196 20.68%	198 5.66%	±
Triamcinolone	15 1.58%	23 0.66%	±
Budesonide	33 3.48%	108 3.09%	±
Melatonin	196 20.68%	319 9.12%	187 25.03%
Dexamethasone	18 1.90%	206 5.89%	81 10.84%
Vasopressin	221 23.31%	283 8.09%	±

Table 3.6: Fraction of oxygen therapy periods where the patient was exposed to the drug during oxygen therapy periods. (± Not Determined) (cont.)

Drug	COVID-19 +	COVID-19 -	2018
Prednisone	60 6.33%	171 4.89%	±
Methylprednisolone	286 30.17%	315 9.01%	±
Levothyroxine	63 6.65%	201 5.75%	±
Salmon Calcitonin	≤10 ≤1.05%	≤10 ≤0.27%	±
Fludrocortisone	≤10 ≤1.05%	15 0.43%	±
Insulin Lispro	362 38.19%	493 14.10%	±
Desmopressin	≤10 ≤1.05%	19 0.54%	±
Clobetasol	≤10 ≤1.05%	≤10 ≤0.27%	±

3.3.6 Univariate analysis of hormone exposure on outcome following oxygen therapy

We used univariate analysis of hormone exposure during oxygen therapy to identify hypotheses for follow up analysis. Among the subset of periods for COVID-19 patients during which the patient required mechanical ventilation, exposure to methylprednisolone (HR: 1.63, 95% CI: 1.07 - 2.47 p-value = 2.37E-02) and levothyroxine (HR: 2.26, 95% CI: 1.13 - 4.51, p-

value = 2.04E-02) prior to the start of the oxygen therapy were significantly associated with a negative outcome (Table 3.7, Figure 3.5).

Exposure to insulin glargine (HR: 0.665, 95% CI: 0.521 - 0.849 p-value = 1.04E-03), budesonide (HR: 0.290, 95% CI: 0.108 - 0.778, p value = 1.40E-02), melatonin (HR: 9.17E-02, 95% CI: 5.43E-02 - 0.155, p-value = 4.81E-19), prednisone (HR: 0.432, 95% CI: 0.230 - 0.812. p-value = 9.11E-03), methylprednisolone (HR: 0.773, 95% CI: 0.603 - 0.991, p-value = 4.25 E-02) and insulin lispro (HR: 0.731, 95% CI: 0.575 - 0.930, p-value = 1.07E-02) between the start day of the oxygen therapy period and the end day were significantly associated with a positive outcome in periods of COVID-19 patients (Table 3.8, Figure 3.6). Among the same oxygen therapy periods, exposure to hydrocortisone (HR: 1.56, 95% CI: 1.22 - 2.00, p-value = 3.54E-04) during the oxygen therapy period was significantly associated with negative outcomes (Table 3.8, Figure 3.6).

Exposure to melatonin (HR: 9.13E-02, 95% CI: 4.40E-02 - 0.189, p-value = 1.32E-10) during the oxygen therapy period was significantly associated with a positive outcome in oxygen therapy periods for COVID-19 patients requiring mechanical ventilation (Table 3.7, Figure 3.7). Conversely, exposure to hydrocortisone (HR: 2.16 95% CI: 1.42 - 3.28, p-value = 2.98E-04), methylprednisolone (HR: 1.73, 95% CI: 1.13 - 2.64, p-value = 1.19E-02) and levothyroxine (HR: 1.89, 95% CI: 1.05 - 3.40, p-value = 3.43E-02) during the oxygen therapy period was significantly associated with a negative outcome (Table 3.7, Figure 3.7).

Table 3.7: Univariate Cox proportional hazards ratios for hormone exposure prior to start of oxygen therapy period for COVID-19 + patients. (± Not Determined)

Drug Name (DrugBank ID)	Oxygen Therapy Periods		Mechanically Ventilated Periods	
	N	Hazards Ratio (95 % CI), P-value	N	Hazards Ratio (95 % CI), P-value
Insulin glargine (DB00047)	138	0.892 (0.653-1.22), p = 0.471	70	1.28 (0.805-2.02), p = 0.299
Insulin Human (DB00030)	150	0.826 (0.606-1.13), p = 0.225	74	0.902 (0.554-0.147), p = 0.679
Hydrocortisone (DB00741)	34	1.04 (0.583-1.86), p = 0.893	19	1.71 (0.826-3.53), p = 0.149
Triamcinolone (DB00620)	6	0.547 (7.67E-02-3.89), p = 0.547		±
Budesonide (DB01222)	18	0.703 (0.262-1.89), p = 0.484	7	0.936 (0.230-3.80), p = 0.927
Melatonin (DB01065)	86	0.635 (0.403-1.00), p = 5.10E-02	39	1.26 (0.701-2.27), p = 0.439
Dexamethasone (DB01234)	7	0.475 (6.66E-02-3.38), p = 0.457	3	1.57 (0.218-11.3), p = 0.654
Vasopressin (DB00067)	37	0.626 (0.322-1.22), p = 0.167	19	1.18 (0.516-2.71), p = 0.692
Prednisone (DB00635)	41	0.716 (0.381-1.35), p = 0.299	16	0.819 (0.300-2.23), p = 0.697
Methylprednisolone (DB00959)	146	1.01 (0.759-1.36), p = 0.922	98	1.63 (1.07-2.47), p = 2.37E-02

Table 3.7: Univariate Cox proportional hazards ratios for hormone exposure prior to start of oxygen therapy period for COVID-19 + patients. (± Not Determined) (cont.)

Drug Name (DrugBank ID)	Oxygen Therapy Periods		Mechanically Ventilated Periods	
	N	Hazards Ratio (95 % CI), P-value	N	Hazards Ratio (95 % CI), P-value
Levothyroxine (DB00451)	36	1.01 (0.550-1.84), $p = 0.984$	20	2.26 (1.13-4.51)), $p = 2.04E-02$
Fludrocortisone (DB00687)	7	0.407 (5.72E-02-2.90)), $p = 0.370$		±
Insulin Lispro (DB00046)	172	0.974 (0.731-1.30)), $p = 0.857$	85	1.14 (0.724-1.78)), $p = 0.580$

Table 3.8: Univariate Cox proportional hazards ratios for hormone exposure during oxygen therapy period for COVID-19 + patients. (± Not Determined)

Drug Name (DrugBank ID)	Oxygen Therapy Periods		Mechanically Ventilated Periods	
	N	Hazards Ratio (95 % CI), P-value	N	Hazards Ratio (95 % CI), P-value
Insulin glargine (DB00047)	341	0.665 (0.521-0.849), p = 1.04E-03	181	0.832 (0.543-1.27), p = 0.398
Insulin Human (DB00030)	486	1.03 (0.790-1.33), p = 0.846	253	0.924 (0.538-1.59), p = 0.776
Hydrocortisone (DB00741)	196	1.56 (1.22-2.00), p = 3.54E-04	112	2.16 (1.42-3.28), p = 2.98E-04
Budesonide (DB01222)	33	0.290 (0.108-0.778), p = 1.40E-02	15	0.165 (2.30E02-1.19), p = 7.37E-02
Melatonin (DB01065)	196	9.17E-02 (5.43E-02-0.155), p = 4.81E-19	112	9.13E-02 (4.40E-0.2-0.189), p = 1.32E-10
Dexamethasone (DB01234)	18	0.167 (2.35E-02-1.19), p = 7.43E-02	4	0.880 (0.122-6.34), p = 0.899
Vasopressin (DB00067)	221	1.02 (0.796-1.32), p = 0.854	135	1.31 (0.865-1.98), p = 0.202
Prednisone (DB00635)	60	0.432 (0.230-0.812), p =9.11E-03	20	0.417 (0.132-1.32), p = 0.136
Methylprednisolone (DB00959)	286	0.773 (0.603-0.991), p = 4.25E-02	154	1.73 (1.13-2.64), p = 1.19E-02
Levothyroxine (DB00451)	63	0.790 (0.484-1.29), p = 0.347	32	1.89 (1.05-3.40), p = 3.43E-02

Table 3.8: Univariate Cox proportional hazards ratios for hormone exposure during oxygen therapy period for COVID-19 + patients. (± Not Determined) (cont.)

Drug Name (DrugBank ID)	Oxygen Therapy Periods		Mechanically Ventilated Periods	
	N	Hazards Ratio (95 % CI), P-value	N	Hazards Ratio (95 % CI), P-value
Salmon Calcitonin (DB00017)	3	0.535 (7.49E-02-3.82), $p = 0.533$	±	
Fludrocortisone (DB00687)	9	0.205 (2.87E-02-1.46), $p = 0.114$	±	
Insulin Lispro (DB00046)	362	0.731 (0.575-0.930), $p = 1.07E-02$	174	0.868 (0.571-1.32), $p = 0.505$
Clobetasol (DB11750)	2	0.937 (0.131-6.68), $p = 0.948$	2	1.81 (0.251-13.0), $p = 0.557$

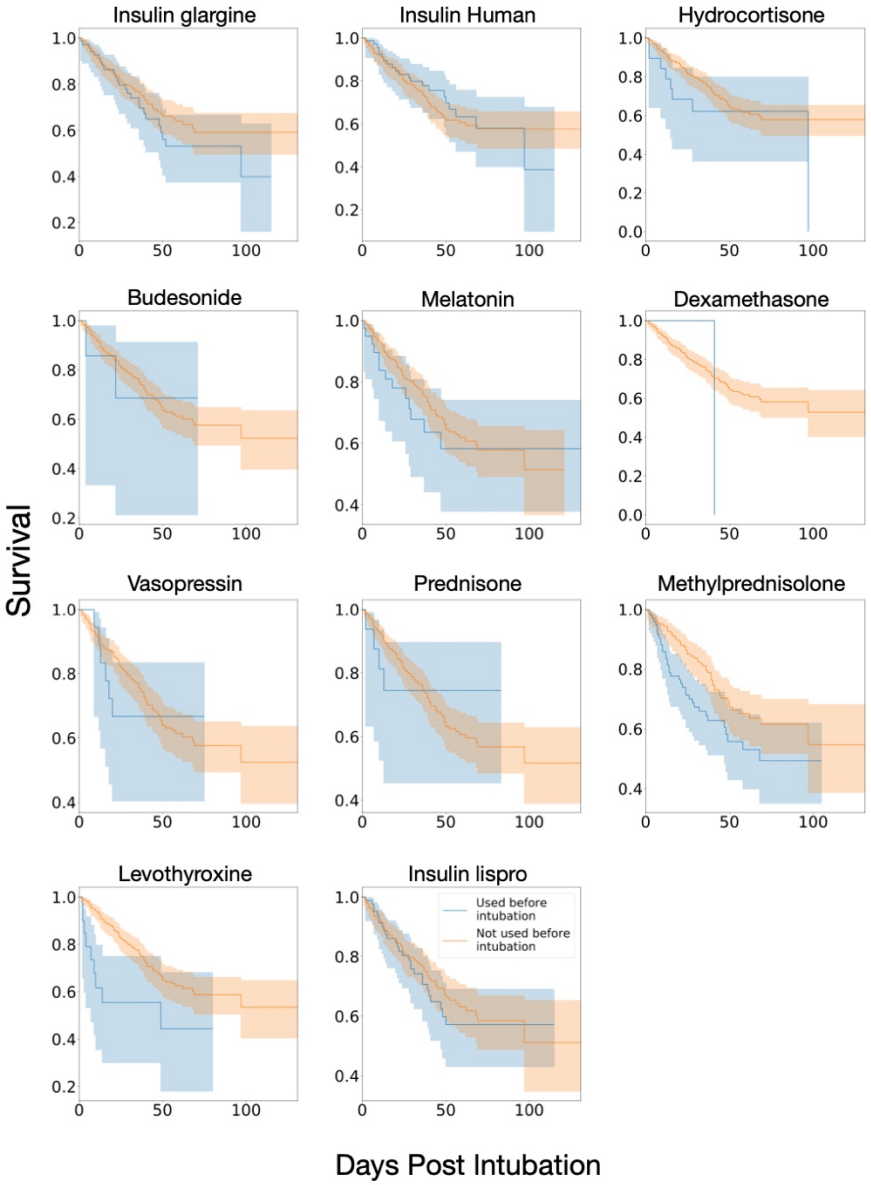


Figure 3.5: Kaplan-Meier curves for hormones exposure prior to oxygen therapy period for COVID-19 patients.

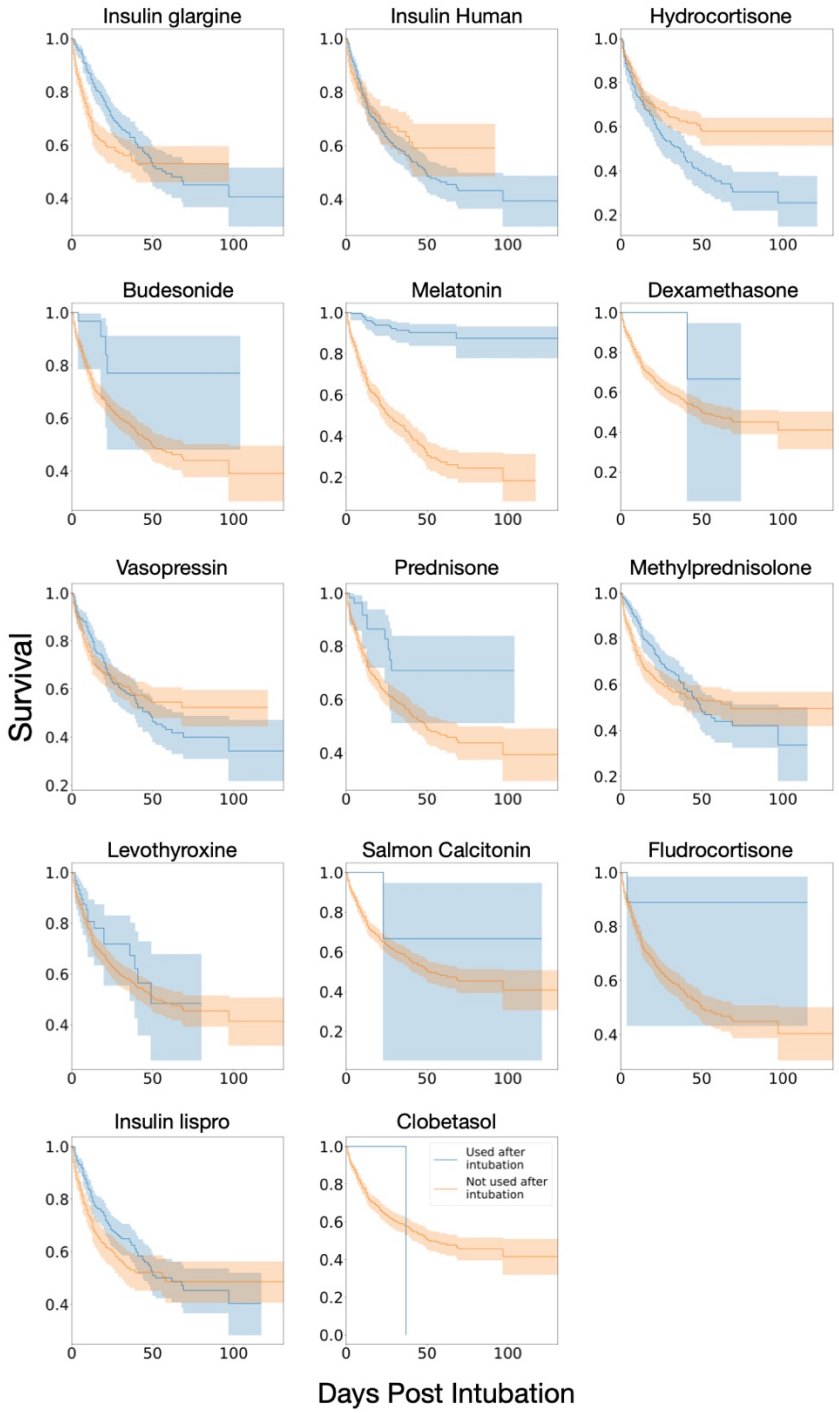


Figure 3.6: Kaplan-Meier curves for hormones exposure during oxygen therapy period for COVID-19 patients.

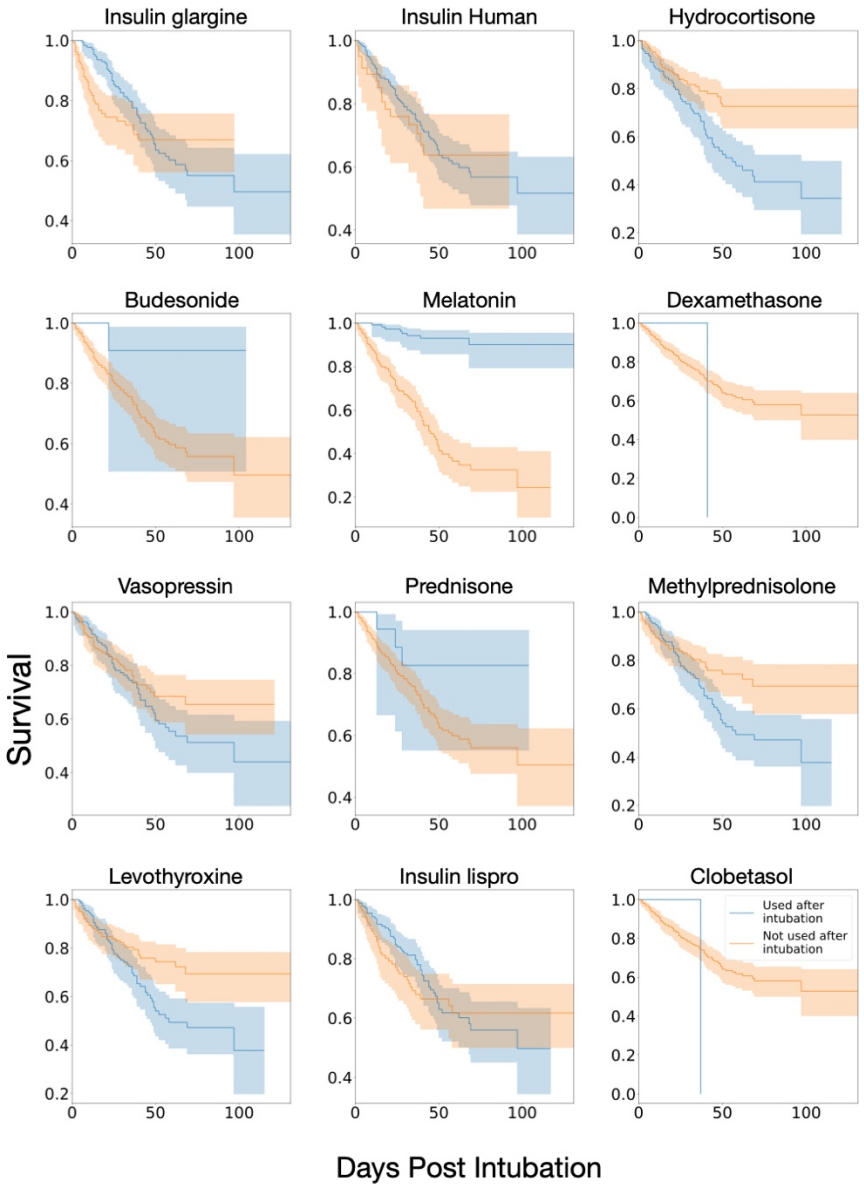


Figure 3.7: Kaplan-Meier curves for hormones exposure during oxygen therapy for COVID-19 oxygen therapy periods requiring mechanical ventilation.

3.3.7 Dexamethasone treatment after oxygen therapy is associated with increased survival among intubated COVID-19 patients

Based on the results of the study from the RECOVERY Collaborative Group in the UK, we were interested in the effects of dexamethasone treatment when accounting for other significant covariates, following the univariate analysis. We fit a Cox proportional hazards model using age (binary for whether the patient was older than 65), whether or not the patient self-identified as American Indian or Alaskan Native, whether or not the patient had a history of chronic kidney disease, chronic obstructive lung disease, coronary artery disease, diabetes mellitus, hypertension, myocardial infarction and respiratory disease, and whether or not a patient was exposed to dexamethasone after being intubated. Being at least 65 years old (HR: 2.45, 95% CI: 1.88 - 3.20, p-value = 4.14E-11) and having a history of chronic kidney disease (HR: 4.05, 95% CI: 2.28 - 7.18, p-value = 1.82E-06) and a history of chronic obstructive lung disease (HR: 1.48, 95% CI: 1.01 - 2.17, p-value = 4.25E-02) were significantly associated with a negative outcome during oxygen therapy in periods for COVID-19 patients (Table 3.9). Having a history of respiratory disease (HR: 0.470, 95% CI: 0.351 - 0.631, p-value = 4.71E-07) was significantly associated with a positive outcome in oxygen therapy periods for COVID-19 patients (Table 3.5). While not significant in the model, exposure to dexamethasone during oxygen therapy period was associated with a positive outcome (HR: 0.235, 95% CI: 3.25E-02 - 1.69, p-value = 0.151) in oxygen therapy for COVID-19 patients (Table 3.9).

In similar model, being at least 65 years old (HR: 1.93, 95% CI: 1.3 4- 2.78, p-value = 3.84E-04), having a history of chronic kidney disease (HR: 4.45, 95% CI: 2.80 - 7.09, p-value = 3.12E-10) and a history of myocardial infarction (HR: 2.55, 95% CI: 1.59 - 4.09, p-value = 1.04E-04) were significantly associated with a negative outcome in oxygen therapy periods for

non- COVID-19 patients (Table 3.9). Being at least 65 years old (HR: 2.48, 95% CI: 1.71 - 3.58, p- value = 1.41E-06) and having a history of chronic kidney disease (HR: 1.44, 95% CI: 1.03 - 2.01, p-value = 3.16E-02) were significantly associated with a negative outcome in oxygen therapy periods from 2018 (Table 3.9).

In a follow up analysis of the oxygen therapy periods where the patient required mechanical ventilation, we fit a Cox proportional hazards model using age (binary for whether the patients was older than 65), whether or not the patient self-identified as Black or African American, whether or not the patient had a history of asthma, chronic kidney disease, chronic obstructive lung disease and respiratory disease and whether or not a patients was exposed to dexamethasone during oxygen therapy. Being at least 65 years old (HR: 2.29, 95% CI: 1.46 - 3.59, p- value = 3.00E-04) was significantly associated with a negative outcome in oxygen therapy periods for COVID-19 patients (Table 3.10). Self-identifying as Black or African American (HR: 0.384, 95% CI: 0.185 - 0.797, p-value = 1.02E-02) and having a history of respiratory disease (HR: 0.534, 95% CI: 0.314 - 0.910, p-value = 2.11E-02) were significantly associated with a positive outcome in oxygen therapy periods for COVID-19 patients (Table 3.10).

In a similar model, being at least 65 years old (HR: 3.11, 95% CI: 1.47 - 6.55, p-value = 2.91E-03) was associated with a negative outcome in oxygen therapy periods for non-COVID-19 patients (Table 3.10). Being at least 65 years old (HR: 2.77, 95% CI: 1.90 - 4.03, p-value = 1.11E-07) was associated with a negative outcome in oxygen therapy periods from 2018 (Table 3.10).

Table 3.9: Dexamethasone exposure during oxygen therapy multivariate model Cox proportional hazards ratios. (± Not Determined)

Covariates	COVID-19 +		COVID-19 -		2018	
	N	Hazards Ratio (95 % CI) p-value	N	Hazards Ratio (95 % CI) p-value	N	Hazards Ratio (95 % CI) p-value
Age >= 65	409	2.45 (1.88-3.20) *p* = 4.14E-11	1125	1.93 (1.34-2.78) 3.84E-04	315	2.48 (1.71-3.58) *p* = 1.41E-06
Dexamethasone used during period	18	0.235 (3.25E-0.2-1.69) *p* = 0.151	206	0.566 (0.247-1.30) *p* = 0.178	81	0.862 (0.514-1.44) *p* = 0.572
American Indian or Alaskan Native	3	3.02 (0.738-12.4) *p* = 0.124		±		±
Chronic kidney disease	671	4.05 (2.28-7.18) *p* = 1.82E-06	1303	4.45 (2.80-7.09) *p* = 3.12E-10	259	1.44 (1.03-2.01) *p* = 3.16E-02
Chronic obstructive lung disease	85	1.48 (1.01-2.17) *p* = 4.25E-02	244	0.786 (0.460-1.34) *p* = 0.379	187	1.07 (0.757-1.52) *p* = 0.698
Coronary artery disease	145	1.24 (0.862-1.78) *p* = 0.247	539	0.655 (0.407-1.05) *p* = 8.10E-02	265	0.790 (0.544-1.15) *p* = 0.217
Diabetes mellitus	408	1.28 (1.00-1.66) *p* = 5.46E-02	722	1.05 (0.701-1.58) *p* = 0.808	283	1.36 (0.963-1.91) *p* = 8.06E-02

Table 3.9: Dexamethasone exposure during oxygen therapy multivariate model Cox proportional hazards ratios. (± Not Determined) (cont.)

Covariates	COVID-19 +		COVID-19 -		2018	
	N	Hazards Ratio (95 % CI) p-value	N	Hazards Ratio (95 % CI) p-value	N	Hazards Ratio (95 % CI) p-value
Hypertension	415	1.03 (0.786 - 1.35) $p = 0.832$	1181	0.952 (0.634-1.43) $p = 0.814$	471	0.815 (0.564-1.18) $p = 0.276$
Myocardial infarction	80	0.969 (0.628-1.50) $p = 0.887$	216	2.55 (1.59-4.09) $p = 1.04E-04$	160	1.39 (0.955-2.04) $p = 8.53E-02$
Respiratory Disease	706	0.470 (0.351-0.631) $p = 4.71E-07$	1819	1.11 (0.750-1.64) $p = 0.606$		±

Table 3.10: Dexamethasone exposure during oxygen therapy multivariate Cox proportional hazards ratios requiring mechanical ventilation. (± Not Determined)

Covariates	COVID-19 +		COVID-19 -		2018	
	N	Hazards Ratio (95 % CI) p-value	N	Hazards Ratio (95 % CI) p-value	N	Hazards Ratio (95 % CI) p-value
Age >= 65	144	2.29 (1.46-3.59) $p = 3.00E-04$	90	3.11 (1.47-6.55) $p = 2.91E-03$	273	2.77 (1.90-4.03) $p = 1.11E-07$
Dexamethasone used during period	4	0.883 (0.121-6.44) $p = 0.903$	10	0.525 (0.124-2.23) $p = 0.383$	67	0.857 (0.494-1.49) $p = 0.585$
Black or African American	62	0.384 (0.185-0.797) $p = 1.02E-02$	30	1.03 (0.393-2.68) $p = 0.956$	96	1.00 (0.639-1.55) $p = 0.985$
Asthma	32	0.347 (0.109-1.11) $p = 7.36E-02$	15	1.20 (0.267-5.42) $p = 0.809$	132	0.886 (0.566-1.39) $p = 0.594$
Chronic kidney disease	270	2.71 (0.977-7.52) $p = 5.56E-02$	127	1.28 (0.598-2.73) $p = 0.526$	223	1.40 (0.994-1.98) $p = 5.39E-02$
Chronic obstructive lung disease	20	1.96 (0.994-3.85) $p = 5.22E-02$	18	0.923 (0.314-2.72) $p = 0.885$	163	1.06 (0.731-1.54) $p = 0.752$
Respiratory Disease	275	0.534 (0.314-0.910) $p = 2.11E-02$	174	0.754 (0.358-1.59) $p = 0.457$		±

3.3.8 Univariate analysis of melatonin, quetiapine, trazodone and benzodiazepines on outcome following oxygen therapy

Following the significant association between melatonin exposure following oxygen therapy and a positive outcome in periods and in oxygen therapy periods requiring mechanical ventilation for COVID-19 patients, we conducted a univariate analysis of exposure to quetiapine, trazodone and benzodiazepines in COVID-19 and non-COVID-19 patients. Exposure to quetiapine (HR: 0.536, 95% CI: 0.293 - 0.980, p-value = 4.28E-02) and benzodiazepines (HR: 0.477, 95% CI: 0.330 - 0.690, p-value = 8.31E-05) between the visit start day and the start of oxygen therapy were significantly associated with a positive outcome following oxygen therapy in periods for COVID-19 patients (Table 3.11, Figure 3.8). Exposure to quetiapine (HR: 0.242, 95% CI: 0.178 - 0.329, p-value = 1.60E-19), trazodone (HR: 8.31E-02, 95% CI: 1.17E-02 - 0.593, p-value = 1.13E-02) and benzodiazepines (HR: 0.418. 95% CI: 0.318 - 0.550, p-value = 4.18E-10) during oxygen therapy were significantly associated with a positive outcome in oxygen therapy periods for COVID-19 patients (Table 3.12, Figure 3.8). Among oxygen therapy periods for non-COVID-19 patients, exposure to benzodiazepines (HR: 0.322. 95% CI: 0.232 - 0.447, p-value = 1.48E-11) during oxygen therapy were significantly associated with a positive outcome for oxygen therapy for non-COVID-19 patients (Table 3.12, Figure 3.8). Exposure to quetiapine (HR: 0.471, 95% CI: 0.328 - 0.676, p-value = 4.46E-05) was significantly associated with a positive outcome in oxygen therapy periods from 2018 (Table 3.12).

Table 3.11: Insomnia and agitation medications (melatonin, quetiapine, trazodone and benzodiazepines) univariate Cox proportional hazards ratios for oxygen therapy periods before the start of the period

Drug	COVID-19 +		COVID-19 -		2018	
	N	Hazards Ratio (95 % CI) p-value	N	Hazards Ratio (95 % CI) p-value	N	Hazards Ratio (95 % CI) p-value
Melatonin	86	0.635 (0.403-1.00) p = 5.10E-02	172	2.20 (1.41-3.42) p = 4.83E-04	93	1.25 (0.841-1.85) p = 0.273
Quetiapine	52	0.536 (0.293-0.980) p = 4.28E-02	53	1.50 (0.699-3.22) p = 0.298	77	1.12 (0.713-1.75) p = 0.632
Trazodone	13	0.771 (0.287-2.07) p = 0.606	28	1.79 (0.570-5.64) p = 0.318	27	0.790 (0.349-1.79) p = 0.571
Benzodiazepines	174	0.477 (0.330-0.690) p = 8.31E-05	544	1.27 (0.888-1.82) p = 0.190	518	1.20 (0.853-1.68) p = 0.300

Table 3.12: Insomnia and agitation medications (melatonin, quetiapine, trazodone and benzodiazepines) univariate Cox proportional hazards ratios for oxygen therapy periods during oxygen therapy

Drug	COVID-19 +		COVID-19 -		2018	
	N	Hazards Ratio (95 % CI) p-value	N	Hazards Ratio (95 % CI) p-value	N	Hazards Ratio (95 % CI) p-value
Melatonin	196	9.17E-02 (5.43E-02-0.155) $p = 4.81E-19$	319	0.322 (0.169-0.615) $p = 5.94E-04$	187	0.407 (0.269-0.614) $p = 1.81E-05$
Quetiapine	252	0.242 (0.178-0.329) $p = 1.60E-19$	119	0.838 (0.450-1.56) $p = 0.577$	217	0.471 (0.328-0.676) $p = 4.46E-05$
Trazodone	20	8.31E-02 (1.17E-02-0.593) $p = 1.13E-02$	55	0.697 (0.222-2.19) $p = 0.537$	49	0.527 (0.268-1.03) $p = 6.22E-02$
Benzodiazepines	729	0.418 (0.318-0.550) $p = 4.18E-10$	2484	0.322 (0.232-0.447) $p = 1.48E-11$	611	1.02 (0.672-1.55) $p = 0.926$

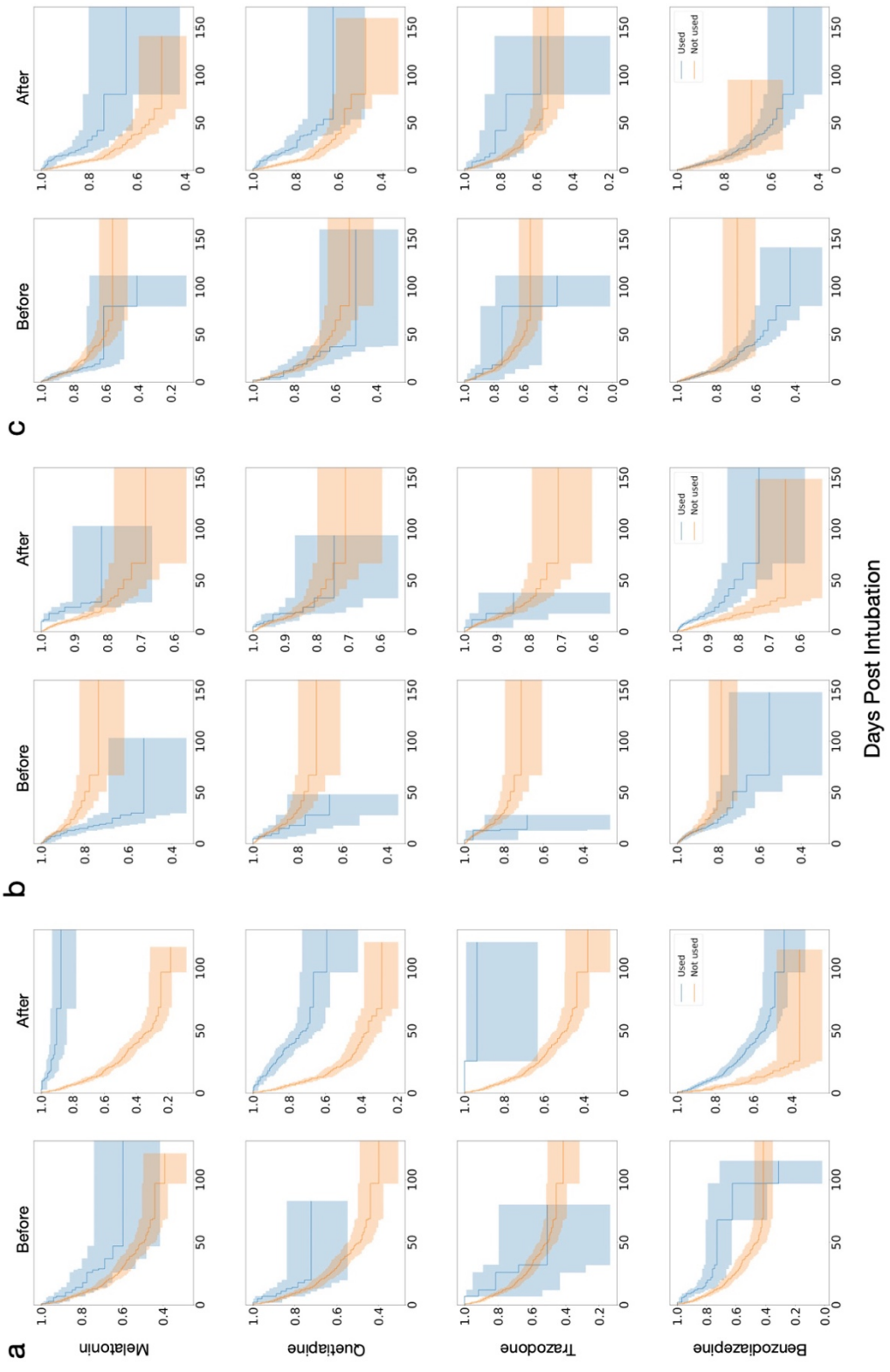


Figure 3.8: Kaplan-Meier curves for melatonin, quetiapine, trazodone and benzodiazepines treatment before and after oxygen therapy for (a) COVID-19 periods, (b) non-COVID-19 periods and (c) periods from 2018.

3.3.9 Melatonin treatment is associated with increased survival among patients receiving oxygen therapy

In order to further understand the association between melatonin during oxygen therapy and survival, we were interested in the effect of melatonin exposure when accounting for other factors significant covariates. We fit a Cox proportional hazards model using age (binary for whether the patients was older than 65), whether or not the patient self-identified as American Indian or Alaskan Native, whether or not the patient had a history of chronic kidney disease, chronic obstructive lung disease, coronary artery disease, diabetes mellitus, hypertension, myocardial infarction and respiratory disease, and whether or not a patient was treated with melatonin, quetiapine, trazodone and benzodiazepines after being intubated. Being at least 65 years old (HR: 1.78, 95% CI: 1.36 - 2.32, p-value = 2.35E-05), having a history of chronic kidney disease (HR: 6.32, 95% CI: 3.54 - 10.3, p-value = 4.32E-10) were significantly associated with a negative outcome in oxygen periods for COVID-19 patients (Table 3.13). Having a history of respiratory disease (HR: 0.493, 95% CI: 0.367 - 0.663, p-value = 2.78E-06) and exposure to quetiapine during oxygen therapy (HR: 0.289, 95% CI: 0.210 - 0.398, p-value = 2.37E-14), benzodiazepines after during oxygen therapy (HR: 0.585, 95% CI: 0.439 - 0.78, p-value = 2.30E-04) and melatonin after during oxygen therapy (HR: 0.131, 95% CI: 7.76E-02 - 0.223, p-value = 8.19E-14) are significantly associated with a positive outcome in during oxygen therapy periods for COVID-19 patients (Table 3.13).

Having a history of chronic kidney disease (HR: 5.14, 95% CI: 3.18 - 8.29, p-value = 2.08E-11) and myocardial infarction (HR: 3.22, 95% CI: 1.95 - 5.33, p-value = 5.13E-06) were significantly associated with a negative outcome following oxygen therapy (Table 3.9). Exposure to benzodiazepines during oxygen therapy (HR: 0.358, 95% CI: 0.250 - 0.513, p-value = 1.95E-

08) and exposure to melatonin during oxygen therapy (HR: 0.278, 95% CI: 0.142 - 0.542, p-value = 1.72E-04) were significantly associated with a positive outcome for non-COVID-19 patients (Table 3.13). Being at least 65 years old (HR: 2.55, 95% CI: 1.77 - 3.67, p-value = 4.84E-07) was significantly associated with a negative outcome in oxygen periods from 2018 (Table 3.9). Exposure to quetiapine during oxygen therapy (HR: 0.520, 95% CI: 0.358 - 0.756, p-value = 6.06E-04) and melatonin during oxygen therapy (HR: 0.423, 95% CI: 0.274 - 0.653, p-value = 1.03E-03) were significantly associated with a positive outcome in oxygen therapy periods from 2018 (Table 3.13).

Table 3.13: Melatonin exposure during oxygen therapy multivariate model Cox proportional hazards ratios for oxygen therapy periods. (± Not Determined)

Covariates	COVID-19 +		COVID-19 -		2018	
	N	Hazards Ratio (95 % CI) P-value	N	Hazards Ratio (95 % CI) P-value	N	Hazards Ratio (95 % CI) P-value
Age >= 65	409	1.78 (1.36-2.33) 2.35E-05	1125	1.31 (0.887-1.95) 0.173	315	2.55 (1.77-3.67) 4.84E-07
Melatonin exposure during oxygen therapy	196	0.131 (7.66E-02-0.223) 8.19E-14	319	0.278 (0.142-0.542) 1.72E-04	187	0.423 (0.274-0.653) 1.03E-04
American Indian or Alaskan Native	3	1.43 (0.346-5.93) 0.621		±		±
Chronic kidney disease	671	6.32 (3.54-11.3) 4.32E-10	1303	6.00 (3.67-9.81) 8.45E-13	259	1.40 (0.995-1.96) 5.38E-02
Chronic obstructive lung disease	85	1.35 (0.921-1.97) 0.125	244	0.765 (0.445-1.31) 0.332	187	1.10 (0.776-1.55) 0.603
Coronary artery disease	145	1.33 (0.937-1.88) 0.110	539	0.774 (0.473-1.27) 0.307	265	0.935 (0.640-1.37) 0.729
Diabetes mellitus	408	1.07 (0.830-1.38) 0.603	722	0.964 (0.638-1.46) 0.864	283	1.25 (0.880-1.78) 0.213
Hypertension	415	0.987 (0.754-1.29) 0.922	1181	1.04 (0.679-1.58) 0.867	471	1.00 (0.690-1.44) 0.991

Table 3.13: Melatonin exposure during oxygen therapy multivariate model Cox proportional hazards ratios for oxygen therapy periods. (± Not Determined) (cont.)

Covariates	COVID-19 +		COVID-19 -		2018	
	N	Hazards Ratio (95 % CI) P-value	N	Hazards Ratio (95 % CI) P-value	N	Hazards Ratio (95 % CI) P-value
Myocardial infarction	80	0.978 (0.645-1.48) 0.916	216	3.22 (1.95-5.33) 5.13E-06	160	1.33 (0.901-1.95) 0.152
Respiratory Disease	706	0.493 (0.367-0.663) 2.78E-06	1819	1.02 (0.685-1.51) 0.936		±
Quetiapine exposure during oxygen therapy	252	0.289 (0.210-0.398) 2.37E-14	119	0.557 (0.293-1.06) 7.54E-02	217	0.520 (0.358-0.756) 6.06E-04
Trazodone exposure during oxygen therapy	20	0.231 (3.20E-02-1.67) 0.146	55	0.772 (0.241-2.47) 0.662	49	0.763 (0.379-1.54) 0.451
Benzodiazepines exposure during oxygen therapy	729	0.585 (0.439-0.778) 2.3E-04	2484	0.358 (0.250-0.513) 1.95E-08	611	1.17 (0.762-1.79) 0.477

3.3.10 Melatonin treatment is associated with increased survival among COVID-19+ patients requiring mechanical ventilation

In a follow up analysis of the oxygen therapy periods where the patient required mechanical ventilation, we fit a Cox proportional hazards model using age (binary for whether the patients was older than 65), whether or not the patient self-identified as Black or African American, whether or not the patient had a history of asthma, chronic kidney disease, chronic obstructive lung disease and respiratory disease and whether or not a patients was treated with

melatonin, quetiapine, trazodone and benzodiazepines during oxygen therapy. Having a history of chronic kidney disease (HR: 3.00, 95% CI: 1.07 - 8.45, p-value = 3.71E-02) was significantly associated with a negative outcome in oxygen therapy periods for COVID-19 patients (Table 3.14). Self-identifying as Black or African American (HR: 0.403, 95% CI: 0.193 - 0.839, p-value - 1.15E-02), having a history of respiratory disease (HR: 0.433, 95% CI: 0.250-0.749, p-value = 2.80E-03) and exposure to quetiapine during oxygen therapy (HR: 0.404, 95% CI: 0.262 - 0.624, p-value = 4.31E-05), benzodiazepines during oxygen therapy (HR: 0.329, 95% CI: 0.187 - 0.580, p-value = 1.19E-04) and melatonin during oxygen therapy (HR: 0.127, 95% CI: 6.01E-02 - 0.269, p-value = 7.15E-08) after oxygen therapy were significantly associated with a positive outcome for oxygen therapy periods for COVID-19 patients where mechanical ventilation was required (Table 3.14).

Being at least 65 years old (HR: 3.07, 95% CI: 1.39 - 6.78, p-value = 5.54E-03) was significantly associated with a negative outcome in oxygen therapy periods for non-COVID-19 patients where mechanical ventilation was required (Table 3.14). Being at least 65 years old (HR: 3.06, 95% CI: 2.10 - 4.45, p-value = 5.44E-09) and having a history of chronic kidney disease (HR: 1.44, 95% CI: 1.02 - 2.04, p-value = 3.86E-02) were significantly associated with a negative outcome in oxygen therapy periods from 2018 (Table 3.14). Exposure to quetiapine during oxygen therapy (HR: 0.482, 95% CI: 0.322 - 0.722, p-value = 4.04E-04) and melatonin during oxygen therapy (HR: 0.492, 95% CI: 0.315 - 0.768, p-value = 1.78E-03) were associated with a positive outcome in oxygen therapy periods from 2018 (Table 3.14).

Table 3.14: Melatonin exposure during oxygen therapy multivariate model Cox proportional hazards ratios for periods requiring mechanical ventilation.

Covariates	COVID-19 +		COVID-19 -		2018	
	N	Hazards Ratio (95 % CI) P-value	N	Hazards Ratio (95 % CI) P-value	N	Hazards Ratio (95 % CI) P-value
Age >= 65	144	1.51 (0.947-2.40) 8.33E-02	90	3.07 (1.39-6.78) 5.54E-03	273	3.06 (2.10-4.45) 5.44E-09
Melatonin exposure during oxygen therapy	112	0.127 (6.01E-02-0.269) 7.15E-08	34	0.689 (0.275-1.72) 0.425	152	0.492 (0.315-0.768) 1.78E-03
Black or African American	62	0.403 (0.193-0.839) 1.15E-02	30	0.988 (0.373-2.61) 0.980	96	0.915 (0.587-1.43) 0.696
Asthma	32	0.452 (0.138-1.38) 0.190	15	1.16 (0.255-5.25) 0.849	132	0.828 (0.526-1.30) 0.415
Chronic kidney disease	270	3.14 (1.12-8.80) 2.98E-02	127	1.55 (0.693-3.48) 0.285	223	1.44 (1.02-2.04) 3.86E-02
Chronic obstructive lung disease	20	1.84 (0.923-3.66) 8.32E-02	18	1.17 (0.386-3.53) 0.785	163	1.20 (0.816-1.75) 0.360
Respiratory Disease	275	0.433 (0.250-0.749) 2.80E-03	174	0.891 (0.414-1.92) 0.768		±
Quetiapine exposure during oxygen therapy	175	0.404 (0.262-0.624) 4.31E-05	39	0.412 (0.156-1.09) 7.42E-02	181	0.482 (0.322-0.722) 4.04E-04

Table 3. 14: Melatonin exposure during oxygen therapy multivariate model Cox proportional hazards ratios for periods requiring mechanical ventilation. (cont.)

Covariates	COVID-19 +		COVID-19 -		2018	
	N	Hazards Ratio (95 % CI) P-value	N	Hazards Ratio (95 % CI) P-value	N	Hazards Ratio (95 % CI) P-value
Trazodone exposure during oxygen therapy	14	0.334 (4.48E-02-2.50) 0.286	9	0.766 (0.163-3.59) 0.735	39	0.871 (0.414-1.83) 0.716
Benzodiazepines exposure during oxygen therapy	272	0.329 (0.187-0.580) 1.19E-04	163	0.754 (0.383-1.49) 0.415	519	1.28 (0.809-2.02) 0.291

3.3.11 Chart review of COVID-19 patients treated with melatonin

Following the consistent significant associations between melatonin exposure following the start of oxygen therapy and a positive outcome in oxygen therapy periods and oxygen therapy periods requiring mechanical ventilation for COVID-19 patients, we were interested in the clinical nature of the melatonin prescription. We conducted a manual chart review of 50 randomly identified intubated COVID-19 patients to identify the justification, if any, for melatonin treatment. Of the 34 patients with justifications accompanying melatonin prescription, 21 patients' charts referenced insomnia, sleep wake cycle or difficulty sleeping for melatonin being prescribed and 18 patients' charts referenced anxiety, delirium, agitation or agitation delirium (Table 3.15). Additionally, five patients' charts referenced sedation, three patients' charts referenced derangement, altered mental status or mood, and 1 patient's chart referenced each difficulty waning sedation, adjuvant for presentation of respiratory disorder and pain (Table 3.15).

Table 3.15: Frequency of terms associated with melatonin treatment.

Term	N	% (of 34)
Derangement, Altered mental status, Mood	3	8.82
Insomnia, Sleep wake cycle, Difficulty sleeping	21	61.8
Anxiety, Delirium, Agitation, Agitated delirium	18	52.9
Difficulty weaning sedation	1	2.94
Sedation	5	14.7
Adjuvant for presentation of respiratory disorder	1	2.94
Pain	1	2.94

3.4 Discussion

In our retrospective analysis of patients who sought care at NYP/CUIMC between February 1st, 2020 and August 1st, 2020, we investigated the effects of hormone exposure in patients requiring oxygen therapy on mortality. For the 948 oxygen therapy periods among 791 patients who were diagnosed with COVID-19 or infected with SARS-CoV-2, there was exposure data for 14 hormone drugs prior to oxygen therapy: insulin glargine, insulin human, dronabinol, hydrocortisone, triamcinolone, budesonide, melatonin, dexamethasone, vasopressin, prednisone, methylprednisolone, levothyroxine, fludrocortisone and insulin lispro (Table 3.5). Additionally, there was exposure data for 17 hormone drugs during oxygen therapy: insulin glargine, insulin human, dronabinol, hydrocortisone, triamcinolone, budesonide, melatonin, dexamethasone, vasopressin, prednisone, methylprednisolone, levothyroxine, salmon calcitonin, fludrocortisone, insulin lispro, desmopressin and clobetasol (Table 3.6).

Univariate survival analysis identified exposure to insulin glargine, budesonide, melatonin, prednisone, methylprednisolone and insulin lispro during oxygen therapy as significantly associated with a positive outcome after oxygen therapy and exposure to hydrocortisone during oxygen therapy is significantly associated with a negative outcome following oxygen therapy in periods for COVID-19 patients (Table 3.7). Additionally, exposure to methylprednisolone and levothyroxine before oxygen therapy are significantly associated with a negative outcome following oxygen therapy in periods requiring mechanical ventilation for COVID-19 patients (Table 3.8). Exposure to hydrocortisone, methylprednisolone and levothyroxine during oxygen therapy are significantly associated with a negative outcome following oxygen therapy in periods requiring mechanical ventilation for COVID-19 patients and exposure to melatonin during oxygen therapy is significantly associated with a positive outcome (Table 3.7).

Univariate survival analysis also identified age (as a continuous variable and binary variable), self-identifying as American Indian or Alaskan Native and a history of chronic kidney disease, chronic obstructive lung disease, coronary artery disease, diabetes mellitus, hypertension and myocardial infarction as significantly associated with a negative outcome following oxygen therapy in periods for COVID-19 patients (Table 3.8). A history of respiratory disease was significantly associated with a positive outcome following oxygen therapy (Table 3.8). Age (as a continuous variable and binary variable) and a history of chronic kidney disease and chronic obstructive lung disease are significantly associated with a negative outcome following oxygen therapy in periods requiring mechanical ventilation for COVID-19 patients (Table 3.8). Having a history of asthma and respiratory disease are significantly associated with a positive outcome following oxygen therapy (Table 3.8).

Treatment with dexamethasone following intubation was associated, though not significantly, with a positive outcome in our univariate analysis of oxygen therapy periods for COVID-19 patients (Table 3.7). Furthermore, the association, though not significant, was observed in oxygen therapy periods for COVID-19 and non-COVID-19 patients when accounting for other covariates (Table 3.9). However, our analysis did not indicate an association between dexamethasone treatment during oxygen therapy and a positive outcome in periods requiring mechanical ventilation unlike the observation from the RECOVERY Collaborative Group's study [69] (Table 3.10). The power of our analysis is most likely limited due to the small sample size (N=4)

Moreover, our results identify exposure to melatonin as significantly associated with a positive outcome after oxygen therapy in univariate analyses of periods for COVID-19 patients and periods where mechanical ventilation was required for COVID-19 patients (Table 3.7) concurring with previous studies on the attenuation of cardiovascular responses following anesthesia [71], duration of mechanical ventilation in hemorrhagic stroke patients [72], and identification of melatonin acting as a regulator of inflammation [73]. The significant association of exposure to melatonin during oxygen therapy with a positive outcome in a multivariate model of COVID-19 and non-COVID-19 patients suggests that melatonin exposure is not specifically attenuating inflammation due to SARS-CoV-2 infection (Table 3.13). However, the multivariate model focusing on oxygen therapy periods where mechanical ventilation was required indicated that exposure to melatonin was only associated with a positive outcome in COVID-19 patients suggesting that melatonin's mechanism of action in the most severe cases of COVID-19 may be targeted to SARS-CoV-2 induced inflammation (Table 3.14).

While steroid hormones can be substituted for each other with adjustments to dosage and length of treatment, the dataset used for this retrospective analysis had incomplete data for dosage and the amount prescribed. Furthermore, the steroid hormones in the analysis have different structures and, as a result different, non-target interacting partners. For example insulin lispro, insulin human, and insulin glargine are different preparations with synthetics molecules. The results of this study can be validated in a controlled clinical trial where the amount of melatonin or any of the other steroid hormones analyzed given to patients is systemically tracked and the amount in the blood before and after oxygen therapy is determined to identify whether or not certain therapeutics are only beneficial if administered. after the onset of oxygen therapy.

A manual chart review of a subset of intubated COVID-19 patients did not reveal any inflammation specific goals for the treatment (Table 3.15). While melatonin is a popular over-the- counter sleep aid, our results lend support to the need for further follow-up into the mechanism of action of how melatonin may attenuate inflammation and specifically more studies into the observed association in severely affected COVID-19 patients.

The analysis done in this study utilized data from patients over seven months during which treatment for COVID-19 changed because of clinical experience with treating the disease. While dexamethasone and other steroids are the first that are used to reduce inflammation, there are other drugs that have been developed that are not account for which may introduce additional bias. Furthermore, melatonin is readily available over the counter to the general public, however the cost associated with them do not necessarily make readily accessible to everyone. This study focuses on drugs administered during hospitalization, patients who had been using the drug of their own fruition may be more likely to ask for it. Covariates, such as socioeconomics and

quality of health insurance, all contribute to confounder biases, which were not explored in the study.

Chapter 4: Identifying effects of COVID-19

The work in this chapter is adapted in part from the following publication:

V. Ramlall, B. May and N. P. Tatonetti. "Using machine learning probabilities to identify COVID-19 effects". medRχiv, Jul 2022. DOI: 10.1101/2022.07.02.22277179

4.1 Introduction

The ongoing COVID-19 pandemic, caused by SARS-CoV2 infection, of which there have been over 500 million cases worldwide, has resulted in more than 6.2 million deaths worldwide [19]. In the more than 30 months since the first infection is purported to have occurred [74] and the 26 months since the start of the pandemic as declared by the World Health Organization [75], the full impact of SARS-CoV-2 and COVID-19 remains to be seen.

Research has been paramount in responding to the COVID-19 pandemic from identifying patients susceptible to infection and at risk for severe disease [49,76,77] to identifying beneficial treatments [69,78,79] and developing prophylactic measures [80,81,82]. While there have been investigations into the long-term effects of COVID-19 [83,84,85,86,87] continual retrospective analyses will be important to identify all the long-term effects and to understand the full scope of the impact of COVID-19.

The long-term effects of viral infections vary greatly. While some viruses, such as certain strains of the seasonal flu and the common cold, have no-to-little impact on the long term health of those who are infected, others can have profound long lasting effects [88,89]. Through long

term analysis, it was determined that varicella zoster, the virus that causes chicken pox, also causes shingles [90], a rash accompanied by pain, itching and tingling, in adults [91]. Retrospective analyses in patients infected with certain strains of human papilloma virus (HPV) have shown that there is an increased risk of developing anal, cervical [92,93], penile, vaginal and vulvar cancers [94]. More recently, researchers have identified that Epstein-Barr virus, which causes mononucleosis, also triggers multiple sclerosis [95, 96], a demyelinating disease affecting the central nervous system [97].

Much of the investigations into COVID-19, as well as varicella zoster, HPV, and Epstein-Barr virus infections, have utilized patients' data sourced from electronic health records (EHRs). While EHRs provide a vast amount of data, such as clinical diagnoses, measurements, and procedures, they were not designed with the intention of being used for research and are incomplete. Research into COVID-19 has been further complicated by the novelty of the disease - the ICD10 code for COVID-19 (U07.1) was not effective until October 2020 [98]. While the diagnosis code was indicated for COVID-19 as early as April 2020, it was not used for all COVID-19 patients nor universally adapted, which hindering differentiating COVID-19 patients from non-COVID-19 visits. To address this, we used a random forest classifier to assign a probability of a patient having had COVID-19 during each of their visits (Training Set AUROC = 0.9867, Training Set OOB AUROC = 0.8957, Evaluation Set AUROC = 0.8958).

Furthermore, we used these probabilities to identify conditions associated with a higher probability of the patient having had COVID-19 by comparing the distributions of COVID-19 probability of visits that were followed with the diagnosis of a conditions at 1 week, 2 weeks, 3 weeks, 4 weeks, 3 months, 6 months, 9 months and 1year using a Mann-Whitney U test.

4.2 Methods

4.2.1 Data Source

We collected data from New York-Presbyterian, Columbia University Irving Medical Center Weill-Cornell Medical Center between February 1st, 2020 and March 31st, 2022 for patients who had at least one interaction with Columbia University Irving Medical Center during that period. We sourced historical data from our clinical data warehouse available through December 31st, 2020.

4.2.2 Identifying visits

We used MySQL 5.7.35 and Python 3.9.10 with numpy 1.19.5, pymysql 1.0.2 and pandas 1.2.3 libraries to extract and prepare data. From admissions data, we identified patients who had a valid admission date, a valid admission discharge date and who were hospitalized on or after February 1st, 2020. To remove duplicate entries, incorrect discharge entries and control for admittance procedure (e.g. patients who seek treatment at the emergency department and are then admitted), we grouped entries by the patients' medical reference number and the admission date and used the latest discharge date for the patient and the admission date as the discharge date for our analysis. We identified 1,844,018 visits for 636,063 patients.

4.2.3 Collecting and processing demographic data

We used MySQL 5.7.35 and Python 3.9.10 with numpy 1.19.5, pymysql 1.0.2 and pandas 1.2.3 libraries to extract and prepare data. We identified sex, race (for which there are up to three entries), ethnicity and date of birth for the patient associated with each visit and excluded visits for patients who were did not have a valid date of birth. We identified 1,573,113 visits for 434,152 patients where the patient had a valid date of birth.

For each visit, we identified the age of the patient at the start of the visit (i.e. admission date) as (i) birth to 13 years old, (ii) 13 to 19 years old, (iii) 19 to 60 years old and (iv) over 60 years old and if the patents indicated their sex as female. For each visit, we identified whether the patient indicated their race(s) as (i) American Indian or Alaskan Native, (ii) Asian, (iii) Black or African American (iv) Native Hawaiian or Other Pacific Islander or (v) White, and whether the patient indicated their ethnicity as of Hispanic or Latino or Spanish Origin. All variables were treated as a binary categorical variables with 1 indicating that the patient was a part of the age group, or self-identified as female or self-identified as the specific race and 0 indicating the inverse. For example, a visit for 27 year old male patient who self identified as Asian and indicated that he was not of Hispanic or Latino or Spanish Origin is represented by [0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0].

Note: In our clinical data the possible values for sex are female, male, nonbinary, unknown, X and null. The prepared date indicated a 0 for these visits in the sex field, which is also used for male, as queried whether or not the patient for that visit identified as female. We identified 3,364 visits for 307 patients where the patient did not indicate male or female for their sex.

Note: Patients could indicate that they are of more than one racial background and our data was prepared such that it queried whether or not each of the five options were identified in any of the patients up to three entries. For example, a patient who self-identified as Asian is represented by [0, 0, 1, 0, 0] for the race component of the demographics matrix and a patient who self-identified as Asian and White is represented by [0, 0, 1, 0, 1].

4.2.4 Collecting and processing temporal data

We used MySQL 5.7.35 and Python 3.9.10 with numpy 1.19.5, pymysql 1.0.2 and pandas 1.2.3 libraries to extract and prepare data. We identified the start date for each visit and represented it as a categorical variable based on whether the visit started in any of the 26 months period from which the data is sourced. All variables were treated as binary categorical variables with 1 indicating that a visit started during a specific month and 0 indicating that it did not. For example, a visit that started on June 14th, 2020 is represented by:

[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0].

4.2.5 Collecting and processing diagnosis data

We used MySQL 5.7.35 and Python 3.9.10 with numpy 1.19.5, pymysql 1.0.2 and pandas 1.2.3 libraries to extract and prepare data. We identified 16,221 distinct ICD10 clinical diagnosis codes listed for patients in the 26 month period of the data sourced and mapped those diagnoses to specific visits using the start and end dates of the visit. For our modeling, we generalized the 16,221 ICD10 clinical diagnoses codes to 1,600 distinct category level diagnoses codes. All variables were treated as a binary categorical variables with 1 indicating that a specific diagnosis code was listed during that visit and 0 indicating that it was not. Our data was prepared such that it queried whether each of the 1,600 category level codes were present in the diagnosis codes indicated for each visit.

Note: U07.1, which was indicated for identifying COVID-19 diagnoses, was omitted prior and generalized the diagnoses codes to 1,600 distinct category levels codes.

Note: U07.0, which is in the same category as U07.1, and not omitted prior to diagnosis indicates vaping-related disorders.

4.2.6 Training and evaluating the random forest classifier

We used Python 3.9.10 with sklearn 0.24.2 and pickle libraries to fit, evaluate and apply a random forest model. For the training and evaluation sets, he input matrix comprised of the demographic, temporal and diagnosis data for each visit with the outcome indicated as 1 if the visit was in the COVID-19 training or evaluation set and the outcome indicated as 0 if the visit was in the non-COVID-19 training or evaluation set. An initial model of 200 estimators and out-of-bag samples to estimate the generalization score was trained and evaluated. The random forest classifier was refined by increasing the number of estimators and the maximum depth to maximize AUROC in the independent evaluation set. The final model was applied to the data for all of the visits.

4.2.7 Identifying previous clinical phenotypes

We used MySQL 5.7.35 and Python 3.9.10 with numpy 1.19.5, pymysql 1.0.2 and pandas 1.2.3 libraries to extract and prepare data. For the patients with historical data in our clinical data warehouse, we mapped SNOMED condition code to PheCodes via ICD10-codes. Additionally, from our current clinical data, we mapped ICD-10 diagnoses codes to PheCodes. For each visit we identified previous clinical phenotypes as those that were indicated prior to the start of the visit.

Note: For patients who had multiple visits, the previous clinical phenotypes of the second visit will account for any new clinical phenotypes identified in the first visit, and the previous clinical phenotypes of the third visit will account for any new clinical phenotypes identified in the first and second visits, and so on.

4.2.8 Identify clinical phenotypes

We used MySQL 5.7.35 and Python 3.9.10 with numpy 1.19.5, pymysql 1.0.2 and pandas 1.2.3 libraries to extract and prepare data. From our current clinical data, we mapped ICD-10 diagnoses codes for each visit to PheCodes to identify clinical phenotypes present at each visit.

4.2.9 Identifying clinical phenotypes that develop

We used Python 3.9.10 with numpy 1.19.5, pandas 1.2.3, and scipy 1.6.2 libraries to statistically evaluate the distributions. For visits with a follow up within each time interval (i.e. within 1 week, 2 weeks, 3 weeks, 4 weeks, 3 months, 6 months, 9 months and 1 year), we discerned the visits where the phenotype was observed in the followup and the visits where the phenotype was not observed and compared the distributions of COVID-19 probability of the initial visit using a Mann-Whitney U test.

Note: p-values of 0 are presented as $p < 2.225E\text{-}308$ (the minimum value for a float object in Python) in the manuscript and tables, while p-values of 0 are recast as half the minimum non-zero p-value per test for stylistic purposes in figures.

4.2.10 Identifying new clinical phenotypes that develop

We used Python 3.9.10 with numpy 1.19.5, pandas 1.2.3, and scipy 1.6.2 libraries to statistically evaluate the distributions. For visits with a follow up within each time interval (i.e. within 1 week, 2 weeks, 3 weeks, 4 weeks, 3 months, 6 months, 9 months and 1 year), we excluded all visits for which the patient already has the clinical phenotype. Then we discerned the visits where the phenotype was observed in the followup and the visits where the phenotype was not observed and compared the distributions of COVID-19 probability of the initial visit using a Mann-Whitney U test.

Note: p-values of 0 are presented as $p < 2.225E\text{-}308$ (the minimum value for a float object in Python) in the manuscript and tables, while p-values of 0 are recast as half the minimum non-zero p-value per test for stylistic purposes in figures.

In evaluating instances where the patient was not previously diagnosed with the condition, we eliminated all patients who had a previous history of the condition (i.e. had the diagnosis prior to the start of the visit).

4.2.11 Cox Proportional Hazards modeling and Kaplan-Meier curve fitting

From our cases visits (those visits where the patient returned with the condition within one year), we identified the time to event as the time from the end of the preceding visit to the first instance of the condition within one year of the visit. In our non-case visits, we censored the data at the final interaction with NYP/CUIMC within the time period. We used Python 3.9.10 with numpy 1.19.5, pandas 1.2.3, and lifelines 0.25.10 libraries to determine and statistically evaluate the hazards ratios associated with COVID-19 probability. In order to build Kaplan-Meier curves, we stratified our data by the COVID-19 probability of the preceding visit (≤ 0.2. > 0.2 and ≤ 0.4, > 0.4 and ≤ 0.6, > 0.6 and ≤ 0.8, and > 0.8) and fit individual curves to each stratified dataset.

4.3 Results

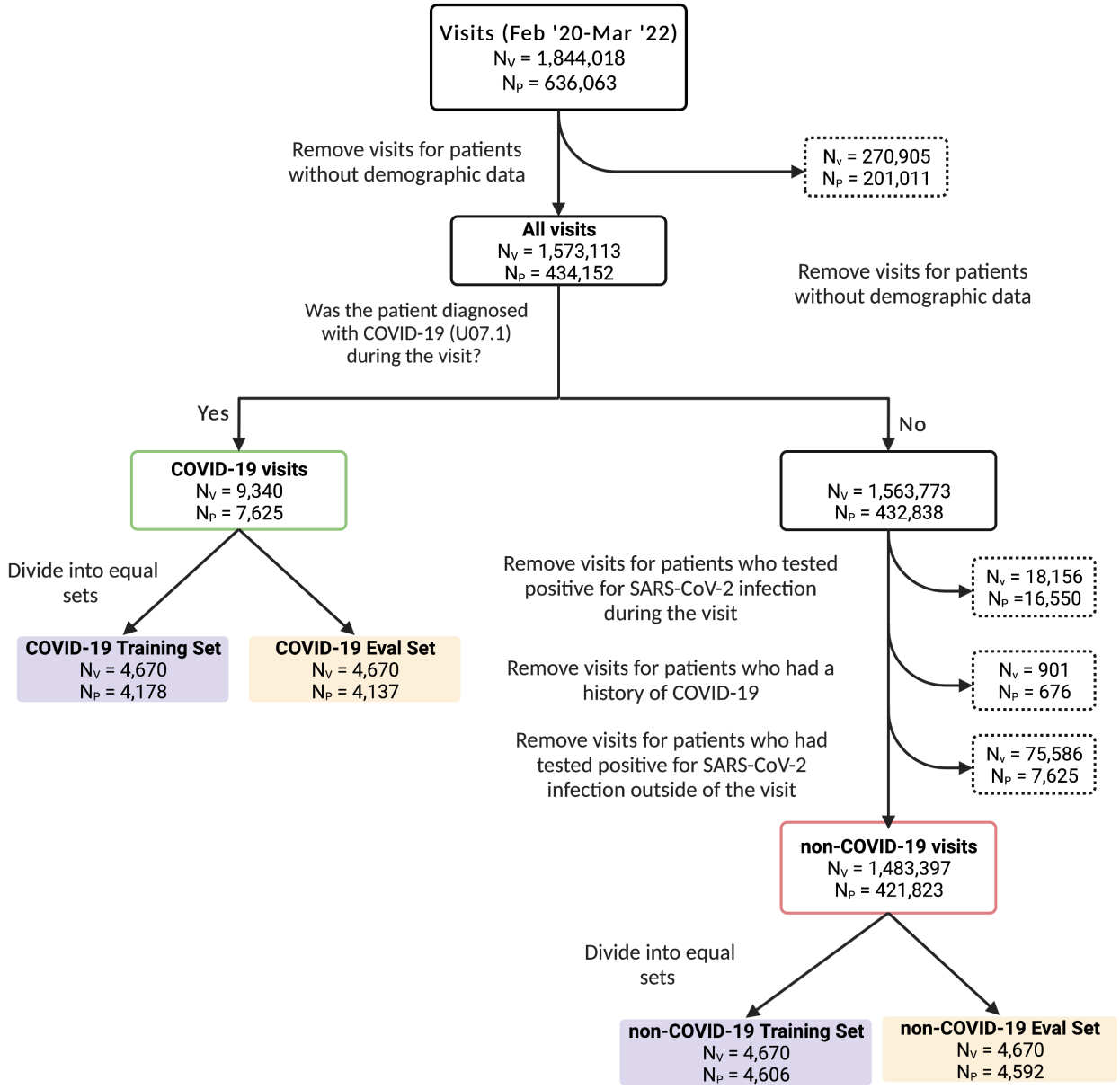


Figure 4.1 Data processing flowchart Identification of COVID-19 and non-COVID-19 training sets (purple) and evaluation sets (orange). N_v indicates the number of visits and N_P indicates the number of patients in each group. Note: the exclusion criteria used to identify non-COVID-19 visits are not mutually exclusive.

From the clinical data at New York-Presbyterian, we identified 1,844,018 visits for 636,063 patients who sought treatment at least once between February 1st, 2020 and March 31st,

2022 at /Columbia University Irving Medical Center (NYP/CUIMC). We omitted 270,905 visits for 201,911 patients who did not have any demographic data available in our clinical data set (Figure 4.1). From these visits, we identified 9,340 visits (COVID-19 visits) where the patient was diagnosed with COVID-19 evidenced by the presence of the COVID-19 ICD-10 diagnosis code (U07.1) (Figure 4.1). Additionally, we identified 1,483,397 visits (non-COVID-19 visits) where the patient did not test positive for SARS-CoV-2 during that visit nor had a history of COVID-19 nor previously tested positive for SARS-CoV-2 infection (Figure 4.1). The set of COVID-19 visit was randomly split into distinct testing and evaluation sets, each with 4,670 visits and from the set of non-COVID-19 visits, we randomly identified distinct testing and evaluation non-COVID-19 sets, each with 4,670 unique visits. Among all visits between February 2020 and March 2022, as well as the COVID-19 and non-COVID-19 training and evaluations sets, more than 50% of the visits were for patients who self-identified as female and more than 85% of the visits were for patients who were at least 19 years old (adults and senior age groups) (Table 4.1). Across all of the groups, more than 35% of the visits were for patients who self-identified as White, more than 15% were for patients who self-identified as Black or African American and more than 29% were visits for patients who self-identified as Hispanic or of Latino or Spanish origin (Table 4.1). In all groups, less than 5% of visits were for patients

Table 4.1 Demographics of patients of visits used for model training, model evaluation and all visits between February 2020 and March 2022

Demographic variable (% of visits)	Model Training Set		Model Evaluation Set		All Visits
	non-COVID-19	COVID-19	non-COVID-19	COVID-19	Feb 20 - Mar 22
N(visits)	4,670	4,670	4,670	4,670	1,573,113
N(patients)	4,606	4,178	4,592	4,137	434,152
Age Child (< 13)	365 7.82%	235 5.03%	326 6.98%	242 5.18%	122,140 7.76%
Age Adol (≥ 13 and < 19)	146 3.13%	115 2.46%	156 3.34%	115 2.46%	49,232 3.13%
Age Adult (≥ 19 and < 60)	2,375 50.86%	2,159 46.23%	2,324 49.76%	2,099 44.95%	772,606 49.11%
Age Senior (≥ 60)	1,784 38.20%	2,161 46.27%	1,864 39.91%	2,214 47.41%	629,135 39.99%
Self Identified Sex as Female	2,756 59.01%	2,412 51.65%	2,885 61.78%	2,403 51.46%	941,558 59.85%
Self Identified as American Indian or Alaskan Native	14 0.30%	18 0.39%	11 0.24%	< 10 < 0.41%	3,994 0.25%
Self Identified as Asian	108 2.31%	124 2.66%	120 2.57%	116 2.48%	39,091 2.48%
Self Identified as Black or African American	728 15.59%	803 17.19%	702 15.03%	811 17.37%	245,104 15.58%
Slf Identified as Native Hawaiian or Other Pacific Islander	< 10 < 0.21%	10 0.21%	< 10 < 0.21%	< 10 < 0.21%	1,520 0.10%

Table 4.1 Demographics of patients of visits used for model training, model evaluation and all visits between February 2020 and March 2022 (cont.)

Demographic variable (% of visits)	Model Training Set		Model Evaluation Set		All Visits
	non-COVID-19	COVID-19	non-COVID-19	COVID-19	Feb 20 - Mar 22
N(visits)	4,670	4,670	4,670	4,670	1,573,113
N(patients)	4,606	4,178	4,592	4,137	434,152
Self Identified as White	1,900 40.69%	1,645 35.22%	1,974 42.27%	1,669 35.74%	643,848 40.93%
Self identified as Hispanic or of Latino or Spanish Origin	1,433 30.69%	1,873 40.11%	1,367 29.27%	1,782 38.16%	473,501 30.10%

who self-identified as American Indian or Alaskan Native, Asian or Native Hawaiian or Other Pacific Islander (Table 4.1).

Among all visits between February 2020 and March 2022, the largest fraction of visits (5.17%) began in March 2021 (Table 4.2). The largest fraction of visits in the COVID-19 training and evaluation sets began in April 2020 (18.29% and 17.99%, respectively), while the smallest fraction of all visits began in April 2020 until March 2022 (1.64%) (Table 4.2). The fraction of visits in the non-COVID-19 training and evaluation sets that began in each month were similar to the faction of all visits that began in each month (Table 4.2).

Table 4.2 The month during which the visits used for model training, model evaluation and all visits began between February 2020 and March 2022

Visit start month (% of visits)	Model Training Set		Model Evaluation Set		All Visits
	non-COVID-19	COVID-19	non-COVID-19	COVID-19	Feb 20 - Mar 22
N(visits)	4,670	4,670	4,670	4,670	1,573,113
N(patients)	4,606	4,178	4,592	4,137	434,152
February 2020	159 3.40%	< 10 < 0.21%	162 3.47%	< 10 < 0.21%	58,627 3.73%
March 2020	151 3.23%	416 8.91%	148 3.17%	425 9.10%	50,069 3.18%
April 2020	75 1.61%	854 18.29%	60 1.28%	840 17.99%	25,857 1.64%
May 2020	117 2.51%	275 5.89%	106 2.27%	275 5.89%	37,202 2.36%
June 2020	176 3.77%	211 4.52%	187 4.00%	181 3.88%	59,374 3.77%
July 2020	191 4.09%	186 3.98%	187 4.00%	208 4.45%	65,739 4.18%
August 2020	185 3.96%	187 4.00%	189 4.05%	191 4.09%	65,999 4.20%
September 2020	200 4.28%	127 2.72%	206 4.41%	186 3.98%	67,833 4.31%
October 2020	242 5.18%	190 4.07%	228 4.88%	196 4.20%	79948 5.08%

Table 4.2 The month during which the visits used for model training, model evaluation and all visits began between February 2020 and March 2022 (cont.)

Visit start month (% of visits)	Model Training Set		Model Evaluation Set		All Visits
	non-COVID-19	COVID-19	non-COVID-19	COVID-19	Feb 20 - Mar 22
N(visits)	4,670	4,670	4,670	4,670	1,573,113
N(patients)	4,606	4,178	4,592	4,137	434,152
November 2020	220 4.71%	208 4.45%	216 4.63%	234 5.01%	74831 4.76%
December 2020	221 4.73%	339 7.26%	237 5.07%	297 6.36%	70,401 4.48%
January 2021	206 4.41%	357 7.64%	206 4.41%	366 7.84%	67,593 4.30%
February 2021	161 3.45%	300 6.42%	184 3.94%	301 6.45%	62,744 3.99%
March 2021	247 5.29%	295 6.32%	220 4.71%	289 6.19%	81,348 5.17%
April 2021	214 4.58%	135 2.89%	220 4.71%	123 2.63%	74,489 4.74%
May 2021	178 3.81%	41 0.88%	191 4.09%	44 0.94%	66,782 4.25%
June 2021	251 5.37%	21 0.45%	224 4.80%	30 0.64%	76,091 4.84%
July 2021	232 4.97%	21 0.45%	210 4.50%	27 0.58%	73,954 4.70%

Table 4.2 The month during which the visits used for model training, model evaluation and all visits began between February 2020 and March 2022 (cont.)

Visit start month (% of visits)	Model Training Set		Model Evaluation Set		All Visits
	non-COVID-19	COVID-19	non-COVID-19	COVID-19	Feb 20 - Mar 22
N(visits)	4,670	4,670	4,670	4,670	1,573,113
N(patients)	4,606	4,178	4,592	4,137	434,152
August 2021	229 4.90%	32 0.69%	227 4.86%	19 0.41%	75,266 4.78%
September 2021	227 4.86%	32 0.69%	226 4.84%	32 0.69%	71,980 4.58%
V October 2021	211 4.52%	27 0.58%	233 4.99%	24 0.51%	73,793 4.69%
November 2021	154 3.30%	41 0.88%	164 3.51%	56 1.20%	52,130 3.31%
December 2021	158 3.38%	117 2.51%	162 3.47%	103 2.21%	49,074 3.12%
January 2022	127 2.72%	168 3.60%	115 2.46%	159 3.40%	41,889 2.66%
February 2022	110 2.36%	73 1.56%	123 2.63%	55 1.18%	40,862 2.60%
March 2022	25 0.54%	10 0.21%	32 0.69%	< 10 < 0.21%	7,947 0.51%

Among all visits between February 2020 and March 2022, the four diagnoses listed in the most visits were encounter for supervision of normal pregnancy (2.38%), transplanted organ and tissue status (2.26%), other symptoms and signs involving the circulatory and respiratory system (2.18%) and essential (primary) hypertension (2.06%) (Table 4.3). Among the COVID-19 visits in the training and evaluation sets, diagnosis of other symptoms and signs involving the circulatory and respiratory system (20.75% and 19.21%, respectively), encounter for screening for malignant neoplasms (19.46% and 19.08%, respectively), essential (primary) hypertension (8.84% and 9.27%, respectively) and transplanted organ and tissue status (8.22% and 8.78%, respectively) were frequently diagnosed (Table 4.3). The fraction of non-COVID-19 visits in the training and evaluation sets with the diagnoses listed was similar to the fraction of all visits with the diagnosis listed (Table 4.3). A complete table of all diagnoses is available at https://github.com/vijendra-cuimc/thesis/blob/main/table_4.3_diag_formatted_git.csv.

Table 4.3 The ten most frequently observed ICD10 diagnoses for visits used for model training, model evaluation and all visits between February 2020 and March 2022

Visit start month (% of visits)	Model Training Set		Model Evaluation Set		All Visits
	non-COVID-19	COVID-19	non-COVID-19	COVID-19	Feb 20 - Mar 22
N(visits)	4,670	4,670	4,670	4,670	1,573,113
N(patients)	4,606	4,178	4,592	4,137	434,152
Encounter for supervision of normal pregnancy.	124 2.66%	118 2.53%	101 2.16%	123 2.63%	37,409 2.38%
Transplanted organ and tissue status	92 1.97%	384 8.22%	95 2.03%	410 8.78%	35,478 2.26%
Other symptoms and signs involving the circulatory and respiratory system	85 1.82%	969 20.75%	77 1.65%	897 19.21%	34,361 2.18%
Essential (Primary) Hypertension	97 2.08%	413 8.84%	79 1.69%	433 9.27%	32,379 2.06%
Other joint disorder, not elsewhere classified	99 2.12%	87 1.86%	86 1.84%	96 2.06%	29,838 1.9%
Abdominal and pelvic pain	79 1.69%	143 3.06%	77 1.65%	180 3.85%	27,083 1.72%
Encounter for screening for malignant neoplasms	69 1.48%	909 19.46%	73 1.56%	891 19.08%	26,545 1.69%
Supervision of high risk pregnancy	65 1.39%	21 0.45%	79 1.69%	26 0.56%	23,503 1.49%
Encounter for other special examination without	55 1.18%	71 1.52%	62 1.33%	48 1.03%	20995 1.33%

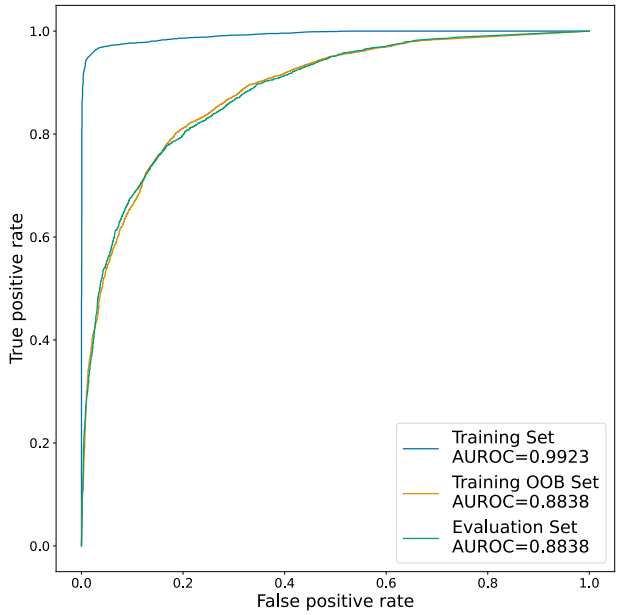


Figure 4.2 ROC curves of training set, training set using out-of-bag estimates, and evaluation set of the original model (n_estimators = 200, max_depth not constrained)

We collected demographic data for the patient in each visit (date of birth, self-identified sex, self-identified race(s) and self-identified ethnicity), temporal data (during what month the visit started) and visit specific diagnosis data. In our dataset, there were 16,220 distinct ICD10 codes used to records diagnoses which we generalized to 1,600 category level ICD10 codes. We decided to use a random forest classifier to predict whether or not a patient was diagnosed with COVID-19 during their visit using demographic, temporal, and visit-specific clinical diagnoses. The diagnosis code for COVID-19 (U07.1) was removed from the data to be used in the training the model prior to generalization. Instead of binary outcome (patient having been diagnosed with COVID-19 during their visit or not), we used the fraction of estimators identifying the visit as one where the patient was diagnosed with COVID-19 as the probability of the patient having COVID-19 during the visit. An initial random forest classifier of 200 estimators was fit using the COVID-19 and non-COVID-19 training sets with bootstrapped sampling and using out-of-bag

sampling (Training AUROC = 0.9923, Training OOB AUROC = 0.8838, Evaluation AUROC = 0.8838) (Figure 4.2).

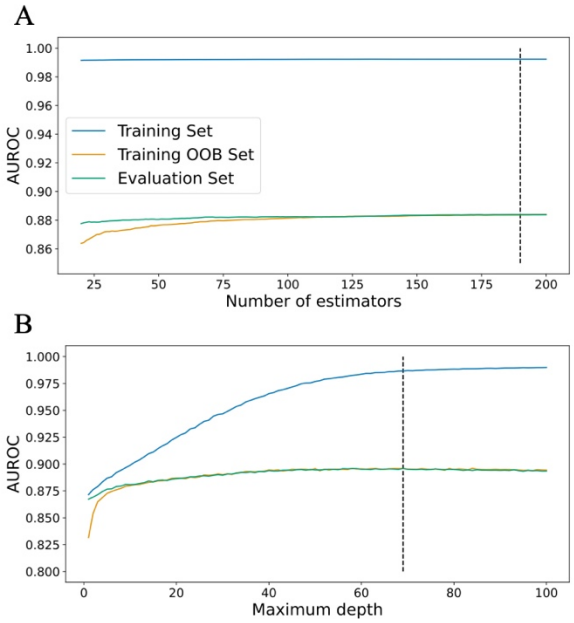


Figure 4.3 Model performance optimization (A) AUROC in training set, training set using out-of-bag estimates, and evaluation set plotted against number of estimators (dashed line indicates maximum AUROC in evaluation set, n_estimators = 190). (B) AUROC in training set, training set using out-of-bag estimates, and evaluation set plotted against maximum depth (dashed line indicates maximum AUROC in evaluation set, max_depth = 69).

In order to optimize the performance of the model, we monitored the AUROC of the training set, the training set using out-of-bag estimates and the evaluation set while increasing the number of estimators from 20 to 200 and achieved a maximum AUROC in the evaluation set with 190 estimators (Training Set AUROC = 0.9924, Training Set OOB AUROC = 0.8836, Evaluation Set AUROC = 0.8839) (Figure 4.3 A). We further optimized the performance of the model by monitoring the AUROC while increasing the maximum depth of the model from 1 to 100 with 190 estimators and achieved a maximum AUROC in the evaluation set with a depth of 69 (Training Set AUROC = 0.9867, Training Set OOB AUROC = 0.8957, Evaluation Set

AUROC = 0.8958) (Figure 4.3 B). The optimized model trained with 190 estimators with a maximum depth of 69 was fit to the data representing all 1,573,113 visits (Figure 4.4 A). The distribution of the COVID-19 and non-COVID-19 training sets are skewed to 1 and 0, respectively, with minor overlap between 0.3 and 0.5 (Figure 4.4 B). The COVID-19 and non-COVID-19 evaluation sets are similarly skewed, though with a wider overlap (Figure 4.4 C).

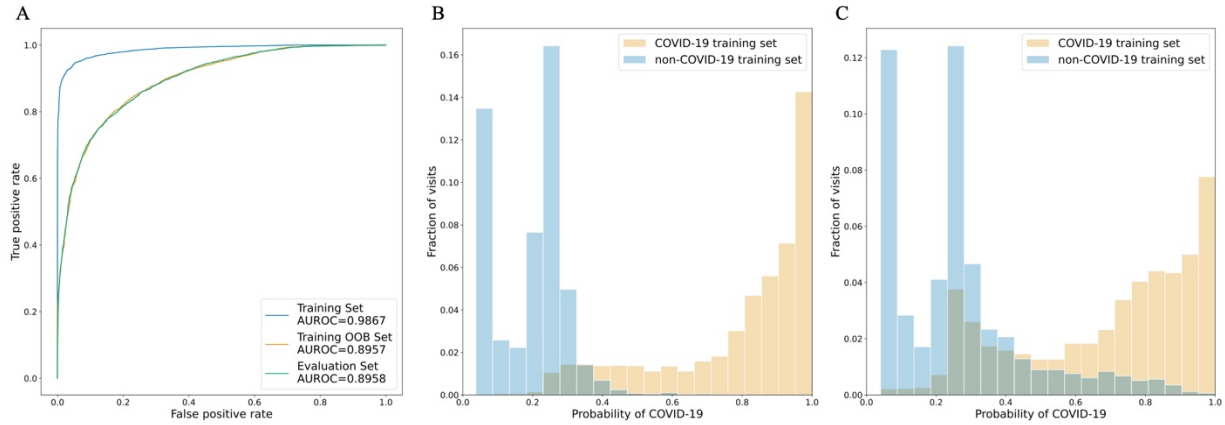
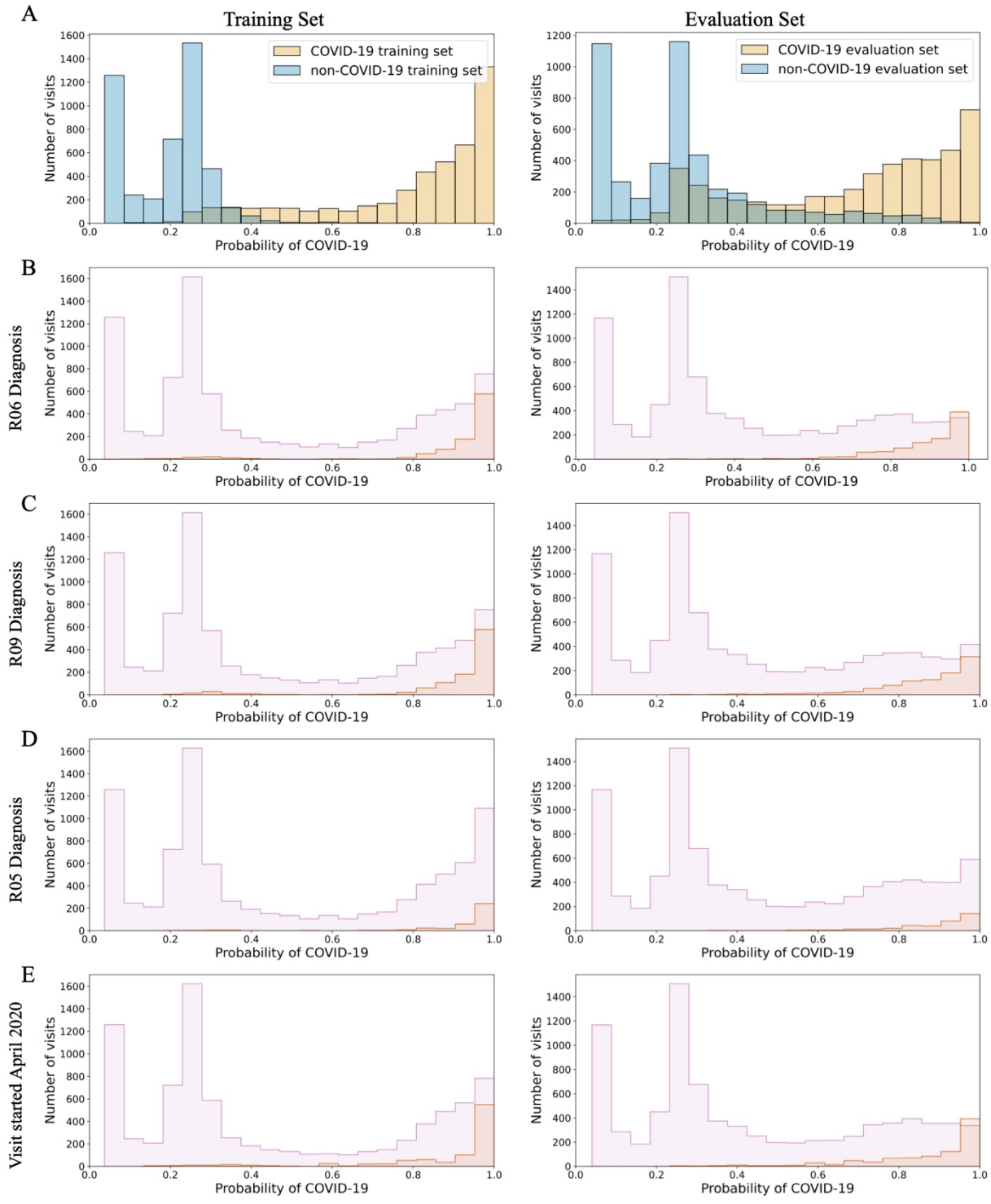


Figure 4.4 (A) ROC curves of training set, training set using out-of-bag estimates, and evaluation set based on the optimized model. COVID-19 probabilities COVID-19 (orange) and non-COVID-19 (blue) visits used in the model training (B) and evaluation (C).

We evaluated the features utilized in the final model using the Gini importance (Table 4.4). Diagnosis of abnormalities of breathing (R06), other symptoms and signs involving the circulatory and respiratory system (R09) and cough (R05) during the visit had the highest importance in the final model (Table 4.4). The distribution of the COVID-19 probabilities of the visits where the diagnoses were noted were skewed to higher COVID-19 probability than those where the diagnosis were not noted in both the training and evaluation sets (Wasserstein distance = 0.4602, 4510, 0.4458, respectively in the training set) (Figure 4.5 B-D, Table 4.4). Visits starting in April 2020, June 2021 and July 2021 were the temporal features with the highest importance in the final model (Table 4.4). The distribution of the COVID-19 probabilities of visits that started in April 2020 were skewed to higher COVID-19 probabilities than those that


A
Training Set
Evaluation Set
COVID-19 training set
non-COVID-19 training set
COVID-19 evaluation set
non-COVID-19 evaluation set
Number of visits
Probability of COVID-19
B
R06 Diagnosis
C
R09 Diagnosis
D
R05 Diagnosis
E
Visit started April 2020

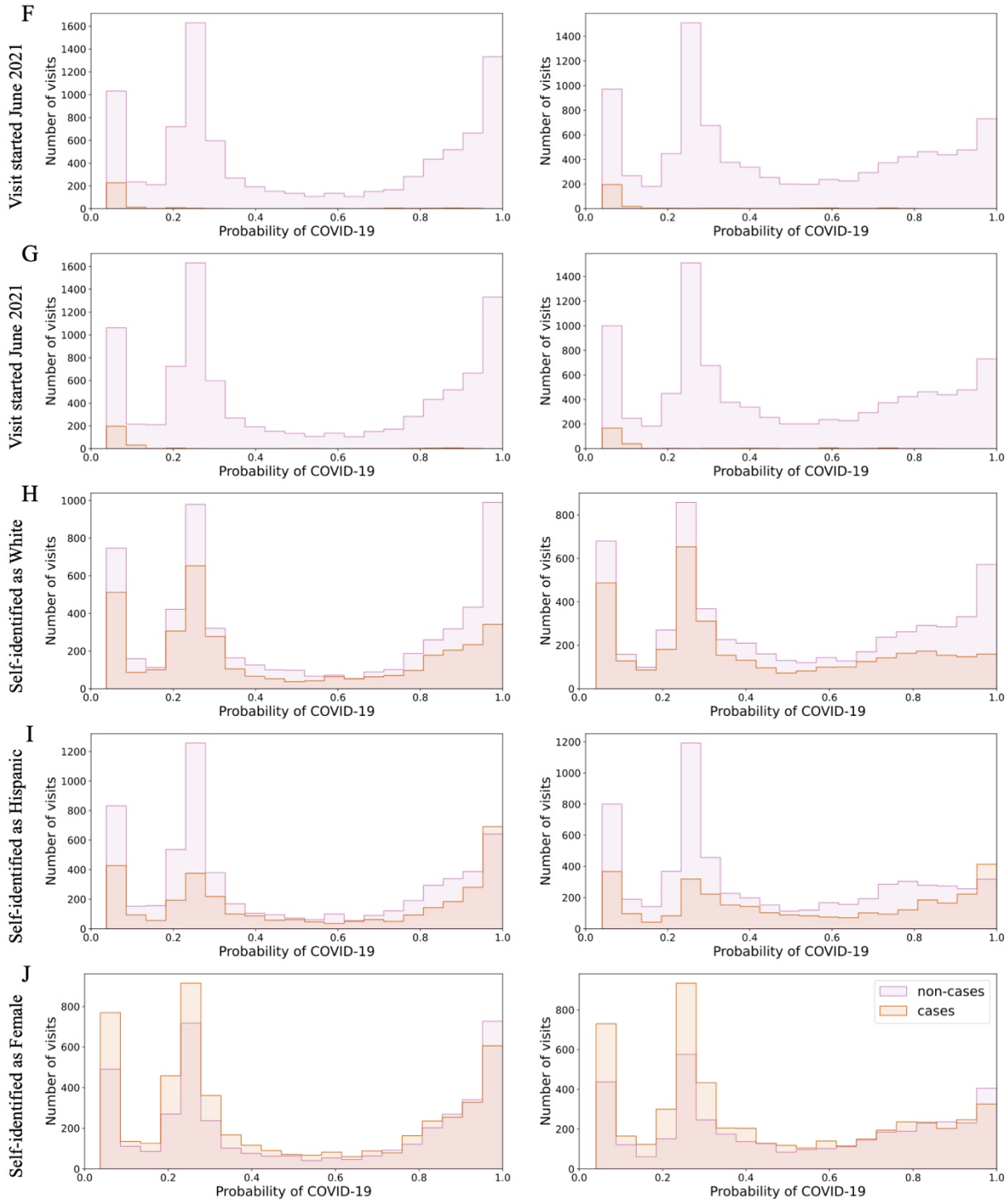


Figure 4.5 Distribution important features in random forest classifier in training and evaluation sets (A) Distribution of COVID-19 probability in COVID-19 (yellow) and non-COVID-19 (blue) training (left) and evaluations (rights) sets (top). Distribution of cases (red) and non-cases (purple) for important diagnoses (B-D), temporal (E-G) and demographic (H-J) features for training and evaluation sets. Note: R06 - abnormalities of breathing, R09 - other circulatory and respiratory system, R05 - cough.

did not start in April 2020 (Wasserstein distance = 0.4353 in the training set) (Figure 4.5 E, Table 4.4). Conversely, the distributions of the COVID-19 probabilities of visits that started in June 2021 and July 2021 were skewed to lower COVID-19 probabilities than those started at other times (Wasserstein distance = 0.3871, 0.3780, respectively in the training set) (Figure 4.5 F-G, Table 4.4). Patients self-identifying as White, of Hispanic or Latino or Spanish origin, and female were the demographic features with the highest importance in the final model (Table 4.4). The distributions of COVID-19 probabilities of visits where the patients self-identified as White or female were skewed to lower COVID-19 probabilities that those where the patient did not (Wasserstein distance = 0.0573, 0.0711, respectively in the training set) (Figure 4.5 H, 4.5 J, Table 4.4). The distribution of COVID-19 probabilities of visits where the patients self-identified as of Hispanic or Latino or Spanish origin were skewed to higher COVID-19 probabilities that those where the patient did not (Wasserstein distance = 0.0920) (Figure 4.5 I, Table 4.4). A complete table of all features is available at https://github.com/vijendra-cuimc/thesis/blob/main/table_4.4_all_features.csv.

Table 4.4 Importance for the top 20 important features and Wasserstein distance between distribution where the feature is observed and the feature is not observed. Negative Wasserstein distance indicates that the average COVID-19 probability in the set of visits where the feature was observed is less than the average of the set where the feature was not observed.

Feature	Importance	Wasserstein Distance		
		Training Set	Evaluation Set	All Visits
Abnormalities of breathing diagnosis noted during visit (R06)	0.0650	0.4602	0.4640	0.5074
Other symptoms and signs involving the circulatory and respiratory system diagnosis noted during visit (R09)	0.0628	0.4510	0.4355	0.4718
Visit started in April 2020	0.0543	0.4353	0.4371	0.4209
Cough diagnosis noted during visit (R05)	0.0259	0.4458	0.4293	0.5160
Viral pneumonia, not elsewhere classified diagnosis noted during visit (J12)	0.0236	0.4979	0.4777	0.6738
Encounter for other special examination without complaint, suspected or reported diagnosis noted during visit (Z01)	0.0234	0.2969	0.3009	0.4270
Transplanted organ and tissue status diagnosis noted during visit (Z94)	0.0229	0.3066	0.3145	0.4175
Fever of other and unknown origin diagnosis noted during visit (R50)	0.0195	0.4400	0.4169	0.5089

Table 4.4 Importance for the top 20 important features and Wasserstein distance between distribution where the feature is observed and the feature is not observed. Negative Wasserstein distance indicates that the average COVID-19 probability in the set of visits where the feature was observed is less than the average of the set where the feature was not observed.

Feature	Importance	Wasserstein Distance		
		Training Set	Evaluation Set	All Visits
Respiratory failure, not elsewhere classified diagnosis noted during visit (J96)	0.0176	0.5022	0.4920	0.6076
Self Identified as White	0.0148	-0.0573	-0.0615	-0.0079
Visit started in June 2021	0.0148	-0.3871	-0.3630	-0.2007
Self identified as of Hispanic or Latino or Spanish Origin	0.0141	0.0920	0.0899	0.0287
Self Identified Sex as Female	0.0141	-0.0711	-0.0759	-0.0187
Visit started in July 2021	0.0130	-0.3780	-0.3570	-0.1868
Visit started in August 2021	0.0126	-0.3432	-0.3643	-0.1722
Visit started in September 2021	0.0117	-0.3439	-0.3266	-0.1826
Visit started in February 2020	0.0115	-0.3903	-0.3221	-0.1221
Visit started in October 2021	0.0112	-0.3539	-0.3586	-0.1855
Type 2 diabetes mellitus diagnosis noted during visit (E11)	0.0106	0.4029	0.3949	0.4329
Acute kidney failure diagnosis noted during visit (N17)	0.0104	0.4751	0.4700	0.5350

We further evaluated the model by evaluating the distributions of COVID-19 probabilities for visits within inclusion and exclusion criteria for the training and evaluation sets (Figure 1). Compared to the distribution of COVID-19 probabilities for all of the visits between February 2020 and March 2022 (Figure 4.6 A), visits where the patient was diagnosed with COVID-19 based on the presence of the U07.1 ICD-10 code (N=9,340) during the visits were skewed to higher COVID-19 probabilities (Wasserstein distance = 0.4695) (Figure 4.6 B). The distribution of COVID-19 probabilities of visits where the patient tested positive for SARS-CoV-2 infection (N=18,156) was bimodal with a skewed to higher COVID-19 probabilities (Wasserstein distance = 0.2319) (Figure 4.6 C). The distribution of COVID-19 probabilities of visits where the patient tested negative for SARS-CoV-2 infection (N=238,438) was marginally skewed to higher COVID-19 probabilities (Wasserstein distance = 0.0550) (Figure 4.6 D). The distribution of COVID-19 probabilities of visits where clinical diagnosis notes indicated that the patient did not have COVID-19 (N=168) was skewed to higher COVID-19 probabilities (Wasserstein distance = 0.4158) (Figure 4.6 E). The distribution of COVID-19 probabilities of visits where the patient had a noted history of COVID-19 (N=899) was skewed to higher COVID-19 probabilities (Wasserstein distance = 0.3547) (Figure 4.6 F).

In order to identify what, if any, conditions are associated with a history COVID-19, we identified visits where the patient returned to the hospital within 7 days, 14 days, 21 days, 28 days, 3 months, 6 months, 9 months and 12 months by comparing the distributions of COVID-19 probabilities of visits where the patient returned within each time period and then segregated the visits into those where a particular condition was observed in the follow-up and those where the condition was not. We used a Mann-Whitney U test to compare between the two distributions for each conditions irrespective of whether or not the patient was previously diagnosed with the

condition (Figures 4.7 left, Table 4.5) and only if the patient was not diagnosed with the condition prior to the visit (Figure 4.7 right, Table 4.6). We identified, among other conditions, the distribution of COVID-19 probability preceding myocardial infarction was significantly different from the distribution of COVID-19 probability not preceding myocardial

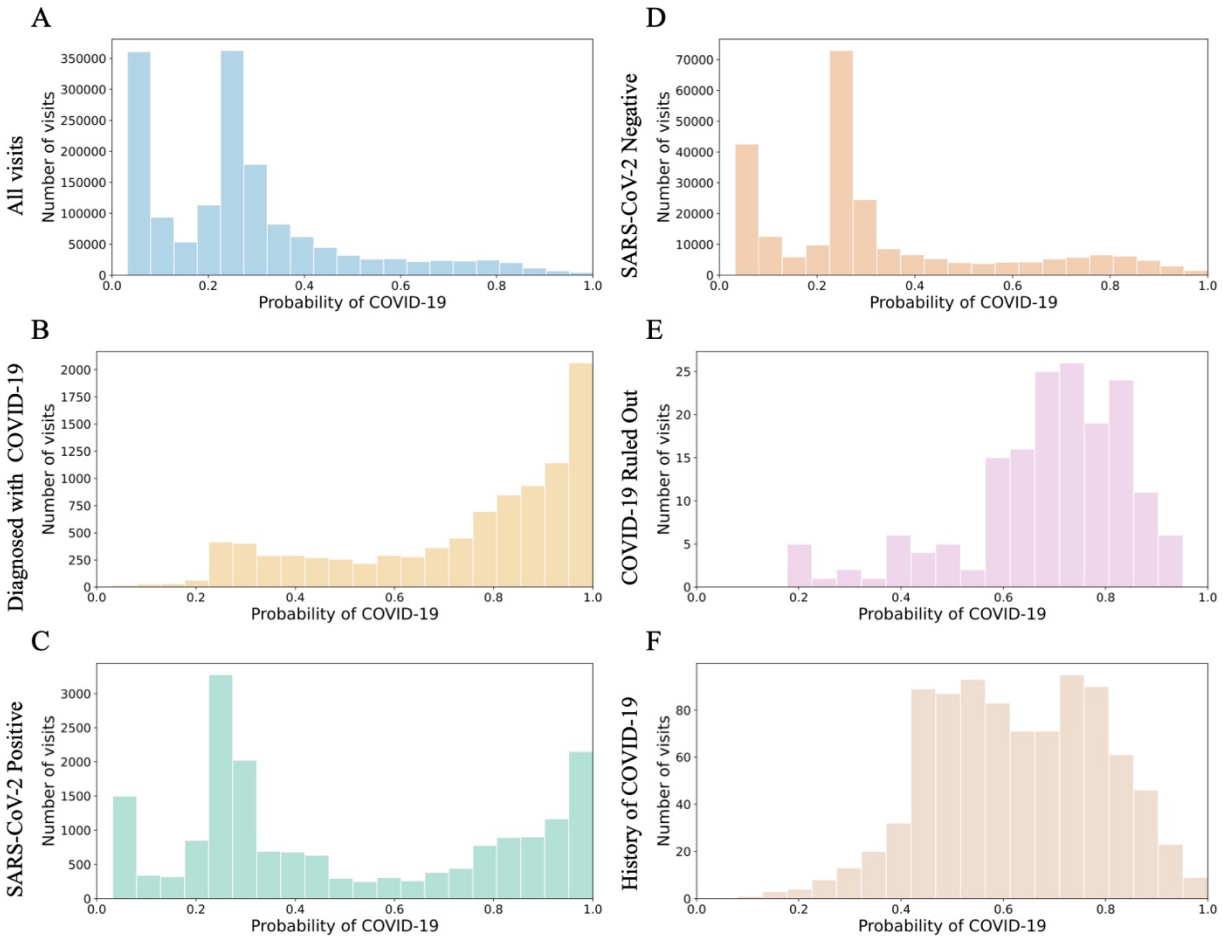
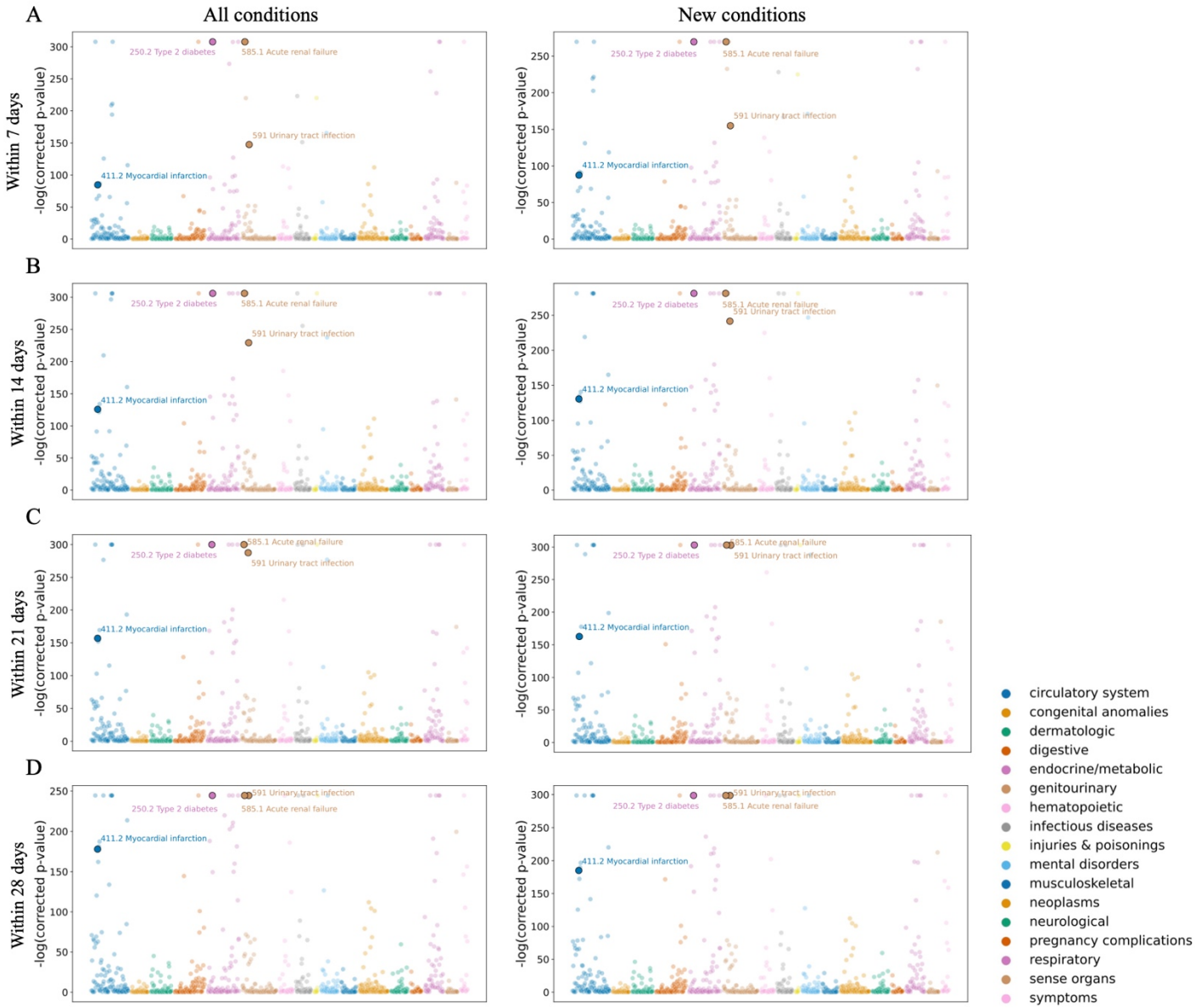


Figure 4.6 Distribution of COVID-19 probability for visits different patient groups (A) Distribution of COVID-19 probability for all visits. Distribution of visits where patients were diagnosed with COVID-19 (B), tested positive for SARS-CoV-2 infection (C), tested negative for SARS-CoV-2 infection (D), where clinical diagnosis note indicated the "COVID-19 was ruled out" (E) and visits where the patient had a history of COVID-19 (F).

infarction both with and without accounting for previous clinical history in all time periods (Mann-Whitney U test statistic = 1.206E8, FDR correct $p < 2.225E-308$, Mann-Whitney U test statistic = 1.339E8, FDR correct $p < 2.225E-308$, respectively within one year) (Figure 4.5). We

observed a similar difference with and without accounting for previous clinical history for urinary tract infection (Mann-Whitney U test statistic = 1.968E8, FDR correct $p < 2.225E-308$, Mann-Whitney U test statistic = 2.562E8, FDR correct $p < 2.225E-308$ within one year), acute renal failure (Mann-Whitney U test statistic = 8.969E7, FDR correct $p < 2.225E-308$, Mann-Whitney U test statistic = 1.234E8, FDR correct $p < 2.225E-308$ within one year), and type 2 diabetes (Mann-Whitney U test statistic = 2.317E8, FDR correct $p < 2.225E-308$, Mann-Whitney U test statistic = 3.273E8, FDR correct $p < 2.225E-308$ within one year) (Figure 4.7). A complete table of all 1,042 phenotypes is available at https://github.com/vijendra-cuimc/thesis/blob/main/table_4.5_4.6_mannwhitney_results_git_formatted.csv.


A
All conditions
New conditions
Within 7 days
Within 14 days
Within 21 days
Within 28 days
-log(corrected p-value)
250.2 Type 2 diabetes
585.1 Acute renal failure
591 Urinary tract infection
411.2 Myocardial infarction
circulatory system
congenital anomalies
dermatologic
digestive
endocrine/metabolic
genitourinary
hematopoietic
infectious diseases
injuries & poisonings
mental disorders
musculoskeletal
neoplasms
neurological
pregnancy complications
respiratory
sense organs
symptoms

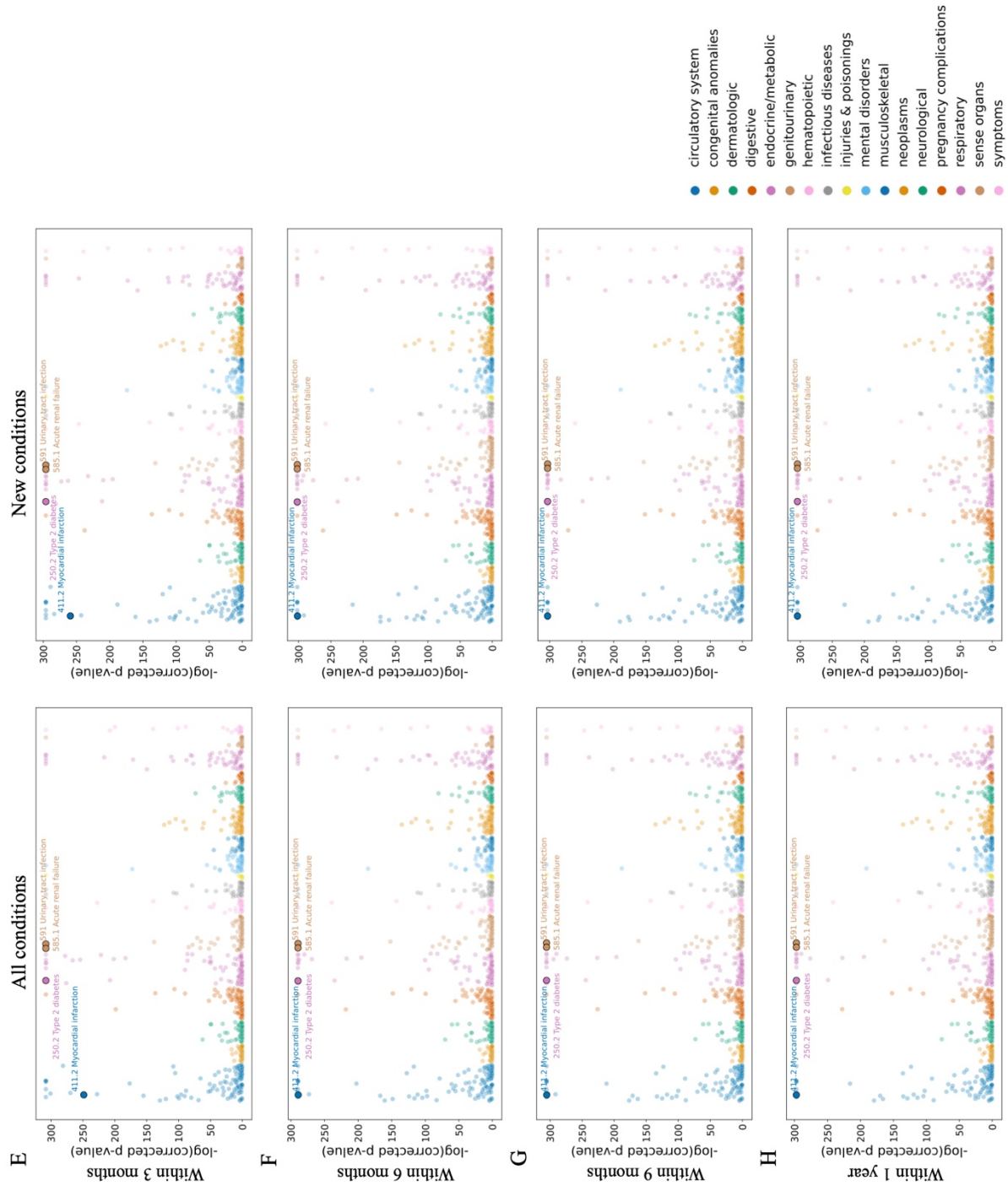


Figure 4.7 Statistical testing of conditions associated with COVID-19 -log$_{10}$(corrected *p*-value) for each phenotype (colored by family) from Mann-Whitney U test between distributions of COVID-19 probabilities of cases and non-cases for each phenotype within (A) 7 days, (B) 14 days, (C) 21 days, (D) 28 days, (E) 3 months, (F) 6 months, (G) 9 months and (H) 1 year irrespective of previous clinical list (left) and when accounting for clinical history (right).

Table 4.5 Mann-Whitney U test statistic, *p*-value, FDR corrected *p*-value (*q*-value) for the diagnosis of myocardial infarction, type 2 diabetes, acute renal failure and urinary tract infection within 8 time periods following discharge.

	Myocardial infarction (411.2)	Type 2 diabetes (250.2)	Acute renal failure (585.1)	Urinary tract infection (591)
Up to 7 days	Stat =1.084E+07 *p* = 1.064E-86 *q* = 2.550E-85	Stat = 3.499E+07 *p* = 2.017E-310 *q* = < 2.225E-308	Stat = 1.774E+07 *p* = 2.017E-310 *q* = < 2.225E-308	Stat = 2.434E+07 *p* = 7.655E-150 *q* = 3.135E-148
Up to 14 days	Stat = 2.375E+07 *p* = 5.440E-128 *q* = 1.361E-126	Stat = 7.228E+07 *p* = 2.497E-308 *q* = < 2.225E-308	Stat = 3.526E+07 *p* = 2.497E-308 *q* = < 2.225E-308	Stat = 5.186E+07 *p* = 1.474E-231 *q* = 6.148E-230
Up to 21 days	Stat = 3.299E+07 *p* = 5.824E-159 *q* = 1.551E-157	Stat = 1.007E+08 *p* = 2.137E-302 *q* = < 2.225E-308	Stat = 4.718E+07 *p* = 2.137E-302 *q* = < 2.225E-308	Stat = 7.445E+07 *p* = 6.110E-290 *q* = 2.688E-288
Up to 28 days	Stat = 4.288E+07 *p* = 4.102E-180 *q* = 1.070E-178	Stat = 1.294E+08 *p* = 7.017E-247 *q* = < 2.225E-308	Stat = 5.805E+07 *p* = 7.017E-247 *q* = < 2.225E-308	Stat = 9.688E+07 *p* = 7.017E-247 *q* = < 2.225E-308
Up to 3 months	Stat = 9.049E+07 *p* = 9.358E-251 *q* = 2.551E-249	Stat = 2.317E+08 *p* = 2.264E-310 *q* = < 2.225E-308	Stat = 9.649E+07 *p* = 2.264E-310 *q* = < 2.225E-308	Stat = 1.829E+08 *p* = 2.264E-310 *q* = < 2.225E-308
Up to 6 months	Stat = 1.161E+08 *p* = 2.472E-291 *q* = 7.156E-290	Stat = 2.914E+08 *p* = 1.236E-291 *q* = < 2.225E-308	Stat = 1.137E+08 *p* = 1.236E-291 *q* = < 2.225E-308	Stat = 2.289E+08 *p* = 1.236E-291 *q* = < 2.225E-308
Up to 9 months	Stat = 1.287E+08 *p* = 1.595E-307 *q* = 4.639E-306	Stat = 3.158E+08 *p* = 7.975E-30 *q* = < 2.225E-308	Stat = 1.205E+08 *p* = 7.975E-308 *q* = < 2.225E-308	Stat = 2.481E+08 *p* = 7.975E-308 *q* = < 2.225E-308
Up to 1 year	Stat = 1.339E+08 *p* = 2.944E-299 *q* = < 2.225E-308	Stat = 3.273E+08 *p* = 2.944E-299 *q* = < 2.225E-308	Stat = 1.234E+08 *p* = 2.944E-299 *q* = < 2.225E-308	Stat = 2.562E+08 *p* = 2.944E-299 *q* = < 2.225E-308

Table 4.6 Mann-Whitney U test statistic, *p*-value, FDR corrected *p*-value (*q*-value) for new diagnosis of myocardial infarction, type 2 diabetes, acute renal failue and urinary tract infection within 8 time periods following discharge.

	Myocardial infarction (411.2)	Type 2 diabetes (250.2)	Acute renal failure (585.1)	Urinary tract infection (591)
Up to 7 days	Stat = 9.875E+06 *p* = 1.542E-89 *q* = 3.789E-88	Stat = 2.548E+07 *p* = 3.976E-272 *q* = < 2.225E-308	Stat = 1.283E+07 *p* = 3.976E-272 *q* = < 2.225E-308	Stat = 1.917E+07 *p* = 2.816E-157 *q* = 1.153E-155
Up to 14 days	Stat = 2.154E+07 *p* = 1.099E-132 *q* = 2.749E-131	Stat = 5.174E+07 *p* = 1.065E-283 *q* = < 2.225E-308	Stat = 2.522E+07 *p* = 1.065E-283 *q* = < 2.225E-308	Stat = 4.045E+07 *p* = 6.999E-244 *q* = 2.919E-242
Up to 21 days	Stat = 2.984E+07 *p* = 6.671E-165 *q* = 1.777E-163	Stat = 7.179E+07 *p* = 1.565E-305 *q* = < 2.225E-308	Stat = 3.378E+07 *p* = 1.565E-305 *q* = < 2.225E-308	Stat = 5.784E+07 *p* = 3.130E-305 *q* = 1.377E-303
Up to 28 days	Stat = 3.873E+07 *p* = 3.986E-187 *q* = 1.014E-185	Stat = 9.160E+07 *p* = 4.561E-301 *q* = < 2.225E-308	Stat = 4.097E+07 *p* = 4.561E-301 *q* = < 2.225E-308	Stat = 7.464E+07 *p* = 4.561E-301 *q* = < 2.225E-308
Up to 3 months	Stat = 8.155E+07 *p* = 1.935E-261 *q* = 5.276E-260	Stat = 1.636E+08 *p* = 1.831E-298 *q* = < 2.225E-308	Stat = 6.900E+07 *p* = 1.831E-298 *q* = < 2.225E-308	Stat = 1.404E+08 *p* = 1.831E-298 *q* = < 2.225E-308
Up to 6 months	Stat = 1.045E+08 *p* = 6.732E-304 *q* = 1.949E-302	Stat = 2.059E+08 *p* = 3.366E-304 *q* = < 2.225E-308	Stat = 8.204E+07 *p* = 3.366E-304 *q* = < 2.225E-308	Stat = 1.755E+08 *p* = 3.366E-304 *q* = < 2.225E-308
Up to 9 months	Stat = 1.159E+08 *p* = 2.082E-305 *q* = < 2.225E-308	Stat = 2.233E+08 *p* = 2.082E-305 *q* = < 2.225E-308	Stat = 8.740E+07 *p* = 2.082E-305 *q* = < 2.225E-308	Stat = 1.905E+08 *p* = 2.082E-305 *q* = < 2.225E-308
Up to 1 year	Stat = 1.206E+08 *p* = 1.793E-306 *q* = < 2.225E-308	Stat = 2.317E+08 *p* = 1.793E-306 *q* = < 2.225E-308	Stat = 8.969E+07 *p* = 1.793E-306 *q* = < 2.225E-308	Stat = 1.968E+08 *p* = 1.793E-306 *q* = < 2.225E-308

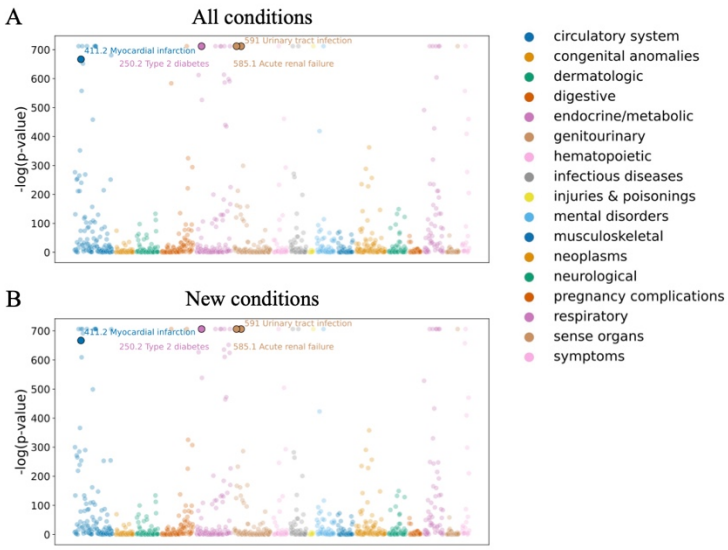


Figure 4.8 Statistical testing of conditions associated with COVID-19 (A)-log$_{10}$(*p*-value) for each phenotype (colored by family) from Cox Proportional Hazards test for COVID-19 probability of the previous visit for conditions developed within 1 year irrespective of clinical history (left) and when accounting for clinical history (right).

To further investigate the association between COVID-19 probability and the onset of conditions, we calculated the hazard ratio using a Cox proportional hazards model for COVID-19 probability irrespective of previous clinical history (Figure 4.8 A) and respective of previous clinical history (Figure 4.8 B). Increasing COVID-19 probability in the preceding visit was associated with increases risk of myocardial infarction within one year with and without accounting for previous clinical history (Hazards ratio = 93.713 (73.906-118.829), p = 2.199E-307 and Hazards ratio = 82.557 (65.102-104.693), p = 2.414E-290, respectively) (Table 4.7). A similar association was observed within one year with and without accounting for previous clinical history for urinary tract infection (Hazards ratio = 75.241 (63.192 -89.587), p < 2.225E-308 and Hazards ratio = 62.038 (52.176 -73.765), p < 2.225E-308, respectively), acute renal failure (Hazards ratio = 7762.722 (6156.997 - 9787.216), p < 2.225E-308 and Hazards ratio = 5488.974 (4345.262 - 6933.722), p < 2.225E-308, respectively) and type 2 diabetes (Hazards ratio = 403.553 (350.901 - 464.106), p < 2.225E-308 and Hazards ratio = 270.035 (235.213 -

310.013), $p < 2.225E-308$, respectively) (Table 4.7). A complete table of all 1,042 phenotypes is available at https://github.com/vijendra-cuimc/thesis/blob/main/table_4.7_coxph_results_git_formatted.csv

Table 4.7 Univariate hazards ratio, 95% confidence interval and *p*-value of COVID-19 probability from Cox proportional hazards tests of myocardial infarction, type 2 diabetes, acute renal failure and urinary tract infection.

	Myocardial infarction (411.2)	Type 2 diabetes (250.2)	Acute renal failure (585.1)	Urinary tract infection (591)
All conditions	82.557 (65.102, 104.693) 2.414E-290	270.035 (235.213, 310.013) < 2.225E-308	5488.974 (4.345E3, 6.934E3) < 2.225E-308	62.038 (52.176, 73.765) < 2.225E-308
New Conditions	93.713 (73.906, 118.829) 2.199E-307	403.553 (350.901, 464.106) < 2.225E-308	7762.723 (6.157E3, 9.787E3) < 2.225E-308	75.241 (63.192, 89.587) < 2.225E-308

Among the visits with a follow-up within one year, the ten most frequently observed phenotypes were essential hypertension (401.1), shortness of breath (512.7), hyperlipidemia (272.1), other complications of pregnancy NEC (646), cough (512.8), back pain (760), injury, NOS (1009), gastroesophageal reflux disease (530.11), other headache syndromes (339), and pulmonary collapse; interstitial and compensatory emphysema (508), respectively. When accounting for demographics and the ten most frequently observed phenotypes in a multivariate Cox proportional hazards model, increasing COVID-19 probability in the preceding visit was associated with increased risk of myocardial infarction within one year with and without accounting for previous clinical history (Hazards ratio = 121.736 (87.375, 169.611), $p = 3.796E-177$ and Hazards ratio = 80.262 (4.134, 4.637), $p = 4.543E-256$, respectively) (Table 4.8). A

similar association was observed within one year with and without accounting for previous clinical history for urinary tract infection (Hazards ratio = 72.021 (58.116 - 89.253), p < 2.225E-308 and Hazards ratio = 61.380 (51.273 - 73.479), p < 2.225E-308, respectively), acute renal failure (Hazards ratio = 1.264E4 (9.278E4 - 1.724E4), p < 2.225E-308 and Hazards ratio = 6.333E3 (4.947E3 - 8.108E3), p < 2.225E-308, respectively) and type 2 diabetes (Hazards ratio = 345.730 (283.180 - 422.098), p < 2.225E-308 and Hazards ratio = 217.271 (187.898 - 251.235), p = 1.39E-22, respectively) (Table 4.9).

Table 4.8 Multivariate hazards ratio, 95% confidence interval and *p*-value from Cox proportional hazards tests myocardial infarction, type 2 diabetes, acute renal failure and urinary tract infection.

	Myocardial infarction (411.2)	Type 2 diabetes (250.2)	Acute renal failure (585.1)	Urinary tract infection (591)
Age Adolescent (≥ 13 and < 19)	0.202 (0.073, 0.559) 2.084E-03	4.538 (1.520, 13.542) 6.708E-03	2.805 (1.956, 4.022) 2.038E-08	0.466 (0.334, 0.651) 7.132E-06
Age Adult (≥ 19 and < 60)	0.964 (0.711, 1.308) 0.814	17.850 (7.396, 43.081) 1.445E-10	2.644 (2.014, 3.470) 2.514E-12	0.606 (0.516, 0.712) 9.451E-10
Age Senior (≥ 60)	2.135 (1.579, 2.888) 8.485E-07	33.457 (13.860, 80.759) 5.843E-15	3.809 (2.900, 5.005) 7.594E-22	0.673 (0.569, 0.796) 3.834E-06
Self Identified Sex as Female	0.510 (0.454, 0.574) 3.249E-29	0.797 (0.735, 0.864) 4.054E-08	0.583 (0.541, 0.629) 4.651E-45	1.407 (1.289, 1.535) 1.595E-14
Self Identified as American Indian or Alaskan Native	3.889 (2.329, 6.495) 2.076E-07	1.007 (0.479, 2.116) 0.986	1.175 (0.610, 2.264) 0.629	1.020 (0.486, 2.145) 0.957
Self Identified as Asian	0.334 (0.192, 0.582) 1.068E-04	1.099 (0.868, 1.391) 0.432	1.750 (1.452, 2.109) 4.367E-09	1.187 (0.910, 1.547) 0.206
Self Identified as Black or African American	0.749 (0.628, 0.895) 1.447E-03	0.941 (0.835, 1.061) 0.321	1.331 (1.198, 1.479) 1.047E-07	1.085 (0.958, 1.228) 0.199
Self Identified as Native Hawaiian or Other Pacific Islander	1.324 (0.330, 5.312) 0.692	1.366 (0.567, 3.294) 0.487	0.642 (0.160, 2.570) 0.531	0.790 (0.197, 3.164) 0.739
Self Identified as White	0.783 (0.687, 0.891) 2.136E-04	0.631 (0.572, 0.696) 2.018E-20	1.029 (0.939, 1.127) 0.537	1.131 (1.026, 1.247) 1.304E-02

Table 4.8 Multivariate hazards ratio, 95% confidence interval and *p*-value from Cox proportional hazards tests myocardial infarction, type 2 diabetes, acute renal failure and urinary tract infection. (cont).

	Myocardial infarction (411.2)	Type 2 diabetes (250.2)	Acute renal failure (585.1)	Urinary tract infection (591)
Self identified as Hispanic or of Latino or Spanish Origin	0.766 (0.667, 0.881) 1.851E-04	1.005, (0.912, 1.109) 0.916	0.726 (0.662, 0.796) 7.626E-12	1.104 (1.002, 1.216) 4.493E-02
COVID-19 Probability	80.262 (62.417, 103.208) 4.543E-256	345.730 (283.180, 422.098) < 2.225E-308	6333.163 (4947.052, 8107.648) < 2.225E-308	61.380 (51.273, 73.479) < 2.225E-308
Hyperlipidemia	1.423 (1.235, 1.640) 1.094E-06	2.413 (2.186, 2.662) 8.284E-69	0.902 (0.825, 0.986) 2.347E-02	1.043 (0.936, 1.162) 0.442
Other headache syndrome	1.004 (0.859, 1.173) 0.963	0.923 (0.832, 1.025) 0.134	0.798 (0.722, 0.881) 8.288E-06	0.943 (0.847, 1.051) 0.289
Essential hypertension	1.043 (0.898, 1.211) 0.585	1.711 (1.537, 1.906) 1.213E-22	1.166 (1.061, 1.282) 1.500E-03	1.027 (0.920, 1.146) 0.638
Pulmonary collapse; interstitial and compensatory emphysema	1.079 (0.937, 1.243) 0.290	1.056 (0.959, 1.164) 0.268	1.348 (1.232, 1.475) 7.775E-11	1.223 (1.101, 1.359) 1.738E-04
Shortness of breath	0.949 (0.833, 1.080) 0.427	0.967 (0.884, 1.058) 0.465	0.875 (0.803, 0.953) 2.130E-03	0.940 (0.850, 1.039) 0.225
Cough	0.673 (0.584, 0.775) 3.906E-08	0.880 (0.801, 0.967) 7.829E-03	0.850 (0.780, 0.927) 2.357E-04	0.877 (0.794, 0.968) 9.381E-03
Gastroesophageal reflux disease	0.921 (0.803, 1.057) 0.242	1.091 (0.996, 1.196) 6.164E-02	1.031 (0.945, 1.125) 0.487	1.186 (1.073, 1.310) 7.950E-04

Table 4.8 Multivariate hazards ratio, 95% confidence interval and *p*-value from Cox proportional hazards tests myocardial infarction, type 2 diabetes, acute renal failure and urinary tract infection. (cont).

	Myocardial infarction (411.2)	Type 2 diabetes (250.2)	Acute renal failure (585.1)	Urinary tract infection (591)
Other complication of pregnancy NEC	0.819 (0.705, 0.952) 9.230E-03	0.956 (0.867, 1.055) 0.374	0.872 (0.796, 0.955) 3.107E-03	0.854 (0.773, 0.943) 1.869E-03
Back pain	0.823 (0.712, 0.951) 8.425E-03	0.964 (0.876, 1.061) 0.451	0.882 (0.804, 0.966) 7.121E-03	1.200 (1.084, 1.329) 4.696E-04
Injury NOS	1.035 (0.897, 1.195) 0.638	0.990 (0.897, 1.092) 0.837	1.042 (0.952, 1.141) 0.370	0.965 (0.873, 1.068) 0.492

Table 4.9 Multivariate hazards ratio, 95% confidence interval and *p*-value from Cox proportional hazards tests for new diagnosis of myocardial infarction, type 2 diabetes, acute renal failure and urinary tract infection.

	Myocardial infarction (411.2)	Type 2 diabetes (250.2)	Acute renal failure (585.1)	Urinary tract infection (591)
Age Adolescent (≥ 13 and < 19)	0.382 (0.086, 1.695) 0.206	4.538 (1.520, 13.542) 6.708E-03	2.968 (1.938, 4.547) 5.690E-07	0.537 (0.365, 0.792) 1.692E-03
Age Adult (≥ 19 and < 60)	1.944 (1.100, 3.436) 2.213E-02	17.850 (7.396, 43.081) 1.445E-10	2.696 (1.952, 3.723) 1.717E-09	0.652 (0.535, 0.796) 2.549E-05
Age Senior (≥ 60)	4.979 (2.829, 8.762) 2.592E-08	33.457 (13.860, 80.759) 5.843E-15	3.831 (2.773, 5.293) 3.877E-16	0.922 (0.750, 1.132) 0.436
Self Identified Sex as Female	0.466 (0.401, 0.541) 2.367E-23	0.797 (0.735, 0.864) 4.054E-08	0.548 (0.500, 0.600) 1.078E-38	1.629 (1.470, 1.806) 1.404E-20
Self Identified as American Indian or Alaskan Native	0.445 (0.062, 3.167) 0.419	1.007 (0.479, 2.116) 0.986	1.242 (0.557, 2.772) 0.596	1.296 (0.581, 2.894) 0.526
Self Identified as Asian	0.381 (0.195, 0.743) 4.658E-03	1.099 (0.868, 1.391) 0.432	1.149 (0.858, 1.540) 0.351	1.231 (0.891, 1.700) 0.207
Self Identified as Black or African American	0.813 (0.648, 1.019) 7.210E-02	0.941 (0.835, 1.061) 0.321	1.461 (1.290, 1.655) 2.627E-09	1.043 (0.897, 1.212) 0.585
Self Identified as Native Hawaiian or Other Pacific Islander	2.143 (0.533, 8.614) 0.283	1.366 (0.567, 3.294) 0.487	0.344 (0.048, 2.446) 0.286	0.936 (0.234, 3.753) 0.926
Self Identified as White	0.801 (0.676, 0.948) 1.008E-02	0.631 (0.572, 0.696) 2.018E-20	0.975 (0.874, 1.086) 0.643	1.174 (1.046, 1.318) 6.604E-03
Self identified as Hispanic or of Latino or Spanish Origin	0.734 (0.612, 0.881) 8.753E-04	1.005 (0.912, 1.109) 0.916	0.749 (0.672, 0.836) 2.336E-07	1.148 (1.023, 1.288) 1.855E-02

Table 4.9 Multivariate hazards ratio, 95% confidence interval and *p*-value from Cox proportional hazards tests for new diagnosis of myocardial infarction, type 2 diabetes, acute renal failure and urinary tract infection. (cont.)

	Myocardial infarction (411.2)	Type 2 diabetes (250.2)	Acute renal failure (585.1)	Urinary tract infection (591)
COVID-19 Probability	121.736 (87.375, 169.611) 3.796E-177	345.730 (283.180, 422.098) < 2.225E-308	12647.836 (9277.739, 17242.105) < 2.225E-308	72.021 (58.116, 89.253) < 2.225E-308
Hyperlipidemia	1.746 (1.462, 2.085) 7.839E-10	2.413 (2.186, 2.662) 8.284E-69	1.055 (0.947, 1.174) 0.330	1.046 (0.920, 1.188) 0.493
Other headache syndrome	1.065 (0.880, 1.290) 0.517	0.923 (0.832, 1.025) 0.134	0.806 (0.715, 0.908) 3.912E-04	1.091 (0.962, 1.237) 0.173
Essential hypertension	1.004 (0.831, 1.213) 0.967	1.711 (1.537, 1.906) 1.213E-22	1.310 (1.170, 1.466) 2.739E-06	1.037 (0.912, 1.180) 0.577
Pulmonary collapse; interstitial and compensatory emphysema	1.209 (1.013, 1.443) 3.515E-02	1.056 (0.959, 1.164) 0.268	1.944 (1.750, 2.161) 5.218E-35	1.378 (1.215, 1.563) 6.025E-07
Shortness of breath	1.032 (0.877, 1.215) 0.706	0.967 (0.884, 1.058) 0.465	0.857 (0.775, 0.947) 2.544E-03	0.822 (0.730, 0.925) 1.192E-03
Cough	0.754 (0.633, 0.898) 1.562E-03	0.880 (0.801, 0.967) 7.829E-03	0.774 (0.697, 0.859) 1.593E-06	1.024 (0.911, 1.151) 0.686
Gastroesophageal reflux disease	1.008 (0.850, 1.194) 0.931	1.091 (0.996, 1.196) 6.164E-02	1.100 (0.991, 1.222) 7.417E-02	1.197 (1.063, 1.347) 2.889E-03
Other complication of pregnancy NEC	0.863 (0.715, 1.041) 0.124	0.956 (0.867, 1.055) 0.374	0.956 (0.856, 1.068) 0.427	1.009 (0.897, 1.136) 0.878

Table 4.9 Multivariate hazards ratio, 95% confidence interval and *p*-value from Cox proportional hazards tests for new diagnosis of myocardial infarction, type 2 diabetes, acute renal failure and urinary tract infection. (cont.)

	Myocardial infarction (411.2)	Type 2 diabetes (250.2)	Acute renal failure (585.1)	Urinary tract infection (591)
Back pain	0.872 (0.729, 1.043) 0.133	0.964 (0.876, 1.061) 0.451	0.910 (0.815, 1.016) 9.337E-02	1.440 (1.279, 1.622) 1.697E-09
Injury NOS	1.062 (0.887, 1.272) 0.510	0.990 (0.897, 1.092) 0.837	1.111 (0.997, 1.238) 5.637E-02	1.258 (1.120, 1.414) 1.076E-04

We further stratified the COVID-19 probabilities into quintiles and generated Kaplan-Meier curves for the data within one year (Figure 6B-E). The Kaplan-Meier curves stratified by COVID-19 probability for myocardial infarction showed three distinct sets, (i) COVID-19 probability greater than 0.6, (ii) COVID-19 probability greater than 0.4 and less than or equal to 0.6 and (iii) COVID-19 probability less than or equal to 0.4, with higher incidence observed in the sets of higher COVID-19 probability (Figure 4.9 A). The Kaplan-Meier curves for urinary tract infection showed three sets, (i) COVID-19 probability greater than 0.8, (ii) COVID-19 probability greater than 0.4 and less than or equal to 0.8 and (iii) COVID-19 probability less than or equal to 0.4, up to 8 months with the higher incidence observed in the sets of higher COVID-19 probability (Figure 4.9 B). The Kaplan-Meier curves for acute renal failure showed four distinct sets, (i) COVID-19 probability greater than 0.8, (ii) COVID-19 probability greater than 0.6 and less than or equal to 0.8, (iii) COVID-19 probability greater than 0.4 and less than or equal to 0.6 and (iii) COVID-19 probability less than or equal to 0.4, with the higher incidence observed in the sets of higher COVID-19 probability (Figure 4.9 C). The Kaplan-Meier curves

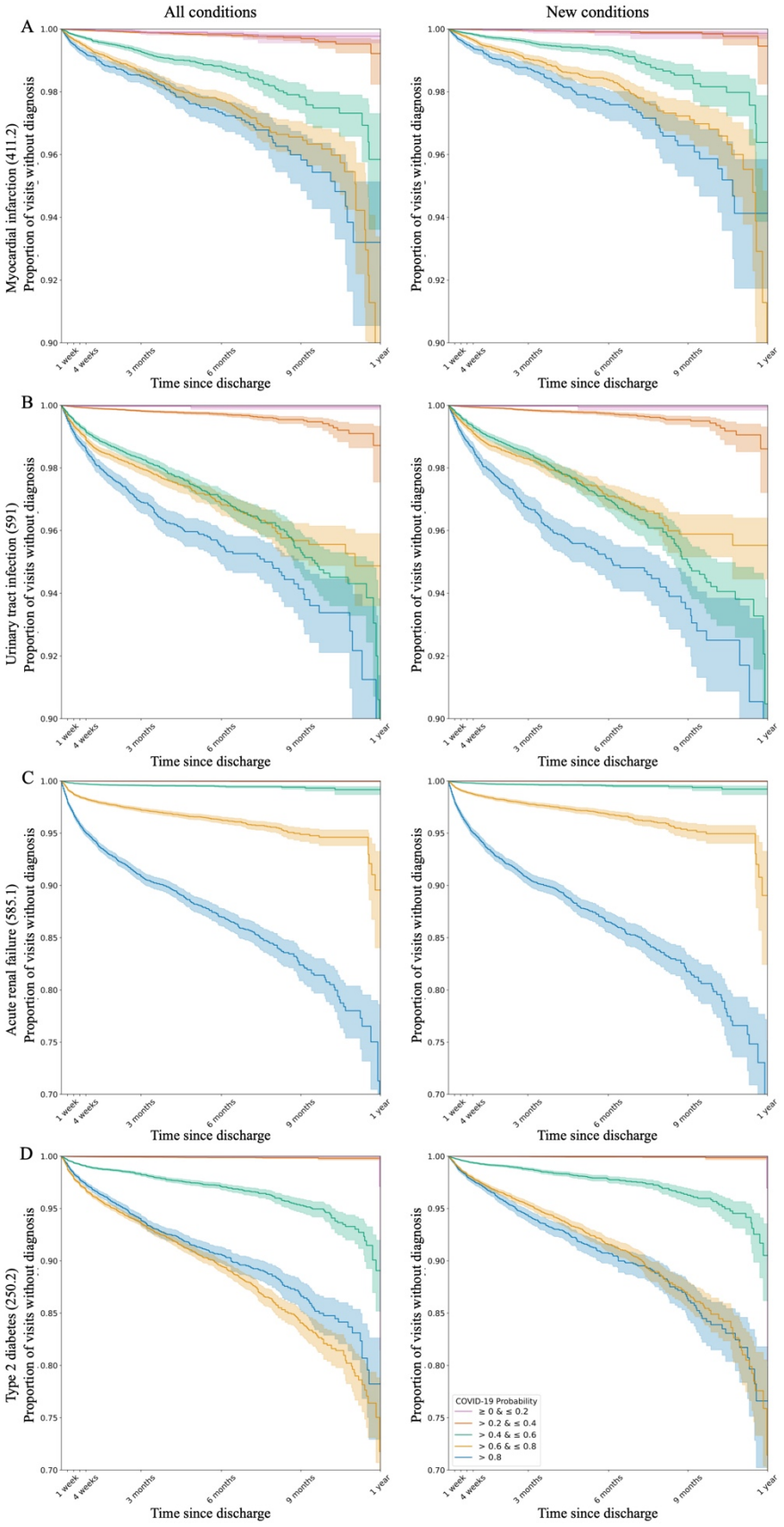

All conditions
New conditions
A
B
C
D
Myocardial infarction (411.2)
Urinary tract infection (591)
Acute renal failure (585.1)
Type 2 diabetes (250.2)
Proportion of visits without diagnosis
Time since discharge
1 week
4 weeks
3 months
6 months
9 months
1 year
COVID-19 Probability
≥ 0 & ≤ 0.2
> 0.2 & ≤ 0.4
> 0.4 & ≤ 0.6
> 0.6 & ≤ 0.8
> 0.8

Figure 4.9 Statistical testing of conditions associated with COVID-19 Kaplan-Meier curves for (A) myocardial infarction, (B) urinary tract infection, (C) acute renal failure, (D) type 2 diabetes stratified by COVID-19 probability quintile within 1 year irrespective of clinical history (left) and when accounting for clinical history (right).

for the onset of type 2 diabetes showed three distinct sets (i) COVID-19 probability greater than 0.6, (ii) COVID-19 probability greater than 0.4 and less than or equal to 0.6 and (iii) COVID-19 probability less than or equal to 0.4, with higher incidence observed in the sets of higher COVID-19 probability (Figure 4.9 D).

4.4 Discussion

In this study, we collected demographic, temporal and clinical data from 434,152 patients who sought treatment at New York-Presbyterian over 1,573,113 visits between February 2020 and March 2022, who had at least one interaction with Columbia University Irving Medical Center, to develop an algorithm to identify conditions that are associated with COVID-19. The 26 month period from which our data is sourced encompasses the height of the first wave of the COVID-19 pandemic (Spring 2020) when New York City was an epicenter in the United States as well as the subsequent Delta and Omicron waves. Additionally, our data encompasses periods, such as summer 2020 when case counts were at some of their lowest levels throughout the pandemic, as well as the period following development of treatments for COVID-19 and prophylactics for SARS-CoV-2 infection.

Using data for patients who had COVID-19 diagnosed (as determined by the presence of the U07.1 ICD-10 diagnosis code) and non-COVID-19 patients, we trained an optimized random forest classifier with high performance as evaluated in an independent data set, and applied it to the full set of 1,573,113 visits. Instead of the binary classification that would result from the random forest classifier, we instead treated the fraction of estimators that identified the visit as a

COVID-19 visit as a probability of a patient having been diagnosed with COVID-19 during that visit. While the random forest classifier is overfitting based on the high AUROC observed in the training set, we were comfortable using it because it performed similarly in the training set using out-of-bag estimates and the evaluation set. Based on the presence of U07.1 ICD-10 diagnosis code, there were only 9,340 where the patient was diagnosed with COVID-19, however our model identified 198,562 visits where the patients had a probability of having been diagnosed with COVID-19 greater than 0.5.

When evaluating our model, the most important features represented previously identified differences between demographic groups, such as those who identify as Hispanic or Latino or of Spanish origin or Black or African American [99,100]. (Table 3.4, Figure 3.5). Important temporal features represented periods of extreme case counts in New York City, such as spring 2020 and summer 2021 [101] (Table 3.4, Figure 3.5). Important clinical diagnoses were reflective of known symptoms of COVID-19, such as abnormalities of breathing (R06), other symptoms and signs involving the circulatory and respiratory system (R09) and cough (R05) (Table 3.4, Figure 3.5).

Using these visit specific probabilities, we identified conditions that developed within different time periods after the visit (up to 7 days, 14 days, 21 days, 28 days, 3 moths, 6 months, 9 months, and 12 months) and used a Mann-Whitney U test to identify conditions that were associated with increased COVID-19 probability. Among others, our analysis identified myocardial infarction, urinary tract infection, acute renal failure and type 2 diabetes as being associated with COVID-19 (Figure 3.8). In further analysis of the results of our results, we estimated the hazards ratio of COVID-19 probability for each of these conditions (Table 3.7). Cox proportional hazards model indicated that higher COVID-19 probability in the preceding

visit was associated with an increased risk of myocardial infarction, urinary tract infection, acute renal failure and type 2 diabetes within one year. Our result for myocardial infarction is consistent with those of researchers who identified a higher risk of heart attack and ischemic stroke in COVID-19 patients using self-controlled case series [84].

Results from a retrospective observational study of patients in early 2020 observed that severe COVID-19 disease is associate with acute kidney injury [87]. The researchers have suggested that the observed inflammation in the kidney is similar to that of the lungs in COVID-19 patients because of the similar structure of the organs. The alveoli of the lungs are like the nephron of the kidney and the interstitium of the lung is like the calyx of the kidney. While not directly investigated in this study, genetic variation, drug exposure and past clinical history can also influence the risk of developing acute renal failure.

Other researchers have identified an increased risk of type 2 diabetes in patients who had been infected with SARS-CoV-2 compared to patient who had not and compared to a historical control [86]. The onset of type 2 diabetes is thought to be the due to increased stress on the pancreas. SARS-CoV-2 is known to infect pancreatic cells, however unlike in type 1 diabetes where islets cells are target by the immune system preventing the pancreas from producing insulin, the pancreas in type 2 diabetes is not producing enough inclusion leading the body to be in a prolonger hyperglycemic state.

There remains a caveat in how to interpret the hazards ratios that range from less than 10 to more than 1,000 due to how the random forest classifier was fit. The random forest classifier assumed a continuous probability between 0 and 1; using proxies for subcategories such as between 0 and 4 for quintiles or between 0 and 9 for deciles would result in smaller hazards ratios with a similar interpretation to those identified in this study.

Furthermore, while the probabilities in this study ae assumed to be continuous, they do not necessarily correlate with severity. Because random forest classifiers are dependent on the amount of data included when the model is applied, the presence of more COVID-19 symptoms diagnosis codes may correlate with a higher probability. As such, while it may be that probability is correlated with severity, this study does not assume that in the analyses.

The main limitation of this method is that it cannot be universally applied – a new random forest classifier would have to be trained for each site. The random forest classifier is also depended on input data, so in order for the model to predict the probability of a patient having had COVID-19 during a visit, they would have need to have diagnoses entered. As such, patients who expired on arrival due to severe COVID-19 were not included in the model and the model should not be used on data from such patients.

If the random forest classifier could be expanded from NYP/CUIMC, to a national dataset, we expect that important features will remain the same; they will reflect know symptoms of COVID-19. However, depending on the coding practices of individual hospitals or healthcare sites, the specific diagnosis code that are important will differ and a combined dataset would include redundancies. The important temporal covariates would change depending on location as case counts and trends differed between cities. The temporal covariates would most likely no longer be important in a model that is trained on data combined from multiple sites.

While this study shows that demographic, temporal and clinical data can be utilized to predict the probability of a patient having COVID-19 during their visit, the model and the important features are specific to NYP/CUIMC. An implementation this model elsewhere is expected to identify important temporal features specific to the site (e.g. periods of extreme case counts varied between New York City and London) and demographic variables depending on the

patient's seeking treatment at those sites. However, it would be expected to identify similar clinical variables that are representative of known symptoms or comorbidities associated with COVID-19. While the results concur with other studies, they are not without their biases as this study relied on patients who sought treatment at New York-Presbyterian on multiple occasions and was unable to incorporate data from patients who may have also sought outside treatment due to the nature of primary care in the United States.

The method developed here is reliant on clinical diagnosis data, which limits the applicability of this model. Based on the coding practices of different physicians or hospitals, suspected diagnoses may be entered as a place holder until it can be confirmed or ruled out, as with initial diagnoses of multiple sclerosis [97]. Additionally, urinary tract infections can be diagnosed based on clinical symptoms, though a laboratory test and culture are important for confirming infections.

The results in here do not identify causality and, while this study attempts to control for a number of factors, those that can be controlled for are dependent on the dataset available. It is expected that patients who had a vaccine before the onset of COVID-19 might also expect different effects than those who had COVID-19 before having a vaccine. Establishment of causality would require recruiting patients specifically by their COVID-19 history and follow them in the long term with follow-up experiments to establish causality between SARS-CoV-2 infection and the clinical manifestation, such as scarring in the cardiac muscle or decreased insulin production. Finally, in identifying effects of COVID-19, we are limited by the novelty of the disease itself since other effects may take years or decades to develop.

Chapter 5: Conclusion

At the start of the COVID-19 pandemic, it was being touted as a "once in a century" type of event due to the time elapsed since the previous global public health situation of this nature, the Flu of 1918 [102]. However, as there have been other virus like SARS-CoV-2, such as H1N1, SARS, monkeypox, that have the potential of leading to a similar situation, situations of this nature might become more frequent. Should a situation like COVID-19 ever arise again, the projects in this thesis are an exemplar of how to identify beneficial treatments and how to utilize machine learning to identify long-term effects.

5.1 Identifying clinical and genetic factors affecting COVID-19 susceptibility, severity and mortality

The work presented in chapter 1 identified that patient's a history of macular degeneration and coagulation disorders were at increased risk for severe disease. The results of the analysis of the comorbidities concurred with the result from other labs. The genetic analysis presented in this study was the first to demonstrate that genetic variants were associated with altered risk for severe disease – the analysis identified variants associated with both increased and decreased risk.

5.2 Investigating steroid hormone exposure on outcome in intubated and mechanically ventilated COVID-19 patients

Our study implemented survival analysis to evaluate the effect of steroid hormone exposure on patients' survival following intubation and intubation with mechanical ventilation. In our data driven approach, we worked to identify intubation-extubation period visits for

patients based on clinical orders and utilized medication administration data to identify whether the patient in each intubation-extubation period was exposed to each drug we have data for before or after intubation. This method can be extrapolated to look at the effects exposure to other classes of drugs for COVID-19 as well as other procedures with a temporal aspect.

5.3 Identifying effects of COVID-19

Our study demonstrated a new method to conduct retrospective analyses for identifying the effects of COVID-19. By implementing a model trained on clinical data at the visit level and using the output from a random forest classifier as a probability instead of a binary outcome, we mitigated the need to definitively distinguish cases. Additionally, the results from our study can be used to direct further investigations into the effects of COVID-19. As the COVID-19 pandemic transitions to an endemic situation, our method can be utilized to understand potential pathophysiological difference in symptoms associated with COVID-19 spikes. Moreover, as this method was designed using concurrent clinical data, it can be adapted to other novel or emerging diseases.

References

1. World Health Organization. “COVID-19 China” Jan. 2020. https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229.

2. United States Centers for Disease Control. “CDC Museum COVID-19 Timeline” Jan. 2022. https://www.cdc.gov/museum/timeline/covid19.html

3. D. L. Roberts, J. S Rossman and I. Jaric, “Dating first cases of COVID-19” PLoS Pathogen vol. 17, no. 6, e1009620, Jun 2021. DOI: 10.1371/journal.ppat.1009620.

4. P. Zhou, X. Yang, X. Wang, B. Hu, L. Zhang, W. Zhang, H. Si, Y. Zhu, B. Li, C. Huang, H. Chen, J. Chen, Y. Luo, H. Guo, R. Jiang, M. Liu, Y. Chen, X. Shen, X. Wang, X. Zheng, K. Zhao, Q. Chen, F. Deng, L. Liu, B. Yan, F. Zhan, Y. Wang, G. Xiao and Z. Shi, “A pneumonia outbreak associated with a new coronavirus of probable bat origin” Nature vol. 579, pp. 270-273, Feb. 2020. DOI: 10.1038/s41586-020-2012-7.

5. F. Wu, S. Zhao, B. Yu, Y. Chen, W. Wang, Z. Song, Y. Hu, Z. Tao, J. Tian, Y. Pei, M. Yuan, Y. Zhang, F. Dai, Y. Liu, Q. Wang, J. Zheng, L. Xu, E. Holmes, and Y. Zhang, "A new coronavirus associated with human respiratory disease in China” Nature vol. 579, pp. 265-269, Feb. 2020. DOI: 10.1038/s41586-020-2008-3.

6. C. Calisher, D. Carroll, R. Colwell, R. Corley, P. Daszak, C. Drosten, L. Enjuanes, J. Farrar, H. Field, J. Golding, A. Gorbalenya, B. Haagmans, J. Hughes, W. Karesh, G. Keusch, S. Lam, J. Lubroth, J. Mackenzie, L. Madoff, J. Mazet, P. Palese, S. Perlman, L. Poon, B. Roizman, L. Saif, K. Subbarao and M. Turner, "Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19” The Lancet vol. 395, no. 10226, pp. E42-E43, Mar 2020. DOI: 10.1016/S0140-6736(20)30418-9.

7. G. Chavarria-Miró, E. Anfruns-Estrada, S Guix, M Paraira, B Galofré, G Sánchez, R. M. Pintó, A. Bosch, “Sentinel surveillance of SARS-CoV-2 in wastewater anticipates the occurrence of COVID-19 cases” medRχiv, Jun. 2022. DOI: 10.1101/2020.06.13.20129627.

8. B. Mueller and C. Zimmer. “G.O.P. Senator’s Report on Covid Origins Suggests Lab Leak, but Offers Little New Evidence” The New York Times October 27, 2022 https://www.nytimes.com/2022/10/27/science/covid-lab-leak-burr-report.html.

9. J. Pekar, A. Magee, E. Parker, N. Moshiri, K. Izhikevich, J. Havens, K. Gangavarapu, L. Malpica Serrano, A. Crits-Christoph, N. Matteson, M. Zeller, J. Levy, J. Wang, S. Hughes, J. Lee, H. Park, M. Park, K. Ching Zi Yan, R. Tzer Pin Lin, M. Mat Isa, Y. Muhammad Noor, T. Vasylyeva, R. Garry, E. Holmes, A. Rambaut, M. Suchard, K. Andersen, M. Worobey and J. Wertheim, "SARS-CoV-2 emergence very likely resulted from at least two zoonotic events" Zenodo Feb. 2022 DOI: 10.5281/zenodo.6342616.

10. G. Gao, W. Liu, P. Liu, W. Lei, Z. Jia, X. He, L. Liu, W. Shi, Y. Tan, S. Zou, X. Zhao, G. Wong, J. Wang, F. Wang, G. Wang, K. Qin, R. Gao, J. Zhang, M. Li, W. Xiao, Y. Guo, Z. Xu, Y. Zhao, J. Song, J. Zhang, W. Zhen, W. Zhou, B. Ye, J. Song, M. Yang, W. Zhou, Y. Bi1, K. Cai1, D. Wang, W. Tan1, J. Han, W. Xu and G. Wu, "Surveillance of SARS-CoV-2 in the environment and animal samples of the Huanan Seafood Market" Research Square Feb. 2022 DOI: 10.21203/rs.3.rs-1370392/v1.

11. M. Worobey, J. Levy, L. Serrano, A. Crits-Christoph, J. Pekar, S. Goldstein, A. Rasmussen, M. Kraemer, C. Newman, M. Koopmans, M. Suchard, J. Wertheim, P. Lemey, D. Robertson, R. Garry, E. Holmes, A. Rambaut and K. Andersen "The Huanan market was the epicenter of SARS-CoV-2 emergence" Zenodo Feb. 2022 DOI: 10.5281/zenodo.6299600.

12 C. Jackson, M. Farzan, B. Chen and H. Choe, "Mechanisms of SARS-CoV-2 entry into cells" Nature Reviews Molecular Cell Biology vol. 23, pp. 3-20, Oct 2021 DOI: 10.1038/s41580-021-00418-x.

13. G. Pascarella, A. Strumia, C. Piliego, F. Bruno, R. Del Buono, F. Costa, S. Scarlata, and F. E. Agrò, "COVID-19 diagnosis and management: a comprehensive review" Journal of Intern Medicine vol. 288, iss. 2, pp. 192-206, Aug. 2020 DOI: 10.1111/joim.13091.

14. L. J. Carter, L V. Garner, J. W. Smoot, Y. Li, Q. Zhou, C. J. Saveson, J. M. Sasso, A. C. Gregg, D. J. Soares, T. R. Beskid, S. R. Jervey, and C. Liu**,** "Assay Techniques and Test Development for COVID-19 Diagnosis" ACS Central Science vol. 6, iss. 5, pp. 591–605, May 2020 DOI: 10.1021/acscentsci.0c00501.

15. S. B. Stoecklin, P. Rolland, Y. Silue, A. Mailles, C. Campese, A. Simondon, M. Mechain, L. Meurice, M. Nguyen, C. Bassi, E. Yamani, S. Behillil, S. Ismael, D. Nguyen, D. Malvy, F. X. Lescure, S. Georges, C. Lazarus, A. Tabaï, M. Stempfelet, V. Enouf, B. Coignard, D. Levy-Bruhl and Investigation team, "First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020" Eurosurveillance vo;l. 25, iss. 6, pp. 2000094, Feb 2020 DOI: 10.2807/1560-7917.ES.2020.25.6.2000094.

16. University of Pennsylvania Medicine. "COVID-19 Testing Sites" Mar 2020 https://www.pennmedicine.org/coronavirus/testing-sites.

17. The New York Times. "Coronavirus in the U.S.: Latest Map and Case Count" Jul 2022 https://www.nytimes.com/interactive/2021/us/new-york-covid-cases.html.

18. University of Pennsylvania Medicine. “WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020” Mar 2020 https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020\.

19. E. Dong, H. Du and L. Gardner “An interactive web-based dashboard to track COVID-19 in real time” Lancet Infectious Diseases vol. 20, no. 5, pp. 533-534 May 2020 DOI: 10.1016/S1473-3099(20)30120-1.

20. J. Chen. “The United States badly bungled coronavirus testing—but things may soon improve” ScienceInsider February 28, 2020 https://www.science.org/content/article/united-states-badly-bungled-coronavirus-testing-things-may-soon-improve.

21. The Associated Press. “China delayed releasing coronavirus info, frustrating WHO” The Associated Press June 2, 2020 https://apnews.com/article/united-nations-health-ap-top-news-virus-outbreak-public-health-3c061794970661042b18d5aeaaed9fae.

22. M. L. Holshue, C DeBolt, S. Lindquist, K. H. Lofy, J. Wiesman, H. Bruce, C. Spitters, K. Ericson, S. Wilkerson, A. Tural, G. Diaz, A. Cohn and Washington State 2019-nCoV Case Investigation Team, “First Case of 2019 Novel Coronavirus in the United States” New England Journal of Medicine vol. 382, pp. 929-936, Mar 2020 DOI: 10.1056/NEJMoa2001191.

23. The New York Times. “Who’s on the U.S. Coronavirus Task Force” The New York Times February 29, 2020 https://www.nytimes.com/2020/02/29/health/Trump-coronavirus-taskforce.html

24. United States Centers for Medicare and Medicaid Services “CMS Takes Action Nationwide to Aggressively Respond to Coronavirus National Emergency” Mar 2020 https://www.cms.gov/newsroom/press-releases/cms-takes-action-nationwide-aggressively-respond-coronavirus-national-emergency.

25. T. Nguyen and C. Animashaun “How the coronavirus is disrupting US air travel, in 2 charts” Vox April 20, 2020 https://www.vox.com/the-goods/2020/4/20/21224080/coronavirus-air-travel-decline-charts.

26. United States Congressional Research Services “COVID-19: Federal Travel Restrictions and Quarantine Measures” May 2020 https://crsreports.congress.gov/product/pdf/LSB/LSB10415.

27. Aljazeera. “Coronavirus: Travel restrictions, border shutdowns by country” Aljazeera June 3, 2020 https://www.aljazeera.com/news/2020/6/3/coronavirus-travel-restrictions-border-shutdowns-by-country

28. J. Goldstein and J. McKinley. “Coronavirus in N.Y.: Manhattan Woman Is First Confirmed Case in State” The New York Times March 5, 2020 https://www.nytimes.com/2020/03/01/nyregion/new-york-coronvirus-confirmed.html.

29. A. Gonzalez-Reiche, , M. Hernandez, , M. Sullivan, , B. Alshammary, , A. Obla, , S. Fabre, , G. Kleiner, , J. Alburquerque, , A. Guchte, , J. Dutta, , N. Francoeur, , B. Oussenko, , G. Deikus, , J. Soto, , S. Sridhar, , Y. Wang, , K. Twyman, , A. Kasarskis, , D. Altman, , M. Smith, , R. Aberg, , F. Krammer, , A. García-Sastre, , M. Patel, , A. Paniz-Mondolfi, , M. Gitman, , E. Sordillo, V. Simon and H. van Bake, "Introductions and early spread of SARS-CoV-2 in the New York City area" Science vol. 369, no. 6501, pp. 297-301 May 2020 DOI: 10.1126/science.abc1917.

30. United States Centers for Disease Control. "Morbidity and Mortality Weekly Report (MMWR) COVID-19 Outbreak - New York City, February 29-June 1, 2020" Nov. 2020. https://www.cdc.gov/mmwr/volumes/69/wr/mm6946a2.htm.

31. The New York Times. "Coronavirus in the U.S.: Latest Map and Case Count" Jul 2022 https://www.nytimes.com/interactive/2021/us/new-york-covid-cases.html.

32. M. Wilson, "As Summer Wanes in N.Y.C., Anxiety Rises Over What Fall May Bring" The New York Times September 1, 2020 https://www.nytimes.com/2020/08/26/nyregion/coronavirus-fall-new-york.html.

33. New York City Office of the Mayor. "New York City to Close All School Buildings and Transition to Remote Learning" Mar. 2020. https://www1.nyc.gov/office-of-the-mayor/news/151-20/new-york-city-close-all-school-buildings-transition-remote-learning.

34. New York City Office of the Mayor. "Mayor de Blasio Issues New Guidance to New Yorkers" Mar. 2020. https://www1.nyc.gov/office-of-the-mayor/news/173-20/mayor-de-blasio-issues-new-guidance-new-yorkers.

35. New York State Office of the Governor. "Governor Cuomo Issues Guidance on Essential Services Under The 'New York State on PAUSE' Executive Order" Mar. 2020. https://www.governor.ny.gov/news/governor-cuomo-issues-guidance-essential-services-under-new-york-state-pause-executive-order.

36 P. E. Morange, P. Suchon and D. A. Tregouet, "Genetics of venous thrombosis: update in 2015" Thrombosis and Haemostasis vol114 pp. 910–919 Sep 2015 DOI: 10.1160/TH15-05-0410.

37 B. Zoller, "Genetics of venous thromboembolism revised" Blood vol. 134, pp. 1568–1570 Nov 2019 DOI: 10.1182/blood.2019002597.

38 E. J. Nascimento, A. M. Silva, M. T. Cordeiro, C. A. Brito, L. H. V. G. Gil, U. Braga-Neto and E. T. A. Marques, "Alternative complement pathway deregulation is correlated with dengue severity. PLoS ONE vol. 4, pp. e6782 (2009) DOI: 10.1371/journal.pone.0006782.

39 A. F.Pastor, L. R. Moura, J. W. D. Neto, E. J. M. Nascimento, C. E. Calzavara-Silva, A. L. V. Gomes, A. M. DaSilva,M. T. Cordeiro, U. Braga-Neto, S. Crovella, L. H. V. G. Gil, E. T. A. Marques Jr. and B. Acioli-Santos "Complement factor H gene (CFH) polymorphisms C-257T, G257A and haplotypes are associated with protection against severe dengue phenotype, possible related with high CFH expression" Human Immunology vol. 74, pp. 1225–1230 Sep 2013 DOI: 10.1016/j.humimm.2013.05.005.

40 A. M. Risitano, D. C. Mastellos, M. Huber-Lang, D. Yancopoulou, C. Garlanda, F. Ciceri and J. D. Lambris, "Complement as a target in COVID-19?." Nature Reviews Immunology vol. 20, pp. 343–344 Apr 2020 DOI: 10.1038/s41577-020-0320-7.

41 A. Ruggeri, A. M. Ristano, P. Angelillo, D. Yancopoulou, D. C. Mastellos, M. Huber-Lanf, S. Piemontese, A. Assaelli C. Garlanda, J. D. Lambris and F. Ciceri "The first case of COVID-19 treated with the complement C3 inhibitor AMY-101" Clinical Immunology vol. 215, pp. 108450 Jun 2020 DOI: 10.1016/j.clim.2020.108450.

42 F. C. G Polubriaginof, P. Ryan, H. Salmasian, A. W. Shapiro, A. Perotte, M. M. Safford, G. Hripcsak, S. Smith, N. P. Tatonetti and D. K. Vawdrey. "Challenges with quality of race and ethnicity data in observational databases" Journal of the American Medical Informatics Association vol. 26, pp. 730-736 Jul 2019 DOI: doi.org/10.1093/jamia/ocz113.

43 D. W. Hosmer, S. Lemeshow and S. May, Applied Survival Analysis: Regression Modeling of Time-to-Event Data. 2nd edn, Wiley-Interscience, 2008.

44 N30 Butler, D. J. et al. "Shotgun transcriptome and isothermal profiling of SARS-CoV-2 infection reveals unique host responses, viral diversification and drug interactions" bioRxiv May 2020 DOI: 10.1101/2020.04.20.048066.

45. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles" Proceedings of the National Academy of Science vol. 102, pp. 15545–15550 (2005) DOI: 10.1073/pnas.0506580102.

46. V. K. Mootha, C. M. Lindgren, K. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler and L. C. Groop. "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes" Nature Genetics vol. 34, pp. 267–273 Jul 2003 DOI: doi.org/10.1038/ng1180.

47. C. C. Chang, C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee "Second-generation PLINK: rising to the challenge of larger and richer datasets" GigaScience vol. 4, iss. 1 Feb 2015, DOI: 10.1186/s13742-015-0047-8.

48. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Baly and P. C. Sham. "PLINK: a toolset for whole-genome association and population-based linkage analysis" American Journal of Human Genetics vol. 81, no. 3, pp. 559-575 Sep 2007 DOI: 10.1086/519795.

49. G. Lasso, B. Honig and S. D. Shaira "A Sweep of Earth's Virome Reveals Host-Guided Viral Protein Structural Mimicry and Points to Determinants of Human Disease" Cell Systems vol. 12, pp. 82-91 Jan 2021 DOI: 10.1016/j.cels.2020.09.006.

50. V. Ramlall, P. M. Thangaraj, C. Meydan, J. Foox, D. Butler, J. Kim, B. May, J. K. de Freitas, B. S. Glickberg, C. E. Mason, N. P. Tatonetti and S. D. Shapira. "Immune complement and coagulation dysfunction in adverse outcomes of SARS-CoV-2 infection". Nature Medicine, vol. 26, pp. 1609–1615, Aug 2020. DOI: 10.1038/s41591-020-1021-2.

51. F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, L. Guan, Y. Wei, H. Li, X. Wu, J. Xu, S. Tu and Y. Zhang. "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study" The Lancet vol. 395, iss. 10229, pp. 1054-1062 Mar 2020 DOI: 10.1016/S0140-6736(20)30566-3.

52. N. L. Smith, et al. "Genetic Predictors of Fibrin D-Dimer Levels in Healthy Adults" Circulation vol. 123, iss. 17, pp.1864-1872, DOI: 10.1161/CIRCULATIONAHA.110.009480.

53. TEx Consortium "Genetic effects on gene expression across human tissues" Nature vol. 550, pp. 204–213 Oct 2017 DOI: 10.1038/nature24277.

54. A. A. Rehman, H. Ahsan and F. H. Khan "α-2-Macroglobulin: a physiological guardian" Journal of Cellular Physiology vol. 228, pp. 1665-1675 Oct 2012 DOI: 10.1002/jcp.24266.

55. M. Gary-Bobo, P. Nirde, A.Jeanjean, A. Morere and M. Garcia, "Mannose 6-phosphate receptor targeting and its applications in human diseases" Current Medicinal Chemistry vol.14, pp. 2945–2953 2007 DOI: 10.2174/092986707782794005.

56. D. Ermert, and A. M. Blom "C4b-binding protein: the good, the bad and the deadly. Novel functions of an old friend" Immunology Letters vol. 169, pp. 82–92 Jan 2016 DOI: 10.1016/j.imlet.2015.11.014.

57. S Jean, P Lee, P Hsueh "Treatment options for COVID-19: The reality and challenges" Journal of Microbiology, Immunology and Infection, vol. 53, no. 3, pp. 436-443 2020 DOI: 10.1016/j.jmii.2020.03.034.

58. A CTT-1 Study Group Members "Remdesivir for the treatment of Covid-19- Preliminary Report" New England Journal of Medicine vol. 383, pp.1813-1826 Nov. 2020 DOI: 10.1056/NEJMoa2007764.

59. J. Geleris, Y. Sun, J. Platt, J. Zucker, M. Baldwin, G. Hripcsak, A. Labella, D. K. Manson, C. Kubin, R. G. Barr, M. E. Sobieszczyk and N.W. Schluger "Observation Study of Hydroxychloroquine in Hospitalized Patients with Covid-19" New England Journal of Medicine vol. 282, pp. 2411-2418 2020 DOI: 10.1056/NEJMoa2012410.

60. C. D. Funk, C. Laferrière and A. Ardakani "Snapshot of the Global Race for Vaccines Targeting SARS-CoV2 and the COVID-19 Pandemic" Frontiers in Pharmacology vol 11 Jun 2020 DOI: 10.3389/fphar.2020.00937.

61. D. Boulware, M. Pullen, A. Bangdiwala, K. Pastick, S. Lofgren, E. Okafor, C. Skipper, A. Nascene, M. Nicol, M. Abassi, N. Engen, M. Cheng, D. LaBar, S. Lother, L. MacKenzie, G. Drobot, N. Marten, R. Zarychanski, L. Kelly, I. Schwartz, E. McDonald, R. Rajasingham, T. Lee, K. H. Hullsiek. "A Randomized Trial of Hydroxychloroquine as Postexposure Prophylaxis for Covid-19", New England Journal of Medicine vol. 383, pp. 517-525 Nov. 2020 DOI: 10.1056/NEJMoa2016638.

62. Moderna. "Moderna's Work on a COVID-9 Vaccine Candidate" Jun 2020 https://www.modernatx.com/modernas-work-potential-vaccine-against-covid-19

63. United States Centers for Disease Control. "Coronavirus Disease 2019 (COVID-19)" Mar 2022 https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

64. M. Lamamri, A. Chebbi, J. Mamane, S. Abbad, M. Munuzzolini, F. Sarfati and S. Legriel "Priapism in a patient with coronavirus disease 2019 (COVID-19): A case report" American Journal of Emergency Medicine vol. 39, pp.251, Jan 2021 DOI: 10.1016/j.ajem.2020.06.027.

65. D. Chakravarty, S. S. Nair, N. Hammouda, P. Ratnani, Y. Gharib, V. Wagaskar, N. Mohamed, D. Lundon, Z. Dovey, N. Kyprianou and A. K. Tewari "Sex differences in SARS-CoV-2 infection rates and the potential link to prostate cancer" Communications Biology vol. 3, pp. 374 Jul 2020 DOI: 10.1038/s42003-020-1088-9.

66. E. Fan, J. R. Beitler, L. Brochard, C. S. Calfee, N. D. Ferguson, A. S. Slutsky and D. Brodie "COVID-19-associated acute respiratory distress syndrome: is a different approach to management warranted?" Lancet Respiratory Medicine vol. 8, no. 8, pp. 816-821 DOI: 10.1016/S2213-2600(20)30304-0.

67. D. Grady D. "Covid-19 Patient Gets Double Lung Transplant, Offering Hope for Others" The New York Times Jun. 2020 https://www.nytimes.com/2020/06/11/health/coronavirus-lung-transplant.html.

68. R. Li, C. Rivers, Q. Tan, M. B. Murray, E. Toner, M. Lipsitch. "Estimated Demand for US Hospital Impatient and Intensive Care Unit Beds for Patients with COVID-19 Based on Comparisons with Wuhan and Guangzhou, China" Journal of the American Medical Association Network Open Access vol. 3, no. 5, pp. e208297 May 2020 DOI: 10.1001/jamanetworkopen.2020.8297.

69. The RECOVERY Collaborative Group “Dexamethasone in Hospitalized Patients with Covid-19 - Preliminary Report” New England Journal of Medicine vol. 384, no. 8, pp. 693-704 Feb 2021 DOI: 10.1056/NEJMoa2021436.

70. GLUCOCOVID investigators “Methylprednisolone in adults hospitalized with COVID-19 pneumonia”, Wiener klinische Wochenschrift vol. 133, pp. 303–311 Feb 2021 DOI: 10.1007/s00508-020-01805-8.

71. P. Gupta, D Jethava, R. Choudhary and D. D. Jethava “Role of melatonin in attenuation of haemodynamic responses to laryngoscopy and intubation” Indian Journal of Anaesthesia vol. 60, no. 10, pp. 712-718 Oct 2016 DOI: 10.4103/0019-5049.191667.

72. M. Dianatkhah, A. Najafi, M. Sharifzadeh, A. Ahmadi, H. Sharifnia, M. Mojtahedzadeh, F. Najmeddin, and A. Moghaddas “Melatonin Supplementation May Improve the Outcome of Patients with Hemorrhagic Stroke in the Intensive Care Unit” Journal of Research in Pharmacy Practice vol. 6, no. 3, pp. 173-177 Sep. 2017 DOI: 10.4103/jrpp.JRPP_17_49.

73. A. Tarocco, N. Caroccia, G. Morciano, M. R. Wieckowski, G. Ancora, G. Garani, and P. Pinton “Melatonin as a master regulator of cell death and inflammation: molecular mechanisms and clinical implications for newborn care” Cell Death Disease vol. 10, no. 4, pp. 317 Apr. 2019 DOI: 10.1038/s41419-019-1556-7.

74. L. van Dorp, M. Acman, D. Richard, L. P. Shaw, C. E. Ford, L. Ormond, C. J. Owen, J. Pang, C. C. S. Tan, F. A. T. Boshier, A. T. Ortiz and F. Balloux “Emergence of genomic diversity and recurrent mutations in SARS-CoV-2” Infections, Genetics and Evolution vol. 83, no. 104351 Sep 2020 DOI: 10.1016/j.meegid.2020.104351.

75. World Health Organization. “Coronavirus disease (COVID-19) pandemic” Jul 2022. https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov

76. W. He, L. Chen. L. Chen. G. Yuan, Y. Fang, W. Chen, D. Wu, B. Liang, X. Lu, Y. Ma, L. Li, H. Wang, Z. Chen Q. Li, and R. P. Gale “COVID-19 in person with haematological cancers” Leukemia vol. 34, pp. 1637-1645 Apr 2020 DOI: 10.1038/s41375-020-0836-7.

77. S. R. Bornstein, R. Dalan, D. Hopkins, G. Mingrone, and B. O. Boehm “Endocrine and metabolic linked to coronavirus infection” Nature Reviews Endocrinology vol. 16, pp. 297-298 Jun 2020 DOI: 10.1038/s41574-020-0353-9.

78. WHO Solidarity Trial Consortium “Remdesivir and three other drugs for hospitalized patients with COVID-19: final results of the WHO Solidarity randomised trial and updated meta-analyses” Lancet vol. 399, no. 10339, pp. 1941-53 May 2022 DOI: 10.1016/S0140-6736(22)00519-0.

79. R. J. Reiter, R. Sharma, F. Simko, A. Dominguez-Rodriguez, J. Tesarik, R. L. Neel, A. T. Slominski, K. Kleszczynski, V. M. Martin-Gimenez, W. Manucha and D. P. Cardinali "Melatonin: highlighting its use as potential treatment for SARS-CoV-2 infection" Cellular and Molecular Life Science vol. 79, pp.143. Feb 2022 DOI: 10.1007/s00018-021-04102-3.

80. Oxford COVID Vaccine Trial Group "Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK" Lancet vol. 397, no. 10269, pp. 99-111 Jan 2021 DOI: 10.1016/S0140-6736(20)32661-1.

81. COVE Study Group "Efficacy and safety go the mRA-1273 SARS-CoV-2 vaccine" New England Journal of Medicine vol. 384, pp. 403-416, Feb 2021 DOI: 10.1056/NEJMoa2035389.

82. C4591001 Clinical Trial Group "Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine" New England Journal of Medicine vol. 284, pp. 2603-2615 Dec 2020 DOI: 10.1056/NEJMoa2034577.

83. S. Lopez-Leon, T. Wegman-Ostrosky, C. Perelman, R. Sepulveda, P. A. Rebolledo, A. Cuapio and S. Villapol. "More than 50 long-term effects of COVID-19: a systematic review and meta-analysis" Scientific Reports vol. 11, pp.16144 Aug 2021 DOI: 10.1038/s41598-021-95565-8.

84. I. Katsoularis, O. Fonseca-Rodriguez, P. Farrington, K. Lindmark and A. F. Connolly "Risk of acute myocardial infarction and ischaemic stroke following COVID-19 in Sweden: a self-controlled case series and matched cohort study" Lancet vol. 398, no. 10300, pp. 599-607 Jul 2021 DOI: 10.1016/S0140-6736(21)00896-5.

85. Y. Xie, E. Xu and Z. Al-Aly "Risk of mental health outcomes in people with covid-19: cohort study" British Medical Journal vol 276, pp. e068993 Feb 2022 DOI: 10.1136/bmj-2021-068993.

86. Xie Y and Al-Aly Z "Risks and burdens of incident diabetes in long COVID: a cohort study" Lancet Diabetes Endocrinology vol. 10, no. 5, pp. 311-321 May 2022 DOI: 10.1016/S2213-8587(22)00044-4.

87. Y. P. See, B. E. Young, L. W. Ang, X. Y. Ooi, C. P. Chan, W. L. Looi, S. C. Yeo and D. C. Lye "Risk factors for development of acute kidney injury in COVID-19 patients: a retrospective observational cohort study" Nephron vol. 145, no. 3, pp. 256-264 May 2021 DOI: 10.1159/000514064.

88. J. Chen, J. Wu, S. Hao, M. Yang, X. Lu, X. Chen and L. Li "Long term outcomes in survivors of epidemic Influenza A (H7N9) virus infection" Scientific Reports vol. 7, pp. 17275 Dec 2017 DOI: 10.1038/s41598-017-17497-6.

89. RVA Study Group. “Long-term outcomes of pandemic 2009 influenza A(H1N1)-Associated Severe ARDS” Chest vol.142, no. 3, pp. 583-592 Sep 2017
DOI: 10.1378/chest.11-2196.

90. H. F. Tseng, N. Smith, R. Harpaz, S. R. Bialek, L. S. Sy and S. J. Jacobsen “Herpes zoster vaccine in older adults and the risk of subsequent herpes zoster disease” Journal of the American Medical Association vol. 305, no. 2, pp.160-166 Jan 2011
DOI: 10.1001/jama.2010.1983.

91. United States Centers for Disease Control and Prevention “Shingles (Herpes Zoster)” Oct 2020 https://www.cdc.gov/shingles/hcp/clinical-overview.html

92. F. X. Bosch, A. Lorincz, N. Munoz, C. J. L. M. Meijer and K. V. Shah “The causal relation between human papillomavirus and cervical cancer” Journal of Clinical Pathology vol. 55, no. 4, pp. 244-265 Apr 2002 DOI: 10.1136/jcp.55.4.244.

93. J. Lei, A. Ploner, K. M. Elfstorm, J. Wang, A. Roth, F. Fang, K. Sundstorm, J. Dillner and P. Sparen “HPV vaccination and the risk of invasive cervical cancer” New England Journal of Medicine vol. 383, pp: 1340-1384 Oct 2020 DOI: 10.1056/NEJMoa1917338.

94. D. M. Parkin and F. Bray “Chapter 2: The burden of HPV-related cancers” Vaccine vol. 24, no. 3, pp. S11-S25. Aug 2006 DOI: 10.1016/j.vaccine.2006.05.111.

95. T. Lanz, , R. Brewer, , P. Ho, , J. Moon, , K. Jude, , D. Fernandez, , R. Fernandes, , A. Gomez, , G. Nadj, , C. Bartley, , R. Schubert, , I. Hawes, , S. Vazquez, , M. Iyer, , J. Zuchero, , B. Teegen, , J. Dunn, , C. Lock, , L. Kipp, , V. Cotham, , B. Ueberheide, , B. Aftab, , M. Anderson, , J. DeRisi, , M. Wilson, , R. Bashford-Rogers, , M. Platten, , K. Garcia, , L. Robinson “Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM” Nature vol. 603, pp. 321-327 Jan 2022 DOI: 10.1038/s41586-022-04432-7.

96. K. Bjornevik, M. Cortese, B. C. Healy, J. Kuhle, M. J. Mina, Y. Leng, S. J. Elldge, D. W. Niebuhr, A. I. Scher, K. L. Munger and A. Ascherio “Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis” Science vol. 375, no. 6578, pp. 296-301 Jan 2022. DOI: 10.1126/science.abj8222.

97. Cleveland Clinic “Multiple Sclerosis (MS)” Feb 2021 https://my.clevelandclinic.org/health/diseases/17248-multiple-sclerosis.

98. United States Centers for Disease Control and Prevention “New ICD-10-CM code of the 2019 Novel Coronavirus (COVID-19)” March 2020
https://www.cdc.gov/nchs/data/icd/Announcement-New- ICD-code-for-coronavirus-3-18-2020.pdf.

99. United States Centers for Disease Control and Prevention “Disparities in COVID-19 Illness” Jul 2022 https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/racial-ethnic-disparities/increased-risk-illness.html.

100. R. A. Oppel Jr, R. Gebeloff, K. K. R. Lai, W. Wright and M. Smith. "The fullest look yet at the radical inquiry of coronavirus" The New York Times Jul 2020 https://www.nytimes.com/interactive/2020/07/05/us/coronavirus-latinos-african- americans-cdc-data.html.

101. The New York Times "Tracking Coronavirus in New York City, N.Y.: Latest Map and Case Count" July 2022 https://www.nytimes.com/interactive/2021/us/new-york-city-new-york-covid- cases.html.

102. B. Gates "Responding to Covid-19 — A Once-in-a-Century Pandemic?" New England Journal of Medicine vol. 382, pp: 1677-1679 Apr 2020 DOI: 10.1056/NEJMp2003762.

Appendix A

Due to the physical size of some of the full tables containing the results for Chapter 4, only the data for selected diseases is presented in this thesis. The tables are presented in full in a Github repository located at: https://github.com/vijendra-cuimc/thesis. For each table in the repository, refer to the below information on understanding the format.

The file named "table_4.3_diag_formatted_git.csv" contain the full data from which a subset is presented in Table 3.3 printed in this thesis. The first non-header row lists the visit counts in each category, the second list the number of patients in each category. All subsequent rows list the number of visits (N_X) and the percent of visits (%_X) where that ICD10 category level diagnosis code was listed. Of note, X indicates a chapter level ICD10 code. The second column (negative_training_set) identifies non-COVID-19 visits used in model training, the third (positive_training_set) identifies COVID-19 visits used in model training, the fourth column (negative_eval_set) identifies non-COVID-19 visits used in model training, the fifth (positive_eval_set) identifies COVID-19 visits used in model training and the sixth (all_visits) identifies all visits between February 2020 and March 2022 for which demographics data is available.

The file name "table_4.4_all_features.csv" contains the full data from which a subset is present in Table 3.4 printed in this thesis. The first column (feature) lists all the feature in the model, the second column (importance) list the Gini importance as outputted by the function, the

third column (wasserstein_distance_training) lists the Wasserstein distance between distributions where the feature is observed and where the feature is not observed in the model training set, the fourth column (wasserstein_distance_eval) lists the Wasserstein distance between distributions where the feature is observed and feature is not observed in the model evaluation set and the fifth column (wasserstein_distance_all_visits) lists the Wasserstein distance between distributions where feature is observed and feature is not observed for all visits between February 2020 and March 2022 for which demographics data is available.

The file name "table_4.5_4.6_mannwhitney_results_git_formatted.csv" contains the full data from which a subset is present in tables 4.5 and 4.6 printed in this thesis. The first column (category) lists the category of the specific phenotype, the second column (phenotype) lists the name of the phenotype and the third column (phe_code) list the PheCode used to identify the phenotype. The fourth, seventh, tenth, 13th, 16th, 19th, 22nd and 25th columns (all_X_mwu_stat) lists the Mann-Whitney U test statistic comparing between distribution of probabilities of visits followed by the phenotype and not followed by the phenotype. The fifth, eight, 11th, 14th, 17th, 20th, 23rd and 26th columns (all_X_pvalue) list the p-value from the Mann-Whitney U test comparison listed in the preceding column. The sixth, ninth, 12th, 15th, 18th, 21st, 24th and 27th columns (all_X_c_pvalue) list the corrected p-value from the Mann-Whitney U test comparison listed in the two columns present using a false discovery Benjamini-Hochbergs calculation. The 28th, 31st, 34th, 37th, 40th, 43rd, 46th and 49th columns (new_X_mwu_stat) lists the Mann-Whitney U test statistic comparing between distribution of probabilities of visits followed by the phenotype and not followed by the phenotype for new phenotypes. The 29th, 32nd, 35th, 38th, 41st, 44th, 47th and 50th columns (new_X_pvalue) list the p-value from the Mann-Whitney U test comparison listed in the preceding column. The 30th, 33rd, 36th, 39th, 42nd, 45th, 48th and 51st

columns (new_X_c_pvalue) list the corrected p-value from the Mann-Whitney U test comparison listed in the two columns present using a false discovery Benjamini-Hochbergs calculation. Of note, X indicates the time period being considered in each test, 7 days, 14 days, 21 days, 28 days, 3 months, 6 months, 9 months and 1 year respectively.

The file named "table_4.7_coxph_results_git_formatted.csv" contain the full data from which a subset is presented in Table 3.7 printed in this thesis. The first column (category) lists the category of the specific phenotype, the second column (phenotype) lists the name of the phenotype and the third column (phe_code) list the PheCode used to identify the phenotype. The fourth column (all_1_year_case_n) lists the number of cases of the phenotype irrespective of the previous health history of the patient occurring within 1 year, the fifth column (all_1_year_noncase_n) lists the number of non-cases of the phenotype irrespective of the previous health history of the patient occurring within 1 year and the sixth column (all_1_year_hazards ratio) lists the Cox Proportional hazards ratio for the phenotype irrespective of the previous health history of the patient occurring within 1 year. The seventh and eight columns (all_1_year_95% confidence interval (lower), all_1_year_95% confidence interval (upper)) list the lower and upper, respectively, 95% confidence interval for the hazard ratio in the preceding columns and the ninth column (all_1_year_pvalue) list the *p*-value of the hazard ratio.

The tenth column (new_1_year_case_n) lists the number of cases of the phenotype when accounting for the previous health history of the patient occurring within 1 year, the 11th column (new_1_year_noncase_n) lists the number of non-cases of the phenotype when accounting for the previous health history of the patient occurring within 1 year and the 12th column (new_1_year_hazards ratio) lists the Cox Proportional hazards ratio for the phenotype when accounting for the previous health history of the patient occurring within 1 year. The 13th and

14th columns (new_1_year_95% confidence interval (lower), new_1_year_95% confidence interval (upper)) list the lower and upper, respectively, 95% confidence interval for the hazard ratio in the preceding columns and the 15th column (new_1_year_pvalue) list the *p*-value of the hazard ratio.