

ABSTRACT

MELE, JESSICA A. Data Analytics and Decision Making: Evaluating Risk and Burden Associated with Infectious Respiratory Diseases. (Under the direction of Julie Swann and Osman Ozaltin.)

Infectious respiratory diseases cause substantial burden on healthcare systems in the United States each year. In recent years, two of the most notable infectious respiratory diseases have been pneumonia and COVID-19. Recent studies suggest that pneumonia causes 1.5 million unique hospitalizations in the United States each year whereby COVID-19 is estimated to have caused 4.6 million hospitalizations nationally since it first emerged in 2020. In addition to the increased risk of severe disease and mortality, pneumonia and COVID-19 both incur substantial healthcare-associated costs. Understanding risk factors and disease characteristics remain critical components to improving health outcomes and reducing overall burden associated with these two diseases.

This work explores the use of data analytics to develop tools to overcome modern challenges surrounding computational modeling and big data in healthcare with the intention to provide interpretable solutions to inform public health. The proposed models are developed to be used as tools to improve population health and facilitate intervention design to reduce severe outcomes for pneumonia and COVID-19. This work proposes new ideas and developments to guide public health policy and draws insights that result from applying these methods to real-world data.

The ideas presented in this work expand upon current literature surrounding infectious respiratory disease risk, prevention, and prevalence, and provides tools to: (i) identify at-risk populations; (ii) recommend targeted intervention policies; and (iii) evaluate and understand disease burden. First, risk will be explored via the development of three predictive models to understand and evaluate 30-day risk factors for pneumonia hospitalization at the individual, community, and provider levels. Next, data-driven models will be employed to provide recommendations for targeted intervention policies to reduce hospitalizations and healthcare-associated costs. Lastly, disease progression models will be developed to inform population susceptibility, understand historical risk, and evaluate mitigation tactics associated with COVID-19.

© Copyright 2022 by Jessica A. Mele

All Rights Reserved

Data Analytics and Decision Making:
Evaluating Risk and Burden Associated with Infectious Respiratory Diseases

by
Jessica A. Mele

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Operations Research

Raleigh, North Carolina

2022

APPROVED BY:

Maria Mayorga

Min Chi

Julie Swann
Co-chair of Advisory Committee

Osman Ozaltin
Co-chair of Advisory Committee

DEDICATION

To Vincent, Joseph, and Nora, I love you.

To my parents, for always supporting me.

To Gianna, Christopher, Koral, and Heather because you're the only reason I made it this far.

To my grandparents, for all that you've done.

BIOGRAPHY

Jessica Mele was born and raised on Long Island, New York where she attended Malverne High School. She began her professional studies at the City College of New York where she discovered her interest in Applied Mathematics, which led to her transfer to Stony Brook University. She realized her passion to conduct research during her summer internship at Brookhaven National Laboratory, where she worked with world-renowned physicists and applied researchers for two summers. She then continued her professional education at North Carolina State University, pursuing a Masters in Applied Mathematics en route to her doctoral degree in Operations Research.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors Dr. Julie Swann and Dr. Osman Ozaltin. This work would not have been possible without their continuous effort, support, and thoughtful insights. I would like thank them for the impact they both have had on my professional development and my growth as an independent researcher and of course, for their patience and for providing me with the opportunity to work on such meaningful projects. I also would like to thank my committee members, Dr. Maria Mayorga and Dr. Min Chi for their guidance, effort, and suggestions towards developing this dissertation and for all of their time and consideration throughout this process.

Secondly, I want to thank every faculty member and student who contributed to the ideas and development of certain areas of this work. I would like to say thank you to Dr. Sara Shaashani, Kimia Vahdat, Lingchao Mao, Justin Lancaster, and others at NC State as well as Dr. Joseph Agor and other students from Oregon State University for their contributions. Furthermore, I would like to give thanks to the faculty and student members of the COVSIM modeling team Dr. Julie Ivy, Dr. Mayorga, Dr. Swann, Erik Rosenstrom, Yiwei Zhang, and others from UNC Chapel Hill and East Carolina University for all of their assistance and continued efforts to enable our work to have an impact during the COVID-19 pandemic.

Last, but not least, I want to thank my family and friends for their continued support throughout this journey. I would not have enjoyed the successes of this degree without the constant love and support from my family to keep me grounded. I also would like to thank my friends who helped get me through qualifying exams, Vishwaraj Doshi, Marwen Zrida, and Matthew Fletcher, and my office mates Gregory Hauser, Margaret Tobey, Danika Dorris, Yiwei Zhang, Erik Rosenstrom, Cameron Lisy, Gimantha Perera, and others from the health systems office, and other good friends, Morgan DiCarlo, and Lee Jones, who have become my support system away from home. Completing this journey would not have been possible without all of their help and the added enjoyment of having a supportive group to celebrate all of our successes along the way.

TABLE OF CONTENTS

LIST OF TABLES	vi
LIST OF FIGURES	ix
Chapter 1 Introduction	1
1.1 Data Analytics and Medical Decision Making	1
1.2 Infectious Diseases in the United States	3
1.3 Objectives and Organization	7
Chapter 2 Predictive Modeling to Assess 30-Day Risk of Unplanned Pneumonia Related Hospitalization	9
2.1 Introduction	9
2.2 Literature Review	10
2.3 Methods	12
2.4 Results	24
2.5 Discussion	27
2.6 Conclusion	32
Chapter 3 Data-driven Interventions to Prevent Pneumonia Hospitalizations	34
3.1 Introduction	34
3.2 Literature Review	36
3.3 Model Formulation	37
3.4 Solution Algorithm	44
3.5 Numerical Experiments	47
3.6 Conclusion	51
Chapter 4 Methods to Estimate True Disease Burden of COVID-19	55
4.1 Introduction	55
4.2 Literature Review	57
4.3 Methods	58
4.4 Results	71
4.5 Discussion	76
4.6 Conclusion	77
Chapter 5 Discussion	80
BIBLIOGRAPHY	85
APPENDICES	100
Appendix A Overview of Computational Models	101
Appendix B Appendix to Chapter 2	108
B.0.1 Variable Coefficients and Feature Importance	132
Appendix C Appendix to Chapter 3	151
C.1 Derivations for Optimization Model	151
Appendix D Appendix to Chapter 4	155

LIST OF TABLES

Table 2.1	This demographic table illustrates the composition of each patient-row group, (i) event or (ii) non-event group for this study. We report whether the two groups have statistically significant differences in counts of each group, according to a two-sided proportions z-test with $\alpha = 0.05$ (*), $\alpha = 0.01$ (**).	15
Table 2.2	CCS Category groupings and corresponding diagnoses used to compare the event and non-event rows in the study population in Table 2.1	16
Table 2.3	Descriptive statistics comparing the event and non-event patient rows of the patient-rows that were filtered out, or excluded, using the above filtering criteria by recent medical history.	18
Table 2.4	List of diagnosis codes provided by AHRQ to compute the numerator for PQI-11 Bacterial Pneumonia Admission Rates. "Nos" refers to not-otherwise specified.	19
Table 2.5	Collection of the top 60 variables, e.g., top 20 by odds ratio, and feature importance for logistic regression and the ensemble methods, respectively. The type indicates "Ind." for individual, "Com." for community, and "Prov." for provider level variables. The coefficient shown is the mean coefficient from the logistic regression model across 5-CV folds where each feature in this table was assigned a nonzero coefficient across all 5 folds using 'l1-regularization'. Last, the model number indicates whether the variable was a part of the top 20 for 1:logistic regression, 2:XG boost, or 3:random forest. Confidence intervals are not generated due to the regularization term used in model training.	25
Table 2.6	Mean value and standard deviation (-) for each metric across all 5 validation testing sets.	28
Table 2.7	Mean and standard deviation for the out of sample testing set across 5 CV-folds for each model trained in the study, compared to the optimal model presented in Uematsu, et al. in 2017.	28
Table 3.1	Testing set evaluation for the mean number of high and low cost interventions, I_1 and I_2 , when using the best x_1, x_2 strategy, along with 95% confidence intervals over 30 replications.	51
Table 3.2	Testing set evaluation for the mean percent savings and 95% CIs compared to doing nothing, all low cost interventions, and all high cost interventions when using the best x_1, x_2 strategy, along with 95% confidence intervals over 30 replications.	52
Table 3.3	Recommended policies for each C_{i1}, p_{e1} pair. Here we show the average value and 95% confidence interval of best x_1, x_2 values for each C_{i1}, p_{e1} pair, over 30 replications.	53
Table 4.1	Estimated percent of total infections by deterministic IFR analysis and corresponding resulting NC-IFR from composition of estimated infections.	59

Table 4.2	Parameter descriptions for the inputs to the state level SEIR model, split into types, e.g., time (days), rates, and fitting parameters. The bounds correspond to the upper and lower bounds used as inputs to create the initial Latin Hypercube Sample. All parameter bounds were determined via literature estimates and model validation.	63
Table 4.3	The allowable percent change (reduction) in disease progression parameters that were assumed to be reduced by available treatments in 2020 assigned by the respective fitting periods the treatments became available.	63
Table 4.4	North Carolina final estimates for the top five trajectories chosen by the end of the final fitting period. Here we report the state level estimates for IFR, cumulative percent infected in 2020, and overall lab multiplier when comparing total cumulative simulation infections and cumulative lab-confirmed cases reported by the New York Times.	73
Table A.1	Visualizing the elements of a confusion matrix.[87]	105
Table B.1	ICD-9 codes for AHRQ definition of PQI-11 inpatient visit [2, 46]	109
Table B.2	Admission codes for transfers [2]	110
Table B.3	Sick cell anemia or HB-S disease diagnosis codes. [2, 46]	110
Table B.4	Immunocompromised State Diagnosis Codes Set 1. [2, 46]	111
Table B.5	Immunocompromised State Diagnosis Codes Set 2. [2, 46]	112
Table B.6	Immunocompromised State Diagnosis Codes Set 3. [2, 46]	113
Table B.7	Immunocompromised State Diagnosis Codes Set 4. [2, 46]	114
Table B.8	Immunocompromised State Procedure Codes. [2, 46]	115
Table B.9	List of ICD-9 Codes for influenza like illnesses.	116
Table B.10	ICD-9 and CPT/HCPCS codes for chest x-rays.	118
Table B.11	List of ICD-9 and CPT/HCPCS codes to identify someone with a medical history of smoking.	119
Table B.12	List of ICD-9 and CPT codes for the administration of an influenza vaccine.	119
Table B.13	CPT and HCPCS codes for the administration of the pneumococcal vaccines.	120
Table B.14	ICD-9 codes for attention deficit and conduct disorders.	120
Table B.15	ICD-9 codes for bronchitis.	121
Table B.16	ICD-9 codes for pneumonia.	121
Table B.17	ICD-9 codes for fatigue.	122
Table B.18	ICD-9 codes for examinations, evaluations, and screenings.	122
Table B.19	ICD-9 codes for chemotherapy, immunotherapy, and radiotherapy.	122
Table B.20	ICD-9 codes for cystic fibrosis.	123
Table B.21	ICD-9 codes for lung disorders and complications due to external agents.	123
Table B.22	ICD-9 codes for skull injuries or fractures.	124
Table B.23	ICD-9 codes for a medical history of obesity.	124
Table B.24	ICD-9 codes for previous illnesses of the respiratory tract[**].	125
Table B.25	Provider speciality codes for cardiology and oncology.	126
Table B.26	ICD-9, CPT, and HCPCS codes for nursing home visits.	127
Table B.27	ICD-9 codes and related frailty coefficients for diagnosis components of frailty computation.	129
Table B.28	CPT and HCPCS codes and related frailty coefficients for diagnosis components of frailty computation.	130

Table D.1 Resulting population-weighted IFR values, by age, for the United States from each study. 155

LIST OF FIGURES

Figure 2.1	This flowchart illustrates the inclusion and exclusion criteria for beneficiaries to be included in our study. The values on the left-hand side of the diagram represent the number of beneficiaries that are excluded during each stage in the process in response to the question listed immediately to the right and with the response listed directly above the connective arrow.	13
Figure 2.2	Filtering criteria to gather beneficiaries based on their medical history, which is applied using an inclusive 'OR' statement to create our prediction set.	17
Figure 2.3	Sampling algorithm for training sets.	21
Figure 2.4	Mean logistic regression coefficients identified by at least two of the three models in the post prediction analysis. All variables identified in this analysis were assigned a nonzero logistic regression coefficient across all five folds.	24
Figure 2.5	Histogram of the logistic regression prediction probability distribution for all patient rows.	26
Figure 2.6	ROC and Precision Recall Curve for the out of sample testing set predictions for each CV fold and each model in the study.	29
Figure 3.1	Risk score distribution and associated hospitalization probability. Here, the yellow block would indicate no interventions, whereby the two red blocks could indicate low and high cost interventions for the left and right red blocks, respectively. The arrows identifies the single period decision point for a given x_1 value.	37
Figure 3.2	Flowchart of psuedo-code for the heuristic to determine the best intervention strategy.	45
Figure 3.3	Cumulative number of hospitalizations versus cumulative number of beneficiaries in a sampled training set (dashed) and approximated quadratic, convex curve (solid) for a selection of x_1 thresholds. Here, the top graph is for when $x_1 = 0$	46
Figure 3.4	[a-d] The average total cost, over 30 replications, on the training data for the recommended strategy for each x_1 value, compared to the 95% confidence interval for the cost of doing nothing (horizontal, solid red line) over 30 replications, ordered by increasing x_1 threshold for varying costs of I_1 . The red dots represent the optimal, recommended strategy over each $C_{i1}, p e_1$ pair for I_1 and include a 95% confidence interval.	49
Figure 4.1	Compartmental model showing the seven disease states considered in this analysis.	61
Figure 4.2	Timeline of nationally available, effective treatments in the United States and the corresponding fitting periods M_i that result for this analysis.	64
Figure 4.3	Overview of weekly fitting to percent of infections that are documented and the transmission rate. The number of people in each disease state, connects each week during a fitting period and each fitting period through updating the y_0 term for each SEIR model.	68

Figure 4.4	Overview of branching algorithm for each fitting period identified by the available treatments in 2020. The fitting periods are connected through a branching procedure where the top disease progression parameter samples, chosen through the efficient frontier, are used to create updated bounds for a new Latin Hypercube (LHC) sample of size 50. The first fitting period is fed with approximately 500 samples described by the bounds in Table 4.2. In the final fitting period, the top samples by death mean squared error are chosen to identify the best trajectories to estimate disease burden.	69
Figure 4.5	These figures illustrate the identified effective frontier for March (a) and April (b) using the branching algorithm. Each point in the plot represents a selection of disease progression parameters for each fitting period. The red samples indicate the effective frontier, whereas the yellow indicate the chosen samples to feed the next fitting period.	70
Figure 4.6	North Carolina cumulative percent infected by (a) age: 0-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80+ years of age and (b) geographic location: urban, suburban, and rural counties, as estimated for 2020 from the simple, deterministic application of IFR values to the NY Times reported deaths. . . .	71
Figure 4.7	North Carolina model fit for estimated daily lab-confirmed cases (a), and daily deaths (b), compared to the 7-day averages reported by the New York Times. Panel (c) shows the estimated lab multiplier over time.	72
Figure 4.8	Comparing branching SEIR estimates for (a) daily lab-confirmed cases, (b) time varying effective reproductive number, and (c) daily confirmed COVID inpatient admissions with other approaches. The branching SEIR model trajectories are provided by the black lines, where other studies have colors indicated in the legend. The United States Department of Health and Human Services (HHS) did not begin reported COVID inpatient admissions until July 2020.	74
Figure 4.9	North Carolina disease progression parameters for top five trajectories chosen by mean squared death error at the end of the final fitting period.	75
Figure B.1	Histogram of the random forest prediction score distributions for all patient rows.	149
Figure B.2	Histogram of the XG boost prediction score distributions for all patient rows.	149
Figure B.3	Hospitalization rates among patient groups that were included in the prediction set because they had at least one of each condition, e.g., Nursing Home (0) indicates that patients in this group could have been included due to a recent nursing home visit, among other conditions.	150
Figure C.1	Sampling distribution of the sampled training data truncated with risk scores on or above x_1 , showing the number of hospitalizations in the set. This distribution is used to approximate $F(q)$. The dashed line represents the actual number of hospitalizations in the testing set with risk scores on or above x_1	154
Figure D.1	Infection Fatality Ratios presented from empirical data from China and six regions in Europe, stratified by age, for the early months (January-May 2020) of the COVID-19 pandemic, and the comparison to other studies and a developed meta regression IFR set by age group.	156

Figure D.2	North Carolina estimated lab multipliers by age group using the IFR values and 95% confidence intervals from [104]	156
Figure D.3	North Carolina lab multiplier estimates by age and geographic location from IFR analysis.	157
Figure D.4	Comparing trends between the weekly branching SEIR estimates for time varying lab multiplier and percent positive tests. The percent positive tests is estimated by the 7-day average of daily lab-confirmed cases by the New York Times over the 7-day average of daily total tests administered collected by John's Hopkins University and availability in the JHU CCI data repository. The legend indicates each trajectory's correlation between weekly lab multipliers and percent positivity, computed using the Spearman correlation test, and the associated p-value in parenthesis.	159
Figure D.5	North Carolina lab-confirmed cases versus average daily mobility (Google mobility reports)	160
Figure D.6	Results from clustering on North Carolina data.	160

CHAPTER 1

INTRODUCTION

1.1 Data Analytics and Medical Decision Making

Description of Big Data

The term “big data” has many definitions, but most agree that the concept centralizes about the fact that the data in question is large enough to be unmanageable using traditional software [57]. Big data is often described through a characterization of what’s known as the Vs: (i) Volume, (ii) Velocity, (iii) Variety, and (iv) Veracity [16, 57, 74], with the list still expanding as more information is gathered on big data. To give a short description of each: (i) Volume refers the size of the data itself; (ii) Velocity refers to the increased speed that data is collected; (iii) Variety refers to the diverse types and forms of organized and disorganized data that any given system can collect; and (iv) Veracity refers to different aspects of data quality, e.g., bias, noise, etc.[57, 74]. Each of these attributes possess unique benefits and associated challenges with these large-scale datasets. As such, the development of automated tools for data visualization, acquisition, and decision-making are in increased demand to aid clinicians, health departments, policy makers, and the public in understanding and communicating key ideas and concepts that can benefit society.

There have been many advancements in the field of big data, machine learning, and artificial intelligence, which have provided pathways for profound research and utilization of such large data sources for various fields and data generation outlets. The utilization of such surpluses of data, however, requires novel approaches to identify or discover patterns and relationships within the data [74]. The studies presented in this thesis explore applications of machine learning, artificial intelligence, and big data with respect to specific questions within the scope of medical decision making.

Health Care Data

When discussing big data in healthcare, it is important to contextualize the various types of data and the aggregate volume and growth of information in order to place emphasis on and fully understand the complexity of the problems which rely on health care data. Health care data comprises a variety of big data that can be collected in a multitude of forms such as medical histories, patient surveys, electronic health records, doctor's notes, lab or imaging results, data collected from wearable devices or smartphone applications, claims data, and more [57, 74]. In recent years, estimates have shown that the volume of health data is expanding at rates upwards of 48% growth annually[9, 120]. Other estimates suggest that healthcare is quickly becoming the fastest growing outlet for data creation, consumption, and utilization among others such as financial services, manufacturing, and media and entertainment [43, 68] .

Due to the rapid accruelement of and complexity surrounding healthcare data, there remains a persistent unknown as to which questions can best be answered as well as which data sources should be collected and used to answer those questions. Presently, there are different techniques such as data mining, machine learning, and other forms of statistical analyses that are studied and utilized, but improvements are necessary to combat the limitations of each approach [74]. We present some of the many challenge and limitations of computational models in healthcare in the proceeding sections.

A prime example of this extensive accumulation of data coupled with the need to answer never before asked questions surrounds the COVID-19 pandemic. Since the beginning of the pandemic in the United States, early 2020, researchers have been presented with the challenges of taking large scale and highly irregular data to perform analyses that can aid communication between local health departments and the general public and guide public policy. The challenges presented during COVID-19 are unique in the sense that data is rapidly accumulating and made publicly available for researchers to use for various problems. As a result, we are gaining insight towards the limitations of the data in the context of aiding public policy decision making as well as defining new bounds on our abilities to rapidly transform data to analyze human behavior and disease spread.

Challenges of Using Healthcare Data

With such existing and continued projected growth unique problems are bound to arise ranging anywhere from data storage to usability. Even with current efforts in place to draw attention to the importance of health data collection and participation, big data in health care falls short of being used to its full potential. The mistranslation from data availability to knowledge or guided clinical practice is due in part to each data variation, complexity, incompleteness, mismatch of scaling, lack or shortcomings of computational technologies, and data privacy concerns. [1, 99]. All of these limitations contribute to the gap between potential and practice. This work will propose new ideas and developments to overcome these obstacles with the intention to guide public health policy and report insights that result from applying these methods to real-world data.

1.2 Infectious Diseases in the United States

The United States has historically faced the onset of threat from several infectious diseases ranging from Smallpox, with a recent outbreak in 1949, to other prevalent diseases such as HIV/AIDS [23, 36]. While some infectious diseases have vaccines or treatments available to lessen their impact, some remain untreated, and there is always the chance that a new outbreak can emerge. Some of the most prevalent infectious diseases that the United States faces annually is pneumonia and influenza, and in recent years, COVID-19 [24, 28]. These illnesses cause substantial burden in terms of prevalence and high risk of severe outcomes where influenza causes approximately 9 million to 41 million illnesses annually [35], pneumonia causes 1.5 unique hospitalizations each year [137], and since it first arrived in the United States COVID-19 has caused approximately 82 million reported cases, 4.6 million hospital admissions, and over 900,000 reported deaths [32]. That means that in recent years, two of the most burdensome infectious respiratory diseases affecting the United States in terms of hospitalizations and deaths are pneumonia and the novel coronavirus [24, 32]. As such, this work focuses on developing computational tools to inform risk factors, intervention strategies, and burden analyses for pneumonia and COVID-19. In some cases, the work presented here can also be generalized to inform policy for other diseases or current problems and these ideas will be presented in each of the chapters where appropriate.

There are multiple problems to solve to improve individual and population health outcomes attributable to infectious diseases, even specifically for pneumonia and COVID-19. Due to the high prevalence of these diseases, it is important to identify which of the population is at greatest risk of severe disease, or even death and how to reduce that risk. Furthermore, targeted intervention policies are necessary to reduce severe outcomes and overall costs related to this risk. Lastly, public health experts and medical decision makers could benefit from the addition of tools to fully capture disease incidence and evaluate transmission dynamics and disease characteristics over time. Presently, there exists a need to identify risk factors, develop intervention policies, and improve the use of computational modeling tools to understand total disease burden. This work focuses on addressing all three of these aspects and provides solutions with regards to infectious respiratory diseases.

The rest of the introduction chapter will provide an overview of the three main areas that this work will address, namely, risk of severe disease, interventions and prevention, and disease burden. Each subsection below introduces the relevant ideas, motivation, and key contributions for the three main objectives of this work.

Risk of Severe Disease

Infectious diseases such as pneumonia and COVID-19 are not only extremely prevalent among the population each year, but are known to cause increased risk of severe disease or mortality risk [28, 30]. Ramirez et al. report that of the 1.5 million unique patients with community-acquired pneumonia (CAP) hospitalized each year, one out of every three die within one year of said hospitalization [137]. Understanding pneumonia causes and risk factors can help inform interventions and improve

health literacy so that individuals can make the best decisions regarding their health to avoid these outcomes. The CDC notes that pneumonia can be caused by a variety of pathogens, bacterial or viral. Some of the viral diseases that are known to cause pneumonia are COVID-19, human metapneumovirus (HMPV), human para-influenza virus (HPIV), influenza, pneumococcal disease, respiratory syncytial virus (RSV), and rhinovirus infection, among others [26]. Ramirez et al. also were able to identify clusters of CAP within certain populations such as low income and African American communities which may indicate that these communities are at increased risk of exposure to a disease that can cause pneumonia [137]. Once an individual acquires a pneumococcal illness, there are certain known risk factors that increase their likelihood of severe disease or death. Some include a person's age or any pre-existing conditions such as alcoholism, chronic heart or lung disease, diabetes, or decreased immune function [39]. There are other known risk factors such as certain racial and ethnic groups like Alaska Native, African American, and certain American Indian populations that can increase an individual's likelihood of severe infections [30].

One objective of this dissertation will be to assess and evaluate risk factors associated with severe instances of respiratory diseases. Specifically, we will utilize administrative claims data to evaluate individual, community, and provider level impacts on an individual's 30-day risk of experiencing a hospitalization with a primary diagnosis of pneumonia. A key contribution of this chapter includes the evaluation of 30-day hospitalization risk, a more immediate time frame compared to other published studies. Other contributions include assessing risk factors via multiple machine learning algorithms to evaluate pneumonia hospitalization risk for the general population and using data sets which are available to multiple stakeholders ranging from hospitals to payers.

Prevention

It is important to identify the most vulnerable populations to target for interventions, or treatments, to reduce the risk of developing severe disease which could lead to a reduced quality of life or increased mortality risk. It is well known that the most vulnerable populations for pneumonia and COVID-19 tend to be those youngest in age, 0-4 years old, and oldest in age, 65+ years old [39]. In addition to age, there are other factors that put an individual at greater risk, such as one or more comorbid conditions, i.e., diabetes, hypertension, obesity, cancer, etc., that have immunocompromising consequences. The notion of identifying at-risk populations is already explored in the first objective of this dissertation. The second objective will explore intervention targeting based on individual risk scores.

In order to consider whom to target with interventions, it is important to understand what interventions are available and recommended to the public. Here are summarized several interventions, varying by accessibility and costs. The easiest and cheapest intervention from an administrative standpoint is to communicate and teach the importance of making conscientious healthy lifestyle choices to protect oneself from infectious diseases and reduce the likelihood of experiencing severe disease. For respiratory diseases, the CDC recommends individuals practice smart hygiene practices such as washing your hands often and correctly, maintaining a safe distance from those who are

currently infectious, and washing high contamination-risk surfaces, such as kitchen counters and bathrooms, regularly [34]. These changes are generally risk-free lifestyle choices that can easily become integrated in an individual's routine which can greatly improve health-related outcomes. In addition to these lifestyle changes, individuals can take other measures to improve their overall health, which lessens the likelihood of severe disease. This could come in the form of attending general check-ups annually to complete appropriate tests, or screenings, and receive medical advice specific to one's health history. Further, if an individual already has been diagnosed with a chronic illness, or other health condition, they can adhere to guidelines and recommendations to alleviate the risk attributable to these conditions. For instance, if an individual is pre-diabetic, then making conscientious diet choices can greatly improve one's health outlook [34], and in turn reduces their risk profile when considering infectious diseases. Outside of daily lifestyle choices one can make, there are also vaccinations and treatments available that greatly reduce one's risk of infection and likelihood of severe disease. Concerning pneumonia, the current CDC vaccination recommendations include vaccines against the pathogens and other diseases that are known to cause pneumonia[30]. These types of interventions require physician assistance to administer and are more costly than the above-mentioned interventions. Therefore, there exist current recommendations about when an individual should receive and how often [34]. Despite the benefits associated with vaccinations, there still exists hesitant populations that do not follow the recommended guidelines. Therefore, it is important to know whom to target with interventions such as supplying information regarding vaccine safety and efficacy, reminders about when to receive certain immunizations, and with the vaccines themselves.

While understanding the types of interventions available is important for health experts, this is only one component of medical decision making processes necessary to improve public health outcomes. For instance, if you have potential interventions to utilize and distribute, e.g., vaccines, how many people should receive the intervention? This decision can be impacted by various factors such as the population at hand or budgetary constraints. If decision makers have an exact budget, then the decision may be obvious to decidedly distribute interventions to those at highest risk. However, this decision becomes more complex when it is desirable to make decisions that balance overall costs and outcomes, whether successful or not, across a set of patients, such as within a hospital system. In a set of patients with available risk scores, is there a better overall solution to reduce population outcomes? From a stakeholder's perspective, such as payers or hospital administrators, it may be necessary to consider the entire set of patients to determine the best decisions at the population level rather than focusing on one individual patient at a time due to resource constraints or ease of decision making. Additionally, the National Association of Community Health Centers notes that care models based on risk allow for customized care and more appropriate resource selection at the provider level [51]. In these scenarios there may be uncertainty, surrounding the individual risk or likelihood of severe outcomes and the overall effectiveness of the intervention to prevent severe outcomes. As noted above, there are many available intervention types and sometimes it may also be beneficial to consider distributing an additional intervention of lower cost. These questions

remain important to public health experts faced with the decision to intervene on a population to reduce severe outcomes, such as hospitalizations, while also considering overall costs.

The second objective of this dissertation is to develop a multi-class method to identify populations to target with interventions to reduce pneumonia hospitalizations. In this model we consider two available interventions, and describe a heuristic to identify risk score thresholds to administer the two types of interventions based on risk scores, efficacy, and unit costs. Some highlighted contributions of this work include the consideration of uncertainty of regarding targeting a population with two interventions of varying efficacy, and the optimization of expected costs to reduce pneumonia hospitalizations. The work presented in this chapter generalizes well to other notions of risk, for other infectious disease as well as instances where risk scores are available and interventions are appropriate.

Burden

A more recent infectious disease, COVID-19, has been widely studied within the medical community to understand and evaluate risk and burden associated with the disease. Similar to pneumonia, it is well known that likelihood of severe disease, hospitalization, or even death, increases with age and the presence of certain comorbid conditions [28]. There are other factors, such as geographic location within the United States, which has been found to significantly change the level of risk an individual might experience, or race and ethnicity where certain population groups have been shown to share a disproportionate burden of the disease compared to others [32]. Other studies have also shown that the total disease burden may have been much greater than what was captured, or officially reported, via lab-confirmed cases, confirmed hospital admissions, and death certificates [10, 32, 42]. As of May 2022, COVID-19 has been responsible for approximately 82 million documented infections, 45 million unique hospitalizations, and over 1 million deaths in the United States alone. In terms of costs, hospitalization expenses, which varied depending on whether an individual developed severe disease, required intensive care or mechanical ventilation, ranged anywhere from \$13,000 to \$30,000 per stay [155].

Due to the rapid spread of the disease over time the United States has suffered from inadequate and insufficient data reporting and thus has relied heavily on the assistance of data analytics and modeling to inform decisions to reduce disease burden. For instance, compartmental and agent-based models have been used to aid policymakers and other primary stakeholders to perform decision analysis on the impact of certain non-pharmaceutical interventions (NPIs) and testing strategies on disease spread, severe disease, and deaths due to the virus [17, 146]. Policymakers worked side by side with researchers to create communication tools to make information widely available and accessible to a variety of populations nationally and globally [54]. Understanding the impact of these interventions and policies can help inform the future of the pandemic and other infectious diseases that can cause substantial burden on the healthcare system at large. Furthermore, understanding under-reporting can inform current issues regarding COVID related outcomes such as identifying long-COVID populations, where individuals across the country are struggling to

receive assistance due to the lack of an official confirmed positive COVID test to document their illness [115, 118].

The third objective of this dissertation is to develop methods towards understanding total burden, i.e., incidence, hospitalizations, and deaths, associated with infectious diseases. We illustrate our methods by estimating the total disease burden associated with COVID-19 prior to vaccination in 2020. The methods considered in this chapter can be utilized to inform similar problems with other infectious respiratory diseases. In particular, it is helpful to characterize true disease burden in any instance when there is under-reporting, under-diagnosis, or any changing disease landscape. Some key contributions of this chapter include disease modeling that dynamically learns and characterizes disease landscape from data and considers the impact of new and effective treatments, changing transmissibility due to variants or human behaviors, and case ascertainment due to infrastructure and testing capacity.

1.3 Objectives and Organization

The main objectives of this work can be summarized into three main components. Namely, this work will explore and discuss the synthesis of multiple data sources to achieve the following objectives specific to infectious respiratory diseases in the United States.

1. Identify at-risk populations.
2. Recommend targeted intervention policies.
3. Evaluate and understand disease burden.

This work expands upon the current literature on infectious respiratory disease risk, prevention, and spread in the following ways. First, the computational models used in these studies are uniquely defined with a purpose to provide interpretable solutions for each chapter. These models produce findings that are consistent with current literature and can be used as tools to aide communications and bridge the gap between large data in health care and clinical practice. Second, large volume data sets are carefully considered and incorporated into the separate models. The exploration of different data sets provides insights upon the usefulness of data and features that help explain individual risk of infectious respiratory diseases as well as points to data that should be considered and collected more frequently. Third, the tools described and developed in this work were created to utilize big data to help inform public health, individual risk, and mitigation. Many of the models can be altered, or transformed, to aide the same purpose for other areas outside of infectious respiratory diseases. That is, this work can be expanded to assess risk of other unplanned or adverse events, provide intervention policy recommendations for other areas where there is well-defined risk and available interventions, and provide modeling tools for other infectious diseases that may cause burden on healthcare systems.

Organization

The remainder of this work is outlined in the following manner. Chapter 2 will discuss the development of three predictive modeling tools to identify at-risk populations for pneumonia hospitalizations within 30 days. Chapter 3 provides policies to target interventions to reduce pneumonia hospitalizations. Chapter 4 offers computational modeling tools and insights to evaluate burden, e.g., incidence, hospitalization risk, and death, associated with COVID-19. Chapter 5 provides concluding remarks. A review of the computational models and necessary nomenclature is provided in Appendix A followed by Appendices for relevant additional information from each chapter.

CHAPTER 2

PREDICTIVE MODELING TO ASSESS 30-DAY RISK OF UNPLANNED PNEUMONIA RELATED HOSPITALIZATION

2.1 Introduction

Pneumonia causes undue burden on the United States healthcare system each year. Even those who are able to receive treatment and recover are still at risk for long term health complications. Grimwood et al. [79] report that a pneumococcal infection during an individual's developmental period, e.g. 0 to 5 years of age, lead to increased risk of chronic lung disease, asthma, and chronic obstructive pulmonary disease in adulthood. Other sources from John Hopkins University note that complications for pneumonia can include acute respiratory distress syndrome, lung abscesses, respiratory failure, and sepsis [117]. These life altering illnesses are not taken lightly by individuals or medical practitioners and in turn research efforts have been dedicated towards understanding and reducing individual risk of pneumonia and developing severe conditions such as requiring hospitalization.

Current literature suggests that there are known groups that are of greater risk of developing pneumonia with serious complications that lead to a higher mortality risk than the general population. Namely, the CDC lists individuals of age 65+, certain racial and ethnic groups such as Alaska Native, African American, and American Indian, and others with conditions such as alcoholism, chronic diseases, a history of smoking, diabetes, or decreased immune function [30]. Knowing these risk factors has led to personalized recommendations for individuals to reduce their risk

of pneumococcal disease or severe complications. Ramirez et al.[137] report that mortality rates following a community-acquired pneumonia hospitalization for 30 days, 6 months, and 1 year were 13%, 23.4%, and 30.6%, respectively. Additionally, they found that areas with high incidence of community acquired pneumonia were associated with low-income, who are likely unable to afford doctor's visits and necessary preventative medications [147], and black/African American populations, which are known to have higher prevalence of high-risk comorbid conditions [179].

Due to the complexity of diagnostic measures and poor prognosis for those with high risk factors, there remains a need to strengthen diagnoses and identify individuals who are in greater need of interventions to inform policies and target interventions to reduce the burden associated with severe instances of pneumonia. The following section summarizes the literature on risk evaluation and efforts for predicting severe outcomes related to pneumonia prognoses and diagnoses.

2.2 Literature Review

Over the years, various prediction and classification models have been developed to understand factors that can help quantify an individual's risk of severe disease or mortality. The pneumonia severity index (PSI) is successful at identifying patients with community-acquired pneumonia who are less likely to experience severe disease outcomes or death [72] using a two-step prediction rule to classify individuals into three risk classes depending on medical histories and physical examination findings. The CURB-65 risk score quantifies pneumonia severity according to increased risk of 30-day mortality captured in a six point metric [106]. An assessment of the PSI, CURB, and CURB-65 scores on evaluating 30-day mortality from pneumonia revealed that the more complex score, PSI, which utilizes specific information on coexisting conditions performs better at identifying low-risk individuals [136].

In 2019, a predictive model which uses biomarkers to stratify patients with acute respiratory infections into one of three risk groups for developing pneumonia, with the highest risk group identified as having greater than 20% chance of a pneumonia diagnosis [80], showcased the importance of chest-x ray results in diagnoses and highlighted an association between certain biomarkers (C-reactive protein existence) and the need for antibiotics for treatment. These results are important in guiding clinical practice for administering antibiotic treatments and chest x-rays, but in practice these tools are not available for everyone, particularly those without adequate insurance coverage. In an analysis aiming to identify leading pathogens causing pneumonia hospitalizations, it was reported that pathogens were only detected in 38% of adults included in the study [88]. This suggests that there remains a need to define other variables with stronger predictive power for community or viral pneumonia related hospitalizations.

These developed models focus on predicting overall severity of disease given that an individual has a diagnosis of pneumonia. Then, there are some models that focus on evaluating pneumonia readmission rates within a population [113]. Other models focus on the prediction of pneumonia, or evaluating risk, based cohorting by pre-existing medical conditions, i.e., those with diagnosed

respiratory tract infections, myocardial infarction, congestive heart failure, chronic obstructive pulmonary disease, etc, to name a few [98, 143]. While these studies provide useful results and help inform how to treat patients at-risk or diagnosed with pneumonia, they do not generalize well to the general public, where it is known that anyone over 65 years of age is at increased risk of severe pneumonia [30].

Moving from symptoms and individual health records, the use of administrative claims data has been explored to understand disease severity. In 2014, administrative claims were used to predict in-hospital mortality for individuals diagnosed with pneumonia within the first two days of hospital admission [148]. Other studies sought to develop predictive models for hospital admissions related to other diseases, which highlight the potential of administrative claims for patient predictions. Yang et al. described a method to aggregate administrative claims data to develop a predictive model to assess an individual's risk of hospital admission given that they have congestive heart failure and found that the dynamic [175]. They present a dynamic random survival forest model to develop dynamic variables solely based on administrative claims data and find that the model was successful at identifying individual risk factors for those with congestive heart failure. Another study used administrative claims to perform personalized predictions to evaluate risk of unplanned urinary tract hospitalizations for hierarchical clusters of Medicare beneficiaries [114]. The authors report that the model achieves reasonable predictive performance and is useful at identifying interpretable features that can aid intervention design.

The above-mentioned models focus on personal medical histories, and current symptoms of pneumonia at an individual level. An assessment of community variables and CMS hospital readmission rates highlights that readmissions for a variety of clinical conditions are influenced by community factors at an individual's place of residence [158]. A separate analysis signifies that social factors such as lower education, low income, and unemployment were highly associated with readmission and mortality following community-acquired pneumonia [20]. This highlights the need for careful consideration of community factors when considering pneumonia hospitalization risk.

The closest known existing model in the literature that attempts to develop a predictive model for pneumonia hospitalization is presented by Uetmatsu et al., where they propose a model to predict pneumonia hospitalization [167]. The study design in Uematsu et al. specifically included patients who had undergone specific health checkups as part of a nationwide screening program during a one year period while the program was active. The follow-up period for whether an individual experienced an event, e.g. pneumonia hospitalization, was 4-5 years. The study included individuals age 40-74 years old who were eligible for the nationwide screening study, though it is not clear what makes an individual eligible for the study and their outcome event rate was approximately 1.68%. While the information gained can prove to be useful to inform lifestyle changes and some interventions, it does not easily allow for identifying and targeting hospitalizations for those who may be at risk in the near future.

To the best of our knowledge, this would be the first published study utilizing administrative claims data to develop a dynamic, predictive model to identify individuals in the general population

at risk for an unplanned hospital admission with a primary diagnosis of pneumonia within 30 days. The other predictive models listed above either evaluate mortality risk for confirmed pneumonia hospitalizations, evaluate readmission risk, or consider risk for cohorts with specific comorbid conditions, for which the populations have much higher event rate for predictive analysis. The model presented in this chapter provides interpretable solutions and is extendable and generalizable to other populations. Furthermore, this study will be the first published to evaluate pneumonia hospitalization risk using only administrative claims data and publicly available information, which is useful from a payer's perspective and can be utilized to evaluate risk at an administration level. This study is also the first to assess the joint impact of individual, community, and provider, level information on the risk of pneumonia hospitalizations, unspecific to readmissions, to inform potential interventions at various levels ranging from the individual to the entire community. The study also identifies new interpretable and important predictive features for pneumonia hospitalization prediction such as frailty and community flu activity that can aid intervention practices.

Study Description

This study aims to create a predictive model using individual, community and provider level variables to assess Medicare beneficiaries risk of unplanned pneumonia hospitalization within 30 days of a set index date, e.g. the first of every month. We use the Agency for Healthcare Research and Quality (AHRQ) Prevention Quality Indicator 11 (PQI-11) [2] definition for the numerator to compute hospital admission rates for Bacterial Pneumonia to create our independent variable. The PQI-11 label is computed using a set of ICD-9 (International Classification of Diseases, Ninth Revision) [46] diagnosis codes after exclusions listed in Table 2.4. Some of these exclusions defined by AHRQ and listed in Appendix B include cases with transfers from other facilities, any patient specific medical history exclusions including claims with any listed diagnosis or procedure codes for sickle cell anemia, HB-S disease, or an immunocompromised state and cases with missing information. Our population exclusion criteria includes the same provisions to be consistent with this label definition.

2.3 Methods

Study Population

Inclusion/Exclusion Criteria

This retrospective case-control study cohort comprises a 5% sample of Medicare beneficiaries from 2008 to 2011. For each year, we have access to claims reported in the carrier, inpatient, outpatient, skilled nursing facility, home health, and durable medical equipment files. To aid in observation completeness, we required that an individual had at least one claim in 2011 to be included in the study. To limit potential biases, we instantiated the study population using several other inclusion/exclusion criteria outlined in Figure 2.1.

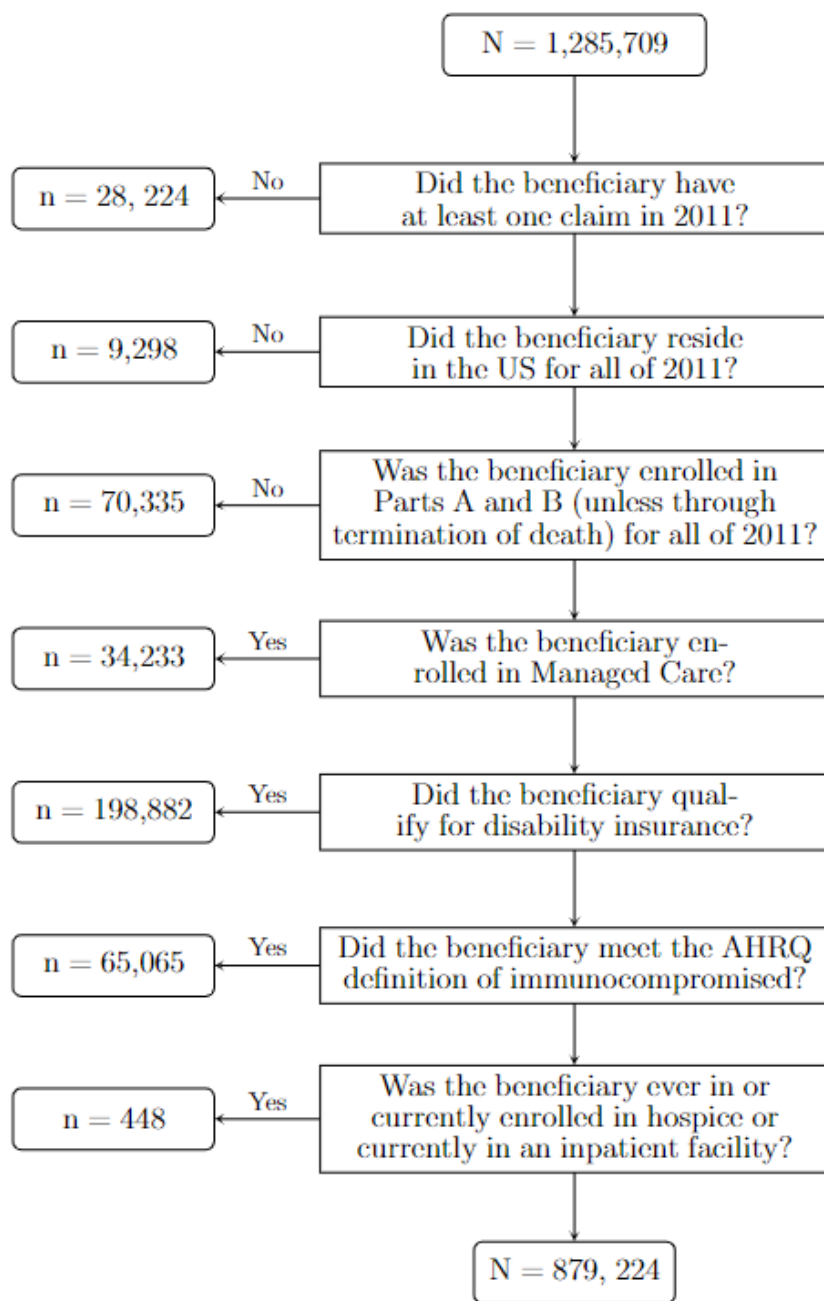


Figure 2.1 This flowchart illustrates the inclusion and exclusion criteria for beneficiaries to be included in our study. The values on the left-hand side of the diagram represent the number of beneficiaries that are excluded during each stage in the process in response to the question listed immediately to the right and with the response listed directly above the connective arrow.

Requiring that the beneficiary resides in the United States during the study period ensures that we have access to community and other provider level information obtained from publicly available datasets including information about quality of care at the beneficiary's most recent inpatient provider, type of facility of the most recent inpatient visit, and information pertaining to community level vaccination rates. Requiring that each beneficiary is enrolled in both parts A and B for all twelve months of the study period, unless through termination of death, ensures that we have access to medical billing information from primary care providers, available for those enrolled in Part B, from which we derive some individual-level medical history variables. We include those who died later in the year because they are still at risk for an unplanned hospitalization in the earlier months. To fully capture healthcare utilization for everyone, we exclude beneficiaries enrolled in Managed Care and those who had conditions that would qualify them for any form of disability insurance. This exclusion is due to the fact that Managed Care facilities do not submit all their claims to CMS and individuals with disabilities may be dually eligible for Medicaid, where either dual-enrollment would result in an incomplete sample of claims data for these individuals.

In accordance with our event definition, we removed all individuals whose claims had diagnosis and/or procedure codes that fell within the categories for sickle cell anemia, HB-S disease, or an immunocompromised state defined by AHRQ [2] (see Appendix C), removing those entirely from the set who had any of these markers in 2008 to 2010 and excluding any later months for beneficiaries who first met any of these conditions in 2011. Lastly, we remove all individuals that were ever in hospice care, or are currently enrolled in hospice, since they would likely not benefit from a prediction by our model, and additionally remove anyone who is currently in an inpatient facility by the beginning of each month since we are aiming to capture new unplanned hospital admissions.

After the study population has been fully instantiated and appropriate exclusions were applied, the population consisted of 879,224 unique patients and correspondingly 10,123,471 rows of patient feature vectors for all index dates for the year 2011. Table 2.1 shows a comparison of demographic data between patient rows with an event (event rows), and patient rows without an event (non-event rows). The proportions within each set, event and non-event rows, are compared using a proportions z-test at two levels of significance where differences are reported at the 0.05 level of significance, indicated by (*) and the 0.01 level of significant, indicated by (**). The descriptions of the chronic diseases and CCS categories listed in this table are defined in Table 2.2.

Filtering Patients by Medical History

Of the 10,123,471 patient rows in our data set, only 11,700 are associated with a PQI-11 event, which yields 0.12% event rate. This rare event occurrence and large class imbalance is likely to lead to poor discriminantory power. Thus, we applied additional filters to the data based on known risk factors identified in the literature. The remaining population gets filtered by medical history described below. The filtering method can be considered as an initial prediction round. Any beneficiary who does not meet these criteria does not require a prediction because they are considered low-risk.

We used the filtering rules in Figure 2.2 to identify individuals who are at risk for hospitalization.

Table 2.1 This demographic table illustrates the composition of each patient-row group, (i) event or (ii) non-event group for this study. We report whether the two groups have statistically significant differences in counts of each group, according to a two-sided proportions z-test with alpha =0.05 (*), alpha = 0.01 (**).

	Count Event Rows	Percent Event Rows (%)	Count Non-Event Rows	Percent Non-Event Rows (%)	
Total Observations	11,700		10,111,771		
Unique Beneficiaries	11,109		879,071		
Age					
Younger than 65	74	0.63	56,375	0.56	
65-69	1,444	12.34	2,470,494	24.43	**
70-74	1,742	14.89	2,340,119	23.14	**
75-79	1,987	16.98	1,960,385	19.39	**
80-84	2,398	20.50	1,624,798	16.07	**
85+	4,055	34.66	1,659,600	16.41	**
Race and Ethnicity					
Unknown Race	26	0.22	31,172	0.31	
White	10,511	89.84	8,909,999	88.12	**
Black	657	5.62	727,159	7.19	**
Other	93	0.79	134,610	1.33	**
Asian	114	0.97	124,604	1.23	*
Hispanic	218	1.86	138,800	1.37	**
North American Native	81	0.69	45,427	0.45	**
Gender					
Male	5,077	43.39	3,962,860	39.19	**
Female	6,623	56.61	6,148,911	60.81	**
Chronic Diseases and Disorders					
Chronic Diseases other than Heart	9,582	81.90	6,443,972	63.73	**
Diseases of the Nervous System	1,069	9.14	392,448	3.88	**
Heart Disease	7,195	61.50	3,550,312	35.11	**
Attention and Personality Disorders	2,385	20.38	794,771	7.86	**
Sum of CCS Categories					
0-2	3,777	32.28	6,380,442	63.10	**
3-5	3,774	32.26	2,534,119	25.06	**
6-8	2,574	22	870,488	8.61	**
9 or more	1,575	13.46	326,722	3.23	**

Table 2.2 CCS Category groupings and corresponding diagnoses used to compare the event and non-event rows in the study population in Table 2.1

Grouping	CCS Category	Diagnoses
Chronic diseases other than heart	19	Cancer of bronchus; lung
	20	Cancer; other respiratory and intrathoracic
	49	Diabetes mellitus without complication
	50	Diabetes mellitus with complications
	98	Essential hypertension
	99	Hypertension with complications and secondary hypertension
	102	Nonspecific chest pain
	127	Chronic obstructive pulmonary disease and bronchiectasis
	128	Asthma
	158	Chronic kidney disease
Diseases of the nervous system	79	Parkinson's disease
	80	Multiple sclerosis
	81	Other hereditary and degenerative nervous system conditions
	85	Coma; stupor; and brain damage
Heart disease	96	Heart valve disorders
	97	Peri-;endo-;and myocarditis; cardiomyopathy
	100	Acute myocardial infarction
	101	Coronary atherosclerosis and other heart disease
	103	Pulmonary heart disease
	104	Other and ill-defined heart disease
	105	Conduction disorders
	106	Cardiac dysrthmias
	107	Cardiac arrest and ventricular fibrillation
	108	Congestive heart failure; nonhypertensive
Attention and Personality Disorders	652	Attention-deficit, conduct, and disruptive behavior disorders
	653	Delirium, dementia, and amnestic and other cognitive disorders
	654	Developmental disorders
	658	Personality disorders
	659	Schizophrenia and other psychotic disorders

Recent claims data in the last 6 months prior to the patient-row index date ensure that we have information from the administrative claims sample about the patient's health history including previous diagnoses and procedures. This is coupled with an inpatient Elixhauser score [64] greater than zero because of literature that suggests certain comorbid conditions that lead to increased risk of severe disease or mortality [72]. Our inpatient Elixhauser score is a comorbidity index computed from diagnosis codes from an individual's most recent visit to an inpatient facility. The Elixhauser score is a comorbidity index computed from a set of 30 comorbid conditions, which has shown to be helpful in prediction of clinical outcomes [64]. Other risk factors identified from the literature include a history of severe pneumonia or pneumonia hospitalization and previous nursing home visits [72]. Other literature also points to the fact that certain biomarkers such as temperature, and respiratory rate [72, 167], are strong indicators of whether someone will experience severe

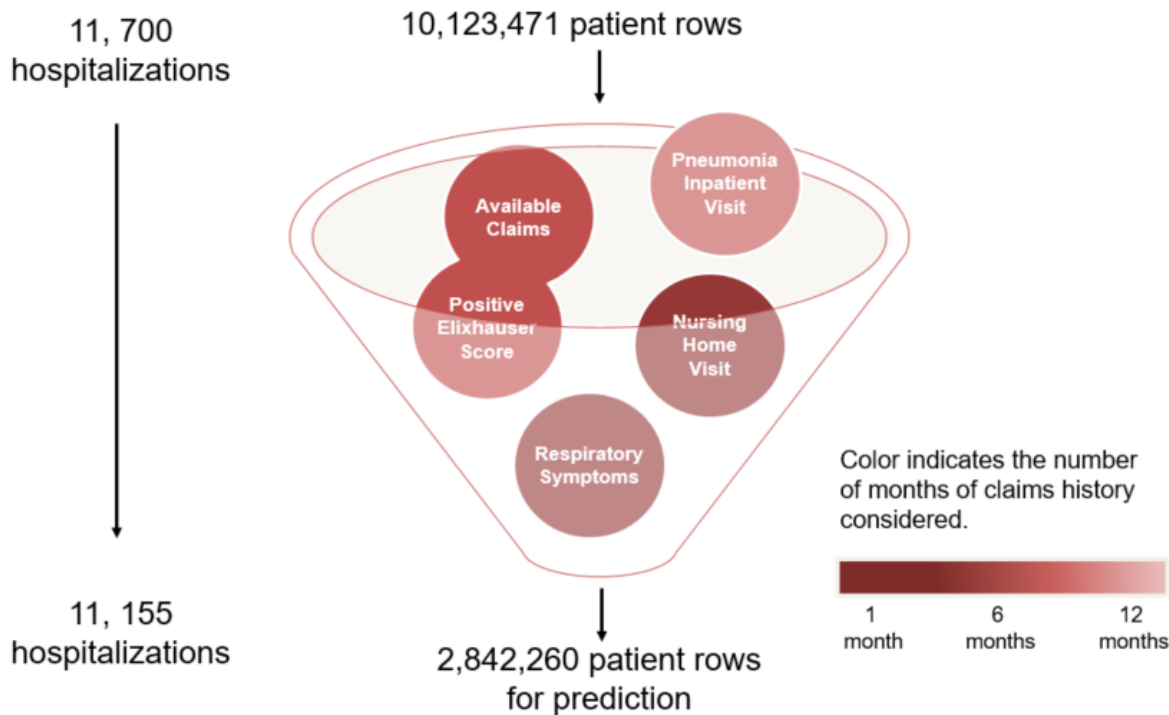


Figure 2.2 Filtering criteria to gather beneficiaries based on their medical history, which is applied using an inclusive 'OR' statement to create our prediction set.

pneumonia. As discussed in the limitations section regarding administrative claims data, we do not have access to these biomarkers which serve as individual risk factors. In lieu of personal biomarkers, we develop a variable to assess whether an individual has respiratory symptoms in the month prior based on ICD-9 codes described in more detail in Appendix B.

As illustrated in Figure 2.2, applying this filtering technique reduces the patient-row dataset to 454,560 unique patients, 2,842,260 patient rows and 11,155 PQI-11 events. We are left with a 0.39% PQI-11 event rate, which corresponds to a 72% reduction in the number of patient rows in the data set and the retention of at least 95% of the PQI-11 events in our original set.

In Table 2.3 we present a few descriptive statistics comparing the event and non-event patient rows in the set that was filtered out using the criteria outlined in Figure 2.2. If the statistic showed that there was a potentially implementable policy to identify greater risk populations then the criteria was considered in addition to the four main filtering criteria. For instance, the mean frailty score is higher in the non-event patient row set and we would not generally consider patients with lower frailty scores to be at greater risk so this filtering criteria is not considered. For the other strategies, such as considering older age groups, smokers, or vaccines such as influenza or pneumonia, we report that the inclusion of these policies significantly reduces the event rate in the final, resulting data structure by another 0.09 to 0.26 percentage points, which is a significant reduction when considering the large size of the overall data structure. Therefore, we conclude that our filtering criteria is the most appropriate consideration given the lack of recent history, i.e., zero value for the

median carrier and inpatient claims in the event group's patient rows, and the riskier non-event patient rows, i.e., higher mean frailty score, higher median number of CCS conditions, and higher percentage of previous influenza-like illnesses, etc.

Table 2.3 Descriptive statistics comparing the event and non-event patient rows of the patient-rows that were filtered out, or excluded, using the above filtering criteria by recent medical history.

	Event Patient Rows	Non-Event Patient Rows
Median age group	80-84	70-74
Percentage		
Previous ILI (6mth)	2.75%	6.10%
Median Number		
CCS Conditions	1	5
Percentage Smokers	26.97%	12.58%
Median Number		
Carrier Claims		
(1, 3, 6 mth)	0, 0, 0	1, 3, 6
Median Number		
Inpatient claims		
(1,3, 6 mth)	0, 0, 0	0, 0, 0
Mean Frailty		
Score	0.0119	0.0340
Percentage		
influenza vaccine	35.96%	50.89%
Percentage		
Pneumonia vaccine	11.93%	17.81%

Study Design

Individual observations, or feature vectors, are aggregated at the monthly level. Variables were summarized up to and including the month prior to the observation window. We define an event to be an unplanned hospitalization with a primary diagnosis code for PQI-11, which is a list of 16 ICD-9 codes for pneumonia, given in Table 2.4. We develop models to predict whether an individual will experience an unplanned hospitalization within 30 days post index date for that patient-row month. The events are modeled as binary indicator variables which inform whether an individual experienced an unplanned hospitalization. Our model does not distinguish between hospital readmissions and first-time admissions.

Letting our index dates be the first of every month in 2011, we define an observation to be a complete set of dynamically computed variables for a beneficiary for a given index date.

Table 2.4 List of diagnosis codes provided by AHRQ to compute the numerator for PQI-11 Bacterial Pneumonia Admission Rates. "Nos" refers to not-otherwise specified.

ICD-9 Code	Description
481	Pneumococcal Pneumonia
4822	H.Influenzae Pneumonia
48230	Strep Pneumonia Unspec
48231	Grp A Strep Pneumonia
48232	Grp B Strep Pneumonia
48239	Oth Strep Pneumonia
48240	Staph Pneu Nos
48241	Meth Sus Pneum D/T Staph
48242	Meth Res Pneu D/T Staph
48249	Staph Pneumonia Nec
4829	Bacterial Pneumonia Nos
4830	Mycoplasma Pneumonia
4831	Chlamydia Pneumonia
4838	Oth spec Org Pneumonia
485	Broncopneumonia Org Nos
486	Pneumonia, Organism Nos

Definition 2.3.1. Let \mathbf{B} be the set of all beneficiaries in our sampled study population. Then we define an observation for an individual i at month \mathbf{m} :

$$\mathbf{B}_{i,m} := \{v_{1,m}, v_{2,m}, \dots, v_{p,m}, E_{i,m}\},$$

where $1 \leq m \leq 12$ and $1 \leq p \leq 195$

$$E_{i,m} := \begin{cases} 1 & \text{Pneumonia-related hospitalization in month } m \\ 0 & \text{Otherwise} \end{cases}$$

The resulting dataset is a set of approximately $N * 12$ observations consisting of one patient-row for each month in 2011 for each beneficiary who was not excluded in our initial assessment for inclusion.

Variable Creation and Selection

Variable creation can extend back as far as 2008, using the earliest data available to us, and as recently as the month prior to the prediction window. The variables are aggregated at a monthly level, i.e. we assume that if a diagnosis code is found for a given month \mathbf{m} , then it will not reflect in an individual's record until month $\mathbf{m}+1$ to ensure accurate recording. The variables include both individual level information available from the administrative claims data and community

and provider level variables from publicly available data sources. Variables were considered either through identification from relevant literature searches or direct analysis of the data and were assessed using statistical analyses to identify variables with significant differences, e.g., means or proportions, between event and non-event patient-rows in the population.

Individual Level Variables

The individual level variables quantify healthcare utilization, comorbidities, and other measures related specifically to pneumonia. These variables were specifically chosen according to what is reported in the literature. For instance, in a 2015 study of pathogens and their association with pneumonia hospitalizations it was reported that of the identified pathogens, viruses were detected in 27% of the patients and human rhinovirus and influenza viruses were the two most commonly detected in patients with pneumonia [88]. Other studies have shown that a history of smoking, obesity, and previous respiratory illnesses are also indicators of increased risk [30, 72]. We include variables about previous influenza like illnesses, previous diagnosis of pneumonia, smoking, nursing home status, influenza and pneumococcal vaccinations, Elixhauser scores, frailty scores, and the presence of comorbidities such as existing cancers and other diseases related to the heart, chest, and altered mental state. We also have information on each beneficiary's age, race, and healthcare utilization in terms of total cost and number of claims. A complete list of the final individual level variables used in this study is included in Appendix B to this document.

Community and Provider Level Variables

The community and provider level variables gauge healthcare access, quality of healthcare, and risk due to community parameters. Some variables, which were created from previously identified community level factors for CAP, include measures of poverty and population demographics [20] We also collected information surrounding community vaccination rates, influenza activity, measures of air quality, hospital overall ranking, and type of inpatient facility for an individual's most recent inpatient stay.

Documentation describing the data sources, algorithms, and diagnosis and procedure codes for the creation of the final variables used in model training is provided in Appendix B.

Addressing Correlation and Multi-Collinearities

We begin with a set of 330 variables that characterize an individual's level of risk up to each index date for each patient- month row in our dataset. We test for multicollinearities and correlations among our 147 continuous, 140 binary, and 43 categorical variables. Continuous variables were iteratively removed from the set until all predictors had variance inflation factors smaller than fifteen. Continuous variables were also assessed for correlations using the Pearson correlation test and variables with absolute correlation coefficients greater than 0.7 were iteratively removed. Binary features were iteratively removed until all non-ordinal predictors achieved a Phi Coefficient of less

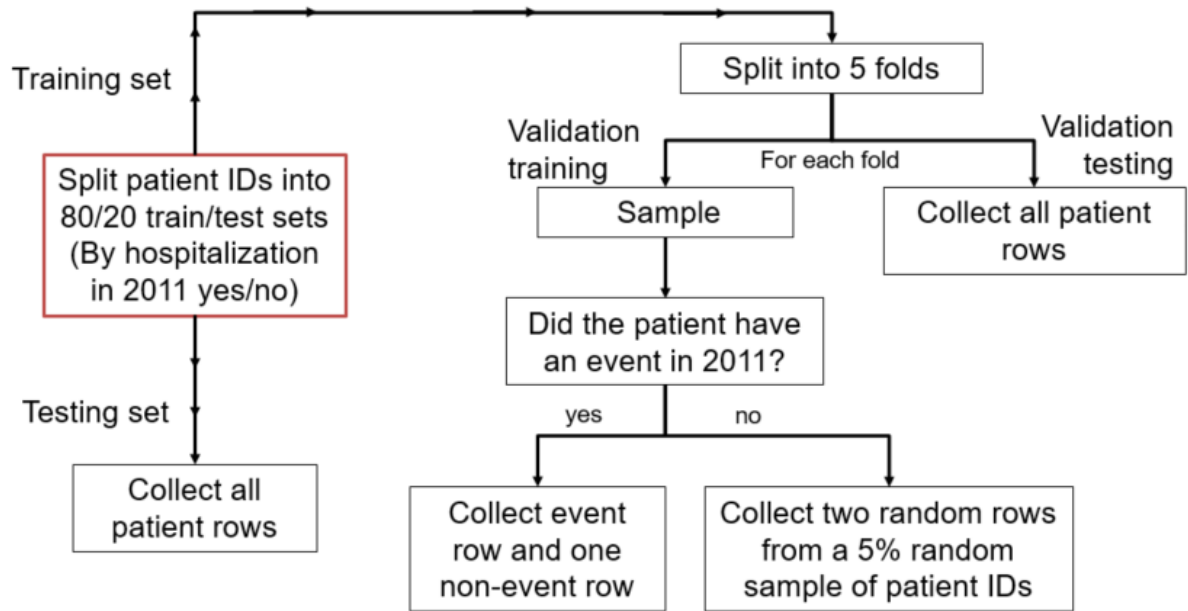


Figure 2.3 Sampling algorithm for training sets.

than 0.70 in magnitude with every feature in the final set. Finally, we perform a point biserial test between all continuous and binary variables and remove any variables with correlations greater than 0.7 in magnitude. After all highly correlated variables were removed from the set, the final prediction data structure had 195 variables: 155 individual level, 24 community level, and 16 provider level variables.

Model Training

First, we transform all of the continuous variables using the Absolute Max Scaler preprocessing function available from SciKit Learn's preprocessing library [130]. This preprocessing step reduces the overall computational complexity of our training time and allows for more interpretable coefficients and feature importances that are generated from each model.

Before performing the train/test split on our patient rows for our dataset, we create a new binary flag for whether a beneficiary experienced a PQI-11 event during 2011 or not. We then perform a split where 80% of the beneficiary IDs (train-IDs) are used to create our training set and 20% of the beneficiary IDs (test-IDs) are used to create our testing set. We further split the train-IDs into five folds to perform cross validation analysis. To create the validation training sets, we first begin with all patient-rows corresponding to the beneficiary IDs in each train-ID fold (validation train-IDs). For every beneficiary who experienced an event in 2011, we take the patient-row corresponding to the event and one other random patient rows. For every beneficiary who did not experience an event in 2011, we select two random patient rows. This sampling method was used to prevent overfitting and enable the models to train on multiple patient rows from the same beneficiaries. No further

sampling was done on the testing set and all patient rows selected from the above filtering method are used for final predictions. An overview of the sampling method to create the validation training sets is illustrated in Figure 2.3.

We trained three classifiers to perform our patient row classifications. Below we describe the settings used for the training of each of the final models. We use 5-fold cross validation to fine tune the hyper parameters described for each classifier and select the parameter with the best mean balanced accuracy score across all five folds. Here, the balanced accuracy measure used in this study is defined as the average of recall obtained on each class, or in other words, the accuracy score applied with class balance sample weights. In this analysis, the inverse event rate in the population is denoted by ρ .

1. **Logistic Regression Classifier**

Our logistic regression classifier is from the SciKit Learn `linear_model` library [130]. We used the `LogisticRegression` class which performs a classification based on a logistic regression for which we can specify the penalty function, class weights, solver, and maximum iterations. We can alter other parameters, but these were the ones focused on for this model in particular. We chose the 'l1' penalty function, which adds a penalty term using L1-normalization for incorrect classifications during the training process. This transforms the model into a LASSO logistic regression. We also report using the `liblinear` solver and find the best class weights using a 5-fold cross validation. For reproducibility, the random state is held constant at 1. Here, we are putting more emphasis, or greater cost, on a False Negative as opposed to a False Positive. These class weights are used to offset some of the issues we encounter when training and evaluating our models on a highly imbalanced dataset.

In these instances the majority weight is constant with a value of 1 and we vary the weight associated with the minority class.

For our hyper parameter search we considered the following possible values.

$$\text{Minority weights} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, \rho\}.$$

By considering the balanced accuracy score, we take the best minority class weight to be a value of 4.

2. **Gradient Boosting Classifier**

The gradient boosting classifier utilizes the `XGBClassifier` from the `xgboost` library in Python [40]. The gradient boosting classifier utilizes decision trees and optimizes the weights used to aggregate each tree in the set using a boosted gradient approach. Therefore, for hyper parameters we felt it was most beneficial to consider the best possible values for the maximum tree depth in the base learners and the number of estimators (gradient boosted trees) used in training.

For our hyper parameter search we considered the following possible values.

Maximum tree depth = {10, 25, 50, 75, 100, 150}.

Number of estimators = {150, 250, 300, 350, 400, 500}.

The best value for XGBoost by balanced accuracy score is a depth of 10 and 400 estimators.

3. Random Forest Classifier

Our random forest classifier is from the SciKit Learn Ensemble library [130]. We allow this method to be trained using a ‘gini’ criterion. The ‘gini’ criterion calls for the use of the Gini impurity measure to evaluate the split quality of a given node. This function aims to minimize the overall Gini impurity measure, which is a function of the probability of incorrectly classifying a patient row within our model. The maximum depth and number of estimators are found using a 5-fold cross validation. For reproducibility, the random state is held constant at 1.

For our hyper parameter search we considered the following possible values.

Maximum tree depth = {10, 25, 50, 75, 100, 150}.

Number of estimators = {150, 250, 300, 350, 400, 500}.

The best value for random forest by balanced accuracy score is a depth of 10 and 500 estimators.

The hyperparameters were chosen by the balanced accuracy score. The balanced accuracy score was chosen because it is likely the best metric to evaluate a dataset with a class imbalance this large. Other metrics such as accuracy, PPV, or NPV are known to be poor performance measures when there is large class imbalance as the models can predict 1s or 0s for everyone and receive close to perfect scores by these metrics. After we choose the best values for each of the hyperparameters, we make final predictions for our out-of-sample testing set across each fold. We report the mean and standard deviation across all five CV-folds for each of the output metrics for our validation testing set in Table 2.6 and in the out of sample testing set shown in Table 2.7.

Post Prediction Feature Analysis

To analyze the most important and relevant features from the three models that are present in this paper we select the top 60 features by the mean odds ratios for the logistic regression model, and feature importances for both the XG boost and random forest models. That is, we take the top 20 features from each model and compare the overlap between the top selections. We present the features that were considered in the top 20 by at least two models in Table 2.5. In this table we also indicate which models considered each feature important and provide the logistic regression coefficient for each feature. All of the variables in Table 2.5 were assigned a non-zero coefficient across all

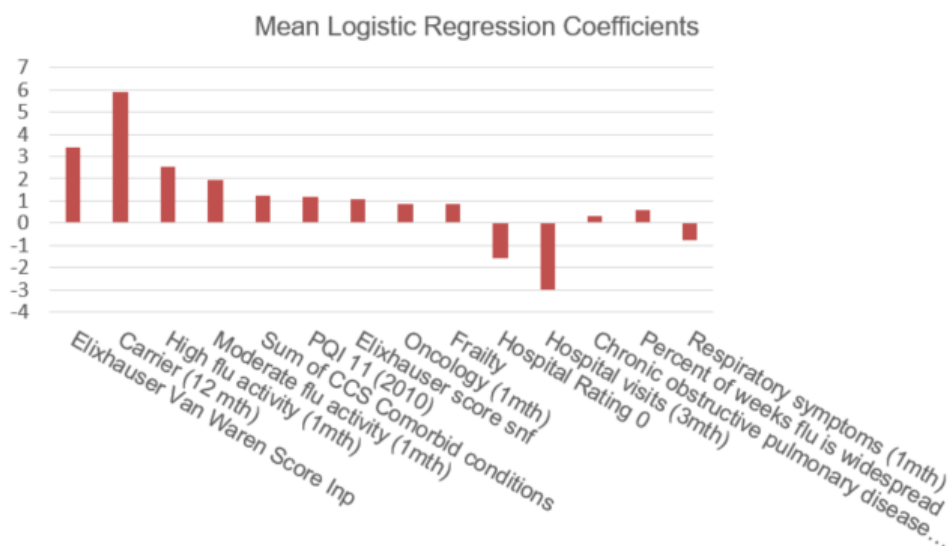


Figure 2.4 Mean logistic regression coefficients identified by at least two of the three models in the post prediction analysis. All variables identified in this analysis were assigned a nonzero logistic regression coefficient across all five folds.

five folds during the logistic regression training iterations. We did not report significance levels or confidence intervals because the logistic regression model was trained using 'L1' regularization or in other words a LASSO algorithm.

2.4 Results

Post-Prediction

Table 2.5 shows the results of the post-prediction analysis where we collect the top 20 variables by odds ratios and feature importances for the logistic regression and XG boost and random forest models, respectively. Here, we report the variables that appeared in the top 20 analysis by at least two of the models. We indicate which model had each variable in the top 20 defined in this manner, whether the variable is individual, community, or provider level and the logistic regression coefficient that each variable was assigned. All of the variables in this table were assigned nonzero logistic regression coefficients across all 5 CV folds and the mean coefficients are given in Figure 2.4. One variable, the Inpatient Elixhauser van Walraven score, was selected by all three models. The thirteen remaining variables were selected by different combinations of the three models. Of the fourteen variables chosen in this analysis 10 were individual, 3 were community, and 1 was a provider level variable. Six of the variables rely on the month prior to each index date and require the most up to date data available. These six are high flu activity, moderate flu activity, oncology visit, percent weeks widespread flu, and respiratory symptoms, and frailty when claims were available to update the frailty score. A complete list of each variable, logistic regression coefficient, and feature importances

Table 2.5 Collection of the top 60 variables, e.g., top 20 by odds ratio, and feature importance for logistic regression and the ensemble methods, respectively. The type indicates "Ind." for individual, "Com." for community, and "Prov." for provider level variables. The coefficient shown is the mean coefficient from the logistic regression model across 5-CV folds where each feature in this table was assigned a nonzero coefficient across all 5 folds using 'l1-regularization'. Last, the model number indicates whether the variable was a part of the top 20 for 1:logistic regression, 2:XG boost, or 3:random forest. Confidence intervals are not generated due to the regularization term used in model training.

Variable	Type	Mean Coefficient (LR)	Model(s)
Inpatient Elixhauser van Walraven Score	"Ind."	3.42	1-3
Carrier (12mths)	"Ind."	5.89	1,3
High flu activity	"Com."	2.56	1,2
Moderate flu activity	"Com."	1.94	1,2
Sum CCS	"Ind."	1.23	1,3
Pneumonia Inpatient Claims (2010)	"Ind."	1.20	1,2
SNF Elixhauser Score	"Ind."	1.09	1,3
Oncology visit (1mth)	"Ind."	0.86	1,2
Frailty Hospital Rating 0	"Ind." "Prov."	0.84 -1.56	1,2 2,3
Hospital visit (3mths)	"Ind."	-2.97	2,3
COPD	"Ind."	0.33	2,3
Percent weeks widespread flu	"Com."	0.61	2,3
Respiratory symptoms (1mth)	"Ind."	-0.78	2,3

is provided in Appendix B.

Next, we can look at the distribution of prediction scores according to final prediction classifications. In Table 2.5 we illustrate this distribution from the logistic regression model. We observe that the true negatives in the model are more uniform, compared to the false positives which yield a

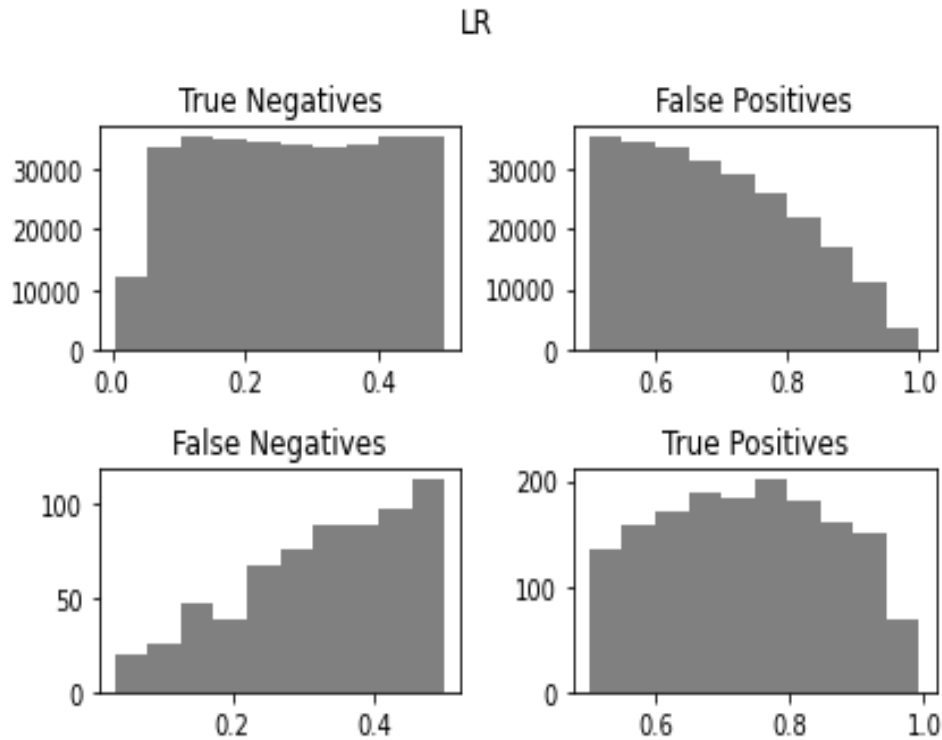


Figure 2.5 Histogram of the logistic regression prediction probability distribution for all patient rows.

more right-tailed distribution. The false negatives appear to form a left-tailed distribution whereas the true positives yield a more uniform distribution. This highlights the model's confusion in trying to discern the false positives from the true positives. The false positive predictions are more concentrated between 0.5 and 0.65 prediction probabilities. Distribution graphs for XG boost and random forest are provided in Appendix B.

Model Performance

We report the mean performance metrics across the 5 validation testing sets in Table 2.6. All performance metrics reported in Table 2.6 - 2.7 are reported for the out of sample testing set where predictions are made from each of the 5 training folds and the standard 0.5 threshold is used to predict a hospitalization or not from the score. We observe that all three models have a relatively low precision, or positive predictive value, due to the rare event occurrence in the data, which also yields similarly low f1-scores. The recall, or sensitivity, was highest for the random forest, followed by the logistic regression model. The recall measure, also known as the true positive rate indicates that the logistic regression and random forest models were able to make more correct hospitalization predictions than the XG boost model. The accuracy measure is about 0.23 and 0.25 times higher for the XG boost model compared to the logistic regression and random forest models, respectively. However, the balanced accuracy measures for all three models within 0.025 points of one another.

The closeness of these scores compared with the differences across recall scores suggests that the logistic regression and random forest models correctly identify more hospitalizations compared to the XG boost model, which correctly labels more non-hospitalizations in the set. This is supported by the specificity, or true negative rate, which is highest for the XG boost model compared to both logistic regression and random forest.

Next, we can consider the discriminatory powers of each of the models. Table 2.6 indicates that all models perform within 0.01 points of one another by ROC AUC score. Similarly, Figure 2.6, which graphs the false positive rate and true positive rate for the out of sample testing set across all 5 CV-folds, illustrates that the ROC curves of all of the models overlap one another. In Figure 2.6(a) we can see that approximately higher than halfway through the curve, or slightly higher than 0.5 probability threshold, there is a switch between which model is leading, or has highest area under the curve. Prior to halfway, the logistic regression model is leading, where after the halfway point the random forest becomes the leading model. Overall the areas under the curve are similar, as indicated in Table 2.6, but we can see that the discriminatory power for the logistic regression slightly decreases with an increase in the probability threshold, whereas the random forest's discriminatory power slightly increases. We also observe in Figure 2.6(b) that all models perform similarly according to the precision recall curve, where the XG Boost model begins as the leading model, and is overturned by the logistic regression model at a recall value of approximately 0.1 according to this curve.

In Figure 2.7 we compare the three models in this study with that of the closest existing model in the literature. It is important to note that the testing sets and variables used in each study are very different from one another, so while we can observe trends we cannot use these numbers to strictly compare the models and determine a "best" model. More about the differences between the study designs is provided in the Discussion section below. Uetmastu, et al. report sensitivity, specificity, PPV, and NPV for the optimal model presented in their study. We observe that no one model outperforms the others by all three metrics. The logistic regression and random forest models yield higher sensitivity and NPV values. The higher sensitivity values indicate that these two models are able to identify more of the events that occurred in the study compared to the model in the comparative study. The mean PPV value reported for all three models in this study is approximately 20% of what was reported in the Uematsu study. We describe below how these differences directly relate to the differences in the event rate in the study population.

2.5 Discussion

Interpretability was a main goal and focus during the design of this study. As such, the features were carefully designed and selected to create an interpretable framework for administrators or others in public health who may utilize the information gained from expanding knowledge of risk factors. We aggregated and synthesized data to understand what individual, community, and provider level variables further our understanding of an individual's 30-day risk of a potentially

Table 2.6 Mean value and standard deviation (·) for each metric across all 5 validation testing sets.

Model	Precision [PPV]	Recall [Sensitivity]	F1 Score	ROC AUC	Accuracy	Balanced Accuracy	NPV	Specificity
Logistic Regression	0.0061 (0.0011)	0.7279 (0.0364)	0.0121 (0.0022)	0.6875 (0.0105)	0.5430 (0.0449)	0.6351 (0.0100)	0.9981 (0.0005)	0.5422 (0.0449)
XG Boost	0.0076 (0.0012)	0.4434 (0.0345)	0.0149 (0.0023)	0.6951 (0.0216)	0.7763 (0.0152)	0.6105 (0.0181)	0.9972 (0.0008)	0.7776 (0.0151)
Random Forest	0.0060 (0.0012)	0.7382 (0.0572)	0.0118 (0.6950)	0.6950 (0.0097)	0.5254 (0.0538)	0.6314 (0.0054)	0.9981 (0.0003)	0.5245 (0.0544)

Table 2.7 Mean and standard deviation for the out of sample testing set across 5 CV-folds for each model trained in the study, compared to the optimal model presented in Uematsu, et al. in 2017.

Model	Sensitivity	Specificity	PPV	NPV
Uematsu, 2017	0.6600	0.6500	0.0320	0.9910
Logistic Regression	0.7119 (0.0055)	0.5612 (0.0072)	0.0064 (0.00)	0.9980 (0.00)
XG Boost	0.4012 (0.0085)	0.7922 (0.0080)	0.0077 (0.0001)	0.9970 (0.00)
Random Forest	0.7475 (0.0118)	0.5216 (0.0060)	0.0062 (0.0001)	0.9981 (0.0001)

preventable pneumonia related hospitalization. First, we will highlight the insights gained from the individual, community, and provider level variables from each of the three models. Then we will relate our model to the closest known existing model in terms of model performance and study design. Lastly, we will describe some challenges and limitations from this work before providing concluding remarks.

Considering the results in Table 2.5 we can discuss the variables selected by each of the three models in the post prediction analysis. Looking at this collection of top 20 variables per model, there is one variable that was grouped by all three models. This variable is a comorbidity score computed from inpatient claims, the Inpatient Elixhauser van Walraven weighted comorbidity index [169]. This feature is updated anytime an individual has a new visit to an inpatient facility. The van Walraven version of the Elixhauser score was created to provide weights to certain comorbidities end pre-existing conditions that can increase an individual's mortality risk. It is not surprising that the weighted version of this score would indicate a higher likelihood of an individual to be at risk for a pneumonia hospitalization in the next 30 days. We also report that the mean logistic regression coefficient for this variable is 3.42, which can be thought of as any unit increase in the van Walraven comorbidity score is associated with an increase in the log-odds of an individual's likelihood of pneumonia hospitalization by 3.42.

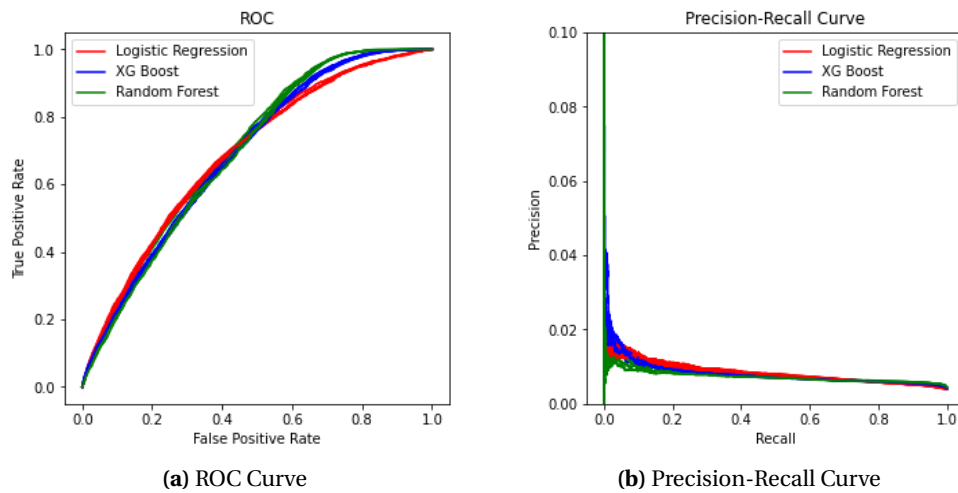


Figure 2.6 ROC and Precision Recall Curve for the out of sample testing set predictions for each CV fold and each model in the study.

There are fourteen total variables that were selected by two of the models where one is provider level, three are community level, and the remaining features are all individual level variables. The community level variables involve the flu mean activity level in the month prior to the index date. Specifically, the variables indicate whether the flu mean activity level within the state was moderate or high, with a reference group of minimal, and the percentage of weeks that flu was widespread in the month prior. All three variables have positive logistic regression coefficients and were assigned nonzero coefficients across all five folds in the training analysis. The provider level variable that was chosen was the hospital overall rating of 0 in comparison to the reference group for this variable which is a hospital overall rating of 5. The Center for Medicare and Medicaid Services (CMS) provides Hospital ratings from all facilities that report CMS. The hospital overall rating of 5 is the highest rating that a facility can receive and indicates top quality of care. In our dataset, a hospital rating of 0 indicates a missing value, either from the CMS rating or because the individual has not had an inpatient visit for us to compute the value. The model includes interaction terms between hospital overall ratings and inpatient visits, for which the interaction term between hospital rating 0 and ever been in an inpatient facility was assigned a positive coefficient across all 5 CV-folds. This could indicate that a hospital rating of 0 is acting as a proxy for the patient’s overall health if they have never had a visit to an inpatient facility.

Other variables mentioned in Table 2.5 indicate that an individual who had similar pneumonia inpatient visit in 2010, COPD, higher inpatient or SNF Elixhauser van Walraven scores, carrier claims in the last 12 months, greater sum of comorbid conditions, or a higher frailty index tend to have higher risk prediction scores. Some others, such as the respiratory symptoms in the past month have a negative logistic regression coefficient. It is important to note that these results are performed on a filtered set from the general population where everyone needs to meet at least one of four filtering criteria, and one of the four is having respiratory symptoms present in your claims diagnoses

in the month prior to each index date. Taking a look at Figure B.3 in Appendix 2 we can see the hospitalization rates for patients who were included in the set due to each condition and observe that patients included due to having a nursing home visit or respiratory symptoms in the month prior, have a smaller hospitalization rate compared to those who may have been included due to meeting one of the other conditions. That is to say that having respiratory symptoms in the month prior to each index date puts an individual at-risk, but not as high of risk as those who may be at risk due to one of the other conditions. The results also indicate that a hospital visit in the previous three months may lower an individual's risk of hospitalization which could be due to the fact that this individual already experienced a hospitalization or has otherwise received treatment for the illness that required them to seek care.

To the best of our knowledge, the closest existing model in the literature for predicting pneumonia hospitalizations comes from a study presented by Uematsu et al. in 2017 [167]. Although we compare the results of our model to the best one presented in Uematsu, et al. there remain significant differences across study design that make a direct comparison difficult without addressing and acknowledging key differences. In the study presented by Uematsu et al., patients participated in a nationwide program for the screening and management of lifestyle-related diseases. Our study was not a part of a nationwide screening program and instead included a 5% sample of the entire population. Furthermore, our model focuses on 30-day prediction versus the five year window allowed in the Uematsu study. It is possible that the individuals who participated in a nationwide screening program were more likely to be health conscious and risk-averse people. This study population selection also greatly reduces the sample size which can relieve models and computers from large data that are difficult to fit due to memory constraints, i.e., the study population in Uematsu et al. was a 55,842 sized cohort compared to our 454,560 unique patients with a total 2,842,260 row dataset. The event rate in Uematsu's population is 1.68% compared to our 0.39% after filtering. The percentage difference between the event rates is roughly equal to the percentage difference between the Uematsu's model's PPV, 0.032, and the average across all three of our models. Additionally, the nationwide program included health checkup data, such as hemoglobin count, body mass index, ECG, fasting blood glucose, blood pressure, and some other patient-specific survey data such as reported symptoms, alcohol consumption, daily walking habits, and if they were a current smoker. These specific variables were of the highest odds ratio for their model, and almost all of them are either not available or difficult to estimate or compute using solely administrative claims data. This suggests that if we reviewed a subset of our population at higher risk and with an event rate closer to that in the Uematsu study, and the ability to compute similar variables or proxies for these biomarkers, our predictions may yield an improved PPV. Additionally, a comparison to other predictive models for specific subgroup populations, e.g., readmissions or congestive heart failure could be possible if we consider a specific subgroup for comparison, although the models would need to be re-calibrated and retrained to before producing predictions. Above all else, the comparative model in the Uematsu study has higher PPV, but is lower across all reported measures compared to each of the three models presented in this work and does not include analysis of

community level variables. This means that our model is more important for short term decision making and informs more about community wide intervention planning.

Challenges and Limitations

Unsurprisingly and consistent with other models in the literature, our results show low positive predictive values due to the naturally occurring imbalance that is observed within the data and in the real world. The class imbalance causes the machine learning models to tend towards predicting all input vectors as the majority class. In order to combat this effect we had to instantiate a bi-level filtering approach; first by current status as of the index date for each prediction month and then by medical histories and more recent medical claims information available as described in Figure 2.2. This method reduced the overall data set by 72% and retained 95% of all events in the data set. Even with this reduction, we needed to develop a sampling algorithm in order to train models that can learn something from the data. To further reduce the impact of the large class imbalance we used hyper parameter tuning to enforce greater weights on incorrectly classifying minority class labels and try to avoid overfitting to the training data.

The information available from administrative claims data is limited to what is billed at the discretion of the provider. We cannot view or infer specifics of each visit such as electronic healthcare data, lab results, prescription drugs ordered or taken, etc. Claims data in general is subject to incomplete, inaccurate, or otherwise missing data as well as nonspecific diagnosis codes that are appropriate for billing, but not necessarily to help with creating a health snapshot of an individual at a point in time. There also exists some time lag between when an individual receives care and when a claim or billing statement is finally processed. This also aids in the missing-ness of available data over time as well as potentially causes differences between reality and billing dates which can have grave impacts on models strictly created to inform an immediate 30-day risk evaluation.

Another limitation lies in the way the data is aggregated. A combination of the above-mentioned limitations paired with data storage and computational complexity led to the approach to develop monthly aggregated patient rows. This limits our abilities to capture immediate changes in an individual's health at a point in time. Storage concerns notwithstanding, it could be interesting to consider and compare more immediate changes, perhaps on a weekly level, to more accurately detect when a significant shift in an individual's health could lead to a pneumonia-related hospitalization. However, as we have concluded from this analysis, more granular level data from an individual's health care provider such as lab results, physical examination findings, bio-markers, and x-ray results could all be useful towards increasing the discriminatory and predictive power of the models we present in this study. This, however, would require the aggregation of multiple data sources to pair results of these findings with recent visits to a healthcare provider. If in fact we found these recent visit markers to be significant for our model development, then we would potentially have to only include individual's who had a recent physician visit to measure the significance of each of these new findings. This would require more recent visits to a doctor's office prior to each event, which we have seen from different attempts at patient filtering could reduce the number

of PQI-11 events which are captured in our set overall. If we had the access, the incorporation of drug data could prove to be a useful indicator of individual health status, comorbid conditions, and health behavior that we did not have access to in this study.

2.6 Conclusion

Overall, we find that we are able to contribute the following:

- A feature-engineering method to develop a dynamic framework for administrative claims data for pneumonia hospitalization risk.
- A multi-level factor approach, i.e., individual, community and provider level analysis on an individual's pneumonia hospitalization risk within 30 days.
- A filtering method based on medical history to identify a population at higher risk for pneumonia hospitalizations.
- A sampling algorithm to simultaneously reduce the training space and sufficiently train the models on data from the event versus non-event population at different points in time.
- Predictive models to assess an individual's 30-day risk of pneumonia hospitalization from the general population without the use of personal biomarkers from specific checkup data.
- Population-level focused analysis, e.g., as appropriate for a stakeholder like a payer or hospital, across a portfolio of patients.

This study supports the development of a feature engineering method to develop a framework that is semi-dynamic using administrative claims and publicly available data. We present the algorithms and scripts to create variables that are used to assess an individual's health at a point in time and report the ones that are statistically significant when comparing our event and non-event groups in terms of their created patient rows in the patient demographic analysis in Table 2.1. These variables can be used to aid other point scores or prediction measures that have been previously used to identify risk of severe disease or mortality from a pneumonia-related illness.

The predictive models explored in this study provide a starting point for the framework development of the large scale of information that is needed to more accurately predict who is most at risk of an unplanned hospitalization within 30 days. Future work to improve the predictions could include the inclusion of more detailed data such as drug and prescription data or health checkup or electronic health record data. The inclusion of these data will allow researchers to compute the variables known to have high informative power on general risk for hospitalizations.

Another possible area for improvement would be to incorporate later years of data to get a more complete health history of an individual. Some comorbidities or other procedures or health histories prior to the years of data available to us could have created a more accurate snapshot of a patient's

health status. Another future direction could be to utilize the same methods and procedures to analyze the usefulness of administrative claims data and other publicly available data sources on other adverse events or potentially preventable outcomes.

With the data available to us for this study, we present a filtering framework that allows for the reduction of beneficiary identifiers and associated most at-risk patient row months dependent on any of the following factors: (i) a history of a pneumonia-related hospitalization; (ii) nursing home visits in the previous month; (iii) respiratory symptoms in the previous month; and (iv) claims within the last 6 months, indicating some sort of medical visit, and a most-recent Elixhauser score computed from inpatient claims diagnosis codes. This filtering approach significantly reduced our original patient-row set and allowed us to develop a framework to allow us to exclude anyone who is not at risk of a pneumonia-related hospitalization. These factors are consistent with what has been presented in the literature and are now scalable in terms of claims data.

The assessment of individual, community, and provider level information from each of the models tells us that while individual level patient history remains crucial to assess risk of a 30-day unplanned hospital admission we still need to be cognizant of other factors that could increase an individual's risk further. Some of these factors identified in this analysis include personal health history such as the presence of comorbidities, as seen in Table 2.5. Some targeted intervention strategies would be to identify populations that are characterized by these risk factors and support vaccine campaigns, health advisory communications and messages, and reminders from health care providers to attend annual visits or checkups, particularly for those who meet any of our filtering criteria.

Pneumonia-related hospitalizations are a serious concern because they are difficult to predict and an individual that is hospitalized with pneumonia is often presented with an increase in mortality risk due to disease-related complications. As such, it is necessary to implement interventions to avoid severe disease in the first place and limit these mortality risks. These models support the literature in terms of important risk factors and can help identify at-risk populations from an administrative perspective without requiring specific health check-up data. As such, the models can be employed on any size data set of Medicare beneficiaries to aid public health communication tactics to inform the public about the dangers, risks, and prevention strategies for severe cases of pneumonia.

Acknowledgements

Ling Mao, Kimia Vhadat, Dr. Sara Shashaani and others from North Carolina State University and James McKenna and Dr. Joseph Agor and others from Oregon State University contributed to the overall study design, framework, data and feature development, and early discussions of the overall problem.

CHAPTER 3

DATA-DRIVEN INTERVENTIONS TO PREVENT PNEUMONIA HOSPITALIZATIONS

3.1 Introduction

As a leading cause of both morbidity and mortality, the Centers for Disease Control and Prevention (CDC) consistently identifies pneumonia as one of the top ten contributors to all-cause mortality for all age groups within the United States, where 2018 data suggests that the percent of total deaths attributable to influenza/pneumonia is higher for more vulnerable populations, i.e., 2.33% for those ages 65 years and over and 3.19% for those ages 1-4 years old, compared to an average of 1.36% to all other age groups [24]. Other studies have shown that pneumonia is responsible for approximately 1.5 million unique hospitalizations each year, where 1 in 3 hospitalized patients dies [111]. Even those who are able to recover from the disease are known to suffer from long-term consequences such as chronic lung diseases, cardiovascular disease, and overall worsened quality of life [38, 79, 119, 128]. Accordingly, economic evaluations suggest that the annual estimate of total healthcare associated costs for pneumococcal disease is approximately \$13.4 billion in the United States [165].

Currently, there are a handful of interventions that can be employed to try to reduce the burden of pneumonia in terms of incidence of severe disease, mortality risk, and overall cost expenditures that require decisions from policy makers about which interventions to utilize and whom to target with them. For stakeholders like payers, hospitals, and some providers managing a set of patients, there is a population of patients with varying risk levels. In the simplest case, if there is one intervention available and an exact budget, policy makers might decide to distribute interventions to those who have the highest risk. However, resources are always limited so it is important to think about not

only the cost of an intervention, but also the cost if an intervention is not applied, and the cost if the intervention is applied to too many people [12]. Considering these factors, it may be desirable for policy makers to make decisions that balance overall costs and outcomes, whether successful or not, across a set of patients. In this case, the decision becomes more complex as there are multiple factors to consider such as a patient's risk, the types of interventions available, and uncertainty regarding whether the intervention will be successful at preventing severe outcomes.

Furthering the complexity of this decision, not only must policy makers consider the probability of success of an intervention, but also whether it is beneficial to consider a second, cheaper intervention. Some examples of intervention options varying in costs and effectiveness include antibiotic or antiviral treatments, vaccinations, sending out communications advocating for a healthy lifestyle, or receiving phone calls from doctors or nurses reminding patients to follow up with their health care providers or complete medication [34, 116, 123, 135, 150]. In all, there exist a wide range of well-known preventative measures that can be administered or utilized to help protect particularly vulnerable populations. While certain preventative measures are extremely cost effective and more accessible to a wide audience, others like vaccines require physician assistance and have a higher unit cost per intervention [48, 77]. Therefore, it is worthwhile to identify and target our most at-risk populations to inform the most appropriate intervention recommendation. It is well known that certain populations remain at increased risk for developing severe pneumonia and there exist a selection of risk-severity scores for patients [6, 30, 73, 76, 94, 107, 149, 167, 170]. While these studies offer insights towards quantifying risk factors, they do not consider specific prevention strategies.

The purpose of this chapter is to provide a one-time decision on who to target with one of two interventions, if any, of varying costs based on risk scores. Specifically, the decision will determine the number of people on which to intervene with the higher cost intervention. Furthermore, we model interventions as one-time, single period purchase products to inform how many interventions an administrator should purchase for a population of arbitrary size, which considers variations that can occur among hospital systems [83, 97], and to whom these interventions should target based on the one-time decision regarding risk scores.

In the one-intervention case, the decision problem shares some similarities with the well known newsvendor problem with unknown demand. A key difference here is that we consider individual risk scores and the translation between the number of people who receive interventions and the corresponding number of hospitalizations, prevented or not. We also assume that risk scores are available and in reality they can be derived in many ways, including from prediction algorithms [73, 76, 107, 167, 170]. We explore data-driven approaches to estimate the unknown demand and provide recommendations on whom to target with one of two interventions.

The rest of the chapter is organized as follows. In section 3.2 we provide a literature review to highlight this chapter's contributions with respect to intervention planning. Sections 3.3 - 3.4 presents the model formulation, the data and methodology used to compute the pneumonia risk scores and hospitalization indicators, and introduces a heuristic to provide an intervention policy recommendation. Section 3.5 provides theoretical model verification, numerical results and insights

from our optimal policy recommendations. All derivations and proofs for sections 3.3 and 4 are provided within each section or in the Appendix C. Section 3.6 contains discussions and concluding remarks.

3.2 Literature Review

Newsvendor models, a subgroup of revenue management models, have been used to inform optimal solutions for a wide variety of problems in healthcare such as resource allocation, capacity planning, and medical inventory planning [151]. Specifically, newsvendor models are used to consider tradeoffs between costs and demand and offer solutions for decision makers who need to consider a one-time upfront purchase of a product before knowing the actual product demand and have been used in healthcare settings to inform intervention and allocation planning [151, 156]. Furthermore, newsvendor decisions have been explored under differing cost parameters, where the optimal newsvendor quantity depends on demand and product prices [153]. Chick et al., model influenza vaccines as one-time newsvendor-type products to minimize costs associated with influenza vaccine supply chain coordination [41]. Other studies use newsvendor models to consider optimal capacity allocation for inpatient beds and balance between operating room reservations and surgery demand [125, 178]. While each of these models incorporate inventory management to improve healthcare none consider intervention planning based on individual risk scores.

Resource allocation problems have been studied to optimize public health outcomes and inform optimal intervention planning among cohorts or population subgroups. Zaric et al., consider the optimal investment in HIV prevention programs in order to either maximize quality-adjusted life years (QALYs) gained or number of infections averted given budgeting constraints, but do not consider individual risk nor do they optimize costs [177]. Hynninen et al., consider patient segments and develop a two phase optimization model to maximize the expected health benefits given a chosen policy-level objective [85]. Their model supports intervention allocation based on population segments, but they do not consider multiple interventions based on risk scores, and while they provide a cost-effectiveness analysis to illustrate which optimal policies should reasonably be considered by decision makers, they do not optimize costs.

In terms of monetary gain, there are several studies that indicate a considerable return on investment over time when interventions of varying costs and efficacy are applied to specific populations. Yildirim et al. quantify the effects of a selection of interventions for a population with asthma and report reductions in utilization and medication costs in post intervention time periods compared to pre-intervention [176]. Another study forecasted reduced healthcare costs and improved outcomes for interventions using specific selection criteria, e.g. high healthcare utilization [159]. These studies highlight the potential cost savings for targeted interventions of varying costs and efficacy.

This chapter expands the newsvendor model to solve a problem with two classes of interventions to be distributed based on derived risk scores to minimize expected costs associated with pneumonia hospitalizations while considering uncertainty from the decision maker's perspective. We summarize

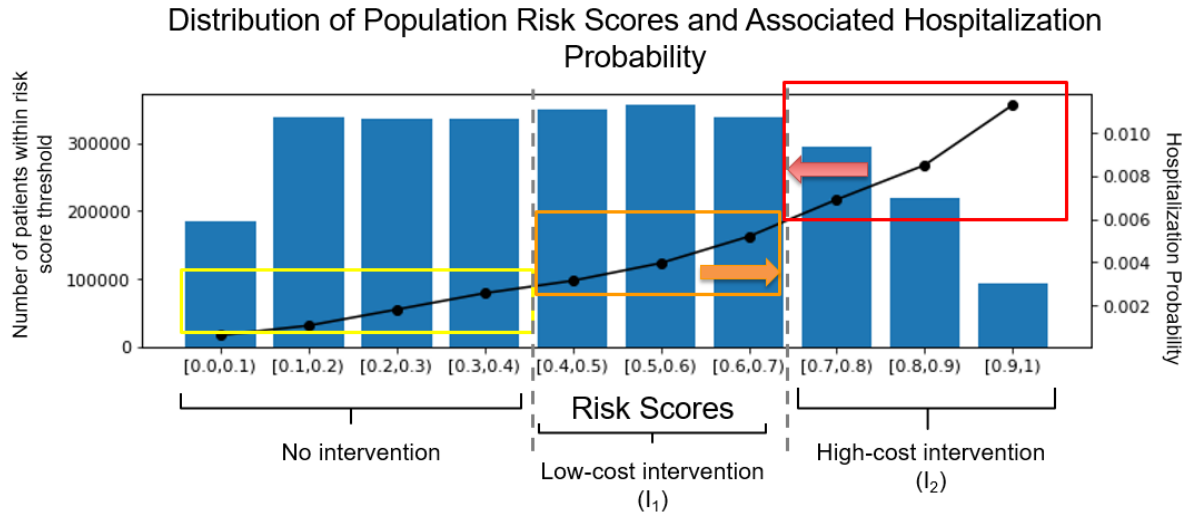


Figure 3.1 Risk score distribution and associated hospitalization probability. Here, the yellow block would indicate no interventions, whereby the two red blocks could indicate low and high cost interventions for the left and right red blocks, respectively. The arrows identifies the single period decision point for a given x_1 value.

the following contributions. First, we extend the newsvendor model to determine a single period decision aimed at minimizing the expected costs associated with two classes of interventions and pneumonia hospitalizations. The model results in a structural solution that closely resembles the well-known newsvendor critical ratio, but requires numerical solutions due to the translation between number of patients given an intervention and the number of hospitalizations. Second, we provide a heuristic to determine the best number of hospitalizations an administrator should target with each intervention and whom to target with each depending on individual risk scores. Finally, we show that our model provides interpretable solutions that are computationally inexpensive to compute and generalizes well to other notions of risk.

3.3 Model Formulation

In this section we focus on problem setting. We first begin with providing a general overview of the main concept of this chapter and draw relationships to revenue management to highlight why the newsvendor model is an appropriate choice for our analysis. Then, we review necessary nomenclature and define the cost function used in our optimization model. We finish with our complete model formulation and properties of the optimal solution.

We consider a multiclass problem where we want to determine the optimal deployment of two interventions based on costs, effective probabilities, and individual risk scores. We consider a population of varying risk scores, as seen in Figure 3.1 where each risk score grouping has an associated hospitalization probability, that is nonlinear increasing with respect to individual risk score. We define our optimal decision by identifying two decision thresholds, x_1 and x_2 , that partition

our population into three groups: individuals with risk scores (i) below x_1 who receive no intervention (ii) between x_1 and x_2 who receive the less effective intervention, I_1 , and (iii) above x_2 who receive the more effective intervention, I_2 . We observe that for any selection of x_1 , the decision x_2 can be found by minimizing the expected costs of distributing the two interventions to the remaining population. We require that the policy targets the more effective intervention towards individuals with higher risk scores first. We discretize over possible values of x_1 and note that for each value of x_1 the remaining decision variable x_2 can be determined using a newsvendor formulation considering the optimal number of hospitalizations targeted with the more effective intervention. We summarize our model's relations to revenue management literature in the following ways.

First, the number of hospitalizations targeted with the more effective intervention can be viewed as the “policymaker’s product” [4], for which the policymaker can control the number of hospitalizations to target. We take into consideration the nonlinear relationship between the number of more effective interventions available and the number of hospitalizations that can effectively be targeted with them given the hospitalization distribution among different risk score groups. We incorporate this tradeoff into the cost function.

Secondly, we observe that the hospitalization demand follows an uncertain, unknown distribution. Due to the uncertainty in the demand, we take the standard newsvendor approach to minimize expected costs [100]. Since our demand distribution is unknown, unlike with the classic newsvendor assumptions, we consider an alternative method to estimate the demand distribution using sampled, empirical data and sample average approximations [101, 154, 178]. We recognize that in reality, administrators need to make a one-time decision regarding intervention planning prior to knowing the true hospitalization demand. In this sense, the newsvendor model can provide an optimal recommendation [4, 156].

Thirdly, we aim to provide an optimal recommendation that takes into consideration the potential cost trade-off between not fully and over satisfying the demand, prior to knowing the demand realization [100]. Note here that the overall cost that we are considering is affected by both the number of interventions given and by the number of hospitalizations prevented. The relationship between the costs and the number of interventions given is not always linear because decision makers should aim to target higher risk individuals first, especially if there is limited supply, to increase the likelihood of preventing a hospitalization as illustrated in Figure 3.1. As such we define a cost function such that it considers these penalties, e.g., giving too many interventions and not preventing hospitalizations, for a single iteration, i.e., specific x_1 value in our model. We note that the pneumonia hospitalization cost could also include indirect cost that is associated with the hospitalization or resulting illness, e.g., loss of productivity or morbidity and mortality. Here the penalty for not satisfying our demand, oftentimes referred to as underage costs, occurs when an individual is targeted with the less effective intervention when they were at risk for pneumonia hospitalization. The penalty for exceeding our demand, e.g., overage costs, occurs when an individual is targeted with the more effective intervention when they are not at risk for a hospitalization. Here, the decision variable, q , is the number of interventions targeted with the high cost intervention. We define a

function $r(q)$ to map the number of interventions with the number of hospitalizations prevented. We assume that the unit costs and effective probabilities for each of these interventions are known and static such that the only variable component in our cost function is the hospitalization demand.

We define the following notation.

Definition 3.3.1. Notation

x_1 = minimum risk score threshold for which all individuals will receive either low or high cost intervention

x_2 = minimum risk score threshold to assign high cost interventions.

C_{i1} = Unit cost of I_1

C_{i2} = Unit cost of I_2

pe_1 = Effective probability of I_1

pe_2 = Effective probability of I_2

C_h = Pneumonia hospitalization cost

P_0 = Population with risk scores below x_1 who will not receive an intervention.

P_{1+2} = Population with risk scores above x_1 .

D_0 = Random variable for number of hospitalizations in P_0 .

D_{1+2} = Random variable representing the number of hospitalizations over population, P_{1+2}

$F_0(\cdot), F_{1+2}(\cdot)$ = Cumulative distribution functions of D_0, D_{1+2}

$f_0(\cdot), f_{1+2}(\cdot)$ = Probability density functions of D_0, D_{1+2}

$r(q)$ = The number of individuals required to intervene upon in order to intervene upon q events when interventions are assigned by decreasing risk score.

Assuming x_1 is identified, our cost function for a given number of hospitalizations targeted with I_2 for our population can be defined in the following manner.

$$TC(q, x_1) = \underbrace{C_h D_0}_{\text{Cost of unprevented hospitalizations for } q \text{ targeted events}} + \underbrace{W_1 C_{i2} + W_2 C_{i1}}_{\text{Cost of interventions for } r(q) \text{ people}} + \underbrace{Z_1(1 - pe_2)C_h + Z_2(1 - pe_1)C_h}_{\text{Cost of unprevented hospitalizations with interventions and } q \text{ targeted events}} \quad (3.1)$$

where

$$\begin{aligned} W_1 &:= \min\{P_{1+2}, r(q)\} \\ W_2 &:= P_{1+2} - W_1 \\ Z_1 &:= \min\{q, D_{1+2}\} \\ Z_2 &:= D_{1+2} - Z_1 \end{aligned}$$

Here we denote Z_1, Z_2 as the number of hospitalizations that that are covered by I_2, I_1 , respectively. In the cost function, we consider the instances where the interventions do not work and define the cost of unprevented hospitalizations. We make the following assumptions regarding the cost function and model formulation.

Assumption 1- $C_{i2} - C_{i1} \geq 0, p e_2 - p e_1 \geq 0$

Assumption 1 follows from an intuitive ordering of the two available interventions where I_2 , the more effective intervention, costs more than I_1 . If this were not the case then the optimal solution will always be to satisfy your at-risk population with the cheaper, more effective intervention.

Assumption 2- $r(q) := a q^2 + b q + c, a \geq 0, q \geq 0$

In order to accurately account for the effect of the changing event rate over different values of x_1 , we rely on $r(q)$ to map the cumulative number of hospitalizations to the cumulative number of patients when patients are ordered by decreasing risk score. Observe in Figure 3.1 that the problem exhibits convex, increasing properties. As such, we assume in that $r(q)$ is a convex, quadratic function in order to best fit the data and maintain convexity in our objective function and we provide additional arguments in Appendix C that $r(q)$ can always be estimated such that it is a convex polynomial of at most degree two. We define $r(q)$ for all non-negative q such that the inverse exists.

Also, we assume that a decision-maker is focused on maximization of the value function with no supply limitations. More information regarding the data and risk scores can be found in the Numerical Experiments section.

Finally, we take into consideration the uncertain distribution, F_{1+2} , of pneumonia hospitalizations and formulate our model to minimize expected costs.

$$\min_{q \geq 0} \mathbb{E}[TC(q, x_1)] = \min_{q \geq 0} \mathbb{E}[C_h Z_0 + W_1 C_{i2} + W_2 C_{i1} + Z_1(1 - p e_2)C_h + Z_2(1 - p e_1)C_h] \quad (3.2)$$

where

$$\begin{aligned}
W_1 &:= \min\{P_{1+2}, r(q)\} \\
W_2 &:= P_{1+2} - W_1 \\
Z_1 &:= \min\{q, D_{1+2}\} \\
Z_2 &:= D_{1+2} - Z_1
\end{aligned}$$

In expectation, when $r(q) \leq P_{1+2}$ yields:

$$\begin{aligned}
\mathbb{E}[TC(q, x_1)] &= C_h \int_{-\infty}^{\infty} x f_0(x) dx + C_{i2} r(q) + C_{i1} (P_{1+2} - r(q)) \\
&\quad + C_h q [1 - F_{1+2}(q)] (p e_1 - p e_2) \\
&\quad + (1 - p e_2) C_h \int_{x \leq q} x f_{1+2}(x) dx + (1 - p e_1) C_h \int_{x > q} x f_{1+2}(x) dx \quad (3.3)
\end{aligned}$$

The full derivation of $\mathbb{E}[C(q)]$ can be found in Appendix C. With this formulation, we can evaluate and utilize the properties of our expected cost function to determine the optimal solution in the following manner.

Theorem 1. $\mathbb{E}[C(q)]$ is convex with respect to q .

Proof.

$$\begin{aligned}
\frac{\partial \mathbb{E}[TC(q, x_1)]}{\partial q} &= C_{i2} r'(q) - C_{i1} r'(q) + C_h (1 - p e_2) - C_h (1 - p e_1) \\
&\quad - [F_{1+2}(q) C_h (1 - p e_2) - F_{1+2}(q) C_h (1 - p e_1) + q f_{1+2}(q) C_h (1 - p e_2) - q f_{1+2}(q) C_h (1 - p e_1)] \\
&\quad + (1 - p e_2) C_h q f_{1+2}(q) + (1 - p e_1) C_h (-q f_{1+2}(q)) \\
&= r'(q) (C_{i2} - C_{i1}) + C_h (1 - p e_2) - C_h (1 - p e_1) + F_{1+2}(q) C_h (p e_2 - p e_1) \\
\frac{\partial^2 \mathbb{E}[TC(q, x_1)]}{\partial^2 q} &= r''(q) (C_{i2} - C_{i1}) + f_{1+2}(q) C_h (p e_2 - p e_1) \geq 0, \forall q
\end{aligned}$$

□

Since we have convexity, the optimal solution to our unconstrained optimization problem, (2), can be found by taking the derivative with respect to q and setting it equal to zero, e.g., finding the stationary points of (3). We remark here that other assumptions on $r(q)$ would also satisfy convexity as long as $r''(q)$ had the appropriate sign. The result yields an optimal solution with a similar form to the well-known newsvendor critical fractile.

Theorem 2. *The optimal solution, q^* , satisfies the following equation.*

$$F(q^*) = \frac{C_h(p e_2 - p e_1) - (C_{i2} - C_{i1})r'(q^*)}{C_h(p e_2 - p e_1)}$$

Proof.

$$\begin{aligned} \frac{\partial \mathbb{E}[C(q)]}{\partial q} &= r'(q)(C_{i2} - C_{i1}) + C_h(1 - p e_2) - C_h(1 - p e_1) + F_{1+2}(q)C_h(p e_2 - p e_1) = 0 \\ \implies F_{1+2}(q) &= \frac{C_h(p e_2 - p e_1) - (C_{i2} - C_{i1})r'(q)}{C_h(p e_2 - p e_1)} \end{aligned}$$

□

Therefore, we can denote the overage, C_O , and underage costs, C_U , for this problem.

$$C_O = (C_{i2} - C_{i1})r'(q) \quad (3.4)$$

$$C_U = C_h(p e_2 - p e_1) - (C_{i2} - C_{i1})r'(q) \quad (3.5)$$

We can interpret these relationships in the following manner. Given q covered hospitalizations, $r'(q)$ is the marginal change in the number of people receiving interventions with respect to changes in q . In other words, the number of additional interventions needed to target one more hospitalization. C_O is then the change in intervention costs to plan to target an additional hospitalization, which is considered overage costs once all patients who will experience a hospitalization have already been targeted.

Similarly, we can interpret C_U , or underage costs. $C_h(p e_2 - p e_1)$ is the change in hospitalization costs covered by the more effective intervention, compared to the less effective intervention. The remaining terms, $(C_{i2} - C_{i1})r'(q)$, same as above, represents the intervention costs to cover an additional hospitalization.

Theorem 2 presents the optimal solution in the form of the critical fractile, which is also known to be of the following form. We note here that if $r(q) = q$ this reduces to the standard newsvendor problem and critical fractile. Given the defined costs, we obtain the following.

$$CR = \frac{C_h(p e_2 - p e_1) - (C_{i2} - C_{i1})r'(q)}{C_h(p e_2 - p e_1)} = \frac{C_U}{C_U + C_O} \quad (3.6)$$

In reality, it is helpful to consider the impacts of changes in the intervention costs and probability of effectiveness. We can determine the effects of changing the values of $p e_1$, $p e_2$, C_{i1} , and C_{i2} on the value of the critical fractile (CR). For the following results, we assume $q \geq \frac{-b}{2a}$. For the data with which we are working with this is true in almost all cases. In these instances, we have the following propositions.

Proposition 3. For all $q \geq \frac{-b}{2a}$ and assuming C_{i1} , C_{i2} , and one of p_{e1} or p_{e2} are constant, we can quantify the impact of the changing intervention's effectiveness on the optimal solution.

1. Per unit increase in p_{e1} , the optimal policy will be to give fewer people I_2 .
2. Per unit increase in p_{e2} , the optimal policy will be to give more people I_2 .

Proof. 3.1

$$\frac{\partial CR}{\partial p_{e1}} = \frac{(C_{i1} - C_{i2})r'(q)}{C_h(p_{e2} - p_{e1})^2} \leq 0^*$$

□

Proof. 3.2

$$\frac{\partial CR}{\partial p_{e2}} = \frac{(C_{i2} - C_{i1})r'(q)}{C_h(p_{e2} - p_{e1})^2} \geq 0^*$$

□

Similarly, we can evaluate the properties of (CR) with respect to changing cost values.

Proposition 4. For all $q \geq \frac{-b}{2a}$ and assuming p_{e1} , p_{e2} , and one of C_{i1} or C_{i2} are constant, we can quantify the impact of the changing intervention's cost on the optimal solution.

1. Per unit increase in C_{i1} , the optimal policy will be to give more people I_2 .
2. Per unit increase in C_{i2} , the optimal policy will be to give fewer people I_2 .

Proof. 4.1

$$\frac{\partial CR}{\partial C_{i1}} = \frac{r'(q)}{C_h(p_{e2} - p_{e1})} \geq 0^*$$

□

Proof. 4.2

$$\frac{\partial CR}{\partial C_{i2}} = \frac{-r'(q)}{C_h(p_{e2} - p_{e1})} \leq 0^*$$

□

Finally, we can observe the limiting behavior of the critical fractile as the cost of I_2 reduces towards the cost of I_1 .

Proposition 5. For all $q \geq \frac{-b}{2a}$ and assuming p_{e1} , p_{e2} , and C_{i1} are constant, as the cost of I_2 reduces towards the cost of I_1 , the optimal solution will be to give everyone I_2 .

Proof.

$$\lim_{C_{i2} \rightarrow C_{i1}} F(q^*) = \lim_{C_{i2} \rightarrow C_{i1}} \frac{C_h(p_{e2} - p_{e2}) - (C_{i2} - C_{i1})r'(q^*)}{C_h(p_{e2} - p_{e1})} \rightarrow 1^*$$

□

(*)Note each of these equations holds true $\forall q \geq \frac{-b}{2a}$. In the context of the numerical example we are working with, most of the time the following holds: $\frac{-b}{2a} < 0, \forall q \leq r^{-1}(N)$, which can be seen in Figure 3.3 below. In the instances where this value is non-negative, it is still very small and indicates a turning point of the function $r(q)$. Since $r(q)$ approximates the cumulative number of people needed to intervene upon in order to target q events, this turning point can be thought of as the value of q for which the event rate in the population drastically decreases when the data is ordered by decreasing risk score.

3.4 Solution Algorithm

In this section we extend our analysis to provide a best intervention policy based on risk scores given two interventions, I_1 and I_2 . While the equation has a nice form, the function mapping interventions to covered hospitalizations, $r'(q)$, and the cumulative distribution function on number of hospitalizations, $F(q)$ means a direct solution is not available. Given a population to receive intervention 1, $F(q)$ can be obtained for each value of q , as a distribution or estimated by sampling data. Since historical data is not readily available, we provide recommendations to estimate the distribution using data-driven approaches. The function $r(q)$ can also be empirically determined from a set of data for each possible size of q . With the distribution $F(q)$ and the function $r(q)$, then q^* can be determined using a solver to satisfy Equation 3.7. However, we note that the number of interventions of I_2 also depends on the size and characteristics of the population set to receive I_1 . Thus, it is useful to integrate the decision over x_1 . We propose an algorithm that searches over the space of potential x_1 values, solving for the optimal q^* for each. Figure 3.2 below provides an outline of the final recommendation heuristic for this policy determination. We conclude by explaining the necessary details of each step in the algorithm.

3.4.0.1 Sample the training data.

We aim to provide a generalizable framework that can be adapted to populations of various sizes and evaluate our final policy using an out-of-sample testing set. We note that our model formulation is dependent upon the size of our population, N , particularly, after we select a starting x_1 value. We perform all iterations of the newsvendor on training data (2,274,045 patient-rows), sampled without replacement such that it is the size of the testing set (568,215 patient-rows) that we will use for final evaluation.

3.4.0.2 Increment x_1 .

We noted above that the newsvendor critical fractile (CR) that we found above is useful for determining the optimal number of I_2 interventions we should target towards our population for a given value of x_1 . Therefore, in order to provide an optimal solution given the I_1, I_2 pair we discretize over all possible values of x_1 and increment x_1 by +0.01. That is, for every possible value in the set of

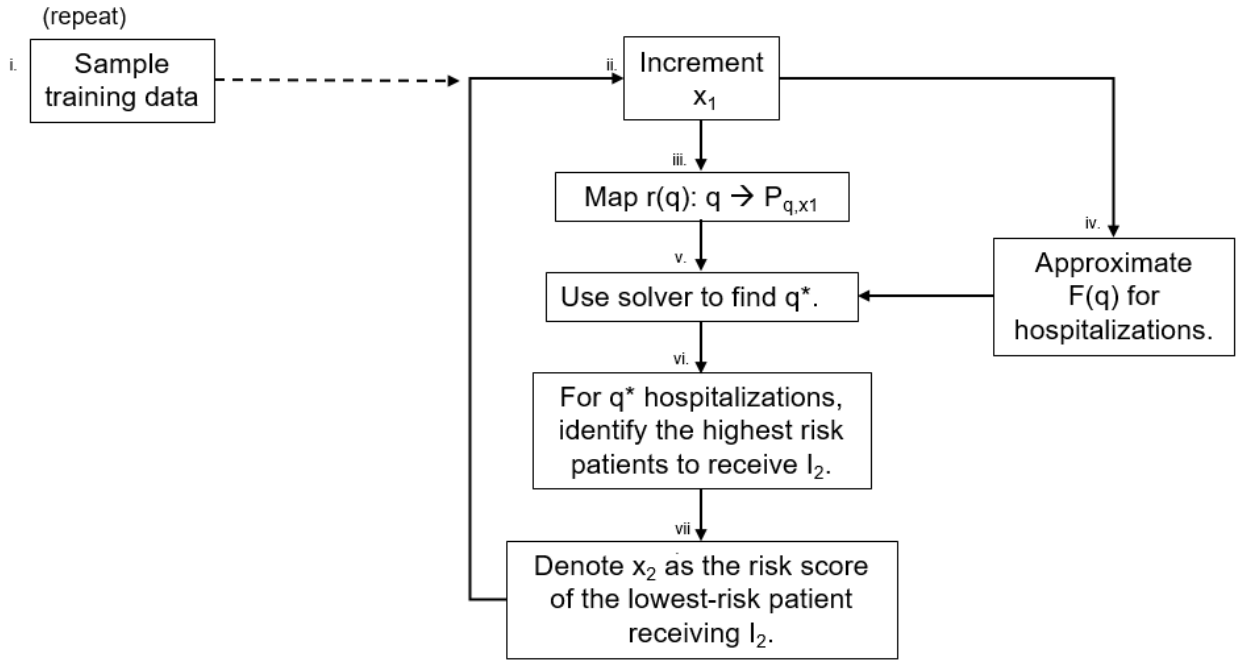


Figure 3.2 Flowchart of psuedo-code for the heuristic to determine the best intervention strategy.

thresholds for which no intervention is give, $x_1 = \{0.00, 0.01, 0.02, \dots, 0.99, 1.00\}$, we evaluate the critical ratio.

3.4.0.3 Approximate $r(q)$.

We consider $r(q)$ a mapping from q , the number of hospitalizations we would like to intervene upon with the more effective intervention, I_2 , to the number of people, P_{q,x_1} we would need to target with I_2 .

For each iteration of the newsvendor applied for each x_1 value we approximate $r(q)$ such that it takes the general form $r(q) = aq^2 + bq + c$. We approximate the best values of a, b, and c using a least squares minimization implemented via the SciPy optimize, curve_fit function [168]. For the numerical results shown below we estimate a, b, and c by setting the independent data to be the cumulative number of pneumonia hospitalizations and the dependent data to be the cumulative number of beneficiaries, with the data ordered from highest to lowest risk score for both data streams. In Appendix C we provide arguments that this assumption holds and that you can always find a convex polynomial of no more than degree two for any set of risk scores even when no information is known about the population and when risk scores are random and uniformly distributed.

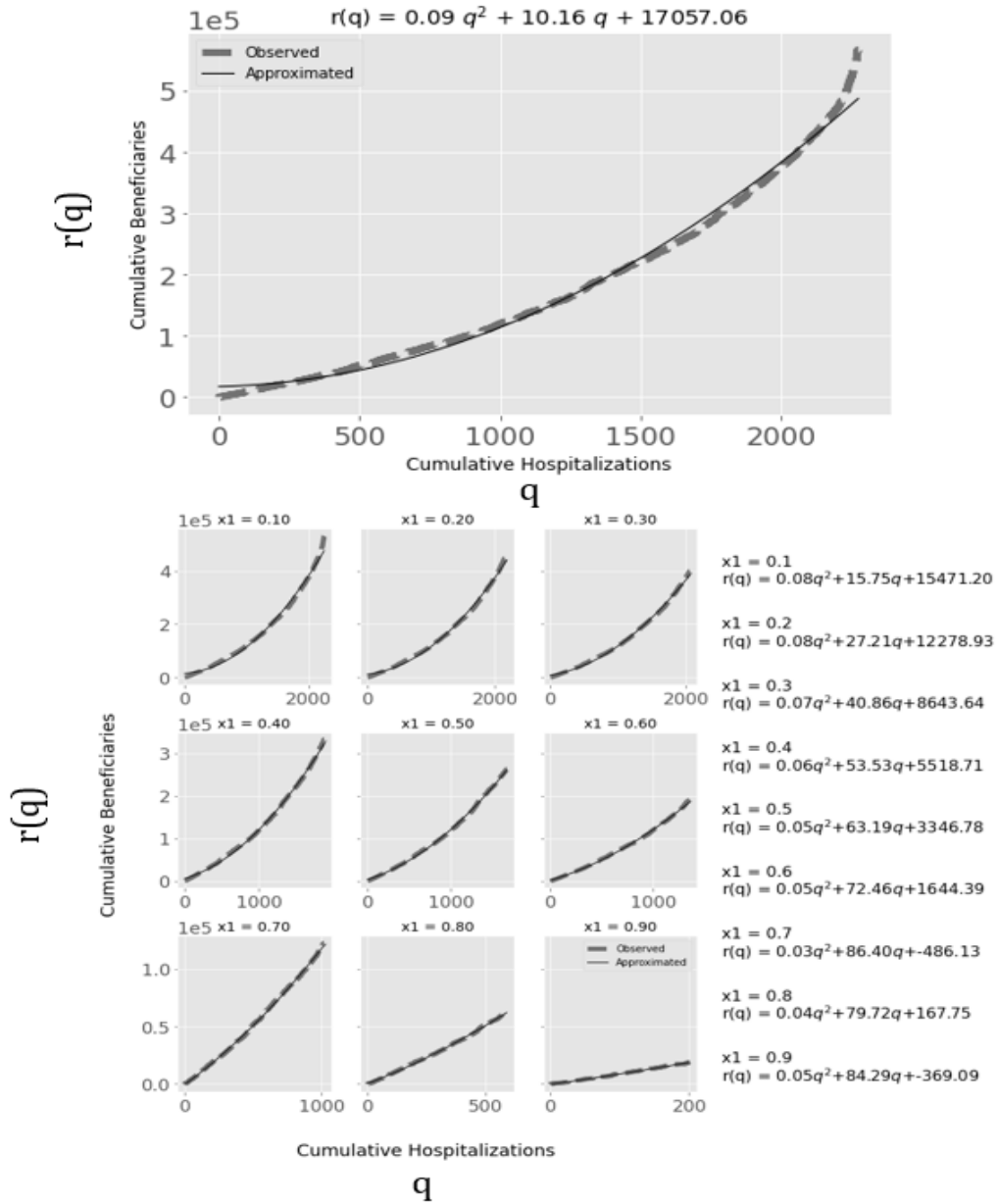


Figure 3.3 Cumulative number of hospitalizations versus cumulative number of beneficiaries in a sampled training set (dashed) and approximated quadratic, convex curve (solid) for a selection of x_1 thresholds. Here, the top graph is for when $x_1 = 0$.

3.4.0.4 Approximate $F_{1+2}^{\wedge}(q)$.

If we know the exact demand distribution for the number of occurrences of the potentially preventable events, then we can directly solve (CR) using a numerical solver. In instances where the demand distribution is unknown, data-driven sampling approaches, such as sampling average approximation over historical or sampled demand distributions have been known to provide near-optimal solutions [101, 154, 178]. For our numerical results we estimate $F_{1+2}^{\wedge}(q)$ using the empirical distribution obtained from collecting 100 random samples of size N_{x_1} , drawn with replacement, from the training data whose risk scores are greater than or equal to x_1 . Here, N_{x_1} , refers to the number of rows in the testing set with risk scores greater than or equal to x_1 . These 100 samples could be thought of as sampling data from 100 hospitals of similar size to estimate the hospitalization distribution for a particular group of patients.

3.4.0.5 Determine q^* .

We use a derivative free, numerical solver to find q^* by minimizing the squared difference of the left and right hand sides of equation (CR), shown below, and limit the search space of q^* to be less than or equal to $r^{-1}(N)$. For every $C_{i1}, p e_1$ pair we are able to achieve a maximum absolute difference between the left and right hand sides of (CR) to be within an epsilon of 0.02235.

$$\min_{q \leq r^{-1}(N)} \left\{ F_{1+2}^{\wedge}(q) - \frac{C_h(p e_2 - p e_1) - (C_{i2} - C_{i1})r'(q^*)}{C_h(p e_2 - p e_1)} \right\}_2 \quad (3.7)$$

3.4.0.6 Identify patients to receive I_2 .

The resulting q^* informs how many high cost interventions an administrator should target for the given x_1 value. For q^* hospitalizations, we identify the highest risk patients to receive I_2 and allow the remaining individuals in the population to receive I_1 . We denote x_2 as the risk score associated with the lowest-risk patient receiving I_2 .

3.5 Numerical Experiments

For the numerical results, we use a 5% sample of Medicare data from 2008-2011 [45] coupled with publicly available data to create a monthly aggregated, patient level dataset to assess an individual's 30-day risk of pneumonia hospitalization. In this section we provide an overview of the data sources, risk scores, and numerical results from applying our solution algorithm.

We define pneumonia hospitalizations using the Agency for Healthcare Research and Quality (AHRQ) Prevention Quality Indicator 11 (PQI-11) definition for hospital admission rates for Bacterial Pneumonia[2]. We consider individual level, community level, and provider level variables to create the feature vectors used to train a logistic regression classifier to create individual risk scores. Our final dataset consists of 2,842,260 patient rows with approximately a 0.39% PQI-11 event rate. The data is split into an 80/20 training and testing set, which we use to numerically solve and evaluate

the results of our model, respectively. More information about the development of the risk scores can be found in the Chapter 2.

The unit cost and effective probability of I_2 are selected as \$120 and 0.65, respectively. The selection of cost and effectiveness could resemble an intervention such as an annual influenza vaccination. Historically, the influenza vaccine's efficacy depends on the circulating strains. Estimates from the 2010 to 2011 influenza season showed that the overall adjusted vaccine efficacy of the influenza vaccine ranged from 53-66% (95% confidence interval) [166]. The Medicare cost per dose is estimated to be around \$20-\$65 [47, 48, 75]. We consider the upper end of this range and more for any possible administration fees [47]. Additionally, we assume that the cost of a pneumonia hospitalization is \$33,380 [171].

We vary the cost and effective probability of I_1 using any combination of a discrete selection of arbitrarily chosen values smaller than I_2 and briefly provide some context of possible interventions that might reasonably fall within these ranges. Specifically, the cost can be \$25, \$50, \$75, or \$100 and the efficacy can be 10%, 20%, 30%, 40%, or 50%. The first type of low-cost intervention can be some form of patient outreach or communication. For instance, automated letters or phone calls are cost-effective measures that have shown to improve immunization rates in under-immunized populations [105]. The second communication could be a more direct, personal phone call from a nurse or physician for which the cost can vary depending on the length of time for the phone call. Furthermore, telehealth visits can improve patient outcomes and recent estimates indicate that a telehealth visit for acute respiratory infections could cost up to \$80 per visit and can reduce overall hospitalization rates and length of stay [5, 133].

We first show the approximations for $r(q)$ and $F_{1+2}^{\wedge}(q)$ for a sample of the discretized x_1 values that we used for this analysis. We show in Figure 3.3 that the approximation of $r(q)$ depends on the choice of x_1 and note that all values of a and b are positive, which implies Propositions 1-3 hold for all $q \geq 0$ for this set of x_1 values. Then we note in Figure C.1 that our sampled distributions to approximate $F_{1+2}^{\wedge}(q)$ capture the true number of hospitalizations that are seen in the out-of-sample testing set truncated above x_1 .

Due to the sampling strategy to obtain estimates on population sizes of our testing set, we repeat the process 30 times for each value of x_1 in order to obtain 95% confidence intervals around the optimal q^* , x_2 values. Figure 3.4 illustrates the average total cost for each x_1 given the recommended x_2 threshold from the heuristic over 30 replications. For each x_1 , we follow the pseudo-code outline in Figure 3.2 and determine the best x_2 . We then compute the total cost associated with targeting individuals with both I_1 and I_2 interventions which includes the costs of hospitalizations that occurred for individuals with risk scores below x_1 . We average these costs across all 30 replications for each x_1 value and show the average across all replications, indicated by the solid lines. We include the 95% confidence intervals around the costs and the best x_1 scores for each $p e_1$, C_{i1} pair across all 30 replications. Figure 3.4d shows a closer look at those confidence intervals. Here, the horizontal red line represents the 95% confidence interval for the cost of doing nothing in 30 sampled training sets during this evaluation.

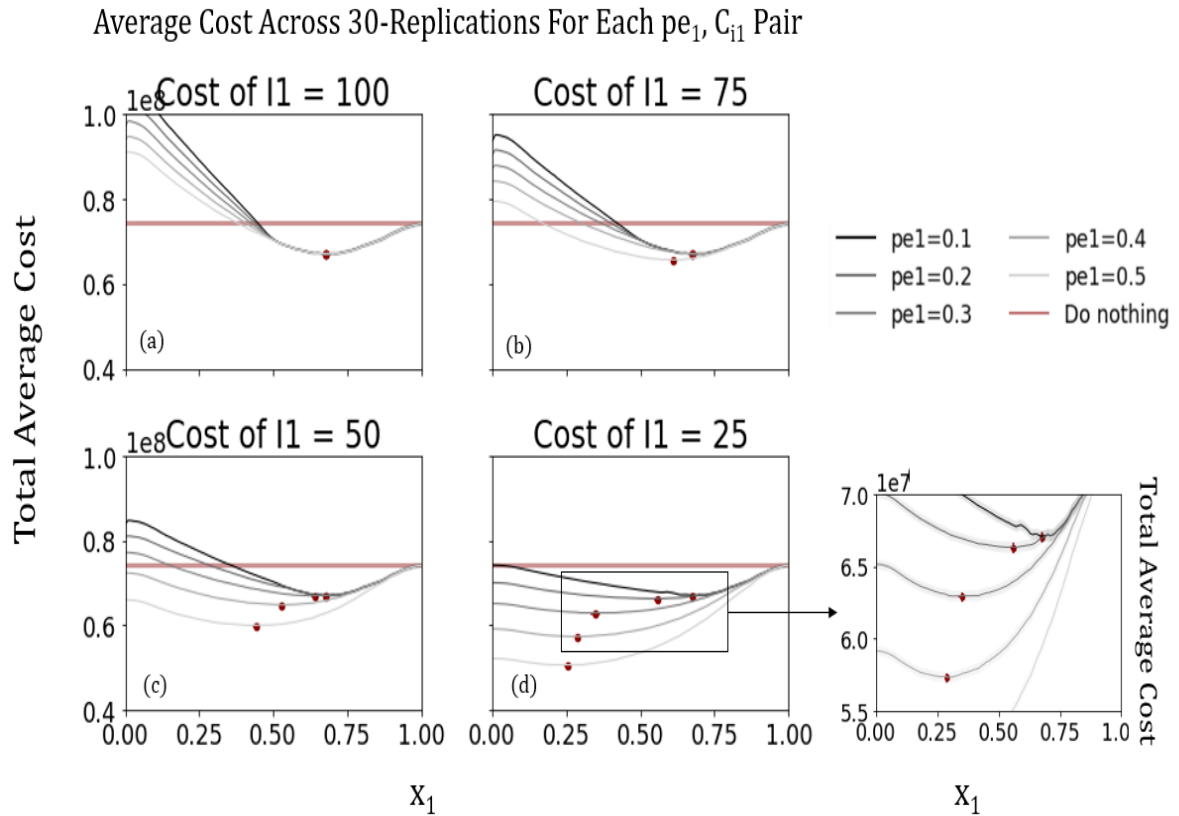


Figure 3.4 [a-d] The average total cost, over 30 replications, on the training data for the recommended strategy for each x_1 value, compared to the 95% confidence interval for the cost of doing nothing (horizontal, solid red line) over 30 replications, ordered by increasing x_1 threshold for varying costs of I_1 . The red dots represent the optimal, recommended strategy over each C_{i1}, pe_1 pair for I_1 and include a 95% confidence interval.

We first observe the impact of the recommended policy on the training and out-of-sample testing set in Figures 3.4 and Table 3.1, respectively. Figure 3.4 shows the average total cost, including hospitalizations that were not covered or occurred due to an intervention not working, compared to the cost of doing nothing, horizontal, red line. We note that all recommended solutions (red points) show potential cost savings compared to doing nothing in our sampled training data across 30 replications. In the out-of-sample testing set, we note that the percent savings compared to doing nothing increases approximately 9.24-9.27%, 9.25-11.27%, 9.26-19.03%, 9.27-31.83%, as the effectiveness of intervention one increases from 0.1 to 0.5, across all tested values for the cost of intervention one indicating that the policies obtained from the training data are effective in the testing data as well.

Table 3.1 illustrates the impact of cost and effectiveness of the low cost intervention on our out-of-sample testing set evaluations. Here we highlight the mean number of low cost interventions, I_1 , and mean number of high cost interventions, I_2 , along with 95% confidence intervals over 30 replications. We note that when the effectiveness of intervention one is lowest, e.g., $p e_1 = 0.1$, all recommended policies suggest that policy makers should consider giving more patients the high cost intervention regardless of the cost of I_1 . Similarly, we highlight that when the cost of intervention one is highest, e.g., $C_{i1} = \$100$, the recommended policy will be the same. We observe that as the effectiveness of intervention one increases we should consider giving more people intervention one depending on the cost.

Next, we can observe the effects of decreasing the cost of intervention one. In Figure 3.4a, we observe that when the cost of $I_1 = \$100$, the recommended policies converge for all of the chosen values for the effectiveness of intervention one. As the cost of intervention one decreases in Figure 2b-d we observe that the recommended strategies depend more on the effectiveness of intervention one. It is clear that in all of these scenarios all decisions lead to an overall cost that results in a net savings compared to doing nothing, i.e., red points are below the horizontal, red line.

Second, we are able to observe the impact of the effectiveness of intervention one on the recommended strategy. We notice that when the cost of intervention one is less than \$100 the overall cost reduces depending on the effectiveness of intervention one. We also note that as the effectiveness of intervention one increases, the recommended strategy increases there are fewer number of people receiving either intervention. We can view the exact policies in Table 3.3 which shows the recommended $\{x_1, x_2\}$ policy for each $C_{i1}, p e_1$ pair along with 95% confidence intervals across 30 replications.

We can see from Table 3.1 that the best policy when the cost of intervention one is high, e.g., $C_{i1} = \$100$, the best policies suggest that we should attempt to cover the hospitalizations with high cost interventions, regardless of the effectiveness of intervention one. As the cost of intervention one decreases and the effectiveness of intervention one increases we observe that the recommendation is to give more people some form of intervention, e.g., lower x_1 threshold in Table 3.3. Across all values for effectiveness of intervention one we note that as the cost of intervention one decreases, the value of x_2 increases, implying that fewer patients should receive intervention two.

Table 3.1 Testing set evaluation for the mean number of high and low cost interventions, I_1 and I_2 , when using the best x_1, x_2 strategy, along with 95% confidence intervals over 30 replications.

		C_{i1}		pe_1		
		0.1	0.2	0.3	0.4	0.5
I_1	100	241	241	241	250	250
		± 144	± 144	± 144	± 149	± 149
	75	241	250	250	250	140,762
		± 144	± 149	± 149	± 149	$\pm 6,482$
	50	250	250	38,159	185,950	283,959
		± 149	± 149	$\pm 9,333$	$\pm 5,201$	$\pm 7,356$
25	250	117,263	300,408	367,273	409,885	
	± 149	$\pm 9,588$	$\pm 10,222$	$\pm 5,183$	$\pm 7,851$	
I_2	100	131,165	131,165	131,165	131,372	131,372
		$\pm 5,792$	$\pm 5,530$	$\pm 4,617$	$\pm 4,112$	$\pm 4,085$
	75	131,165	131,372	131,372	131,372	33,571
		$\pm 5,596$	$\pm 4,426$	$\pm 4,135$	$\pm 4,341$	$\pm 1,866$
	50	131,372	131,372	117,369	45,366	8,139
		$\pm 5,542$	$\pm 5,542$	$\pm 4,223$	$\pm 1,392$	$\pm 1,740$
	25	131,372	93,003	54,174	28,327	8,520
		$\pm 5,542$	$\pm 1,923$	$\pm 1,094$	$\pm 1,105$	$\pm 1,327$

3.6 Conclusion

In summary, we propose a model formulation that allows us to provide interpretable solutions and recommendations to targeting individuals based on risk scores. We consider that in reality pneumonia hospitalizations can vary depending on factors such as overall population health, demographics, location, and presently circulating diseases and provide the best recommendations that consider this uncertainty and minimize overall expected costs [13, 82, 83, 145, 157]. This framework can aid policy makers in targeting at-risk populations with interventions and reduce hospitalizations and attributable costs. In practice, one could take the recommended x_1 and x_2 values for each pair of interventions and target individuals within each of the three groups created by this partition. We note that the recommended policies will depend on the types of interventions available for a population. For pneumonia hospitalizations these interventions could include vaccinations, preventative drugs, smoking cessation sessions, or communication regarding healthy lifestyle practices [34]. Our policy can inform how many high and low cost interventions an administration should consider and whom to target with them based on individual risk scores.

Although we provide recommendations specifically for interventions to target pneumonia hospitalizations, we note that our model does not require any information about pneumonia or associated hospitalizations, other than attributable costs. Therefore, our model can be adapted and transformed for any other discipline where risk scores are available. That is, our model generalizes well to other notions of risk for any population. Aside from pneumonia hospitalizations, this strategy can be used to inform intervention strategies for other types of preventable hospitalizations or

Table 3.2 Testing set evaluation for the mean percent savings and 95% CIs compared to doing nothing, all low cost interventions, and all high cost interventions when using the best x_1, x_2 strategy, along with 95% confidence intervals over 30 replications.

Doing Nothing	100	9.24 ±0.12	9.25 ±0.12	9.26 ±0.12	9.26 ±0.12	9.27 ±0.11
	75	9.25 ±0.04	9.26 ±0.04	9.26 ±0.04	9.27 ±0.04	11.27 ±0.08
		50	9.26 ±0.12	9.26 ±0.12	9.39 ±0.09	12.46 ±0.02
	25		9.27 ±0.12	9.95 ±0.04	15.16 ±0.06	22.69 ±0.04
All low (I_1)	100	45.13 ±0.07	41.60 ±0.08	37.59 ±0.08	32.98 ±0.09	27.64 ±0.09
	75	38.07 ±0.08	33.54 ±0.09	28.29 ±0.09	22.15 ±0.10	16.72 ±0.07
		50	28.94 ±0.095	22.90 ±0.01	15.87 ±0.08	10.40 ±0.02
	25		16.64 ±0.11	8.90 ±0.04	4.51 ±0.07	1.95 ±0.06
All high (I_2)	100	27.66 ±0.10	27.67 ±0.09	27.68 ±0.09	27.68 ±0.09	27.68 ±0.09
	75	27.67 ±0.09	27.67 ±0.09	27.68 ±0.09	27.68 ±0.09	29.28 ±0.06
		50	27.68 ±0.09	27.68 ±0.09	27.78 ±0.07	30.23 ±0.02
	25		27.68 ±0.09	28.22 ±0.03	32.38 ±0.05	38.38 ±0.04

Table 3.3 Recommended policies for each C_{i1} , $p e_1$ pair. Here we show the average value and 95% confidence interval of best x_1 , x_2 values for each C_{i1} , $p e_1$ pair, over 30 replications.

C_{i1}	$p e_1$	X_1	X_2
100	0.1	0.6767 \pm 0.0086	0.6770 \pm 0.0085
	0.2	0.6767 \pm 0.0086	0.6770 \pm 0.0085
	0.3	0.6767 \pm 0.0086	0.6770 \pm 0.0085
	0.4	0.6763 \pm 0.0088	0.6767 \pm 0.0087
	0.5	0.6763 \pm 0.0088	0.6767 \pm 0.0087
75	0.1	0.6767 \pm 0.0086	0.6770 \pm 0.0085
	0.2	0.6763 \pm 0.0088	0.6767 \pm 0.0087
	0.3	0.6763 \pm 0.0088	0.6767 \pm 0.0087
	0.4	0.6763 \pm 0.0088	0.6767 \pm 0.0087
	0.5	0.6113 \pm 0.0091	0.8565 \pm 0.0036
50	0.1	0.6763 \pm 0.0088	0.6767 \pm 0.0087
	0.2	0.6763 \pm 0.0088	0.6767 \pm 0.0088
	0.3	0.6397 \pm 0.0110	0.6990 \pm 0.0069
	0.4	0.5287 \pm 0.0080	0.8294 \pm 0.0031
	0.5	0.4413 \pm 0.0122	0.9364 \pm 0.0081
25	0.1	0.6763 \pm 0.0088	0.6767 \pm 0.0087
	0.2	0.5590 \pm 0.0126	0.7395 \pm 0.0033
	0.3	0.3493 \pm 0.0158	0.8108 \pm 0.0022
	0.4	0.2883 \pm 0.0085	0.8697 \pm 0.0029
	0.5	0.2550 \pm 0.0127	0.9330 \pm 0.0057

other adverse events related to diseases such as influenza or COVID-19, especially where risk scores are widely available [89, 108]. Outside of healthcare, this strategy can be useful for policy makers looking to implement interventions to reduce costs incurred from any potentially avoidable event. This could be in areas where policymakers are considering interventions to prevent crime [172] or improving graduation or retention rates in educational settings [139], just to name a few.

It is important to note that this formulation is imperfect and has some limitations. The first limitation is that this model only considers the decision at a single point in time. However, the newsvendor single-period approach can also be used as a building block on multi-period decision models [3]. Second, decision makers should always optimize over x_1 and x_2 simultaneously. As shown in Figure 3.4a-d there are some values of x_1 that yield optimal solutions that are worse than doing nothing, meaning that the recommendation from a cost perspective should be to give no interventions. Nevertheless, our functional form allows us to evaluate more scenarios and determine the optimal policy for a specific value of cost and effectiveness of I_2 . Third, the results depend on the discretization chosen for x_1 . We chose to increment by 0.01 to gain a sense of the entire search space. Fourth, we are limited in this problem by the lack of historical data to estimate the underlying demand distribution. We circumvent this by applying a random sampling technique and repeat our sampling for 30 replications to provide confidence intervals around our recommended policies and results. Lastly, the numerical solver sometimes faces numerical complications due to the

approximation of the empirical distribution function to find q^* . We found that q^* yielded an error within an epsilon of 0.02235 to ensure that we found the best possible solution given the limitations of using a numerical solver.

We note that this work considers the best strategy between two interventions under unconstrained supply, but in the future can be extended to consider these aspects. Some approaches, such as dynamic programming, may be useful to determine the best strategy for more than two interventions, but may not necessarily provide interpretable results for policymakers. If interpretability remains important, one could consider grouping interventions into two groups to ascertain the best policy given the expected costs and effectiveness of interventions of similar types. In order to account for constrained intervention supply, one could consider adding resource constraints to the minimization problem formulated above.

In all, we provide a generalizable framework to solving a multiclass newsvendor based model to determine an optimal intervention strategy in a population with available risk scores. We show that our policy can reduce overall pneumonia hospitalization attributable costs up to 32%, depending on the available interventions. If given a complete list of possible interventions, one could use this method to create a lookup table of recommended intervention strategies. This would be easy to compute as our model adapts well to problems of various sizes. Other optimization models, such as integer programming approaches, do not scale well and are known to be computationally complex and time intensive [85]. For our model, it takes less than 60 seconds per replication to evaluate the best policy for each C_{i1}, p_{e1} pair over our entire discretization of x_1 . That is, this model can quickly adjust the costs and effective probabilities to recommend a new policy for alternative sets of interventions and could become a useful tool for any area where risk scores are available.

CHAPTER 4

METHODS TO ESTIMATE TRUE DISEASE BURDEN OF COVID-19

4.1 Introduction

The true number of infections, hospitalizations, and deaths, henceforth referred to as disease burden, attributable to the COVID-19 pandemic has remained unknown since the World Health Organization (WHO) first declared the emergence of COVID-19 as a global pandemic. [141]. Over time, and especially during the earlier months of the pandemic, there were substantial barriers such as insufficient testing capacity and low implementation of and adherence to the use of mitigation measures due to the unknown impacts of masks and social distancing at the time. These barriers contributed to the unmitigated spread of the highly transmissible COVID-19 virus across the United States. Especially during the first few months, and even throughout the first year prior to vaccine deployment, the majority of the COVID-19 infections went undetected [141, 174]. During this beginning period, the guidelines recommended by the Centers of Disease Control and Prevention (CDC) suggested that people stay home unless severely ill and in need of medical intervention from a health-care professional and did not include provisions for acquiring tests [28]. With most individuals staying home to quarantine and before testing capacity was expanded to the public, the true number of infections still remains unknown.

Understanding the true disease burden associated with COVID-19 is imperative to understand important measures surrounding the current state of the pandemic such as herd immunity, the emergence of variants, breakthrough infections and reinfections, and fatality rates over time [42, 141]. If we can estimate who already had the disease and determine whether current infections and severe disease are likely to be new infections or breakthrough or reinfections, we can better understand who remains at risk and the likelihood of experiencing another wave of incidence,

hospitalizations, or deaths. This is particularly important as there remains a gap in knowledge surrounding the lasting immune response to someone who is fully recovered from COVID-19 [104]. This also helps us gain insight towards reinfections as we consider the impacts of new and more transmissible and severe variants which may have immune-escaping [27, 92, 104] properties that affect even those who may have already gained some immunity from a prior infection.

Internationally, nationally, and sub-nationally, it is known that there exist significant differences in total burden where it is documented certain countries and subgroup populations have suffered more hardships than others, especially in the early months when entire populations were susceptible to the virus. Levin et al. [102] estimated that the age-specific infection fatality rates (IFRs) were roughly two times higher in developing countries compared to high income countries. Their study also indicated that the seroprevalence in developing countries was roughly the same in older age groups as compared to younger age groups, pointing to these countries' difficulties regarding providing protection to their most vulnerable populations. Within the United States multiple studies and analyses indicated that certain sub-populations experienced disproportional amounts of burden compared to the general population. These populations include certain racial and ethnic groups [146], working classes such as healthcare workers, first responders, and other essential workers [173], and residents of specific geographic locations where certain states, dense cities, and rural areas with limited access to health resources suffered more than in other areas of the country [10, 42, 131].

Identifying the remaining susceptible population at a point in time can improve health related outcomes by targeting mitigation resources and interventions to prevent severe disease. For instance, understanding susceptibility could have helped inform vaccine prioritization schemes when vaccines were first available in early 2021 in limited supply [146]. Additionally, understanding disease burden across subgroups can further efforts towards bridging the equity gap by identifying populations that were hit the hardest in terms of all incidence, severity, and mortality [146]. Understanding disease patterns can also help public health experts analyze the impact of mitigation measures, such as the use of certain non-pharmaceutical interventions, i.e., wearing masks, social distancing, closing schools, working from home, etc., on overall disease incidence and mortality rates for a period of time. Therefore, the development of methods to estimate population susceptibility at a point in time can prove to be a beneficial tool for public health experts, decision makers, and other stake holders when it comes to COVID-19 and potentially other infectious and fatal diseases.

Identifying populations that have already experienced a prior infection can further inform infection hospitalization rates and mortality risk. Aside from these metrics, the public health community can benefit from population estimates of long COVID, which has been linked to deaths even months after initial COVID-19 infections. Public health experts are working to learn more about risk factors that can contribute to the likelihood of severe disease or mortality associated with long COVID [8], but the reality is that in meantime these post-infection factors are affecting many people all over the country. A recent report from the CDC that reports 1 in 5 adults age 18 years or older [18], are estimated to experience the effects of long COVID when considering both the prevalence of long COVID among the infected and the underestimation in overall infections compared to doc-

umented cases [115]. Even if the effects of long COVID are not publicly advertised, many experts believe that we are just beginning to see the impacts of this long term illness. For instance, Nicole Maestas from Harvard Medical School, suggests that the country has yet to see the impact of long COVID in disability claims due to the paperwork and bureaucratic process that individuals have to navigate in order to file claims when they cannot prove that they ever had the virus. There currently exists no diagnostic test for long COVID and individuals are required to build up their own medical documentation in order to prove that they are impaired for a length of time before they qualify for disability assistance [115]. Other anecdotal accounts suggest that individuals are also struggling to receive assistance and necessary prescriptions from their insurance companies without a diagnosis or laboratory-confirmed documentation of the disease [163]. Pin-pointing geographic areas or subgroups that were likely to have been hit with a surge of infections at any point throughout the course of the pandemic can inform disability assistance, insurance claims, and resource allocation to treat the symptoms experienced by many with long COVID.

To summarize, understanding the past disease burden can help to inform the analysis of mitigation efforts, infer changes in behavioural patterns and relations to disease spread, evaluate health equity, and provide additional tools to help identify those who are at risk of experiencing the negative effects of long COVID.

4.2 Literature Review

Understanding the true disease burden has been of interest to many public health experts throughout the course of the pandemic, particularly those in forecasting and public health policy. The CDC has estimated true disease burden for the United States in efforts to capture how much of the population has already been infected with the disease [25]. The CDC estimated that in 2020 approximately 1 in 7, infections were reported through September 2020, and 1 in 4 infections were reported through December 2020, indicating that these rates of underreporting improved over time[25, 141]. The CDC analysis considers values by age group, but does not consider the differences in disease burden or unreported cases at the sub-national level.

Pei, et al. used a meta-population, Bayesian inference model to estimate disease progression across counties within the country, with a specific focus on metropolitan areas. The study estimated that the total population susceptibility in the United States by the end of 2020 was around 69% with considerable variability between major metropolitan areas such as 48% in Los Angeles compared to 73% in Phoenix[131]. The model is validated on reported cases and out-of-sample validated on seroprevalence estimates from conducted surveys. The surveys available in the earlier months in year only include estimates from 10 sites. Thus, the use of seroprevalence estimates during this time period and limited data is likely to yield bias in the model inference parameters. Additionally, the model does not report or evaluate hospitalization need or risk of severe disease.

Similarly, a comprehensive, statistical analysis performed by Barber, et al. relied on seroprevalence survey data, which could potentially be impacted by the same biases [10]. This analysis

was performed at the national and regional level and also provides estimates for the infection fatality ratio and cumulative percent infected over time by age [10]. The authors note that this study ends upon the introduction of the Omicron variant due to the likely impact on the estimated values computed from the seroprevalence surveys and that new results await the addition of updated survey results. In that sense, even though the results can typically be run to provide daily estimates, the methodology will not adapt well to new variants, which is a cause of major concern when considering all of the new and more severe variants of COVID-19 [84].

Chitwood, et al. developed a Bayesian model to estimate cumulative incidence across states and counties over time in a "backcasting" approach. The model comprises a deterministic, compartmental model that accounts for diagnosed and undiagnosed cases of COVID. However, the author's report that fitting the Bayesian model requires fixed distributions to reduce overall computational complexity [42]. Next, the model assumes a population weighted IFR value to fit to empirical death data whereby the IFR values are highly uncertain and can lead to increased sensitivity to model performance.

While these models provide useful insights and calibrate well to confirmed cases and deaths, the underlying assumptions make the models highly sensitive to uncertainty surrounding parameters such as the IFR values and biases in seroprevalence surveys. This chapter will provide a thorough analysis of IFR applications to estimate overall disease incidence and introduce a novel method to fit empirical data and estimate disease characterizing parameters. In the end, estimates from these approaches will be compared to estimates provided by the aforementioned models, where applicable.

4.3 Methods

This chapter describes two approaches to evaluate disease burden and estimate the total number of true infections of COVID-19 experienced in 2020, prior to vaccine distribution: (i) the deterministic application of infection fatality rates; and (ii) the use of deterministic, compartmental epidemiological modeling. This section will describe all of the necessary assumptions, data sources, and procedures used for each method.

Deterministic Application of Infection Fatality Rates (IFR)

In the early stages of the pandemic many studies sought to inform population infection fatality rates [29, 81, 102]. The infection fatality rates differ by age and geographic location as seen in Figure D.1 due to the different prevalence of disease-related risk factors that are present in these populations. Due to the uncertainty surrounding these IFRs, and their unknown true values, we compute two versions of the population-weighted infection fatality rate for the United States. In the first method, shown in Appendix Table D.1, we use population estimates gathered from the 2019 Census data [66] to perform a population weighted IFR value for the United States for each of the

sets of infection fatality ratios that have been peer reviewed. Upon closer inspection of the resulting US Age population-weighted IFR values presented in D.1, and comparing to what was reporting in the literature for the United States, the rest of this analysis assumes the IFRs reported in Levin et. al [104]. In the second US-population IFR calculation, we perform a population weighted IFR value dependent on the composition of true estimated infections in the United States where the assumption is that the infection fatality rate is dependent of the infection fatality rate of those who are reported infected, the results of which are provided in Table 4.1. The method for computing the true estimated infections is provided below.

Table 4.1 Estimated percent of total infections by deterministic IFR analysis and corresponding resulting NC-IFR from composition of estimated infections.

Age	% true cases (end of 2020)	Levin IFRs
0-9	0.1452	0.00001
10-19	0.09571	0.00003
20-29	0.1862	0.00011
30-39	0.162	0.00037
40-49	0.1527	0.00123
50-59	0.1192	0.00413
60-69	0.0804	0.0138
70-79	0.0401	0.0462
80+	0.01851	0.1546
		0.0066

Gathering Estimates of Undocumented Infections

Using the COVID-19 Restricted Use Case Surveillance Data [21] gathered from the CDC we collect all documented lab confirmed cases and deaths by age group and geographic location, i.e., urban, suburban, or rural residencies. We estimate the true number of infections by applying the IFR to the number of reported deaths for each age group, through division. We define the time-varying lab multiplier to the ratio of true infections, estimated from the IFR analysis, to the number of lab-confirmed infections. To eliminate noise in the data we use 13-week average values of documented cases and deaths for each stratified population for this analysis. Since the younger age groups had little to no deaths reported throughout 2020, age groups 0 - 29 Years are grouped together for nontrivial estimates. If there are no reported deaths over a 13-week period, the lab multiplier is

assumed to be one, and the number of estimated true infections is equal to the number of lab-confirmed cases. For the age group analysis, we use the following formula.

$$IFR_i = \frac{\mu_i}{\pi_i} \text{ where,}$$

$$\mu_i := \text{reported deaths for age group } i,$$

$$\pi_i := \text{true infections for age group } i,$$

$$\forall i \in 1, \dots, 7$$

and define our time-dependent lab multiplier to be:

$$LM_{i,t} := \frac{\pi_i}{\bar{\gamma}_{i,t}}, \text{ where}$$

$$\pi_i := \frac{\overline{\mu_{i,t}}}{IFR_i}$$

$$\bar{\gamma}_{i,t} := \text{13 week average of lab-confirmed cases for age group } i$$

$$\overline{\mu_{i,t}} := \text{13 week average of COVID-19 deaths for age group } i$$

Using the deterministic approach to compute the lab multipliers for different age groups over time, we can view the true estimate of prevalence of COVID-19 by age group for different states, and at the sub-geographic populations, e.g., urban, suburban, and rural locations. We present the results of this initial analysis in Figures D.2 - 4.6 as the time-varying lab multiplier for each age group. The lab multipliers are computed by age group, for each geographic location in the state as presented in Figure D.3. The resulting lab multipliers are used to provide a final case estimate for each week in 2020. The results by age and rurality are aggregated weekly via summation and used to inform the state level estimated number of infections. The results in Figure D.2 show the estimated state-wide lab multipliers compared to the CDC Case Surveillance Data [21].

Epidemiological Model

The second approach to estimate disease burden is to fit a deterministic, compartmental Susceptible-Exposed-Infected-Recovered (SEIR) model to empirical data. The model is fit using a backcasting analysis to estimate documented and undocumented infections, hospitalization need, and deaths over time. This section will first describe the model description including all governing equations and assumptions, then explain the different fitting periods used in this analysis, and finally, provide the fitting algorithm to produce the final results.

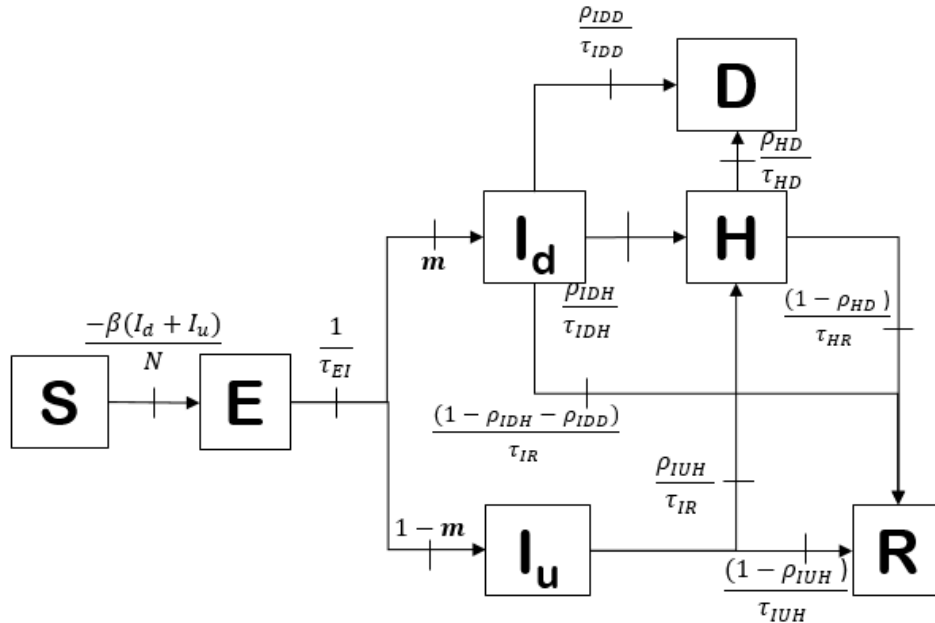


Figure 4.1 Compartmental model showing the seven disease states considered in this analysis.

Model Description

The SEIR model comprises seven compartments, among which there exists a compartment for both documented and undocumented infections to track what these values might be over time. To fit this model, we focus on two parameter sets: (i) disease progression parameters and; (ii) fitting parameters, e.g., the percent of documented infections, and the transmissibility of the virus. The model comprises seven disease states: (S)usceptible, (E)xposed, (I_d)Documented infected, (I_u)Undocumented infected, (H)ospitalized, (R)ecovered, and (D)ead, and the flowchart of disease progression is illustrated in Figure 4.1. The model assumes complete mixing within the population and the parameter labels and descriptions are shown in Table 4.2.

Note. Undiagnosed infections cannot die without hospitalization because we assumed that these individuals had less severe instances of the disease, as they were not ordered a test by a doctor or physician, unless they entered the hospital due to the development of severe disease. We validate the hospitalized compartment using "confirmed and suspected inpatient admissions" reported by HHS. For this reason, lab-confirmed and non-lab-confirmed cases are both able to reach the hospitalized compartment.

$$\frac{dS}{dt} = -S\beta * \frac{I_d + I_u}{N} \quad (4.1)$$

$$\frac{dE}{dt} = S\beta * \frac{I_d + I_u}{N} - E \frac{1}{\tau_{EI}} \quad (4.2)$$

$$\frac{dI_d}{dt} = (m)E \frac{1}{\tau_{EI}} - I_d \left[\frac{(1 - \rho_{IDH} - \rho_{IDD})}{\tau_{IR}} + \frac{\rho_{IDH}}{\tau_{IDH}} + \frac{\rho_{IDD}}{\tau_{IDD}} \right] \quad (4.3)$$

$$\frac{dI_u}{dt} = (1 - m)E \frac{1}{\tau_{EI}} - I_u \left[\frac{1 - \rho_{IUH}}{\tau_{IR}} + \frac{\rho_{IUH}}{\tau_{IUH}} \right] \quad (4.4)$$

$$\frac{dH}{dt} = \frac{\rho_{IDH}}{\tau_{IDH}} I_d + \frac{\rho_{IUH}}{\tau_{IUH}} I_u - H \left[\frac{\rho_{HD}}{\tau_{HD}} - \frac{1 - \rho_{HD}}{\tau_{HR}} \right] \quad (4.5)$$

$$\frac{dR}{dt} = \frac{(1 - \rho_{IDH} - \rho_{IDD})}{\tau_{IR}} I_d + \frac{1 - \rho_{IUH}}{\tau_{IR}} I_u + \frac{1 - \rho_{HD}}{\tau_{HR}} H \quad (4.6)$$

$$\frac{dD}{dt} = \frac{\rho_{IDD}}{\tau_{IDD}} I_d + \frac{\rho_{HD}}{\tau_{HD}} H \quad (4.7)$$

$$(4.8)$$

Model Fitting

Fitting Periods

Retrospectively, throughout 2020 there were many treatments recommended and tested to try to reduce the risk of severe outcomes from COVID-19. In this backcasting analysis, a review was conducted to inform which treatments were considered effective, either through case studies or via the status of any Federal Emergency Use Authorizations (EUA). In order to account for the available treatments that impacted disease progression, the empirical data is split into five fitting periods as illustrated in Figure 4.2.

The first fitting period is prior to April, 2020, which is the most unknown time period throughout the year. Since COVID-19 was first declared a pandemic in March of 2020, the first fitting period will be from the time the first case is documented to the first week in April. In April, CNN Health published an article on the benefits of prone positioning [49], and in May 2020 an editorial was published suggesting that trials were underway to confirm the preliminary conjectures that prone positioning could potentially prevent or delay intubation in some patients [160]. A study published in late August, 2021 confirmed that the early application of prone positioning showed improvements in patient outcomes [7], and another in early 2022 indicated the potential to reduce mortality rates by 10% in patients with COVID-19 related acute respiratory distress syndrome (ARDS) [63].

The remaining fitting periods were determined using a joint assessment of model calibration and treatments archived by the New York Times [180]. The first treatment, Remdesivir, first received its EUA for critically ill COVID patients in May 2020 [95]. The drug has been documented to reduce recovery time in hospitalized patients with lower respiratory tract infections by about four days

Table 4.2 Parameter descriptions for the inputs to the state level SEIR model, split into types, e.g., time (days), rates, and fitting parameters. The bounds correspond to the upper and lower bounds used as inputs to create the initial Latin Hypercube Sample. All parameter bounds were determined via literature estimates and model validation.

Type	Parameter	Description	Bounds	Source
Time (days)	τ_{EI}	Latent period	[4, 8]	[31]
	τ_{IR}	Infectious duration without hospitalization	[4, 10]	[19, 22]
	τ_{IH}	Infection time to hospitalization	[2, 12]	[31, 67]
	τ_{HD}	Time from hospitalization to death	[5, 13]	[31, 67]
	τ_{IDD}	Time to death from a documented infection without hospitalization	[4, 10]	[31, 67]
	τ_{HR}	Hospitalization duration	[2, 14]	[31, 67]
Rates	ρ_{IDH}	Documented infection hospitalization rate	[0.05, 0.2]	[21, 31]
	ρ_{IUH}	Undocumented infection hospitalization rate	[0.05, 0.2]	Assumed
	ρ_{HD}	Hospitalization fatality rate	[0.15, 0.3]	[31, 33, 67]
	ρ_{IDD}	Documented infection fatality rate without hospitalization	[0.0, 0.10]	Assumed
Fitting	β	Transmissibility	[0,1]	Fitted
	m	Percent of infections that are documented	[0,1]	Fitted

Table 4.3 The allowable percent change (reduction) in disease progression parameters that were assumed to be reduced by available treatments in 2020 assigned by the respective fitting periods the treatments became available.

Parameter	Description	Fitting Period	Allowable Change (%)
ρ_{IUH}	Undocumented infection hospitalization rate	May-June	30
		Nov-Dec	30
ρ_{IDH}	Documented infection hospitalization rate	May-June	30
		Nov-Dec	30
ρ_{HD}	Hospitalization fatality rate	May-June	
		June-Nov	30

and reduce mortality rates by about 3.8 to 5.2 percentage points depending on the time from treatment (or placebo) [11], and in later trials in 2021, was found to reduce risk of hospitalization

Timeline of Effective Treatments in 2020

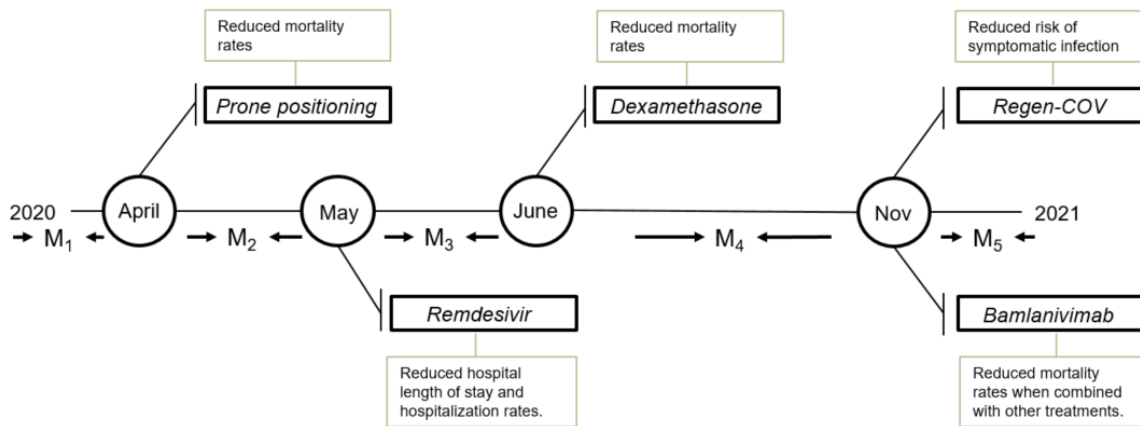


Figure 4.2 Timeline of nationally available, effective treatments in the United States and the corresponding fitting periods M_i that result for this analysis.

and death by 87% among non-hospitalized patients at risk for disease progression [78]. In early June, Dexamethasone, a commonly used drug prior to 2020 was found to reduce deaths by up to one third in a clinical trial involving hospitalized COVID patients [121, 140]. Later, Regen-COV and Bamlanivimab received EUAs in November, 2020 [70, 161]. Regen-COV was reported to reduce the risk of symptomatic infection by up to 81% in a phase three clinical trial [142] and Bamlanivimab, a monoclonal antibody treatment was issued an EUA in November 2020 for distribution mild-to-moderate COVID-19 infection in adults and certain pediatric patients [69], was found to reduce incidence of COVID-19 by approximately 6.7 percentage points among residents and staff in skilled nursing and assisted living facilities [50].

Disease Progression Parameters

For the state level model, there are a total of ten disease progression parameters, six relating to the time, in days, spent in each disease state and four pertaining to transition rates. These parameters are sampled using the bounds listed in Table 4.2 and a Latin Hypercube sampling technique. The bounds for the time parameters are estimated from the values reported by the CDC Pandemic Planning Scenarios [31] where the rate parameters are estimated from CDC Case Surveillance Data [21], NY Times [162], and national reports on hospitalization and fatality rates provided by the CDC especially during early 2020 [33].

The disease progression parameters are sampled 1,000 times using the general bounds imposed in Table 4.2. After generating the 1,000 samples of input parameters, the following filters are applied to avoid over-fitting due to unlikely scenarios. That is, input samples are collected provided they

meet the following assumptions.

Assumption 1- $\tau_{HD} \geq \tau_{IDD}$

The time to death without a hospitalization must be no greater than the time to death with a hospitalization. This assumption implies that those who need hospitalization are severely ill and are likely to have increased mortality risk throughout the entire infection.

Pathak et. al, report out-of-hospital mortality during the time period from January 2020 to March 2021 in terms of the percent of deaths occurring in hospitals, long term care facilities. The authors find that deaths occurring outside of hospitals varied by age group, for the United States with the highest percentage occurring in those age 0 to 17 years [129]. The study highlighted that for those 85+ years and older, 12.3% of deaths occurred outside of the hospital and emergency department, 42.9% occurred in long-term care facilities and 44.8% occurred in hospital inpatient setting [129]. Aggregating their reported values taken from death reports from the National Center for Health Statistics, we find that 12.14%, 22.73%, and 65.13% of all COVID-19 deaths occurred outside of the hospital and emergency departments from January 2020 to March 2021, in long-term care facilities, and within inpatient hospital settings, respectively [129]. Currently, it is not clear whether these deaths would have occurred at a shorter, or longer, length of time. However, since most deaths occurred in inpatient hospital settings, this is unlikely to cause major differences in the final results of this chapter.

Assumption 2- $\tau_{IH} \leq \tau_{IDD}$

The time to hospitalization should be no greater than the days to death without a hospitalization. This assumption accounts for the fact that if an individual were feeling extremely unwell, they would either seek medical help and hospitalization, and if unable to, would be at increased risk of death at that point in time.

Assumption 3- $\rho_{IUH} \leq \rho_{IDH}$

The hospitalization rate for an undocumented case is no greater than the hospitalization rate for a lab-confirmed case. This assumption accounts for the fact that undocumented cases were more likely to be asymptomatic and less severe since individuals did not require a test for treatment.

After all filtering is complete, there remain 499 samples to feed into the first fitting period, corresponding to March of 2020.

Fitting Parameters

The remaining parameters, β , the transmissibility of the virus, and m , the percent of infections that are documented relate to the overall force of infection which directly impacts the estimation of the time-varying time-varying reproductive number $R_e(t)$. An effective contact in this model is estimated assuming complete mixing. The transmissibility of the virus determines the likelihood that an effective contact results in a new infection, i.e., probability of transmission from an infected, effective contact. The percent of infections that are documented will affect the number of undocumented

infections, hospitalizations, and deaths, as the rates differ among these disease states. Both of these parameters are dynamically estimated each week across all of the fitting periods, where the assumption is that they are unconstrained, e.g., continuous between 0 and 1. Here, the definitions and calculations are provided to process the results from the fitting algorithm into time varying lab-multipliers and time-varying reproductive numbers.

For a given week j in fitting period i , $1 \leq i \leq 5$, $1 \leq j \leq W_i$ where W_i is the number of weeks in fitting period M_i the following estimates are computed.

1. Time Varying Lab-Multiplier, $LM^{i,j}(t)$

$$LM^{i,j}(t) = \frac{1}{m_{i,j}}$$

2. Time Varying Reproductive Number $R_e^{i,j}(t)$

The time varying reproductive number is estimated by the spectral radius of the next generation matrix method [60, 61, 122]. The details of the derivation are provided in Appendix D.

$$R^{i,j}(t) = \beta(t)[m_{i,j}(C_d - C_u) - 1]$$

And therefore, the time-varying effective reproductive number can be computed using the following.

$$R_e^{i,j}(t) = \frac{\mathcal{S}(t)}{N} R^{i,j}(t)$$

Where $\frac{\mathcal{S}(t)}{N}$ is the proportion of the population that is susceptible at time t .

$$C_d = \frac{1}{\frac{(1-\rho_{IDH}-\rho_{IDD})}{\tau_{IR}} + \frac{\rho_{IDH}}{\tau_{IDH}} + \frac{\rho_{IDD}}{\tau_{IDD}}}$$

$$C_u = \frac{1}{\frac{1-\rho_{IUH}}{\tau_{IR}} + \frac{\rho_{IUH}}{\tau_{IUH}}}$$

Branching Algorithm

The overarching fitting algorithm, denoted the branching algorithm, for the entire backcasted analysis in 2020 is illustrated in Figure 4.4. The algorithm can be explained in four main components.

- i. Initial Latin Hypercube Sampling

The initial Latin Hypercube sampling uses the bounds given in Table 4.2. After the samples have been filtered according to the assumptions previously mentioned, e.g. 499 remaining samples, the parameters are used to feed 499 different SEIR models. The initial sampling is done once at the beginning of the algorithm. Each sample is then used to feed the branching

SEIR algorithm to find the best weekly estimates of $LM^{i,j}(t)$ and $R_e^{i,j}(t)$.

ii. Branching SEIR

The overview of the weekly fitting to the fitting parameters, β and m , of the branching SEIR algorithm is shown in Figure 4.3. The initial condition $y_0(1)$ for the first time period is estimated using the IFR analysis described in the first method of this Chapter. The number of deaths by each age, approximately 25 days from the beginning of the simulation, is divided by the age-specific IFR values to feed the exposed state as a summation. Then, for each week, k in the fitting period M_i the optimal $\{\beta_{i,k}, m_{i,k}\}$ pairing is found via the argmin of the following minimization function.

$$\{\beta_{i,k}, m_{i,k}\} = \operatorname{argmin}\{WMSE(C_k, SC_k) + \gamma_k WMSE(D_{k:k+3}, SD_{k:k+3})\}$$

C_k := 7-day average values of daily lab-confirmed cases by the New York times for week k .

SC_k := simulation estimated daily lab-confirmed cases for week k

$D_{k:k+3}$:= 7-day average values of daily deaths by the New York times for weeks $k:k+3$.

$SD_{k:k+3}$:= simulation estimated daily deaths by the New York times for weeks $k:k+3$.

WMSE := weighted mean squared error function with weights chosen as 1:2

for under and overestimation, respectively.

$$\gamma_k = \frac{\max\{C_k\}}{\max\{D_{k:k+3}\}}$$

The weighted mean squared error function is used to penalize values that result in overestimation greater than values that result in underestimation [132]. Specifically, the weights are 1 for any underestimation and 2 for overestimation, resulting in overestimation being doubly penalized compared to underestimation. Here, the weighted mean squared error function was chosen to allow for this extra penalization and was experimentally determined to avoid the model fitting to unrealistic values when trying to "correct" an overestimation between subsequent fitting periods. The status of each disease state, i.e., number of documented infection, exposed, undocumented infected, etc. is fed into the subsequent SEIR model. After the first week of the year, the new percent of infections that are documented, m are constrained to within 10% of the previous week's optimal values by this minimization function, aside for the first week in April, which is unconstrained due to the amount of uncertainty in March.

iii. Identify feed parameters

Before progression to the next fitting period, the algorithm performs a final evaluation of the model output for each parameter sample through a mean squared error of daily lab-confirmed

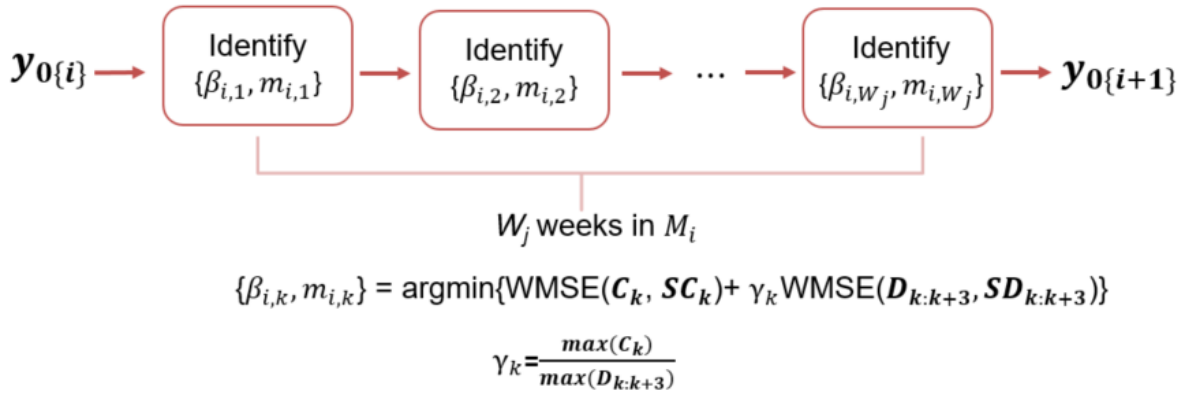


Figure 4.3 Overview of weekly fitting to percent of infections that are documented and the transmission rate. The number of people in each disease state, connects each week during a fitting period and each fitting period through updating the y_0 term for each SEIR model.

cases and daily deaths, compared with 7-day average empirical data reported by the New York Times. The error values are evaluated through a pseudo-algorithm to identify, approximately fifty samples that are a part of the effective frontier of sample parameters, after outliers by case error are removed, illustrated in Figure 4.5. The effective frontier is visualized in red. The yellow samples represent the top ten samples chosen from within the effective frontier to feed the next fitting period, or begin new branches. The branches are chosen based on the minimum death error within the set.

iv. Branch to the next fitting period

Once the top ten samples are identified, the algorithm branches to conduct ten more similar analyses. First, the new disease progression parameter samples are generated with updated bounds based on the selection of disease progression parameters from the previous fitting period. For the parameters that are estimated to change, based on literature surrounding the effective treatments, the new bounds are chosen with an upper bound of the input sample, and a lower bound that is 30% lower than the input sample value for each parameter. These special considerations are highlighted in Table 4.3. All other parameters are sampled such that the new bounds are within 25% of the incoming branch's parameters, where fifty new samples are generated for each branch, yielding 500 new samples for the next fitting period. This 25% constraint parameter was chosen due to uncertainty around the parameters particularly during the early months where it is easy to overfit to documented cases. The branching SEIR analysis is repeated for each sample and the rest of the algorithm repeats from step (ii) until the end of the year.

Branching Algorithm Between Fitting Periods

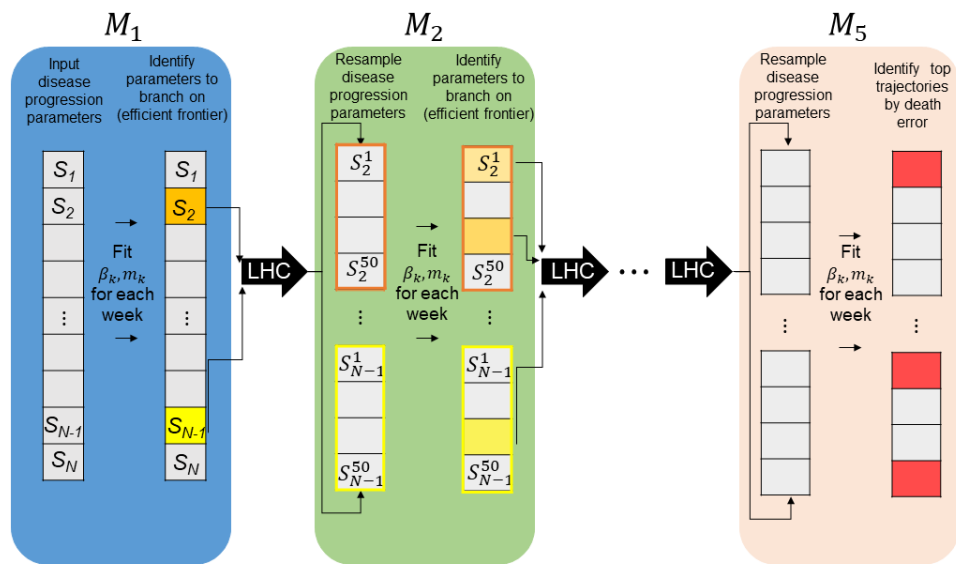
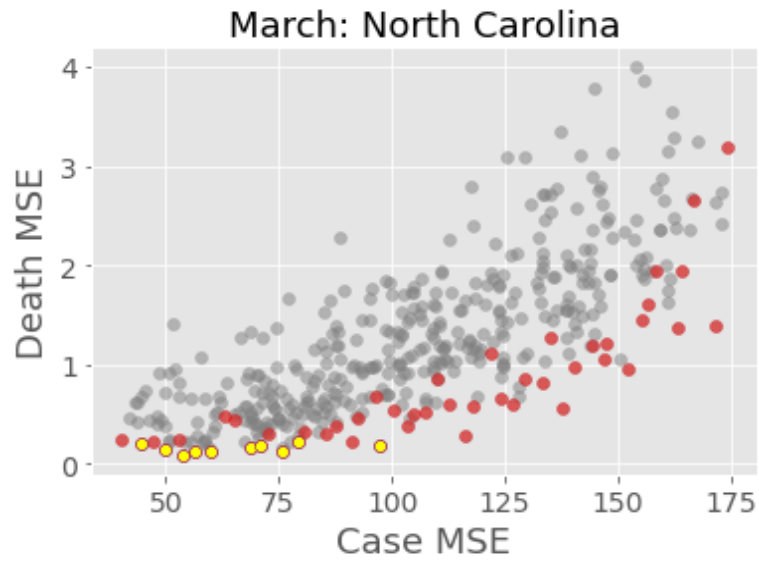
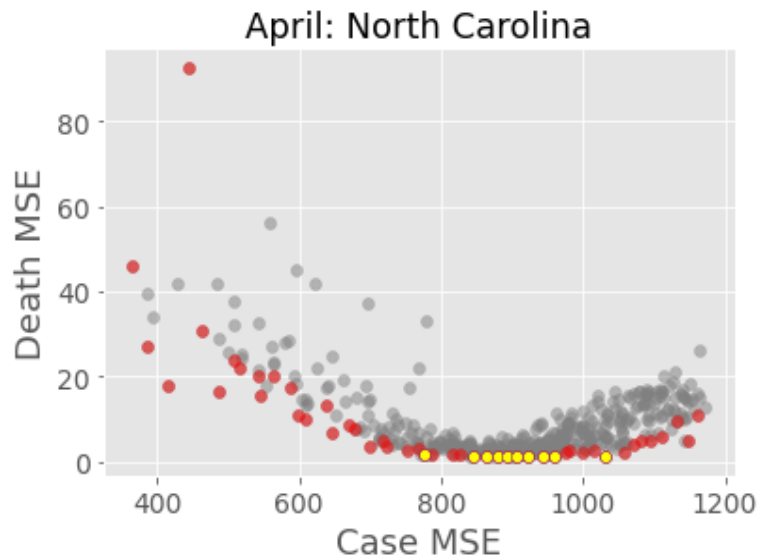


Figure 4.4 Overview of branching algorithm for each fitting period identified by the available treatments in 2020. The fitting periods are connected through a branching procedure where the top disease progression parameter samples, chosen through the efficient frontier, are used to create updated bounds for a new Latin Hypercube (LHC) sample of size 50. The first fitting period is fed with approximately 500 samples described by the bounds in Table 4.2. In the final fitting period, the top samples by death mean squared error are chosen to identify the best trajectories to estimate disease burden.



(a) March



(b) April

Figure 4.5 These figures illustrate the identified effective frontier for March (a) and April (b) using the branching algorithm. Each point in the plot represents a selection of disease progression parameters for each fitting period. The red samples indicate the effective frontier, whereas the yellow indicate the chosen samples to feed the next fitting period.

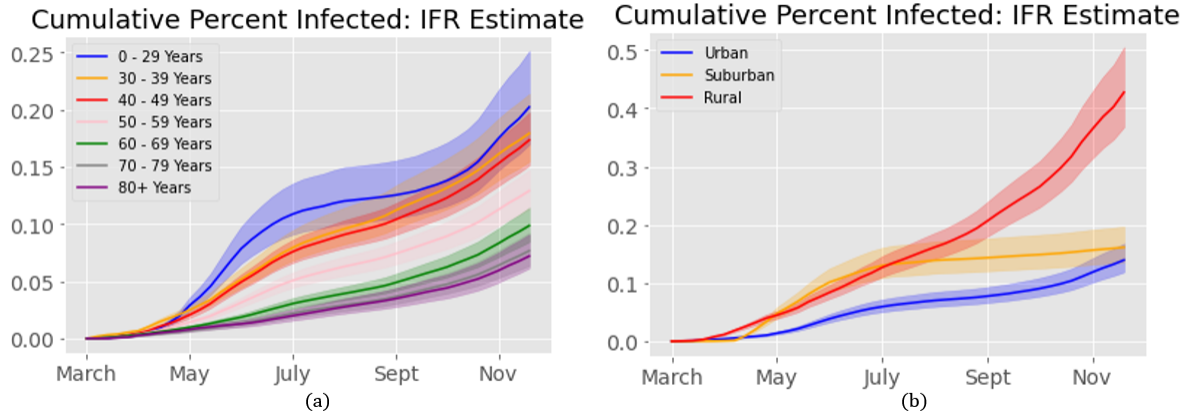


Figure 4.6 North Carolina cumulative percent infected by (a) age: 0-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80+ years of age and (b) geographic location: urban, suburban, and rural counties, as estimated for 2020 from the simple, deterministic application of IFR values to the NY Times reported deaths.

4.4 Results

IFR Analysis

The deterministic application of infection fatality rates to reported deaths is presented in Figures D.2- D.3, where the first shows the estimated lab multiplier at the state level across seven age groups, and the second breaks the analysis down further by geographic location. These figures show the estimates from the application of the IFR value reported in Levin et al. [104] and bands indicating the estimates from using the 95% confidence intervals around these reported IFR values. In both figures, we can see that younger age groups, particularly 0 - 29 years and 30 - 39 years tend to have higher rates of under-reporting, where the statewide lab multiplier for those 0 - 29 years of age reaches as high as 40+ during the early months of 2020. The black, dotted line indicates the first week of 2020 where the youngest age group, 0 - 29 years had a reported death. Before this date you can see that the estimated lab multiplier cannot be computed, and thus is assigned to be the value. This pattern occurs again in late August, where the youngest age group's lab multiplier drops to one because of insufficient data when using this method. The other age groups have lab multipliers that tend to decrease over the year. Furthermore, we can view the estimated cumulative percent infected by each age group in Figure 4.6. We can see that the youngest age group, 0 - 29 Years had the highest estimated cumulative percent infected until late summer, where the confidence intervals overlap with 30 - 39 and 40 - 49 years of age.

The rates of underreporting also vary significantly by geographic location. In Figure D.3 the age distributions indicate that there were higher rates of underreporting among rural areas, particularly for age groups 0 - 29 years and 30 - 39 years, compared to suburban and urban locations. The youngest age group in the rural locations had its highest peak in May, 2020, compared to earlier months in

suburban and urban locations. In suburban and urban locations we see that in October-November, the youngest age groups had a spike in lab multiplier values. Taking a look at the cumulative percent infected by the end of the year it is clear that rural areas had the highest estimated percent infected, with values ranging from 35-50% infected. Suburban areas experience a sharp incline in infection from April to July, reaching top values of approximately 15% infected with rates slowing down considerable soon thereafter and finally reaching approximately a 17-18% cumulative infection rate by the end of the year. The urban location trends indicate an infection spike in May-June and November - December.

Compartmental Model

The results of the compartmental model branching algorithm can be viewed in Figure 4.7a-c. Each graph shows five trajectories, indicated by five unique colors. The five trajectories are selected based on the mean squared death error from the final fitting period. The five samples with the smallest death error get traced back each fitting period to find the respecting connecting branch that fed into the final five and are plotted on these graphs.

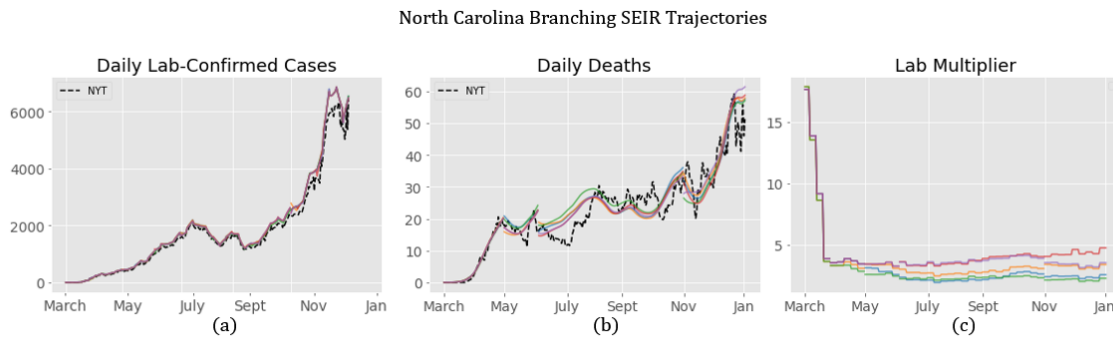


Figure 4.7 North Carolina model fit for estimated daily lab-confirmed cases (a), and daily deaths (b), compared to the 7-day averages reported by the New York Times. Panel (c) shows the estimated lab multiplier over time.

The final fits for each of the top five trajectories can be viewed in Figure 4.7 (a-b). It is clear that all five trajectories find similar, optimal fits on lab-confirmed cases throughout 2020, although yielding slightly different estimates on daily deaths. All top trajectories by death error yield an overestimation in cases in December, when the daily death tend positively upwards. This could indicate that the fatality rates have changed during this time period that this method is unable to catch due to imposing the the 25% requirement in the branching algorithm.

The state-wide estimated lab multiplier from this approach is shown in Figure 4.7(c). The lab multiplier in the first fitting period ranges from 20 in the beginning of the simulation period, down

to 5 towards the end of the first fitting period. This time period is crucial towards understanding the true disease burden before all age groups experienced deaths which is required for the IFR analysis. Similar to the trend in the IFR analysis, the simulation results suggest that there was additional underreporting in May 2020, where the lab multiplier tends above five for some trajectories.

Table 4.4 North Carolina final estimates for the top five trajectories chosen by the end of the final fitting period. Here we report the state level estimates for IFR, cumulative percent infected in 2020, and overall lab multiplier when comparing total cumulative simulation infections and cumulative lab-confirmed cases reported by the New York Times.

Trajectory	Estimated IFR (%)	Cumulative Percent Infected	Overall Lab Multiplier
1	0.2954	23.00	4.46
2	0.5271	13.14	2.55
3	0.4136	16.83	3.26
4	0.3078	22.19	4.30
5	0.3585	19.29	3.74

Table 4.4 holds the estimated infection fatality ratio, overall cumulative percent infected, and the resulting lab multiplier when compared to total cases reported by the New York Times by the end of 2020. The IFRs range from 0.2954% to 0.5271%, which is smaller than the population weighted estimate from the IFR analysis yielding an NC IFR of 0.66% in 4.1. The cumulative percent infected ranges from 13.14% to 23% and the corresponding lab multipliers compared to the New York times are 2.55 to 4.46, which is within the range reported by the CDC nationally from February to December of 2020 [32].

Figure 4.9 shows the disease progression parameters chosen by each of the top five trajectories chosen by minimum mean squared death error by the end of 2020. Each colored box represents a different selection of samples, from a unique trajectory, for which you can observe the disease progression parameters from March through November. The hospital fatality rate appears to decrease over time, especially from April to June across all five trajectories. This approach also enables us to observe different combinations of estimates for parameters that are unknown in the literature such as the documented infection fatality rate without hospitalization and the undocumented infection hospitalization rate, which varies in magnitude across all trajectories. The hospitalization duration decreases from April to June across all but one trajectory. The infectious period tends to decrease over time, with the lowest values reported in the November fitting period for each trajectory, which could indicate the effectiveness of certain stay at home orders or other mitigation techniques to reduce infectivity. Lastly, the hospital fatality rate slightly increases from June to November in some trajectories, which could indicate an overloading on the health care system which causes a higher death rate despite available treatments.

Finally, the results of the branching algorithm are compared to estimates produced by other studies using different methods to compute disease incidence and reproductive number over time.

North Carolina Branching SEIR Compared to Other Approaches

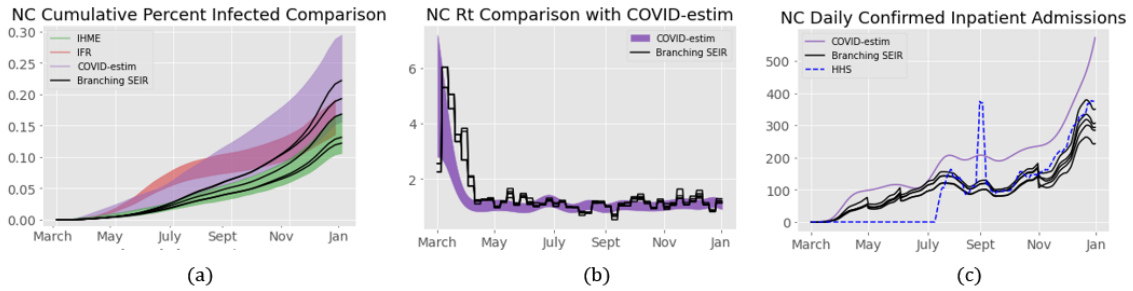


Figure 4.8 Comparing branching SEIR estimates for (a) daily lab-confirmed cases, (b) time varying effective reproductive number, and (c) daily confirmed COVID inpatient admissions with other approaches. The branching SEIR model trajectories are provided by the black lines, where other studies have colors indicated in the legend. The United States Department of Health and Human Services (HHS) did not begin reported COVID inpatient admissions until July 2020.

A comparison of the branching SEIR and IFR analysis of this chapter, along with two other models, one from IHME and another called COVID-estim, is provided in Figure 4.8. First, note that the IFR analysis estimates higher levels of cumulative infections around June. Afterwards, the IFR bounds fall within the bounds estimated by both IMHE (green) and COVID-estim (purple). The black lines indicate the results from the branching SEIR approach. Observe here that the estimates fall within the ranges estimated by both IHME, and COVID-estim. In Figure 4.8 the time-varying reproductive number is compared for the entire year. In April, the branching SEIR reports a higher reproductive number, 3-4 versus 1.5 estimated by COVID-estim. Furthermore, we can compare the weekly lab multiplier estimates from the branching algorithm and the weekly percent test positivity, illustrated in Appendix Figure D.4, where we observe in most trajectories the weekly lab multipliers are moderately positively correlated with weekly test positivity estimates.

Extensions and Reproducibility

The algorithms presented in this Chapter are not built solely for the state of North Carolina. This analysis can be repeated for any state for which there exists sufficient and decent data reporting. Furthermore, this model can be run serially, or in parallel, using a high performance computing system, which makes the analysis scalable for more granular models as well. The models can be expanded to include more compartments, such as age or vaccine compartments to understand more about disease dynamics over time. The model can also be transformed into a "nowcasting" model instead of the current "backcasting" version that was intended for this analysis. The current implementation requires a minimum of three weeks of death data in order to estimate the best fits for percent of documented infected and transmissibility over time. In order to transform to a "nowcasting" model, certain assumptions would have to be made on parameter limits and how to

Latent period	Infectious period	Time from hospital to death	Time to death	Time to hospital	Hospital duration	Doc. Infection fatality rate without hospitalization	Undoc. Infection hospital rate	Doc. Infection hospitalization rate	Hospital fatality rate	
6.5484	9.8311	10.4477	9.0468	8.7	13.1742	0.0004	0.0638	0.1795	0.207	March
5.0473	7.5736	11.0681	8.7937	9.8722	11.8434	0.0004	0.0767	0.1816	0.2154	April
6.1595	5.6922	8.6323	10.7895	9.0199	9.2845	0.0004	0.0711	0.1297	0.1736	May
6.7259	4.506	10.0507	12.1932	10.1442	8.1642	0.0004	0.0824	0.1564	0.1395	June
7.7589	3.7816	9.0832	10.7227	8.9218	7.4116	0.0005	0.0776	0.1111	0.1112	Nov
6.4079	8.7812	12.9644	9.2033	7.0429	11.8896	0.0076	0.0992	0.1472	0.16	March
7.3423	8.8468	15.373	9.3665	6.3152	10.3538	0.007	0.103	0.1561	0.1664	April
6.4253	6.636	19.1438	10.6824	7.055	9.445	0.0077	0.0884	0.1265	0.1563	May
6.5643	5.2379	22.8044	11.5911	7.6926	8.8437	0.0078	0.077	0.0994	0.1489	June
5.522	4.2096	18.0498	12.3261	9.1679	7.8765	0.009	0.0611	0.0984	0.1172	Nov
6.4079	8.7812	12.9644	9.2033	7.0429	11.8896	0.0076	0.0992	0.1472	0.16	March
6.5988	9.8547	14.9175	11.4472	7.8859	14.2294	0.0079	0.0828	0.1172	0.1966	April
7.9959	7.5868	11.7169	10.8761	9.0074	10.9457	0.0099	0.0679	0.1148	0.1465	May
9.638	6.0624	12.4973	11.4051	10.1604	9.56	0.0088	0.0838	0.0982	0.1088	June
10.5342	6.2506	12.9575	13.3306	10.376	7.6753	0.0071	0.0838	0.0743	0.0906	Nov
6.5484	9.8311	10.4477	9.0468	8.7	13.1742	0.0004	0.0638	0.1795	0.207	March
5.0473	7.5736	11.0681	8.7937	9.8722	11.8434	0.0004	0.0767	0.1816	0.2154	April
5.3823	5.8416	9.4402	9.0195	9.1289	10.0725	0.0003	0.0541	0.1498	0.1564	May
4.4711	5.2829	10.8215	9.7226	9.3892	7.7258	0.0002	0.0621	0.1203	0.1165	June
4.9511	4.0564	13.3223	11.1677	9.4253	6.5236	0.0003	0.0634	0.0878	0.1294	Nov
6.5484	9.8311	10.4477	9.0468	8.7	13.1742	0.0004	0.0638	0.1795	0.207	March
5.0473	7.5736	11.0681	8.7937	9.8722	11.8434	0.0004	0.0767	0.1816	0.2154	April
5.3823	5.8416	9.4402	9.0195	9.1289	10.0725	0.0003	0.0541	0.1498	0.1564	May
4.8103	4.9069	10.7117	8.2231	10.0071	9.0154	0.0003	0.0572	0.1154	0.121	June
5.9437	4.3154	13.3533	6.4675	8.8596	7.9872	0.0002	0.0587	0.1109	0.124	Nov

Figure 4.9 North Carolina disease progression parameters for top five trajectories chosen by mean squared death error at the end of the final fitting period.

fit on the most recent data in order to make the best estimates for the future time periods.

The methods proposed, defined, and explored in this chapter provide a data-driven framework towards understanding disease spread. These approaches are particularly useful in instances where a lot is unknown about the disease, particularly in the early months of disease spread, or in the face of new variants which have shown to have immune-escaping properties, such as what the United States is currently facing under the uncertainty surrounding the latest subvariants as of July, 2022 that are suggested to have immune-escaping properties from previous infection and vaccination [84]. These methods can help estimate new disease parameters in the face of uncertainty of new strains, especially if immune-escape suggests an effectively fully susceptible population. In addition to COVID modeling, the framework presented in this chapter can be adapted and transformed to inform public health policy regarding any infectious disease with unmitigated and undetected spread. This tool can be used to identify populations that have historically been subject to disproportional amounts of disease burden and inform policy to improve population health outcomes.

4.5 Discussion

This analysis supports the literature that indicates that there was a substantial amount of under reporting of true number of infections of COVID-19 during 2020 [10, 42, 131, 141]. Both methods described in this chapter yield that under reporting was most severe during the first few months of the pandemic when testing was unavailable and inaccessible and measures such as mask wearing and social distancing were not fully enforced or utilized.

The IFR approach allows for the lab multiplier analysis by age and geographic location and tells us that under reporting was more severe for younger age groups, particularly those 0-39 years of age where we see that by the end of 2020 held the highest values of cumulative percent infected. Additionally, this analysis allows for the comparison of ages across geographic locations. The results indicate that lab multipliers vary, and were significantly higher, 200 versus 45 for 0 - 29 years of age compared to urban locations. The peaks also occur during different parts of the year. Anecdotally, this makes sense considering the fact that urban areas were more likely to have a case importation followed by disease spread, especially in higher density areas. We also observe from the cumulative percent infected that by the end of the year the rural areas held the highest estimated cumulative percent infected. Other studies conjecture that this increase in disease risk in rural areas from August 2020 onwards may be explained by socio-economic differences, such as the ability to work from home [65].

The IFR approach provides useful insights towards understanding disease burden, but has limitations. First, the results are dependent upon the specific IFR values from the literature. As noted earlier, many studies have provided estimates of IFR values which have shown to change considerably over age groups and for different countries. To the best of my knowledge, I have not seen IFR estimates at the state level, aside from population weighting in other studies. Selecting a misrepresenting selection of IFR values could yield misleading results. Additionally, assuming

the same IFR across all states, counties, etc. could be further misleading as the population health composition and medical care access are known to vary considerably across state lines [138]. The IFR analysis is also dependent upon accurate and sufficient death reporting data. As noted above and indicated by the black line in Figure D.2, even with grouping the younger age groups there were no deaths reported for 0 - 29 years in the first month of the pandemic so the analysis is unable to provide an appropriate estimate during the most uncertain time of the year.

To make up for these dependency limitations, this chapter proposes and explores the use of the branching SEIR model to estimate disease progression parameters and disease spread over time. The model allows estimates of documented and undocumented cases of COVID-19 as well as deaths and hospitalizations. The hospitalizations were not shown due to insufficient data reporting in the early months. The branching SEIR model suggests that there are multiple solutions which provide reasonable fits to cases and deaths from an epidemiological model. This further supports the notion of uncertainty surrounding these estimates. In addition you can observe that in comparison to other models in Figure 4.8 the bounds on the estimated percent of cumulative infections vary considerably depending on the assumptions of the model. The branching SEIR model does not account for differences across age or geographic location, but can be extended to estimate those values for areas where data reporting is sufficient enough to estimate the necessary parameters and provide the model with enough data for multiple minimization fits. Fitting on cases and deaths allows for data-driven approaches, e.g., the branching method, to find appropriate estimates for the reproductive number especially during times when disease transmission was so uncertain.

The branching SEIR model approach is not without limitations itself. First, the current implementation assumes that all diagnoses cases and COVID related deaths are reported in the empirical data, similar to other studies where the true values are unknown [42]. The model also does not incorporate excess mortality. Second, the model assumes that disease progression parameters can only vary within 25% between fitting periods. This could not allow for appropriate changes in mortality rates when treatments were more effective and in turn may not fully account for changes in disease severity over time, and will likely need to be less strict when new variants are introduced. Lastly, the model currently does not include reinfection, which is likely not to have an impact in 2020, since COVID immunity was estimated to last at least six months for the early, wild type strain [56].

4.6 Conclusion

Overall, the branching SEIR model estimates that the true number of estimated cumulative infections is approximately 3-4 times higher than what was reported for North Carolina and that the overall IFR values for the state could range anywhere from 0.2954% to 0.5271%, where the population weighted IFR suggested by the application of age specific Levin's IFR values falls outside of this range. This implies that this data-driven approach suggests lower IFR values for the state compared to what is reported in the literature. This suggests that treatments that lowered mortality risk for patients with severe COVID-19 should be considered when modeling disease dynamics over time. The branching

method on fitting periods dependent upon the available effective treatments allows to explore this possibility and capture differences among disease progression over 2020.

This model allows us to determine time periods where the reproductive number of the disease changes over time, and compare with other estimates. In the early months, when not much was known about the disease, the estimates vary considerably as seen in Figure 4.8. Over time, the reproductive number can provide estimates of disease spread and inform tradeoffs between policies and human behavior. Such identified changes in this value can point to multiple mitigation measures or lack thereof, e.g., increased mask compliance or mask mandates, decreased mobility, school closures, and social distancing, and other concerns such as the introduction of more infectious variants, outbreaks due to other events or confined or dense living spaces [124, 126]. Understanding how each of these policies impacts disease dynamics can help inform public health policy. In addition to the reproductive number, we can view the changes in disease progression parameters. The resulting non-linear parameters, indicate that overall hospitalization and fatality rates decreased until the November fitting period. This is consistent with national findings reported by Pei et al., where the national IFR and CFR were estimated to reduce from 0.77% to 0.31% and 7.1% to 1.29%, respectively [131].

This analysis is also important to identify more potential utilization of the multiple data sources that have been made available to researchers throughout the pandemic. An important conclusion is that data reporting is crucial to identify and evaluate disease progression trends over time. Further analysis These measures are necessary to consider to understand (i) which populations remain the most vulnerable; (ii) which populations were disproportionately affected; and (iii) which mitigation measures had the most substantial impact on curbing the spread of disease. More granular results are dependent upon the availability of data. For instance, finding data to perform this analysis by race and ethnicity have proven to be difficult, particularly at the state and sub-regional levels. More data reporting on severe outcomes and IFR values by race and ethnicity would be required in order to complete the same analyses.

Finally, the results of this analysis can inform the following. The number of true infections over time, and particularly during the first few months of the pandemic, can be estimated for a given location. This number can aid those suffering from symptoms of long-COVID who are still unable to receive assistance due to not being able to prove they had the disease. For instance, as of July 2022, the COVID Recovery Clinic at the University of North Carolina Medical Clinic requires that individuals have a documented COVID positive test in addition to a month of lingering symptoms before they can be considered for evaluation [118]. This puts many at odds considering the estimated number of unreported infections was likely 3 to 4 times higher than the reported number of lab-confirmed cases, and higher during times when testing was unavailable. This analysis is also relevant as the number of undocumented infections is likely worsening as the use of at-home antigen tests are becoming more widely available without any way to report test results. Lastly, the methods proposed here allow for the changes of disease progression parameters as new treatments became available in 2020. The results indicate that the hospitalization rate was likely declining over the early stages of the

pandemic due to effective strategies and available treatments [11, 49]. This analysis can be used as a modeling recommendation tool for other disease modelers. The results expressed from this study here strongly suggest that modelers take into account new treatments as they become available over time to ensure proper model calibration. On a similar note, the branching algorithm can be used to sample possible parameters as an alternative method to estimating disease parameters, particularly in the face of uncertainty.

CHAPTER 5

DISCUSSION

The methods proposed in this thesis support the use of data-driven solutions to aide, inform, and communication public health policy. Each chapter worked to build upon the notion that big data can be utilized to (i) assess and evaluate risk; (ii) recommend intervention policies; and (iii) evaluate overall disease burden associated with infectious, respiratory diseases in the United States. In the first chapter, two infectious respiratory diseases were introduced and described by their respective current challenges within the United States healthcare system. For instance, even though pneumonia has historically been known to cause substantial burden on healthcare systems each year, the disease still manages to evade medical experts who are unable to pinpoint exactly who will fall ill and require subsequent hospitalization [30, 167]. On the other hand, COVID-19, a novel coronavirus that has swept the nation since early 2020, has brought on a plethora of new challenges in the realm of infectious disease knowledge concerning disease characteristics, transmissions dynamics, and population susceptibility, just to name a few [10, 42, 131]. The ideas presented contribute to the literature and provide the groundwork for developing data-driven tools relying on the intersection of health systems and data analytics to draw insights and create interpretable, practical solutions. The contributions for each chapter are summarized as follows.

The second chapter explored the use of administrative claims data to identify at-risk populations for pneumonia hospitalizations. The analysis centered upon the notion of 30-day risk to identify risk factors that can flag individuals to target with interventions and potentially avoid hospitalizations. Intervening before hospitalization is important due to the associated risk of mortality and long-term complications that follow a severe case of pneumonia [38, 137]. These long term complications can include chronic illness that can oftentimes lead to overall worsened quality of life. To combat this, three models were developed to assess the use of administrative claims and publicly available information in evaluating an individual's 30-day risk of pneumonia hospitalization. Unsurprisingly, the results indicated that the while predicting these rare, but serious events is difficult from a machine learning perspective, it is possible with current advancements to produce meaningful features that

can be computed from an administrative standpoint. Developing the model for administrative claims data means that the results are reproducible for any set of individuals with available medical history. The models presented in this work did not depend on cohorting based on specific pre-existing conditions, like other studies have in the past [98, 143], and instead were created to be generalizable to the public. Additionally, the models all signified an importance to estimating influenza activity level in assessing a 30-day hospitalization risk, which is a community level feature that can be further generalized. Considering that influenza undergoes strict surveillance every year by the CDC, it could be beneficial to improve or expand public health communication to areas that are historically known to have high incidence of influenza. In all the results of these models can be used to inform public health communication and intervention targeting for pneumonia.

In addition to evaluating pneumonia risk, the results of chapter two can be altered or used to inform and identify risk populations for other infectious respiratory diseases. For instance, the risk factors for COVID-19 are similar to those reported for severe pneumonia. It is known that the risk of death increases greatly with age as well as with the presence of certain medical conditions [28, 103]. The CDC identifies high risk groups as those having one or more of the following comorbid conditions: cancer, chronic kidney disease, chronic liver disease, chronic lung disease, cystic fibrosis, dementia, diabetes, disabilities, heart conditions, HIV infection, immunocompromising conditions, obesity, sick cell, and more [37]. These risk factors were used to identify vulnerable populations to guide public health communication and support the adoption of policies such as mask wearing and social distancing to reduce overall mortality rates. For instance, knowing that the risk of death greatly increases with age, several grocery stores across the country adopted policies to allow special hours for the elderly to complete their grocery shopping before opening to the general public [91], which are still in place in Summer 2022. In addition, several schools and workplaces transitioned to virtual activity to reduce interactions and lessen the likelihood of workplace or school exposure. Others incorporated some accommodations to help protect workers from exposure such as enforcing mask mandates, installing air filtration systems, and scheduling more frequent disinfecting procedures and so on. These examples highlight how understanding risk factors of severe disease can help the public make more informed decisions to improve health outcomes and reduce mortality. The methods presented in chapter two can help identify risk factors for other infectious respiratory disease and highlight how predictive models can be used to inform individual risk as the pandemic continues to evolve.

Chapter three describes a novel method to provide intervention policy recommendations to target an at-risk population. Resource allocation problems in the literature, specific to healthcare, have not considered risk within a population [62, 85, 178]. Using data-driven techniques, risk scores can be incorporated into well-known optimization problem formulations such as with the newsvendor model discussed within the chapter. The structure of the revenue management type model allows for quick computations and policy recommendations, which means that computational power and time is not a concern for implementing this model for any hospital system, as was the intention. The model was designed to create an additional motive for the incorporation of data-driven models

in medical decision making by making it financially desirable for administrations to successfully target interventions towards the at-risk and vulnerable populations. Furthermore, as the results indicate, the best policy recommendations are dependent upon the effectiveness of the available interventions. To further reduce hospitalizations, and health-related costs, more work should be focused on the development of more effective interventions. Even more, efforts should be placed on educating the public on available interventions and prevention tools so that they can make the most informed decision regarding their own health and in turn, population health will improve.

Finally, chapter four introduces a data-driven algorithm to inform population transmission dynamics and disease characterization. A deterministic, compartmental model is developed to answer a long withstanding question, how many people have been infected with COVID-19? The framework designed in this chapter is generalizable to any state, barring decent data reporting required to fit the model, e.g., lab-confirmed cases and deaths. The model dynamically estimates disease parameters and ascertainment rates and provides estimates of true cumulative infections over time. In comparison to other studies, which rely on seroprevalence surveys and a computationally expensive Bayesian framework [10, 42, 131] in addition to empirical data for fitting and calibration, the algorithm designed in this study only relies on empirical data and "some" knowledge regarding available treatments. In the backcasting approach we can take for granted that we know what treatments were available, but otherwise there were no assumptions on the impact each treatment had on overall disease parameters, so in essence these changes could be estimated over time via model fit evaluation. Splitting the backcasting year into fitting periods defined by the available treatments provides another tool for public health experts to evaluate the impact of mitigation measures over time. This can help inform the impact of certain treatments, even after an initial EUA was rescinded. The results presented in this chapter can aide individuals who are currently facing the issue of having to prove that long-lasting ailments are due to a prior COVID infection. This tool can be useful for the public and administrators in deciding where to focus relief efforts based on the population who may be susceptible to long-COVID and intervention efforts towards the remaining susceptible population.

Health Analytics Insights

The algorithms and methods used in this thesis draw insights towards the current status of data analytics in medical decision making. Particularly, the use of high-performance computing and distributed computing made the aggregation, synthesis, and evaluation of large and complex volumes of data possible to be transformed into tools to guide public health analyses. These recent state-of-the-art advancements are rapidly transforming the way that researchers can approach medical decision making. These tools can be utilized to transform unique datasets to draw new insights and develop dynamic tools in real time. This work also points to which datasets were insightful and worth pursuing in terms of prolonged data collection, availability, and analyses. Below are some insights gained from all of the work completed in this thesis.

i. Claims Data

Administrative claims data has been used by insurance companies for fraud detection and by researchers to perform cohorting analyses. The use of administrative claims data can be explored further to define meaningful features. Over the course of this work, I have noticed that there is a need for meaningful groupings of administrative claims codes. Chapter 2 tries to identify certain grouping through variable creation, but specific groupings were not testing beyond pneumonia hospitalization risk. Identifying diagnosis and procedure codes and the time it takes between their appearance and an adverse event could be extremely useful for researchers.

ii. Mobility Data

Although the algorithm defined in chapter 4 does not rely on mobility data, several useful insights were drawn from the data availability. Pictured in Appendix D, you can view the relationship between the publicly report Google mobility data and lab-confirmed cases over time. These preliminary results were used to draw insights on human behavior and contact patterns that govern transmission dynamics as seen in [146]. The continued collection and distribution of this type of data could prove useful in a long term longitudinal analysis regarding behaviors over time with regards to new variants, influenza, and other infectious diseases.

iii. Race and Ethnicity Data

A major letdown in the vast amounts of available data regarding the COVID-19 pandemic was the unavailability of accessible data by race and ethnicity. In my experience, very few datasets reported these metrics and those that did yielded highly sparse information. For future data collection aficionados, the collection of race and ethnicity is imperative and crucial for the conduction of fair research practices concerning public health. Data suggests that there are significant health inequities both historically, and most notably with COVID-19 [32, 146] present within our society. It is unfair to consider best public health outcomes and trajectories without giving special consideration to those who are historically underrepresented. Yet, it is impossible to sufficiently consider inequities when race and ethnicity data is not reported or available.

Future Work

The work presented in this thesis can be extended in a variety of ways. First, the models can be transformed to study similar research questions for other infectious diseases or general topics. The predictive models can be altered to consider other hospitalizations or adverse events. The intervention policy recommendation can be used to suggest intervention strategies for other areas of risk such as with other diseases, or other disciplines such as education retention or crime. The modeling algorithm can be used for other infectious diseases with undocumented infections, for instance HIV/AIDS, measles, etc. Second, the analyses could be repeated for alternative data sets to evaluate the differences in effectiveness for instance, a Medicare versus Medicaid population for

risk evaluation. Lastly, all of the the analyses can be expanded to evaluate multiple intervention strategies and recommend the best mitigation practices for a specific problem.

BIBLIOGRAPHY

- [1] Adibuzzaman, M. et al. “Big data in healthcare – the promises, challenges and opportunities from a research perspective: A case study with a model database”. *AMIA Annual Symposium Proceedings* **2017** (2018), pp. 384–392. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977694/> (visited on 10/08/2021).
- [2] AHRQ. *Prevention Quality Indicator 11 (PQI 11) Bacterial Pneumonia Admission Rate*. 2016. URL: https://qualityindicators.ahrq.gov/Downloads/Modules/PQI/V60-ICD09/TechSpecs/PQI_11_Bacterial_Pneumonia_Admission_Rate.pdf.
- [3] Arikan, E. *Single Period Inventory Control and Pricing*. 2011. URL: <file:///C:/Users/jmele/Downloads/1003180.pdf>.
- [4] Arrow, K. J. et al. “OPTIMAL INVENTORY POLICY”. *ECONOMETRICA* **19** (3 1951), pp. 250–272.
- [5] Ashwood, J. S. et al. “Direct-To-Consumer Telehealth May Increase Access To Care But Does Not Decrease Spending”. *Health Affairs* **36.3** (2017).
- [6] Aujesky, D et al. “Prospective comparison of three validated prediction rules for prognosis in community-acquired pneumonia”. *AMERICAN JOURNAL OF MEDICINE* **118** (4 2005), pp. 384–392.
- [7] Bahloul, M. et al. “Impact of prone position on outcomes of COVID-19 patients with spontaneous breathing”. *ACUTE AND CRITICAL CARE* **36.3** (2021), pp. 208–214.
- [8] Banco, E. *Can long Covid lead to death? A new analysis suggests it could*. 2022. URL: <https://www.politico.com/news/2022/06/03/can-long-covid-lead-to-death-a-new-analysis-suggests-it-could-00036845>.
- [9] Banks, M. A. “Sizing up big data”. en. *Nature Medicine* **26.1** (2020), pp. 5–6. URL: <https://www.nature.com/articles/s41591-019-0703-0> (visited on 10/08/2021).
- [10] Barber, R. M. et al. “Estimating global, regional, and national daily and cumulative infections with SARS-CoV-2 through Nov 14, 2021: a statistical analysis”. *The Lancet* **399**.10344 (2022), pp. 2351–2380. URL: <https://www.sciencedirect.com/science/article/pii/S0140673622004846>.
- [11] Beigel, J. H. et al. “Remdesivir for the Treatment of Covid-19 — Final Report”. *New England Journal of Medicine* **383**.19 (2020), pp. 1813–1826. eprint: <https://doi.org/10.1056/NEJMoa2007764>. URL: <https://doi.org/10.1056/NEJMoa2007764>.
- [12] Bortoletto, P. & Little, S. E. “Assess the ‘value’ of a healthcare intervention, not just its price”. *BJOG: An International Journal of Obstetrics & Gynaecology* **126.4** (2019), pp. 525–525. eprint: <https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1111/1471-0528.15201>. URL: <https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1111/1471-0528.15201>.

- [13] Bosco, E. et al. “Geographic Variation in Pneumonia and Influenza in Long-Term Care Facilities: A National Study”. *Clinical Infectious Diseases* **71** (8 2020), e202–e205. URL: <https://doi.org/10.1093/cid/ciaa081>.
- [14] Brauer, F. & Castillo-Chavez, C. “Discrete Population Models”. *Mathematical Models in Population Biology and Epidemiology*. Ed. by Brauer, F. & Castillo-Chavez, C. Texts in Applied Mathematics. Springer, 2012, 49–90. URL: https://doi.org/10.1007/978-1-4614-1686-9_2.
- [15] Breiman, L. “Random Forests”. en. *Machine Learning* **45.1** (2001), pp. 5–32. URL: <https://doi.org/10.1023/A:1010933404324> (visited on 10/08/2021).
- [16] Bresnick, J. *Understanding the Many V's of Healthcare Big Data Analytics*. 2017. URL: <https://healthitanalytics.com/news/understanding-the-many-vs-of-healthcare-big-data-analytics>.
- [17] Bubar, K. M. et al. “Model-informed COVID-19 vaccine prioritization strategies by age and serostatus”. *Science* **371**.6532 (2021), pp. 916–921. URL: <https://www.science.org/doi/10.1126/science.abe6959> (visited on 10/08/2021).
- [18] Bull-Otterson, L. et al. “Post-COVID Conditions Among Adult COVID-19 Survivors Aged 18–64 and 65 Years — United States, March 2020–November 2021” (2022).
- [19] Byrne, A. W. et al. “Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases”. *BMJ Open* **10.8** (2020). eprint: <https://bmjopen.bmj.com/content/10/8/e039856.full.pdf>. URL: <https://bmjopen.bmj.com/content/10/8/e039856>.
- [20] Calvillo-King, L. et al. “Impact of Social Factors on Risk of Readmission or Mortality in Pneumonia and Heart Failure: Systematic Review”. en. *Journal of General Internal Medicine* **28.2** (2013), pp. 269–282. URL: <https://doi.org/10.1007/s11606-012-2235-x> (visited on 10/08/2021).
- [21] CDC. *COVID-19 Case Surveillance Restricted Access Detailed Data Data*. URL: <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t> (visited on 10/08/2021).
- [22] CDC. *Ending Isolation and Precautions for People with COVID-19: Interim Guidance*. <https://www.cdc.gov/coronavirus/2019-ncov/hcp/duration-isolation.html>. (Visited on 07/16/2022).
- [23] CDC. *Smallpox*. <https://www.cdc.gov/smallpox/index.html>. 2017. (Visited on 07/13/2022).
- [24] CDC. *Health, United States, 2019 – Data Finder*. 2018. URL: https://www.cdc.gov/nchs/healthus/contents2019.htm?search=Influenza_and_pneumonia.
- [25] CDC. *Cases, Data, and Surveillance*. en-us. 2020. URL: <https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html> (visited on 10/08/2021).

- [26] CDC. *Causes of Pneumonia*. <https://www.cdc.gov/pneumonia/causes.html>. 2020.
- [27] CDC. *Coronavirus Disease 2019 (COVID-19)*. en-us. 2020. URL: <https://www.cdc.gov/coronavirus/2019-ncov/variants/delta-variant.html> (visited on 10/09/2021).
- [28] CDC. *COVID-19 and Your Health*. en-us. <https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/diy-cloth-face-coverings.html>. 2020. (Visited on 10/09/2021).
- [29] CDC. *Healthcare Workers*. en-us. 2020. URL: <https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html> (visited on 10/08/2021).
- [30] CDC. *Pneumococcal Disease - Risk Factors and How It Spreads*. <https://www.cdc.gov/pneumococcal/about/risk-transmission.html>. 2020.
- [31] CDC. *COVID-19 Pandemic Planning Scenarios*. 2021. URL: <https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html>.
- [32] CDC. *Estimated COVID-19 Burden*. 2021. URL: <https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html>.
- [33] CDC. *In-hospital Mortality Among Hospital Confirmed COVID-19 Encounters by Week From Selected Hospitals*. 2021. URL: <https://www.cdc.gov/nchs/covid19/nhcs/hospital-mortality-by-week.htm>.
- [34] CDC. *Pneumococcal Disease - Pneumonia Can Be Prevented - Vaccines Can Help*. 2021. URL: <https://www.cdc.gov/pneumonia/prevention.html>.
- [35] CDC. *Disease Burden of Flu*. <https://www.cdc.gov/flu/about/burden/index.html>. 2022. (Visited on 07/13/2022).
- [36] CDC. *HIV*. <https://www.cdc.gov/hiv/default.html>. 2022. (Visited on 07/13/2022).
- [37] CDC. *People with Certain Medical Conditions*. <https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html#:~:text=Like%20adults%2C%20children%20with%20obesity,very%20sick%20from%20COVID%2D19.> 2022.
- [38] CDC. *Pneumococcal Disease - Symptoms and Complications*. 2022. URL: <https://www.cdc.gov/pneumococcal/about/symptoms-complications.html>.
- [39] CDC. *Pneumococcal Disease: Risk Factors*. <https://www.cdc.gov/pneumococcal/clinicians/risk-factors.html>. 2022.
- [40] Chen, T. & Guestrin, C. "XGBoost: A Scalable Tree Boosting System". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016). arXiv: 1603.02754, pp. 785–794. URL: <http://arxiv.org/abs/1603.02754> (visited on 10/08/2021).

- [41] Chick, S. E. et al. "Supply Chain Coordination and Influenza Vaccination". *Operations Research* **56** (6 2008). doi: 10.1287/opre.1080.0527, pp. 1493–1506. URL: <https://doi.org/10.1287/opre.1080.0527>.
- [42] Chitwood, M. H. et al. "Reconstructing the course of the COVID-19 epidemic over 2020 for US states and counties: results of a Bayesian evidence synthesis model". *medRxiv* (2021). eprint: <https://www.medrxiv.org/content/early/2021/07/22/2020.06.17.20133983.full.pdf>. URL: <https://www.medrxiv.org/content/early/2021/07/22/2020.06.17.20133983>.
- [43] CIO&Leader. *More Than Half Of Healthcare Organizations Expect An Increase In Demand For AI-Based Solutions During And After The Pandemic: IDC*. 2020. URL: <https://www.cioandleader.com/article/2020/09/23/more-half-healthcare-organizations-expect-increase-demand-ai-based-solutions> (visited on 10/08/2021).
- [44] CMS. *Flu Shot Coding*. <https://www.cms.gov/medicare/preventive-services/flu-shot-coding>. (Visited on 10/08/2021).
- [45] CMS. *Non-Identifiable Data Files*. 2008-2011. URL: <https://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/NonIdentifiableDataFiles>.
- [46] CMS. <https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes>. 2010.
- [47] CMS. *Seasonal Influenza Vaccines Pricing*. <https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Part-B-Drugs/McrPartBDrugAvgSalesPrice/VaccinesPricing>. 2011.
- [48] CMS. *Medicare Part B Drug Average Sales Price*. 2022. URL: <https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Part-B-Drugs/McrPartBDrugAvgSalesPrice>.
- [49] Cohen, E. '*Such a simple thing to do*': *Why positioning Covid-19 patients on their stomachs can save lives*. 2020. URL: <https://www.cnn.com/2020/04/14/health/coronavirus-prone-positioning/index.html>.
- [50] Cohen, M. S. et al. "Effect of Bamlanivimab vs Placebo on Incidence of COVID-19 Among Residents and Staff of Skilled Nursing and Assisted Living Facilities: A Randomized Clinical Trial". *JAMA* **326**.1 (2021), pp. 46–55. eprint: https://jamanetwork.com/journals/jama/articlepdf/2780870/jama_cohen_2021_oi_210061_1625172038.59986.pdf. URL: <https://doi.org/10.1001/jama.2021.8828>.
- [51] Community Health Centers, N. A. of. *Population Health Management: Risk Stratification*. 2019. URL: <https://www.nachc.org/wp-content/uploads/2019/03/Risk-Stratification-Action-Guide-Mar-2019.pdf>.
- [52] Cook, N. R. "Statistical Evaluation of Prognostic versus Diagnostic Models: Beyond the ROC Curve". *Clinical Chemistry* **54**.1 (2008), pp. 17–23. URL: <https://doi.org/10.1373/clinchem.2007.096529> (visited on 10/08/2021).

- [53] Cook, S. F. & Bies, R. R. “Disease Progression Modeling: Key Concepts and Recent Developments”. *Current pharmacology reports* **2.5** (2016), pp. 221–230. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5602534/> (visited on 10/08/2021).
- [54] Cramer, E. Y. et al. “The United States COVID-19 Forecast Hub dataset”. *medRxiv* (2021). URL: <https://www.medrxiv.org/content/10.1101/2021.11.04.21265886v1>.
- [55] *Critical Access Hospitals | CMS*. URL: <https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/CertificationandComplianc/CAHs> (visited on 10/08/2021).
- [56] Dan, J. M. et al. “Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection”. *Science* **371**.6529 (2021), eabf4063. eprint: <https://www.science.org/doi/pdf/10.1126/science.abf4063>. URL: <https://www.science.org/doi/abs/10.1126/science.abf4063>.
- [57] Dash, S. et al. “Big data in healthcare: management, analysis and future prospects”. *Journal of Big Data* **6.1** (2019), p. 54. URL: <https://doi.org/10.1186/s40537-019-0217-0> (visited on 10/08/2021).
- [58] Delua, J. *Supervised vs. Unsupervised Learning: What's the Difference?* <https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning>.
- [59] Desai, R. J. et al. “Identification of smoking using Medicare data- A validation study of claims-based algorithms”. *Pharmacoepidemiology and drug safety* **25.4** (2016), pp. 472–475. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4826837/> (visited on 10/08/2021).
- [60] Diekmann, O. et al. “The construction of next-generation matrices for compartmental epidemic models”. *JOURNAL OF THE ROYAL SOCIETY INTERFACE* **7.47** (2010), pp. 873–885.
- [61] Driessche, P. van den. “Reproduction numbers of infectious disease models”. *Infectious Disease Modelling* (2017).
- [62] Earnshaw, S. R. et al. “Optimal allocation of resources across four interventions for type 2 diabetes”. *MEDICAL DECISION MAKING* **22** (5 2002), S80–S91.
- [63] Elferjani, B. et al. “THE EFFECT OF PRONE POSITIONING ON MORTALITY IN COVID-19-RELATED ARDS: A RETROSPECTIVE ANALYSIS”. *CRITICAL CARE MEDICINE* **50.1**, S (2022), p. 135.
- [64] Elixhauser, A et al. “Comorbidity measures for use with administrative data”. *MEDICAL CARE* **36.1** (1998), pp. 8–27.
- [65] Entrepreneurship, O. C. for. *Addressing COVID-19 and Moving to Net Zero Greenhouse Emissions*.
- [66] *Explore Census Data*. = <https://data.census.gov>. (Visited on 10/09/2021).

- [67] Faes, C. et al. “Time between Symptom Onset, Hospitalisation and Recovery or Death: Statistical Analysis of Belgian COVID-19 Patients”. *INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH* **17.20** (2020).
- [68] Farrier, H. *Seagate Launches New Data-Readiness Index Revealing Impact Across Four Global Industries as 30 Percent of Data Forecasted to Be Real-Time by 2025*. en. 2018. URL: <https://www.businesswire.com/news/home/20181126005585/en/Seagate-Launches-New-Data-Readiness-Index-Revealing-Impact-Across-Four-Global-Industries-as-30-Percent-of-Data-Forecasted-to-Be-Real-Time-by-2025> (visited on 10/08/2021).
- [69] FDA. *Coronavirus (COVID-19) Update: FDA Authorizes Monoclonal Antibodies for Treatment of COVID-19*. <https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-authorizes-monoclonal-antibodies-treatment-covid-19-0>. 2021.
- [70] FDA. *Emergency Use Authorization 091*. <https://www.fda.gov/media/145610/download>. 2022.
- [71] Ferranna, M. et al. “COVID-19 Vaccine Allocation: Modeling Health Outcomes and Equity Implications of Alternative Strategies”. en. *Engineering* (2021). URL: <https://www.sciencedirect.com/science/article/pii/S2095809921001934> (visited on 10/08/2021).
- [72] Fine, M. J. et al. “A Prediction Rule to Identify Low-Risk Patients with Community-Acquired Pneumonia”. *New England Journal of Medicine* **336.4** (1997), pp. 243–250. URL: <https://doi.org/10.1056/NEJM199701233360402> (visited on 10/08/2021).
- [73] Fine, M. J. et al. “A Prediction Rule to Identify Low-Risk Patients with Community-Acquired Pneumonia”. *New England Journal of Medicine* **336** (4 1997). doi: 10.1056/NEJM199701233360402, pp. 243–250. URL: <https://doi.org/10.1056/NEJM199701233360402>.
- [74] Firouzi, F. et al. “Internet-of-Things and big data for smarter healthcare: From device to architecture, applications and analytics”. *Future Generation Computer Systems* **78** (2018), 583–586.
- [75] Foundation, K. F. *Vaccine Coverage, Pricing, and Reimbursement in the U.S.* <https://www.kff.org/report-section/vaccine-coverage-pricing-and-reimbursement-in-the-u-s-tables/>. 2020.
- [76] Gao, J. et al. “Predicting Potentially Avoidable Hospitalizations”. *Medical Care* **52** (2 2014). URL: https://journals.lww.com/lww-medicalcare/Fulltext/2014/02000/Predicting_Potentially_Avoidable_Hospitalizations.12.aspx.
- [77] Gatewood, J. et al. “Social Media in Public Health: Strategies to Distill, Package, and Disseminate Public Health Research”. *Journal of Public Health Management and Practice* **26** (5 2020). URL: https://journals.lww.com/jphmp/Fulltext/2020/09000/Social_Media_in_Public_Health__Strategies_to.14.aspx.

- [78] Gottlieb, R. L. et al. “Early Remdesivir to Prevent Progression to Severe Covid-19 in Outpatients”. *New England Journal of Medicine* **386.4** (2022), pp. 305–315. eprint: <https://doi.org/10.1056/NEJMoa2116846>. URL: <https://doi.org/10.1056/NEJMoa2116846>.
- [79] Grimwood, K & Chang, A. B. “Long-term effects of pneumonia in young children”. *PNEUMONIA* **6** (2015), pp. 101–114.
- [80] Groeneveld, G. H. et al. “Prediction model for pneumonia in primary care patients with an acute respiratory tract infection: role of symptoms, signs, and biomarkers”. *BMC Infectious Diseases* **19.1** (2019), p. 976. URL: <https://doi.org/10.1186/s12879-019-4611-1> (visited on 10/08/2021).
- [81] Hauser, A. et al. “Estimation of SARS-CoV-2 mortality during the early stages of an epidemic: A modeling study in Hubei, China, and six regions in Europe”. en. *PLOS Medicine* **17.7** (2020), e1003189. URL: <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003189> (visited on 10/08/2021).
- [82] Hayes, B. H. et al. “Burden of Pneumonia-Associated Hospitalizations: United States, 2001–2014”. *Chest* **153** (2 2018), pp. 427–437. URL: <https://pubmed.ncbi.nlm.nih.gov/29017956https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6556777/>.
- [83] Hines, S. & Joshi, M. S. “Variation in Quality of Care Within Health Systems”. *The Joint Commission Journal on Quality and Patient Safety* **34** (6 2008), pp. 326–332. URL: <https://www.sciencedirect.com/science/article/pii/S1553725008340410>.
- [84] Howard, J. *New coronavirus subvariants escape antibodies from vaccination and prior Omicron infection, studies suggest*. 2022. URL: <https://www.cnn.com/2022/06/22/health/ba4-ba5-escape-antibodies-covid-vaccine/index.html>.
- [85] Hynninen, Y. et al. “Operationalization of Utilitarian and Egalitarian Objectives for Optimal Allocation of Health Care Resources”. *Decision Sciences* **52** (5 2021). <https://doi.org/10.1111/dec.12448>, pp. 1169–1208. URL: <https://doi.org/10.1111/dec.12448>.
- [86] *ICD-9-CM Diagnosis and Procedure Codes: Abbreviated and Full Code Titles | CMS*. URL: <https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes> (visited on 10/08/2021).
- [87] *Introduction to Machine Learning with Python [Book]*. en. URL: <https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/> (visited on 10/08/2021).
- [88] Jain, S. et al. “Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults”. en. *New England Journal of Medicine* **373.5** (2015), pp. 415–427. URL: <http://www.nejm.org/doi/10.1056/NEJMoa1500245> (visited on 10/08/2021).
- [89] Ji, D. et al. “Prediction for Progression Risk in Patients With COVID-19 Pneumonia: The CALL Score”. *Clinical Infectious Diseases* **71** (6 2020), pp. 1393–1399. URL: <https://doi.org/10.1093/cid/ciaa414>.
- [90] July 28, a. a. Posted by Stephanie Glen on & Blog, V. *Decision Tree vs Random Forest vs Gradient Boosting Machines: Explained Simply*. en. URL: <https://www.datasciencecentral.com/>

com/profiles/blogs/decision-tree-vs-random-forest-vs-boosted-trees-explained (visited on 10/08/2021).

- [91] Kassraie, A. *Supermarkets Offer Special Hours for Older Shoppers*. 2022. URL: <https://www.aarp.org/home-family/your-home/info-2020/coronavirus-supermarkets.html>.
- [92] Katella, K. *5 Things To Know About the Delta Variant*. en. URL: <https://www.yalemedicine.org/news/5-things-to-know-delta-variant-covid> (visited on 10/09/2021).
- [93] Kim, D. H. et al. "Measuring Frailty in Medicare Data: Development and Validation of a Claims-Based Frailty Index". *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **73.7** (2018), pp. 980–987. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001883/> (visited on 10/08/2021).
- [94] Koivula, I. et al. "Risk factors for pneumonia in the elderly". *The American Journal of Medicine* **96** (4 1994), pp. 313–320. URL: <https://www.sciencedirect.com/science/article/pii/S002934394900604>.
- [95] Kolata, G. 2020. URL: <https://www.nytimes.com/2020/05/01/health/coronavirus-remdesivir.html>.
- [96] Koroukian, S. M. et al. "Ability of Medicare Claims Data to Identify Nursing Home Patients". *Medical care* **46.11** (2008), pp. 1184–1187. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3178883/> (visited on 10/08/2021).
- [97] Krumholz, H. M. "Variations in health care, patient preferences, and high-quality decision making". *JAMA* **310** (2 2013), pp. 151–152. URL: <https://pubmed.ncbi.nlm.nih.gov/23839747><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5459397/>.
- [98] Kuo, K. M. et al. "Predicting hospital-acquired pneumonia among schizophrenic patients: a machine learning approach". *BMC MEDICAL INFORMATICS AND DECISION MAKING* **19** (2019).
- [99] Lee, C. H. & Yoon, H.-J. "Medical big data: promise and challenges". *Kidney Research and Clinical Practice* **36.1** (2017), pp. 3–11. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5331970/> (visited on 10/08/2021).
- [100] Lee, H. L. & Nahmias, S. *Chapter 1 Single-Product, single-Location models*. Vol. 4. Elsevier, 1993, pp. 3–55. URL: <https://www.sciencedirect.com/science/article/pii/S0927050705801811>.
- [101] Levi, R. et al. "The Data-Driven Newsvendor Problem: New Bounds and Insights". *Operations Research* **63.6** (2015), pp. 1294–1306. eprint: <https://doi.org/10.1287/opre.2015.1422>. URL: <https://doi.org/10.1287/opre.2015.1422>.
- [102] Levin, A. T. et al. "Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications". en. *European Journal of Epidemiology* **35.12** (2020), pp. 1123–1138. URL: <https://doi.org/10.1007/s10654-020-00698-1> (visited on 10/08/2021).

- [103] Levin, A. T. et al. “Assessing the burden of COVID-19 in developing countries: systematic review, meta-analysis and public policy implications”. *BMJ Global Health* **7.5** (2022). eprint: <https://gh.bmj.com/content/7/5/e008477.full.pdf>. URL: <https://gh.bmj.com/content/7/5/e008477>.
- [104] Levin, E. G. et al. “Waning Immune Humoral Response to BNT162b2 Covid-19 Vaccine over 6 Months”. *New England Journal of Medicine* **0.0** (2021), null.
- [105] Lieu, T. et al. “Effectiveness and cost-effectiveness of letters, automated telephone messages, or both for underimmunized children in a health maintenance organization”. *PEDIATRICS* **101.4** (1998).
- [106] Lim, W. S. et al. “Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study”. en. *Thorax* **58.5** (2003), pp. 377–382. URL: <https://thorax.bmj.com/content/58/5/377> (visited on 10/08/2021).
- [107] Lim, W. S. et al. “Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study”. *Thorax* **58** (5 2003), p. 377. URL: <http://thorax.bmj.com/content/58/5/377.abstract>.
- [108] Lin, S. et al. “Establishment of a Risk Score Model for Early Prediction of Severe H1N1 Influenza”. *Frontiers in cellular and infection microbiology* **11** (2022), p. 776840. URL: <https://pubmed.ncbi.nlm.nih.gov/35059324https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8764189/>.
- [109] Liu, X. & Stechlin, P. *Infectious Disease Modeling: A Hybrid System Approach*. en. Nonlinear Systems and Complexity. Springer International Publishing, 2017. URL: <https://www.springer.com/gp/book/9783319532066> (visited on 10/08/2021).
- [110] Loaiza, S. *Gini Impurity Measure*. en. 2020. URL: <https://towardsdatascience.com/gini-impurity-measure-dbd3878ead33> (visited on 10/08/2021).
- [111] Louisville, U. et al. “Adults Hospitalized With Pneumonia in the United States: Incidence, Epidemiology, and Mortality”. *CLINICAL INFECTIOUS DISEASES* **65** (11 2017), pp. 1806–1812.
- [112] Luque, A. et al. “The impact of class imbalance in classification performance metrics based on the binary confusion matrix”. *Pattern Recognition* **91** (2019), pp. 216–231. URL: <https://www.sciencedirect.com/science/article/pii/S0031320319300950>.
- [113] Makam, A. N. et al. “Predicting 30-Day Pneumonia Readmissions Using Electronic Health Record Data”. *Journal of Hospital Medicine* **12.4** (2017), pp. 209–216. eprint: <https://shmpublications.onlinelibrary.wiley.com/doi/pdf/10.12788/jhm.2711>. URL: <https://shmpublications.onlinelibrary.wiley.com/doi/abs/10.12788/jhm.2711>.
- [114] Mao, L. et al. “Personalized Predictions for Unplanned Urinary Tract Infection Hospitalizations with Hierarchical Clustering”. *AI and Analytics for Public Health*. Ed. by Yang, H. et al. Cham: Springer International Publishing, 2022, pp. 453–465.

- [115] Mazer, B. *Long COVID Could Be a 'Mass Deterioration Event'*. 2022. URL: https://www.theatlantic.com/health/archive/2022/06/long-covid-chronic-illness-disability/661285/?utm_source=apple_news.
- [116] McCullers, J. A. "Effect of Antiviral Treatment on the Outcome of Secondary Bacterial Pneumonia after Influenza". *The Journal of Infectious Diseases* **190** (3 2004), pp. 519–526. URL: <https://doi.org/10.1086/421525>.
- [117] Medicine, J. H. URL: <https://www.hopkinsmedicine.org/health/conditions-and-diseases/pneumonia>.
- [118] Medicine, U. of North Carolina School of. *COVID Recovery Clinic*. URL: <https://www.med.unc.edu/phyrehab/patient-services/services-and-specialties/covid-recovery-clinic/>.
- [119] Mendez, R. et al. *Cardiovascular Complications of Respiratory Diseases*. European Respiratory Society, 2020. Chap. Cardiovascular consequences of community-acquired pneumonia and other pulmonary infections, pp. 212–218.
- [120] Minor, L. *Harnessing the Power of Data in Health*. <https://med.stanford.edu>. 2017.
- [121] Mueller, B. & Rabin, R. C. *Common Drug Reduces Coronavirus Deaths, Scientists Report*. 2020. URL: <https://www.nytimes.com/2020/06/16/world/europe/dexamethasone-coronavirus-covid.html?searchResultPosition=5>.
- [122] Mwalili, S. et al. "SEIR model for COVID-19 dynamics incorporating the environment and social distancing". *BMC RESEARCH NOTES* **13.1** (2020).
- [123] Niederman, M. S. et al. "Guidelines for the Management of Adults with Community-acquired Pneumonia". *American Journal of Respiratory and Critical Care Medicine* **163** (7 2001). doi: 10.1164/ajrccm.163.7.at1010, pp. 1730–1754. URL: <https://doi.org/10.1164/ajrccm.163.7.at1010>.
- [124] Noland, R. B. "Mobility and the effective reproduction rate of COVID-19". en. *Journal of Transport & Health* **20** (2021), p. 101016. URL: <https://www.sciencedirect.com/science/article/pii/S2214140521000104> (visited on 10/09/2021).
- [125] Olivares, M et al. "Structural estimation of the newsvendor model: An application to reserving operating room time". *MANAGEMENT SCIENCE* **54** (1 2008), pp. 41–55.
- [126] *Outbreaks and Cluster NC COVID-19*. <https://covid19.ncdhhs.gov/dashboard/outbreaks-and-clusters>. (Visited on 10/09/2021).
- [127] Pant, A. *Introduction to Logistic Regression*. en. 2019. URL: <https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148> (visited on 10/08/2021).
- [128] Pasquale, C. B. et al. "Patient-Reported Consequences of Community-Acquired Pneumonia in Patients with Chronic Obstructive Pulmonary Disease". *Chronic obstructive pulmonary diseases (Miami, Fla.)* **6** (2 2019), pp. 132–144. URL: <https://pubmed.ncbi.nlm.nih.gov/30974053https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6596434/>.

- [129] Pathak, E. B. et al. “Out-of-Hospital COVID-19 Deaths: Consequences for Quality of Medical Care and Accuracy of Cause of Death Coding”. *American Journal of Public Health* **111**.S2 (2021). PMID: 34314208, S101–S106. eprint: <https://doi.org/10.2105/AJPH.2021.306428>. URL: <https://doi.org/10.2105/AJPH.2021.306428>.
- [130] Pedregosa, F. et al. “Scikit-learn: Machine Learning in Python”. *Journal of Machine Learning Research* **12** (2011), pp. 2825–2830.
- [131] Pei, S. et al. “Burden and characteristics of COVID-19 in the United States during 2020”. *Nature* **598** (2021). URL: <https://doi.org/10.1038/s41586-021-03914-4>.
- [132] Perini, T. et al. “Agent-Based Simulation for Seasonal Guinea Worm Disease in Chad Dogs”. *AMERICAN JOURNAL OF TROPICAL MEDICINE AND HYGIENE* **103**.5 (2020), pp. 1942–1950.
- [133] Peters, G. M. et al. “The Effect of Telehealth on Hospital Services Use: Systematic Review and Meta-analysis”. *JOURNAL OF MEDICAL INTERNET RESEARCH* **23**.9 (2021).
- [134] *Pneumococcal Vaccination: Who and When to Vaccinate* / CDC. en-us. 2021. URL: <https://www.cdc.gov/vaccines/vpd/pneumo/hcp/who-when-to-vaccinate.html> (visited on 10/08/2021).
- [135] Postma, D. F. et al. “Antibiotic Treatment Strategies for Community-Acquired Pneumonia in Adults”. *New England Journal of Medicine* **372** (14 2015). doi: 10.1056/NEJMoa1406330, pp. 1312–1323. URL: <https://doi.org/10.1056/NEJMoa1406330>.
- [136] *Prospective comparison of three validated prediction rules for prognosis in community-acquired pneumonia*. en. URL: <https://read.qxmd.com/read/15808136/prospective-comparison-of-three-validated-prediction-rules-for-prognosis-in-community-acquired-pneumonia> (visited on 10/08/2021).
- [137] Ramirez, J. A. et al. “Adults Hospitalized With Pneumonia in the United States: Incidence, Epidemiology, and Mortality”. *Clinical Infectious Diseases* **65**.11 (2017), pp. 1806–1812. URL: <https://doi.org/10.1093/cid/cix647> (visited on 10/08/2021).
- [138] Rankings, A. H. <https://www.americashealthrankings.org/>.
- [139] Rasco, D. et al. “Student Retention: Fostering Peer Relationships Through a Brief Experimental Intervention”. *Journal of College Student Retention: Research, Theory & Practice* (2020). doi: 10.1177/1521025120972962, p. 1521025120972962. URL: <https://doi.org/10.1177/1521025120972962>.
- [140] (RECOVERY), R. E. of COVid-19 thERapY. *Low-cost dexamethasone reduces death by up to one third in hospitalised patients with severe respiratory complications of COVID-19*. <https://www.recoverytrial.net/news/low-cost-dexamethasone-reduces-death-by-up-to-one-third-in-hospitalised-patients-with-severe-respiratory-complications-of-covid-19>. 2020.
- [141] Reese, H. et al. “Estimated incidence of COVID-19 illness and hospitalization — United States, February–September, 2020”. *Clinical Infectious Diseases: An Official Publication of*

- the Infectious Diseases Society of America* (2020), cial1780. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7717219/> (visited on 10/08/2021).
- [142] Regeneron Pharmaceuticals, I. *PHASE 3 PREVENTION TRIAL SHOWED 81% REDUCED RISK OF SYMPTOMATIC SARS-COV-2 INFECTIONS WITH SUBCUTANEOUS ADMINISTRATION OF REGEN-COV™ (CASIRIVIMAB WITH IMDEVIMAB)*. 2021. URL: <https://newsroom.regeneron.com/news-releases/news-release-details/phase-3-prevention-trial-showed-81-reduced-risk-symptomatic-sars>.
- [143] Restrepo, M. I. et al. “Pneumonia in Patients with Chronic Obstructive Pulmonary Disease”. *TUBERCULOSIS AND RESPIRATORY DISEASES* **81.3** (2018), pp. 187–197.
- [144] *Revenue Center Code (FFS)/ResDAC*. URL: <https://resdac.org/cms-data/variables/revenue-center-code-ffs> (visited on 10/08/2021).
- [145] Rodrigues, E. et al. “Excess pneumonia and influenza hospitalizations associated with influenza epidemics in Portugal from season 1998/1999 to 2014/2015”. *Influenza and other respiratory viruses* **12** (1 2018), pp. 153–160. URL: <https://pubmed.ncbi.nlm.nih.gov/29460423https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5818339/>.
- [146] Rosenstrom, E. T. et al. “Can vaccine prioritization reduce disparities in COVID-19 burden for historically marginalized populations?” *PNAS Nexus* **1.1** (2022). pgab004. eprint: <https://academic.oup.com/pnasnexus/article-pdf/1/1/pgab004/42757305/pgab004.pdf>. URL: <https://doi.org/10.1093/pnasnexus/pgab004>.
- [147] Ross, J. S. et al. “Use of preventive care by the working poor in the United States”. *Preventive medicine* **44.3** (2007), pp. 254–259. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1810564/> (visited on 10/08/2021).
- [148] Rothberg, M. B. et al. “Using Highly Detailed Administrative Data to Predict Pneumonia Mortality”. en. *PLOS ONE* **9.1** (2014), e87382. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0087382> (visited on 10/08/2021).
- [149] Ruiz, M. et al. “Severe Community-acquired Pneumonia”. *American Journal of Respiratory and Critical Care Medicine* **160** (3 1999). doi: 10.1164/ajrccm.160.3.9901107, pp. 923–929. URL: <https://doi.org/10.1164/ajrccm.160.3.9901107>.
- [150] Ruuskanen, O. et al. “Viral pneumonia”. *The Lancet* **377** (9773 2011), pp. 1264–1275. URL: <https://www.sciencedirect.com/science/article/pii/S0140673610614596>.
- [151] Saha, E. & Ray, P. K. “Modelling and analysis of inventory management systems in healthcare: A review and reflections”. *Computers & Industrial Engineering* **137** (2019), p. 106051. URL: <https://www.sciencedirect.com/science/article/pii/S0360835219305108>.
- [152] Saito, T. & Rehmsmeier, M. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets”. *PLOS ONE* **10.3** (2015). Ed. by Brock, G., e0118432.
- [153] Schweitzer, M. & Cachon, G. “Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence”. *MANAGEMENT SCIENCE* **46.3** (2000), pp. 404–420.

- [154] Shapiro, A. et al. *Lectures on Stochastic Programming Modeling and Theory*. SIAM, 1949. Chap. Statistical Properties of Sample Average Approximation Estimators, pp. 155–170.
- [155] Shrestha, S. S. et al. “Estimation of Coronavirus Disease 2019 Hospitalization Costs From a Large Electronic Administrative Discharge Database, March 2020–July 2021”. *Open Forum Infectious Diseases* **8.12** (2021). ofab561. eprint: <https://academic.oup.com/ofid/article-pdf/8/12/ofab561/41826186/ofab561.pdf>. URL: <https://doi.org/10.1093/ofid/ofab561>.
- [156] Silver, E. et al. *Inventory Management and Production Planning and Scheduling*. Wiley, 1998.
- [157] Simonsen, L. et al. “The Impact of Influenza Epidemics on Hospitalizations”. *The Journal of Infectious Diseases* **181** (3 2000), pp. 831–837. URL: <https://doi.org/10.1086/315320>.
- [158] Spatz, E. S. et al. “Community factors and hospital wide readmission rates: Does context matter?” en. *PLOS ONE* **15.10** (2020), e0240222. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240222> (visited on 10/08/2021).
- [159] Swann, J. L. et al. “Return on investment of self-management education and home visits for children with asthma”. *JOURNAL OF ASTHMA* **58.3** (2021), pp. 360–369.
- [160] Telias, I. et al. “Is the Prone Position Helpful During Spontaneous Breathing in Patients With COVID-19?” *JAMA* **323.22** (2020), pp. 2265–2267. eprint: https://jamanetwork.com/journals/jama/articlepdf/2766290/jama_telias_2020_ed_200045.pdf. URL: <https://doi.org/10.1001/jama.2020.8539>.
- [161] Thomas, K. & Weiland, N. *Eli Lilly’s Antibody Treatment Gets Emergency F.D.A. Approval*. <https://www.nytimes.com/2020/11/09/health/covid-antibody-treatment-eli-lilly.html>. 9.
- [162] Times, T. N. Y. “Coronavirus in the U.S.: Latest Map and Case Count”. en-US. *The New York Times* (2020). URL: <https://www.nytimes.com/interactive/2021/us/covid-cases.html> (visited on 10/08/2021).
- [163] Timmins, A. 2022. URL: <https://newhampshirebulletin.com/2022/06/01/from-skepticism-to-insurance-denials-long-covid-patients-face-more-than-only-health-challenges/>.
- [164] Tolles, J. & Luong, T. “Modeling Epidemics With Compartmental Models”. *JAMA* **323.24** (2020), pp. 2515–2516. URL: <https://doi.org/10.1001/jama.2020.8420> (visited on 10/08/2021).
- [165] Tong, S. et al. “Trends in healthcare utilization and costs associated with pneumonia in the United States during 2008–2014”. *BMC Health Services Research* **18** (1 2018), p. 715. URL: <https://doi.org/10.1186/s12913-018-3529-4>.
- [166] Treanor, J. J. et al. “Effectiveness of Seasonal Influenza Vaccines in the United States During a Season With Circulation of All Three Vaccine Strains”. *CLINICAL INFECTIOUS DISEASES* **55.7** (2012), pp. 951–959.

- [167] Uematsu, H. et al. “Prediction of pneumonia hospitalization in adults using health checkup data”. *PLOS ONE* **12** (6 2017), e0180159-. URL: <https://doi.org/10.1371/journal.pone.0180159>.
- [168] Virtanen, P. et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. *Nature Methods* **17** (2020), pp. 261–272.
- [169] Walraven, C. van et al. “A Modification of the Elixhauser Comorbidity Measures Into a Point System for Hospital Death Using Administrative Data”. *Medical Care* (2009). URL: https://journals.lww.com/lww-medicalcare/Fulltext/2009/06000/A_Modification_of_the_Elixhauser_Comorbidity.4.aspx.
- [170] Wang, L. et al. “Predicting Risk of Hospitalization or Death Among Patients Receiving Primary Care in the Veterans Health Administration”. *Medical Care* **51** (4 2013). URL: https://journals.lww.com/lww-medicalcare/Fulltext/2013/04000/Predicting_Risk_of_Hospitalization_or_Death_Among.12.aspx.
- [171] Weycker, D. et al. “Attributable Cost of Adult Hospitalized Pneumonia Beyond the Acute Phase”. *PharmacoEconomics - Open* **5** (2 2021), pp. 275–284. URL: <https://doi.org/10.1007/s41669-020-00240-9>.
- [172] Wijck, P. van. “The economics of pre-crime interventions”. *European Journal of Law and Economics* **35** (3 2013), pp. 441–458. URL: <https://doi.org/10.1007/s10657-011-9229-8>.
- [173] Wolfe, R. et al. “Inequalities At Work and the Toll of COVID-19”. *Health Affairs* (2021). URL: <https://www.healthaffairs.org/doi/10.1377/hpb20210428.863621/full/>.
- [174] Wu, S. L. et al. “Substantial underestimation of SARS-CoV-2 infection in the United States”. *Nature Communications* **11.1** (2020), p. 4507.
- [175] Yang, T. et al. “Dynamic prediction of hospital admission with medical claim data”. *BMC MEDICAL INFORMATICS AND DECISION MAKING* **19.1** (2019). International Conference on Intelligent Biology and Medicine - Medical Informatics and Decision Making, Los Angeles, CA, JUN 10-12, 2018.
- [176] Yildirim, M. et al. “Estimating the impact of self-management education, influenza vaccines, nebulizers, and spacers on health utilization and expenditures for Medicaid-enrolled children with asthma”. *JOURNAL OF ASTHMA* **58.12** (2021), pp. 1637–1647.
- [177] Zaric, G. S. & Brandeau, M. L. “Optimal investment in a portfolio of HIV prevention programs”. *MEDICAL DECISION MAKING* **21** (5 2001), pp. 391–408.
- [178] Zhu, T. et al. “Data-Driven Models for Capacity Allocation of Inpatient Beds in a Chinese Public Hospital”. *Computational and mathematical methods in medicine* **2020** (2020), p. 8740457. URL: <https://pubmed.ncbi.nlm.nih.gov/32377227https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7199538/>.
- [179] Zhu, Y. et al. “Racial/Ethnic Disparities in the Prevalence of Diabetes and Prediabetes by BMI: Patient Outcomes Research To Advance Learning (PORTAL) Multisite Cohort of

Adults in the U.S.” en. *Diabetes Care* **42.12** (2019), pp. 2211–2219. URL: <https://care.diabetesjournals.org/content/42/12/2211> (visited on 10/08/2021).

- [180] Zimmer, C. et al. *Coronavirus Drug and Treatment Tracker*. 2022. URL: <https://www.nytimes.com/interactive/2020/science/coronavirus-drugs-treatments.html> (visited on 07/01/2022).

APPENDICES

APPENDIX A

OVERVIEW OF COMPUTATIONAL MODELS

Overview

In this thesis we explore various techniques for the development, calibration, and use of computational models with respect to health care, medical decision making, and intervention analyses. This section is intended to provide a high level summary of the fundamental assumptions, limitations, uses, and advantages/disadvantages of the different types of models that will be discussed in the various sections of this study.

Computational Models in Health Care

Different types of computational models can be used to answer questions relating to medical decision making. One major type of model used in the realm of healthcare analysis involves predictive modeling, which can be either prognostic or diagnostic. Prognostic models are concerned with predicting a response over time, such as the future development of a disease, whereas diagnostic models are developed to characterize the current health state of an individual including the identification of any present disease [52].

Model training and calibration, for either prognostic or diagnostic studies, can be either supervised, in which there is some applicable gold standard to the set, or unsupervised where we aim to identify existing patterns or other relationships within the data itself. Supervised learning can take the form of classification, or regression, whereas unsupervised learning involves either clustering, association, or dimensionality reduction [58]. The model development technique depends largely on the question being asked, whether a gold standard exists, and the types of data that are available

to the researcher.

A separate form of computational modeling that will be presented in this study involves modeling infectious diseases. Disease progression models rely on mathematical functions that govern disease spread as well as measures to indicate overall severity and capture outcomes of interest such as infection rates, hospitalizations, and mortality [53]. We couple standard, well-studied epidemiological models with training and calibration techniques to improve fit and model performance.

Evaluation and Reliability

A major concern, or criticism of computational models in healthcare, is how well we can rely on the output or information gained from each individual technique. To date, there exists no universal metric or performance measure to compare, or benchmark the multitude of model development strategies and subsequent analyses [99]. It is up to researchers and decision makers to determine, (i) which assumptions are necessary and worth making for model development, and (ii) which conclusions can be drawn from the insights gained from the data analysis.

We have seen during the COVID-19 pandemic that one metric, at least for the public perception of trust in computational models, has been their ability to match empirical data. This is particularly important for forecasting efforts, where rapid decision making is reliant on dependable output to gauge what could happen in the near future. Throughout the work presented in this study we place emphasis on the model's abilities to match empirical data when considering overall performance and significance. This metric is particularly important for healthcare data analyses due largely to the stakes associated with medical decision making in terms of public health including quality of life and substantial costs.

Aside from model validation, other metrics such as computational complexity, processing time, and the ability to integrate multiple sources and forms of data are other metrics that can help guide model development for clinical practice. Even a perfectly calibrated model can have limitations if it cannot be reproduced, scaled, and utilized. We consider the roles of each (i) reproducibility, (ii) scalability, and (iii) interpretability in model assessment as we present each of our findings.

Predictive Modeling

The first study we will consider in Chapter 3 involves the development of a predictive model to assess an individual's risk of an unplanned hospitalization. To capture this risk, we propose the use of supervised learning methods to develop binary classifiers given the data we were able to collect and transform. The study considers and evaluates the model performance on this binary outcome for three separate classifiers: (i) logistic regression classifier, (ii) random forest classifier, and (iii) extreme gradient boosting classifier. Here we present a high level overview of the different underlying assumptions, parameters, strengths, and weaknesses of each.

Most machine learning tools, packages, and functions used in our analysis are from the SciKit Learn Package in Python, with the addition of the XGBoost package.[40]

i. Logistic Regression Classifier

Logistic regression classifiers are type of parametric model that assume a linear relationship between the feature vectors and the predictive outcome. To avoid overfitting to the training data, we use a regularization term when training our model.

In general, linear regressions take the form:

$$h(X) = \beta_0 + \sum_{i=1}^p \beta_i X_i$$

, where X is a feature vector. [127]

For logistic regression, this linear representation of weights and regressors gets applied to a Sigmoid function, σ , where

$$\sigma(z) = \frac{1}{1 + e^{-z}} \in [0, 1]$$

Thus resulting in the following,

$$H_i(X) = \frac{1}{1 + e^{-\beta_0 + \sum_{j=1}^p \beta_j X_j}} \in [0, 1]$$

which becomes our probability of predicting a 1 for set of features, i .

[127]

Using ℓ_1 regularization results in a generation of a sparse weight matrix and allows us to view and interpret important regressors from our regression.

The ability to generate weights, and coefficients, allows us to actively interpret the regressors and their values in the context of the problem. Of course, the relationship between the predictors and the binary outcome need not be linear and thus we may be imposing an unneeded bias or limitation to our model by assuming this linearity.

ii. Random Forests

Random forest models utilize an ensemble of decision trees to form non-linear relationships between features and their outcomes. The general idea is that you form a forest of k trees, each of which are trained on independently gathered feature vectors, and utilize their outcomes to vote for the predicted outcome, which is determined via the most popular predictive value across all trees in the forest [130].

Each new tree is formed via a bootstrap sample of predictors, i.e., a random selection of features, taken with replacement, and trained. One of the most important aspects of the formation of each tree is how to determine the number of features, and the associated split, at each node in your tree. The default measure for random forests is the Gini impurity index, or probability of misclassification of predictive label [15, 110].

Compared to the outputs from logistic regression, random forests do not provide linear relationships or associations between the predictors and the outcome at hand. Instead, we can view what is known as feature importances, which tell us the importance of each predictor towards the predictive power of the model. It is related to the computation of the Gini impurity and is thought of the quantitative evaluation of how much each feature contributed to decreasing the impurity measure during training [110].

Random forests are quick to train, and can handle large datasets, however, the interpretability of the feature importances needs to be taken with caution. Feature importances can be biased and the relative evaluation of the importance value needs to be carefully considered in the context of every unique problem.

iii. **Extreme Gradient Boosting**

Similar to random forests, gradient boosting models are ensembles developed through a collection of classifiers such as decision trees. Similar to training for decision trees, tree ensemble methods seek to optimize a regularized objective function to decrease overall loss and model complexity. [90]

Compared to random forests, which randomly build trees independently of one another, gradient boosting builds concurrent trees dependent upon the current stage of the training process and seek to improve the trees that are already formed. Gradient boosting does not take a vote of a number of trees as the final predictor, rather it iteratively evaluates the performance of each tree during the training process to improve the loss function. [90]

The pitfalls of gradient boosting methods are the increased difficulty in fine tuning the model-specific parameters, and the inability to perform well when there is a lot of noisy data [90].

Performance Measures

In the proceeding chapters you will see multiple evaluation metrics being used to compare the performance of each of our models for the problem at hand. Here we give a brief overview of what these metrics look like and how one should interpret them for performance measure.

i. **Confusion Matrix**

We present results of both our training and testing data through a confusion matrix, which takes the following form for binary classification problems:

Ideally, we would like the confusion matrix to be diagonally dominant, with the largest number of predictions falling within the "True Negatives" and "True Positives" cells.

From here on throughout the text we will use the following interchangeably:

- TP = "True Positives"
- FP = "False Positives"

Table A.1 Visualizing the elements of a confusion matrix.[87]

		Predicted Label	
		0	1
Actual Label	0	"True Negatives"	"False Positives"
	1	"False Negatives"	"True Positives"

- TN = "True Negatives"
- FN = "False Negatives"

ii. **Accuracy** Accuracy is an overall measure of classification error. For binary problems, the accuracy measure is the rate at which we correctly classify each element in our set. In other words, if N is the number of elements in our set, then

$$\text{Accuracy} = \frac{\sum TP + TN}{N}$$

Although classification error is important to consider, it is important to note that accuracy scores are usually not considered the most relevant piece of information when considering model evaluation. Accuracy is a biased estimator of model performance, particularly in the scenarios where there is a large data imbalance between your minority and majority classes. If the data imbalance is large enough, any classifier training on accuracy alone will likely result in a model that classifies everything as part of the majority class and showcase deceptively high accuracy scores [112].

iii. **Precision, Recall, and F1 Score**

Precision, also known as **positive predictive value**, is a measure of the total relevant or true values of classifications that were predicted, or labeled, as a 1 from our model.[152]

Recall, also known as **sensitivity**, is the rate at which we are able to capture these events in the context of our problem. In other words, of all of the events we are trying to predict, at what rate were we successful?[152]

F1 Score, the harmonic mean between precision and recall. The F1 score is used simultaneously evaluate a model's ability to capture all events, and the how meaningful each predicted 1 is for our problem. [152]

We summarize the above mentioned measures as follows:

$$\text{Precision} = \frac{TP}{\sum TP + FP}$$

$$\text{Recall} = \frac{TP}{\sum TP + FN}$$

$$F1 = \frac{1}{2}(\text{Precision} + \text{Recall})$$

It is also common to view these measures in a precision-recall curve, which plots the precision along the y-axis and recall along the x-axis, for various probability thresholds to determine whether the model will predict a 1 for any given feature set. A no-skill classifier will have a curve that is a horizontal line about the number of the event-rate in the dataset. These plots are recommended for imbalanced data evaluation.

iv. Receiver-Operator Characteristic Curve (ROC) and Area Under the Curve (AUC)

The ROC curve is a measure of False Positive Rate versus True Positive Rate for various probability thresholds to predict whether a set of features will be associated with a label of 1 [52]. A no-skill classifier typically results in a straight line along $y = x$, whereas, a perfect classifier will have a high point close to (0,1) with a bend towards (1,1) [87].

The information in the ROC curve can be summarized using the the area under the curve. Here, a score is a measure between 0 and 1, where a measure of 1 is associated with a perfect classifier. The measure of the ROC-AUC can be thought of as a model's ability to correctly classify an event or label 1, over a non-event or label 0, as a 1 in our output [52, 152].

Disease Modeling

Epidemiological models are used to estimate the spread of infectious diseases and evaluate clinical outcomes associated with the disease. These models can be adapted to simulate disease progression values and behavioral changes [164] and used for forecasting and understanding overall disease burden. These models are widely used, particularly now during the COVID-19 pandemic where a plethora of models ranging from simple to highly detailed and sophisticated have been used to address public health concerns due to the rapidly changing and unknown trajectory of the disease spread. Compartmental models are used often due to their simplicity and computational effectiveness with relatively small numbers of disease parameters [164].

Various compartmental models have been employed to assess disease transmission, effects of public policies such as mask wearing and social distancing, impacts of vaccine prioritization, and many more [17, 71]. These models are governed by a set of disease equations, which can be solved for different time steps.

The simplest form of compartmental models is the SIR: Susceptible-Infected-Recovered model, but a more appropriate choice for COVID-19 spread is the SEIR model, which includes an exposed

period [109].

The equations for a general SEIR model are as follows [14]:

$$N = S + E + I + R$$

$$S' = -\beta SI$$

$$E' = \beta SI - \kappa E$$

$$I' = \kappa E - \alpha I$$

$$R' = \alpha I$$

where our mean exposed period is $1/\kappa$.

The basic reproductive number, \mathcal{R}_0 governs the severity of disease spread and we have that if $\mathcal{R}_0 < 1$ the disease dies out, whereas if $\mathcal{R}_0 > 1$ there is an epidemic. [14]

We expand upon a version of the SEIR model to assess past disease burden relating to COVID-19 in the early months of the pandemic.

APPENDIX B

APPENDIX TO CHAPTER 2

Variable Creation: A Collection of Codes, Data Sources, and Algorithms

i. PQI 11: Inpatient

Sources: Inpatient base claims

For the main flag in Chapter 2, we identify patients who experienced an inpatient visit with a primary diagnosis code of pneumonia, identified by one of the ICD-9 codes listed in Table B.1. We follow the AHRQ definition for Prevention Quality Indicator (PQI 11) Bacterial Pneumonia Admission Rate.

For the event label creation, we follow the algorithm outlined in [2], which excludes pneumonia admission rates for individuals with immunocompromised conditions or transfers from other facilities. The exclusions are listed below.

For our filtering variable, PQI 11 history in the last 12 months prior to each index date, we do not make the same exclusions, rather we identify patients who had a primary diagnoses code listed in Table B.1.

- Diagnosis code: ICD_DGNS_CD1 (primary)

Exclusions for numerator computation for PQI 11 taken from [2]:

- Transfers from a hospital (different facility)
- Transfers from a skilled nursing facility (SNF) or immediate care facility (ICF)
- Transfers from another health care facility Transfer codes are listed in Table B.2.
- Any-listed ICD-9-CM diagnosis codes for sickle cell anemia or HB-S disease listed in Table B.3.

Table B.1 ICD-9 codes for AHRQ definition of PQI-11 inpatient visit [2, 46]

ICD-9 Code	CMS Description
481	Pneumococcal pneumonia [Streptococcus pneumoniae pneumonia]
48242	Methicillin resistant pneumonia due to Staphylococcus aureus
4822	Pneumonia due to Hemophilus influenzae [H. influenzae]
48249	Other Staphylococcus pneumonia
48230	Pneumonia due to Streptococcus, unspecified
4829	Bacterial pneumonia, unspecified
48231	Pneumonia due to Streptococcus, group A
4830	Pneumonia due to mycoplasma pneumoniae
48232	Pneumonia due to Streptococcus, group B
4831	Pneumonia due to chlamydia
48239	Pneumonia due to other Streptococcus
4838	Pneumonia due to other specified organism
48240	Pneumonia due to Staphylococcus, unspecified
485	Bronchopneumonia, organism unspecified
48241	Methicillin susceptible pneumonia due to Staphylococcus aureus
486	Pneumonia, organism unspecified

- Any-listed ICD-9-CM diagnosis code or any ICD-9-CM procedure codes for immunocompromised state listed in Tables B.4-B.8.
- Missing gender, age, quarter, year, principal diagnosis, or county.

Table B.2 Admission codes for transfers [2]

Code	Description
2	Another hospital
3	Another facility, including long-term care
4	Transfer from a hospital
5	Transfer from a Skilled Nursing Facility (SNF) or Immediate Care Facility (ICF)
6	Transfer from another health care facility
If admission type is newborn, 5	Born inside this hospital
If admission type is newborn, 6	Born outside this hospital

Table B.3 Sick cell anemia or HB-S disease diagnosis codes. [2, 46]

Codes	Description
28241	Sickle-cell thalassemia without crisis
28242	Sickle-cell thalassemia with crisis
28260	Sickle-cell disease, unspecified
28261	Hb-SS disease without crisis
28262	Hb-SS disease with crisis
28263	Sickle-cell/Hb-C disease without crisis
28264	Sickle-cell/Hb-C disease with crisis
28268	Other sickle-cell disease without crisis
28269	Other sickle-cell disease with crisis

ii. **Prev_ILI: Previous Influenza like illness**

Table B.4 Immunocompromised State Diagnosis Codes Set 1. [2, 46]

Diagnosis Codes	42	Human immunodeficiency virus disease
	1363	Pneumocytosis
	1992	Malignant neoplasm associated with transplant organ
	23873	High grade myelodysplastic syndrome lesions
	23876	Myelofibrosis with myeloid metaplasia
	23877	Post-transplant lymphoproliferative disorder (PTLD)
	23879	Other lymphatic and hematopoietic tissues
	260	Kwashiorkor
	261	Nutritional marasmus
	262	Other severe protein-calorie malnutrition
	27900	Hypogammaglobulinemia, unspecified
	27901	Selective IgA immunodeficiency
	27902	Selective IgM immunodeficiency
	27903	Other selective immunoglobulin deficiencies
	27904	Congenital hypogammaglobulinemia
	27905	Immunodeficiency with increased IgM
	27906	Common variable immunodeficiency
	27909	Other deficiency of humoral immunity
	27910	Immunodeficiency with predominant T-cell defect, unspecified
	27911	Digeorge's syndrome
	27912	Wiskott-aldrich syndrome
	27913	Nezelof's syndrome
	27919	Other deficiency of cell-mediated immunity
	2792	Combined immunity deficiency
	2793	Unspecified immunity deficiency

Table B.5 Immunocompromised State Diagnosis Codes Set 2. [2, 46]

Diagnosis Codes	2794	Autoimmune disease, not elsewhere classified
	27941	Autoimmune lymphoproliferative syndrome
	27949	Autoimmune disease, not elsewhere classified
	27950	Graft-versus-host disease, unspecified
	27951	Acute graft-versus-host disease
	27952	Chronic graft-versus-host disease
	27953	Acute on chronic graft-versus-host disease
	2798	Other specified disorders involving the immune mechanism
	2799	Unspecified disorder of immune mechanism
	28409	Other constitutional aplastic anemia
	2840	Pancytopenia
	28411	Antineoplastic chemotherapy induced pancytopenia
	28412	Other drug induced pancytopenia
	28419	Other pancytopenia
	2880	Neutropenia not elsewhere specified
	28800	Neutropenia, unspecified
	28801	Congenital neutropenia
	28802	Cyclic neutropenia
	28803	Drug induced neutropenia
	28809	Other neutropenia
	2881	Functional disorders of polymorphonuclear neutrophils
	2882	Genetic anomalies of leukocytes
	2884	Hemophagocytic syndromes
	28850	Leukocytopenia, unspecified
	28851	Lymphocytopenia
	28859	Other decreased white blood cell count

Table B.6 Immunocompromised State Diagnosis Codes Set 3. [2, 46]

	28953	Neutropenic splenomegaly
	28983	Myelofibrosis
Diagnosis Codes	40301	Hypertensive chronic kidney disease, malignant, with chronic kidney disease stage V or end stage renal disease
	40311	Hypertensive chronic kidney disease, benign, with chronic kidney disease stage V or end stage renal disease
	40391	Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage V or end stage renal disease
	40402	Hypertensive heart and chronic kidney disease, malignant, without heart failure and with chronic kidney disease stage V or end stage renal disease
	40403	Hypertensive heart and chronic kidney disease, malignant, with heart failure and with chronic kidney disease stage V or end stage renal disease
	40412	Hypertensive heart and chronic kidney disease, benign, without heart failure and with chronic kidney disease stage V or end stage renal disease
	40413	Hypertensive heart and chronic kidney disease, benign, with heart failure and chronic kidney disease stage V or end stage renal disease
	40492	Hypertensive heart and chronic kidney disease, unspecified, without heart failure and with chronic kidney disease stage V or end stage renal disease
	40493	Hypertensive heart and chronic kidney disease, unspecified, with heart failure and chronic kidney disease stage V or end stage renal disease
		5793

Table B.7 Immunocompromised State Diagnosis Codes Set 4. [2, 46]

	585	Chronic renal failure
	5855	Chronic kidney disease, Stage V
	5856	End stage renal disease
Diagnosis Codes	9968	Complications of transplanted organ
	99680	Complications of transplanted organ, unspecified
	99681	Complications of transplanted kidney
	99682	Complications of transplanted liver
	99683	Complications of transplanted heart
	99684	Complications of transplanted lung
	99685	Complications of transplanted bone marrow
	99686	Complications of transplanted pancreas
	99687	Complications of transplanted intestine
	99688	Stem cell transplant complications
	99689	Complications of other specified transplanted organ
	V420	Kidney replaced by transplant
	V421	Heart replaced by transplant
	V426	Lung replaced by transplant
	V427	Liver replaced by transplant
	V428	Other specified organ or tissue replaced by transplant
	V4281	Bone marrow replaced by transplant
	V4282	Peripheral stem cells replaced by transplant
	V4283	Pancreas replaced by transplant
	V4284	Organ or tissue replaced by transplant, intestines
V4289	Other specified organ or tissue replaced by transplant	
V51	Renal dialysis status	
V511	renal dialysis status	
V560	Encounter for extracorporeal dialysis	
V561	Fitting and adjustment of extracorporeal dialysis catheter	
V562	Fitting and adjustment of peritoneal dialysis catheter	

Table B.8 Immunocompromised State Procedure Codes. [2, 46]

Procedure Codes	0018	Infusion of immunosuppressive antibody therapy
	335	Lung transplantation
	3350	Lung transplantation, not otherwise specified
	3351	Unilateral lung transplantation
	3352	Bilateral lung transplantation
	336	Combined heart-lung transplantation
	3751	Heart transplantation
	410	Operations on bone marrow and spleen
	4100	Bone marrow transplant, not otherwise specified
	4101	Autologous bone marrow transplant without purging
	4102	Allogeneic bone marrow transplant with purging
	4103	Allogeneic bone marrow transplant without purging
	4104	Autologous hematopoietic stem cell transplant without purging
	4105	Allogeneic hematopoietic stem cell transplant without purging
	4106	Cord blood stem cell transplant
	4107	Autologous hematopoietic stem cell transplant with purging
	4108	Allogeneic hematopoietic stem cell transplant with purging
	4109	Autologous bone marrow transplant with purging
	4697	Transplant of intestine
	5051	Auxiliary liver transplant
	5059	Other transplant of liver
	5280	Pancreatic transplant, not otherwise specified
	5281	Reimplantation of pancreatic tissue
	5282	Homotransplant of pancreas
	5283	Heterotransplant of pancreas
	5285	Allotransplantation of cells of Islets of Langerhans
	5286	Transplantation of cells of Islets of Langerhans, not otherwise specified
	5569	Other kidney transplantation

Sources: Carrier base claims, Outpatient base claims, Inpatient (emergency department only) base claims, skilled nursing facility base claims
 Diagnosis code trailer: ICD_DGNS_CD_X, X = 1, 2, ..., 12

- codesA = {4871, 4878, 07999}
- codesB = {78600}
- codesC = {V462}
- codesD = {7862}
- codesE = {0796, 4659, 4660, 46611, 46619, 4789}

Table B.9 List of ICD-9 Codes for influenza like illnesses.

ICD-9 Code	CMS Description
4871	Influenza with other respiratory manifestations
4878	Influenza with other manifestations
7999	Unspecified viral infection
78600	Respiratory abnormality, unspecified
V462	Other dependence on machines, supplemental oxygen
7862	Cough
796	Respiratory syncytial virus (RSV)
4659	Acute upper respiratory infections of unspecified site
4660	Acute bronchitis
46611	Acute bronchiolitis due to respiratory syncytial virus (RSV)
46619	Acute bronchiolitis due to other infectious organisms
4789	Other and unspecified diseases of upper respiratory tract

Previous influenza-like-illness (ILI) is any diagnosis code from the individual's diagnosis code trailer using the above defined data sources and code groupings (A-E) captured in the following manner: Binary flag of one (1) for previous ILI if a beneficiary meets one of the following criteria:

- (a) dgns_cd_i in codesA or;
- (b) dgns_cd_i in codesB and dgns_cd_j (i≠j) in codesC or;

(c) dgns_cd_i in codesB and dgns_cd_j(i≠j) in codesD or;

(d) dgns_cd_i in codesE

We also specify that any claims coming from the inpatient file must be coming from the Emergency Room. We capture this by comparing claim numbers from the inpatient base claims and the inpatient revenue centers and ensure that each previous ILI flag from an inpatient file is matched with a revenue center code from the following set: Emergency room alias::rev_cntr codes:

```
{0981 : professional fees emergency room,  
 0450 : emergency room – general classification,  
 0451 : emergency room – EMTALA emergency room screening services,  
 0452 : emergency room – ER beyond EMTALA screening,  
 0453 : reserved emergency room,  
 0454 : reserved emergency room,  
 0455 : reserved emergency room,  
 0456 : emergency room – urgent care,  
 0457 : reserved emergency room,  
 0458 : reserved emergency room,  
 0459 : emergency room - other}
```

ref: <https://resdac.org/cms-data/variables/revenue-center-code-ffs>, <https://static.cigna.com/assets/chcp/pdf/code-list-requiring-cpt-and-hcps-codes-for-outpatient-facility-claims.pdf> [144]

iii. **Prev_ILI_with_XRAY: Previous Influenza like illness with chest XRAY**

Sources: Carrier base claims, Outpatient base claims, Inpatient (emergency department only) base claims, skilled nursing facility base claims, Carrier revenue, Outpatient revenue, Inpatient revenue, Skilled nursing facility revenue

- Diagnosis code trailer: ICD_DGNS_CDX, X = 1, 2, ..., 12
- Procedure code trailer: ICD_PRCDR_CDX, X = 1, 2, ..., 25
- HCPCS_CD

To compute the previous influenza like illness with chest XRAY we begin with the same definition of a previous influenza-like-illness defined in (1. Prev_ILI). Similar to how we extract all

inpatient visits that are associated with an emergency room visit, we perform the same cross-check on claim numbers associated with a flag for a previous ILI and those with a procedure code for a chest XRAY.

The procedure codes for chest XRAYs:

Table B.10 ICD-9 and CPT/HCPCS codes for chest x-rays.

CPT/HCPCS	71010-71035	Chest x-ray
ICD-9	8744	Routine chest x-ray, so described

iv. Smoking

Sources: Carrier base claims, Outpatient base claims, Inpatient (emergency department only) base claims, skilled nursing facility base claims, Carrier revenue, Outpatient revenue, Inpatient revenue, Skilled nursing facility revenue

- Diagnosis code trailer: ICD_DGNS_CDX, X = 1, 2, ..., 12
- HCPCS_CD

This variable encompasses whether an individual has a known or reported history of tobacco use. We define this variable to be a binary indicator if an individual has any of the below mentioned codes in any of their claims from the [Smoking] sources listed above. Since this variable is more about the patient’s history, we look as far back as 2008 to make the determination.

ref: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4826837/> [59]

v. Flu_vaccine: Administration of influenza vaccine

Sources: Carrier base claims, Outpatient base claims, Inpatient (emergency department only) base claims, skilled nursing facility base claims, Carrier revenue, Outpatient revenue, Inpatient revenue, Skilled nursing facility revenue

- Procedure code trailer: ICD_PRCR_CD, X = 1, 2, ..., 25
- HCPCS_CD

To compute whether an individual received their influenza vaccination during the current flu season of the respective index date we consider those who received their influenza vaccine July 2010-June2011 [season1] and July2011-December2011 [season2]. Here, an individual will be marked as having their flu vaccine for January-June if they had a vaccine season1 and July-December if they had a vaccine season2.

ref: <https://www.cms.gov/medicare/preventive-services/flu-shot-coding> [44]

Table B.11 List of ICD-9 and CPT/HCPCS codes to identify someone with a medical history of smoking.

	99406	
	99407	
	G0436	Smoking counseling visit
	G0437	
	G9016	
CPT/HCPCS	S9453	Smoking cessation classes
	S4995	Smoking cessation gum
	9276	Disposable censor
	9458	Tobacco cessation intervention
	1034F	Patient history
	4004F	Tobacco screening
	4001F	Therapeutic, Preventive, or Other Interventions
ICD-9	3051	Tobacco use disorder
	64900	Tobacco use disorder complicating pregnancy, childbirth, or the puerperium, unspecified as to episode of care or not applicable
	64901	Tobacco use disorder complicating pregnancy, childbirth, or the puerperium, delivered, with or without mention of antepartum condition
	64902	Tobacco use disorder complicating pregnancy, childbirth, or the puerperium, delivered, with mention of postpartum complication
	64903	Tobacco use disorder complicating pregnancy, childbirth, or the puerperium, antepartum condition or complication
	64904	Tobacco use disorder complicating pregnancy, childbirth, or the puerperium, postpartum condition or complication
	98984	Toxic effect of tobacco
	V1582	Personal history of tobacco use

Table B.12 List of ICD-9 and CPT codes for the administration of an influenza vaccine.

	G0008	Administration of influenza virus vaccine
CPT/HCPCS	90630	
	90653-90756	Influenza virus vaccine[*]
	Q2034-Q2039	
ICD-9	9952	Prophylactic vaccination against influenza

vi. **Pneumococcal_vaccine: Administration of pneumococcal vaccine**

Sources: Carrier revenue, Outpatient revenue, Inpatient revenue, Skilled nursing facility revenue

- HCPCS_CD

The pneumococcal vaccine has different recommendations from the CDC, compared to the influenza vaccine ref: <https://www.cdc.gov/vaccines/vpd/pneumo/hcp/who-when-to-vaccinate.html>. [134] Since we only have data from 2008-2011 we consider an individual to have received an pneumococcal vaccine (and thus receiving preventive care for pneumococcal disease) if they have any of the below CPT or HCPCS codes anywhere in their claims data from 2008 up until the index date for our prediction window.

Table B.13 CPT and HCPCS codes for the administration of the pneumococcal vaccines.

CPT/HCPCS	G0009	Administration of the pneumococcal vaccine
	90670	Pneumococcal conjugate vaccine, 13 valent, for intramuscular use
	90732	Pneumococcal polysaccharide vaccine, 23-valent

vii. **Attention-Deficit Disorders**

Sources: Carrier base claims, Outpatient base claims, Inpatient base claims, skilled nursing facility base claims

- Diagnosis code trailer: ICD_DGNS_CDX, X = 1, 2, ..., 12

Table B.14 ICD-9 codes for attention deficit and conduct disorders.

	ICD-9 Codes	CCS-Category-Description
ICD-9	31200-3149	'Attention-deficit/conduct/disruptive beha'

viii. **Bronchitis**

Sources: Carrier base claims, Outpatient base claims, Inpatient base claims, skilled nursing facility base claims

- Diagnosis code trailer: ICD_DGNS_CDX, X = 1, 2, ..., 12

Table B.15 ICD-9 codes for bronchitis.

	ICD-9 Codes	CCS-Category-Description
ICD-9	4660-46619	'Bronchitis'

ix. Pneumonia

Sources: Carrier base claims, Outpatient base claims, Inpatient base claims, skilled nursing facility base claims

- Diagnosis code trailer: ICD_DGNS_CDX, X = 1, 2, ..., 12

Table B.16 ICD-9 codes for pneumonia.

	ICD-9 Codes	CCS-Category-Description
	322	'Pneumonia'
	0203-0830	
	1124-11595	
ICD-9	1304	
	1363	
	4800-5171	

x. Fatigue

Sources: Carrier base claims, Outpatient base claims, Inpatient base claims, skilled nursing facility base claims

- Diagnosis code trailer: ICD_DGNS_CDX, X = 1, 2, ..., 12

Table B.17 ICD-9 codes for fatigue.

	ICD-9 Codes	CCS-Category-Description
ICD-9	78071-78079	'Fatigue'

xi. **Exam_eval**

Sources: Carrier base claims, Outpatient base claims, Inpatient base claims, skilled nursing facility base claims

- Diagnosis code trailer: ICD_DGNS_CDX, X = 1, 2, ..., 12

Table B.18 ICD-9 codes for examinations, evaluations, and screenings.

	ICD-9 Codes	CCS-Category-Description
ICD-9	V290-V299	'Exam/eval'
	V6801-V6809	
	V700-V709	
	V718-V729	

xii. **Maint_chem_r: Chemotherapy, Radiotherapy, Immunotherapy**

Sources: Carrier base claims, Outpatient base claims, Inpatient base claims, skilled nursing facility base claims

- Diagnosis code trailer: ICD_DGNS_CDX, X = 1, 2, ..., 12

Table B.19 ICD-9 codes for chemotherapy, immunotherapy, and radiotherapy.

	ICD-9 Codes	CCS-Category-Description
ICD-9	V580-V672	'Maint chem/r'

xiii. **Cystic_fibro: Cystic Fibrosis**

Sources: Carrier base claims, Outpatient base claims, Inpatient base claims, skilled nursing facility base claims

- Diagnosis code trailer: ICD_DGNS_CD, X = 1, 2, ..., 12

Table B.20 ICD-9 codes for cystic fibrosis.

	ICD-9 Codes	CCS-Category-Description
ICD-9	27700-27709	'Cystic fibro'

xiv. **Lung_external: Lung disorders or other complications due to external factors**

Sources: Carrier base claims, Outpatient base claims, Inpatient base claims, skilled nursing facility base claims

- Diagnosis code trailer: ICD_DGNS_CD, X = 1, 2, ..., 12

Table B.21 ICD-9 codes for lung disorders and complications due to external agents.

	ICD-9 Codes	CCS-Category-Description
ICD-9	4950-5089	'Lung external'

xv. **Fx_skull_fac: Skull injury or fracture**

Sources: Carrier base claims, Outpatient base claims, Inpatient base claims, skilled nursing facility base claims

- Diagnosis code trailer: ICD_DGNS_CD, X = 1, 2, ..., 12

Table B.22 ICD-9 codes for skull injuries or fractures.

	ICD-9 Codes	CCS-Category-Description
ICD-9	80000-9050	'Fx skull fac'

xvi. **Obesity**

Sources: Carrier base claims, Outpatient base claims, Inpatient base claims, skilled nursing facility base claims

- Diagnosis code trailer: ICD_DGNS_CDX, X = 1, 2, ..., 12

Table B.23 ICD-9 codes for a medical history of obesity.

	ICD-9 Codes	Description
ICD-9	27800	Obesity, unspecified
	27801	Morbid obesity

xvii. **Resp_symptoms: Respiratory symptoms**

Sources: Carrier base claims, Outpatient base claims, Inpatient base claims, skilled nursing facility base claims

- Diagnosis code trailer: ICD_DGNS_CDX, X = 1, 2, ..., 12

ref: <https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes> [86][**]

Table B.24 ICD-9 codes for previous illnesses of the respiratory tract^[**].

ICD-9 Code	CMS Description
78600	Respiratory abnormality, unspecified
78601	Hyperventilation
78602	Orthopnea
78603	Apnea
78604	Cheyne-Stokes respiration
78605	Shortness of breath
78606	Tachypnea
78607	Wheezing
78609	Other respiratory abnormalities
7861	Stridor
7862	Cough
78630	Hemoptysis, unspecified
78631	Acute idiopathic pulmonary hemorrhage in infants [AIPHI]
78639	Other hemoptysis
7864	Abnormal sputum
78650	Chest pain, unspecified
78651	Precordial pain
78652	Painful respiration
78659	Other chest pain
7866	Swelling, mass, or lump in chest
7867	Abnormal chest sounds
7868	Hiccough
7869	Other symptoms involving respiratory system and chest

xviii. **Specialty visits: Oncologist, cardiologist**

Sources: Carrier revenue

- Provider_speciality (alias::PRVDR_SPCLTY)

Table B.25 Provider speciality codes for cardiology and oncology.

	Code	CMS_PRVDR_SPCLTY_TB
Cardiologist	06, C3, C7	Cardiology, Interventional cardiology, advanced heart failure and transplant cardiology
Oncologist	90-92	Medical oncology, surgical oncology, radiation oncology

xix. **Most recent Inpatient visit type: Acute care, critical access, psychiatric and Ever been in Critical Access Facility and Ever been in Psychiatric facility**

Sources: Inpatient base claims, CMS Hospital Compare

- Provider ID

CMS Hospital Compare provides a list of all inpatient facilities and groupings by acute care, critical access, and psychiatric facilities. To compute this variable we first pull all claims from the inpatient files from 2009-2011. Then, on a rolling basis for 2011, we match beneficiaries and their respective patient row index month with their most recent inpatient provider ID up until the month prior to each index date. We lookup this provider ID in the CMS Hospital Compare file and note the type of facility for each visit.

We also keep track of two variables, `ever_been_in_critical_access` and `ever_been_in_psychiatric_facility`. These variables are proxies for things such as access to care and mental health conditions. CMS declares a facility to be critical access if it is:

- Located in a rural area;
- Located greater than 35 miles from the nearest inpatient facility (generally) or 15 miles for areas in the mountains or only accessible via secondary roads;
- Comprised of at most 25 inpatient beds

- Average length of stay of 96 hours
- Service 24/7 emergency care.

[55]

xx. **Total costs**

Sources: Outpatient base claims, Inpatient base claims, Skilled nursing facility base claims, Carrier base claims, Home health base claims, and Durable medical equipment base claims

To compute these variables for each outpatient, inpatient, snf, carrier, homehealth, dme, and overall total cost we consider the sum of the alias::CLM_TOT_CHRG_AMT from each of these files for each patient row. We compute the sum of the charges from each facility over the previous 1, 3, and 6 months.

xxi. **Recent emergency room visits**

Sources: Inpatient revenue centers, outpatient revenue centers

Using the same alias::rev_cntr codes defined and outlined in (1.Previous influenza like illness), we track whether a beneficiary had a recent ER visit in the last 1, 3, or 6 months.

xxii. **Number of claims by facility type**

Sources: Outpatient base claims, Inpatient base claims, Skilled nursing facility base claims, Carrier base claims, Home health base claims, and Durable medical equipment base claims

To compute these variables for each outpatient, inpatient, snf, carrier, homehealth, dme, and overall total cost we consider the sum of the total number of claims from each of these files for each patient row. We compute the sum of the charges from each facility over the previous 1, 3, and 6 months.

xxiii. **Nursing home visits**

Sources: Carrier revenue, Inpatient revenue, Outpatient revenue, Skilled nursing facility revenue

- HCPCS_CD

Table B.26 ICD-9, CPT, and HCPCS codes for nursing home visits.

CPT/HCPCS	99304-99306	Initial nursing facility assessments
	99307-99310	Subsequent nursing facility assessments

To compute the most recent information about nursing home status, we consider claims from the above sources that are marked with an hcpcs code belonging to either an initial nursing facility assessment or any subsequent nursing facility assessments. We consider whether a beneficiary has any of these codes present in their data in the previous 1, 3, or 6 months prior to each index date. ref: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3178883/> [96]

xxiv. **Frailty**

Sources: Carrier base claims, Inpatient base claims, Outpatient base claims, Skilled nursing facility base claims, Home health base claims, Durable medical equipment base claims, Carrier revenue, Inpatient revenue, Outpatient revenue, Skilled nursing facility revenue, Home health revenue, Durable medical equipment revenue

The paper used as a reference for this variable lists variables that results in at least 0.010-value increase in claims-based frailty scores according to results from their lasso regression model. We chose all values from their supplementary file associated with an absolute change in in frailty score of 0.010, which include those with a decrease in frailty score. We felt these were important to include since they encompass a person's willingness to receive care or take preventive measures such as receiving screening.

We compute our frailty index on a rolling basis from 2010-2011 and update the score as new information is presented in the individual's claims data. That is, the frailty index is a dynamic variable that changes any time a new code in the below table is found. When we find a new code, the associated coefficient from the supporting reference is added to the individual's frailty score. If the code is found in 2010, it is present in the individual's frailty score for all months in 2011 and if found in January 2011, it is applied to all months February –December 2011 and so on.

A list of codes, descriptions, and associated frailty coefficients is outline in the following table. ref:<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6001883/> [93]

Final List of Variables Included in Model Training

For model training we utilize 195 variables: 155 Individual, 24 Community, and 16 Provider variables listed as their short descriptions below. For each of the variables, if it is dynamic and considers a certain amount of history the number of months of data considered is denoted (Xmth) meaning that this variable considers the previous X months of data prior to the index date.

155 Individual Variables

Interaction inpatient elixhauser score and ever in inpatient, Interaction hospital overall rating 5 and ever in inpatient, Interaction hospital overall rating 4 and ever in inpatient, Interaction hospital overall rating 3 and ever in inpatient, Interaction hospital overall rating 2 and ever in inpatient,

Table B.27 ICD-9 codes and related frailty coefficients for diagnosis components of frailty computation.

Code Type	Codes	Description	Frailty Coefficient
ICD-9	030-041	Other bacterial diseases	0.01188
ICD-9	250-259	Diseases of other endocrine glands	0.01078
ICD-9	290-294	Organic psychotic conditions	0.04745
ICD-9	295-299	Other psychoses	0.02112
ICD-9	300-316	Neurotic disorders, personality disorders, and other nonpsychotic mental disorders	0.01353
ICD-9	330-338	Hereditary and degenerative diseases of the central nervous system	0.04031
ICD-9	401-405	Hypertensive disease	0.0174
ICD-9	410-414	Ischemic heart disease	0.0192
ICD-9	420-429	Other forms of heart disease	0.01989
ICD-9	430-438	Cerebrovascular disease	0.01552
ICD-9	451-459	Diseases of veins and lymphatics, and other disease of circulatory system	0.01234
ICD-9	480-487	Pneumonia and influenza	0.01185
ICD-9	490-496	Chronic obstructive pulmonary disease and allied conditions	0.01295
ICD-9	580-589	Nephritis, nephrotic syndrome, and nephrosis	0.01047
ICD-9	590-599	Other diseases of the urinary system	0.01062
ICD-9	600-608	Diseases of male genital organs	-0.02131
ICD-9	710-719	Arthropathies and related disorders	0.01408
ICD-9	797-799	Ill-defined and unknown causes of morbidity and mortality	0.01119
ICD-9	890-897	Open wound of lower limb	0.02042
ICD-9	920-924	Contusion with intact skin surface	0.01074
ICD-9	V70-V82	Persons without reported diagnosis encountered during examination and investigation of individuals and populations	-0.01208

Table B.28 CPT and HCPCS codes and related frailty coefficients for diagnosis components of frailty computation.

Code Type	Codes	Description	Frailty Coefficient
CPT	99308	Nursing facility care - subsequent	0.01389
CPT	00500-00580	Anesthesia-intrathoracic	-0.01568
CPT	00800-00882	Anesthesia-lower abdomen	-0.01053
CPT	01320-01444	Anesthesia-knee and popliteal area	-0.01215
CPT	88104-8819	Cytopathology	-0.01082
CPT	98940-98943	Chiropractic manipulative treatment	-0.01426
HCPCS	A0021-A0999	Transportation services including ambulance	0.01022
HCPCS	A4244-A4290	Other supplies including diabetes supplies and contraceptives	0.0236
HCPCS	A5500-A5513	Diabetic footwear	0.02375
HCPCS	E0100-E0159	Walking aids and attachments	0.02821
HCPCS	E0250-E0373	Hospital beds and associated supplies	0.08649
HCPCS	E1353-E1406	Accessories for oxygen delivery devices	0.02702
HCPCS	G0101-G0124	Screening examinations and disease management training	-0.01136
HCPCS	K0001-K0462	Wheelchairs, components, and accessories	0.07802

Interaction hospital overall rating 1 and ever in inpatient, Race, Gender, Age, Pqi (1mth), CLAIMS (6mth), Obesity, ER (6mth), ER (3mth), ER (1mth), Prev ILI XRAY (6mth), Prev ILI XRAY (3mth), Prev ILI XRAY (1mth), Prev ILI (6mth), Prev ILI(3mth), Prev ILI (1mth), NH (1mth), EVER inpatient, Cardiology (1mth), Cardiology (1mth), Oncology (1mth), Oncology (1mth), Oncology (1mth), INP (6mth), INP (6mth), INP (6mth), CARRIER (6mth), CARRIER (6mth), CARRIER (6mth), CLAIMS (6mth), CLAIMS (6mth), CLAIMS (6mth), Fx skull fracture (1mth), Lung (1mth), Cystic (1mth), Maintenance chemotherapy (1mth), BRONCHITIS (1mth), Exam (1mth), FATIGUE (1mth), PNEUMONIA (1mth), Attention disorder (1mth), PQI 11 (2010), NH (6mth), Respiratory symptoms (1mth), pneumococcal vaccine, flu vaccine, Allergy (1mth), Neurology (1mth), Endocrinomology (1mth), Low income, Post transplant, Smoker, Rural, Homehealth now, High socioeconomic insurance, SNF now, Still in NH, Schizophrenia and other psychotic disorders, Personality disorders, Developmental disorders, Delirium, dementia, and amnestic and other cognitive disorders, Attention-deficit, conduct, and disruptive behavior disorders, Rheumatoid arthritis and related disease, Infective arthritis and osteomyelitis, Chronic ulcer of the skin, Chronic kidney disease, Asthma, Chronic obstructive pulmonary disease and bronchiectasis, Confestive heart failure; nonhypertensive, Cardiac arrest and ventricular fibrillation, Cardiac dysrhythmias, Conduction disorders, Other and ill-defined heart disease, Pulmonary heart disease, Nonspecific chest pain, Coronary atherosclerosis and other heart disease, Acute myocardial infarction, Hypertension with complications and secondary hypertension, Essential hypertension, Peri-; endo-; and myocarditis; cardiomyopathy, Heart valve disorders, Coma; stupor; and brain damage, Headache; including migraine, Epilepsy; convulsions, Paralysis, Other hereditary and degenerative nervous system conditions, Multiple sclerosis, Parkinson's disease, Cystic fibrosis, Diabetes mellitus with complications, Diabetes mellitus without complication, Cancer; other and unspecified primary, Cancer of thyroid, Cancer of brain and nervous system, Cancer of other urinary organs, Cancer of kidney and renal pelvis, Cancer of bladder, Cancer of other male genital organs, Cancer of testis, Cancer of prostate, Cancer of other female genital organs, Cancer of ovary, Cancer of cervix, Cancer of uterus, Cancer of breast, Other non-epithelial cancer of skin, Melanomas of skin, Cancer of bone and connective tissue, Cancer; other respiratory and intrathoracic, Cancer of bronchus; lung, Cancer of other GI organs; petioneum, Cancer of pancreas, Cancer of liver and intraheptic bile duct, Cancer of recum and anus, Cancer of colon, Cancer of stomach, Cancer of esophagus, Cancer of head and neck, Cancer diagnoses, Neurological disorders, Elixhauser score snf, Cognitive disorders, Cost inpatient (1mth), Cost snf (1mth), Cost carrier (1mth), DME (1mth), ER (1mth), SNF (12), Outpatient cost (3mth), Outpatient (1mth), ER (3mth), HH (1mth), SNF cost (6mth), Elixhauser Van Waren Score Inp, SNF (1mth), Number of heart conditions, Primary Care (1mth), Carrier (12 mth), Cost inpatient (6mth), Outpatient (12mth), HH (6mth), DME (1mth), Hospital visits (1mth), Frailty, Carrier (1mth), Inpatient (12mth), Cost carrier (6mth), Homehealth (6mth), Hospital visits (3mth), Sum of CCS Comorbid conditions

24 Community Variables

Household income, Percentage 65+ Poverty, Percentage of Medicare Enrollees, Percent Smokers, Percent Illiterate, Percent Rural, Population per primary care provider, Percent Asian Population, Percent African American, Unhealthy Ozone Days, Hospital bed ratio, Air pollution particular matter, Percent of weeks flu is widespread, Percent American Indian/Alaskan Native/Native Hawaiian/Other pacific islander, Intensity of flu, Activity spread of flu, PQI05 Region Score, PQI07 Region Score, PQI08 Region Score, PQI11 Region Score, PQI12 Region Score, PQI14 Region Score, PQI16 Region Score, Mean activity level flu (1mth)

16 Provider Variables

CABG death rate, Postoperative Respiratory Failure Rate, C.diff llaboratory identified events, CAUTI rate, CLABSI rate, Colon Surgery Infection rate, MRSA rate, Risk-standardized potentially preventable hospital readmissions , 30-day rates of readmission and excess dates in acute care due to pneumonia, Unplanned hospital visits due to heart failure, Pressure ulcer rate, Unplanned hospital visits due to heart attack, Percent of SNF residents who experience one or more falls, Surgical site infection from abdominal hysterectomy, Count of outpatient procedures, Hospital overall rating

Post Prediction Analysis

B.0.1 Variable Coefficients and Feature Importance

Variable	Des	Long Des	I/C /P	Mean LR Coefficient	Nonzero LR	Mean XGB FI	Mean RF FI
age_1	<65		I	ref			
age_2	65-69		I	0.11332 9877	5	0.00660 2149	0.00153 8948
age_3	70-74		I	0.19085 6663	4	0.00499 5759	0.00105 4538
age_4	75-79		I	0.39808 7017	5	0.00327 9587	0.00077 8684
age_5	80-84		I	0.56098 4094	5	0.00392 9764	0.00075 1642
age_6	>84		I	0.80656 7809	5	0.00611 306	0.00450 8876
Aggregate count of outpatient procedures	Count of outpatient procedures	Sum of outpatient procedures including gastrointestinal, eye, nervous system, musculoskeletal, skin, genitourinary, cardiovascular, respiratory	P	- 0.59946 277	5	0.00326 9106	0.03336 3669
AirPollutionOzoneDays	Unhealthy Ozone Days	Annual number of unhealthy air quality days due to ozone from county health ranking	C	0.53807 2465	5	0.00306 8327	0.00509 566
AirPollutionParticulateMatterDays	Air pollution particulate matter	Annual number of unhealthy air quality days due to fine particulate matter	C	0.19708 3652	5	0.00315 9322	0.00377 2248
allergy_prev_one_month	Allergy (1mth)		I	0.03342 4318	5	0.00133 7003	6.63E-05
Attention_deficit_conduct_disruptive_behavior_prev_month	Attention disorder (1mth)		I	0.87361 2528	3	0	2.06E-05
Bronchitis_prev_month	BRONCHITIS (1mth)		I	0.05356 791	4	0	2.81E-05
cancer	Cancer diagnoses	Grouped 26 cancer diagnoses through summation	I	0	0	0.00321 6537	0.00116 3073

cardiology_visit_prev_3_month	Cardiology (1mth)		I	- 0.13058 3189	5	0.00292 2545	0.00198 1068
cardiology_visit_prev_6_month	Cardiology (1mth)		I	- 0.12069 0867	5	0.00373 8137	0.00149 2825
carrier_cost_1mth	Cost carrier (1mth)		I	- 1.39346 5556	3	0.00267 9561	0.01531 8542
carrier_cost_6mth	Cost carrier (6mth)	cumulative cost from carrier claims in the last 6 months	I	0.05986 4742	3	0.00322 6858	0.01536 6843
carrier_prev_1_month	CARRIER (6mth)		I	- 0.96627 4095	5	0	0.00602 522
carrier_prev_3_month	CARRIER (6mth)		I	- 1.31696 9977	5	0.00489 914	0.00378 5492
carrier_prev_6_month	CARRIER (6mth)		I	0.80417 9556	5	0.00323 2714	0.00392 6123
ccs_category_100	Acute myocardial infarction		I	- 0.10117 8319	5	0.00467 1941	0.00039 1516
ccs_category_101	Coronary atherosclerosis and other heart disease		I	- 0.05927 843	5	0.00238 437	0.00096 9448
ccs_category_102	Nonspecific chest pain		I	0.03814 1444	5	0.00382 8146	0.00067 5351
ccs_category_103	Pulmonary heart disease		I	- 0.04013 6246	5	0.00381 0239	0.00067 5547
ccs_category_104	Other and ill-defined heart disease		I	- 0.03130 5194	5	0.00550 6682	0.00042 3313
ccs_category_105	Conduction disorders		I	- 0.06131 6974	5	0.00415 5671	0.00060 5615
ccs_category_106	Cardiac dysrhythmias		I	- 0.08955 0386	5	0.00226 9534	0.00145 5048
ccs_category_107	Cardiac arrest and ventricular fibrillation		I	- 0.06916 3991	5	0.01263 8027	9.96E- 05
ccs_category_108	Confestive heart failure; nonhypertensive		I	- 0.00228 683	5	0.00272 3598	0.00574 3877

ccs_category_11	Cancer of head and neck		I	0.17142 863	5	0.00252 7859	0.00015 9528
ccs_category_12	Cancer of esophagus		I	0.30804 0997	5	0.00113 1277	6.22E- 05
ccs_category_127	Chronic obstructive pulmonary disease and bronchiectasis		I	0.33423 7936	5	0.00969 9354	0.01670 0666
ccs_category_128	Asthma		I	0.17608 8728	5	0.00502 5356	0.00086 9412
ccs_category_13	Cancer of stomach		I	0.17643 1487	3	0	3.92E- 05
ccs_category_14	Cancer of colon		I	- 0.14539 7781	5	0.00347 2131	0.00030 0428
ccs_category_15	Cancer of rectum and anus		I	0.00896 6907	5	0.00147 5359	8.59E- 05
ccs_category_158	Chronic kidney disease		I	- 0.10959 403	5	0.00322 0377	0.00088 5548
ccs_category_16	Cancer of liver and intrahepatic bile duct		I	- 0.82664 4694	5	0.00124 5792	2.77E- 05
ccs_category_17	Cancer of pancreas		I	- 0.51526 7027	5	0	3.16E- 05
ccs_category_18	Cancer of other GI organs; peritoneum		I	- 0.16744 0593	5	0.00359 705	4.01E- 05
ccs_category_19	Cancer of bronchus; lung		I	0.28411 5021	5	0.00614 915	0.00088 8644
ccs_category_199	Chronic ulcer of the skin		I	- 0.06119 7755	5	0.00346 4812	0.00059 2145
ccs_category_20	Cancer; other respiratory and intrathoracic		I	0.25704 7776	3	0	8.84E- 06
ccs_category_201	Infective arthritis and osteomyelitis		I	- 0.04525 664	5	0.00293 7539	0.00020 0795
ccs_category_202	Rheumatoid arthritis and related disease		I	0.15617 4537	5	0.00469 8444	0.00039 4623

ccs_category_21	Cancer of bone and connective tissue		I	- 0.01304 2052	3	0.00022 7702	1.74E- 05
ccs_category_22	Melanomas of skin		I	- 0.41826 2662	5	0.00413 7457	0.00010 478
ccs_category_23	Other non-epithelial cancer of skin		I	- 0.04013 8	5	0.00408 7659	0.00024 9476
ccs_category_24	Cancer of breast		I	- 0.00306 8478	5	0.00436 5318	0.00040 0172
ccs_category_25	Cancer of uterus		I	- 0.05094 4846	5	0.00306 401	0.00010 0104
ccs_category_26	Cancer of cervix		I	0.25520 5286	5	0.01386 3626	4.62E- 05
ccs_category_27	Cancer of ovary		I	- 0.13601 6592	3	0.00261 194	4.75E- 05
ccs_category_28	Cancer of other female genital organs		I	- 0.25474 092	5	0	1.76E- 05
ccs_category_29	Cancer of prostate		I	- 0.12240 4113	5	0.00516 593	0.00042 3988
ccs_category_30	Cancer of testis		I	- 0.59226 0093	2	0	1.19E- 06
ccs_category_31	Cancer of other male genital organs		I	0.23348 5919	4	0	5.62E- 06
ccs_category_32	Cancer of bladder		I	- 0.00792 7627	3	0.00169 4525	0.00019 2967
ccs_category_33	Cancer of kidney and renal pelvis		I	0.03245 0175	5	0.00393 5864	0.00010 6194
ccs_category_34	Cancer of other urinary organs		I	0.18342 3991	4	0	1.55E- 05
ccs_category_35	Cancer of brain and nervous system		I	0.17239 2206	4	0	3.78E- 05
ccs_category_36	Cancer of thyroid		I	- 0.13675 9757	4	0.00173 5553	4.97E- 05
ccs_category_41	Cancer; other and unspecified primary		I	- 0.18992 3837	5	0.00277 3971	8.44E- 05

ccs_category_49	Diabetes mellitus without complication		I	0.02503 318	5	0.00245 2704	0.00082 1483
ccs_category_50	Diabetes mellitus with complications		I	- 0.02367 4478	5	0.00334 8722	0.00068 6683
ccs_category_56	Cystic fibrosis		I	0	0	0	1.13E- 06
ccs_category_652	Attention-deficit, conduct, and disruptive behavior disorders		I	0.28486 7888	5	0.00131 4399	9.87E- 05
ccs_category_653	Delirium, dementia, and amnesic and other cognitive disorders		I	- 0.11757 9976	5	0.00357 4755	0.00106 383
ccs_category_654	Developmental disorders		I	0.18812 178	5	0.00536 3953	0.00014 3701
ccs_category_658	Personality disorders		I	- 0.11909 0298	3	0.00061 182	6.53E- 05
ccs_category_659	Schizophrenia and other psychotic disorders		I	- 0.04381 8752	5	0.00431 3163	0.00041 1371
ccs_category_79	Parkinson's disease		I	- 0.05773 1248	5	0.00604 4322	0.00033 8107
ccs_category_80	Multiple sclerosis		I	- 0.04417 5298	3	0.00307 6741	4.81E- 05
ccs_category_81	Other hereditary and degenerative nervous system conditions		I	- 0.07474 5132	5	0.00334 5345	0.00039 5984
ccs_category_82	Paralysis		I	- 0.23888 6876	5	0.00585 0259	0.00028 7808
ccs_category_83	Epilepsy; convulsions		I	- 0.04519 4678	5	0.00620 2542	0.00041 4871
ccs_category_84	Headache; including migraine		I	- 0.05570 3109	5	0.00447 8454	0.00036 8459

ccs_category_85	Coma; stupor; and brain damage		I	0.09838 861	5	0.00782 972	0.00031 1325
ccs_category_96	Heart valve disorders		I	- 0.04808 7793	5	0.00298 4534	0.00063 014
ccs_category_97	Peri-; endo-; and myocarditis; cardiomyopathy		I	- 0.07547 9568	5	0.00453 4843	0.00048 857
ccs_category_98	Essential hypertension		I	- 0.07981 19	5	0.00288 0329	0.00089 7544
ccs_category_99	Hypertension with complications and secondary hypertension		I	- 0.08900 3542	4	0.00285 8586	0.00076 6368
claims_carrier_12mth	Carrier (12 mth)		I	5.89205 8732	5	0.00525 3677	0.02018 6735
claims_inpatient_12mth	Inpatient (12mth)	Number of inpatient claims from the last 12 months	I	- 0.48467 3054	5	0.00493 5695	0.00891 2581
claims_outpatient_12mth	Outpatient (12mth)		I	0.12731 3636	4	0.00297 237	0.00708 118
claims_prev_1_month	CLAIMS (6mth)		I	0.01485 2208	5	0.00219 6613	0.00206 1349
claims_prev_3_month	CLAIMS (6mth)		I	- 0.10541 4001	5	0.00299 2575	0.00045 3428
claims_prev_6_month	CLAIMS (6mth)		I	0.23034 5209	5	0.00198 3849	8.09E- 05
claims_prev_6mth	CLAIMS (6mth)		I	- 0.01152 7121	3	0	1.78E- 05
claims_snf_12mth	SNF (12)		I	- 1.06431 4511	5	0.00360 259	0.00240 8597
cog	Cognitive disorders	Grouped 11 mental disorders through summation	I	0.69907 6479	5	0.00284 899	0.00293 5822
Cystic_fibro_prev_mont h	Cystic (1mth)		I	0	0	0	0
dme_cost_1mth	DME (1mth)		I	0	0	0.00375 9271	0.01118 9821

EDAC_30_AMI	Unplanned hospital visits due to heart attack		P	0.08830 5038	4	0.00391 2942	0.00686 3091
EDAC_30_HF	Unplanned hospital visits due to heart failure	Unplanned hospital visits due to heart failure	P	0.04672 2281	4	0.00334 5113	0.01797 8401
EDAC_30_PN	30-day rates of readmission and excess dates in acute care due to pneumonia		P	- 0.45044 784	5	0.00348 6448	0.02397 1122
endocrinomology_prev_one_month	Endocrinomology (1mth)		I	- 0.04053 1281	4	0.00301 6663	0.00017 9255
er_prev_1_month	ER (1mth)		I	0.10179 2344	5	0.00359 8707	0.00081 4736
er_prev_3_month	ER (3mth)		I	- 0.04904 3622	5	0.00301 6895	0.00181 8253
er_prev_6_month	ER (6mth)		I	0.02050 3632	5	0.00228 704	0.00128 428
er_prev_one_month	ER (1mth)		I	2.90295 2884	5	0.00450 7293	0.00101 0427
er_prev_three_month	ER (3mth)		I	- 0.44761 5458	2	0.00339 3852	0.00300 534
ever_in_inpatient	EVER inpatient		I	0.34525 6924	5	0.00267 3134	0.00365 8128
Exam_eval_prev_month	Exam (1mth)		I	- 0.49883 7284	5	0.00039 71	2.93E- 05
Fatigue_prev_month	FATIGUE (1mth)		I	0.29915 3584	5	0.00323 813	8.31E- 05
Flu_High_Intensity_Last_Month	Intensity of flu	Binary from CDC dataset	C	- 0.80719 9415	5	0.01409 8957	0.00502 3879
Flu_Mean_Activity_Level_Last_Month_high			C	2.56315 2008	5	0.02675 8725	0.00706 7129
Flu_Mean_Activity_Level_Last_Month_low			C	0.59731 6594	5	0.00581 334	0.00944 4376
Flu_Mean_Activity_Level_Last_Month_minimal	Mean activity level flu (1mth)		C	ref			

Flu_Mean_Activity_Level_Last_Month_moderate			C	1.93539 3442	5	0.01433 2088	0.00208 29
Flu_Vaccine	flu vaccine		I	- 0.07230 4453	5	0.00231 3049	0.00091 1169
Flu_Widespread_Last_Month	Activity spread of flu	Binary from CDC dataset	C	- 1.59517 7162	5	0	0.01920 6717
frailty	Frailty	Computed using algorithm described in Appendix	I	0.84175 5046	5	0.00310 1413	0.01444 8052
Fx_skull_fac_prev_month	Fx skull fracture (1mth)		I	1.10710 6363	4	0	7.55E- 06
gender_1	Male		I	ref			
gender_2	Female		I	- 0.07217 6285	5	0.00239 7	0.00092 4936
HAI_1_SIR	CLABSI rate	Central line-associated bloodstream infections (CLABSI) in ICUs and select wards	P	0.06354 9465	4	0.00355 8615	0.00762 3737
HAI_2_SIR	CAUTI rate	Catheter-associated urinary tract infections	P	- 0.26229 3573	5	0.00328 9097	0.01201 6962
HAI_3_SIR	Colon Surgery Infection rate	Surgical site infection from abdominal hysterectomy	P	- 0.08100 0734	5	0.00344 1618	0.00674 9828
HAI_4_SIR	Surgical site infection from abdominal hysterectomy		P	- 0.36315 6528	5	0.00429 5486	0.00276 5498
HAI_5_SIR	MRSA rate	Methicillin-resistant Staphylococcus aureus blood laboratory-identified rates	P	- 0.00352 085	5	0.00377 3912	0.00604 4835
HAI_6_SIR	C.diff laboratory identified events	Clostridium difficile	P	- 0.22892 155	4	0.00317 6211	0.02100 321

heart	Number of heart conditions		I	0	0	0.00294 8107	0.00604 7586
HH_now	Homehealth now	Indicates whether a patient was in a home healthy facility at the beginning of the currenrt month	I	- 0.08936 2611	5	0.00353 2159	0.00057 9234
high_socioeconomic_insurance	High socioeconomic insurance	Identifying how a patients claims are being paid	I	0.25225 8935	5	0.00449 6845	0.00051 1287
homehealth_cost_6mth	HH (6mth)		I	0	0	0.00325 4639	0.00415 8679
hosp_prev_one_month	Hospital visits (1mth)	Hospital visits in the previous month	I	1.75208 8934	5	0.00427 0607	0.00252 6052
hosp_prev_three_month	Hospital visits (3mth)	Hospital claims in the last three months	I	- 2.96987 3822	5	0.01052 7929	0.01675 1449
Hospital overall rating 0			P	- 1.55629 5544	5	0.07113 1572	0.03995 5528
Hospital overall rating 1			P	0.12103 9953	5	0.00445 743	0.00041 2972
Hospital overall rating 2			P	0.18624 1102	5	0.00370 44	0.00146 3269
Hospital overall rating 3			P	0.21396 4598	5	0.00329 4986	0.00415 1935
Hospital overall rating 4			P	0.15129 166	5	0.00305 309	0.00373 3446
Hospital overall rating 5	Hospital overall rating		P	ref			
HospitalBedratio	Hospital bed ratio		C	- 0.24099 6141	4	0.00294 9986	0.00667 6321
Householdincome	Household income	Median household income in each county from County Health Rankings.	C	- 0.71020 9866	5	0.00312 8588	0.00847 3241
inp_cost_1mth	Cost inpatient (1mth)		I	0.54246 3445	1	0.00324 304	0.00374 0453
inp_cost_6mth	Cost inpatient (6mth)		I	- 3.17261 1577	5	0.00409 4014	0.01361 8965

inp_score_vw	Elixhauser Van Waren Score Inp	Weighted Elixhauser van Waren score from inpatient conditions	I	3.41696 4573	5	0.01917 6755	0.16074 0175
inp_prev_1_month	INP (6mth)		I	0.26783 8524	5	0.00592 6146	0.00197 1566
inp_prev_3_month	INP (6mth)		I	- 0.31491 8066	5	0.03534 4534	0.01192 6603
inp_prev_6_month	INP (6mth)		I	- 0.21318 1151	5	0.00636 0108	0.00982 4235
interaction_everinp_hospitalrating_0.0	Interaction hospital overall rating 2 and ever in inpatient		I	0.46647 7453	5	0.00223 9652	0.01127 148
interaction_everinp_hospitalrating_1.0	Interaction hospital overall rating 3 and ever in inpatient		I	- 0.04338 1926	5	0.00686 3595	0.00028 1282
interaction_everinp_hospitalrating_2.0	Interaction hospital overall rating 4 and ever in inpatient		I	- 0.06134 8586	5	0.00439 3442	0.00056 4232
interaction_everinp_hospitalrating_3.0	Interaction hospital overall rating 5 and ever in inpatient		I	- 0.15593 5363	5	0.00346 9408	0.00129 202
interaction_everinp_hospitalrating_4.0			I	- 0.16164 9205	5	0.00379 0259	0.00100 1924
interaction_everinp_hospitalrating_5.0	Interaction hospital overall rating 1 and ever in inpatient		I	ref			
low_income	Low income	Entitlement buy in indicator was used as a proxy for low income, codes 1 and A	I	0	0	0	0
Lung_externl_prev_month	Lung (1mth)		I	0.41252 3486	3	0	2.85E-05
Maint_chem_r_prev_month	Maintenance chemotherapy (1mth)		I	1.12638 8867	5	0.00184 9412	2.39E-05

MORT_30_CABG	CABG death rate	Death rate for CABG surgery patients at hospital level	P	- 0.48718 7795	5	0.00420 8522	0.00442 4282
neuro	Neurological disorders	Grouped 7 neurological disorders through summation	I	0	0	0.00308 4692	0.00107 5916
neurology_prev_one_month	Neurology (1mth)		I	- 0.14879 7333	5	0.00548 0601	0.00038 5655
nh_prev_6_month	NH (6mth)		I	0.06923 2519	5	0.00374 8219	0.00303 4171
number_carrier_claims_1mth	Carrier (1mth)	Number of carrier claims from the previous month.	I	- 2.78996 6252	5	0.00382 3108	0.01029 2021
number_dme_claims_1mth	DME (1mth)		I	5.84870 7282	5	0.00347 6843	0.00789 4628
number_homehealth_claims_1mth	HH (1mth)		I	0.67396 5831	2	0.00442 0849	0.00062 4673
number_homehealth_claims_6mth	Homehealth (6mth)	Number of homehealth claims in the last 6 months	I	- 0.45574 5417	5	0.00399 1295	0.00222 8118
number_outpatient_claims_1mth	Outpatient (1mth)		I	- 0.55311 4746	4	0.00249 9208	0.00289 4622
number_snf_claims_1mth	SNF (1mth)		I	0.32243 9776	5	0.00613 6478	0.00093 7333
nursing_home_prev_1_month	NH (1mth)		I	- 0.39300 8564	5	0.00867 5164	0.00348 0535
Obesity	Obesity		I	0.11254 2024	5	0.00387 1212	0.00050 421
oncology_vist_prev_1_month	Oncology (1mth)		I	0.85780 7843	5	0.00811 4268	0.00457 3266
oncology_vist_prev_3_month	Oncology (1mth)		I	1.02390 9662	5	0.00523 9015	0.00099 8717
oncology_vist_prev_6_month	Oncology (1mth)		I	- 0.60873 1948	5	0.00237 2363	0.00019 4163
out_cost_3mth	Outpatient cost (3mth)		I	- 0.17976 2367	3	0.00264 9256	0.00686 0624
pc_prev_one_month	Primary Care (1mth)		I	- 0.53439 7141	4	0.00255 5257	0.00339 5933

PCPRatio	Population per primary care provider		C	- 0.02802 4944	4	0.00301 6625	0.00759 4728
Pct_Flu_Widespread_Last_Month	Percent of weeks flu is widespread	CDC state	C	0.61248 7869	5	0.00906 7604	0.02306 74
Percentage of SNF residents who experience one or more falls with major injury during their SNF stay	Percent of SNF residents who experience one or more falls		P	0.17472 026	5	0.00416 9868	0.00273 5507
PercentageAmericanIndian/AlaskaNative/NativeHawaiian/OtherPacificIslander	Percent American Indian/Alaska Native/Native Hawaiian/Other Pacific Islander		C	1.41170 8445	5	0.00316 5654	0.00667 8223
PercentageAsianPopulation	Percent Asian Population		C	0.41121 7688	5	0.00331 8737	0.00763 5805
PercentageBlackAfricanAmerican	Percent African American		C	- 0.30043 8063	5	0.00321 8225	0.00767 9309
Percentageilliterate	Percent Illiterate	Percent of illiterate people obtained from county health ranking.	C	- 0.12417 6841	5	0.00332 098	0.00771 9514
PercentageofMedicareEnrollees	Percentage of Medicare Enrollees	Computed from Census and County Health Ranking	C	- 0.47215 8411	5	0.00307 849	0.00790 4172
PercentagePersonover65inDeepPoverty	Percentage 65+ Poverty	Calculated from census and Area health resource file.	C	- 0.32156 3397	4	0.00294 5067	0.00570 1835
PercentageRural	Percent Rural	Percent of population in each county living in rural areas	C	0.30125 5667	5	0.00306 0511	0.00824 8502

PercentageSmokers	Percent Smokers	Percent of adults that report smoking at least 100 cigarettes and currently smoke from county health rankings.	C	- 0.19984 4226	5	0.00287 267	0.00682 2793
Pneumonia_prev_month	PNEUMONIA (1mth)		I	0.35614 4172	5	0.00432 6973	0.00011 5331
Pneumonia_Vaccine	pneumococcal vaccine		I	0.05157 8179	5	0.00358 7568	0.00069 5353
post_transplant	Post transplant	Indicates whether a patient claim had major organ transplants (kidney, liver, heart, lungs, intestines, pancreas) from codes V420, V427, V421, V426, V4284, V4283	I	0	0	0	1.59E-07
PQI_11_inp_2010	PQI 11 (2010)		I	1.20401 0718	5	0.01012 2847	0.00327 7929
PQI_11_inp_prev_12months	Pqi (1mth)		I	- 1.05647 5296	5	0.00998 0473	0.00382 0528
Prev_ILI_prev_1_month	Prev ILI (1mth)		I	0.28087 3597	5	0.00533 9934	0.00062 4812
Prev_ILI_prev_3_month	Prev ILI(3mth)		I	- 0.26456 1252	5	0.00376 1839	0.00062 8395
Prev_ILI_prev_6_month	Prev ILI (6mth)		I	0.31190 8211	5	0.00423 9154	0.00070 4526
Prev_ILI_XRAY_prev_1_month	Prev ILI XRAY (1mth)		I	- 0.09353 0318	4	0.00348 1328	0.00010 2487
Prev_ILI_XRAY_prev_3_month	Prev ILI XRAY (3mth)		I	0.21335 3724	5	0.00713 5577	0.00022 5991
Prev_ILI_XRAY_prev_6_month	Prev ILI XRAY (6mth)		I	- 0.09304 1699	5	0.00724 0144	0.00029 6762
PSI_11_POST_RESP	Postoperative Respiratory Failure Rate	Hospital level postoperative respiratory failure rate	P	- 0.20059 6946	5	0.00319 4575	0.01993 1814

PSI_3_ULCER	Pressure ulcer rate		P	- 0.34328 0171	5	0.00309 4366	0.02274 0299
race_0	Unknown		I	- 0.29696 2992	5	0.00179 5426	3.25E- 05
race_1	White		I	ref			
race_2	Black		I	- 0.17258 4189	5	0.00514 3532	0.00067 3345
race_3	Other		I	- 0.17142 8248	5	0.00212 0759	0.00013 7683
race_4	Asian		I	0.04099 5874	4	0.00634 6689	0.00017 3189
race_5	Hispanic		I	0.27707 7584	5	0.00422 9582	0.00028 3254
race_6	North American Native		I	0.01998 1377	4	0.00246 6062	0.00015 3692
region_score_PQI05_0	PQI rates have not been considered, SSA codes 60-66,97-99	COPD or asthma	C	ref			
region_score_PQI05_1	Bottom 10% PQI rates		C	- 0.07069 9158	4	0.01026 5612	0.00039 3577
region_score_PQI05_2	Middle 80% PQI Rates		C	- 0.00343 647	4	0.00320 0607	0.00065 9259
region_score_PQI05_3	Top 10% PQI Rates		C	- 0.01333 01	3	0.00504 892	0.00057 6721
region_score_PQI07_0	PQI rates have not been considered, SSA codes 60-66,97-99	Hypertension	C	ref			
region_score_PQI07_1	Bottom 10% PQI rates		C	- 0.20655 0881	5	0.00626 5996	0.00045 1295
region_score_PQI07_2	Middle 80% PQI Rates		C	0.02691 5372	3	0.00653 8275	0.00043 4662
region_score_PQI08_0	PQI rates have not been considered, SSA codes 60-66,97-99	Heart Failure	C	ref			
region_score_PQI08_1	Bottom 10% PQI rates		C	0.05093 242	5	0.00810 3839	0.00021 3899

region_score_PQI08_2	Middle 80% PQI Rates		C	- 0.00876 4113	1	0.00408 0388	0.00057 6305
region_score_PQI08_3	Top 10% PQI Rates		C	- 0.09353 6004	5	0.00452 4846	0.00049 242
region_score_PQI11_0	PQI rates have not been considered, SSA codes 60-66,97-99	Community-acquired pneumonia	C	ref			
region_score_PQI11_1	Bottom 10% PQI rates		C	- 0.30559 999	5	0.00552 5039	0.00042 7099
region_score_PQI11_2	Middle 80% PQI Rates		C	- 0.01627 4026	5	0.00425 1228	0.00057 4678
region_score_PQI11_3	Top 10% PQI Rates		C	0.01449 3091	2	0.00567 7508	0.00050 66
region_score_PQI12_0	PQI rates have not been considered, SSA codes 60-66,97-99	Urinary tract infection	C	ref			
region_score_PQI12_1	Bottom 10% PQI rates		C	- 0.02287 2382	5	0.00691 7583	0.00022 9665
region_score_PQI12_2	Middle 80% PQI Rates		C	0.03190 8552	4	0.00410 7699	0.00055 4354
region_score_PQI12_3	Top 10% PQI Rates		C	- 0.02274 0502	3	0.00479 1816	0.00048 8559
region_score_PQI14_0	PQI rates have not been considered, SSA codes 60-66,97-99	Uncontrolled diabetes	C	ref			
region_score_PQI14_1	Bottom 10% PQI rates		C	- 0.12872 5209	5	0.00487 9085	0.00040 0619
region_score_PQI14_2	Middle 80% PQI Rates		C	- 0.05006 9956	3	0.00342 3633	0.00074 9312
region_score_PQI14_3	Top 10% PQI Rates		C	0.06397 4169	4	0.00371 5334	0.00067 3353
region_score_PQI16_0	PQI rates have not been considered, SSA codes 60-66,97-99	Lower-extremity amputation	C	ref			

region_score_PQI16_1	Bottom 10% PQI rates		C	- 0.08740 6298	4	0.00708 2933	0.00027 0697
region_score_PQI16_2	Middle 80% PQI Rates		C	0.02970 8641	4	0.00508 1249	0.00046 9548
region_score_PQI16_3	Top 10% PQI Rates		C	- 0.05843 562	4	0.00570 869	0.00034 7666
resp_symptoms_prev_1 _month	Respiratory symptoms (1mth)		I	- 0.77931 2848	5	0.00814 0147	0.05024 8494
Risk-Standardized Potentially Preventable Readmission\nRate	Risk- standardized potentially preventable hospital readmissions	Risk-standardized	P	- 0.19795 8669	5	0.00375 8385	0.00567 6478
RuralLabel	Rural	This variable takes the value of 1 is Percentage Rural is 99% or higher.	I	0.06681 7874	5	0	0.00048 1247
smoking	Smoker		I	0.45650 8865	5	0.00702 558	0.00670 1705
snf_cost_1mth	Cost snf (1mth)		I	4.33522 8816	3	0.00411 5473	0.00211 3296
snf_cost_6mth	SNF cost (6mth)		I	0.18066 8558	1	0.00367 5605	0.00307 1154
snf_score	Elixhauser score snf		I	1.08585 5657	5	0.00389 3672	0.01661 1554
snf_now	SNF now		I	0.15113 1533	5	0.00348 7285	0.00044 8807
still_in_nh	Still in NH	Nursing home by the index date	I	0.02737 4431	4	0.00531 3541	0.00038 5965
sum.ccs	Sum of CCS Comorbid conditions	The number of comorbidities a patient has ever had for each month.	I	1.23132 3523	5	0.00391 2187	0.01899 0361

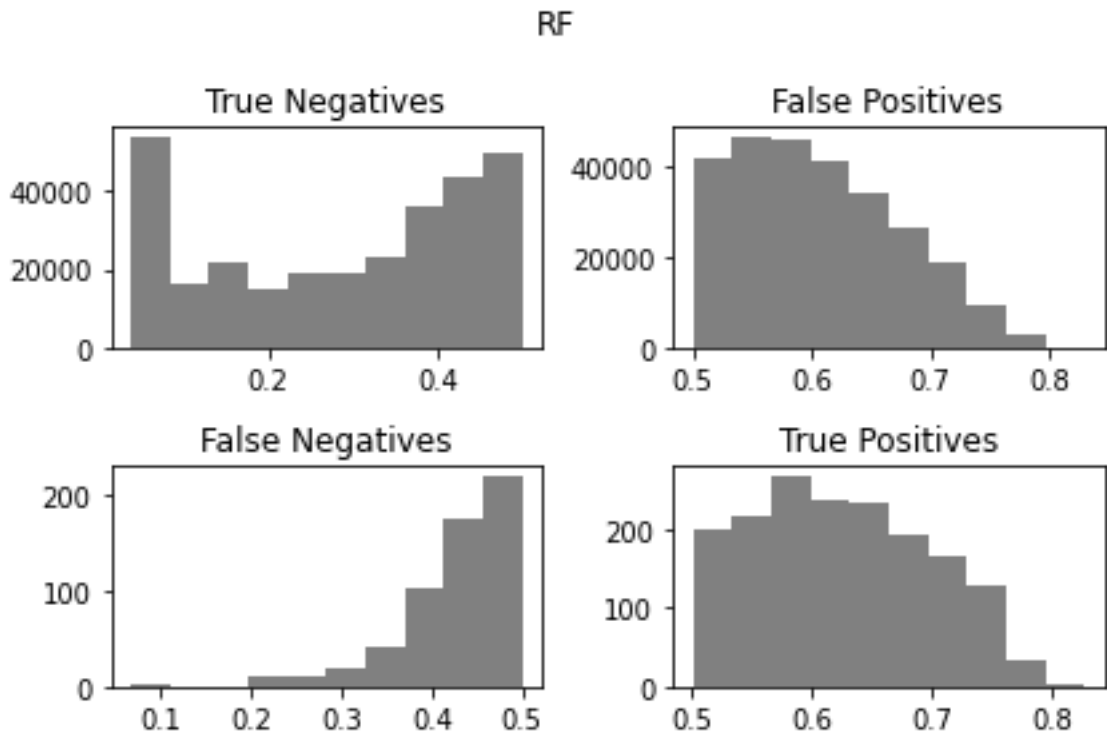


Figure B.1 Histogram of the random forest prediction score distributions for all patient rows.

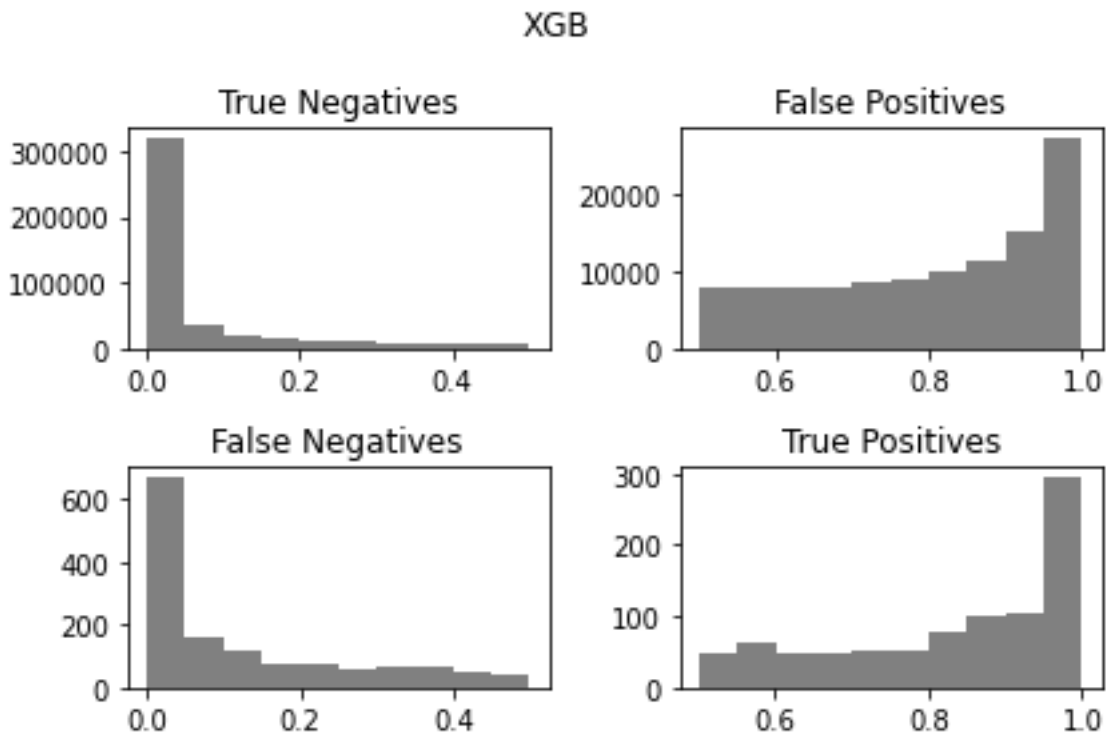


Figure B.2 Histogram of the XG boost prediction score distributions for all patient rows.

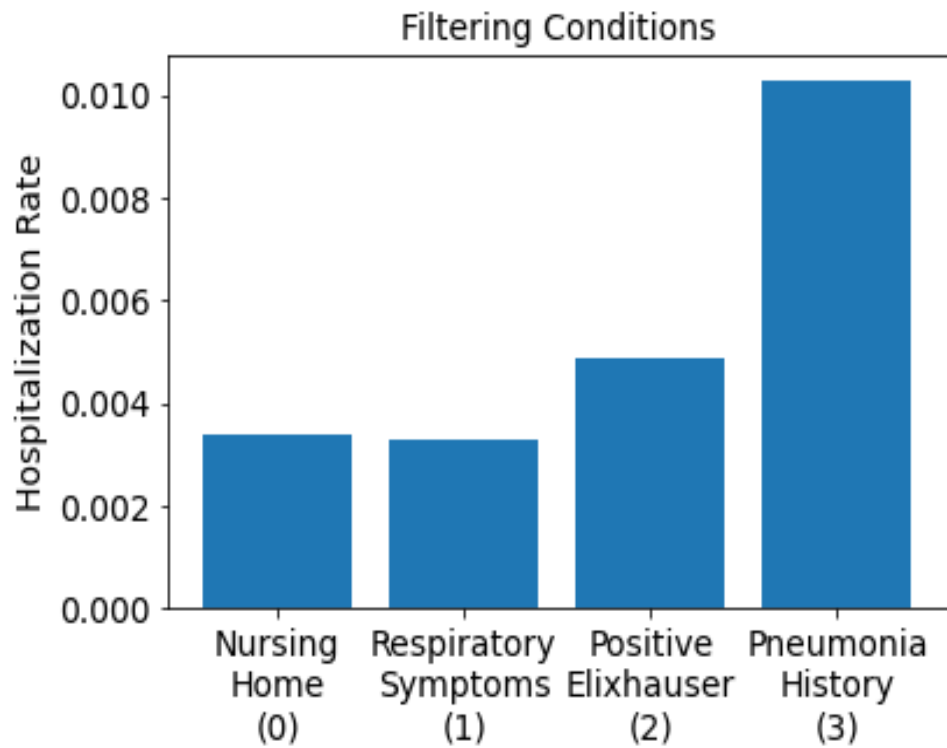


Figure B.3 Hospitalization rates among patient groups that were included in the prediction set because they had at least one of each condition, e.g., Nursing Home (0) indicates that patients in this group could have been included due to a recent nursing home visit, among other conditions.

APPENDIX C

APPENDIX TO CHAPTER 3

C.1 Derivations for Optimization Model

We begin with the cost function described in the main body of the paper.

$$TC(q, x_1) = C_h D_0 + \min\{P_{1+2}, r(q)\} C_{i2} + (P_{1+2} - W_1) C_{i1} + (1 - p e_2) C_h \min\{q, D_{1+2}\} + (1 - p e_1) C_h (D_{1+2} - Z_1)$$

Assuming $P_{1+2} - r(q) \geq 0$, and utilizing the linearity of the expectation function yields the following.

$$\begin{aligned}\mathbb{E}[TC(q, x_1)] &= \mathbb{E}[C_h D_0] + \mathbb{E}[C_{i2} r(q)] + \mathbb{E}[C_{i1}(P_{1+2} - \min\{P_{1+2}, r(q)\})] + \mathbb{E}[(1 - p e_2) C_h \min\{q, D_{1+2}\}] \\ &\quad + \mathbb{E}[(1 - p e_1) C_h (D_{1+2} - \min\{q, D_{1+2}\})]\end{aligned}\tag{C.1}$$

$$\begin{aligned}&= C_h \mathbb{E}[D_0] + C_{i2} r(q) + C_{i1}(P_{1+2} - r(q)) + (1 - p e_2) C_h \mathbb{E}[\min\{q, D_{1+2}\}] \\ &\quad + (1 - p e_1) C_h \mathbb{E}[D_{1+2} - \min\{q, D_{1+2}\}]\end{aligned}\tag{C.2}$$

$$\begin{aligned}&= C_h \int_{-\infty}^{\infty} x f_0(x) dx + C_{i2} r(q) + C_{i1}(P_{1+2} - r(q)) \\ &\quad + (1 - p e_2) C_h \mathbb{E}[q | q < D_{1+2}] p(q < D_{1+2}) + (1 - p e_2) C_h \mathbb{E}[D_{1+2} | D_{1+2} \leq q] p(D_{1+2} \leq q) \\ &\quad + (1 - p e_1) C_h \mathbb{E}[D_{1+2} - q | D_{1+2} - q > 0] p(D_{1+2} - q > 0)\end{aligned}\tag{C.3}$$

$$\begin{aligned}&= C_h \int_{-\infty}^{\infty} x f_0(x) dx + C_{i2} r(q) + C_{i1}(P_{1+2} - r(q)) \\ &\quad + (1 - p e_2) C_h q \int_{x>q} f_{1+2}(x) dx + (1 - p e_2) C_h \int_{x \leq q} x f_{1+2}(x) dx \\ &\quad + (1 - p e_1) C_h \int_{x>q} x f_{1+2}(x) dx - (1 - p e_1) C_h q \int_{x>q} f_{1+2}(x) dx\end{aligned}\tag{C.4}$$

$$\begin{aligned}&= C_h \int_{-\infty}^{\infty} x f_0(x) dx + C_{i2} r(q) + C_{i1}(P_{1+2} - r(q)) \\ &\quad + (1 - p e_2) C_h q [1 - F_{1+2}(q)] + (1 - p e_2) C_h \int_{x \leq q} x f_{1+2}(x) dx \\ &\quad + (1 - p e_1) C_h \int_{x>q} x f_{1+2}(x) dx - (1 - p e_1) C_h q [1 - F_{1+2}(q)]\end{aligned}\tag{C.5}$$

$$\begin{aligned}&= C_h \int_{-\infty}^{\infty} x f_0(x) dx + C_{i2} r(q) + C_{i1}(P_{1+2} - r(q)) + C_h q [1 - F_{1+2}(q)] (p e_1 - p e_2) \\ &\quad + (1 - p e_2) C_h \int_{x \leq q} x f_{1+2}(x) dx + (1 - p e_1) C_h \int_{x>q} x f_{1+2}(x) dx\end{aligned}\tag{C.6}$$

Here, the progression from (8) to (9) relies on the definition of condition expectation. That is, for a random variable X we have the following definitions.

In the discrete case:

$$\mathbb{E}[X | Y = y] = \sum_x P(X = x | Y = y) = \sum_x \frac{P(X = x, Y = x)}{P(Y = y)}$$

In the continuous case:

$$\mathbb{E}[X | Y = y] = \int_{-\infty}^{\infty} f_{X|Y}(x|y) dx = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)}{f_Y(y)} dx$$

Generalizations for Other Classifiers

In the best case of a perfect classifier.

If you further assume that you have a perfect classifier, for unbalanced and balanced datasets, where all of those who will experience a potentially preventable event have a higher risk score than those who will not experience one, then the function $r(q)$ becomes linear with a slope of 1. In this best case, the model becomes the classical newsvendor formulation with critical fractile:

$$F(q) = \frac{C_h(p e_2 - p e_1) - (C_{i2} - C_{i1})}{C_h(p e_2 - p e_1)} = \frac{C_U}{C_U + C_O} = \frac{1}{1 + \frac{C_O}{C_U}}$$

where

$$C_U = C_h(p e_2 - p e_1) - (C_{i2} - C_{i1})$$

$$C_O = C_{i2} - C_{i1}$$

Here the optimal intervention quantity for a given x_1 value is driven solely by the ratio between the overage and underage costs associated with the given problem.

In the worst case of a random classifier.

In the worst case scenario where you do not have any information to calculate the risk scores, you can always randomly assign risk scores uniformly between 0 and 1. In this case, the function $r(q)$ again becomes linear, but now $r'(q)$ is equal to the inverse of the event rate in the population. Therefore, the expected cost function remains convex and the critical fractile found (CR), found in the main body still holds.

When $r'(q) = k$ for any constant k , the optimal q^* can be found by applying the inverse empirical cumulative distribution function, or known cumulative distribution function. For any case in between the perfect and random classifier you can approximate $r(q)$ using a convex, quadratic function whose slope is determined by the changing event rate per risk score bucket in the population.

Sampling Distributions for Number of Hospitalizations (N=100)

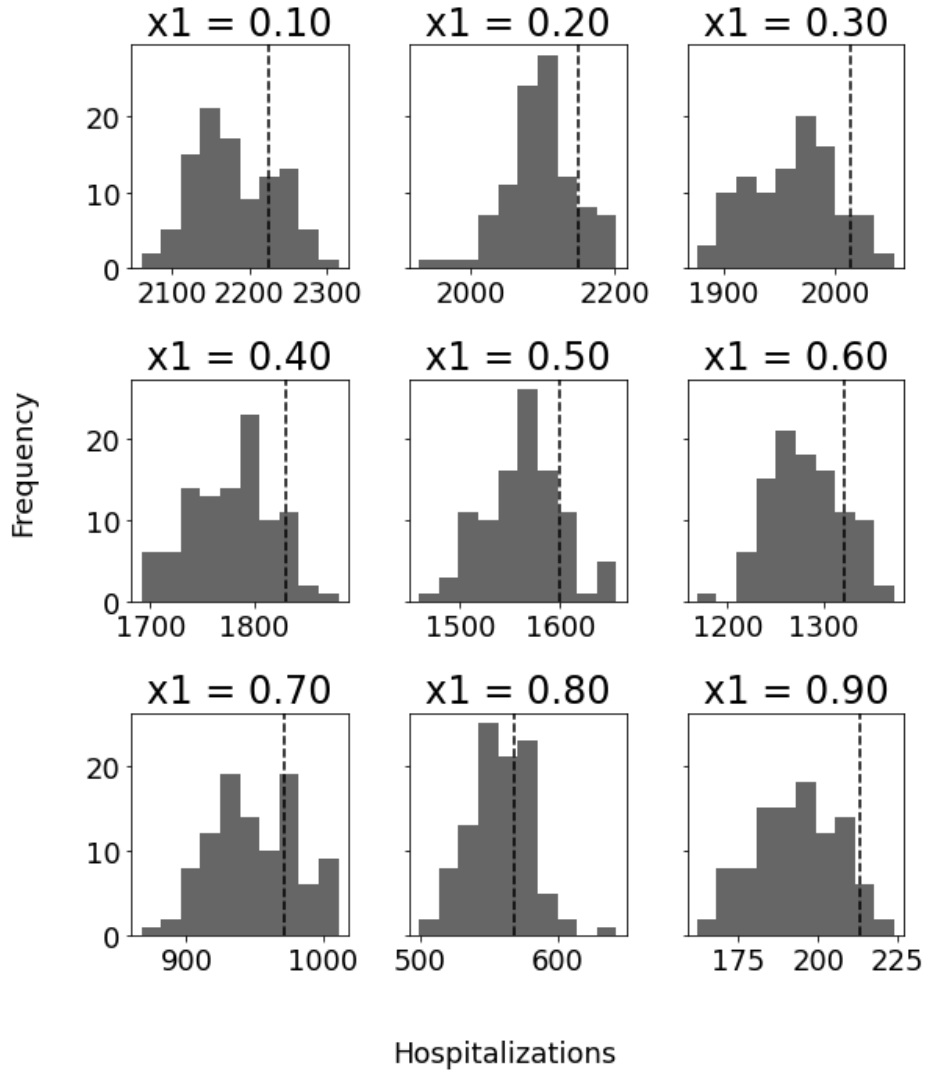


Figure C.1 Sampling distribution of the sampled training data truncated with risk scores on or above x_1 , showing the number of hospitalizations in the set. This distribution is used to approximate $F(q)$. The dashed line represents the actual number of hospitalizations in the testing set with risk scores on or above x_1 .

APPENDIX D

APPENDIX TO CHAPTER 4

Infection Fatality Rates from the Literature

Table D.1 Resulting population-weighted IFR values, by age, for the United States from each study.

Study	US Age Population-Weighted IFR
Verity et al. (2020) IFR	0.94%
Ferguson et al. (2020) IFR	1.08%
Metaregression IFR	1.15%
CDC Pandemic Planning IFR	0.72%
Hubei, China	3.46%
Austria	1.23%
Baden-Wurttemberg, Germany	1.34%
Bavaria, Germany	1.40%
Lombardy, Italy	2.22%
Spain	2.03%
Switzerland	1.02%

Estimated Lab Multipliers from Deterministic Application of Infection Fatality Rates

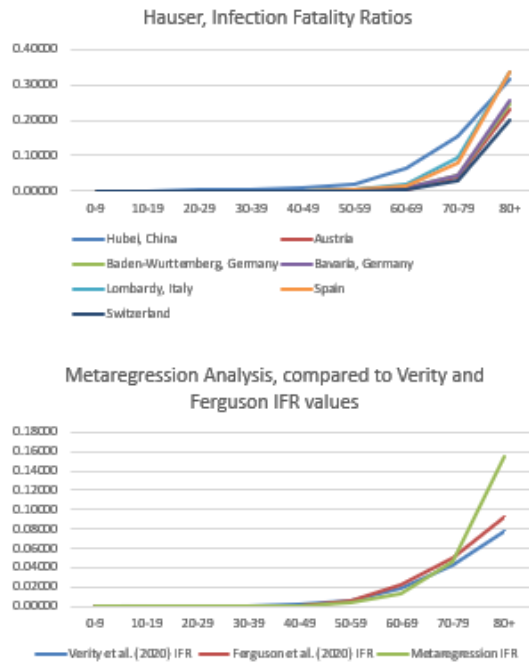


Figure D.1 Infection Fatality Ratios presented from empirical data from China and six regions in Europe, stratified by age, for the early months (January-May 2020) of the COVID-19 pandemic, and the comparison to other studies and a developed meta regression IFR set by age group.

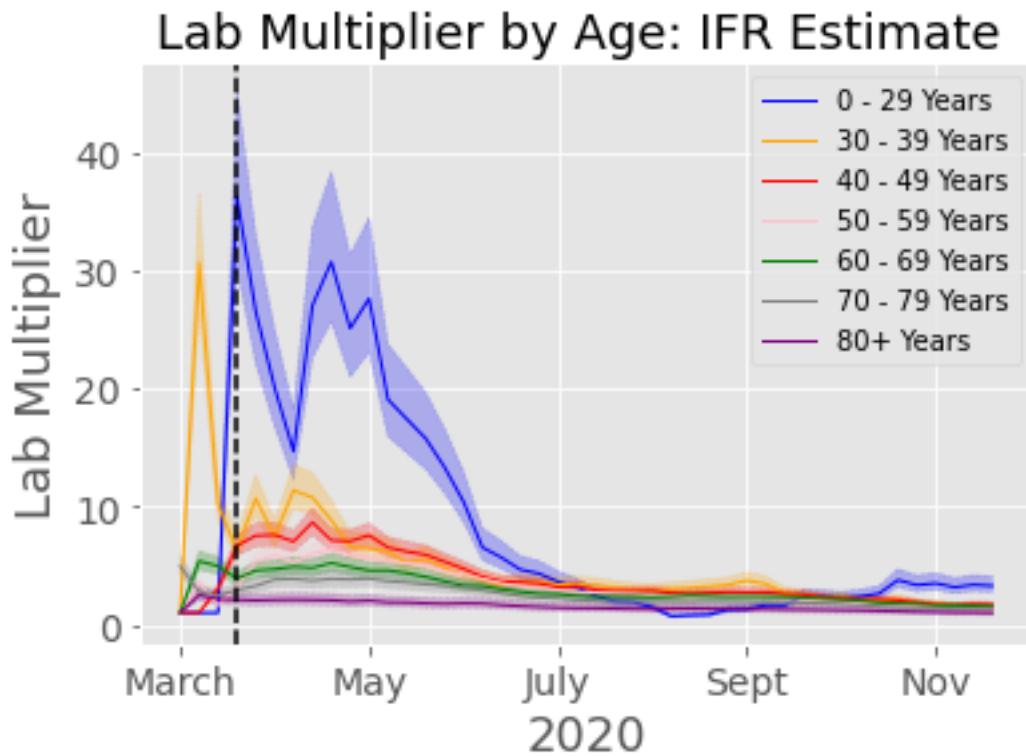


Figure D.2 North Carolina estimated lab multipliers by age group using the IFR values and 95% confidence intervals from [104]

North Carolina Deterministic IFR Lab Multiplier by Age and Geographic Location

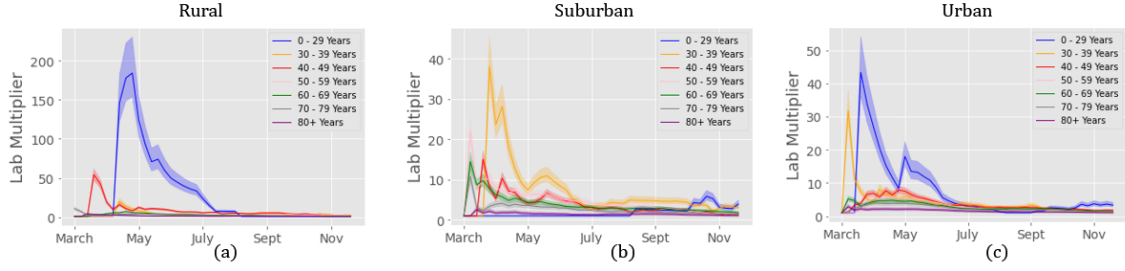


Figure D.3 North Carolina lab multiplier estimates by age and geographic location from IFR analysis.

Derivation of Time Varying Reproductive Number

$$R^{i,j}(t) = \beta(t) [m_{i,j}(C_d - C_u) - 1]$$

Proof. Consider the vector, $(S(t), E(t), I_d(t), I_u(t), H(t), R(t), D(t))$

Then observe that the disease-free equilibrium point exists and when $\mathcal{S}_0 = N$ and $\beta(t = 0) = \beta$ is given by

$$(\mathcal{S}_0, 0, 0, 0, 0, 0, 0)$$

Let $\hat{x} = (E(t), I_d(t), I_u(t))$ Consider the form.

$$\frac{dx}{dt} = \mathcal{F}(x) - \mathcal{V}(x)$$

Where

$$F(\hat{x}) = \begin{pmatrix} \beta(I_d + I_u) \\ 0 \\ 0 \end{pmatrix}$$

$$V(\hat{x}) = \begin{pmatrix} \frac{E}{\tau_{EI}} \\ k_1 I_d + k_2 I_d + k_3 I_d - m \frac{1}{\tau_{EI}} E \\ k_4 I_u + k_5 I_u \end{pmatrix}$$

$$\begin{aligned}
k_1 &= \frac{(1 - \rho_{IDH} - \rho_{IDD})}{\tau_{IR}} \\
k_2 &= \frac{\rho_{IDH}}{\tau_{DH}} \\
k_3 &= \frac{\rho_{IDD}}{\tau_{IDD}} \\
k_4 &= \frac{(1 - \rho_{IUH})}{\tau_{IR}} \\
k_5 &= \frac{(\rho_{IUH})}{\tau_{IUH}}
\end{aligned}$$

Then take the gradient.

$$\hat{V} = \nabla(V)_x = \begin{pmatrix} \frac{1}{\tau_{EI}} & 0 & 0 \\ -m \frac{1}{\tau_{EI}} & k_1 + k_2 + k_3 & 0 \\ -(1-m) \frac{1-m}{\tau_{EI}} & 0 & k_4 + k_5 \end{pmatrix}$$

Now the inverse.

$$\hat{V}^{-1} = \begin{pmatrix} \tau_{EI} & 0 & 0 \\ \frac{m}{k_1 + k_2 + k_3} & \frac{1}{k_1 + k_2 + k_3} & 0 \\ \frac{1-m}{k_4 + k_5} & 0 & \frac{1}{k_4 + k_5} \end{pmatrix}$$

Then take the gradient of F .

$$\hat{F} = \nabla(F)_x = \begin{pmatrix} 0 & \beta & \beta \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Let $C_d = \frac{1}{k_1 + k_2 + k_3}$ and $C_u = \frac{1}{k_4 + k_5}$
Then compute $\hat{F} \hat{V}^{-1}$

$$\hat{F} \hat{V}^{-1} = \begin{pmatrix} m C_d \beta + (1-m) C_u \beta & \frac{C_d \beta \mathcal{S}_0}{N} & C_u \beta \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Then the spectral radius $\rho(FV^{-1})$ is given by:

$$\rho(\hat{F} \hat{V}^{-1}) = m C_d \beta + (1-m) C_u \beta = \beta [m(C_d - C_u) - 1]$$

□

NC Weekly Lab Multiplier vs Weekly Percent Positive Tests

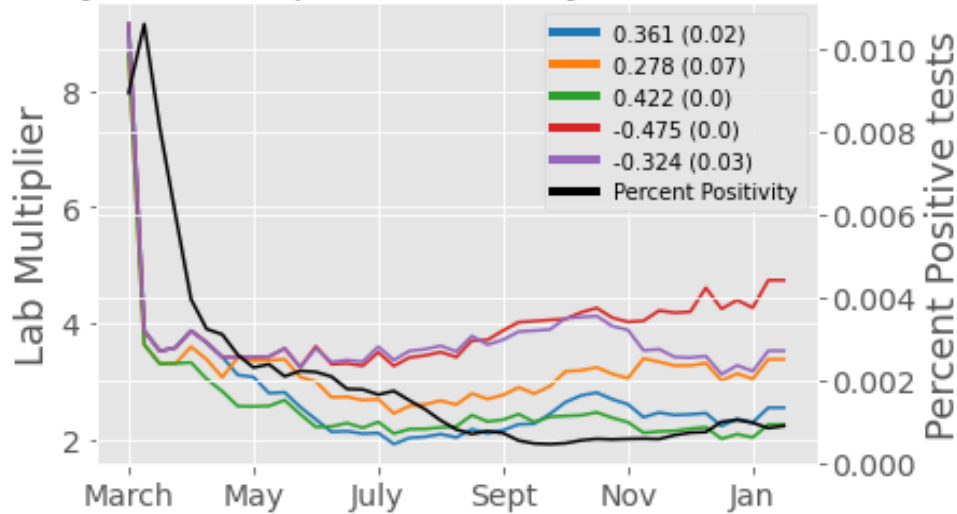


Figure D.4 Comparing trends between the weekly branching SEIR estimates for time varying lab multiplier and percent positive tests. The percent positive tests is estimated by the 7-day average of daily lab-confirmed cases by the New York Times over the 7-day average of daily total tests administered collected by John's Hopkins University and availability in the JHU CCI data repository. The legend indicates each trajectory's correlation between weekly lab multipliers and percent positivity, computed using the Spearman correlation test, and the associated p-value in parenthesis.

Clustering and Relevant Trends in Data

We collect highly informative data such as lab-confirmed cases, COVID-19 related deaths, positivity rate, and mobility data illustrated in Figures D.6 and D.5. We compute the daily percent change in each of these measure values to create feature vectors that capture the change in disease progression throughout 2020. We use these vectors to develop a clustering algorithm to infer where these vectors change significantly over time. We are not concerned with the values of each cluster, or whether the effective reproductive number fit for each cluster is equal, rather we are just interested in if we can identify changes in disease spread over time for which the SEIR model will need to adjust and account for.

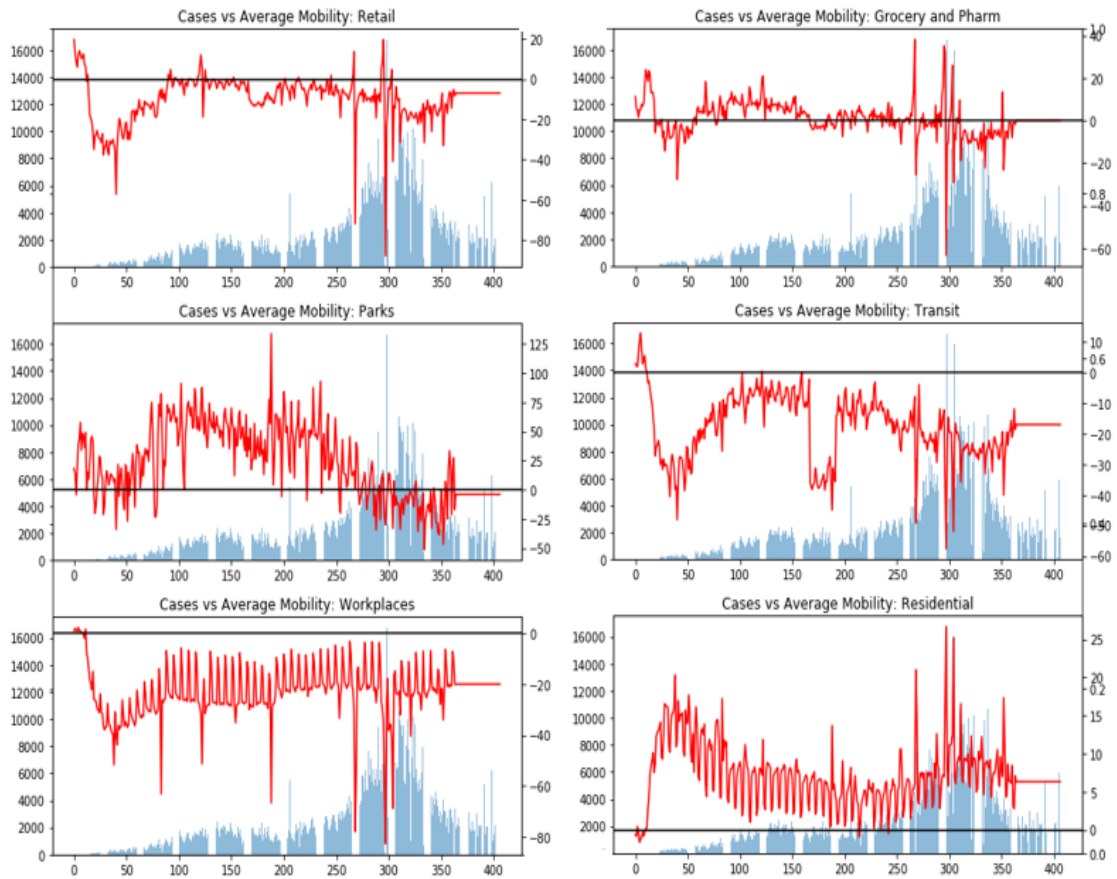


Figure D.5 North Carolina lab-confirmed cases versus average daily mobility (Google mobility reports)

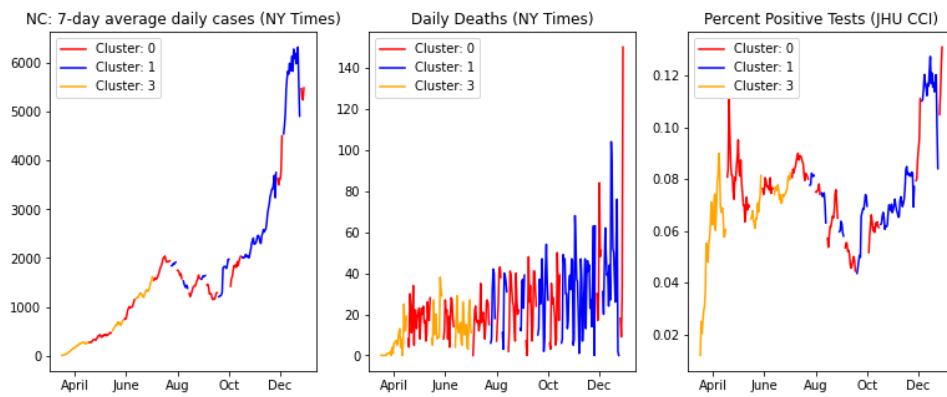


Figure D.6 Results from clustering on North Carolina data.