2024

Data driven models in health care management

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

DATA DRIVEN MODELS IN HEALTH CARE MANAGEMENT

by

YANG HU

B.E., National University of Singapore, 2018

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2024

Approved by

First Reader

Ioannis Ch. Paschalidis, Ph.D.
Distinguished Professor of Engineering
Professor of Electrical & Computer Engineering
Professor of Systems Engineering
Professor of Biomedical Engineering
Professor of Computing & Data Sciences

Second Reader

Christos Cassandras, Ph.D.
Professor of Electrical and Computer Engineering
Professor and Division Head of Systems Engineering

Third Reader

Ayse Coskun, Ph.D.
Professor of Electrical and Computer Engineering
Professor of Systems Engineering

Fourth Reader

William G. Adams, MD
Professor of Pediatrics
Aram V. Chobanian and Edward Avedisian School of Medicine

Acknowledgments

First and foremost, I wish to express my deepest gratitude to my advisor, Prof. Ioannis Ch. Paschalidis, for his unwavering support and guidance. His vision and passion for cutting-edge scientific research have always inspired me to tackle and solve complex problems. I have been profoundly influenced by his extensive theoretical knowledge and enthusiasm for addressing practical issues, which have set a stellar example for me. Prof. Paschalidis has been incredibly patient and supportive, allowing me ample time to learn and grow. I have gained so much from his mentorship. Overall, a heartfelt thank you to my advisor; I am fortunate to have had such an outstanding mentor throughout my Ph.D. journey. You are the perfect advisor for me!

I would also like to extend my sincere thanks to my committee members, Prof. Christos Cassandras, Prof. Ayse Coskun, and Dr. William G. Adams. Their invaluable feedback, insightful comments, and constructive criticism have greatly enhanced the quality and success of my research. Their expertise and encouragement have been immensely beneficial.

Furthermore, I am deeply grateful to all the members of my labs. To my lab mates and colleagues, Boran Hao, Ruidi Chen, Shahabeddin Sotudian, Taiyao Wang, Tingting Xu, and Zahra Zad, thank you for the stimulating discussions, the collaborative projects, and the enjoyable times we have shared over the past few years. Your camaraderie and support have made my Ph.D. journey both rewarding and memorable.

Lastly, I want to express my heartfelt gratitude to my family and all my friends in Boston for their unwavering love, understanding, and encouragement. To all my friends in Boston, thank you for the joyful moments we have spent together; I have learned so much from each of you. Your belief in me has been a constant source of motivation and strength.

Thank you all for being an integral part of this remarkable journey.

DATA DRIVEN MODELS IN HEALTH CARE MANAGEMENT

YANG HU

Boston University, College of Engineering, 2024

Major Professor: Ioannis Ch. Paschalidis, Ph.D.
Distinguished Professor of Engineering
Professor of Electrical & Computer Engineering
Professor of Systems Engineering
Professor of Biomedical Engineering
Professor of Computing & Data Sciences

ABSTRACT

The increasing role of Machine Learning (ML) in scientific and biomedical fields has significantly advanced health and medical informatics. ML-based predictive models have shown great potential in identifying critical features and enabling early-stage treatment. This research explores the increasing worldwide shift towards using data-driven models to support decision-making in healthcare.

This work investigates data-driven techniques to predict health-related events and elucidates important predictive variables. Firstly, it identifies key factors in predicting poorly controlled hypertension using demographic and socioeconomic data, revealing that age, race, social determinants of health, mental health, marital status, cigarette use, and gender are predictive of high systolic blood pressure (SBP). Secondly, it develops a personalized hypertension management model that optimizes medication outcomes using Distributionally Robust Optimization (DRO) regularized regression and K-Nearest Neighbors (K-NN) regression, achieving a significant reduction in SBP compared to standard care. Thirdly, it predicts critical COVID-19 treatment outcomes including hospitalization, ICU care, mechanical

ventilation, and mortality using both linear and nonlinear classification methods, achieving high predictive accuracy. Lastly, it emphasizes the importance of model interpretability and alignment with medical knowledge to ensure adoption and trust in clinical settings. By introducing the Wasserstein DRO formulation and the Grouped LASSO (GLASSO) algorithm, the research demonstrates enhanced interpretability and credibility, effectively merging model performance with expert knowledge.

Together, these contributions demonstrate the potential of data-driven models to enhance healthcare delivery and patient care by providing a comprehensive approach that balances interpretability, credibility, and expert insight integration.

Contents

List of Tables

List of Figures

List of Abbreviations

ACEi	. . . . .	ACE inhibitors
AI	. . . . .	Artificial Intelligence
ARBs	. . . . .	Angiotensin Receptor Blockers
AUC	. . . . .	Area Under the Receiver Operating Characteristic Curve
BMC	. . . . .	Boston Medical Center
BP	. . . . .	Blood Pressure
BUN	. . . . .	Blood Urea nitrogen
CAD	. . . . .	Coronary Artery disease
CCBs	. . . . .	Calcium Channel Blockers
CHF	. . . . .	Congestive Heart failure
CI	. . . . .	Confidence Intervals
CKD	. . . . .	Chronic Kidney Disease
CRP	. . . . .	C-reactive Protein
DL	. . . . .	Deep Learning

DRLR	. . . . .	Distributionally Robust linear regression
DRO	. . . . .	Distributionally Robust Optimization
DVT	. . . . .	Deep Vein Thrombosis
EHRs	. . . . .	Electronic Health Records
FNR	. . . . .	False Negative Rate
FPR	. . . . .	False Positive Rate
GLASSO	. . . . .	Grouped LASSO
HLD	. . . . .	Hypersensitivity Lung Disease
ICU	. . . . .	Intensive Care unit
KNN	. . . . .	K-Nearest Neighbor
KS	. . . . .	Kolmogorov-Smirnov
LDH	. . . . .	Lactate Dehydrogenase
LR	. . . . .	Logistic Regression
MAE	. . . . .	Mean Absolute Error
ML	. . . . .	Machine Learning
MRAs	. . . . .	Mineralocorticoid-Receptor-Antagonists
NLP	. . . . .	Natural Language Processing
NRBC	. . . . .	Nucleated Red blood cell

OLS	. . . . .	Ordinary Least Square Regression
PHQ2	. . . . .	Patient Health Questionnaire-2
PHQ9	. . . . .	Patient Health Questionnaire-9
PMH	. . . . .	Past Medical History
PVE	. . . . .	Proportion of Variance Explained
RDW	. . . . .	Red Blood Cell Distribution Width
RF	. . . . .	Random Forests
RFE	. . . . .	Recursive Feature Elimination
RR	. . . . .	Relative Risk
RT-PCR	. . . . .	Reverse Transcription Polymerase Chain reaction
RTE	. . . . .	Relative Test Error
SARS-CoV-2	. . . . .	Severe Acute Respiratory Syndrome Coronavirus 2
SBP	. . . . .	Systolic Blood Pressure
SDoH	. . . . .	Social Determinants of Health
SFS	. . . . .	Statistical Feature Selection
SNR	. . . . .	Signal-to-Noise Ratio
SVM	. . . . .	Support Vector Machine
XGBoost	. . . . .	Gradient Boosted Decision Trees

Chapter 1

Introduction

The global healthcare system has been under significant strain, particularly highlighted by the pandemic, which exposed the shortages of medical resources and skilled clinicians. Hospitals and clinics have encountered unprecedented challenges, including overwhelming patient loads, shortage of medical supplies, and the need to rapidly adapt to new treatment protocols. These pressures have revealed existing weaknesses in healthcare infrastructure and underscored the necessity for more robust and adaptable systems [Ranney Megan L. et al., 2020, Chopra et al., 2020].

1.1 A Context of Data-Driven Methods in Healthcare

Artificial Intelligence (AI) and Machine Learning (ML) methods have emerged as crucial tools in addressing these challenges by facilitating the development of more efficient and responsive healthcare systems. ML can quickly and accurately analyze vast amounts of data, aiding in predicting patient outcomes, optimizing resource allocation, and personalizing treatment plans. These capabilities are essential for managing current and future health crises, enabling healthcare providers to make data-driven decisions that enhance care quality and operational efficiency [Esteva et al., 2021, Beheshti et al., 2022]. Integrating ML into healthcare practices can lead to more effective disease surveillance, early detection of outbreaks, and improved patient management, thereby strengthening the resilience of healthcare systems against ongoing and future challenges [Topol, 2019, Rajkomar et al., 2019].

Our work aims to support the healthcare industry by enabling efficient resource allocation and staffing decisions through algorithmic, data-driven approaches. The increasing availability of Electronic Health Records (EHRs) has facilitated the use of ML techniques in healthcare, enhancing the accuracy of intelligent decision-making processes in both primary and tertiary patient care [Ludwick and Doucette, 2009].

This dissertation has three primary objectives. First, we aim to develop prognostic models to predict patient outcomes for specific conditions, enabling better health management and guiding the allocation of medical resources, ultimately reducing overall healthcare costs. Second, we plan to design an ML algorithmic framework to enhance and refine clinicians' prescriptions. These models can provide personalized prescriptions, improving treatment quality and efficiency. Third, we will focus on the credibility and interpretability of data-driven models used in healthcare. Research emphasizes the importance of aligning a model's predictions with established medical knowledge to foster trust and acceptance among practitioners [Wang et al., 2018, Yoon et al., 2022]. By introducing a special norm within the context of Distributionally Robust Optimization (DRO) formulation, our resulting model will balance interpretability and credibility, ensuring robustness and alignment with expert insights.

In this dissertation, four studies have been included: developing prognostic model to identify not well-controlled hypertension; designing an algorithmic framework for personalized hypertension prescription; developing predictive models for COVID-19 outcomes; and Distributionally Robust Optimization in Credible Learning. All of the models are trained and evaluated using a large dataset from a safety-net hospital, which provides access to extensive clinical information and includes a substantial proportion of racially and ethnically diverse patients. This ensures broad applicability and generalization of the models in practice.

1.2 Accounting for Racial Bias and Social Determinants of Health in a Model of Hypertension Control

Hypertension is one of the most prevalent chronic cardiovascular disorders. While nearly half of the adults with hypertension in the U.S. have Systolic Blood Pressure (SBP) over 140 mmHg, only 22.5% of them have their hypertension under control [CDC, 2024a]. Lack of control can lead to adverse outcomes, such as ischemic heart disease [Grassi et al., 2011], stroke, and heart failure [Chobanian, 2003]. Severely hypertensive patients typically need long-term medical care with significant cost implications.

Further, it is hard to ignore longstanding health inequity issues of hypertension. There is evidence linking high blood pressure and Social Determinants of Health (SDoH), including income measures [Leng et al., 2015, Mills et al., 2016, Mensah et al., 2005]. Multiple studies have found that Black people have higher incidence of hypertension than any other racial group in the U.S. [Hajjar and Kotchen, 2003, Kramer et al., 2004, Carnethon et al., 2017]. The confluence of all these factors creates persistent health disparities, impacting hypertension control. It becomes important to identify severe hypertension and target patients with a more urgent need of better BP management.

This study developed a prognostic model to identify poorly controlled hypertension incidences. It utilized only non-clinical variables for easy implementation and access. Sociodemographic factors like race and household income are found to be predictive to the outcome. Racial bias analysis has been conducted between Black patients and White patients. Clinical predictive models have been adjusted to mitigate the existing racial disparity in the outcome prediction. We identified a host of structural SDoH factors that can be targeted to help mitigate disparities in hypertension control.

1.3 Personalized Hypertension Treatment Recommended by a Data-Driven Model

Following the successful identification of patients with uncontrolled hypertension, the next step will be effective management of hypertension control issue. The major causes of uncontrolled hypertension are complicated, since inappropriate or inadequate prescriptions, patient noncompliance, and high therapy cost can all lead to treatment failure [Oparil and Calhoun, 1998]. Unlike other medical conditions, therapy for hypertension is easily influenced by peripheral components in the biological network [Melville and Byrd, 2019]. Genetic predisposition, physiological systems, and time varying environmental factors all play important roles in the pathophysiology of hypertension [Mozaffarian et al., 2016].

Despite the consensus that different patient characteristics and distinct underlying mechanisms of high blood pressure would lead to differential responses towards antihypertensive drugs, personalized hypertension treatment has not been widely adopted in clinical practice. Most clinicians are still following evidence-based clinical guidelines that utilize more generalized standard therapies [Melville and Byrd, 2019]. There is a need to incorporate various considerations in clinical decision making, such as medication interactions, comorbidities, white coat effects, obesity, pseudohypertension, and personal lifestyle choices including food and alcohol intake [Savoia et al., 2017, Calhoun et al., 2008]. However, the unknown underlying traits for each patient remain elusive to decipher and account for simply by manually integrating various data sources. Consequently, precise personalized treatment is difficult to attain given the considerable variety of antihypertensive drug options. Several existing studies that model hypertension prescriptions are either restricted to special settings and do not offer full personalization [Backenroth et al., 2017] or are limited to certain hypertension prescription types [Bertsimas et al., 2021].

This study proposed a personalization approach for hypertension treatment based on ML algorithms, seeking to maximize the effectiveness of hypertensive medications at the

individual level.

1.4 Development and Validation of Predictive Models for COVID-19 Outcomes in a Diverse, Safety-net Hospital Population

As a result of the SARS-CoV-2 pandemic, many hospitals across the world have resorted to drastic measures: canceling elective procedures, switching to remote patient consultations, designating most beds to COVID-19 care, expanding Intensive Care Unit (ICU) capacity, and re-purposing doctors and nurses to support the clinic and ICU function. In the US alone, the CDC estimated more than 7.5 million COVID-19 hospitalizations and 92,100 total deaths during the period of February 2020 to September 2021 [CDC, 2020].

As of 2024, about 69.5% of the U.S. population has been fully vaccinated against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and 22.5% of them have received an updated 2023-24 vaccine as of May 11 2024 [CDC, 2024b, Franklin, 2024]. However, vaccination rates remain uneven across different racial/ethnic groups and between rural and urban communities. For instance, vaccination coverage is lower among adults residing in rural areas (17.9%) compared to those in urban areas (24.0%) [Murthy, 2021]. Internationally, vaccination rates also vary significantly. In Nigeria, approximately 20.4% of the population has been fully vaccinated, while in Afghanistan, the rate is around 27.1% [Franklin, 2024]. These disparities highlight the ongoing challenge of achieving equitable vaccine distribution globally. The uneven vaccination rates and the emergence of new COVID-19 variants suggest that the pandemic will continue to pose a significant concern for health systems worldwide [Peiffer-Smadja et al., 2020].

Making predictions about disease severity is important in clinical triage, resource allocation, staffing, and overall planning, within a hospital system, and at the state/country scale. Artificial Intelligence (AI) methods have been used to that end, [Hilton et al., 2020] including the prediction of patient outcomes for COVID-19 [Liang et al., 2020, Hao et al.,

2020, Yan et al., 2020, Ji et al., 2020, Wang et al., 2020, Gong et al., 2020, Bhargava et al., 2020]. However, these studies used data from relatively few patients (the largest used 2,5006 [Liang et al., 2020]) and a limited collection of pre-existing conditions, laboratories, and in-hospital data.

In our study we developed strong predictive models of hospitalization, ICU care, mechanical ventilation, and mortality among all patients (n=7,102) who tested positive for SARS-CoV-2 at the Boston Medical Center (BMC) between January 1st and December 31st, 2020. We derived parsimonious versions of the models with strong predictive power that are interpretable and easy to implement.

1.5 Distributionally Robust Optimization in Credible Learning

Globally, the healthcare industry is experiencing a rising trend towards utilizing data-driven models as tools for decision-making support [Sutton et al., 2020, Kononenko, 2001, Beheshti et al., 2022]. These models are vital for guiding clinical decisions, with their effectiveness in the healthcare community. However, such effectiveness is highly dependent on professionals' clear understanding of the models' predictive mechanisms and rationales [ElShawi et al., 2021, Zhang et al., 2022].

While Deep learning (DL) approaches are superior to traditional approaches in handling images and text data at disposal [Shomirov and Zhang, 2021, Savova et al., 2019, Zhou and Hripcsak, 2007], it is hard to interpret and rationalize the results. Research indicates that clinicians often struggle to understand and trust complex models like neural networks, favoring more interpretable models like linear regression and decision trees [London, 2019, ElShawi et al., 2021].

However, mere interpretability does not suffice for the widespread integration of ML models into clinical practice. As research underscores, it is crucial for the logic behind a model's predictions to align, at least to some extent, with established medical knowl-

edge [Wang et al., 2018, Yoon et al., 2022]. This alignment fosters trust and acceptance among practitioners. A model's success in healthcare hinges on not just its accuracy and interpretability but also its alignment with medical expertise and credibility.

It is important to note that a model's interpretability does not inherently guarantee its credibility. There often exists a delicate balance between interpretability and credibility. While some regularization techniques can lead to sparsity in feature selection and result in feature interpretability, like the l_1 norm [Tibshirani, 1996], it may trade off the predictive performance and risk prioritizing some unimportant features [Wang et al., 2018, Zhao and Yu, 2006]. Similarly, sole reliance on domain knowledge negates the essence of data-driven methodologies and can lead to suboptimal performance. As a result, we need a model that validates its conclusions with existing expertise while ensuring accuracy is not sacrificed.

In this study, we will show that by applying the Wasserstein DRO [**?**] formulation in the context of credible learning, we can recover the Group Least Absolute Shrinkage and Selection Operator (GLASSO) penalty [Yuan and Lin, 2006] and enhance the interpretability and credibility of the model. The Distributionally Robust Optimization (DRO) approach effectively reduces the influence of outliers by considering a family of probability distributions over the observed data, some of which assign minimal probabilities to outliers. It aims to minimize the worst-case absolute loss within a distributional ambiguity set, defined using the Wasserstein metric [Mohajerin Esfahani and Kuhn, 2018, Gao and Kleywegt, 2022], encompassing all potential perturbations in the data distribution. This robustness is achieved by hedging against this diverse family of distributions, ensuring resilience against variations and outliers in the data. Domain expertise is employed to identify a group of covariates, which aids the model in discerning among highly correlated features while promoting sparsity in the selection process. By introducing a special norm $\|\cdot\|_{2,\infty}$, the resulting model could achieve a balance between interpretability and credibility, ensuring that the models are not only robust and understandable but also aligned with expert insights. We further

adapt this norm for scenarios with overlapping group structures, which is more typical in medical applications. Our work is essential for fostering trust and acceptance in healthcare applications, as it integrates expert knowledge with advanced machine learning techniques. Our empirical evidence, drawn from both synthetic and real-world datasets, demonstrates the superiority of the DRO-expert norm in overlapping scenarios. This approach effectively merges model performance with expert knowledge, excelling in both Area Under Curve (AUC) and F1-score metrics.

1.6 Contributions of the Thesis

1.6.1 Enhancing Interpretability and Credibility of Data-Driven Models in Healthcare

In the healthcare system, the interpretability and credibility of data-driven models are critical factors influencing their adoption and effectiveness [ElShawi et al., 2021, Yoon et al., 2022]. Interpretability refers to the extent to which human users can understand the model's predictions and underlying mechanisms [Lipton, 2017, ElShawi et al., 2021, Zhang et al., 2022]. Models with high interpretability, such as linear models, allow clinicians to trace and comprehend the relationship between input variables and outcomes, facilitating more informed and transparent decision-making processes.

Credibility pertains to the degree of confidence that users have in a model's predictions and its overall reliability [Wang et al., 2018, Lipton, 2017]. It is closely correlated with interpretability; when a model is interpretable, clinicians can better understand its functioning and rationale, which enhances their trust in its predictions. Conversely, models that lack interpretability often face credibility issues, as clinicians find it challenging to validate and justify their use in patient care. This skepticism can lead to hesitation in integrating advanced predictive analytics into routine practice. Ensuring that models are interpretable and credible is essential for gaining the confidence of healthcare professionals and ensuring

the ethical application of data-driven technologies [Yoon et al., 2022, Rudin, 2019].

There is also a notable gap between the requirements of interpretability. The technical definitions of interpretability given by model designers often include complex mathematical formulations and visualizations, which are not easily comprehensible to non-experts like clinicians. This discrepancy can hinder the effective implementation of advanced models in clinical practice [Lipton, 2017, Rudin, 2019].

In this dissertation, we focus on designing interpretable and credible models that can be readily understood and adopted by clinical practitioners in the healthcare system. By quantifying model interpretability through sparsity levels and measuring credibility by assessing the percentage of non-sparse features identified by the model that align with established clinical expert knowledge, we have demonstrated that our proposed method using Distributionally Robust Optimization (DRO) achieves enhanced interpretability and credibility without compromising predictive accuracy. This approach ensures that the models remain robust and understandable while aligning closely with expert insights. This ensures that the models are not only robust and understandable but also aligned with expert insights, facilitating their integration into clinical practice.

1.6.2 Optimizing Linear Prognostic Models for Robust and Interpretable Healthcare Predictions

Due to the superior interpretability and ease of use offered by linear models, clinicians frequently favor simple linear models over non-linear or Deep Learning (DL) models [ElShawi et al., 2021, Zhang et al., 2022]. Linear models provide a straightforward and transparent relationship between input variables and outcomes, making it easier for clinicians to understand and trust the predictions. While linear models are appreciated for their simplicity and interpretability, they often fall short in performance compared to more complex models, particularly in capturing intricate patterns and relationships within the data [Perez et al., 2022].

In this dissertation, we investigated the design of linear prognostic models capable of achieving performance comparable to that of complex non-linear models. Our objective was to create parsimonious models, utilizing only the top 10 to 20 important features for health outcome prediction, that can be easily implemented in healthcare systems and interpreted by practitioners. The techniques employed included feature engineering, statistical and recursive feature elimination to select significant features, and the application of distributionally robust optimization (DRO) formulation to enhance model robustness. Additionally, we utilized advanced deep learning methods such as Long Short-Term Memory (LSTM) networks to capture the dynamic evolution of features and generate new feature scores for our linear models. Our study demonstrated that these designed linear predictive models can produce results comparable to those of non-linear models like Random Forest (RF) and Gradient Boosted Decision Trees (XGBoost), as well as to similar work using DL methods.

1.6.3 Evolution from Predictive Models to Prescriptive Models

ML methodologies have been extensively utilized for predictive tasks aimed at forecasting patient outcomes based on historical data [Parikh et al., 2008, Paynter et al., 2009, Liang et al., 2020, Hao et al., 2020, Yan et al., 2020, Ji et al., 2020]. These models, while useful, were limited to providing insights without actionable recommendations. In this dissertation, we have shifted the focus towards prescriptive models, which not only predict outcomes but also suggest optimal interventions to achieve desired results. This shift has been driven by the integration of more complex algorithms and data processing, enabling a more proactive approach to patient care. More importantly, our prescriptive models have been evaluated by clinicians for their credibility and have been selected by the Boston Medical Center (BMC) for an ongoing clinical trial. This recognition demonstrates the performance and reliability of our work, highlighting its acceptance by the clinical community and its potential for immediate integration into the healthcare system.

1.7 Notational Conventions

We use boldfaced lowercase letters to represent vectors, regular lowercase letters to denote scalars, boldfaced uppercase letters to signify matrices, and calligraphic capital letters to indicate sets. Vectors are column vectors unless otherwise specified. Let $\mathbf{x}$ denote the column vector $\mathbf{x} = (x_1, \ldots, x_{\dim(\mathbf{x})})$, where $\dim(\mathbf{x})$ is the dimension of $\mathbf{x}$. $\mathbb{E}$ denotes the expectation and $\mathbb{P}$ probability of an event. $\mathbb{P}(\mathcal{Z})$ denotes the set of probability measures supported on $\mathcal{Z}$. The prime symbol ($'$) denotes the transpose of a matrix or vector. The p-norm of a vector $\mathbf{x}$, denoted by $\|\mathbf{x}\|_p$, is defined as $(\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$, where $p \geq 1$, and $\|\cdot\|$ for the general vector norm that satisfies the following properties:

1. $\|\mathbf{x}\| = 0$ implies $\mathbf{x} = \mathbf{0}$;
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$, for any scalar α;
3. $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$;
4. $\|\mathbf{x}\| = \||\mathbf{x}|\|$, where $|\mathbf{x}| = (|x_1|, \ldots, |x_{\dim(\mathbf{x})}|)$.

For a vector $\mathbf{x}$, the infinity norm, denoted as $\|\mathbf{x}\|_\infty$, is given by $\|\mathbf{x}\|_\infty = \max(|x_1|, |x_2|, \ldots, |x_{dim(\mathbf{x})}|)$. The zero norm of a vector $\|\mathbf{x}\|_0$ counts the number of nonzero elements in the vector. For a matrix $\mathbf{P}$, $\|\mathbf{P}\|_\infty$ denotes the maximum absolute row sum. We use $\emptyset$ to denote the empty set, $[N]$ for the set $\{1, \ldots, N\}$, and $|S|$ for the cardinality of the set S.

1.8 Outline

This dissertation is organized as follows. Chapter 2 introduces work about predicting uncontrolled hypertension. Chapter 3 introduces work on a personalized hypertension prescription recommendation model. Chapter 4 introduces the work on COVID-19 related predictive

models. Chapter 5 introduces the Wasserstein DRO formulation inducing GLASSO algorithm, and demonstrating their enhanced interpretability and credibility. Chapter 6 concludes the work and outlines the future work.

Chapter 2

Accounting for Racial Bias and Social Determinants of Health in a Model of Hypertension Control

2.1 Introduction

Most of the current hypertension predictive models focus on predicting the emergence of hypertension [Parikh et al., 2008, Paynter et al., 2009, Teixeira et al., 2017, Muntner et al., 2010]. Less attention has been given to predicting poorly managed hypertension. Several of the existing prognostic models use clinical variables and laboratory results to improve prediction accuracy [Teixeira et al., 2017, Ye et al., 2020]. A potential drawback to using such variables is that model applicability is restricted to resource-rich settings, where lack of good hypertension control is less prevalent. Also, models with more features and higher complexity may become difficult to implement in clinical practice [John et al., 2022]. The goal of the present study is to develop models that rely on readily available and self-reported sociodemographic information and assess whether such information can predict poor hypertension management and enable targeted interventions.

An additional goal is to elucidate the role of racial and SDoH variables [de la Vega et al., 2019]. Whereas these factors are important to study, to the best of our knowledge no work has examined whether existing disparities affect predictive models and introduce biases in the way predictions are made. Furthermore, no work has considered an array of SDoH and sought to understand their relative importance in making predictions to inform subsequent

support programs that promote health equity.

This chapter develops an easily implementable and interpretable predictive model of very high blood pressure among hypertensive patients, based only on demographic features and socioeconomic factors. We defined very high blood pressure as SBP exceeding 160 mmHg. Historically, that was the threshold for Stage-2 hypertension, but it has been lowered to 140 mmHg with the most recent guidelines [Whelton et al., 2018]. We opted for not using clinical variables so that the model could be used for most individuals, even those with a limited clinical record in a health care system. Using such predictions could enable preventive actions that could range from contact with a care management professional, inviting the patient in the clinic for a hypertension consultation, to broader efforts in partnership with community support groups to address patients' individual social needs.

2.2 Materials and Methods

2.2.1 Cohort Design

We extracted data from Boston Medical Center (BMC) Electronic Health Records (EHRs) from January 1, 2012 to January 1, 2020. The dataset included all patients who satisfied one or more of the following conditions: (1) patients who had a hypertension diagnosis; (2) patients with high blood pressure in their problem list; and (3) patients with at least two recorded SBP measurement exceeding 130 mmHg or a Diastolic Blood Pressure (DBP) measurement exceeding 90 mmHg.

In total, the dataset included 164,041 patients, with 85,924 female (52.38%) and 78,110 male (47.62%). Features extracted included: age, sex, race, language, marital status, SDoH factors, depression scale, cigarette use, and ZIP code. SDoH factors were extracted from the THRIVE survey, a custom screening program created by BMC which surveys patients on their unmet social needs in eight different domains: transportation, ability to secure caregiving for family members, ability to pay for utilities, education, food, housing,

employment issues, and ability to pay for medications [de la Vega et al., 2019]. All records were de-identified before analysis. The study was approved by the Boston University Medical Campus and Boston University Institutional Review Boards.

2.2.2 Pre-processing

We extracted all patient answers for the THRIVE survey and created a binary indicator variable for each of the 8 domains; specifically, a value of '1' implies that the patient reports the corresponding social need and a value of '0' otherwise. We combined and encoded the patients' depression test scores into a mental health indicator variable as follows. If a patient had a record of a Patient Health Questionnaire-2 (PHQ2) score larger than 2 or a Patient Health Questionnaire-9 (PHQ9) score larger than 4, their 'Depression' feature was recorded as '1', otherwise it was set to '0'. These thresholds on the PHQ2/PHQ9 scores are consistent with the scoring system originally described by Spitzer et al [Spitzer et al., 1999], All categorical features were then encoded into indicator variables for each category (in what is often referred to as 'one-hot' encoding). Features were standardized by subtracting their mean and dividing with their standard deviation. We also estimated the median household income and the distance to BMC for each patient based on the provided ZIP code. This was based on the USA ZIP code database 'uszipcode' [noa,], which utilizes up-to-date census information. Median values were used to impute the missing values for each numerical feature. These pre-processing procedures yielded 40 features for each patient.

2.2.3 Models and Metrics

We developed machine-learning models to predict whether a patient with hypertension has at least one SBP record exceeding 160 mmHg. Predictor variables include: age, sex, race, language, marital status, estimated median household income, distance to BMC, depression scale, cigarette use, and SDoH variables.

We trained both nonlinear models and linear models. Nonlinear ensemble methods

included Random Forests (RF) [Breiman, 2001] and gradient boosted decision trees (XGBoost). [Chen and Guestrin, 2016] Both typically provide excellent performance but do not yield interpretable models as the classifier may combine hundreds of decision trees. Linear models included Support Vector Machine (SVM) [Cortes and Vapnik, 1995] and Logistic Regression (LR); [Hastie et al., 2001] both yield interpretable models and produce feature weights that can be used to elucidate the relative predictive power of different features. Features with higher absolute coefficient values can be viewed as more significant in predicting high blood pressure. The sign of the coefficient indicates whether the corresponding feature is positively or negatively correlated with the outcome. Regularization (with l_1 or l_2 norm) was added to the linear models to boost the robustness of the model against outliers/noise [Chen and Paschalidis, 2020, Chen and Paschalidis, 2018] and avoid models that involve arbitrary linear combinations of highly correlated features that may limit interpretability. In addition to the coefficients, we also reported the p-value and the odds ratio for the top features with the highest absolute LR coefficients. The p-value was computed using a chi-squared test for categorical variables and a Kolmogorov-Smirnov (KS) test for continuous variables. The null hypothesis was that the distribution is the same in the two cohorts. A low p-value supports rejection of the null hypothesis, implying that the corresponding variable is statistically different in the cohort of those with high blood pressure compared to its complement.

To evaluate the models' ability to predict high blood pressure, the Area Under the Receiver Operating Characteristic Curve (AUC) and the weighted F1 score was calculated out-of-sample (i.e., in a test set not used for training the models). AUC can be interpreted as the probability that a randomly chosen sample from the positive class will score higher than a randomly chosen sample from the negative class. The F1 score is the harmonic mean of precision and recall. Precision (or positive predictive value) is defined as the ratio of true positives among all true and false positives. Recall (a.k.a. sensitivity) is the ratio of true

Table 2.1: Dataset summary statistics. The 'high blood pressure group' refers to patients with systolic blood pressure (SBP) over 160 mmHg. Patients who report SDoH refers to patients who indicate a social need in any of the 8 domains surveyed by Thrive. Lower income patients are those whose household income is below the median among all patients in the dataset.

Demographics	**Value**		**P-value**
	All patients	Patients with SBP record >160 mmHg M=59,306 (36.15% of N)	
Mean age	57.7	64.2	0
Median age	59	64	
Female	85,924 (52.38% of N)	31,677 (53.41% of M)	<0.001
Male	78,110 (47.62% of N)	27,628 (46.59% of M)	<.001
White Patients	56,218 (34.27% of N)	16,443 (27.73% of M)	0
Black Patients	59,353 (36.18% of N)	27,339 (46.10% of M)	0
Patients who report SDoH	9,317 (5.68% of N)	5,296 (8.93% of M)	0
Patients with lower income (<$53,798 per year)	65,967 (40.21% of N)	26,937 (45.42% of M)	<0.001

positives over true positives and false negatives. The weighted-F1 score is computed by weighting the F1-score of each class by the number of patients in that class. The prediction thresholds were optimized to achieve the best weighted-F1 score.

The data were randomly split into training (80%) and test sets (20%). Algorithm parameters were optimized on the training set using 5-fold cross-validation, and performance metrics were computed on the test set. We report the mean and standard deviation of the test performance metrics over 5 random splits.

2.3 Results

Among 164,041 patients in our dataset, 59,306 (36.15%) have a record of SBP over 160 mmHg ("very high blood pressure"). Table 2.1 provides basic statistics on the entire dataset and on specific groups identified.

There were 10 different races considered by the predictive model, with the majority

being Black or White patients. Black patients comprised 36.18% of all patients but their percentage in the very high SBP cohort was 46.10%. White patients comprise 34.27% of all patients but only 27.73% of the very high SBP group. The computed p-value (chi-square test) compares the distribution of Black/White patients in the two cohorts (patients with no SBP record $>$ 160 mmHg vs. patients with SBP record $>$ 160 mmHg). The p-values were 0 for both variables, which does refute the null hypothesis (the same distribution in the two cohorts).

Similar disparities appear in patients who reported SDoH in the THRIVE survey. Specifically, 5.68% of the patients reported SDoH among all patients but the percentage rose to 8.93% among the very high SBP group. Specifically, 56.84% of those with reported SDoH exhibit very high SBP. The p-value associated with SDoH among the two cohorts was 0, indicating statistical significance.

The estimated median household income of all patients in the dataset was $53,798 U.S. dollars per year. 40.21% of all patients had an estimated household income strictly lower than the median, with that percentage rising to 45.42% among the very high SBP group. The p-value also indicates statistical significance of the influence of household income.

Our algorithms predict who has very high SBP ($>$160 mmHg) reached 74.49%±0.23% AUC for nonlinear models (XGBoost) and 73.42%±0.23% AUC for linear models. Additional performance metrics are listed in Table 2.2.

Table 2.3 reports the top 20 variables, ranked according to their corresponding absolute LR model coefficient. In accordance with the mean age difference observed between all patients and the ones in the high SBP cohort (57.7 vs. 64.2), the most important variable is 'Age'.

The second top feature is cigarette smoking history with positive coefficient 0.41. Smokers had 5.06 times the odds (CI: 4.89—5.22) of exhibiting very high SBP relative to nonsmokers, indicating strong association between smoking status and very high blood

Table 2.2: High SBP record over 160 mmHg predictive model: validation performance metrics and top variables. SD refers to the standard deviation of the corresponding metric. LR-L2 and SVM-L2 refer to the l_2-norm regularized LR and SVM models.

Algorithm	**AUC**		**weighted F1-score**	
	Mean	**SD**	**Mean**	**SD**
LR-L2	73.42%	0.23%	69.17%	0.24%
SVM-L2	73.40%	0.23%	69.14%	0.19%
XGBoost	74.49%	0.23%	70.07%	0.20%
RF	73.53%	0.24%	69.27%	0.24%

Table 2.3: The top 20 features of the predictive model are listed and ranked by the absolute LR coefficients (Coef). We also listed the p-value, correlation of the variable with the outcome (Y-corr), the mean of the variable (Y1-mean) in the patient cohort with high SBP records over 160 mmHg, and the mean of the variable (Y0-mean) in the patients with no SBP record above 160 mmHg. We report the corresponding odds ratios (OR) and their 95% confidence intervals.

Top 20 Variables in the LR model									
	Variable	**Coef**	**p-value**	**Y1-mean**	**Y0-mean**	**Y-corr**	**OR**	**95% OR CI**	**5% OR CI**
1	Age	0.67	0	64.2	54.02	0.28	1.04	1.04	1.04
2	Ever Cigarette User-YES	0.41	0	0.23	0.05	0.26	5.06	5.22	4.89
3	Race_Black	0.21	0	0.46	0.31	0.16	1.94	1.98	1.9
4	Race_White	-0.15	0	0.28	0.38	-0.1	0.63	0.64	0.61
5	Depression	0.09	<0.001	0.09	0.05	0.07	1.81	1.88	1.74
6	Language_English	-0.09	<0.001	0.67	0.71	-0.04	0.85	0.87	0.84
7	Race_Hispanic	-0.09	<0.001	0.01	0.02	-0.05	0.42	0.46	0.38
8	Marital_status:other	-0.07	<0.001	0.03	0.05	-0.04	0.66	0.69	0.62
9	Race_Other	-0.07	<0.001	0.01	0.02	-0.03	0.48	0.53	0.43
10	Household_income	-0.06	<0.001	55586.96	58078.01	-0.06	0.99	0.99	0.99
11	SDoH_transportation	0.05	<0.001	0.03	0.01	0.08	3.46	3.76	3.19
12	Race_Asian	-0.05	<0.001	0.03	0.03	-0.03	0.72	0.77	0.68
13	SDoH_food	0.05	<0.001	0.05	0.02	0.08	2.47	2.62	2.34
14	Race_Middle Eastern	-0.05	<0.001	0	0	-0.02	0.37	0.48	0.29
15	Language_Other	-0.04	<0.001	0.18	0.17	0.02	1.12	1.15	1.09
16	Language_African	0.03	<0.001	0.06	0.03	0.06	1.76	1.85	1.68
17	SDoH_housing	0.03	<0.001	0.01	0	0.05	3.05	3.44	2.7
18	Marital_status:divorced	0.03	<0.001	0.07	0.05	0.04	1.48	1.54	1.42
19	Marital_status:separated	0.02	<0.001	0.03	0.02	0.04	1.68	1.79	1.58
20	SDoH_job	0.02	<0.001	0.02	0.01	0.05	2.09	2.27	1.93

pressure.

Race is also significant, with 'Race_Black' being the third most important feature and 'Race_White' the fourth; each with LR coefficients 0.21 (p-value 0) and -0.15 (p-value 0). Being Black has an associated Odds Ratio (OR) of 1.94 (CI: 1.90—1.98), implying the ratio of the odds for very high SBP versus not is almost twice as high in Black patients compared to other races.

Mental health, and specifically 'Depression', plays a significant role in the model with a coefficient equal to 0.09. The corresponding OR is 1.81 (CI: 1.74—1.88). Among the eight domains of SDoH variables, 'transportation' and 'food' were the most significant with the same coefficient 0.05 (p-value$<$0.001 respectively). Additional SDoH variables included housing and employment needs, both inducing a higher likelihood of very high SBP. Further, 'Household_income' also appeared in the top contributing variables with coefficient -0.06 (p-value $<$ 0.001).

The analysis revealed disparities in very high blood pressure between Black and White patients. In the very high SBP cohort, 46.10% (27,339 out of 59,306) are Black and only 27.73% (16,443 out of 59,306) are White. As shown in Table 2.3, 'Race_Black' and 'Race_White' are top 5 predictive features. A more detailed comparison between Black and White patients is given in Table 2.4. The incidence of very high SBP among Black patients is significantly higher than among White patients. Similarly, Black patients have a larger fraction of lower income patients and a higher percentage of patients with SDoH needs. However, Black patients live, on average, a shorter distance to BMC. This may be influenced by patients experiencing homelessness who provide a BMC address and ZIP code. This explanation is also consistent with the fact that 'Distance_to_BMC' has a negative coefficient in our models.

To further elucidate racial differences, we trained models on Black and White patients separately. The interesting finding is that for both AUC and weighted F1 score, the perfor-

Table 2.4: Comparison between Black and White patients. Percentages in the entire dataset are calculated as a fraction of the total number of patients (N=164,041). The population in the high SBP cohort consists of Black or White patients with SBP records exceeding 160 mmHg and the percentages are computed as fractions of M=59,306, which is the size of the high SBP cohort. The population below the median household income level consists of Black or White patients who have household income below the overall annual median level in the entire dataset ($53,798); the corresponding percentages are computed as fractions of Black (respectively, White) patients in the entire dataset. The population with SDoH consists of patients who answered 'Yes' in any of the 8 domains of the Thrive survey, and percentages are calculated as fractions of Black (respectively, White) patients in the entire dataset.

	Black patients	**White patients**
Population in the entire dataset	59,353 (36.18%)	56,218 (34.27%)
Population in the high SBP cohort	27,339 (46.10%)	16,443 (27.73%)
Population below the median household income level	30,537 (51.45%)	12,628 (22.46%)
Population with SDoH	5,941 (10.01%)	1,215 (2.16%)
Average distance to BMC (miles)	16.99	47.25

mance of the model for Black patients is higher. Specifically, the model for Black patients achieves AUC of 74.07% whereas the corresponding model for White patients achieves AUC of 70.11%. More detailed results are in Table 2.5.

Table 2.5 indicates that when we use the standard decision threshold on the predicted probability of very high SBP (i.e., patients with a predicted probability above 0.5 are classified in the high SBP group), the corresponding false positive rates (FPR) and false negative rates (FNR) are substantially different between the two models. In particular, the FPR is 27.71% for Black patients and 5.92% for White patients (a factor of 4.7). Correspondingly, the FNR for Black patients is 43.11% compared to 73.82% for White patients (a factor of 1.7). We also computed a so called 'Treatment Equality' metric, defined as the ratio FPR/FNR, [Caton and Haas, 2020] to evaluate the model's racial disparity. Having this metric equalized among the two cohorts is considered desirable. Under the standard 0.5 prediction threshold, FPR/FNR for Black patients is 8 times higher than the ratio for White patients. These statistics demonstrate that the models are biased and are more likely to classify Black patients as belonging to the very high SBP group. In essence,

Table 2.5: High SBP prediction models for Black and White patients trained separately: a comparison of validation performance metrics. LR-L2 refers to the l_2-norm regularized LR model. FNR and FPR refer to the model's false positive and false negative rate when the decision threshold (on the predicted probability of high SBP) is set to 0.5.

LR-L2	Black patients	White patients
AUC	74.07%	70.11%
Weighted F1-score	67.51%	68.11%
FPR (Threshold=0.5)	27.71%	5.92%
FNR (Threshold=0.5)	43.11%	73.82%
FPR/FNR (Threshold=0.5)	0.64	0.08

the models "pick up" and exploit racial disparities present in the data.

To resolve the algorithmic racial disparity, we modified the decision thresholds used in the two models. Clearly there is a trade-off between FPR and FNR; higher FPR implies lower FNR and vice versa. One way to resolve the bias is to equate the FNR rates. Arguably, false negatives are more critical to hypertension control because they lead to worse longer-term outcomes and, presumably, greater burden associated with long-term poorly controlled hypertension. On the other hand, false positives only imply providing more support to patients who may not need it. The associated resource consumption is likely to be lower costs of primary prevention and care versus the consequences of longer-term poorly controlled disease.

We elected to set the FNR for either model to 25%, which is attained by a threshold for Black patients equal to 0.397 and a corresponding threshold for White patients equal to 0.245. Table 2.6 shows the performance metrics for both models after modifying the decision thresholds. The resulting FPR of the model for Black patients becomes 47.17% and the FPR for White patient model becomes 47.34% (factor close to 1.0). Similarly, the treatment equality metric of FPR/FNR becomes equal to 1.89 for both cohorts.

Table 2.6: LR prediction thresholds used to tune the validation FNR of the separate models (set to 25%). Validation model performance metrics after tuning the thresholds to equate FNR.

Setting FNR to 25%	
	Prediction Thresholds
White Patients	0.245
Black Patients	0.397

	Black patients	**White patients**
Weighted F1	66.18%	60.85%
FPR	47.17%	47.34%
FPR/FNR	1.89	1.89

2.4 Discussion

The discriminatory power of our models is consistent with the performance of existing hypertension prediction models which use additional information, such as familial history of hypertension and blood pressure (BP) measurements [Parikh et al., 2008]. Our results, using a model that relies only on self-reported demographics and socioeconomic variables, outperform models that use BP measurements, laboratory results, and genetic variables (achieving an AUC of 66.4%) [Fava et al., 2013]. Other models achieve AUCs higher than 0.75 but incorporate multiple BP measurements over time [Kivimäki et al., 2010] and risk scores [Parikh et al., 2008]. Considering that we only used demographics and socioeconomic variables, the discriminatory power of our proposed model is noteworthy.

The importance of age is not surprising since numerous studies have shown such association, providing plausible physiological mechanisms [Dohi et al., 1990].

The harmful effects of cigarette smoking in hypertension control have been previously demonstrated. Moreover, hypertensive smokers have an increased risk of developing severe forms of hypertension which may result from accelerated atherosclerosis due to smoking [Virdis et al., 2010].

It is known that Black patients have significantly higher risk of hypertension than other racial groups [Hajjar and Kotchen, 2003]. Currently, there is no clear physiological explanation. The disparity may reflect the impact of systemic racism, with barriers such as limited access to care and higher incidence of SDoH driving more proximate factors

for hypertension control like diet, stress, and medication prescribing and adherence. Our study and several other studies indicate that the disparity in hypertension persists even after adjusting for individual socioeconomic factors [Mensah et al., 2005, Hertz et al., 2005].

We sought to understand whether there are discernible differences in the predictive variables that lead to differences between Black and White patients and shed some light into the etiology of these differences. Table 2.7 lists the top 20 predictive variables for each model (ranked by the absolute value of their corresponding LR coefficient). Other than age, cigarette use and depression, which are common, the model for Black patients emphasizes SDoH needs in transportation, food, and housing resources. Transportation needs may explain poorer control of hypertension as patients face barriers in physical access to care. Further, needing help to pay for food may correlate with consumption of low-cost, high-calorie, high-sodium fast food. Similarly, housing needs are a cause for stress and overall lower quality of life which may contribute to hypertension. Other neighborhood-level socioeconomic factors like annual household income and distance to BMC also play significant roles in the Black patient predictive model, and they are negatively correlated with the outcome. In particular, 51.45% of Black patients have median household income less than the overall median value in the dataset. This percentage is significantly higher than the average low-income percentage of 40.21% computed over all patients in the dataset. This finding confirms previous research on the association between neighborhood environment and hypertension [Morenoff et al., 2007, Mujahid et al., 2008]. Namely, living in a an area of high resource deprivation is often linked to various deleterious factors such as limited access to medical institutions, decreased neighborhood safety, and lower education level. All of these factors can in turn exacerbate health disparities. Considering the model for White patients, the top features contain marital status (divorced and separated increase the likelihood of high SBP) and language. Arguably, the top predictive variables for Black patients correspond to structural social needs (food, transportation), whereas top predictive

Table 2.7: Top 20 features with largest absolute LR coefficients (Coef) in the predictive models for Black and White patients, respectively.

	TOP 20 Predictive Variables for Black patients		TOP 20 Predictive Variables for White patients	
	Variable	**Coef**	**Variable**	**Coef**
1	Age	0.71	Age	0.56
2	Ever Cigarette User-YES	0.48	Ever Cigarette User-YES	0.41
3	Depression	0.10	Language_English	-0.19
4	SDoH_transportation	0.07	Depression	0.11
5	Marital_status:other	-0.06	Language_Other	-0.10
6	SDoH_food	0.06	Marital_status:other	-0.08
7	Household_income	-0.05	Marital_status:divorced	0.05
8	SDoH_housing	0.04	Household_income	-0.05
9	SDoH_caring	-0.03	Marital_status:married	-0.04
10	Language_African	0.03	Language_European	-0.04
11	SDoH_medication	0.02	Marital_status:separated	0.03
12	SDoH_utilities	0.02	SDoH_food	0.03
13	Distance_to_BMC	-0.02	SDoH_job	0.03
14	SDoH_education	0.02	SDoH_housing	0.03
15	Language_Spanish	-0.02	SDoH_transportation	0.02
16	Language_English	-0.02	SDoH_utilities	0.02
17	Language_Asian	-0.02	Sex_F	-0.02
18	Marital_status:single	-0.01	Sex_M	0.02
19	Marital_status:separated	0.01	Marital_status:single	-0.01
20	Marital_status:married	-0.01	Marital_status:widow	0.01

variables for White patients correspond to other types of demographic or personal issues.

2.5 Conclusion

This study developed an easy to implement, interpretable prognostic model for identifying patients with high SBP exceeding 160 mmHg among hypertensive or pre-hypertensive patients. The derived models exhibit strong predictive power based solely on non-clinical variables. The top predictive features identified are age, cigarette use, race, mental health, and SDoH factors. We found implicit racial disparity in the non-stratified predictive model. Race-specific models were trained for Black and White patients to investigate what underlies

this bias. While Black patients may have a greater prevalence of high blood pressure, their conditions were more significantly associated with structural SDoH variables corresponding to food, transportation, housing, and lower household income. This finding may facilitate the targeted intervention on specific SDoH factors for Black patients in need. Addressing the most important social needs for Black and White patients alike may improve the overall quality of care for patients with hypertension. Moreover, to resolve the algorithmic bias and equate false positive and false negative predictions across racial groups, one needs to select different decision thresholds for Black and White patients.

Predictive analytics of the type we developed can be used in various ways: (1) understanding the risk profiles can lead to better clinical decision making and more effective risk communication with patients; (2) tailoring prescriptions based not only on a BP measurement during a visit but also on the risk profile suggested by the model may lead to better hypertension control; (3) targeted interventions in partnership with national, state, and neighborhood support programs may address SDoH which are modifiable, leading to better hypertension control and lower medical care costs.

Chapter 3

Personalized Hypertension Treatment Recommended by a Data-Driven Model

3.1 Introduction

Being increasingly adopted in the cardiovascular field, ML is expected to facilitate the precision and personalization of hypertension therapies [Krittanawong et al., 2017]. However, there are several challenges for ML-driven personalization:

1. The sparse and incomplete nature of the EHRs [Weiskopf et al., 2013, Goldstein et al., 2017]. Integrating heterogenous data from different sources, the process of medical data collection is always complex. This usually leads to data 'outliers' caused by entry errors, incomplete information, and irregular lab tests, which can significantly reduce the accuracy of ML models. To the best of our knowledge no work on personalized treatment has dealt with 'outliers'.

2. To make personalized recommendations, an ML model cannot see into the future and can only rely on the prediction of counterfactual outcomes under each prescription. Only when the prescription recommended by the model is the same as the administered one, we have ground truth assessment of the outcome. There are several models built for personalized chronic disease treatment using a so-called contextual bandits approach [Zhu et al., 2018, Tewari and Murphy, 2017], which can learn from historical data of the same individual and make predictions. However, this methodology highly relies on frequent patient visits and rich historical information that is hard to find in

sparse EHRs. There is a need to make recommendations even for patients with sparse history.

3. Interpretability of the ML prescriptive model. While ML is superior to traditional approaches in handling large amounts of data at our disposal, ML adoption in health-care settings is limited by the lack of interpretability and comprehensiveness. Clinicians and patients cannot make medical decisions just based on a black-box model designed and evaluated in a narrower scenario, which may not have access to all information about a specific patient [Stiglic et al., 2020]. It is important for the decision makers to understand not just the mechanism of the algorithm, but how and why the model proposes each option. For instance, recent work has developed hypertension treatment prediction models using DL methods [Zhang et al., 2019, Ye et al., 2020]. While DL models often result in accurate predictions, it is hard to interpret and rationalize the results. Moreover, study also shows that for harmonized EHRs data, DL model did not show superior benefits compared with traditional methods [Mandair et al., 2020].

In this study, our objective is to design a model able to generate a personalized hypertension prescription based on each individual patient profile, including demographics, vital signs, past medical history, and clinical test records. To that end, we predict the future SBP by using the past outcomes of patients that had the most similar profiles. We adopt a robustified regression procedure to immunize predictions against outliers in the EHR data that could bias the result [Chen and Paschalidis, 2020]. Our algorithm is based on a regression-informed K-Nearest Neighbor (K-NN) [Cover and Hart, 1967] approach. Unlike other non-linear approaches such as DL or random forests [Breiman, 2001], this method is more reliable than a simple linear model but maintains interpretability and comprehensibility. The predictive power of the prescriptive model has been shown to satisfy certain theoretical guarantees [Chen and Paschalidis, 2018]. We incorporated a randomized prescriptive policy to add robustness to the model and only adopt the recommended prescription if the SBP

improvement over the previous regimen exceeds a certain threshold.

To help the prescribers interpret the model, we randomly sampled 350 cases, listed the recommended prescriptions, and the corresponding patient profiles. Moreover, we summarized affinity profiles built by the model. These cases were manually reviewed by clinicians at Boston Medical Center (BMC) to make sure that the prescriptive model is capable of maintaining a balance between contraindications and antihypertensive efficacy.

3.2 Materials and Methods

3.2.1 Dataset

Among the hypertensive patients' data we extracted in uncontrolled hypertension prediction work, we identified 42,752 patients who met the following criteria: 1) Patients have in the system at least 2 SBP measurements in 180 days (2 defined periods). 2) Received at least one type of cardiovascular medication, including Calcium Channel Blockers (CCBs), Thiazides, Angiotensin Receptor Blockers (ARBs), ACE inhibitors (ACEi), Beta-blockers, Loops, and Mineralocorticoid-Receptor-Antagonists (MRAs).

3.2.2 Features

For each qualified patient, we collected their demographics (age, sex, race, language, marital status, ZIP code), past blood pressure records, past medical history, vital signs, symptoms, medication history, laboratory tests results, cigarette usage, information for their social needs and depression test scores. Information on social needs was extracted from the 'THRIVE' survey program implemented at BMC. All records were de-identified before analysis. The study was approved by the Boston University Medical Campus and Boston University Institutional Review Boards.

3.2.3 Records Timeline Design

We split patients' treatment history based on visits to reflect the effect of prescriptions. Each 'visit' is defined to occur every 90 days up to the period that includes a patient's last recorded SBP measurement. The SBP measurements, vital signs and all other time-varying features were all captured and averaged over each visit (90-day period). The current prescription of each visit is the drug initiated at the current period. The future outcome is defined as the averaged SBP measurement, in mmHg 90 to 180 days after it. This effect period is chosen because previous studies have defined drug persistence as consistently refilling antihypertensive prescriptions in the subsequent clinic visits within 90-180 days of a previous dispensing [Bourgault et al., 2005, Mazzaglia et al., 2005, **?**]. Patient visits without valid future outcomes are dropped. In total, we obtained 432,096 valid visits, each with 144 features.

3.2.4 Prescriptive Policy Design

Our prescriptive model considers monotherapy of 7 types of antihypertensive drugs: Calcium Channel Blockers (CCBs), Thiazides, Angiotensin II Receptor Blockers (ARBs), ACE Inhibitors (ACEi), Beta-Blockers, Loops, Mineralocorticoid Receptor Antagonists (MRAs). We did not consider specific combination of drugs but instead grouping them all together as one option for the recommendation. This is because drug combinations are always involved with complex medical constraints and complications, and there are many options of different combination therapies for hypertension. Effective antihypertensive drug combination usually requires all the classes having different and complementary mechanisms of actions [Taddei, 2015].

The patient's current prescription is also included in the menu of options for the future prescription. Since frequently changed prescriptions can cause concerns about high health care costs and slow transient periods, we opted to maintain the current treatment where

possible. The new prescription will be considered only if the improvement on SBP control is over a pre-specified threshold.

To achieve best performance on future SBP reduction, the most straightforward idea is to choose the drug that yields the minimum predicted outcome. However, this approach may let potential prediction errors mislead clinicians and discard comparable alternatives that could be a better selection in some cases. As a result, we propose a new randomized policy to prescribe a drug with a probability inversely proportional to its exponentiated predicted SBP, which is aimed to improve the robustness of the model by exploration. It has been shown statistically that such randomized strategy can achieve nearly optimal future outcomes if an appropriate parameter is chosen [Chen and Paschalidis, 2018].

For the purpose of this analysis, a few common relative contradictions and disease specific preferences were incorporated in our prescriptive model. All the rules followed the professional suggestions from BMC clinicians and based on ACC/AHA hypertension treatment guidelines [Whelton et al., 2018]:

- If pulse$<$60, do not use Beta-Blockers;
- If potassium $>$4.5, do not use ACEi or ARBs;
- If creatinine of 2 or greater, do not use ACEi or ARBs or Thiazides;
- If potassium $>$5, do not use MRAs;
- If patient has diabetes, should likely be on an ACEi or ARBs;
- If patient has systolic heart failure, should likely be on a Beta-Blocker.

3.2.5 Models and Evaluation Scheme

We developed ML models to predict optimal prescriptions for patients with hypertension. A prediction-based prescriptive model is developed to analyze outcomes under each possible

medication. The goal is to find the treatment that minimizes future SBP based on the medical history of a group of similar patients. Our methods incorporate K-Nearest Neighbors (K-NN) on a regression-weighted metric, which can capture the similarity between patients' most predictive characteristics and model the future SBP accordingly. For example, if we want to estimate one patient's future SBP under thiazides, we will first train a regression model with all patient visits that were under thiazides to predict the future SBP. Then, the coefficients of the regression model are used to identify the importance of each feature in predicting the SBP outcome. The absolute values of coefficients form weights in measuring the distance between this patient and all the other patient visits that were under thiazides. Once the closest visits from the treatment group are identified, their average past outcomes under thiazides are used to estimate the potential SBP for the target patient under thiazides. The same procedure is repeated for all the prescription options included in the model.

Four algorithms were compared to test the best performance in future SBP reduction: Distributionally Robust Linear Regression (DRLR)-informed K-NN (the proposed model), Ordinary Least Square Regression (OLS)-informed K-NN, the Least Absolute Shrinkage and Regression Operator (LASSO) regression [Tibshirani, 1996], and Classification and Regression Trees (CART) [Breiman, 2001].

Under each prescription group, data was randomly split into a training set (80%), a validation set (10%), and a test set (10%). To avoid contamination of the validation/test sets with training data, visits from the same individual were included only into a single set. The predicted SBP reduction under each prescription type was compared with the true outcome under the standard of care and the counterfactual outcome corresponding to following the previous regimen.

DRLR Informed K-NN DRLR informed K-NN method was first formulated in [Chen and Paschalidis, 2019]. Given a feature vector $\mathbf{x} \in \mathbb{R}^p$, and a set of M available actions $[M]$, the goal is to predict the future outcome $y_m(\mathbf{x})$ under each possible action $m \in [M]$. Assume

the following relationship between the features and the outcome:

$$y_m = \mathbf{x}_m{}'\boldsymbol{\beta}_m^* + h_m(\mathbf{x}_m) + \varepsilon_m,$$

where prime donates transpose, $(\mathbf{x}_m, y_m)$ represents the feature-outcome pair of an individual taking action m; $\boldsymbol{\beta}_m^*$ is the coefficient that captures the linear trend; $h_m(\cdot)$ is an unknown Lipschitz continuous nonlinear function describing the nonlinear fluctuation in y_m, and ε_m is the noise term with zero mean and standard deviation η_m that expresses the intrinsic randomness of y_m and is assumed to be independent of $\mathbf{x}_m$.

Suppose for each $m \in [M]$, we observe N_m training samples $(\mathbf{x}_{mi}, y_{mi})$, $i = 1, \cdots, N_m$, that take action m. To estimate $\boldsymbol{\beta}_m^*$, we adopt the DRLR formulation developed in [Chen and Paschalidis, 2018]. The DRLR model minimizes the worst-case absolute loss within a distributional ambiguity set defined using the Wasserstein metric [Gao and Kleywegt, 2022, Mohajerin Esfahani and Kuhn, 2018] that contains all possible perturbations on the distribution of the data. The robustness is achieved by hedging against this family of distributions. The problem is formulated as:

$$\inf_{\beta_m} \sup_{\mathbb{Q}_m \in \Omega_m} \mathbb{E}^{\mathbb{Q}_m}[|y_m - \mathbf{x}_m{}'\boldsymbol{\beta}_m|], \tag{3.1}$$

where $\mathbb{Q}_m$ is the probability distribution of $(\mathbf{x}_m, y_m)$, belonging to some set Ω_m defined as:

$$\Omega_m \triangleq \mathbb{Q}_m \in \mathcal{M}(\mathcal{Z}_m) : W_1(\mathbb{Q}_m, \widehat{\mathbb{P}}_{N_m}) \leq r_m,$$

where $\mathcal{Z}_m$ is the set of all possible values for $(\mathbf{x}_m, y_m)$; $\mathcal{M}(\mathcal{Z}_m)$ is the space of probability distributions supported on $\mathcal{Z}_m$; $\widehat{\mathbb{P}}_{N_m}$ is the uniform empirical distribution on the N_m observed samples $(\mathbf{x}_{mi}, y_{mi}), i = 1, \cdots, N_m$; r_m is a parameter pre-specified to indicate the amount of ambiguity allowed. $W_1(\mathbb{Q}_m, \widehat{\mathbb{P}}_{N_m})$ is the order-1 Wasserstein distance between $\mathbb{Q}_m$ and $\widehat{\mathbb{P}}_{N_m}$ defined as:

$$W_1(\mathbb{Q}_m, \widehat{\mathbb{P}}_{N_m}) = \sup_{f \in \mathcal{L}} \left\{ \int_{\mathcal{Z}_m} f(\mathbf{z}_m)\mathbb{Q}_m(d\mathbf{z}_m) - \int_{\mathcal{Z}_{\Updownarrow}} f(\mathbf{z}_m)\widehat{\mathbb{P}}_{N_m}(d\mathbf{z}_m) \right\},$$

where $\mathbf{z}_m = (\mathbf{x}_m, y_m)$, and $\mathcal{L}$ is the space of all Lipschitz continuous functions satisfying :

$$|f(\mathbf{z}_{m1}) - f(\mathbf{z}_{m2})| \leq \|\mathbf{z}_{m1} - \mathbf{z}_{m2}\|_2, \forall \mathbf{z}_{m1}, \mathbf{z}_{m2} \in \mathcal{Z}_m.$$

Based on proof in [Chen and Paschalidis, 2018], with N_m independently and identically distributed samples $(\mathbf{x}_{mi}, y_{mi}), i = 1, \cdots, N_m$, the problem (3.1) can be reformulated as:

$$\inf_{\beta_m} \frac{1}{N_m} \sum_{i=1}^{N_m} |y_{mi} - \mathbf{x}_{mi}{}'\boldsymbol{\beta}_m| + r_m \|(-\boldsymbol{\beta}_m, 1)\|_2. \tag{3.2}$$

Solving Eq. (3.2) can give us a robust estimation of the linear regression coefficients $\boldsymbol{\beta}_m^*$, which denotes by $\hat{\boldsymbol{\beta}}_m$. The elements of $\hat{\boldsymbol{\beta}}_m$ can be used to elucidate the relative predictive power of different features. Features with higher absolute coefficient values can be viewed as more significant in predicting the outcome. Then, we consider the nonlinear non-parametric K-NN regression model by a $\hat{\boldsymbol{\beta}}_m$-weighted metric:

$$\|\mathbf{x} - \mathbf{x}_{mi}\|_{\hat{\mathbf{W}}_m} = \sqrt{(\mathbf{x} - \mathbf{x}_{mi})'\hat{\mathbf{W}}_m(\mathbf{x} - \mathbf{x}_{mi})}, \tag{3.3}$$

where $\hat{\mathbf{W}}_m$ is a diagonal matrix with elements $(\hat{\boldsymbol{\beta}}_{m1})^2, \cdots, (\hat{\boldsymbol{\beta}}_{mp})^2$, with $\hat{\boldsymbol{\beta}}_{mi}$ the i_th element in $\hat{\boldsymbol{\beta}}_m$. For each new test sample $\mathbf{x}$, within each group m, we find its K_m nearest neighbor using equation below:

$$\hat{y}_m(\mathbf{x}) = \frac{1}{K_m} \sum_{i=1}^{K_m} y_{m(i)}, \tag{3.4}$$

where $y_{m(i)}$ is the i-th closest sample to $\mathbf{x}$ in group m. The K-NN estimation of the future outcome in Eq. (3.4) is computed using the distance metric defined by the linear regression (DRLR in this case) coefficients. This can be viewed as a locally smoothed estimator in the neighborhood of $\mathbf{x}$. Features that are most predictive of the outcome will be amplified

by using the $\hat{\boldsymbol{\beta}}_m$ weighted metric. Thus, the selected samples will be close to $\mathbf{x}$ in terms of those important features, whose response values can be used as a good approximation of the future outcome $\hat{y}_m(\mathbf{x})$.

OLS informed K-NN OLS infomed K-NN was first applied in personalized diabetes management model by Bertsimas, et al. [Bertsimas et al., 2017]. Instead of using DRLR to solve the linear regression coefficient $\boldsymbol{\beta}^*$, Ordinary Least Square regression (OLS) is used to obtain $\hat{\boldsymbol{\beta}}_m$. Assume the following relationship between the features and the outcome:

$$y_m = \mathbf{x}_m{}'\boldsymbol{\beta}_m^* + \varepsilon_m,$$

where ε_m is the noise term that is assumed to be independent of $\mathbf{x}_m$. The problem can be formulated as:

$$\min_{\beta_m} \frac{1}{N_m} \sum_{i=1}^{N_m} (y_{mi} - \mathbf{x}_{mi}{}'\boldsymbol{\beta}_m)^2. \tag{3.5}$$

Then the OLS-learned $\hat{\boldsymbol{\beta}}_m$ will be fed into the nonlinear non-parametric K-NN regression model. The K-NN estimation of the future outcome in Eq. (3.4) will be computed using OLS coefficients weighted distance function.

3.3 Results

Table 3.1 summarizes the basic statistics of the entire dataset and the specific subgroups identified. Most of the patients in our dataset are older adults with a mean age of 60.93. The average SBP among all patient visits was above 130 mmHg. The dataset included patients with 10 different races, with the majority being Black patients (49.68%) or White patients (26.10%). The percentage of Black patients is significantly higher than the other races. Through all patient visits, ACEi was the most common active prescription. While almost 1/3 of patient visits were on ACEi, only 2.21% of them were on MRA.

The performance of the prescriptive model is shown in Table 3.2. Two strategies with

Table 3.1: Dataset summary statistics. Mean age and Mean SBP data is calculated through all patient visits. Hypertension drug used accounted for both monotherapies and drug combinations

Features	**All patients N=42,792**
Mean age	60.93
Mean SBP (mmHg)	136.11
Female	22,445 (52.45% of N)
Male	20,345 (47.54% of N)
White Patients	11,170 (26.10% of N)
Black Patients	21,260 (49.68% of N)
Hispanic Patients	3,946 (9.22% of N)
Ever Cigarette User	11,790 (27.55% of N)
Hypertension drugs used	**All Patient Visits (M=432,096)**
Calcium Channel Blockers (CCBs)	115,450 (26.72% of M)
Thiazides	92,290 (21.36% of M)
Angiotensin II Receptor Blockers (ARBs)	43,072 (9.97% of M)
ACE Inhibitors (ACEi)	129,800 (30.04% of M)
Beta-Blockers	86,171 (19.94% of M)
Loops	35,316 (8.17% of M)
Mineralocorticoid Receptor Antagonists (MRAs)	9,550 (2.21% of M)

Table 3.2: The reduction in future SBP in mmHg. Current regimen refers to the predicted outcome if the patient continues on the current prescription on record. Standard of care refers to the true outcome under the regimen prescribed by clinicians; mean (standard deviation).

Algorithm	**SBP reduction (mmHg)**	
	Deterministic	**Randomized**
LASSO	-10.71 (0.34)	-10.73 (0.35)
CART	-10.24 (0.55)	-10.28 (0.54)
OLS + K-NN	-13.28 (0.34)	-13.27 (0.34)
DRLR + K-NN	-14.22 (0.5)	-14.21 (0.49)
Current regimen	-8.31 (0.08)	
Standard of care	-8.35 (0.07)	

different policies to prescribe actions were compared (deterministic vs. randomized). The performance was tested under five different (random) splits of the data into training and test sets. We calculated the mean (over the five splits) SBP reduction under the model recommended prescription and the true reduction under standard of care. All models under two specific strategies outperform the current prescription and standard of care. The best outcome is achieved by DRLR-informed K-NN, which can achieve mean SBP reduction of 14.22 ± 0.5 mmHg. This amount of reduction is 70.30% better than the SBP reduction under standard of care. The other models also gained better control than the standard prescription with 59.04%-22.99% higher SBP reduction. However, even the second-best model, OLS-informed K-NN, achieved 6.61% lower reduction compared to the best model. This shows the superiority of our robustified procedure DRLR against outliers in terms of the improvement in outcomes. The performance of the randomized strategy did not differ much from the performance of the deterministic one. Nevertheless, considering that the randomized rule gives more flexibility for the prescriber to explore suboptimal options in the model to avoid potential medical contradictions in the deterministic results, the randomized policy is still preferable. Sub-models were developed for each prescription in the patients'

menu of treatment options. The potential outcome under a particular treatment was estimated by the corresponding sub-model, which is trained on patient visits that were previously on this treatment. We summarized the top 10 predictive features by assembling the feature importance scores (regression coefficient magnitudes) from all the sub-models in our best prescriptive model (DRLR-informed K-NN). The results are shown in Figure 3·1, with the features shown on the x-axis and the corresponding importance scores shown on the y-axis. All the features in the figure were positively correlated with the future predicted SBP. The most predictive feature is the SBP value in the current period, which has significantly higher importance score than others. The SBP values of the past two periods and the current period DBP also showed great influence on the prediction. Among all the clinical features, 'Oxygen saturation' is the most important one in predicting SBP. The larger variation of 'Oxygen saturation' may represent those medically fragile patients who have other comorbidities or even acute or subacute illnesses that impact their oxygen saturation. Moreover, the older population generally were more likely to have high SBP since 'Age' is the fourth significant feature with positive correlation with the outcome. There is also evidence supporting the possible roles of neutrophils and Red Blood Cell Distribution Width (RDW) in arterial hypertension and other cardiovascular diseases. [Araos et al., 2020, Li et al., 2017, Tanindi et al., 2012] One explanation is the possible mechanism via the overall inflammatory milieu causing arterial stiffness or subtle changes of myocardial or endocrine function. [Araos et al., 2020]

Since the model is prescribing monotherapy, only 30.68% of the recommendations from the DRLR+K-NN algorithm indicated continuing the current prescription, and 28.91% of the algorithmic recommendations matched the standard of care. In the cases where the prescriptive model recommended switching from the given therapy, the average reduction of SBP under algorithmic therapy is 14.65 mmHg, which is 13.48% better than the average reduction achieved by continuing the current medication in the record.

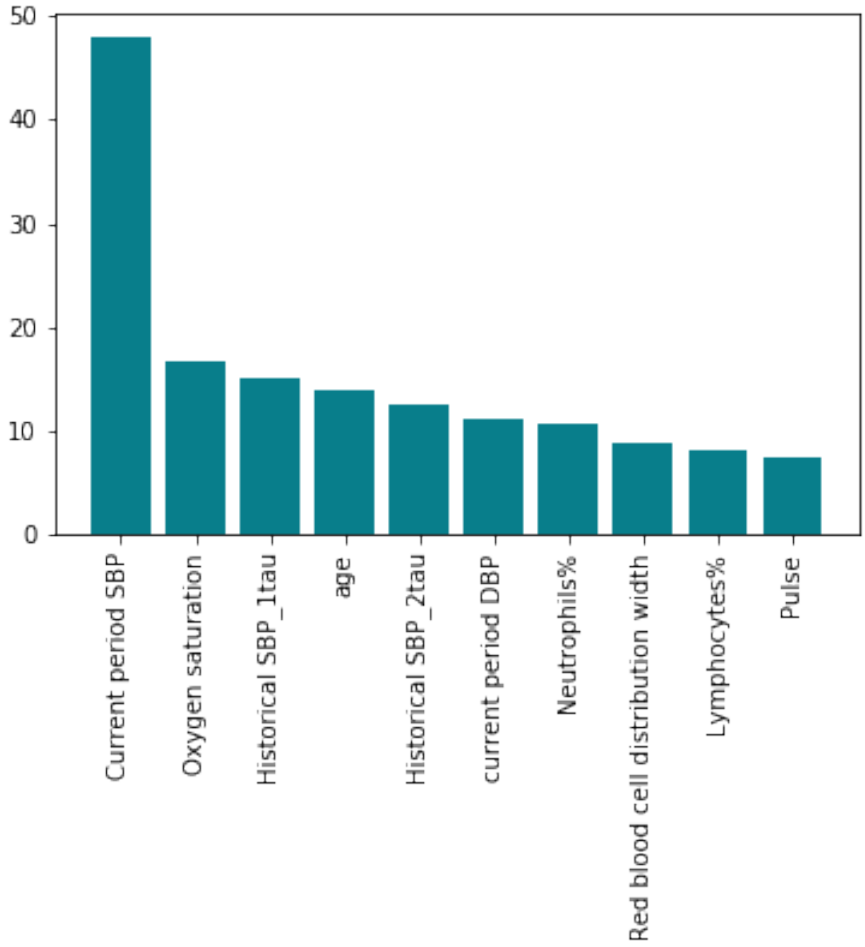


Figure 3·1: Top 10 predictive features summarized over all the sub-models in DRLR-informed K-NN. Historical SBP_1tau and Historical SBP_2tau refer to the Systolic Blood Pressure (SBP) values in previous periods. DBP: Diastolic Blood Pressure. Neutrophils%: Neutrophils as percent of blood leukocytes. Lymphocytes%: Lymphocytes as percent of blood leukocytes.

We randomly generated 350 samples from our best prescriptive algorithm (DRLR-informed K-NN model), each with the corresponding patient features and the model's recommended treatment. For each recommended prescription with the best SBP outcome, we summarized the corresponding neighborhood profile the model has learnt and used to generate the optimal decision. For example, in 'ACEi' profile, we found all patients who were recommended to take ACEi in the future and summarized the features over all their corresponding neighborhoods found by the model. In this way, we seek to interpret the decision-making process of the model and generalize the patient characteristics that point to a certain prescription type. All these results were validated through a sanity check by BMC clinician, to ensure that the recommendations from our model are clinically rational. Among the 350 generated cases, 307 of them (87.71%) have passed the sanity check. 19 out of the 43 failed cases are due to the nature of monotherapy setting of the model, which reduced the recommendation to one of the drugs from the current drug combination. This process is also known as deprescribing, and clinicians decided not to take this into account since it is still rare in the clinic and the effects may vary on a case by case basis.

3.4 Discussion

3.4.1 Feasibility of Deprescribing

Deprescribing refers to the process of tapering off or stopping medications to prevent adverse drug reactions and achieve better health management [Scott et al., 2015, Krishnaswami et al., 2019]. Among all the patient visits, 9196 of them were originally on multiple antihypertensive medication types, where they were able to achieve on average -8.07 mmHg SBP reduction with standard of care. However, with the prescriptive model recommendations, they can actually achieve -14.33 mmHg SBP reduction, on average, by monotherapy, which is 77.57% better than the standard outcomes. Among these 9,196 patients on combo drugs, 7,241 of them (78.74%) indeed achieved better SBP reduction than the standard of care. We

Table 3.3: Summarized statistics of patients with effective deprescribing. Percentages in all patient visits are calculated as a fraction of the total number of patient visits in the test set (N=31,758). The population in the effective deprescribing cohort consists of patients with better SBP control under deprescribing (M=7,241). All diseases refer to the corresponding diagnosis in ICD9/ICD10 (International Classification of Diseases) codes.

	All patient visits	**Visits with effective deprescribing**
Mean age	60.93	65.05
Diabetes	44.69%	54.07%
Chronic systolic heart failure	2.32%	3.25%
Atherosclerotic heart disease	13.23%	18.86%
Peripheral vascular disease	7.29%	11.02%

summarized a profile for these 7,241 patients in Table 3.3. The mean age of them is 65.05, which is higher than the mean age over the entire dataset (60.93). Compared with other patients, they have higher chance to be subject to medical contradictions, since more of them are having diagnosis of other diseases including diabetes and chronic systolic heart failure. The percentage of patients with diabetes diagnosis is 20.99% higher than the percentage in the entire dataset.

The potential benefits shown on deprescribing are interesting, since it is recognized by both the 8th report of the Joint National Committee for the Prevention, Detection, Evaluation and Treatment of Hypertension (JNC8) [James et al., 2014] and the 2013 guidelines of the European Society of Hypertension and the European Society of Cardiology (ESHESC) [Mancia et al., 2013] that patients with SBP above 20 mmHg should take combinations of antihypertensive agents for efficient hypertension control. However, deprescribing is increasingly encouraged among patients under control and older people with multi-morbidity, for whom the risk and burden of the drugs may outweigh their effectiveness [Krishnaswami et al., 2019, Benetos et al., 2019]. In fact, a clinical trial has shown that antihypertensive medication reduction was noninferior compared with the usual care among older patient, with respect to the proportion of patients with SBP lower than 150 mmHg at 12 weeks [Sheppard

et al., 2020]. Moreover, studies have shown that deprescribing of sedatives and nonsteroidal anti-inflammatory drugs is cost-effective [Sanyal et al., 2020, Turner et al., 2021]. In fact, the potential harms and long-term benefits of many cardiovascular drugs are still unknown [Krishnaswami et al., 2019]. Thus, clinicians should tailor appropriate treatment type more carefully, especially with the addition of new medications.

3.4.2 Neighborhood Profile Analysis

Seven neighborhood profiles were built to summarize patients' characteristics the model has captured for each prescription group. Figure 3·2 shows the corresponding percentage of certain subgroups in each neighborhood profile. If the percentage of group A in the drug X neighborhood is higher than the overall percentage of A in the whole test set, it implies that group A may generally get better response towards drug X. Most of the male patients fit into ACEi and Beta-Blockers, as the corresponding neighborhoods both comprise around 48% of male patients, which is 14.29% higher than the overall male percentage. On the other hand, female patients are more likely to be recommended an ARB since 77.26% of the ARB neighborhood are female.

Black patients generally got better response on MRAs compared with other medication types, where 86.14% of the MRAs neighborhood consists of Black patients. Whereas, White patients are better on a Beta-Blocker, and the percentage of them in the corresponding neighborhood is 83.2% higher than the overall percentage. Some hypotheses have been put forward to explain the favor of different prescriptions for Black and White patients [Carey et al., 2019, Cushman et al., 2002, Howard et al., 2015, Sousa et al., 2022], However, race is a social construct that usually comes with other environmental and socioeconomic factors that could be a proxy of unmeasured variables, as opposed to some essential physiologic difference. Thus, it is hard to get a clear explanation towards such racial difference.

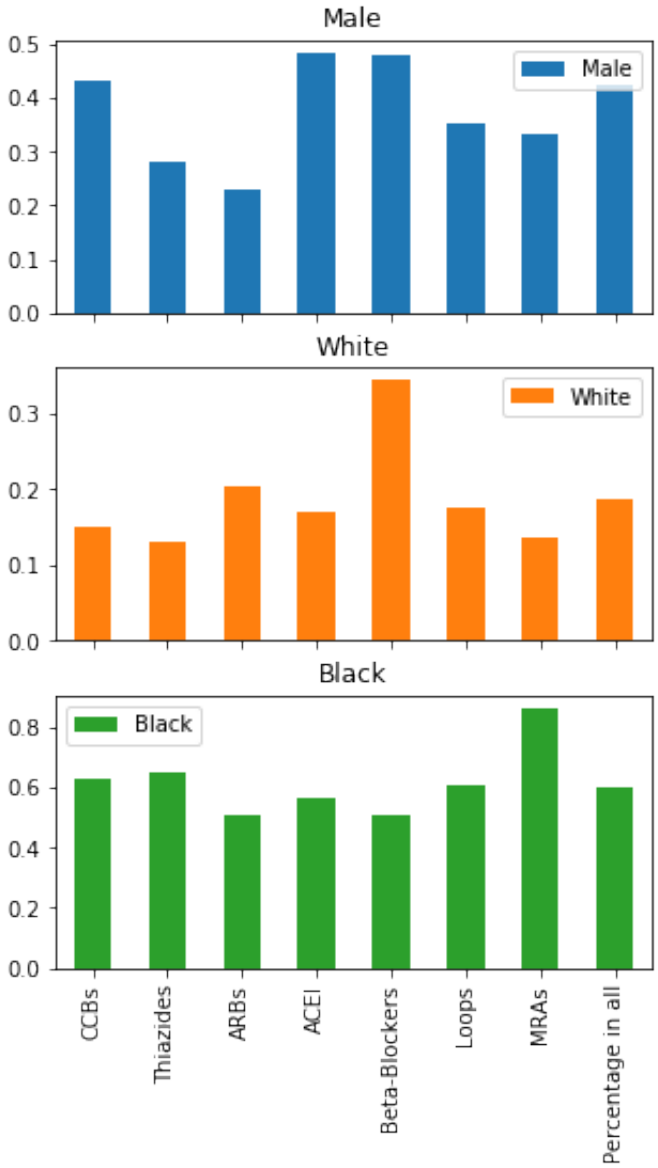


Figure 3·2: Subgroup analysis in the neighborhood profiles for each recommended prescription type. CCBs: Calcium Channel Blockers; ARBs: Angiotensin Receptor Blockers; ACEi: ACE inhibitors; MRAs: Mineralocorticoid-Receptor-Antagonists.

3.4.3 Hypertensive Treatment Initiation for Non-diabetic Drug Naïve Patients

Designed for personalized hypertension treatment, our model aims to make suggestions in real-world scenarios that have true clinical equipoise. One of the cases is the prescribing scenario with non-diabetic and antihypertensive drug naïve patients. We found 7,945 non-diabetic patient visits among the test set who did not take any antihypertensive medication previously. The benefit of using the algorithm amounted to a 15.56 mmHg SBP reduction on average, which is 9.42% higher than the average gain from the model. We summarized the optimal prescription types recommended by the algorithm for these patients in Figure 3·3. Among all the therapies, most of the patients were recommended to initialize with monotherapy of Thiazides or CCBs. This is consistent with the ACC/AHA hypertension treatment guideline, where prescription for drug-naïve patients is recommended to come from one of the four classes: Thiazides, CCBs, ACEi or ARBs [Whelton et al., 2018]. It has also been proved in randomized trials that CCBs and Thiazides are more effective as the initial treatment for Black subjects [Sareli et al., 2001, Mukhtar et al., 2018]. Interestingly, 9.47% of patients were recommended to take combination therapies, for whom the current averaged SBP (152.50 mmHg) is 12.04% higher than the mean SBP of the whole dataset. The very high blood pressure may explain their requirement of more intense treatment.

3.4.4 Limitations

Although the model can significantly improve hypertension control on standard of care, this study has several limitations. First, since it is a retrospective study, only the outcomes with administered prescriptions were observed. The counterfactual outcomes under other medications are unknown and can only be estimated by the predictive model. All the results and improvements we showed here are still worth to be tested in a real clinical trial. Second, there are actually a lot of factors aside from the magnitude of SBP reduction that need to be considered in the routine HTN management, including tolerance of past regimens, allergies,

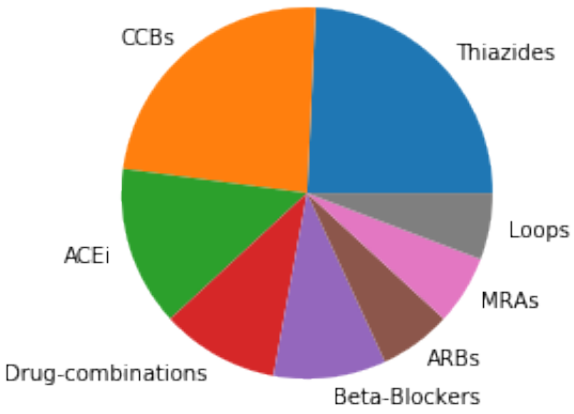


Figure 3·3: Model recommended prescription type for non-diabetic medication-naïve patients. CCBs: Calcium Channel Blockers; ARBs: Angiotensin Receptor Blockers; ACEi: ACE inhibitors; MRAs: Mineralocorticoid-Receptor-Antagonists.

dosage for current medications, best medication for other chronic conditions, etc. Given the nature of the dataset, we cannot account for some of these, which could limit pragmatic clinical utility in an all-hypertensive patients use case.

3.5 Conclusions

This study developed a prescriptive model that determines the optimal therapies for patients based on their specific characteristics. Our proposed robustified algorithm DRLR-informed KNN is developed to accommodate outliers in the EHR data, and it achieved 70.30% larger (-14.22 ± 0.5 mmHg) SBP reduction than standard of care, with the results being 7.08% better than the 2nd best model. Neighborhood profiles and patients' recommendations samples were generated to interpret the decision-making process of the model. 87.71% of the randomly sampled cases passed the professional sanity check by BMC clinicians. The results imply that: 1) personalized hypertension treatment by ML methods can provide good support on medical decision-making to improve drug efficacy; 2) the feasibility of deprescribing may be underestimated since it shows considerable benefits computationally; 3) although developed in limited scenarios, prescriptive algorithm can still give promising

insights on situations with clinical equipoise like drug initializations.

Chapter 4

Development and Validation of Predictive Models for COVID-19 Outcomes in a Diverse, Safety-net Hospital Population

4.1 Introduction

The existing predictive models of hospitalization, disease severity, and mortality were not developed using data from a safety-net hospital caring for a large percentage of racially/ethnically diverse patients, including many lower-income individuals with pressing needs associated with SDoH [Liang et al., 2020, Hao et al., 2020, Yan et al., 2020, Ji et al., 2020, Wang et al., 2020, Gong et al., 2020, Bhargava et al., 2020]. In addition, no models have leveraged SDoH for patients receiving clinical care. The present study includes a large percentage of Black and Hispanic patients and person-level information on SDoH, enabling a characterization of specific race/ethnicity and SDoH variables that influence the predictive models.

An additional characteristic of the current study is the availability of rich information on the daily/hourly evolution of vitals for hospitalized patients. Most of the previously published predictive models were static – they considered a snapshot of the patient's condition and made a forward prediction. In the current work, we leveraged neural networks with long short-term memory cells [Hochreiter and Schmidhuber, 1997] and a transformer [Vaswani et al., 2017] encoder to build a score of vitals that captures their dynamic evolution. Models based just on this score perform surprisingly well compared to more complex models that

also use a host of laboratories. In addition, access to hospital occupancy data reveals how they may influence care decisions.

4.2 Materials and Methods

4.2.1 Data Description

We de-identified data for all 7,102 patients with a positive reverse transcription polymerase chain reaction (RT-PCR) SARS-CoV-2 test at the BMC between January 1st and December 31st, 2020. As a tertiary care academic medical center, BMC is the largest safety-net hospital in New England, providing care for about 30% of Boston residents.18 Features extracted included demographics, SDoH variables, depression status, travel-contact information, vital signs, radiological findings, past medical history, symptoms, medications, laboratory tests, hospital occupancy, hospitalization course, admission to the Intensive Care Unit (ICU), mechanical ventilation, and mortality. SDoH variables were based on answers to the THRIVE survey administered at BMC, which identifies social needs in 8 domains: housing, food, medication, transportation, utilities, childcare, employment, and education [de la Vega et al., 2019]. We also used self-reported race and ethnicity in the electronic health records and hospital occupancy, which was measured by the daily bed usage percentage for surgeries, COVID and non-COVID patients. The study was approved by the BMC Institutional Review Board.

4.2.2 Pre-processing and Variable Selection

We developed predictive models for the following outcomes: 1) hospitalization, 2) ICU care, 3) mechanical ventilation, and 4) mortality. For each patient, we built a profile containing all outcome labels and extracted features. We used natural language processing (NLP) to extract radiology findings from text. We applied “one-hot” encoding to represent categorical features as 0 and 1. We retained variables for which we had values from at least 350 patients,

and we imputed the missing values in a continuous-valued feature using the mean of its non-missing values. All features were standardized to zero mean and unit standard deviation.

For the hospitalization model, we used the admission date as reference date, and the earliest positive SARS-CoV-2 test as the reference date for non-admitted patients. Features were extracted according to the reference dates. We utilized all features except laboratory results, medication, and radiological findings, which were typically not available for non-hospitalized patients. The features we used for predicting hospitalization include pre-existing conditions and SDoH information from the patient's hospital record, symptoms, and observed vital signs, and would all have been readily available to physicians in either the emergency room or the outpatient clinics making these decisions.

For admitted patients, the closest records before their reference date were extracted and we only included records that were within 48 hours before the reference date. For non-admitted patients, we only included records that were within 48 hours before or after the reference date. For ICU, mechanical ventilation and mortality prediction models, we only considered admitted patients. In addition to the features utilized in the hospitalization model, laboratory results and radiological findings were used, and we excluded symptoms since severe COVID-19 patients were less likely to describe their symptoms. The earliest and latest vitals and lab results to be used for ICU/Intubation/Mortality models depend on the timeline settings for models. For patients with the identified outcomes (ICU care, mechanical ventilation, death), the date of the outcome was used as the reference date. For patients with absent outcomes, a random date during their hospitalization was used as their reference date. In general, all input features for predicting the various in-hospital outcomes would have been readily available to physicians in advance of the predicted outcome.

4.2.3 Timeline Strategy

We introduced a timeline strategy to capture the dynamic evolution of vital signs, labs, and radiological findings for predicting ICU, mechanical ventilation, and mortality. Given the

reference date t and a desired "drop time" τ_0, we first eliminated all features during the interval $[t-\tau_0, t]$. Then, we defined k consecutive time windows with length τ each, tracing back from $t-\tau_0$. The mean (for continuous features) or maximum (for categorical features) of all feature records in the i-th time window $[t-\tau_0-i\tau, t-\tau_0-(i-1)\tau]$ was computed and defined as "feature - $i\tau$". We used the sequence "feature - 1τ",...,"feature - $k\tau$" as a feature timeline to train the models. We did not implement timelines for the hospitalization model, because laboratory and radiology findings were not used and vital sign records were sparse.

4.2.4 Classification Methods

We applied linear and nonlinear classifiers to predict outcomes. Linear methods included LR and SVM [Cortes and Vapnik, 1995]. Nonlinear methods included XGBoost [Chen and Guestrin, 2016] and Random Forest (RF). [Breiman, 2001] We introduced regularizations to prevent the influence of outliers in data [Chen and Paschalidis, 2020, Chen et al., 2019]. Furthermore, we used LSTM-Transformer neural networks to compute a score capturing the dynamic evolution of vitals over the timeline.

We applied statistical feature selection (SFS), removing variables with high p-value. We removed one from each pair of features with absolute correlation coefficient > 0.8. We further implemented l_1-regularized LR recursive feature elimination (RFE). Features retained from RFE were used to derive a parsimonious LR model.

4.2.5 Model Evaluation

We evaluated model performance using two metrics: AUC and Weighted-F1 score. We split patients into a training (80%) and test set (20%). We trained the models on the training set and evaluated them on the test set. We repeated this procedure 5 times, each with a different random split. The average and standard deviation on the test set over the 5 random splits are reported. We performed external validation to assess the generalizability

of our hospitalization models. We trained hospitalization models using all BMC samples and evaluated their performance on data from Mass General Brigham used in our earlier work. [Hao et al., 2020] We did not attempt external validation for other models because they rely on clinical variables and it was not possible to match those across the two data sets.

We compared our models with the NEWS226 score for predicting deterioration and the sepsis score qSOFA [Seymour et al., 2016]. These are computed from vital signs, so we compared them with our LSTM-Transformer vital score. In addition, we trained the 'BMC protocol', which is a classifier using a group of labs and vital signs chosen for evaluating COVID-19 severity by BMC physicians.

4.3 Results

Among 7,102 patients, 19.5% were admitted. Among the hospitalized, 23.3% required ICU care, 13.8% received mechanical ventilation, and 9.65% died. The mean age of all patients was 47.9 and 36.5% were Black.

4.3.1 Hospitalization Models

The hospitalization model used the entire data set, labeling patients as hospitalized (class 1) or non-hospitalized (class 0). 126 variables for each patient were retained after pre-processing. The average of the obtained metrics over 5 random splits is reported in Table 4.2. We compared the performance of linear (i.e., best performing SVM and LR) and nonlinear (i.e., XGBoost and RF) methods using all 126 variables. After SFS, 70 variables were retained and RFE retained 20 variables. The latter 'parsimonious' model was enhanced by adding two hospital utilization variables, while controlling for additional relevant variables. Specifically, for each patient we added 'Total Non-COVID Percentage' and 'Total COVID Percentage', indicating the ratio of the number of patients treated for non-COVID diseases and COVID, respectively, over the total number of BMC beds, computed at the patient's

Table 4.1: Hospitalization prediction models. The values inside the parentheses denote the standard deviation of the corresponding metric. SVM-L1 and LR-L1 refer to the l_1-norm regularized SVM and LR models.

Algorithm	AUC	F1-weighted
	Models using all 126 features	
LR-L1	86.5% (0.9%)	87.0% (0.5%)
SVM-L1	86.3% (0.9%)	86.8% (0.8%)
XGBoost	93.1% (0.6%)	90.1% (0.8%)
RF	92.4% (0.5%)	90.0% (0.7%)
	Models using 70 statistically selected features	
LR-L1	86.1% (1.0%)	87.3% (0.7%)
SVM-L1	86.0% (1.0%)	87.1% (0.6%)
XGBoost	92.5% (0.5%)	90.2% (0.7%)
RF	92.2% (0.5%)	89.9% (0.7%)
	Parsimonious Model using 22 features	
LR-L1	83.5% (1.7%)	85.3% (0.7%)
SVM-L1	83.4% (1.8%)	84.7% (1.0%)
XGBoost	92.0% (0.6%)	89.6% (0.6%)
RF	91.3% (0.6%)	89.4% (0.5%)

reference time. This resulted into parsimonious models with 22 variables. The parsimonious models performed almost as well as the models with all 126 features.

4.3.2 ICU Models

The ICU prediction results are in Table 4.3 and Table 4.4. We first trained one immediate model (0-drop), using the features in the past 36 hours to predict need for immediate ICU care. For vitals we used a k=6, τ=6h timeline, while for laboratory and radiology findings we only used one τ=36h window, since most laboratory data and imaging were taken at most once a day. After combining the vitals into the LSTM-Transformer score, we selected at

Table 4.2: Hospitalization prediction models. We report the composition of an l_2-norm regularized LR model, including the coefficient of each variable (Coef), the correlation of the variable with the outcome (Y-corr), the mean of the variable (Y1-mean) in the hospitalized, and the mean of the variable (Y0-mean) in the non-hospitalized. For each variable, we also report the corresponding p-value, the odds ratio (OR) and its 95% confidence interval (CI).

Variables for the Parsimonious Model								
Variable	***Coef***	**OR**	**OR 95% CI**		**Y1-mean**	**Y0-mean**	**Y-corr**	**P-value**
Intercept	-0.7	-	-	-	-	-	-	-
Dyspnea	1.28	3.98	3.3	4.8	0.36	0.08	0.32	1.60E-47
PMH-Transplantation	0.72	3.36	2.06	5.5	0.04	0.01	0.09	1.32E-06
Vomiting	0.52	1.86	1.26	2.74	0.05	0.02	0.07	1.88E-03
PMH-CHF	0.42	1.64	1.17	2.3	0.1	0.02	0.16	4.50E-03
PMH-Hepatitis	0.55	1.54	1.14	2.09	0.07	0.04	0.05	5.05E-03
SpO2	-0.83	0.68	0.65	0.72	96.09	98	-0.33	8.06E-56
SDOH-Food	0.28	1.4	0.87	2.24	0.03	0.02	0.02	1.61E-01
PMH-CKD	0.39	1.35	1.02	1.79	0.14	0.04	0.16	3.65E-02
PMH-Stroke	0.35	1.32	0.9	1.94	0.06	0.02	0.1	1.50E-01
PMH-COPD	-0.16	0.79	0.56	1.11	0.06	0.03	0.06	1.75E-01
Abdominal Pain	0.23	1.26	0.9	1.77	0.06	0.03	0.07	1.75E-01
PMH-Arrhythmia	0.28	1.25	0.9	1.75	0.08	0.03	0.12	1.88E-01
Contact-COVID-Travel	0.23	1.24	1.06	1.45	0.37	0.28	0.08	6.24E-03
Respiratory Rate	0.45	1.18	1.14	1.22	20.07	18.07	0.28	4.87E-25
Diarrhea	0.14	1.1	0.81	1.5	0.07	0.04	0.06	5.39E-01
SDOH-Transportation	0.23	1.06	0.58	1.95	0.02	0.01	0.03	8.43E-01
BMI	0.08	1.04	1.02	1.06	30.01	29.12	0.1	1.65E-05
Diastolic BP	-0.15	0.97	0.96	0.98	75.28	77.63	-0.13	2.56E-12
Age	0.52	1.03	1.03	1.04	57.31	45.62	0.26	1.13E-44
PMH-Cirrhosis	0.1	1.02	0.52	2.01	0.02	0.01	0.03	9.52E-01
Total non-COVID percentage	-0.55	0.991	0.989	0.994	69.05	73.71	-0.08	1.11E-10
Total COVID percentage	-0.59	0.992	0.99	0.994	40.72	38.26	0.04	2.43E-11

PMH: Past Medical History; CKD: Chronic Kidney Disease; COPD: Chronic Obstructive Pulmonary Disease;

Total non-COVID percentage: (Total number of non-COVID patients at the hospital / Total number of beds) * 100;

Total COVID percentage: (Total number of COVID patients at the hospital / Total number of beds) * 100.

most 10 features using SFS and RFE (reported in Table 2) and trained a parsimonious model. The parsimonious model yielded an average AUC of 92.7%, which is close to the best full model AUC of 94.8%. Using only NEWS2 or qSOFA scores as the feature yielded AUC of 84.3% and 71%, respectively, lower than the AUC of 90.9% using the LSTM-Transformer vital score as the only feature.

We further trained two extreme models to predict if a patient would need ICU care after 12 hours (12h-drop model) and 24 hours (24h-drop model). By changing the drop time in the timeline to 12h and 24h, respectively, the model is not using any information for the patient in the 12/24 hours before ICU admission. The parsimonious models maintained a high AUC of 86.5% and 81.1%, respectively, which match and exceed the corresponding best nonlinear full models with AUC of 86.6% and 79.9%, respectively. While for these early predictions NEWS2 and qSOFA-based models performed poorly, the LSTM-Transformer score remained a strong predictor. Apparently, for immediate predictions all models did relatively well, whereas for longer-term predictions the LSTM-Transformer score and other models including it show significant advantage.

Table 4.3: ICU prediction models. For each full model, we only report results from the algorithm with the highest AUC out of LR, SVM, XGBoost and RF.

	0-drop		**12h-drop**		**24h-drop**	
	AUC	**F1-weighted**	**AUC**	**F1-weighted**	**AUC**	**F1-weighted**
Best full models	XGBoost		XGBoost		XGBoost	
before SFS	94.8% (1.2%)	89.2% (1.8%)	86.6% (1.2%)	82.4% (0.9%)	78.3% (2.0%)	76.0% (2.3%)
Best full models	XGBoost (112 features)		XGBoost (104 features)		XGBoost (95 features)	
after SFS	94.3% (1.4%)	88.0% (2.3%)	86.3% (1.8%)	82.2% (1.2%)	79.9% (3.2%)	76.8% (3.4%)
Parsimonious models LR-L1-LSTM-Transformer	92.7% (1.5%)	86.7% (1.3%)	86.5% (1.6%)	81.6% (2.1%)	81.1% (2.5%)	79.4% (0.8%)
BMC-Protocol LR-L1	89.0% (2.2%)	86.1% (2.0%)	73.0% (2.0%)	73.7% (1.0%)	67.4% (1.7%)	70.8% (1.7%)
BMC-Protocol XGBoost	94.8% (0.8%)	89.1% (1.7%)	86.2% (1.0%)	83.2% (1.0%)	77.1% (2.5%)	74.0% (1.4%)
NEWS2 score	84.3% (2.0%)	83.5% (1.8%)	48.6% (2.5%)	72.2% (1.4%)	46.5% (2.3%)	70.4% (1.6%)
qSOFA score	71.7% (3.0%)	79.3% (1.7%)	54.3% (2.1%)	69.1% (1.8%)	52.1% (2.4%)	68.1% (2.0%)
LSTM-Transformer score	90.9% (2.0%)	85.2% (1.8%)	84.5% (2.0%)	81.3% (1.2%)	76.8% (3.9%)	73.6% (1.5%)

4.3.3 Mechanical Ventilation Models

The mechanical ventilation prediction results are in Table 4.5 and Table 4.6. As with the ICU models, we trained one immediate model (0-drop), using the past 36h features to predict if a patient needs to be intubated immediately. For vitals, we used a k=6, τ=6h timeline, while for laboratory and radiology findings we only used one τ=36h window. After combining vitals into the LSTM-Transformer score, we selected 10 features using RFE and trained a parsimonious model. The top features are reported in Table 4.5. The parsimonious model obtained an average AUC of 91.2%, close to the AUC of the best full model (93.8%). Using only NEWS2 or qSOFA scores yields an AUC of 66.0% and 63.1%, respectively, lower than the AUC of 90.0% obtained by using just the LSTM-Transformer vital score.

We further trained two extreme models to predict if a patient would need intubation after 12 hours (12h-drop model) and 24 hours (24h-drop model); the corresponding parsimonious models have AUC of 90.3% and 84.9%, respectively. NEWS2- and qSOFA-based models do considerably worse in these advance predictions.

Table 4.4: ICU prediction models. We present the LR coefficients of each variable (Coef), the correlation of the variable with the outcome (Y-corr), the p-value, the mean of the variable (Y1-mean) in the ICU patients, and the mean of the variable (Y0-mean) in the non-ICU patients.

Variable	Coef	Y1-mean	Y0-mean	P-value	Y-corr
0-drop parsimonious model 10 features					
LSTM-Transformer vital score	3.42	3.44	-4.70	5.63E-163	0.82
Intercept	-2.44	-	-	-	-
Glucose	0.52	182.20	129.24	5.52E-13	0.26
PMH-DVT	0.46	0.09	0.04	2.12E-02	0.08
LDH	0.45	496.37	356.83	2.19E-32	0.24
Total COVID percentage	-0.41	40.76	42.16	3.41E-01	-0.03
PMH-HLD	0.31	0.32	0.24	2.94E-02	0.08
CO2	-0.24	22.62	24.13	2.53E-07	-0.19
NRBC Percentage	0.23	0.23	0.22	4.25E-03	0.01
Lung Opacity on X-Ray	0.2	0.68	0.29	2.35E-33	0.33
CRP	0.1	123.53	70.81	1.87E-28	0.30
12h-drop parsimonious model 10 features					
LSTM-Transformer vital score	1.81	0.83	-1.18	1.79E-74	0.50
Intercept	-0.88	-	-	-	-
Total COVID percentage	-0.33	40.45	41.97	3.16E-01	-0.03
PMH-Arrhythmia	0.22	0.12	0.07	2.71E-02	0.08
LDH	0.22	431.34	358.87	1.49E-30	0.19
BUN	0.2	21.32	18.41	5.47E-20	0.09
Age	0.16	60.09	56.47	2.20E-03	0.09
CRP	0.14	100.25	72.09	4.97E-33	0.19
Ferritin	0.12	1758.64	990.79	4.78E-29	0.15
PMH-CKD	0.11	0.20	0.12	1.14E-02	0.09
PMH-CAD	0.11	0.09	0.04	1.19E-02	0.09

LDH: Lactate Dehydrogenase; BUN: Blood Urea Nitrogen; NRBC: Nucleated Red Blood Cell;

CKD: Chronic Kidney Disease; PMH: Past Medical History; CAD: Coronary Artery Disease;

DVT: Deep Vein Thrombosis; HLD: Hypersensitivity Lung Disease;

Total COVID percentage: (Total number of COVID patients at the hospital / Total number of beds) * 100.

Table 4.5: Mechanical ventilation prediction models. For each full model, we only report results from the algorithm with the highest AUC out of LR, SVM, XGBoost and RF.

	0-drop		**12h-drop**		**24h-drop**	
	AUC	**F1-weighted**	**AUC**	**F1-weighted**	**AUC**	**F1-weighted**
Best full models	XGBoost		RF		RF	
before SFS	93.6% (0.9%)	91.6% (0.8%)	90.6% (1.2%)	87.8% (1.0%)	86.3% (1.1%)	85.0% (2.1%)
Best full models	XGBoost (90 features)		XGBoost (104 features)		XGBoost (107 features)	
after SFS	93.8% (1.2%)	91.8% (1.5%)	90.3% (1.4%)	88.8% (0.7%)	86.1% (1.1%)	85.9% (1.4%)
Parsimonious models LR-L1-LSTM-Transformer	91.2% (2.0%)	90.7% (0.9%)	90.3% (1.4%)	87.6% (1.0%)	84.9% (1.5%)	84.7% (0.5%)
BMC-Protocol LR-L1	82.3% (0.9%)	85.0% (1.7%)	63.4% (3.1%)	79.9% (0.6%)	56.7% (1.5%)	79.9% (0.7%)
BMC-Protocol XGBoost	90.0% (1.0%)	88.7% (0.9%)	88.1% (1.1%)	88.3% (0.3%)	84.0% (1.1%)	80.3% (0.8%)
NEWS2 score	66.0% (6.5%)	84.2% (1.3%)	65.7% (2.8%)	80.0% (0.0%)	67.6% (2.8%)	80.0% (0.0%)
qSOFA score	63.1% (5.1%)	80.0% (0.4%)	52.3% (2.0%)	80.0% (0.0%)	52.3% (2.0%)	80.0% (0.0%)
LSTM-Transformer score	90.0% (2.3%)	88.4% (1.6%)	85.9% (2.8%)	86.3% (1.2%)	80.0% (2.5%)	79.9% (0.2%)

4.3.4 Mortality Models

Due to the relatively longer mean time gap between hospitalization and death, we built different timelines for the mortality models. The first mortality model only uses features within 3 days after admission (adm-based model), and k=3, τ=24h are applied in this timeline. Consequently, we can predict a patient's mortality at the very early stage of hospitalization. Another model uses a drop time of 24h prior to death (24h-drop model), using k=7 and τ=48h for the timeline. For both settings, the LSTM-Transformer vital score is used in parsimonious models. Performance and top features are reported in Table 4.7 and Table 4.8.

For adm-based models, the best full model achieved 91.4% AUC, while the parsimonious model using LR with only 13 features did better (AUC of 92.0%). The AUC of qSOFA and NEWS2 models did not exceed 69%, and the LSTM-Transformer score yielded a model with AUC of 84.3%. For the 24h-drop model, the best nonlinear model achieved 96.2% AUC, and the parsimonious model using 13 features achieved an AUC of 94.0%. When the outcome draws near, the advantage of the LSTM-Transformer score over the NEWS2 score remains significant.

4.4 Discussion

The best AUCs achieved by the four models are between 93% and 96%, indicating strong predictive power. Strong predictions are achieved with relatively few features used by parsimonious models. These models use no more than 22 features each for hospitalization and mortality prediction, and no more than 10 features each for ICU and ventilation prediction, yielding similar (or better) performance with an AUC differential of -2.6% to +1.2% compared to the best models. This indicates the possibility of implementing simple, actionable predictive models to aid triage, staffing and resource planning. The models produced outperformed related models in the literature (e.g., the ventilation model outperforms an earlier model with a 74% AUC [Rodriguez et al., 2021]).

Table 4.6: Mechanical ventilation prediction models. We present the LR coefficients of each variable (Coef), the correlation of the variable with the outcome (Y-corr), the p-value, the mean of the variable (Y1-mean) in the intubated patients, and the mean of the variable (Y0-mean) in the non-intubated patients.

Variable	Coef	Y1-mean	Y0-mean	P-value	Y-corr
0-drop parsimonious model 10 features					
LSTM-Transformer vital score	2.26	2.25	-3.44	4.97E-104	0.68
Intercept	-2.15	-	-	-	-
Immature Granulocytes 0 Percentage	-0.63	0.26	0.50	1.79E-08	-0.17
Gender (male)	0.43	0.68	0.53	3.17E-03	0.10
Venous PH	-0.42	7.32	7.35	6.15E-06	-0.11
Absolute Lymphocytes	-0.4	1.35	1.47	7.22E-07	-0.06
Red cell Distribution Width	0.3	14.54	14.25	1.14E-03	0.06
Glucose	0.3	166.23	130.88	5.32E-21	0.18
Total Elective Surgery	-0.3	6.68	9.65	3.05E-03	-0.08
CRP	0.21	118.04	71.76	1.06E-29	0.22
Age	0.19	59.76	56.92	4.72E-02	0.05
12h-drop parsimonious model 10 features					
LSTM-Transformer vital score	1.54	2.42	0.18	5.05E-62	0.52
Intercept	-1.41	-	-	-	-
CRP	0.6	92.16	72.83	3.32E-29	0.11
Immature Granulocytes 0 Percentage	-0.55	0.09	0.44	2.76E-18	-0.25
Hemoglobin	0.47	12.45	12.34	1.96E-12	0.02
Red Cell Distribution Width	0.41	14.38	14.29	2.86E-14	0.02
Total Elective Surgery percentage	-0.4	6.61	9.75	3.56E-03	-0.08
Calcium	-0.37	8.82	8.87	1.52E-10	-0.04
CO2	-0.36	23.65	24.04	1.51E-12	-0.05
Age	0.31	59.76	56.92	4.72E-02	0.05
Gender (male)	0.26	0.68	0.53	3.17E-03	0.10

CRP: C-reactive protein;

Total Elective Surgery percentage: (Total number of Elective Surgeries / Total number of beds) * 100.

Table 4.7: Mortality prediction models with features extracted within 3 days after admission. For each full model, we only report results from the algorithm with the highest AUC out of LR, SVM, XGBoost and RF.

	Adm-based		**24h-drop**	
	AUC	**F1-weighted**	**AUC**	**F1-weighted**
Best full models before SFS	XGBoost 91.4% (0.8%)	91.9% (0.7%)	XGBoost 96.2% (0.7%)	94.6% (0.6%)
Best full models after SFS	XGBoost 91.1% (2.0%)	91.8% (1.3%)	XGBoost 94.7% (1.2%)	94.3% (1.0)
Parsimonious models LR-L1-LSTM-Transformer	92.0% (2.9%)	92.4% (1.3%)	94.0% (0.6%)	92.7% (0.9%)
BMC-Protocol LR-L1	88.9% (4.1%)	91.2% (0.7%)	89.5% (2.2%)	91.5% (0.6%)
BMC-Protocol XGBoost	89.3% (2.4%)	90.8% (0.7%)	92.3% (1.1%)	92.6% (1.4%)
NEWS2 score	66.8% (5.0%)	85.8% (0.9%)	72.4% (2.9%)	87.2% (1.1%)
qSOFA score	68.2% (3.0%)	86.6% (0.9%)	73.3% (5.1%)	87.1% (0.6%)
LSTM-Transformer score	84.3% (4.4%)	89.0% (0.9%)	87.1% (2.0%)	91.0% (1.5%)

Table 4.8: Mortality prediction models with features extracted within 3 days after admission. We present the LR coefficients of each variable (Coef), the correlation of the variable with the outcome (Y-corr), the p-value, the mean of the variable (Y1-mean) in the deceased, and the mean of the variable (Y0-mean) in the non-deceased.

Variable	Coef	Y1-mean	Y0-mean	P-value	Y-corr
Adm-based parsimonious model 13 features					
LSTM-Transformer vital score	1.55	1.88	-3.32	0.00E+00	0.53
Intercept	-1.98	-	-	-	-
Age	0.48	69.96	55.96	1.11E-16	0.23
CRP - 3τ	0.36	140.83	80.83	1.55E-08	0.26
PMH-CAD	0.34	0.19	0.04	2.74E-11	0.19
Red Cell Distribution Width - 1τ	0.31	15.56	14.27	3.06E-06	0.19
Poly Neutrophils - 3τ	0.29	74.55	65.77	1.42E-09	0.22
Alkaline Phosphatase (total) - 2τ	0.27	111.91	87.40	3.36E-03	0.13
Blood Component Type (Red Cell Group) - 1τ	0.23	0.31	0.10	9.77E-12	0.20
Calcium - 3τ	-0.22	8.51	8.76	4.85E-05	-0.16
Potassium - 1τ	0.22	4.44	4.01	2.89E-15	0.23
Ferritin - 1τ	0.16	2846.19	1026.75	2.41E-06	0.20
Platelets - 2τ	-0.15	226.14	245.33	1.13E-02	-0.06
Glucose - 2τ	0.13	149.06	129.09	4.39E-05	0.11
24h-drop parsimonious model 13 features					
LSTM-Transformer vital score	2.67	2.81	-3.01	0.00E+00	0.63
Intercept	-3.67	-	-	-	-
PH Arteria - 1τ	-1.32	7.34	7.38	1.16E-12	-0.28
White Blood Cells - 3τ	1.13	10.53	8.27	8.88E-16	0.21
Platelets - 3τ	-1.07	230.18	253.84	4.71E-14	-0.11
LDH - 4τ	0.93	691.01	537.40	2.57E-06	0.18
CRP - 3τ	0.83	159.66	112.24	9.11E-10	0.27
Age	0.81	69.96	55.96	1.11E-16	0.23
PHM-CHF	0.67	0.23	0.08	6.86E-07	0.15
Absolute Lymphocytes - 3τ	-0.59	1.29	1.45	5.06E-05	-0.03
Fibrinogen - 3τ	-0.56	512.74	521.48	5.66E-04	-0.03
PCO2 Arterial - 2τ	0.42	46.14	42.43	1.27E-09	0.26
Poly Neutrophils - 6τ	0.20	74.76	72.14	1.36E-09	0.18
Hemoglobin Blood - 7τ	-0.10	10.88	11.16	4.41E-07	-0.15

PMH: past medical history; CHF: Congestive Heart Failure

CAD: Coronary Artery Disease; CRP: C-reactive protein; LDH: Lactate Dehydrogenase.

Patients' vital signs were the most important factors for ICU, ventilation, and mortality prediction. These vital signs imply the severity of the disease and the potential need for cardiorespiratory resuscitation. Most of the prior studies use vital signs as "static" independent predictive variables [Hao et al., 2020, Yan et al., 2020, Ji et al., 2020, Wang et al., 2020]. In this study, we used an LSTM + Transformer encoder deep neural network to develop a single score combining all vitals and capturing their dynamic evolution over time. Models for ICU and ventilation (short-term and longer-term) predictions using just the LSTM-Transformer vital score have an AUC within 1.2% to 4.9% from the corresponding parsimonious models which also use other clinical variables; essentially, for these models, vital sign trends alone suffice. For mortality predictions, the LSTM-Transformer score is the top variable but other clinical variables significantly enhance performance.

Long-term predictions are more challenging than short-term: ICU, ventilation, and mortality predictions deteriorate as we move further from the time of the outcome. While most models do relatively well for short-term predictions, the parsimonious models which include the LSTM-Transformer score increase their advantage to baseline models (e.g., NEWS2 and qSOFA) when longer-term predictions are sought. Specifically, the AUC differential between the parsimonious model and the best of the NEWS2 and qSOFA-based models increases from 8.4% to 25.2% for short-term predictions and 17.3% to 32.2% for longer-term predictions. Incidentally, the protocol used at BMC fares better and is closer to the parsimonious model for both short and long-term predictions.

Some of the variables included in the ICU prediction model have previously been identified in the literature. Patient age and past medical history like renal disease (CKD), cardiac disease (CAD) have extensively been described as factors influencing disease severity. [Kantroo et al., 2021] Laboratory data such as CRP and ferritin [Huang et al., 2020] are acute phase reactants and have also previously been associated with COVID-19 disease severity. The large and diverse population used in our work strongly supported these

findings, and our interpretable LR model coefficients further numerically show their relative importance. The mortality model also includes laboratory data that have also previously been identified in the literature as being associated with disease severity such as CRP, ferritin and LDH. [Henry et al., 2020, Hao et al., 2020] Since mortality prediction models use multiple time windows for labs as well, the most informative period of a certain lab is further revealed.

Analyzing data from a safety net hospital with a high proportion of Black patients and information on SDoH needs, gave us an opportunity to assess the effect of racial bias and socioeconomic variables. Food insecurity and need for Transportation became the top predictive features in the hospitalization model, possibly because the need for food and transportation serve as a marker for economic hardship. Food insecurity is a very basic need related to patients' lifestyle and status of health. In the meantime, COVID pandemic further expanded the worldwide food insecurity, which makes it harder for vulnerable households to get enough healthy food. Similarly, patients with transportation needs rely more on most affordable public transit, which increases their risk for exposure to SARS-CoV-2 virus. While people with private cars and those who work from home can avoid such virus spreading during traveling. Further, delayed access to care leads to worse clinical condition when arriving in acute settings. The other SDoH variables like Housing insecurity were not as predictive as 'Food' and 'Transportation' since the homeless rate in Boston dropped sharply in recent years, 97% to 98% of the homeless population has been sheltered according to the latest homeless census. [202,]

Hospital census, such as the percentage of COVID-19 and non-COVID-19 patients and elective surgeries performed, affect the prediction results. This implies that an oversaturated hospital does affect resource allocation for new patients and further exacerbates the risk of future decompensation without adequate medical support.

Our hospitalization prediction model can be useful in any outpatient or emergency care

setting given that the variables used are readily available to clinicians. This includes information on SDoH which is regularly collected at BMC. We note that such SDoH information gathering practices are becoming more widespread. Patients from underrepresented groups with potential SDoH needs are in fact more likely to present to an emergency care setting with ambulatory sensitive conditions compared to others [Brown et al., 2012].

A potential limitation of the study is that even though the hospitalization model has been externally validated, it has not been possible to do the same with the remaining models, particularly using data from other safety-net hospitals.

4.5 Conclusion

Our COVID-19 prediction models that are based on a large diverse patient population can accurately predict outcomes, potentially aiding in triage, resource allocation, and staffing determinations. Additionally, the use of dynamic variables such as vital signs improve the predictive ability of models and should be considered in future model development. This study highlights the importance of ensuring diverse patient populations are represented in advanced analytics development and suggests how to carefully consider and interpret race within predictive models.

Chapter 5

Distributionally Robust Optimization in Credible Learning

5.1 Introduction

The concept of using regularization for group-level variable selection was first introduced by Bakin [Bakin, 1999], and later expanded upon by Yuan and Lin [Yuan and Lin, 2006]. They extended the original LASSO technique [Tibshirani, 1996] by imposing a block-wise l_2-normed penalty for the grouped coefficient vectors. It is particularly useful in scenarios where predictors or features are naturally grouped, making it highly relevant in areas such as healthcare and genomics [Zhang et al., 2018, Rao et al., 2013]. Group LASSO (GLASSO) [Yuan and Lin, 2006] facilitates interpretability by allowing discussions around groups of predictors instead of lengthy individual lists. Various adaptations of GLASSO, including those for logistic regression [Meier et al., 2008], adaptive GLASSO [Wang and Leng, 2008], overlapping GLASSO [Yuan et al., 2011], and multivariate sparse GLASSO [Li et al., 2015], have been proposed to broaden its applicability.

Several studies illustrated the diverse applications and enhancements in interpretability achieved through the use of LASSO and GLASSO in different research areas [Yang et al., 2024, Huang and Zhang, 2010, Rao et al., 2013, Frost and Amos, 2017]. The study by Li shows how the sparse GLASSO method enhances interpretability in the analysis of brain regions [Li et al., 2020]. Zhang employed the group LASSO to improve interpretability in the context of feature interaction learning [Zhang et al., 2018]. However, to the best of our

knowledge, few studies using GLASSO technique have focused on addressing the credibility and comprehensiveness of models. Models that incorporate additional expert knowledge often do not yield sparse and credible results; instead, they are primarily designed to expedite the learning process or enhance predictive accuracy [Vapnik and Izmailov, 2015, Helleputte and Dupont, 2009, Choi et al., 2017]. Some studies also focus on rule-based systems and incorporate monotone functions identified by experts [Ben-David, 1995, Pazzani et al., 2001]. Perhaps the closest proposed approach to our study is by Wang [Wang et al., 2018], where they suggested a regularization penalty within a linear framework, integrating domain knowledge about factors considered (or presumed) to be significant. However, their focus was limited to a basic group structure in the dataset, merely determining whether an expert had identified a covariate. This approach might not easily extend to more intricate scenarios where features exhibit natural grouping and overlapping groups, due to the complexity of these relationships.

In this chapter, we extend the notion of GLASSO from the model interpretability and credibility aspect. We introduced a regularization penalty called DRO-expert, showing how GLASSO can be adapted and used in conjunction with expert knowledge, enhancing its applicability in various contexts, especially where the grouping of predictors can be informed by domain expertise like healthcare. Additionally, we investigate how this credible learning version of GLASSO can be adjusted for scenarios involving overlapping groups, a situation frequently encountered in healthcare where a single medical trait may be analyzed from multiple perspectives. This exploration underscores the method's versatility and adaptability in integrating expert insights.

5.2 DRO Credible Learning Formulation

5.2.1 Continuous variables with Non-overlapping groups

In this section, we demonstrate how the Wasserstein DRO framework under a new expert norm recovers the Group LASSO (GLASSO) formulation and enhances both the interpretability and credibility of the model. We focus on the scenario with non-overlapping feature groups.

Consider a linear regression model where we are given a predictor vector $\mathbf{x} \in \mathbb{R}^p$, which can be decomposed into two categories $\mathbf{x} \triangleq (\mathbf{x}_K, \mathbf{x}_N)$, where $\mathbf{x}_K \in \mathbb{R}^{p_K}$ is one expert-recognized group of variables, and $\mathbf{x}_N \triangleq (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ consists of n groups of other variables with $\mathbf{x}_i \in \mathbb{R}^{p_i}$, $i \in [n]$. There is no overlap between any pair of feature groups, i.e., $p_K + \sum_{i=1}^{n} p_i = p$. The linear regression model can be expressed as:

$$y = \boldsymbol{\beta}'\mathbf{x} + \eta,$$

where y is the response variable and η is a random noise term. The regression coefficient $\boldsymbol{\beta}$ can be decomposed in a similar fashion as $\mathbf{x}$, i.e., $\boldsymbol{\beta} \triangleq (\boldsymbol{\beta}_K, \boldsymbol{\beta}_N)$, and $\boldsymbol{\beta}_N \triangleq (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n)$. To inject robustness into the linear model, we formulate a distributionally robust learning problem as follows:

$$\inf_{\boldsymbol{\beta}} \sup_{\mathbb{Q} \in \Omega} \mathbb{E}^{\mathbb{Q}} \left[|y - \boldsymbol{\beta}'\mathbf{x}| \right], \tag{5.1}$$

where $\mathbb{E}$ denotes expectation, and $\mathbb{Q}$ is the probability distribution of $(\mathbf{x}, y)$, lying within a distributional uncertainty set Ω defined as:

$$\Omega \triangleq \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) : W_1(\mathbb{Q}, \hat{\mathbb{P}}_M) \leq \varepsilon\},$$

where $\varepsilon > 0$ specifies the size of the ambiguity set Ω, $\mathcal{Z}$ is the set of possible values for $(\mathbf{x}, y)$, $\mathcal{P}(\mathcal{Z})$ is the space of all probability distributions supported on $\mathcal{Z}$, $\hat{\mathbb{P}}_M$ is the empirical probability distribution that assigns equal probability to each training sample point $(\mathbf{x}^{(i)}, y^{(i)})$,

$i \in [M]$, and $W_1(\mathbb{Q}, \hat{\mathbb{P}}_M)$ is the order-one Wasserstein distance between $\mathbb{Q}$ and $\hat{\mathbb{P}}_M$ defined on the metric space $(\mathcal{Z}, s)$ by:

$$W_1(\mathbb{Q}, \hat{\mathbb{P}}_M) \triangleq \min_{\Pi \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z})} \left\{ \int_{\mathcal{Z}\times\mathcal{Z}} s((\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)})) \Pi\big(d(\mathbf{x}^{(1)}, y^{(1)}), d(\mathbf{x}^{(2)}, y^{(2)})\big) \right\},$$

where Π is the joint distribution of $(\mathbf{x}^{(1)}, y^{(1)})$ and $(\mathbf{x}^{(2)}, y^{(2)})$ with marginals $\mathbb{Q}$ and $\hat{\mathbb{P}}_M$, respectively. For the metric $s((\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}))$, we introduce a DRO-expert norm that encourages density in the expert feature group but sparsity in the other feature groups.

Definition 1 (DRO-expert norm)**.** *Define the* δ*-weighted DRO expert norm of* $(\mathbf{x}, y)$ *as*

$$\|(\mathbf{x}, y)\|_{\delta, expert} \triangleq \|\mathbf{x}_K\|_2 + \|(\mathbf{x}_N, \delta y)\|_{2,\infty},$$

where δ *is a very large positive weight assigned to the response variable, and* $\|\cdot\|_{2,\infty}$ *is the* $(2,\infty)$*-norm introduced in [Chen and Paschalidis, 2020] and defined as*

$$\|(\mathbf{x}_N, \delta y)\|_{2,\infty} \triangleq \max(\|\mathbf{x}_1\|_2, \ldots, \|\mathbf{x}_n\|_2, \delta|y|).$$

According to [Chen and Paschalidis, 2020], in order to derive a tractable reformulation of Eq. (5.1), we need to obtain the dual norm of the DRO-expert norm. Here we present the definition of dual norm.

Definition 2 (Dual norm)**.** *Given a norm* $\|\cdot\|$ *on* $\mathbb{R}^d$*, the dual norm* $\|\cdot\|^*$ *for any vectors* $\boldsymbol{\theta}$ *and* $\boldsymbol{z}$ *is defined as:*

$$\|\boldsymbol{\theta}\|^* \triangleq \sup_{\|z\|\leq 1} \boldsymbol{\theta}' z. \tag{5.2}$$

The following theorem presents the results to obtain the dual norm of the DRO-expert norm.

Theorem 5.2.1. *The dual norm of the* δ*-weighted DRO expert norm evaluated at some vector* $\tilde{\boldsymbol{\beta}} \triangleq (\boldsymbol{\beta}, 1) = (\boldsymbol{\beta}_K, \boldsymbol{\beta}_N, 1)$ *can be expressed as*

$$\|\tilde{\boldsymbol{\beta}}\|^*_{\delta, expert} \triangleq \|\boldsymbol{\beta}_K\|_2 + \|\boldsymbol{\beta}_N\|_{2,1} + \frac{1}{\delta}, \tag{5.3}$$

where $\|\boldsymbol{\beta}_N\|_{2,1} \triangleq \sum_{i=1}^{n} \|\boldsymbol{\beta}_i\|_2$. *As* $\delta \to \infty$, *it reduces to*

$$\lim_{\delta\to\infty} \|\tilde{\boldsymbol{\beta}}\|^*_{\delta,expert} = \|\boldsymbol{\beta}_K\|_2 + \|\boldsymbol{\beta}_N\|_{2,1}.$$

Proof. Using the definition of the dual norm, $\|\tilde{\boldsymbol{\beta}}\|^*_{\delta,\text{expert}}$ is the optimal value of the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{x}} \ \mathbf{x}_K'\boldsymbol{\beta}_K + \mathbf{x}_N'\boldsymbol{\beta}_N + y \\ &\text{s.t. } \|\mathbf{x}_K\|_2 + \|(\mathbf{x}_N, \delta y)\|_{2,\infty} \le 1. \end{aligned}$$

Using Hölder's inequality and Theorem 1 from [Chen and Paschalidis, 2020], we have:

$$\begin{aligned} \mathbf{x}_K'\boldsymbol{\beta}_K + \mathbf{x}_N'\boldsymbol{\beta}_N + y &= \mathbf{x}_K'\boldsymbol{\beta}_K + (\mathbf{x}_N, \delta y)'(\boldsymbol{\beta}_N, \frac{1}{\delta}) \\ &\le \|\mathbf{x}_K\|_2\|\boldsymbol{\beta}_K\|_2 + \|(\mathbf{x}_N, \delta y)\|_{2,\infty}\|(\boldsymbol{\beta}_N, \frac{1}{\delta})\|_{2,1} \\ &\le \|\boldsymbol{\beta}_K\|_2 + \|\boldsymbol{\beta}_N\|_{2,1} + \frac{1}{\delta}, \end{aligned}$$

where the last inequality is due to $\|\mathbf{x}_K\|_2 + \|(\mathbf{x}_N, \delta y)\|_{2,\infty} \le 1$. Thus, we have derived the dual norm of the δ-weighted DRO expert norm in the form of Eq. (5.3). □

Using the results of Theorem 5.2.1, when $\delta \to \infty$, with M i.i.d samples $(\mathbf{x}^{(i)}, y^{(i)}), i \in [M]$, Eq. (5.1) under the DRO-expert norm can be reformulated as:

$$\inf_{\boldsymbol{\beta}} \frac{1}{M} \sum_{i=1}^{M} |y^{(i)} - \boldsymbol{\beta}'\mathbf{x}^{(i)}| + \varepsilon(\|\boldsymbol{\beta}_K\|_2 + \|\boldsymbol{\beta}_N\|_{2,1}). \tag{5.4}$$

5.2.2 Continuous Variables with Overlapping Groups

In this subsection, we explore the scenario with overlapping feature groups. Different from Section 5.2.1, here we make the assumption that the feature groups $\mathbf{x}_K, \mathbf{x}_i, i \in [n]$ can overlap with each other. According to [Chen and Paschalidis, 2020], applying (5.4) to the overlapping groups could lead to a scenario where setting one group to zero shrinks its covariates to zero even if they belong to other groups, in which case these other groups will not be entirely selected. To address this issue, the concept of latent variables was introduced

in [Obozinski et al., 2011], and it was shown that GLASSO with the latent variables can induce a solution that would select entire groups of covariates.

Define the latent variables $\mathbf{v}_K, \mathbf{v}_i \in \mathbb{R}^p$ for the feature groups $\mathbf{x}_K$ and $\mathbf{x}_i, i \in [n]$. The non-zero elements of $\mathbf{v}_i$ are a subset of the features that are in a group $\mathbf{x}_i$. We assume that $\exists\ i_1, i_2$ such that $\mathbf{x}_{i_1} \cap \mathbf{x}_{i_2} \neq \varnothing$. For simplicity of notation, we use $\mathbf{x}_i$ to denote both the features in group i and the feature values.

Similar to [Chen and Paschalidis, 2020], we treat the response y as a deterministic quantity for simplicity. The following DRO expert norm is defined in the predictor space:

$$\|\mathbf{x}\|_{\text{expert}} \triangleq \|\mathbf{x}_K\|_2 + \|\mathbf{x}_N\|_{2,\infty}. \tag{5.5}$$

Using the latent variables, the dual norm of (5.5) under overlapping feature groups can be expressed as:

$$\|\boldsymbol{\beta}\|^*_{\text{expert}} \triangleq \|\mathbf{v}_K\|_2 + \|\mathbf{v}_N\|_{2,1}, \tag{5.6}$$

where $\boldsymbol{\beta} = \mathbf{v}_K + \sum_{i=1}^n \mathbf{v}_i$.

The following theorem proves the dual norm formulation of the DRO-expert norm in the overlapping case.

Theorem 5.2.2. *The dual norm of* $\|\boldsymbol{\beta}\|^*_{expert}$ *evaluated at some vector* $\mathbf{x}$ *can be expressed as:*

$$(\|\boldsymbol{\beta}\|^*_{expert})^* \triangleq \|\mathbf{x}_K\|_2 + \|\mathbf{x}_N\|_{2,\infty} \tag{5.7}$$

.

Proof. $\|\boldsymbol{\beta}\|^*_{\text{expert}}$ being a norm, we can consider its dual norm evaluated at some vector $\mathbf{x}$ defined by:

$$\begin{aligned} (\|\boldsymbol{\beta}\|^*_{\text{expert}})^* = \sup_{\boldsymbol{\beta} \in \mathbb{R}^p} & \{\boldsymbol{\beta}'\mathbf{x}\} \\ \text{s.t.} & \|\boldsymbol{\beta}\|^*_{\text{expert}} \leq 1. \end{aligned} \tag{5.8}$$

Using Hölder's inequality, Theorem 1 from [Chen and Paschalidis, 2020], and Lemma 3

from [Obozinski et al., 2011], we can derive:

$$\begin{aligned}\boldsymbol{\beta}'\mathbf{x} &= (\mathbf{v}_K + \sum_{i=1}^{n}\mathbf{v}_i)'\mathbf{x} \\ &= \mathbf{v}_K'\mathbf{x}_K + (\sum_{i=1}^{n}\mathbf{v}_i)'\mathbf{x}_N \\ &\leq \|\mathbf{v}_K\|_2\|\mathbf{x}_K\|_2 + \|\mathbf{v}_N\|_{2,1}\|\mathbf{x}_N\|_{2,\infty} \\ &\leq \|\mathbf{x}_K\|_2 + \|\mathbf{x}_N\|_{2,\infty},\end{aligned}$$

where the first equality is due to the definition of the latent variables, and the last equality is due to $\|\mathbf{v}_K\|_2 + \|\mathbf{v}_N\|_{2,1} \leq 1$. Since for reflexive spaces, the dual norm of the dual norm is the original norm. Thus, we have proved that the dual norm of the Eq. (5.5) is in the form of Eq. (5.6). □

Therefore, with M i.i.d. samples $(\mathbf{x}^{(i)}, y^{(i)}), i \in [M]$, Eq. (5.1) under overlapping feature groups can be reformulated as:

$$\begin{aligned}&\inf_{\boldsymbol{\beta}} \frac{1}{M}\sum_{i=1}^{M}|y^{(i)} - \boldsymbol{\beta}'\mathbf{x}^{(i)}| + \varepsilon\mathcal{B}(\boldsymbol{\beta}) \\ &\text{where:} \quad \mathcal{B}(\boldsymbol{\beta}) = \min_{\substack{\mathbf{v}_1,\ldots,\mathbf{v}_n,\mathbf{v}_K \\ \boldsymbol{\beta}=\sum_{i=1}^{n}\mathbf{v}_i+\mathbf{v}_K}} (\|\mathbf{v}_K\|_2 + \|\mathbf{v}_N\|_{2,1}).\end{aligned} \tag{5.9}$$

5.3 Simulation Experiments on Synthetic Datasets

In this section, we will test DRO expert group regression on several synthetic datasets, and compare the results with other regularization methods. The data generation process follows the setting in [Chen and Paschalidis, 2018]:

1. Generate $\boldsymbol{\beta}^*$ based on the following rule:

$$(\boldsymbol{\beta}^*)^l = \begin{cases} 1, & \text{if } l \text{ is expert group,} \\ 0.5, & \text{if } l \text{ is non-sparse group,} \\ \approx 0, & \text{otherwise.} \end{cases} \tag{5.10}$$

2. Generate the predictor $\mathbf{x} \in \mathbb{R}^p$ from Gaussian distribution $\mathcal{N}(0,\Sigma)$, where $\Sigma = (\sigma_{i,j})_{i,j=1}^{p}$has diagonal elements equal to 1, and off-diagonal elements equal to:

$$(\sigma_{i,j})^l = \begin{cases} \rho_w, & \text{if predictors i and j are in the same group,} \\ 0, & \text{otherwise.} \end{cases} \tag{5.11}$$

   Here ρ_w represents within-group correlation which is the correlation between predictors in the same group. The between-group correlation is 0.

3. Generate the response variable $\mathbf{y}$ with a probability of being drawn from the outlying distribution (the proportion of outliers). Generate a uniform random variable r on [0, 1]:

$$(\mathbf{y})^l \sim \begin{cases} \mathcal{N}(x'\beta^*, \sigma^2), & \text{if } r \leq 1-q, \\ \mathcal{N}(x'\beta^*, \sigma^2) + 5\sigma, & \text{otherwise.} \end{cases} \tag{5.12}$$

   where σ^2 is the intrinsic variance of $\mathbf{y}$.

The dataset is divided into a training set and a test set. Within the training set, cross-validation is employed to optimize the regularization parameters. The optimal regularization parameter is selected based on its ability to achieve the lowest Median Absolute Deviation (MAD) on the validation set. The selection of values for tuning the parameters is guided by the methodologies proposed by Hastie [Hastie et al., 2017] with adjustments to apply MAD. For this study, the specified range for these parameters is set as follows:

$$\sqrt{exp(lin(log(0.005 * \|X'y\|_\infty), log(0.05 * \|X'y\|_\infty), 50))}$$

In this context, $lin(a,b,n)$ is a function that accepts scalars a and b, along with an integer n, and generates n evenly spaced values between a and b.

The performance of all the considering methods is evaluated as the following factors are varied:

- *Signal to Noise Ratio (SNR)*,

$$SNR = \frac{(\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}}{\sigma^2}.$$

- The correlation between predictors: ρ, which takes values in (0.1, 0.2,...,0.9). The performance metrics we use include:

- *Mean Absolute Error (MAE)* of the test set $(\mathbf{x}_i, y_i), i \in [M]$, which is defined to be $\sum_{i=1}^{M} |y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}'|/M$, where $\hat{\boldsymbol{\beta}}$ is the estimation of the true coefficients $\boldsymbol{\beta}^*$ using the training set.

- *Relative Risk (RR)* of $\hat{\boldsymbol{\beta}}$ defined as:

$$RR(\hat{\boldsymbol{\beta}}) \triangleq \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{(\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta}},$$

  with a perfect estimation to be score 0 and the worst estimation to be score 1.

- *Relative Test Error (RTE)* of $\hat{\boldsymbol{\beta}}$ defined as:

$$RTE(\hat{\boldsymbol{\beta}}) \triangleq \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \sigma^2}{\sigma^2},$$

  with a perfect perfect estimation to be score 1 and the worst estimation to be score = SNR+1.

- *Proportion of Variance Explained (PVE)* of $\hat{\boldsymbol{\beta}}$ defined as:

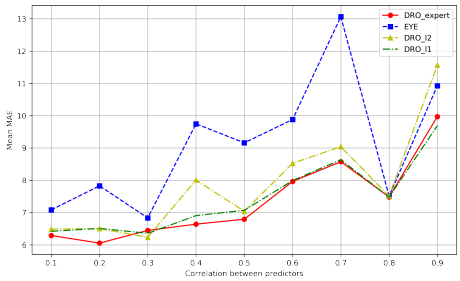
$$PVE(\hat{\boldsymbol{\beta}}) \triangleq \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \sigma^2}{(\boldsymbol{\beta}^*)'\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma^2},$$

  with a perfect perfect estimation to be score = $\frac{SNR}{SNR+1}$ and the worst estimation to be score 0.

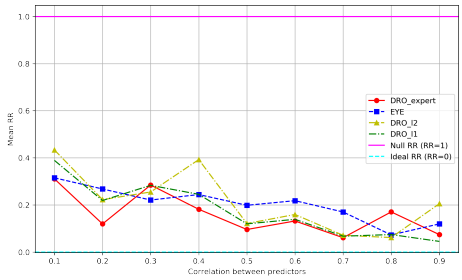
5.3.1 Non-overlapping Group with Continuous Variables

We generated 20 datasets, each with N = 100 data points. Each data point contains M = 28 variables, which can be classified into 5 non-overlapped groups with size The group size is: $p_1 = 3, p_2 = 5, p_3 = 7, p_4 = 4, p_5 = 9$. The coefficients are set to include both dense and sparse coefficients groups. The expert group is chosen to be the second group and the fourth group with relatively dense features. The goal is to enable the model to identify and select the densely populated expert group while excluding the sparser groups that could potentially impair the model's performance. The average values of the performance metrics over the 20 datasets were measured under varying SNRs and varying correlations between predictors. When SNR is varied, the correlation between predictors is set to be 0.8 times a random noise generated from a uniform distribution from [0.4, 0.7]. When the correlation is varied the SNR is fixed to be 1. The performances in terms of all four metrics are shown in Figure 5·1 and Figure 5·2. We compared our proposed method with several other regularization approaches, including, the l_1-norm and l_2-norm induced Wasserstein formulations (DRO-l_1, DRO-l_2) [Chen and Paschalidis, 2018], and the EYE penalty [Wang et al., 2018].

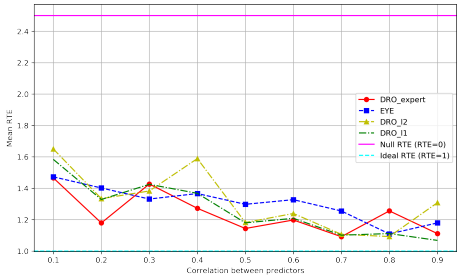
As the correlation increases or the SNR decreases, thereby complicating the estimation challenge, all performance metrics tend to deteriorate. Among them, the DRO-expert norm consistently exhibits the superior and most stable performance across the four assessed metrics. When the predictor correlation is either excessively low (≤ 0.2) or exceedingly high (≥ 0.8), none of the evaluated methods perform particularly well. Nonetheless, the DRO-expert norm still shows a more stable estimation performance. It also remains distinctly advantageous at moderate correlation levels, clearly surpassing the alternative approaches and contributing to a model closer to the ideal estimator. Variations in SNR reveal that the performance of the EYE penalty and DRO-l_2 is more erratic in comparison to the DRO-expert and DRO-l_1. This inconsistency can be attributed to the deliberate introduction of feature sparsity within the synthetic dataset. The sparse coefficients suggest that only a



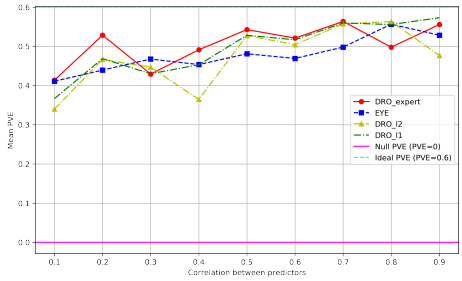
(a) Mean Absolute Error.



(b) Relative Risk.



(c) Relative Test Error.



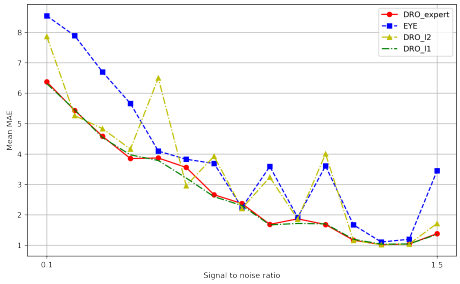
(d) Proportion of Variance Explained.

Figure 5·1: The impact of predictor correlation on the performance metrics: non-overlapping cases

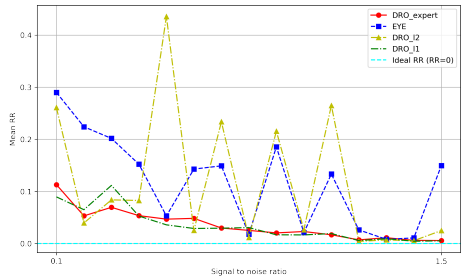
limited number of predictors are relevant to our predictive model. Therefore, we require a metric capable of isolating just this subset of pertinent predictors. The l_2 norm, which considers all the coordinates, might lead to the inclusion of numerous redundant features. Consequently, the l_1-induced approach that focuses specifically on the relevant predictors can enhance the model's efficiency and accuracy. In contrast to the l_2-induced methodologies, the DRO-expert norm demonstrates its effectiveness by steering the model toward group-wise sparser solutions and preserving feature groups that encapsulate more pertinent information.

5.3.2 Overlapping Group with Continuous Variables

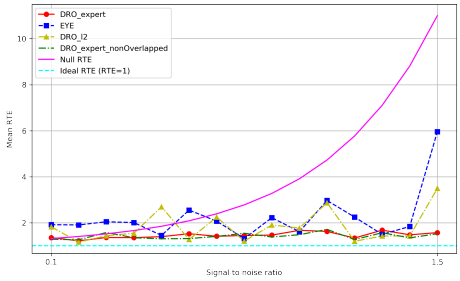
We generated 20 datasets each with N = 100 data points. Each data point contains M = 20 variables, which are the union of 5 overlapped groups with size $p_1 = 3, p_2 = 5, p_3 = 7, p_4 = 4, p_5 = 9$. Each group has variables that overlap with the successive groups. The designated expert group comprises the second group and the fourth group. Each of them overlaps with



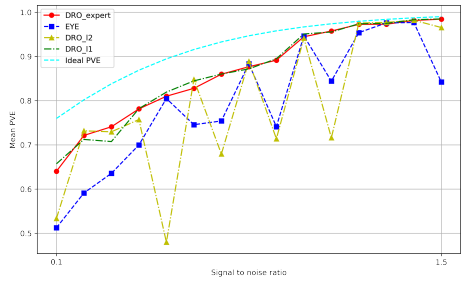
(a) Mean Absolute Error.



(b) Relative Risk.



(c) Relative Test Error.

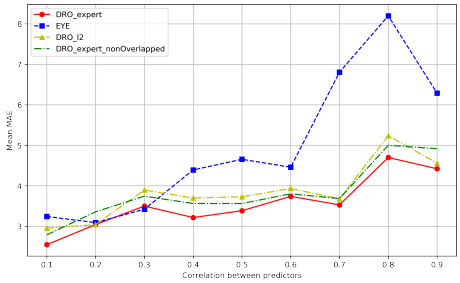


(d) Proportion of Variance Explained.

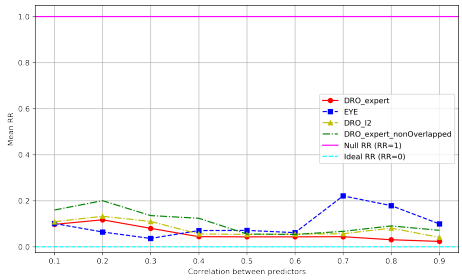
Figure 5·2: The impact of SNR on the performance metrics: non-overlapping cases

the successive non-expert groups. Performance metrics were averaged over 20 datasets to assess the impact of varying SNR and correlations between predictors. Equation (5.9) was solved by decomposing the coefficient vector into the sum of latent vectors, each supported by one of the groups. For varied SNR scenarios, the correlation between predictors was set at 0.8 multiplied by random noise, drawn from a uniform distribution ranging between 0.4 and 0.7. Conversely, when varying the correlation, the SNR was held constant at 1. To compare with the DRO-expert norm in the non-overlapping case, we implemented a strategy to segregate overlapping variables between groups. This involved assigning each overlapping variable to the first group that contains it, thereby mimicking a non-overlapping scenario.

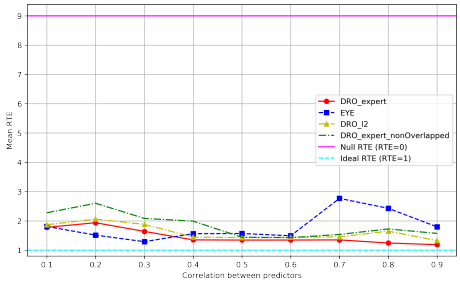
Mirroring the trend observed in non-overlapping experiments, all performance metrics deteriorate with increasing correlation or decreasing SNR. Overall, in cases of variable overlap, all four metrics displayed improved performance compared to non-overlapping



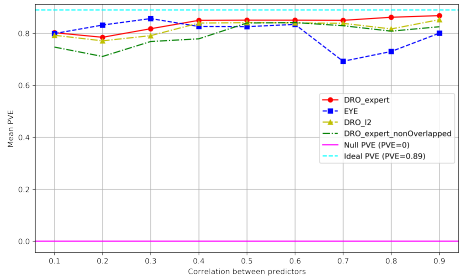
(a) Mean Absolute Error.



(b) Relative Risk.



(c) Relative Test Error.



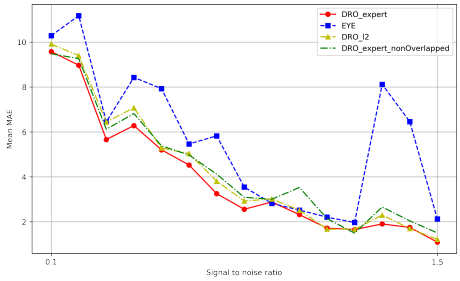
(d) Proportion of Variance Explained.

Figure 5·3: The impact of predictor correlation on the performance metrics: overlapping cases

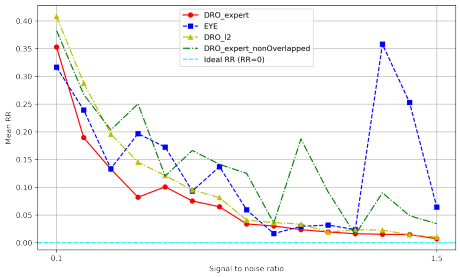
scenarios, due to reduced sparsity in the setting. Nonetheless, the DRO-expert norm's dominance remained evident, particularly in the MAE and RR metrics, underscoring its robustness even with overlapping variable groups.

We note that the DRO_l_1 norm, despite not considering overlapping structures, achieves results similar to the DRO-expert overlapping norm. Both approaches lead to sparse solutions. Nonetheless, in situations characterized by high correlation among predictors or a low SNR, the overlapping configuration outperforms the l_1 norm. This is evident when we consider the group structure, as reflected in the outcomes of the four metrics. This suggests that the overlapping norm's ability to account for group dynamics provides an advantage in these specific contexts.

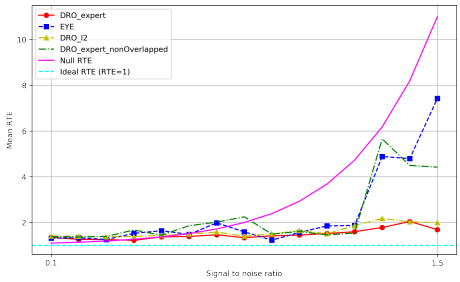
Below, we present a summary of our key findings from the various experiments conducted:



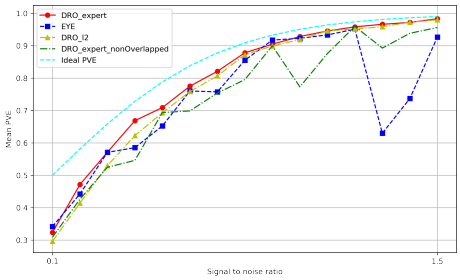
(a) Mean Absolute Error.



(b) Relative Risk.



(c) Relative Test Error.



(d) Proportion of Variance Explained.

Figure 5·4: The impact of SNR on the performance metrics: overlapping cases

- The DRO-expert formulation demonstrates greater stability in estimation compared to other methods, particularly in scenarios where there is a high correlation between predictors.

- The DRO-expert norm consistently outperforms other methods across all four assessed metrics. It maintains stable performance even in challenging conditions of high correlation or low SNR.

- The DRO-expert norm leads to sparser solutions at the group level and preserves feature groups containing more relevant information, unlike the l_2-based methods which exhibit more erratic performance.

- In settings with overlapping variables, The DRO-expert norm stands out, especially in MAE and RR metrics, proving its effectiveness even with overlapping groups. Although the DRO_l_1 norm performs comparably to the DRO-expert norm, it is less

effective in cases of high predictor correlation or low SNR.

5.4 Experiments on Real Datasets

After verifying the DRO expert group norm on the synthetic datasets, we applied it to two different scales of real classification problems. Both of the datasets are used for tackling healthcare problems, where credibility and interpretability are crucial for doctors and patients to adopt the model results.

5.4.1 Application on COVID ICU Risk Prediction Problem

The study aimed to develop predictive models that forecast the likelihood of COVID-19 patients requiring intensive care unit (ICU) admission, thereby providing an indicator of the disease's severity. These models are designed to predict whether a patient with COVID-19 will need ICU care, using various health data and patient-specific factors.

Dataset

We analyzed anonymized data from 7,102 patients diagnosed with SARS-CoV-2 using RT-PCR at BMC in 2020, spanning from January 1st to December 31st. Additional details about the dataset and its features are provided in Chapter 3.

Analysis

The dataset was divided into a training set (80%) and a test set (20%). The training set was used to optimize algorithm parameters through 5-fold cross-validation and a grid search approach. The regularization strength parameters were set on a logarithmic scale ranging from 0.001 to 100. The test set was used to compute performance metrics, and we report the mean and standard deviation of these metrics across five random splits of the data. The performance is evaluated using two metrics: AUC and F1 score. Additionally, we have reported the overall sparsity and credibility for each model. Here, we followed the

definition given in [Wang et al., 2018] for sparsity, the percentage of features whose weights are smaller than 0.01 of the largest absolute feature weight. Credibility is defined as the percentage of expert features whose weights exceed 0.01 of the largest absolute feature weight.

Table 5.1 summarizes the performance metrics on the test set for different regularization methods.

In comparing various regularization methods, the DRO-expert and DRO-expert (overlapped) approaches notably excel in predictive performance. DRO-expert achieves an impressive average AUC of 94.74%, but is slightly outperformed by DRO-expert (overlapped), which boasts a higher AUC of 95.85%. This superiority extends to other metrics as well, with DRO-expert attaining a mean F1-score of 78.03%, indicative of a balanced precision-recall trade-off. However, it is marginally surpassed by DRO-expert (overlapped) which leads with an F1-score of 81.27%.

Not only does the DRO-expert (overlapped) excel in predictive performance, but it also achieves a commendable balance between sparsity and credibility. With a sparsity of 0.41, it ensures that the model remains interpretable by not overloading it with non-contributing features. Furthermore, it retains the highest level of credibility at 0.86, signifying that 86% of the expert features are weighted above the defined importance threshold. Compared with the high sparsity achieved by DRO non-overlapped norm, the decreased sparsity in the overlapped variant suggests a slightly more complex feature set which, interestingly, does not compromise its interpretability, contributing instead to its superior performance.

In essence, while DRO-expert signifies simplicity and parsimony in model design, the overlapping approach offers a nuanced balance, boosting predictive capabilities without significantly affecting interpretability or simplicity. This positions the overlapping model as superior in predictive performance.

Table 5.1: COVID ICU prediction model. 'std' refers to the standard deviation of the corresponding metric. Sparsity is defined as the percentage of features with weights smaller than 0.01 of the largest absolute feature weight. Credibility is defined as the percentage of expert features whose weights exceed 0.01 of the largest absolute feature weight.

Method	AUC		F1-score		Sparsity	Credibility
	mean	std	mean	std		
DRO-$l1$	93.72%	0.46%	77.47%	1.89%	0.18	0.85
DRO-$l2$	92.93%	1.06%	74.38%	3.62%	0	1.00
EYE	93.99%	0.54%	74.38%	0.79%	0.23	0.76
DRO-expert	94.74%	0.55%	78.03%	1.57%	0.67	0.42
DRO-expert (overlapped)	95.85%	0.61%	81.28%	2.88%	0.41	0.86

5.4.2 Application on High Hypertension Identification Problem

To further assess the efficacy of our method, we applied the DRO-expert norm to a hypertensive dataset. The aim was to develop models predicting patients with uncontrolled hypertension, defined as having systolic blood pressure (SBP) records above 160 mmHg.

Dataset

We utilized data from the Boston Medical Center, spanning from January 1, 2012, to January 1, 2020. Additional details about the dataset and its features are provided in Chapter 4.

Analysis

The dataset was divided into a training set (80%) and a test set (20%). We optimized algorithm parameters in the same way as in the COVID dataset. For each approach, we reported the AUC and F1-score of the test set. Table 5.2 summarizes the results.

Overall, we observed a similar trend as what we observed for the COVID dataset, with DRO-expert (overlapped) achieving slightly better than the non-overlapped case. Specifically, DRO-expert (overlapped) leads with the mean AUC at 74.90%, which is indicative of its superior capability to distinguish between patient outcomes. It also achieves the second

highest mean F1-score at 82.35%, reinforcing its status as the model with the best harmonic mean of precision and recall, essential for a balanced binary classification.

It's important to highlight that by implementing recursive feature selection in this dataset and consequently retaining only the most contributing features, we inevitably reduced the inherent sparsity of the original dataset. As a result, methodologies that typically yield sparser models, such as the DRO-expert and EYE penalty, experienced a decrease in performance. Despite this, the DRO-expert method still managed to achieve a comparable AUC at 74.21% with sparsity 0.87, suggesting a parsimonious model. This suggests that while the reduction in sparsity impacts the overall effectiveness of these approaches, the DRO-expert technique maintains a relative advantage in terms of performance.

In comparison to DRO-l_1, DRO-l_2, and DRO-expert, DRO-expert (overlapped) maintains a moderate sparsity value of 0.52. This is in contrast to DRO-l_2, which, despite a comparable AUC, has a sparsity of 0, suggesting a fully dense model that may be less generalized. The EYE penalty, while having the lowest sparsity value at 0.93, indicating a very simplified model, falls short in both AUC and F1-score, highlighting a potential trade-off between sparsity and predictive performance.

In terms of credibility, DRO-l_1 and DRO-l_2 retain a higher percentage of dense expert features, attributable to the inherently lower sparsity of these models. Notably, the credibility ratio for DRO-l_1 is lower than its overall dense feature percentage 88%, indicating the loss of some expert features. However, among the other sparse models, DRO-expert (overlapped) achieves the highest credibility while also maintaining strong predictive performance. Additionally, it achieves a sparsity level of 0.48, which enhances the model's interpretability.

Overall, DRO-expert (overlapped) emerges as the most potent model, with a commendable balance between model credibility and predictive accuracy. Its moderate sparsity value suggests a balanced approach, ensuring sufficient model complexity to capture essential

Table 5.2: Uncontrolled Hypertension prediction model. 'std' refers to the standard deviation of the corresponding metric. Sparsity is defined as the percentage of features with weights smaller than 0.01 of the largest absolute feature weight. Credibility is defined as the percentage of expert features whose weights exceed 0.01 of the largest absolute feature weight.

Method	**AUC**		**F1-score**		**Sparsity**	**Credibility**
	mean	**std**	**mean**	**std**		
DRO-$l1$	74.85%	0.40%	82.23%	0.14%	0.12	0.81
DRO-$l2$	75.27%	0.34%	82.40%	0.13%	0	1.00
EYE	72.46%	0.58%	82.04%	0.12%	0.93	0.19
DRO-expert	74.21%	0.44%	82.25%	0.13%	0.87	0.23
DRO-expert (overlapped)	74.90%	0.40%	82.35%	0.23%	0.52	0.58

features while avoiding overfitting. This positions DRO-expert (overlapped) as a preferable choice for classification problems with requirements on both reliability and interpretability.

5.5 Limitations

The proposed approach, while demonstrating significant advantages in both synthetic and real-world experiments, particularly with the DRO-expert norm in overlapping cases, does have its limitations. Although our method has been adapted to accommodate complex overlapping group structures, it remains confined to a linear setting. Future research could explore extending this approach to non-linear models and incorporating various forms of expert knowledge to enhance its applicability.

5.6 Conclusion

This study introduces the Wasserstein DRO formulation in the context of credible learning and grouped feature selection. We defined a special norm that expands upon the concept of GLASSO, with a specific focus on improving model's interpretability and credibility. To achieve this, we propose the DRO-expert regularization penalty, demonstrating how GLASSO can be effectively adapted for use in tandem with expert knowledge. Additionally,

we explore its application in scenarios featuring overlapping group structures, a common occurrence in healthcare where a single medical trait may require analysis from multiple angles. Using empirical data from both synthetic and real-world datasets, we showcase the efficacy of the DRO-expert norm in these contexts, highlighting its ability to seamlessly combine model performance with expert knowledge.

Our study contributes to the field by providing a comprehensive approach that balances interpretability and credibility, with the integration of expert insights. This balance is essential for crafting effective and reliable models in healthcare analytics. The DRO-expert model sets a new benchmark in the realm of healthcare ML models by integrating domain knowledge and expertise into the predictive modeling process. The findings and methodologies of this study could pave the way for future applications of ML in various domains requiring expertise. We focus on developing models that are not only accurate and interpretable but also credible and in line with established medical knowledge, potentially revolutionizing the application of ML in fields necessitating expert input.

Chapter 6

Conclusions

6.1 Summary of the Thesis

In this dissertation, we explore and develop a comprehensive range of models and methods aimed at improving predictive analytics, prescriptive analytics, and the interpretability and credibility of ML models in healthcare. Our models are developed and tested on a comprehensive dataset from a safety-net hospital, rich with clinical information and encompassing a diverse patient population. This diversity ensures that our models are widely generalizable and effective in real-world clinical settings. Predictive analytics involves creating models that forecast future patient outcomes based on historical and real-time data, aiming to enhance accuracy and reliability through advanced ML algorithms for early diagnosis and timely intervention. Prescriptive analytics goes further by providing actionable recommendations, designing ML frameworks that not only predict outcomes but also suggest optimal actions for clinicians, personalizing patient care, improving treatment protocols, and efficiently allocating resources. Significantly, our prescriptive models have undergone clinical evaluation and have been selected by the hospital for an ongoing clinical trial. This endorsement from the clinical community underscores the robustness and reliability of our models, indicating their potential for immediate implementation within the healthcare system. Another crucial aspect of our research is enhancing the interpretability and credibility of ML models in healthcare, ensuring they are accurate, transparent, and understandable to medical practitioners. This involves developing methods that align model predictions with established medical knowledge and introducing specialized norms within DRO to balance

robustness with interpretability.

By integrating these approaches, our work aims to contribute to the development of ML models that are not only powerful in their predictive and prescriptive capabilities but also trusted and adopted by healthcare professionals, ultimately leading to improved patient outcomes and more efficient healthcare systems.

The first study in the thesis aims to identify patients with not well-controlled hypertension using readily available demographic and socioeconomic features and elucidate important predictive variables. Models were developed to predict which patients had uncontrolled hypertension defined as SBP records exceeding 160 mmHg. The predictive model of high SBP reached an AUC of $74.49\% \pm 0.23\%$. Age, race, SDoH, mental health, and cigarette use were predictive of high SBP. Being Black or having critical social needs led to higher probability of uncontrolled SBP. To mitigate model bias and elucidate differences in predictive variables, two separate models were trained for Black and White patients. Black patients face a 4.7 times higher FPR and a 0.58 times lower FNR compared to White patients. Decision threshold differentiation was implemented to equalize FNR. Race-specific models reveal a different set of social variables predicting high SBP, with Black patients being affected by structural barriers (e.g., food and transportation) and White patients by personal and demographic factors (e.g., marital status). The study shows that models using non-clinical factors can predict which patients exhibit poorly controlled hypertension. Racial and SDoH variables are significant predictors but lead to biased predictive models. Race-specific models are not sufficient to resolve such biases and require further decision threshold tuning. A host of structural socioeconomic factors are identified to be targeted to reduce disparities in hypertension control.

The second part of the thesis aims to facilitate personalized treatment that accounts for a richer set of patient characteristics through data-driven models. Models were developed to recommend a class of antihypertensive medications for each hypertensive patient based on

their characteristics. Regression immunized against outliers was combined with a nearest neighbor approach to associate with each patient an affinity group of other patients. This group was then used to make predictions of future SBP under each prescription type. For each patient, we leveraged these predictions to select the class of medication that minimized their future predicted SBP. The proposed model, built with a distributionally robust learning procedure, leads to a reduction of 14.28 mmHg in SBP, on average. This reduction is 70.30% larger than the reduction achieved by the standard-of-care and 7.08% better than the corresponding reduction achieved by the second best model which uses ordinary least squares regression. All derived models outperform following the previous prescription or the current ground truth prescription in the record. We randomly sampled and manually reviewed 350 patient records; 87.71% of these model-generated prescription recommendations passed a sanity check by clinicians. Our data-driven approach for personalized hypertension treatment yielded significant improvement compared to the standard of care. The model implied the potential benefits of computationally deprescribing and can support situations with clinical equipoise.

The third part of the thesis is to develop predictive models of COVID-19 outcomes, elucidate the influence of socioeconomic factors, and assess algorithmic racial fairness using a racially diverse patient population with high social needs. Linear and nonlinear classification methods were applied. A score based on a recurrent neural network and a transformer architecture was developed to capture the dynamic evolution of vital signs. Combined with patient characteristics, clinical variables, and hospital occupancy measures, this dynamic vital score was used to train predictive models. Hospitalizations can be predicted with an AUC of 92% using symptoms, hospital occupancy, and patient characteristics, including social determinants of health. Parsimonious models to predict intensive care, mechanical ventilation, and mortality that used the most recent labs and vitals exhibited AUCs of 92.7%, 91.2%, and 94%, respectively. Early predictive models, using labs and vital signs closer to

admission had AUCs of 81.1%, 84.9%, and 92%, respectively. The most accurate models exhibit racial bias, being more likely to falsely predict that Black patients will be hospitalized. Models that are only based on the dynamic vital score exhibited accuracies close to the best parsimonious models, although the latter also used laboratories.

The fourth part of the thesis investigates methods to align ML model predictions with existing medical knowledge, aiming to build trust and facilitate acceptance among healthcare providers. This study introduces the Wasserstein DRO formulation within the context of credible learning. By employing a special norm, we highlight the enhanced interpretability provided by the GLASSO algorithm and emphasize how the expert norm bolsters credibility based on that. We expanded the application of the norm in scenarios with overlapping group structures. Our empirical evidence from synthetic and real-world datasets demonstrates the superiority of the DRO-expert norm in these scenarios, effectively merging model performance with expert knowledge. This approach enhances GLASSO's applicability, particularly in healthcare, where predictor groupings are informed by domain expertise. The paper contributes to the field by providing a comprehensive approach that balances interpretability and credibility with expert insight integration, essential for effective and trustworthy models in healthcare analytics.

6.2 Future Work

A primary motivation for this dissertation is to develop interpretable prediction models and ensure their reliability for use in actual healthcare settings. In partnership with Boston Medical Center, we are launching a clinical trial to test the personalized hypertension prescription model. Clinicians will gather patient data for the forthcoming week, and the trained model will generate recommended prescriptions for each patient. Clinicians will then assess the model's feedback, finalize prescription decisions, and monitor patient outcomes based on these recommendations. This clinical trial is crucial for evaluating the performance

of our data-driven models in a practical healthcare environment. Moreover, it seeks to illustrate the potential advantages of incorporating these models into clinical practice, such as increased accuracy in hypertension management, more individualized treatment plans, and improved patient outcomes. The trial will also shed light on how these models can streamline clinical workflows, alleviate the burden on healthcare providers, and ultimately enhance disease management and patient care.

This dissertation presents models developed utilizing the Wasserstein Distributionally Robust Optimization (DRO) framework to enhance model credibility. Although it has shown significant performance benefits, they are currently constrained to linear applications. Future research should explore extending these methods to non-linear and deep learning models to expand their applicability.

Recent research underscores the potential of combining DRO with deep learning, particularly in healthcare contexts. Integrating neural networks with DRO can improve the robustness of predictive models by effectively managing distributional uncertainties. This combination can enhance the accuracy and reliability of clinical decision-making tools [**?**,Xu et al., 2024]. Additionally, incorporating expert knowledge into these models can improve their interpretability and practical application. In healthcare, this integration means developing models that not only predict outcomes with greater accuracy but also provide insights that are comprehensible and actionable for clinicians. Aligning model recommendations with clinical guidelines ensures that they are more readily adopted in practice.

References

2020 census memo.pdf. [Online; accessed 2022-02-24].

Welcome to uszipcode Documentation — uszipcode 0.2.6 documentation.

Araos, P., Figueroa, S., and Amador, C. A. (2020). The role of neutrophils in hypertension. *International Journal of Molecular Sciences*, 21(22):8536.

Backenroth, D., Chase, H. S., Wei, Y., and Friedman, C. (2017). Monitoring prescribing patterns using regression and electronic health records. *BMC Medical Informatics and Decision Making*, 17(1):175.

Bakin, S. (1999). Adaptive regression and model selection in data mining problems. (Doctoral dissertation, Australian National University). http://hdl.handle.net/1885/9449.

Beheshti, I., Ganaie, M. A., Paliwal, V., Rastogi, A., Razzak, I., and Tanveer, M. (2022). Predicting Brain Age Using Machine Learning Algorithms: A Comprehensive Evaluation. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1432–1440.

Ben-David, A. (1995). Monotonicity Maintenance in Information-Theoretic Machine Learning Algorithms. *Machine Learning*, 19(1):29–43.

Benetos, A., Petrovic, M., and Strandberg, T. (2019). Hypertension Management in Older and Frail Older Patients. *Circulation Research*, 124(7):1045–1060.

Bertsimas, D., Borenstein, A., Mingardi, L., Nohadani, O., Orfanoudaki, A., Stellato, B., Wiberg, H., Sarin, P., Varelmann, D. J., Estrada, V., Macaya, C., and Gil, I. J. N. (2021). Personalized prescription of ACEI/ARBs for hypertensive COVID-19 patients. *Health Care Management Science*, 24(2):339–355.

Bertsimas, D., Kallus, N., Weinstein, A. M., and Zhuo, Y. D. (2017). Personalized Diabetes Management Using Electronic Medical Records. *Diabetes Care*, 40(2):210–217.

Bhargava, A., Fukushima, E. A., Levine, M., Zhao, W., Tanveer, F., Szpunar, S. M., and Saravolatz, L. (2020). Predictors for severe COVID-19 infection. *Clinical Infectious Diseases*, 71(8):1962–1968. https://doi.org/10.1093/cid/ciaa674.

Bourgault, C., Sénécal, M., Brisson, M., Marentette, M. A., and Grégoire, J.-P. (2005). Persistence and discontinuation patterns of antihypertensive therapy among newly treated patients: a population-based study. *Journal of Human Hypertension*, 19(8).

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brown, L. E., Burton, R., Hixon, B., Kakade, M., Bhagalia, P., Vick, C., Edwards, A., and Hawn, M. T. (2012). Factors influencing emergency department preference for access to healthcare. *Western Journal of Emergency Medicine*, 13(5):410.

Calhoun, D. A., Jones, D., Textor, S., Goff, D. C., Murphy, T. P., Toto, R. D., White, A., Cushman, W. C., White, W., Sica, D., Ferdinand, K., Giles, T. D., Falkner, B., and Carey, R. M. (2008). Resistant Hypertension: Diagnosis, Evaluation, and Treatment. *Hypertension*, 51(6):1403–1419.

Carey, R. M., Sakhuja, S., Calhoun, D. A., Whelton, P. K., and Muntner, P. (2019). Prevalence of Apparent Treatment-Resistant Hypertension in the United States. *Hypertension*, 73(2):424–431.

Carnethon, M. R., Pu, J., Howard, G., Albert, M. A., Anderson, C. A., Bertoni, A. G., Mujahid, M. S., Palaniappan, L., Taylor, H. A., Willis, M., and Yancy, C. W. (2017). Cardiovascular Health in African Americans: A Scientific Statement From the American Heart Association. *Circulation*, 136(21):e393–e423.

Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

CDC (2020). Cases, Data, and Surveillance. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html.

CDC (2024a). High Blood Pressure Facts. https://www.cdc.gov/high-blood-pressure/data-research/facts-stats/index.html.

CDC (2024b). Weekly COVID-19 Vaccination Dashboard | COVIDVaxView | CDC. https://www.cdc.gov/vaccines/imz-managers/coverage/covidvaxview/interactive/vaccination-dashboard.html.

Chen, R. and Paschalidis, I. (2019). Selecting Optimal Decisions via Distributionally Robust Nearest-Neighbor Regression. In *Advances in Neural Information Processing Systems*, volume 32.

Chen, R. and Paschalidis, I. C. (2018). A Robust Learning Approach for Regression Models Based on Distributionally Robust Optimization. *Journal of machine learning research: JMLR*, 19(1):517–564.

Chen, R. and Paschalidis, I. C. (2020). Distributionally robust learning. *Foundations and Trends® in Optimization*, 4(1-2):1–243.

Chen, R., Paschalidis, I. C., Hatabu, H., Valtchinov, V. I., and Siegelman, J. (2019). Detection of unwarranted ct radiation exposure from patient and imaging protocol meta-data using regularized regression. *European journal of radiology open*, 6:206–211.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, San Francisco, California, USA. Association for Computing Machinery.

Chobanian, A. V. (2003). The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood PressureThe JNC 7 Report. *JAMA: The Journal of the American Medical Association*, 289(19):2560.

Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., and Sun, J. (2017). GRAM: Graph-based Attention Model for Healthcare Representation Learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 787–795, New York, NY, USA. Association for Computing Machinery.

Chopra, V., Toner, E., Waldhorn, R., and Washer, L. (2020). How Should U.S. Hospitals Prepare for Coronavirus Disease 2019 (COVID-19)? *Annals of Internal Medicine*, 172(9):621–622. Publisher: American College of Physicians.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

Cushman, W. C., Ford, C. E., Cutler, J. A., Margolis, K. L., Davis, B. R., Grimm, R. H., Black, H. R., Hamilton, B. P., Holland, J., Nwachuku, C., Papademetriou, V., Probstfield, J., Wright Jr., J. I., Alderman, M. H., Weiss, R. J., Piller, L., Bettencourt, J., Walsh, S. M., and Group, F. T. A. C. R. (2002). Original Papers. Success and Predictors of Blood Pressure Control in Diverse North American Settings: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *The Journal of Clinical Hypertension*, 4(6):393–404.

de la Vega, P. B., Losi, S., Martinez, L. S., Bovell-Ammon, A., Garg, A., James, T., Ewen, A. M., Stack, M., DeCarvalho, H., and Sandel, M. (2019). Implementing an EHR-based screening and referral system to address social determinants of health in primary care. *Medical care*, 57:S133–S139.

Dohi, Y., Thiel, M. A., Bühler, F. R., and Lüscher, T. F. (1990). Activation of endothelial L-arginine pathway in resistance arteries. Effect of age and hypertension. *Hypertension*, 16(2):170–179.

ElShawi, R., Sherif, Y., Al-Mallah, M., and Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, 37(4):1633–1650.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., and Socher, R. (2021). Deep learning-enabled medical computer vision. *npj Digital Medicine*, 4(1):1–9. Publisher: Nature Publishing Group.

Fava, C., Sjögren, M., Montagnana, M., Danese, E., Almgren, P., Engström, G., Nilsson, P., Hedblad, B., Guidi, G. C., Minuz, P., and Melander, O. (2013). Prediction of Blood Pressure Changes Over Time and Incidence of Hypertension by a Genetic Risk Score in Swedes. *Hypertension*, 61(2):319–326.

Franklin, D. (2024). COVID-19 Vaccination Coverage, and Rates of SARS-CoV-2 Infection and COVID-19–Associated Hospitalization Among Residents in Nursing Homes — National Healthcare Safety Network, United States, October 2023–February 2024. *MMWR. Morbidity and Mortality Weekly Report*, 73.

Frost, H. R. and Amos, C. I. (2017). Gene set selection via LASSO penalized regression (SLPR). *Nucleic Acids Research*, 45(12):e114.

Gao, R. and Kleywegt, A. J. (2022). Distributionally Robust Stochastic Optimization with Wasserstein Distance. arXiv:1604.02199 [math].

Goldstein, B. A., Navar, A. M., Pencina, M. J., and Ioannidis, J. P. A. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 24(1):198–208.

Gong, J., Ou, J., Qiu, X., Jie, Y., Chen, Y., Yuan, L., Cao, J., Tan, M., Xu, W., Zheng, F., Shi, Y., and Hu, B. (2020). A tool for early prediction of severe coronavirus disease 2019 (COVID-19): A multicenter study using the risk nomogram in wuhan and guangdong, china. *Clinical Infectious Diseases*. [Online; accessed 2020-06-20].

Grassi, G., Seravalle, G., and Mancia, G. (2011). Cardiovascular consequences of poor compliance to antihypertensive therapy. *Blood Pressure*, 20(4):196–203.

Hajjar, I. and Kotchen, T. A. (2003). Trends in prevalence, awareness, treatment, and control of hypertension in the United States, 1988-2000. *JAMA: The Journal of the American Medical Association*, 290(2):199–206.

Hao, B., Sotudian, S., Wang, T., Xu, T., Hu, Y., Gaitanidis, A., Breen, K., Velmahos, G. C., and Paschalidis, I. C. (2020). Early prediction of level-of-care requirements in patients with COVID-19. *eLife*, 9:e60519.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning: data mining, inference and prediction*. Springer series in statistics Springer, Berlin.

Hastie, T., Tibshirani, R., and Tibshirani, R. J. (2017). Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso. arXiv:1707.08692 [stat].

Helleputte, T. and Dupont, P. (2009). Partially supervised feature selection with regularized linear models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 409–416, New York, NY, USA. Association for Computing Machinery.

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

Hertz, R. P., Unger, A. N., Cornell, J. A., and Saunders, E. (2005). Racial disparities in hypertension prevalence, awareness, and management. *Archives of internal medicine*, 165(18):2098–2104.

Hilton, C. B., Milinovich, A., Felix, C., Vakharia, N., Crone, T., Donovan, C., Proctor, A., and Nazha, A. (2020). Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence. *NPJ digital medicine*, 3(1):1–8.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Howard, V. J., Tanner, R. M., Anderson, A., Irvin, M. R., Calhoun, D. A., Lackland, D. T., Oparil, S., and Muntner, P. (2015). Apparent Treatment-resistant Hypertension Among Individuals with History of Stroke or Transient Ischemic Attack. *The American Journal of Medicine*, 128(7):707–714.e2.

Huang, I., Pranata, R., Lim, M. A., Oehadian, A., and Alisjahbana, B. (2020). C-reactive protein, procalcitonin, d-dimer, and ferritin in severe coronavirus disease-2019: a meta-analysis. *Therapeutic Advances in Respiratory Disease*, 14:175346662093717.

Huang, J. and Zhang, T. (2010). The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004. Publisher: Institute of Mathematical Statistics.

James, P. A., Oparil, S., Carter, B. L., Cushman, W. C., Dennison-Himmelfarb, C., Handler, J., Lackland, D. T., LeFevre, M. L., MacKenzie, T. D., Ogedegbe, O., Smith, Jr, S. C., Svetkey, L. P., Taler, S. J., Townsend, R. R., Wright, Jr, J. T., Narva, A. S., and Ortiz, E. (2014). 2014 Evidence-Based Guideline for the Management of High Blood Pressure in Adults: Report From the Panel Members Appointed to the Eighth Joint National Committee (JNC 8). *JAMA: The Journal of the American Medical Association*, 311(5):507–520.

Ji, D., Zhang, D., Xu, J., Chen, Z., Yang, T., Zhao, P., Chen, G., Cheng, G., Wang, Y., and Bi, J. (2020). Prediction for progression risk in patients with COVID-19 pneumonia: the call score. *Clinical Infectious Diseases*, 71(6):1393–1399. https://doi.org/10.1093/cid/ciaa414.

John, L. H., Kors, J. A., Reps, J. M., Ryan, P. B., and Rijnbeek, P. R. (2022). Logistic regression models for patient-level prediction based on massive observational data: Do we need all data? *International Journal of Medical Informatics*, 163:104762.

Kantroo, V., Kanwar, M. S., Goyal, P., Rosha, D., Modi, N., Bansal, A., Ansari, A. P., Wangnoo, S. K., Sobti, S., Kansal, S., Chawla, R., Jasuja, S., and Gupta, I. (2021). Mortality and clinical outcomes among patients with COVID-19 and diabetes. *Medical Sciences*, 9(4):65.

Kivimäki, M., Tabak, A. G., Batty, G. D., Ferrie, J. E., Nabi, H., Marmot, M. G., Witte, D. R., Singh-Manoux, A., and Shipley, M. J. (2010). Incremental predictive value of adding past blood pressure measurements to the Framingham hypertension risk equation: the Whitehall II Study. *Hypertension*, 55(4):1058–1062.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109.

Kramer, H., Han, C., Post, W., Goff, D., Diez-Roux, A., Cooper, R., Jinagouda, S., and Shea, S. (2004). Racial/ethnic differences in hypertension and hypertension treatment and control in the multi-ethnic study of atherosclerosis (MESA). *American journal of hypertension*, 17(10):963–970.

Krishnaswami, A., Steinman, M. A., Goyal, P., Zullo, A. R., Anderson, T. S., Birtcher, K. K., Goodlin, S. J., Maurer, M. S., Alexander, K. P., Rich, M. W., and Tjia, J. (2019). Deprescribing in Older Adults With Cardiovascular Disease. *Journal of the American College of Cardiology*, 73(20):2584–2595.

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., and Kitai, T. (2017). Artificial Intelligence in Precision Cardiovascular Medicine. *Journal of the American College of Cardiology*, 69(21):2657–2664.

Leng, B., Jin, Y., Li, G., Chen, L., and Jin, N. (2015). Socioeconomic status and hypertension: a meta-analysis. *Journal of hypertension*, 33(2):221–229.

Li, N., Zhou, H., and Tang, Q. (2017). Red Blood Cell Distribution Width: A Novel Predictive Indicator for Cardiovascular and Cerebrovascular Diseases. *Disease Markers*, 2017:7089493.

Li, Q., Zhao, L., Lee, Y.-C., and Lin, J. (2020). Contrast Pattern Mining in Paired Multivariate Time Series of a Controlled Driving Behavior Experiment. *ACM Transactions on Spatial Algorithms and Systems*, 6(4):25:1–25:28.

Li, Y., Nan, B., and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71(2):354–363.

Liang, W., Liang, H., Ou, L., Chen, B., Chen, A., Li, C., Li, Y., Guan, W., Sang, L., Lu, J., Xu, Y., Chen, G., Guo, H., Guo, J., Chen, Z., Zhao, Y., Li, S., Zhang, N., Zhong, N., He, J., and for the China Medical Treatment Expert Group for COVID-19 (2020). Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Internal Medicine*, 180(8):1081–1089.

Lipton, Z. C. (2017). The Mythos of Model Interpretability. arXiv:1606.03490 [cs, stat].

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, 49(1):15–21.

Ludwick, D. A. and Doucette, J. (2009). Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. *International Journal of Medical Informatics*, 78(1):22–31.

Mancia, G., Fagard, R., Narkiewicz, K., Redon, J., Zanchetti, A., Böhm, M., Christiaens, T., Cifkova, R., De Backer, G., Dominiczak, A., Galderisi, M., Grobbee, D. E., Jaarsma, T., Kirchhof, P., Kjeldsen, S. E., Laurent, S., Manolis, A. J., Nilsson, P. M., Ruilope, L. M., Schmieder, R. E., Sirnes, P. A., Sleight, P., Viigimaa, M., Waeber, B., and Zannad, F. (2013). 2013 ESH/ESC Guidelines for the management of arterial hypertension. *Blood Pressure*, 22(4):193–278.

Mandair, D., Tiwari, P., Simon, S., Colborn, K. L., and Rosenberg, M. A. (2020). Prediction of incident myocardial infarction using machine learning applied to harmonized electronic health record data. *BMC Medical Informatics and Decision Making*, 20(1):252.

Mazzaglia, G., Mantovani, L. G., Sturkenboom, M. C., Filippi, A., Trifirò, G., Cricelli, C., Brignoli, O., and Caputi, A. P. (2005). Patterns of persistence with antihypertensive medications in newly diagnosed hypertensive patients in Italy: a retrospective cohort study in primary care. *Journal of Hypertension*, 23(11):2093–2100.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):53–71.

Melville, S. and Byrd, J. B. (2019). Personalized Medicine and the Treatment of Hypertension. *Current Hypertension Reports*, 21(2):13.

Mensah, G. A., Mokdad, A. H., Ford, E. S., Greenlund, K. J., and Croft, J. B. (2005). State of disparities in cardiovascular health in the United States. *Circulation*, 111(10):1233–1241.

Mills, K. T., Bundy, J. D., Kelly, T. N., Reed, J. E., Kearney, P. M., Reynolds, K., Chen, J., and He, J. (2016). Global Disparities of Hypertension Prevalence and Control. *Circulation*, 134(6):441–450.

Mohajerin Esfahani, P. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166. https://doi.org/10.1007/s10107-017-1172-1.

Morenoff, J. D., House, J. S., Hansen, B. B., Williams, D. R., Kaplan, G. A., and Hunte, H. E. (2007). Understanding social disparities in hypertension prevalence, awareness, treatment, and control: the role of neighborhood context. *Social science & medicine*, 65(9):1853–1866.

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., Das, S. R., de Ferranti, S., Després, J.-P., Fullerton, H. J., Howard, V. J., Huffman, M. D., Isasi, C. R., Jiménez, M. C., Judd, S. E., Kissela, B. M., Lichtman, J. H., Lisabeth, L. D., Liu, S., Mackey, R. H., Magid, D. J., McGuire, D. K., Mohler, E. R., Moy, C. S., Muntner, P., Mussolino, M. E., Nasir, K., Neumar, R. W., Nichol, G., Palaniappan, L., Pandey, D. K., Reeves, M. J., Rodriguez, C. J., Rosamond, W., Sorlie, P. D., Stein, J., Towfighi, A., Turan, T. N., Virani, S. S., Woo, D., Yeh, R. W., and Turner, M. B. (2016). Executive Summary: Heart Disease and Stroke Statistics—2016 Update. *Circulation*, 133(4):447–454. Publisher: American Heart Association.

Mujahid, M. S., Roux, A. V. D., Morenoff, J. D., Raghunathan, T. E., Cooper, R. S., Ni, H., and Shea, S. (2008). Neighborhood characteristics and hypertension. *Epidemiology*, pages 590–598.

Mukhtar, O., Cheriyan, J., Cockcroft, J. R., Collier, D., Coulson, J. M., Dasgupta, I., Faconti, L., Glover, M., Heagerty, A. M., Khong, T. K., Lip, G. Y. H., Mander, A. P., Marchong, M. N., Martin, U., McDonnell, B. J., McEniery, C. M., Padmanabhan, S., Saxena, M., Sever, P. J., Shiel, J. I., Wych, J., Chowienczyk, P. J., and Wilkinson, I. B. (2018). A randomized controlled crossover trial evaluating differential responses to antihypertensive drugs (used as mono- or dual therapy) on the basis of ethnicity: The comparIsoN oF Optimal Hypertension RegiMens; part of the Ancestry Informative Markers in HYpertension program—AIM-HY INFORM trial. *American Heart Journal*, 204:102–108.

Muntner, P., Woodward, M., Mann, D. M., Shimbo, D., Michos, E. D., Blumenthal, R. S., Carson, A. P., Chen, H., and Arnett, D. K. (2010). Comparison of the Framingham Heart Study hypertension model with blood pressure alone in the prediction of risk of hypertension: the Multi-Ethnic Study of Atherosclerosis. *Hypertension*, 55(6):1339–1345.

Murthy, B. P. (2021). Disparities in COVID-19 vaccination coverage between urban and rural counties — United States, December 14, 2020–April 10, 2021. *MMWR. Morbidity and Mortality Weekly Report*, 70. [Online; accessed 2021-06-20].

Obozinski, G., Jacob, L., and Vert, J.-P. (2011). Group Lasso with Overlaps: the Latent Group Lasso approach. https://arxiv.org/abs/1110.0413.

Oparil, S. and Calhoun, D. A. (1998). Managing the Patient with Hard-to-Control Hypertension. *American Family Physician*, 57(5):1007.

Parikh, N. I., Pencina, M. J., Wang, T. J., Benjamin, E. J., Lanier, K. J., Levy, D., D'Agostino Sr, R. B., Kannel, W. B., and Vasan, R. S. (2008). A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study. *Annals of internal medicine*, 148(2):102–110.

Paynter, N. P., Cook, N. R., Everett, B. M., Sesso, H. D., Buring, J. E., and Ridker, P. M. (2009). Prediction of incident hypertension risk in women with currently normal blood pressure. *The American journal of medicine*, 122(5):464–471.

Pazzani, M. J., Mani, S., and Shankle, W. R. (2001). Acceptance of Rules Generated by Machine Learning among Medical Experts. *Methods of Information in Medicine*, 40(5):380–385. Publisher: Schattauer GmbH.

Peiffer-Smadja, N., Lucet, J.-C., Bendjelloul, G., Bouadma, L., Gerard, S., Choquet, C., Jacques, S., Khalil, A., Maisani, P., and Casalino, E. (2020). Challenges and issues about organizing a hospital to respond to the COVID-19 outbreak: experience from a french reference centre. *Clinical Microbiology and Infection*, 26(6):669–672. publisher: Elsevier.

Perez, B. C., Bink, M. C. A. M., Svenson, K. L., Churchill, G. A., and Calus, M. P. L. (2022). Prediction performance of linear models and gradient boosting machine on complex phenotypes in outbred mice. *G3 Genes|Genomes|Genetics*, 12(4):jkac039.

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*. Publisher: Massachusetts Medical Society.

Ranney Megan L., Griffeth Valerie, and Jha Ashish K. (2020). Critical Supply Shortages — The Need for Ventilators and Personal Protective Equipment during the COVID-19 Pandemic. *New England Journal of Medicine*, 382(18):e41.

Rao, N., Cox, C., Nowak, R., and Rogers, T. T. (2013). Sparse Overlapping Sets Lasso for Multitask Learning and its Application to fMRI Analysis. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Rodriguez, V. A., Bhave, S., Chen, R., Pang, C., Hripcsak, G., Sengupta, S., Elhadad, N., Green, R., Adelman, J., Metitiri, K. S., Elias, P., Groves, H., Mohan, S., Natarajan, K., and Perotte, A. (2021). Development and validation of prediction models for mechanical ventilation, renal replacement therapy, and readmission in COVID-19 patients. *Journal of the American Medical Informatics Association : JAMIA*. PMID: 33706377 PMCID: PMC7989331.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. arXiv:1811.10154 [cs, stat].

Sanyal, C., Turner, J. P., Martin, P., and Tannenbaum, C. (2020). Cost-Effectiveness of Pharmacist-Led Deprescribing of NSAIDs in Community-Dwelling Older Adults. *Journal of the American Geriatrics Society*, 68(5):1090–1097.

Sareli, P., Radevski, I. V., Valtchanova, Z. P., Libhaber, E., Candy, G. P., Den Hond, E., Libhaber, C., Skudicky, D., Wang, J. G., and Staessen, J. A. (2001). Efficacy of different drug classes used to initiate antihypertensive treatment in black subjects: Results of a randomized trial in johannesburg, south africa. *Archives of Internal Medicine*, 161(7):965–971.

Savoia, C., Volpe, M., Grassi, G., Borghi, C., Agabiti Rosei, E., and Touyz, R. (2017). Personalized medicine—a modern approach for the diagnosis and management of hypertension. *Clinical Science*, 131(22):2671–2685.

Savova, G. K., Danciu, I., Alamudun, F., Miller, T., Lin, C., Bitterman, D. S., Tourassi, G., and Warner, J. L. (2019). Use of Natural Language Processing to Extract Clinical Cancer Phenotypes from Electronic Medical Records. *Cancer Research*, 79(21):5463–5470.

Scott, I. A., Hilmer, S. N., Reeve, E., Potter, K., Le Couteur, D., Rigby, D., Gnjidic, D., Del Mar, C. B., Roughead, E. E., Page, A., Jansen, J., and Martin, J. H. (2015). Reducing inappropriate polypharmacy: The process of deprescribing. *JAMA Internal Medicine*, 175(5):827–834.

Seymour, C. W., Liu, V. X., Iwashyna, T. J., Brunkhorst, F. M., Rea, T. D., Scherag, A., Rubenfeld, G., Kahn, J. M., Shankar-Hari, M., and Singer, M. (2016). Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA: The Journal of the American Medical Association*, 315(8):762–774.

Sheppard, J. P., Burt, J., Lown, M., Temple, E., Lowe, R., Fraser, R., Allen, J., Ford, G. A., Heneghan, C., Hobbs, F. D. R., Jowett, S., Kodabuckus, S., Little, P., Mant, J., Mollison, J., Payne, R. A., Williams, M., Yu, L.-M., McManus, R. J., and for the OPTIMISE Investigators (2020). Effect of Antihypertensive Medication Reduction vs Usual Care on Short-term Blood Pressure Control in Patients With Hypertension Aged 80 Years and Older: The OPTIMISE Randomized Clinical Trial. *JAMA: The Journal of the American Medical Association*, 323(20):2039–2051.

Shomirov, A. and Zhang, J. (2021). An Overview of Deep Learning in MRI and CT Medical Image Processing. In *Proceedings of the 2021 3rd International Symposium on Signal Processing Systems*, SSPS '21, pages 72–78, New York, NY, USA. Association for Computing Machinery.

Sousa, C. T., Ribeiro, A., Barreto, S. M., Giatti, L., Brant, L., Lotufo, P., Chor, D., Lopes, A. A., Mengue, S. S., Baldoni, A. O., and Figueiredo, R. C. (2022). Racial differences in blood pressure control from users of antihypertensive monotherapy: Results from the elsa-brasil study. *Arquivos Brasileiros de Cardiologia*, 118:614–622.

Spitzer, R. L., Kroenke, K., Williams, J. B., Group, P. H. Q. P. C. S., and Group, P. H. Q. P. C. S. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA: The Journal of the American Medical Association*, 282(18):1737–1744.

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., and Cilar, L. (2020). Interpretability of machine learning based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5). arXiv: 2002.08596.

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., and Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1):1–10. Number: 1 Publisher: Nature Publishing Group.

Taddei, S. (2015). Combination Therapy in Hypertension: What Are the Best Options According to Clinical Pharmacology Principles and Controlled Clinical Trial Evidence? *American Journal of Cardiovascular Drugs*, 15(3):185–194.

Tanindi, A., Topal, F. E., Topal, F., and Celik, B. (2012). Red cell distribution width in patients with prehypertension and hypertension. *Blood Pressure*, 21(3):177–181.

Teixeira, P. L., Wei, W.-Q., Cronin, R. M., Mo, H., VanHouten, J. P., Carroll, R. J., LaRose, E., Bastarache, L. A., Rosenbloom, S. T., and Edwards, T. L. (2017). Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *Journal of the American Medical Informatics Association*, 24(1):162–171.

Tewari, A. and Murphy, S. A. (2017). *From Ads to Interventions: Contextual Bandits in Mobile Health*, pages 495–517. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-51394-$2_2$5.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56. Publisher: Nature Publishing Group.

Turner, J. P., Sanyal, C., Martin, P., and Tannenbaum, C. (2021). Economic Evaluation of Sedative Deprescribing in Older Adults by Community Pharmacists. *The Journals of Gerontology: Series A*, 76(6):1061–1067.

Vapnik, V. and Izmailov, R. (2015). Learning Using Privileged Information: Similarity Control and Knowledge Transfer. *Journal of Machine Learning Research*, 16(61):2023–2049.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Virdis, A., Giannarelli, C., Fritsch Neves, M., Taddei, S., and Ghiadoni, L. (2010). Cigarette smoking and hypertension. *Current pharmaceutical design*, 16(23):2518–2525.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis*, 52(12):5277–5286.

Wang, J., Oh, J., Wang, H., and Wiens, J. (2018). Learning Credible Models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 2417–2426, New York, NY, USA. Association for Computing Machinery.

Wang, K., Zuo, P., Liu, Y., Zhang, M., Zhao, X., Xie, S., Zhang, H., Chen, X., and Liu, C. (2020). Clinical and laboratory predictors of in-hospital mortality in patients with coronavirus disease-2019: A cohort study in wuhan, china. *Clinical infectious diseases*, 71(16):2079–2088.

Weiskopf, N. G., Hripcsak, G., Swaminathan, S., and Weng, C. (2013). Defining and measuring completeness of electronic health records for secondary use. *Journal of Biomedical Informatics*, 46(5):830–836.

Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, D. E., Collins, K. J., Dennison, H. C., DePalma, S. M., Gidding, S., Jamerson, K. A., Jones, D. W., MacLaughlin, E. J., Muntner, P., Ovbiagele, B., Smith, S. C., Spencer, C. C., Stafford, R. S., Taler, S. J., Thomas, R. J., Williams, K. A., Williamson, J. D., and Wright, J. T. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults. *Journal of the American College of Cardiology*, 71(19):e127–e248.

Xu, C., Lee, J., Cheng, X., and Xie, Y. (2024). Flow-based Distributionally Robust Optimization. arXiv:2310.19253 [cs, stat].

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., Huang, N., Jiao, B., Cheng, C., Zhang, Y., Luo, A., Mombaerts, L., Jin, J., Cao, Z., Li, S., Xu, H., and Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature Machine Intelligence*, 2:283–288. https://doi.org/10.1038/s42256-020-0180-7.

Yang, S., Chen, S., Wang, P., Chen, A., and Tian, T. (2024). TSPLASSO: A Two-Stage Prior LASSO Algorithm for Gene Selection Using Omics Data. *IEEE Journal of Biomedical and Health Informatics*, 28(1):526–537.

Ye, X., Zeng, Q. T., Facelli, J. C., Brixner, D. I., Conway, M., and Bray, B. E. (2020). Predicting Optimal Hypertension Treatment Pathways Using Recurrent Neural Networks. *International Journal of Medical Informatics*, 139:104122.

Yoon, C. H., Torrance, R., and Scheinerman, N. (2022). Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics*, 48(9):581–585. Publisher: Institute of Medical Ethics Section: Clinical ethics.

Yuan, L., Liu, J., and Ye, J. (2011). Efficient Methods for Overlapping Group Lasso. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

Zhang, L., Zhao, L., Zhang, X., Kong, W., Sheng, Z., and Lu, C.-T. (2018). Situation-Based Interpretable Learning for Personality Prediction in Social Media. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1554–1562.

Zhang, Q., Bai, C., Chen, Z., Li, P., Wang, S., and Gao, H. (2019). Smart Chinese medicine for hypertension treatment with a deep learning model. *Journal of Network and Computer Applications*, 129:1–8.

Zhang, S., Wang, J., Pei, L., Liu, K., Gao, Y., Fang, H., Zhang, R., Zhao, L., Sun, S., Wu, J., Song, B., Dai, H., Li, R., and Xu, Y. (2022). Interpretability Analysis of One-Year Mortality Prediction for Stroke Patients Based on Deep Neural Network. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1903–1910.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7(90):2541–2563.

Zhou, L. and Hripcsak, G. (2007). Temporal reasoning with medical data—A review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, 40(2):183–202.

Zhu, F., Guo, J., Li, R., and Huang, J. (2018). Robust Actor-Critic Contextual Bandit for Mobile Health (mHealth) Interventions. In BCB '18: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics August 2018, pp. 492 - 501. https://doi.org/10.1145/3233547.323355.

CURRICULUM VITAE