DEEP SEMI-SUPERVISED AND MULTI-STAGE LEARNING FOR MEDICAL APPLICATIONS

A Dissertation

by

NATHAN CLINTON HURLEY

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Bobak J. Mortazavi
Committee Members,	Carolyn L. Cannon
	Theodora Chaspari
	Zhangyang Wang
Head of Department,	Scott Schaefer

May 2022

Major Subject: Computer Science

ABSTRACT

Machine learning techniques are widely used to build models for applications in healthcare. These models typically predict likelihood of a particular patient outcome in a given setting. For clinical utility, these models are often used to derive parsimonious models that predict outcome risks of certain populations. Training these models on a specific patient population, their demonstrated utility is confined to patients with characteristics similar to the original derivation cohort. However, these simpler machine learning techniques may lack the discriminatory power to recognize subpopulations within a population that behave or respond differently to identical interventions. Conversely, while more complex machine learning techniques and complex data streams may possess the sophistication necessary to recognize and appropriately predict outcomes of these subpopulations, the training sizes necessary to achieve good results are prohibitively large. Correctly understanding and identifying the differences and similarities that separate and unify various subpopulations is key to building a model that is sufficiently extensible to explain population variance while minimizing unnecessary complexity.

This dissertation applies and advances machine learning for healthcare through three approaches. First, it utilizes advanced machine learning techniques for clinical modeling. This is done while predicting harmful outcomes such as mortality in vulnerable patient populations. Second, it describes advanced machine learning techniques to handle heterogeneity in retrospective analyses. It develops a novel application of a deep mixture of experts to describe this heterogeneity, learning phenotypes in a risk-driven method. Finally, it describes needs and opportunities in harnessing remote sensors for health monitoring and details two specific approaches to extracting useful health data from longitudinal sensors.

DEDICATION

To Mom and Dad, for always helping me to enjoy learning.

To Will, Ben, and Katie, for always being ready to share a laugh.

And to Ayrea, for always keeping me going and for being there for me.

ACKNOWLEDGMENTS

First, I would like to thank my mentor, Dr. Bobak Mortazavi. He has provided a wonderful lab in which to grow and learn over the past four years. I am very grateful that I found your lab when I did, and that you were willing to take me in. I look forward to many years of future collaborations and continued friendship.

I would also like to thank my wonderful labmates, both present and past. You give a great deal of depth to collaborations, and I have enjoyed working with you all. I have learned a lot from all of you, and I hope that I have returned the favor. I wish you all the best.

Thank you to the many collaborators who have been willing and able to help me learn. Thanks to Sanket and Nihar for your frequent meetings, advice, and direction. Thanks to Adrian for providing mentorship as I started the PhD.

Thank you to my friends and colleagues in the MD/PhD program. We've been through a lot together. I know that some of you will be among my best friends for years to come. You've let me stay with you when I needed places to stay, you've answered phone calls about trigonometry at 2 a.m., and I know I can always count on you. You're great people, and I look forward to seeing where we all end up in life.

Thank you to my family. You've been a huge support, and I definitely wouldn't be where I am without that continuing support and encouragement. Thank you.

Finally, my biggest thanks to Ayrea. You've kept me sane, you've taken care of me, and I can't ever thank you enough for that. I'm excited to see where we go next, and looking forward to whatever the future brings.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Bobak Mortazavi [advisor], Professor Zhangyang "Atlas" Wang, and Professor Theodora Chaspari of the Department of Computer Science and Engineering and Professor Carolyn Cannon of the Department of Microbial Pathogenesis and Immunology.

The data analyzed for Chapter 2 was provided by Professor Wade Schulz. The narrative in that chapter was overseen and drafted by Jacob McPadden. Analyses and narrative in Chapter 4 were conducted in part by Justin Lovelace. The data used in Chapters 5, 6, 7, 8, 9, and 10 was provided by the National Cardiovascular Data Registry (NCDR). The manuscript for Chapter 5 was drafted primarily by Dr. Rohan Khera. Chapters 6 and 7 were drafted primarily by Dr. Sanket Dhruva and Dr. Nihar Desai. The model design and implementation for Chapter 12 was by Lida Zhang.

Funding Sources

This project is in part supported by the Defense Advanced Research Projects Agency under grant FA8750-18-2-0027 and National Institutes of Health under grant 1R21EB028486-01.

This work was also made possible in part by the Texas A&M University M.D./Ph.D. Program.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Data regarding healthcare is rich and complex. Between electronic health record (EHR) data and sensor streams from wearable smart devices, an incredible amount of data is generated relating to health. This data takes up many different forms. MIMIC-III, the most widely used freely available EHR dataset incorporates many different types of data including demographics, microbiological results, diagnoses, clinical events, procedures, and free text narrative. [2]. Supplements to MIMIC-III include continuous waveform data [3] and imaging data [4]. Building machine learning models from EHR data requires a great deal of data preparation and standardization. This is complicated by the variety of possible data types and the high levels of sparsity in EHR datasets [5, 6]. Various machine learning techniques including both deep and classical techniques have been applied to EHR datasets [7, 8, 6].

Health data can also extend to data generated outside of the hospital. Sensing technologies have rapidly advanced in the past decades. With the advent of technologies such as smartphones and smartwatches, many people carry devices which provide rich streams of noisy data. In the hospital, monitoring patients is part of routine clinical practice. Providers are able to monitor cardiac status and basic vitals from anywhere in the hospital at any time. Slight deterioration in health can be observed and interventions put into place before patients suffer worsening harm. However, length of stay in these acute care settings is often quite short [9, 10], representing only a small portion of a patient's life despite the prolonged impact that the decision making in these settings have. Given the rise of sensors commonly included in consumer electronics, there exist many opportunities to expand monitoring in the outpatient setting.

Longitudinal monitoring of physiologic parameters and symptoms outside of the hospital can enable better detection and response systems before a person becomes acutely ill and requires hospitalization. After hospitalization, monitoring these signals could help to prevent early readmission to the hospital. However, many commercial devices today are targeted

to healthy people. With the prevalence and ubiquitous nature of remote and wearable sensors, opportunities exist to broaden the applications of sensing and for adapting analytic techniques to enhance diagnosis, monitoring, and treatment of risk factors for primary and secondary prevention of cardiovascular disease. In particular, the ability to capture these measurements is only the first step. Indeed, end-to-end smart health systems are needed that couple the hardware development with advanced analytic techniques to provide both patient and clinical provider necessary confidence in data and risk prediction based upon the measured risk factors.

Integrating EHR data and remote monitoring data through smart devices is a largely underutilized paradigm. This integration has largely been explored in the context of specific disease studies. For instance, [11] utilized smartphones connected to an EHR system to provide pulmonologists with remote measures of inhaler utilization. However, there is a great opportunity for further integration of remote data, particularly in conjunction with machine learning techniques [8].

The many disparate forms of data pose a daunting task for machine learning: how can algorithms and models be built that best take advantage of that data to improve well-being? This dissertation seeks to transform data into more useful forms, find similarities between patients in heterogeneous populations, and make estimates of treatment effects of various interventions, using top level medical knowledge for causal inference. Through all of this, the goal is to better equip physicians with models explaining patient health and providing support for clinical decision making.

This work describes three main goals in improving data utilization. First, it describes using advanced clinical modeling for machine learning predictions. This is shown through characterizing outcomes among patients infected with SARS-CoV-2 (Ch. 2); discovering, characterizing, and visualizing phenotypes among patients presenting to an emergency department (Ch. 3), using natural language processing to extract high level concepts describing patient health (Ch. 4), and finally by characterizing advanced machine learning models to

predict mortality following acute myocardial infarction (Ch. 5).

Next, this work allows for the advancement of clinical decision making through application and advancement of machine learning techniques. This is shown through a series of chapters utilizing cardiac registry data. Utilization changes in mechanical circulatory support devices over time are described (Ch. 6), and then a propensity matching analysis is performed to find the association of two mechanical circulatory support devices with major bleeding and mortality (Ch. 7). A dynamic model to estimate changing risk of major bleeding is described (Ch. 8) and a novel deep mixture of experts for outcome-driven phenotyping characterizes patients undergoing acute myocardial infarction complicated by cardiogenic shock (Ch. 9). This mixture of experts approach is then developed further by inspecting the latent space to describe phenotypes and to allow for soft assignments between them (Ch. 10).

Finally, this work describes the expansion of these techniques into free living environments. Ways in which cardiac sensors are and can be applied outside of clinical settings are detailed (Ch 11). Neural architecture search is applied for implementing a multitask model to derive blood pressure from a wrist-worn bioimpedance device (Ch. 12). Finally, the deep mixture of experts model is applied to remote sensors to aid in human activity recognition (Ch. 13).

Through these three goals, this dissertation advances machine learning for medical applications, utilizing these techniques to aid in clinical decision making, and finally describes ways in which these techniques can be applied outside of clinical settings. There is a technical gap in bridging rich data sources to useful clinical information. While collecting data is relatively cheap, using it intelligently is difficult. This work allows for improved refinement of that data to useful clinical metrics.

2. CHARACTERISTICS AND OUTCOMES AFTER SARS-COV-2 INFECTION*

Machine learning techniques offer a range of tools for discovering and analyzing patterns in data. This chapter was written in the early months of the pandemic, and the numbers below reflect that. The goal of this work was to provide an understanding of the course of infection, and to determine pertinent factors relating to admission and to mortality. The techniques used here are able to show that age and sex are factors with strong statistical significance for predicting both admission and mortality, while other comorbidities had a weaker association. Over a year into the pandemic, these findings are less urgent than they were at its start. However, the work here shows that advanced machine learning is a valuable tool in providing useful and interpretable clinical understanding.

2.1 Introduction

Severe acute respiratory syndrome virus (SARS-CoV-2) has infected over 110 million people with nearly 2.5 million deaths worldwide [12]. Despite the global impact, key gaps in knowledge persist. A comprehensive assessment of patients evaluated for SARS-CoV-2, from testing to outcome, is needed to guide public health recommendations and scientific investigations into the mechanisms of disease pathogenesis.

Prior studies have identified many risk factors for SARS-CoV-2 infections and complications [13, 14, 15, 16]. Older age and male sex have been consistently associated with worse outcomes, as have many chronic cardiovascular and respiratory diseases [14, 15, 16, 17]. Despite some consistent themes, reports from different geographic locations have reported variation in both risks and mortality rates [18, 19, 20, 21, 22]. No study yet exists that describes the characteristics and outcomes of a single cohort from testing to outcome and

*This chapter is reprinted with permission from "Clinical characteristics and outcomes for 7,995 patients with SARS-CoV-2 infection" by McPadden, J., Warner, F., Young, H.P., Hurley, N.C., Pulk, R.A., Singh, A., Durant, T.J., Gong, G., Desai, N., Haimovich, A., Taylor, R.A., Gunel, M., Cruz, C.S.D., Farhadian, S.F., Siner, J., Villanueva, M., Churchwell, K., Hsiao, Al, Torre, C.J., Velazquez, E.J., Herbst, R.S., Iwasaki, A., Ko., A.I., Mortazavi, B.J., Krumholz, H.M., and Schulz, W.L., 2021. PLOS ONE. Copyright 2021 by McPadden, J. et al., CC BY 4.0.

with detailed information on treatments in a racially and ethnically diverse population.

Drawing from a highly curated real-world data set, we describe a diverse cohort from a catchment area that represents the diversity of the nation located in an early epicenter of the US outbreak. We extend the current literature with a detailed assessment of the characteristics of patients tested, and the clinical courses and outcomes of those testing positive, and among those admitted with SARS-CoV-2. We sought to identify risk factors for admission among those with SARS-CoV-2 and in-hospital mortality among discharged patients. We also characterize the patterns of treatment to provide the context to guide interpretation of these results.

2.2 Methods

2.2.1 Study setting and data collection

This was an observational, retrospective study of patients who were tested for SARS-CoV-2 within the Yale New Haven Health (YNHH) system, located within one of the US epicenters of Covid-19. The healthcare system is comprised of a mix of pediatric, suburban community, urban community, and urban academic inpatient facilities at five sites with a total of 2,681 licensed beds and 124,668 inpatient discharges in 2018 [23]. The system also includes associated outpatient facilities that had 2.4 million outpatient encounters in 2018. YNHH uses a single electronic health record (EHR) across the health system. Patient demographics, past medical histories, medications, and clinical outcomes were extracted from our local Observational Medical Outcomes Partnership (OMOP) [24] data repository and analyzed within our computational health platform [25, 26]. Data were extracted with custom PySpark (version 2.4.5) scripts that were reviewed by an independent analyst. The study was approved by the Yale University Institutional Review Board (protocol #2000027747).

2.2.2 Study cohort

The study cohort consists of all adult patients ($\geq$18 years old) at YNHH who had an order for SARS-CoV-2 RT-PCR testing and a test result documented within the medical

Table 2.1: Race and ethnicity as noted in the EHR and mapped to the OMOP CDM.

EHR-Recorded Race or Ethnicity	**OMOP Mapping**	**Abbreviation**
American Indian or Alaska Native	American Indian or Alaska Native	American Indian or Alaska Native
Asian	Asian	Asian
Black or African American	Black or African American	Black
Native Hawaiian	Native Hawaiian or Other Pacific Islander	Native Hawaiian or Other Pacific Islander
Other Pacific Islander	Native Hawaiian or Other Pacific Islander	Native Hawaiian or Other Pacific Islander
Other/Not Listed	Other	Other
Patient Refused	Unknown/Not Stated	Unknown/Not Stated
Unknown	Unknown/Not Stated	Unknown/Not Stated
White or Caucasian	White	White
Hispanic or Latino	Hispanic or Latino	Hispanic
Not Hispanic or Latino	Not Hispanic or Latino	Not Hispanic

record between March 1, 2020 and April 30, 2020 (Figure 2.1). SARS-CoV-2 testing in our health system was limited to symptomatic patients for whom the provider had a concern for respiratory tract infection in the month of March. Testing increased to include a wider breadth of symptoms deemed clinically concerning during the month of April. By the end of April, all patients admitted to the health system were tested for Covid-19. Outpatient testing required a physician order and was primarily sent to external reference laboratories. The decision to test was ultimately left to the ordering provider. Testing was first made available to order within the health system on March 13th, 2020.

Patients admitted more than 24 hours prior to testing were excluded from the admissions group to reduce the likelihood of including hospital acquired infections. Data and outcomes were limited to those collected between March 1, 2020 and April 30, 2020. An extract of our local OMOP data repository from September 13, 2020 was used to allow for final discharge disposition and vendor-provided transformations of the clinical data warehouse to complete. For patients with multiple admissions in the study period, only data from the first admission was used. Race and ethnicity were extracted from the demographics section of the EHR and mapped to the OMOP common data model (Table 2.1). For demographic fields that had selected values or responses, individual counts were further anonymized to remove any counts $\leq$3.

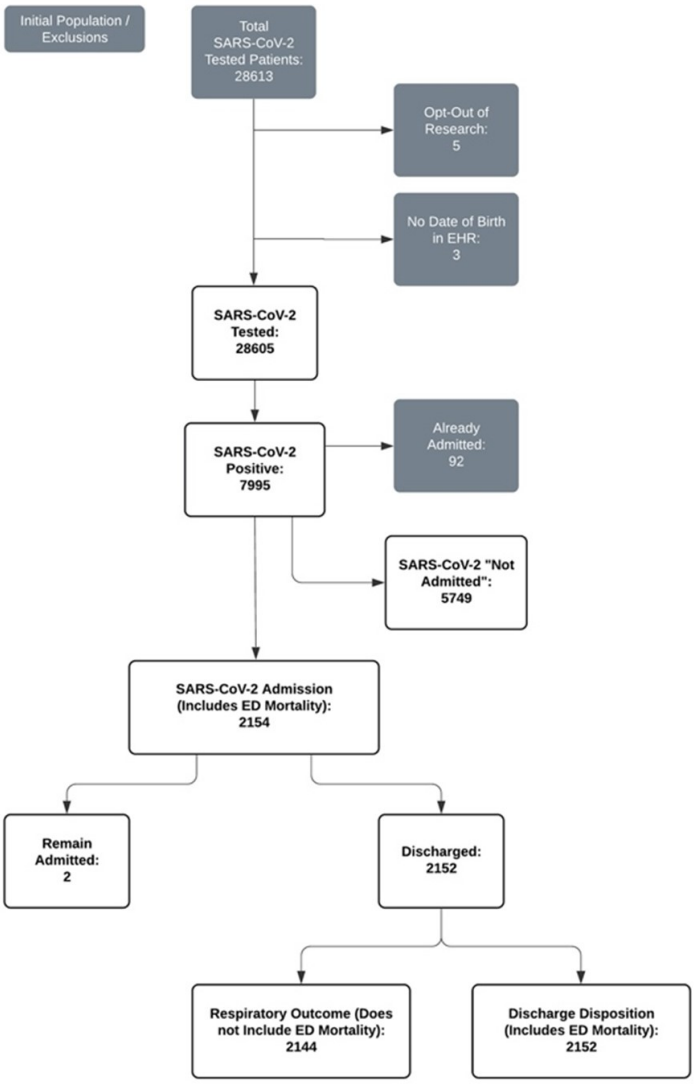


Figure 2.1: Patient counts and exclusions based on computed phenotyping criteria.

2.2.3 Outcome ascertainment

We extracted primary outcomes of admission and discharge disposition along with secondary outcomes of supplemental oxygen use and mechanical respiratory support. The maximum respiratory requirement during admission was used. Covid-19 related admissions were identified by extracting data from each patient's first inpatient admission that had a visit start time within a window 14 days following or 24 hours before a positive SARS-CoV-2

test was ordered for a patient. For patients with a transfer to another facility (n = 44), the outcome from the first visit was used. Visit-related data and in-hospital mortality were directly extracted from our OMOP data repository. Supplemental oxygen requirements were computed based on presence of clinical documentation in flowsheets or vitals measurements and were mapped to one of four categorical variables: low-flow oxygen, high-flow oxygen, noninvasive mechanical ventilation, and/or invasive mechanical ventilation. Outcomes were limited to patients who were discharged and were therefore not extracted for patients who were still admitted at the end of the study period. Digitally extracted outcomes were validated for 30 patients via medical record review by a clinician. All ages were calculated relative to the time of SARS-CoV-2 test order.

2.2.4 Treatment pathways

To document clinical treatment pathways, we extracted medication administration records of all admitted patients for their initial visit. Medications related to Covid-19 treatment based on institutional guidelines were grouped by calendar day of first administration. All forms of corticosteroids were mapped to a single drug class rather than their individual active ingredients. The order of medication initiation defined the separate treatment regimens and final treatment pathway. Treatment pathway visualizations were created with the JavaScript library Data Driven Documents (D3, version 4) [27].

2.2.5 Statistical analyses

The tables of demographic data and outcome data were built using the R (version 3.5.1) package tableone. Logistic regressions were performed using the core R function glm. Model 1 was among those testing positive to identify risk factors associated with admission. Candidate variables included the features described in Figure 2.2. Before computing the final model, the variables for "Other" race and ethnicity of "Not Hispanic" were removed in order to ensure that all variables in the model had variance inflation factor less than 3. Model 2 was among those with a final discharge disposition at the end of the study period

(right-censored for patients who were still admitted) to identify risk factors associated with in-hospital mortality. We began with the variables used in the admission model and removed the race variables for "American Indian or Alaska Native" and "Native Hawaiian or Other Pacific Islander" and the age variable "Age 35–44". This was done to ensure the variation inflation factors would all be less than 10. A value of $p<0.05$ was used as the threshold for significance without adjustment for multiple comparisons.

Elixhauser comorbidity [28] analysis was performed using the R comorbidity package (version 0.5.3) [29]. Briefly, ICD-10 codes from each patient's medical history taken from the OMOP database were used to generate presence or absence of the 31 Elixhauser comorbidity categories, as well as weighted scores using the AHRQ and van Walraven algorithms [30, 31].

Age-adjusted in-hospital mortality was calculated with direct standardization [32] based on the discharge population. In this method, age-specific rates are weighted according to the prevalence of age groups within an a priori standard population. This converts the observed age-specific rates of some process into a rate which would be observed had that same process acted upon the standard population. The 2000 US population was used as the standard population for age adjustment [32]. We used weights for five-year age groupings from ages 15 to 84 and a final group of 85 and over.

2.3 Results

The number of patients positive for SARS-CoV-2 increased rapidly beginning in March 2020 (Figure 2.3). A total of 28605 patients were tested for SARS-CoV-2 with 7995 patients (27.9%) who had at least one positive result during the observation period. Of those with positive tests, 2154 (26.9%) had an associated hospital admission. Of admitted patients, 2152 (99.9%) had a final discharge disposition and 2 (0.1%) remained hospitalized at the time of data extraction. For SARS-CoV-2 infected patients who were not admitted, the median number of days elapsed between testing and the study end date was 23.4 days (IQR 14.6–30.6).

	Tested	Positive	Admitted
n	28605	7995	2154
Sex (%)			
Female	17191 (60.1)	4435 (55.5)	1031 (47.9)
Male	11404 (39.9)	3558 (44.5)	1123 (52.1)
Unknown	10 (0.0)	2 (0.0)	0 (0.0)
Race (%)			
American Indian	60 (0.2)	12 (0.2)	<10 (<0.5)
Asian	721 (2.5)	175 (2.2)	44 (2.0)
Black	5093 (17.8)	1856 (23.2)	546 (25.3)
Hawaiian/Pacific Islander	79 (0.3)	30 (0.4)	<10 (<0.5)
Other	4464 (15.6)	1874 (23.4)	497 (23.1)
Unknown/Not Stated	1363 (4.8)	432 (5.4)	37 (1.7)
White	16825 (58.8)	3616 (45.2)	1021 (47.4)
Ethnicity (%)			
Hispanic or Latino	5468 (19.1)	2245 (28.1)	560 (26.0)
Not Hispanic or Latino	21540 (75.3)	5265 (65.9)	1559 (72.4)
Unknown/Not Stated	1597 (5.6)	485 (6.1)	35 (1.6)
Age (%)			
18–34	6581 (23.0)	1579 (19.7)	146 (6.8)
35–44	4880 (17.1)	1321 (16.5)	181 (8.4)
45–54	5207 (18.2)	1546 (19.3)	258 (12.0)
55–64	5480 (19.2)	1583 (19.8)	435 (20.2)
65–74	3194 (11.2)	885 (11.1)	402 (18.7)
75–84	1928 (6.7)	565 (7.1)	374 (17.4)
85+	1335 (4.7)	516 (6.5)	358 (16.6)
Elixhauser Comorbidities (%)			
AIDS/HIV	248 (0.9)	75 (0.9)	32 (1.5)
Alcohol abuse	2141 (7.5)	398 (5.0)	175 (8.1)
Blood loss anemia	1228 (4.3)	305 (3.8)	144 (6.7)
Cardiac arrhythmias	7242 (25.3)	1691 (21.2)	803 (37.3)
Chronic pulmonary disease	9043 (31.6)	2005 (25.1)	717 (33.3)
Coagulopathy	2450 (8.6)	511 (6.4)	265 (12.3)
Congestive heart failure	3155 (11.0)	805 (10.1)	494 (22.9)
Deficiency anemia	3787 (13.2)	978 (12.2)	409 (19.0)
Depression	7697 (26.9)	1664 (20.8)	637 (29.6)
Diabetes, complicated	3817 (13.3)	1190 (14.9)	624 (29.0)
Diabetes, uncomplicated	5502 (19.2)	1738 (21.7)	810 (37.6)
Drug abuse	2548 (8.9)	401 (5.0)	173 (8.0)
Fluid and electrolyte disorders	6540 (22.9)	1623 (20.3)	905 (42.0)
Hypertension, complicated	3666 (12.8)	988 (12.4)	616 (28.6)
Hypertension, uncomplicated	11950 (41.8)	3334 (41.7)	1387 (64.4)
Hypothyroidism	4467 (15.6)	1154 (14.4)	448 (20.8)
Liver disease	3417 (11.9)	755 (9.4)	283 (13.1)
Lymphoma	421 (1.5)	71 (0.9)	29 (1.3)
Metastatic cancer	1557 (5.4)	271 (3.4)	140 (6.5)
Obesity	7654 (26.8)	2195 (27.5)	684 (31.8)
Other neurological disorders	3481 (12.2)	939 (11.7)	542 (25.2)
Paralysis	672 (2.3)	203 (2.5)	118 (5.5)
Peptic ulcer disease, excluding bleeding	989 (3.5)	217 (2.7)	113 (5.2)
Peripheral vascular disorders	3416 (11.9)	875 (10.9)	523 (24.3)
Psychoses	1259 (4.4)	330 (4.1)	206 (9.6)
Pulmonary circulation disorders	1568 (5.5)	358 (4.5)	226 (10.5)
Renal failure	2888 (10.1)	809 (10.1)	515 (23.9)
Rheumatoid arthritis/collagen vascular diseases	2145 (7.5)	432 (5.4)	174 (8.1)
Solid tumor without metastasis	3023 (10.6)	658 (8.2)	313 (14.5)
Valvular disease	4188 (14.6)	1011 (12.6)	528 (24.5)
Weight loss	2864 (10.0)	664 (8.3)	357 (16.6)

https://doi.org/10.1371/journal.pone.0243291.t001

Figure 2.2: Demographics and Elixhauser comorbidities of all patients tested, tested positive, and admitted for SARS-CoV-2.

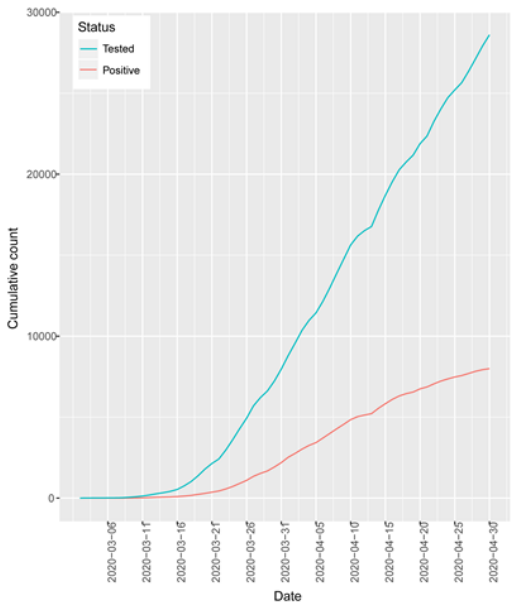


Figure 2.3: Cumulative patients tested (blue) and positive (red) for SARS-CoV-2.

2.3.1 Characteristics of individuals tested for SARS-CoV-2

Of the patients tested for SARS-CoV-2, a majority (n = 17191; 60.1%) were female (Figure 2.2). The most common comorbidities were uncomplicated hypertension (n = 11950; 41.8%), chronic pulmonary disease (n = 9043; 31.6%), and depression (n = 7697; 26.9%). The median age of tested adults was 50.8 years (IQR 36.1–63.5). In those tested for SARS-CoV-2, 4.8% did not have a reported race within the demographics section of the EHR. The majority of tested patients were reported as White (n = 16825; 58.8%), followed by Black (n = 5093; 17.8%) and Other race (n = 4464; 15.6%). Those who self-identified as Hispanic ethnicity represented 19.1% (n = 5468) of the tested population. Testing frequency by race and ethnicity showed slight overrepresentation of minority groups based on the census numbers for Connecticut, which has a demographic breakdown of 66.9% White, 12.2% Black, 5.0% Asian, 0.6% American Indian or Alaskan Native, 0.1% Native Hawaiian or Pacific Islander, and 16.9% Hispanic [33].

Age was similarly distributed between the SARS-CoV-2 tested and positive populations. Of those who tested positive, the median age was 52.3 years (IQR 38.3–64.8). Patients

with a positive test were more frequently female (n = 4435, 55.5%) with uncomplicated hypertension (n = 3334, 41.7%), obesity (n = 2195, 27.5%), and chronic pulmonary disease (n = 2005, 25.1%) as the most common comorbidities. Patients with a positive test were most frequently reported as White (n = 3616, 45.2%), followed by Other race (n = 1874, 23.4%) and Black (n = 1856, 23.2%). Those who were reported as Hispanic ethnicity accounted for 28.1% (n = 2245) of SARS-CoV-2 positive patients.

2.3.2 Features associated with admission in patients with Covid-19

The median age of SARS-CoV-2 positive patients admitted to the hospital was 66.2 years (IQR 53.7–79.9) and a majority were male (n = 1123, 52.1%) as shown in Figure 2.2. The most common Elixhauser comorbidities for admitted patients included uncomplicated hypertension (n = 1387, 64.4%), fluid & electrolyte disorders (n = 905, 42.0%), and diabetes without complications (n = 810, 37.6%). Minority groups were overrepresented in the admitted population compared to census numbers, particularly for those with a recorded race of Black (n = 546, 25.3%) or Other race (n = 497, 23.1%). Those recorded as Hispanic ethnicity accounted for 26.0% (n = 560) of admitted patients.

In multivariable analyses, older age was significantly associated with risk of admission (Figure 2.4, Table 2.2). Age ≥85 years had the highest risk of admission (OR 22.03, 95%CI = 16.10–30.30). Male sex was also associated with increased risk of admission (OR 1.68, 95%CI = 1.48–1.90). The comorbidities associated with increased risk of admission included fluid & electrolyte disorders (OR 1.99, 95%CI = 1.67–2.37), psychoses (OR 1.98, 95%CI = 1.47–2.69), metastatic cancer (OR 1.55, 95%CI = 1.11–2.15), pulmonary circulation disorders (OR 1.53, 95%CI = 1.14–2.06), peptic ulcer disease (OR 1.47, 95%CI = 1.04–2.07), drug abuse (OR 1.46, 95%CI = 1.11–1.92), renal failure (OR 1.38, 95%CI = 1.08–1.75), other neurological disorders (OR 1.31, 95%CI = 1.07–1.61), and obesity (OR 1.18, 95%CI = 1.02–1.37). Of note, complicated hypertension (OR 1.14, 95%CI = 0.88–1.48), uncomplicated hypertension (OR 0.97, 95%CI = 0.83–1.13), and chronic pulmonary disease (OR 0.94, 95%CI = 0.81–1.09) were not found to significantly increase the odds of admission. Recorded

races with increased odds of admission included Asian (OR 1.58, 95%CI = 1.02–2.41) and Black (OR 1.43, 95%CI = 1.14–1.78). Hispanic ethnicity was also associated with increased risk of admission (OR 1.81, 95%CI = 1.50–2.18).

2.3.3 Outcomes in discharged patients with Covid-19

Of the patients admitted for COVID-19 the majority (n = 2152, 99.9%) had a known disposition and therefore had complete outcomes available at the time of data extraction (Figure 2.5). The median length of stay for discharged patients was 8.1 days (IQR 4.3–14.8). Mortality occurred in the emergency department for 8 of these patients who were excluded from respiratory analysis as they did not have complete respiratory outcomes reported.

The majority of patients with respiratory outcomes did not require invasive ventilation (n = 1823, 84.7%). For these patients, male (49.8%) and female (50.2%) sex were similar in frequency with most frequently self-reported races of White (n = 908, 49.8%), Black (n = 444, 24.4%), and Other race (n = 398, 21.8%). The most prevalent comorbidities included uncomplicated hypertension (n = 1182, 64.8%), fluid & electrolyte disorders (n = 770, 42.2%), and cardiac arrhythmia (n = 694, 38.1%). Invasive ventilatory support was required for 15.3% (n = 329) of patients with respiratory outcomes. The majority of those who required invasive ventilatory support were male (n = 215, 65.3%) with self-reported race of White (n = 113, 34.3%), Black (n = 100, 30.4%), and Other race (n = 99, 30.1%). The most prevalent comorbidities included uncomplicated hypertension (n = 204, 62.0%), diabetes without complication (n = 145, 44.1%), and fluid & electrolyte disorders (n = 134, 40.7%).

In-hospital mortality was 14.2% (n = 305) of patients with a discharge disposition and these patients had a median length of stay of 7.9 days (IQR 3.5–15.1). The majority of patients who experienced in-hospital mortality were male (n = 175, 57.4%) and mortality increased with age (Figure 2.6, panel A); the median age of those who experienced in-hospital mortality was 80.7 (IQR 70.5–88.6) years. Those with older age, particularly those $\geq$85 years old, predominantly self-reported a race of White (Figure 2.6, panel B). The comorbidities

Table 2.2: Multivariable analysis with odds ratios for admission in patients with a positive SARS-CoV-2 test compared to patients who were not admitted.

	Odds Ratio	CI 2.50%	CI 97.50%	p
(Intercept)	0.052	0.040	0.067	<0.01
Sex				
Male	1.68	1.48	1.90	<0.01
Race or Ethnicity				
American Indian	0.44	0.02	2.68	0.47
Asian	1.58	1.02	2.41	0.04
Black	1.43	1.14	1.78	<0.01
Hawaiian/Pacific Islander	1.45	0.53	3.52	0.44
Hispanic	1.81	1.50	2.18	<0.01
White	0.85	0.70	1.03	0.09
Age				
35-44	1.43	1.13	1.81	<0.01
45-54	1.76	1.41	2.21	<0.01
55-64	3.24	2.60	4.04	<0.01
65-74	6.95	5.45	8.91	<0.01
75-84	15.91	11.92	21.33	<0.01
>85	22.03	16.10	30.30	<0.01
Elixhauser Comorbidities				
AIDS/HIV	1.26	0.71	2.20	0.43
Alcohol abuse	1.24	0.95	1.62	0.12
Blood loss anemia	0.98	0.70	1.35	0.88
Cardiac arrhythmias	1.13	0.96	1.33	0.13
Chronic pulmonary disease	0.94	0.81	1.09	0.43
Coagulopathy	1.11	0.87	1.43	0.39
Congestive heart failure	1.06	0.83	1.36	0.63
Deficiency anemia	0.82	0.67	1.00	0.05
Depression	0.85	0.72	1.01	0.06
Diabetes, complicated	1.18	0.94	1.47	0.15
Diabetes, uncomplicated	1.16	0.95	1.40	0.15
Drug abuse	1.46	1.11	1.92	0.01
Fluid and electrolyte disorders	1.99	1.67	2.37	<0.01
Hypertension, complicated	1.14	0.88	1.48	0.31
Hypertension, uncomplicated	0.97	0.83	1.13	0.68
Hypothyroidism	0.89	0.75	1.06	0.21
Liver disease	1.01	0.83	1.23	0.91
Lymphoma	1.07	0.58	1.92	0.83
Metastatic cancer	1.55	1.11	2.15	0.01
Obesity	1.18	1.02	1.37	0.02
Other neurological disorders	1.31	1.07	1.61	0.01
Paralysis	1.00	0.70	1.44	1.00
Peptic ulcer disease, excluding bleeding	1.47	1.04	2.07	0.03
Peripheral vascular disorders	0.84	0.68	1.04	0.12
Psychoses	1.98	1.47	2.69	<0.01
Pulmonary circulation disorders	1.53	1.14	2.06	0.01
Renal failure	1.38	1.08	1.75	0.01
Rheumatoid arthritis/collagen vascular diseases	0.89	0.69	1.15	0.38
Solid tumor without metastasis	0.87	0.69	1.09	0.22
Valvular disease	1.01	0.83	1.23	0.93
Weight loss	0.98	0.78	1.23	0.88

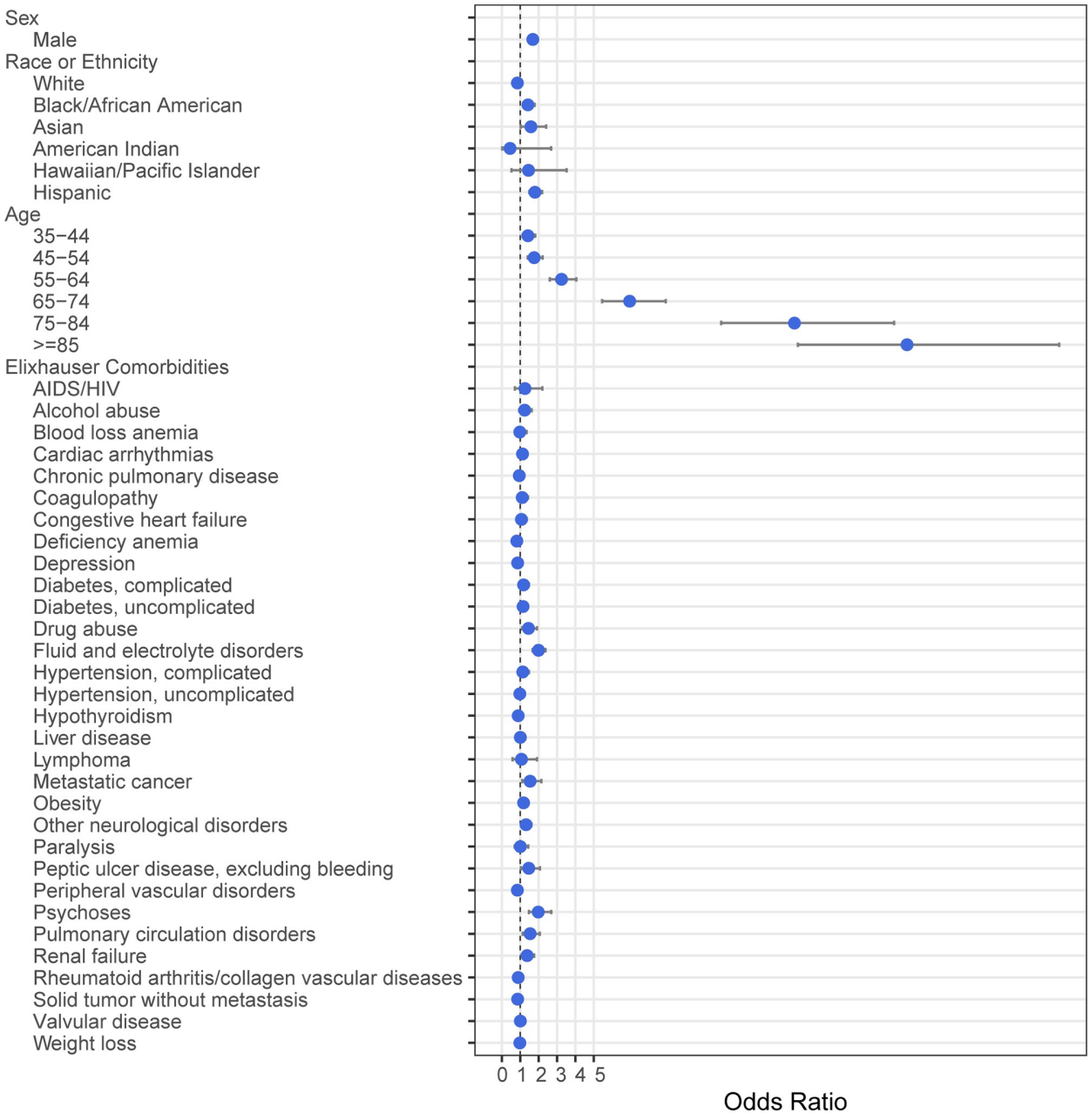


Figure 2.4: Multivariable analysis with odds ratios for admission in patients with a positive SARS-CoV-2 test.

	Discharged	No Oxygen	Low Flow	High Flow	Noninvasive	Invasive	Discharged Alive	Expired
n	2152	430	809	511	73	329	1847	305
Sex								
Male (%)	1122 (52.1)	196 (45.6)	379 (46.8)	296 (57.9)	36 (49.3)	215 (65.3)	947 (51.3)	175 (57.4)
Race (%)								
American Indian or Alaska Native	<10 (<0.5)	0 (0.0)	<5 (<0.6)	0 (0.0)	0 (0.0)	0 (0.0)	<10 (<0.5)	0 (0.0)
Asian	44 (2.0)	8 (1.9)	13 (1.6)	13 (2.5)	<5 (<6.8)	<10 (<3.0)	38 (2.1)	<10 (<3.3)
Black	544 (25.3)	110 (25.6)	209 (25.8)	111 (21.7)	14 (19.2)	100 (30.4)	475 (25.7)	69 (22.6)
Hawaiian/Pacific Islander	<10 (<0.5)	0 (0.0)	<5 (<0.6)	<5 (<1.0)	0 (0.0)	<10 (<3.0)	<10 (<0.5)	<10 (<3.3)
Other	497 (23.1)	107 (24.9)	166 (20.5)	110 (21.5)	15 (20.5)	99 (30.1)	454 (24.6)	43 (14.1)
Unknown/Not Stated	37 (1.7)	7 (1.6)	19 (2.3)	<5 (<1.0)	<5 (<6.8))	<10 (<3.0)	37 (2.0)	0 (0.0)
White	1021 (47.4)	198 (46.0)	397 (49.1)	272 (53.2)	41 (56.2)	113 (34.3)	836 (45.3)	185 (60.7)
Ethnicity (%)								
Hispanic or Latino	560 (26.0)	125 (29.1)	196 (24.2)	111 (21.7)	18 (24.7)	110 (33.4)	511 (27.7)	49 (16.1)
Not Hispanic or Latino	1557 (72.4)	302 (70.2)	596 (73.7)	393 (76.9)	55 (75.3)	211 (64.1)	1308 (70.8)	249 (81.6)
Unknown/Not Stated	35 (1.6)	3 (0.7)	17 (2.1)	7 (1.4)	0 (0.0)	8 (2.4)	28 (1.5)	7 (2.3)
Age (%)								
18–34	146 (6.8)	73 (17.0)	35 (4.3)	18 (3.5)	4 (5.5)	16 (4.9)	145 (7.9)	1 (0.3)
35–44	181 (8.4)	61 (14.2)	58 (7.2)	28 (5.5)	4 (5.5)	30 (9.1)	176 (9.5)	5 (1.6)
45–54	257 (11.9)	54 (12.6)	104 (12.9)	48 (9.4)	8 (11.0)	43 (13.1)	248 (13.4)	9 (3.0)
55–64	435 (20.2)	67 (15.6)	166 (20.5)	101 (19.8)	14 (19.2)	87 (26.4)	399 (21.6)	36 (11.8)
65–74	402 (18.7)	60 (14.0)	171 (21.1)	81 (15.9)	17 (23.3)	73 (22.2)	351 (19.0)	51 (16.7)
75–84	373 (17.3)	58 (13.5)	137 (16.9)	104 (20.4)	18 (24.7)	56 (17.0)	278 (15.1)	95 (31.1)
85+	358 (16.6)	57 (13.3)	138 (17.1)	131 (25.6)	8 (11.0)	24 (7.3)	250 (13.5)	108 (35.4)
Comorbidity (%)								
AIDS/HIV	32 (1.5)	3 (0.7)	13 (1.6)	8 (1.6)	3 (4.1)	5 (1.5)	28 (1.5)	4 (1.3)
Alcohol abuse	175 (8.1)	35 (8.1)	68 (8.4)	39 (7.6)	6 (8.2)	27 (8.2)	151 (8.2)	24 (7.9)
Blood loss anemia	142 (6.6)	16 (3.7)	62 (7.7)	37 (7.2)	5 (6.8)	22 (6.7)	98 (5.3)	44 (14.4)
Cardiac arrhythmias	801 (37.2)	134 (31.2)	318 (39.3)	215 (42.1)	27 (37.0)	107 (32.5)	639 (34.6)	162 (53.1)
Chronic pulmonary disease	716 (33.3)	110 (25.6)	281 (34.7)	195 (38.2)	32 (43.8)	98 (29.8)	587 (31.8)	129 (42.3)
Coagulopathy	264 (12.3)	39 (9.1)	109 (13.5)	64 (12.5)	7 (9.6)	45 (13.7)	201 (10.9)	63 (20.7)
Congestive heart failure	493 (22.9)	61 (14.2)	197 (24.4)	145 (28.4)	19 (26.0)	71 (21.6)	380 (20.6)	113 (37.0)
Deficiency anemia	407 (18.9)	65 (15.1)	162 (20.0)	106 (20.7)	19 (26.0)	55 (16.7)	323 (17.5)	84 (27.5)
Depression	636 (29.6)	101 (23.5)	253 (31.3)	182 (35.6)	27 (37.0)	73 (22.2)	526 (28.5)	110 (36.1)
Diabetes, complicated	623 (28.9)	87 (20.2)	232 (28.7)	159 (31.1)	27 (37.0)	118 (35.9)	499 (27.0)	124 (40.7)
Diabetes, uncomplicated	809 (37.6)	127 (29.5)	305 (37.7)	199 (38.9)	33 (45.2)	145 (44.1)	658 (35.6)	151 (49.5)
Drug abuse	172 (8.0)	34 (7.9)	60 (7.4)	41 (8.0)	8 (11.0)	29 (8.8)	151 (8.2)	21 (6.9)
Fluid and electrolyte disorders	904 (42.0)	149 (34.7)	338 (41.8)	242 (47.4)	41 (56.2)	134 (40.7)	721 (39.0)	183 (60.0)
Hypertension, complicated	615 (28.6)	78 (18.1)	237 (29.3)	176 (34.4)	25 (34.2)	99 (30.1)	478 (25.9)	137 (44.9)
Hypertension, uncomplicated	1386 (64.4)	234 (54.4)	537 (66.4)	365 (71.4)	46 (63.0)	204 (62.0)	1133 (61.3)	253 (83.0)
Hypothyroidism	448 (20.8)	66 (15.3)	183 (22.6)	121 (23.7)	14 (19.2)	64 (19.5)	353 (19.1)	95 (31.1)
Liver disease	283 (13.2)	56 (13.0)	104 (12.9)	69 (13.5)	10 (13.7)	44 (13.4)	232 (12.6)	51 (16.7)
Lymphoma	29 (1.3)	6 (1.4)	10 (1.2)	10 (2.0)	73 (100.0)	3 (0.9)	23 (1.2)	6 (2.0)
Metastatic cancer	140 (6.5)	24 (5.6)	55 (6.8)	42 (8.2)	4 (5.5)	15 (4.6)	111 (6.0)	29 (9.5)
Obesity	682 (31.7)	112 (26.0)	264 (32.6)	161 (31.5)	29 (39.7)	116 (35.3)	582 (31.5)	100 (32.8)
Other neurological disorders	541 (25.1)	66 (15.3)	225 (27.8)	153 (29.9)	20 (27.4)	77 (23.4)	414 (22.4)	127 (41.6)
Paralysis	117 (5.4)	9 (2.1)	46 (5.7)	35 (6.8)	7 (9.6)	20 (6.1)	96 (5.2)	21 (6.9)
Peptic ulcer disease, excluding bleeding	113 (5.3)	22 (5.1)	46 (5.7)	32 (6.3)	2 (2.7)	11 (3.3)	90 (4.9)	23 (7.5)
Peripheral vascular disorders	521 (24.2)	69 (16.0)	200 (24.7)	155 (30.3)	25 (34.2)	72 (21.9)	400 (21.7)	121 (39.7)
Psychoses	206 (9.6)	32 (7.4)	80 (9.9)	67 (13.1)	7 (9.6)	20 (6.1)	166 (9.0)	40 (13.1)
Pulmonary circulation disorders	225 (10.5)	36 (8.4)	73 (9.0)	65 (12.7)	10 (13.7)	41 (12.5)	165 (8.9)	60 (19.7)
Renal failure	513 (23.8)	79 (18.4)	186 (23.0)	144 (28.2)	19 (26.0)	85 (25.8)	392 (21.2)	121 (39.7)
Rheumatoid arthritis/collagen vascular diseases	174 (8.1)	27 (6.3)	74 (9.1)	40 (7.8)	8 (11.0)	25 (7.6)	139 (7.5)	35 (11.5)
Solid tumor without metastasis	313 (14.5)	43 (10.0)	122 (15.1)	88 (17.2)	14 (19.2)	46 (14.0)	249 (13.5)	64 (21.0)
Valvular disease	527 (24.5)	74 (17.2)	229 (28.3)	141 (27.6)	17 (23.3)	66 (20.1)	420 (22.7)	107 (35.1)
Weight loss	356 (16.5)	59 (13.7)	140 (17.3)	107 (20.9)	8 (11.0)	42 (12.8)	277 (15.0)	79 (25.9)

Expired includes those that expired in the ED without known respiratory outcomes.

https://doi.org/10.1371/journal.pone.0243291.t002

Figure 2.5: Discharge and respiratory outcomes (highest requirement during admission) for all patients with known disposition categorized by sex, race, ethnicity and Elixhauser comorbidities.

most common among those who expired were uncomplicated hypertension (n = 253, 83.0%), fluid & electrolyte disorders (n = 183, 60.0%), and cardiac arrhythmia (n = 162, 53.1%). Those who were admitted and/or experienced in-hospital mortality compared to those who tested positive for SARS-CoV-2 in all racial and ethnic groups had an increased comorbidity burden as determined by weighted Elixhauser comorbidity scores (Figure 2.6, panel C and D), with the exception of those with Unknown ethnicity which represented a small number of patients. For those who expired, the most common recorded races were White (n = 185, 60.7%), Black (n = 69, 22.6%), and Other race (n = 43, 14.1%). Those who reported Hispanic ethnicity accounted for 16.1% (n = 49) of in-hospital mortality. In-hospital, age-adjusted mortality rates were 4.1%, 3.8%, 5.3%, 4.0%, and 4.3% for those who reported a race of White, Black, Asian, Hawaiian or Pacific Islander, and Other race, respectively (Figure 2.7). Those who reported Hispanic ethnicity had an age-adjusted in-hospital mortality rate of 4.4%.

As seen with admission, regression analysis demonstrated that increased age had the highest risk for in-hospital mortality (Figure 2.8, Table 2.3), with the largest risk seen for those ≥85 years old (OR 23.3, 95%CI = 10.1–64.1). Male sex was also associated with increased odds of in-hospital mortality (OR 1.76, 95%CI = 1.33–2.35). Of the comorbidities present within the medical history and problem list of the EHR, only a history of blood loss anemia (OR 1.72, 95%CI = 1.07–2.74) and other neurological disorders (OR 1.47, 95%CI = 1.06–2.05) were significant. Race was not statistically associated with a risk of in-hospital mortality in this cohort.

2.3.4 Treatment pathways for admitted patients with Covid-19

Of patients with known outcomes, 1895 (88.1%) received medications for Covid-19 treatment while admitted. We assessed treatment pathways for 13 Covid-19 related medications. Patients were treated with 188 different possible medication regimen permutations with 50 unique combinations (Figure 2.9). The most common first line regimens included hydroxychloroquine (88.3% of patients), tocilizumab (23.8%) and azithromycin (22.7%). The most

Table 2.3: Multivariable analysis with odds ratios for mortality in discharged patients.

	Odds Ratio	CI 2.50%	CI 97.50%	p
(Intercept)	0.008	0.003	0.021	<0.001
Sex				
Male	1.76	1.33	2.35	<0.01
Race or Ethnicity				
Asian	1.76	0.54	5.18	0.32
Black/African-American	1.22	0.64	2.36	0.55
Hispanic	1.34	0.75	2.37	0.32
White	1.18	0.65	2.14	0.59
Age				
45-54	1.78	0.63	5.42	0.28
55-64	4.20	1.83	11.38	<0.01
65-74	7.10	3.13	19.16	<0.01
75-84	15.66	6.91	42.34	<0.01
>85	23.34	10.06	64.09	<0.01
Elixhauser Comorbidities				
AIDS/HIV	1.00	0.27	2.84	0.99
Alcohol abuse	0.91	0.52	1.53	0.73
Blood loss anemia	1.72	1.07	2.74	0.02
Cardiac arrhythmias	1.03	0.74	1.43	0.85
Chronic pulmonary disease	1.01	0.74	1.39	0.93
Coagulopathy	1.31	0.89	1.92	0.16
Congestive heart failure	1.06	0.71	1.56	0.79
Deficiency anemia	0.83	0.58	1.20	0.33
Depression	0.78	0.56	1.08	0.14
Diabetes, complicated	1.10	0.72	1.68	0.66
Diabetes, uncomplicated	0.99	0.66	1.47	0.96
Drug abuse	0.89	0.49	1.55	0.68
Fluid and electrolyte disorders	1.22	0.86	1.72	0.27
Hypertension, complicated	0.88	0.56	1.36	0.56
Hypertension, uncomplicated	1.10	0.73	1.67	0.65
Hypothyroidism	1.12	0.81	1.54	0.51
Liver disease	1.32	0.88	1.94	0.17
Lymphoma	1.21	0.39	3.32	0.72
Metastatic cancer	1.15	0.67	1.93	0.61
Obesity	1.22	0.89	1.69	0.22
Other neurological disorders	1.47	1.06	2.05	0.02
Paralysis	0.83	0.47	1.42	0.51
Peptic ulcer disease, excluding bleeding	0.82	0.46	1.40	0.47
Peripheral vascular disorders	0.94	0.67	1.33	0.74
Psychoses	1.11	0.71	1.70	0.65
Pulmonary circulation disorders	1.45	0.96	2.19	0.07
Renal failure	1.31	0.89	1.93	0.16
Rheumatoid arthritis/collagen vascular diseases	1.01	0.64	1.55	0.96
Solid tumor without metastasis	0.84	0.57	1.21	0.35
Valvular disease	0.75	0.53	1.06	0.11
Weight loss	1.04	0.73	1.48	0.83

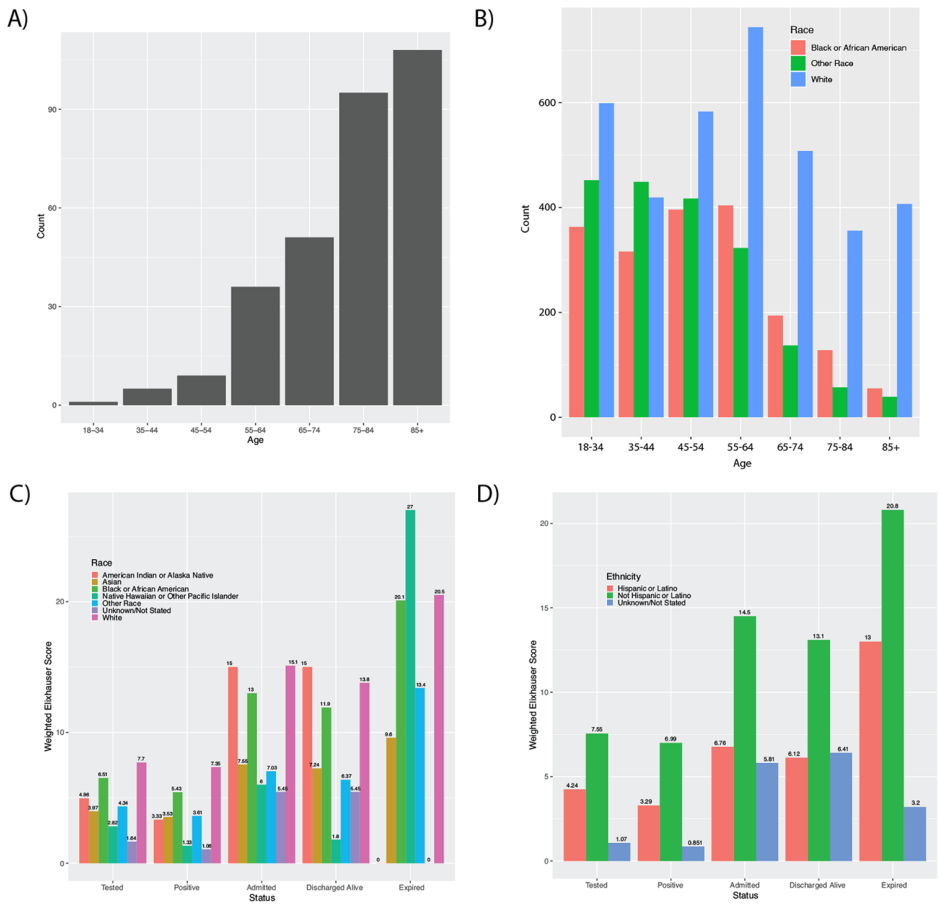


Figure 2.6: A) Frequency of in-hospital mortality by age, B) distribution of age by self-reported race in patients positive for SARS-CoV-2, and weighted Elixhauser comorbidity scores by patient status grouped by C) recorded race and D) recorded ethnicity.

frequent second-line regimens, aside from the most frequent first line agents, included the addition of steroids (21.3%), atazanavir (6.5%), and lopinavir/ritonavir (3.4%). The most common treatment permutations were hydroxychloroquine alone (25.2%), hydroxychloroquine in combination with tocilizumab (18.8%), or hydroxychloroquine in combination with azithromycin (8.1%). A total of just six Covid-19 related medications were given to more than 1% of admitted patients in our cohort: hydroxychloroquine sulfate (94.7%), tocilizumab (51.0%), azithromycin (28.8%), steroids (24.3%), atazanavir (15.5%), and lopinavir/ritonavir (7.8%). All race and ethnicity groups were prescribed hydroxychloroquine most frequently, with patients who self-reported as Asian having the lowest rate (92.7%). Tocilizumab was

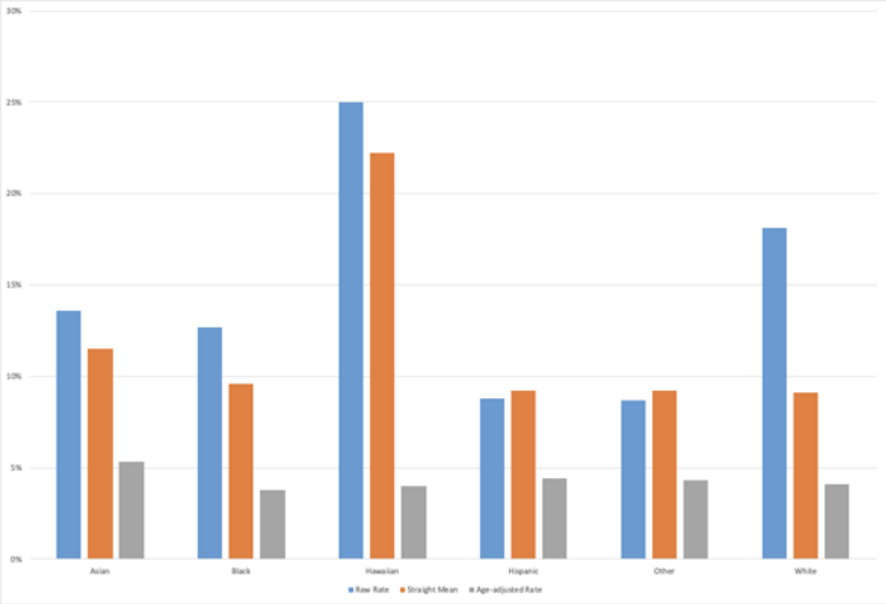


Figure 2.7: In-hospital, age-adjusted mortality in discharged patients with SARS-CoV-2.

the second most frequently prescribed medication in all groups. The use of azithromycin had the most notable variation among groups: it was the second most common medication in those identifying as Other race, with a frequency of 46.7% of patients, but was fifth most common among those who identified as Black, with only 16.7% of patients receiving azithromycin.

2.4 Discussion

In one of the largest real-world analyses of risk factors associated with Covid-19 infection and disease severity, we identified age as the primary risk factor associated with both admission and in-hospital mortality in those infected with SARS-CoV-2. Black race and Hispanic ethnicity were associated with increased risk of admission in our cohort and had increased disease and mortality burden, but age-adjusted in-hospital mortality was similar among all reported races and ethnicities. Comorbidities had much less impact on risk for either admission or in-hospital mortality in our study.

Our work extends the literature in several important ways. Firstly, we followed a single large cohort to identify risks associated with infection and severe disease from the time of

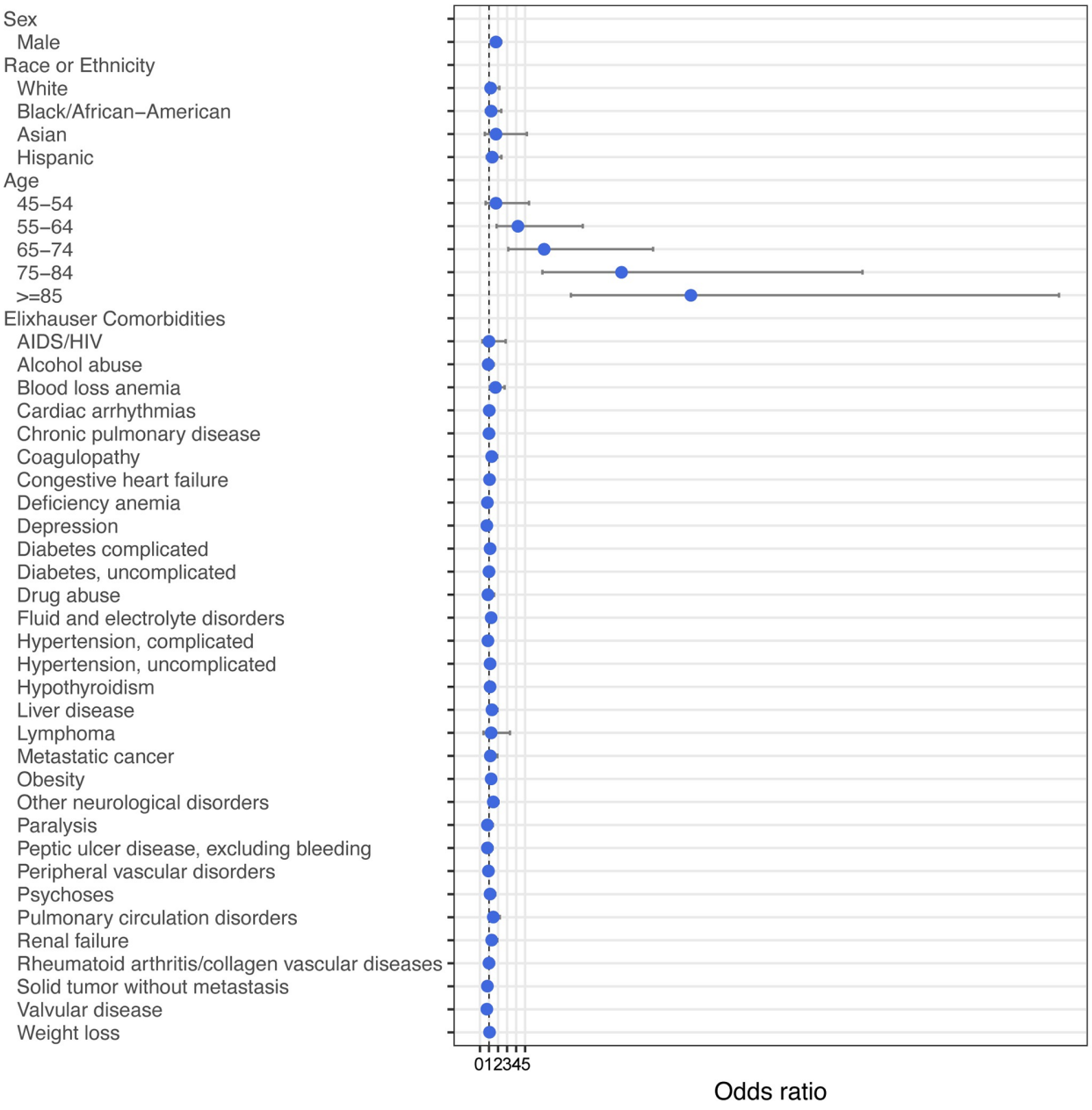


Figure 2.8: Multivariable analysis with odds ratios for mortality in discharged patients.

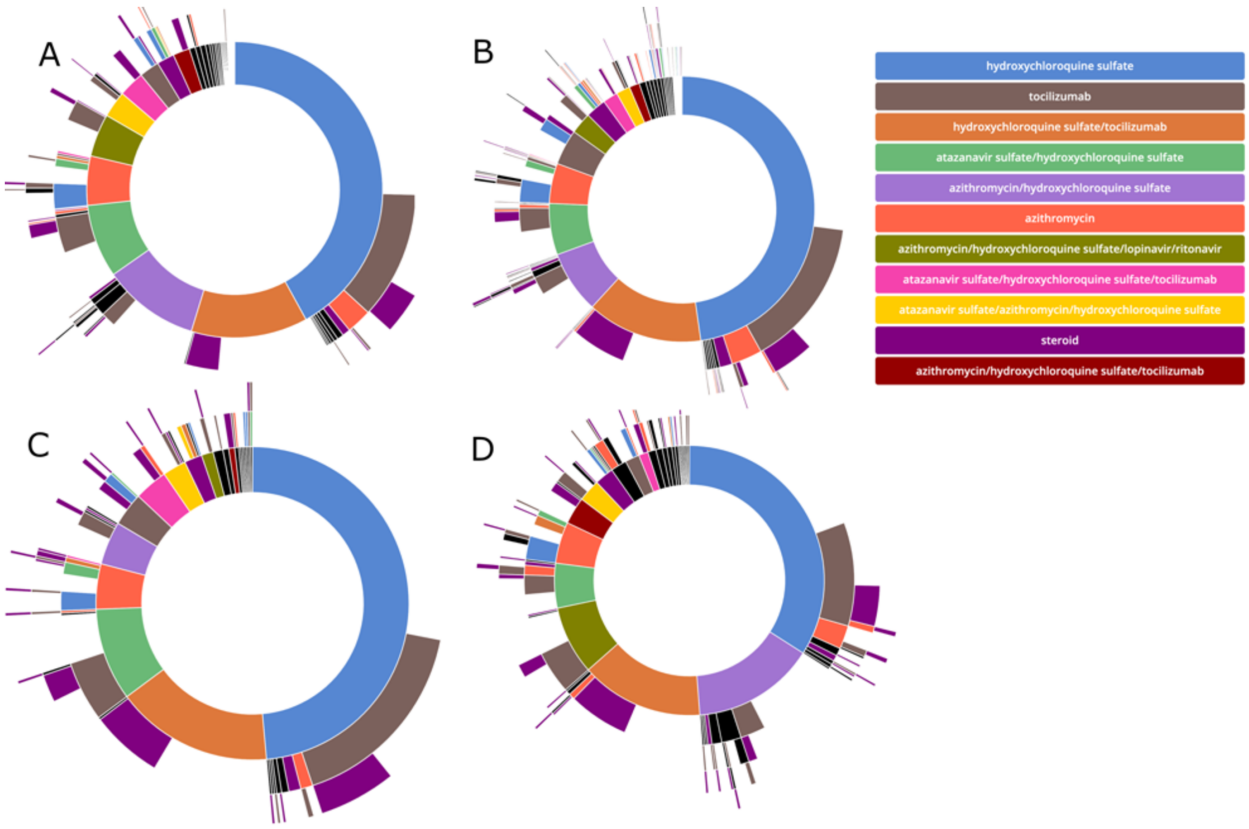


Figure 2.9: Sunburst diagram of medication pathways with individual regimens grouped by order of initiation.

testing through discharge. Secondly, we provide further evidence that age and male sex are significantly associated risk factors for both admission and in-hospital mortality. Thirdly, we found that comorbidities, while common in those with SARS-CoV-2, were not strongly associated with either admission or in-hospital mortality based on multivariable analysis. Fourthly, we found race and ethnicity to be associated with infection and admission in this cohort, but with in-hospital mortality that was similar among these groups in our discharged population. Finally, we identified consistent use of medications within our admitted population, but with many possible treatment pathways for any individual patient and with frequent use of investigational therapies for Covid-19. Further investigation is needed to characterize potential benefits or risks associated with the various treatment pathways that have been available over the course of the pandemic.

Our data confirm findings in other studies that show age as a primary risk factor for admission and in-hospital mortality in adult patients and that male sex is also highly associated with these outcomes [16, 18, 34, 20, 35, 36]. While the mechanisms that may lead to more severe disease in men have not been definitively elucidated, several potential mechanisms have been proposed to explain the demonstrated differences, including increased comorbidities and changes in the immune response in those who are male and older, along with possible genetic/biologic differences that may increase disease severity in men [37]. Similarly, immune senescence, with dysregulated inflammation and decreased adaptive immune response, has been hypothesized as a possible reason for worse disease in older populations [38].

Many studies have shown that Covid-19 has disproportionately affected minority populations across the US [39, 40, 41, 42]. Within our cohort, Black and Hispanic populations were overrepresented in those who were tested, positive, and admitted for SARS-CoV-2 compared to census data for Connecticut. Studies based on regional mortality data, which have included out-of-hospital mortality, have shown that severe disease may also be more prevalent in minority populations [21, 40]. In our study, Black race was overrepresented in those with more severe outcomes compared to state census numbers. However, in the discharged population, we found that age-adjusted, in-hospital mortality was similar among all racial and ethnic groups, with rates ranging from 3.8% to 5.3%. This finding is consistent with other studies of in-hospital mortality related to Covid-19 [14, 34], but also demonstrates that minority populations experience a higher overall burden of disease. While a small percentage of this cohort did not have race or ethnicity data provided, it remains limited by the potential for errors during patient registration and the possibility of provider-reported responses.

Our data reflect the prominence of comorbidities in those with SARS-CoV-2 infection. While comorbidities were common, some of the most commonly reported risks for severe disease [13, 43, 44] were not identified as risks in this study and multivariable analysis did not find a history of hypertension or diabetes to be significantly associated with admission. An increased comorbidity burden was noted in those with in-hospital mortality compared to

those who were discharged alive. However, multivariable analysis only identified a history of blood loss anemia to be significantly associated with in-hospital mortality. Other comorbidities, such as obesity, were associated with admission but not in-hospital mortality. It is unclear if these patients required admission due to more severe disease or were admitted due to perceived risk based on early reports of Covid-19 risk factors. Similarly, a history of drug abuse and psychoses were associated with admission, but likely represented more frequent testing in these populations with limited ability to discharge patients to shared facilities following a positive SARS-CoV-2 test. These findings highlight the fact that age and sex appear to be the predominant drivers of severe disease. Additional studies will be needed to further characterize the risk of underlying disease on the severity of Covid-19.

The risks and outcomes reported here should be assessed in the context of the treatment protocols used during this period of the epidemic. Treatment standards based on early recommendations led to a majority of patients receiving disease related therapy, often with investigational treatments. Of patients who received a Covid-19 targeted therapy, 94.7% received hydroxychloroquine, 51.0% received tocilizumab, and 28.8% received azithromycin. The use of Covid-19 directed treatments was consistent among races and ethnicities in our cohort. But despite an early push to use promising medications from in vitro studies, such as hydroxychloroquine and azithromycin, evidence now demonstrates that neither is likely beneficial for admitted patients. As such, the treatment context of future studies should be similarly assessed to determine whether changes in treatment pathways impact the reported risks and outcomes in those with Covid-19.

Our analysis leveraged real-world data derived from the EHR to assess all patients tested for SARS-CoV-2 within our health system. We implemented computed phenotypes to identify cases and clinically relevant outcomes, with a subset manually reviewed for accuracy. Our findings add to a growing base of evidence related to Covid-19 risk factors and outcomes. However, as an observational study based on real-world data, this study also has several limitations. First, while standardized testing protocols were in place, testing was of-

ten limited to symptomatic individuals or those with known exposure risks, thus potentially biasing our cohort to those who were symptomatic and sought care. The study was also limited to a single health system, but one that consists of a mixture of academic, urban, and suburban care facilities with a diverse patient population. In addition, while our health system implemented standardized treatment protocols, patients received therapies that were investigational for Covid-19 at the time of the study and use of these medications may not be similar at all institutions, especially as Covid-19 treatment protocols rapidly evolve as new evidence is obtained. Another limitation is that features associated with risk of admission may not correlate to risk of disease severity, as the decision to admit can be impacted based on discharge options or perceived clinical risks by healthcare providers. Finally, due to the timeline of the current outbreak, this study was limited to the initial admission and only assessed in-hospital mortality. Therefore, additional studies are needed to assess the impact of disease on patients not admitted to the hospital and the long-term effects of SARS-CoV-2 infection.

There is an ongoing need to rapidly generate and communicate evidence, while also being cautious that only high-quality data are used to inform policy and develop clinical recommendations. While waiting for larger, more comprehensive case control and population-scale studies to define COVID-19 specific risks, prevalence, treatment, and outcomes, providers and public health officials need the best available evidence for clinical use. The data presented here provide findings from a large cohort that was followed from testing through discharge, identified increased age and male sex as the strongest risk factors for admission and in-hospital mortality, and found that in-hospital mortality was similar in racial and ethnic groups within our health system. Ongoing studies that further elucidate the risk of comorbidities, particularly given rapidly evolving treatment guidelines, remain needed as the Covid-19 pandemic continues to grow.

2.5 Conclusion

The early COVID-19 experience at YNHH demonstrated that increasing age and male sex are the risks most strongly associated with admission and in-hospital mortality in those with SARS-CoV-2 infection. Minority racial and ethnic groups had increased risk of admission and higher disease burden, including mortality. But, for discharged patients, in-hospital mortality rates were similar in all racial and ethnic groups. While comorbidities were frequently observed in patients with SARS-CoV-2, few were associated with admission or in-hospital mortality in our cohort. Despite the limitations, this dataset from a multi-hospital health system with a diverse patient population presents valuable information related to risk factors for SARS-CoV-2 infection and short-term outcomes.

3. VISUALIZATION OF EMERGENCY DEPARTMENT CLINICAL DATA FOR INTERPRETABLE PATIENT PHENOTYPING*

While analysis of data can allow for specific and direct relationships to outcomes to be found, it can also allow for finding more nebulous relationships. Whereas the work outlined in Chapter 2 dealt with supervised learning, the work presented in this chapter examines unsupervised learning. Here, we use a combination of techniques in order to assess and characterize heterogeneity in patients presenting to an emergency department. We do this as an aid for rapid clinical understanding. The approach allows for newly presenting patients to rapidly be mapped to an existing phenotype of patients, and allows for visual inspection of a 2D representation of those phenotypes.

3.1 Introduction

Electronic health records (EHRs) include heterogeneous data that represent past and ongoing patient care episodes. The EHR is accessed as both a real-time information transfer environment as well as a medium for retrospective analysis. In the emergency department (ED) setting, patients are, in the vast majority of cases, first seen by medical professionals for registration and triage. This process links a digital record to the individual waiting to be seen and enables the rapid assessment of patient complexity as well as the visit urgency.

ED triage is made more challenging by increases in patient volume [45] and relative subjectivity in the triage process [46]. The emergency severity index (ESI) is a five-level triage system developed to improve robustness in nurse-driven triage assessments and has been shown to correlate with admission rate and mortality [47]. However, this approach is not designed to leverage the breadth of data available in the EHR. More recently, various machine learning approaches to augmenting triage have been described [48][49][50] and there

*This chapter is reprinted with permission from "Visualization of Emergency Department Clinical Data for Interpretable Patient Phenotyping" by Hurley, N. C., Haimovich, A. D., Taylor, R. A., & Mortazavi, B. J., 2019. arXiv preprint arXiv:1907.11039. Copyright 2019 by Nathan C. Hurley.

is mounting evidence to suggest that incorporating heterogenous data into initial patient assessments enables a more refined and accurate prediction of critical patient outcomes.

While risk and event prediction is now a mature field within medical informatics [51], there is growing interest in leveraging massive EHR databases to uncover phenotypes of complex disease processes. These efforts have multiple aims which include the automated discovery of patient populations for retrospective analyses and the identification of specific subgroups that may benefit from particular interventions or therapies [52]. Prior work has shown potential utility in varied fields including hematology [53] and cardiology [52].

In the ED setting, however, there is a pressing need for tools that enable prospective and interpretable phenotyping. Visualization offers one appealing approach to interpretability and techniques like t-distributed stochastic neighbor embedding (t-SNE) [54] have been used in varied settings with great success [55]. However, a given visualization produced by t-SNE is unable to be expanded to future data points, as t-SNE produces a non-parametric visualization [54]. In contrast, uniform manifold approximation and projection (UMAP) is a parametric visualization technique that preserves more global structure than does t-SNE while also allowing future data points to be fit to an existing model without recreating the entire model [56].

Here, we describe the first application of EHR phenotype visualization to the ED triage process. Using a database containing 560,486 anonymized patient visits with 972 sparsely-populated features, we have previously shown the ability to robustly predict patient disposition (hospital admission or discharge) using a very small subset (n = 15) of these features[48]. In that work, we found that an XGBoost model trained on triage score, medication counts, demographics, and hospital usage statistics could predict hospital admission with an AUC of 0.91. That work found that although models trained on triage data or history data performed similarly, models trained on both triage and history data showed a marked improvement in predicting admission. However, that work focused on the binary outcome of admission or discharge. In the present work, we expand on this prior work by implementing and validat-

ing a technique to visualize subpopulations within this dataset. We show that new patients can be mapped into this visualization, and we show that patient similarities and differences to local subpopulations can give information about likely clinical outcomes in order to aid clinical assessment. We anticipate that this work could be utilized in a clinical setting in order to aid in understanding relationships between patients and to aid in clinical decision making.

The contributions of this work are:

- A method of using UMAP for non-linear dimensionality reduction, Gaussian mixture models (GMMs) for clustering data, and a combination for data visualization.
- A metric utilizing the adjusted Rand index (ARI) between different folds of data clustering to determine appropriate model hyperparameters and cluster stability in random subsets of the data.
- An application in real-world clinical data of patients presenting with five common clinical chief complaints. The emergent properties of these clusters are described, and clinically-relevant attributes are discussed to show clinical decision support potential.

The rest of this paper is organized as follows. In section 3.2, we discuss related work in patient phenotyping and dimensionality reduction. We also discuss the utilization of ARI as a metric for measuring partition similarity. In section 3.3, we discuss the data used and the process of building our model to find and validate clusters within the clinical dataset. In section 3.4 we show the model embeddings and results on both the synthetic and clinical dataset. We discuss the clinical characteristics of the clusters discovered within the clinical dataset. In section 3.5, we discuss the significance of the results and some of the clinical pictures that can be drawn from the results. We then discuss directions for future applications of this work.

3.2 Related Work

Previous patient phenotyping work has focused on identifying phenotypes among patients with a given disease state, such as heart disease [57], sepsis [58], or amyotrophic lateral sclerosis [59]. In these works, authors have used various clustering techniques, such as hierarchical agglomeration after reducing dimensionality with principal component analysis (PCA) [57] or by using K-means clustering on patient clinical severity scores [58]. Clusters have also been discovered through training a semi-supervised denoising autoencoder, and then using PCA or t-SNE to represent the hidden nodes of the autoencoder and identifying the clusters manually [59]. In contrast, our approach is not specifically tailored to a given disease state, but rather is applied to any patient presenting for emergency care with a particular chief complaint.

Other work has looked at enumerating a wide range of specific phenotypes, and then identifying those phenotypes within a patient population [60]. However, these approaches are supervised techniques; physicians with top-level domain knowledge drive the phenotype discovery. In these approaches, the phenotype is identified clinically, and new patients are matched to the phenotype clusters using custom rules and logic. In contrast, our approach is not trained with a particular clinical condition or outcome in mind, but instead searches for phenotype clusters within a given patient set. This allows for application of our technique to new and unseen phenotypes without the need for expert evaluation.

Dimensionality reduction techniques are often used in visualizing multidimensional data. Earlier approaches have used PCA and GMMs to visualize populations of samples in a 2D space in order to aid in clinical diagnosis [61]. PCA has also been used in conjunction with other clustering methods such as agglomerative clustering or K-means clustering [62]. However, in sparse datasets PCA often reaches a limit where it is unable to express data without losing a significant amount of information about the dataset variability.

Other sparse clinical datasets have been visualized with t-SNE [55]. t-SNE is a powerful tool that embeds data into lower dimensions while maintaining structure present at higher

dimensions [54]. However, t-SNE is non-parametric; while it is effective at embedding a given dataset, it is difficult to embed new, previously unseen members of that dataset. UMAP is a newer dimensionality reduction technique that is able to scale beyond t-SNE while also expressing relations that t-SNE is unable to express [56]. UMAP is a parametric approach, and as such is able to embed new data without necessitating a retraining of the model. Although several studies have utilized UMAP for phenotyping cell populations [63][64], we did not find any studies that utilize UMAP for patient phenotyping with EHR data.

The adjusted Rand index (ARI) is a widely-used metric of similarity between two partitions of a given set of objects [65]. The ARI for two partitions of a set ranges from a value of 1, when the two partitions are equivalent, to near 0 when the two partitions are chosen at random. The closer the ARI of two partitions is to 1, then the more similarly partitioned the set is. ARI is robust to partitions of different sizes, and has been used for evaluating the results of various clustering techniques [66].

3.3 Methods

In this section, we describe our visualization technique. The objective is to visualize an embedding of EHR data which preserves global structure so that physicians can rapidly infer relationships between new patients and well-defined subpopulations of patients. We first filter our data by chief complaint upon presentation to the ED. Next, we split our data into a training set and a validation set. The training set is split into five partitions which are embedded in two dimensions using UMAP. GMMs are trained on each partition, and then the model trained on each partition is applied to the validation set. The number of clusters in the final model is determined by maximizing the ARI among labels for the validation set. A diagram of this process in shown in Figure 3.1. We perform this process both on synthetic data to demonstrate the method's viability, and on a clinical dataset to demonstrate the clinical applicability. All code used here is available online at `https://github.com/nch08a/EDVizPhenotyping`.

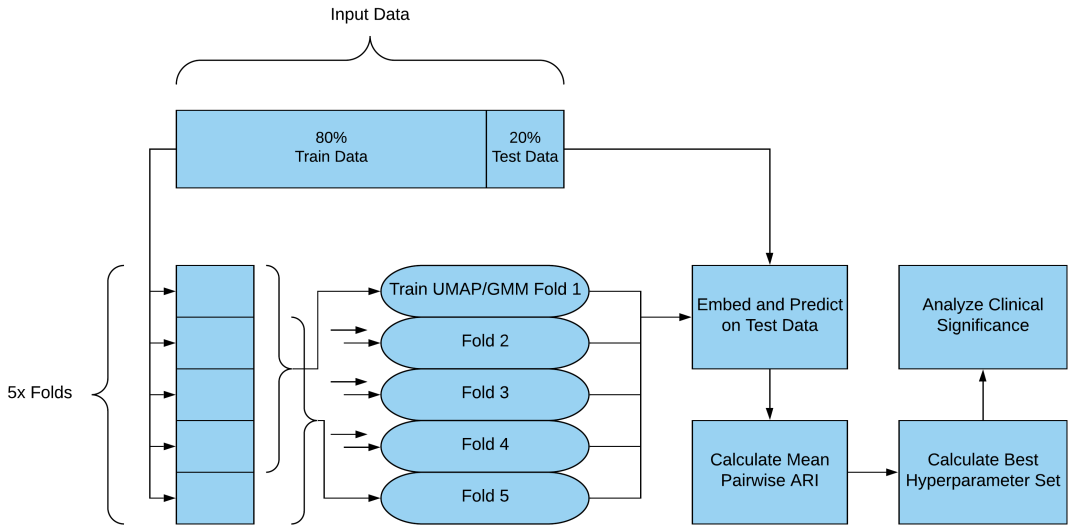


Figure 3.1: A diagram of the method presented here. The data is randomly split into training data (80%) and testing data (20%). The training data is split into five folds. Each combination of four folds is used to train a separate UMAP -> GMM model, which is then applied to the testing data. The mean pairwise ARI is calculated between each test data cluster prediction, and the set of hyperparameters giving the best agreement is selected for clinical analysis.

3.3.1 Datasets

3.3.1.1 Synthetic Data

A synthetic dataset was generated using the make_classification package in scikit-learn [67]. The dataset was generated with 100,000 samples of 100 features, 50 of which were informative. The dataset was generated to have 10 classes, and for the class separation to be 0.75. These parameters were chosen to generate a dataset with noise and a large amount of high-dimensional structure. Having 50 informative features prevents the data from being easily explained by any simple projection to a 2D space. Utilizing non-informative features adds noise to the model in a way similar to correlated patient data. A class separation of 0.75 results in overlapping distributions, so that samples do not clearly align with one cluster or another. With this overlap, there may be some samples which are not able to be correctly classified even with full model information. This noise allows the dataset to reflect

Table 3.1: Selected Patient Characteristics

	Abdominal Pain		Chest Pain		Shortness of Breath		Back Pain		Falls	
Count	54,315		35,778		24,652		20,633		19,012	
Disposition- Admit (%)	19,482	35.9%	16,065	44.9%	15,791	64.1%	3,061	14.8%	5,642	29.7%
Gender— Male (%)	19,169	35.3%	16,587	46.4%	10,231	41.5%	9,010	43.7%	7,823	41.2%
Insurance Status— Medicare (%)	20,833	38.4%	11,778	32.9%	6,530	26.5%	8,690	42.1%	4,677	24.6%
Insurance Status— Medicaid (%)	8,957	16.5%	8,376	23.4%	9,426	38.2%	3,234	15.7%	6,890	36.2%
Language— English (%)	48,560	89.4%	32,479	90.8%	22,861	92.7%	18,537	89.8%	17,811	93.7%
Arrival via Ambulance (%)	12,532	23.1%	13,603	38.0%	11,470	46.5%	4,168	20.2%	10,688	56.2%
Mean Age in Years (range)	45.7	18-105	53.4	18-105	61.2	18-107	47.3	18-103	61.1	18-107
Mean Triage Heart Rate (range)	85.8	35-187	84.6	30-240	88.8	30-199	84.3	44-205	83.6	30-180
Mean Triage Systolic BP(range)	132.5	59-248	135.7	59-274	134.7	60-312	134.0	63-246	135.0	59-261
Mean Triage Diastolic BP (range)	80.6	27-194	81.8	28-172	80.2	30-214	81.3	32-157	80.2	28-156
Mean Triage Respiratory Rate (range)	17.6	8-61	17.7	8-69	18.6	10-66	17.6	8-64	17.5	18-57
Mean Triage O_2 Saturation (range)	97.4	67-99	97.3	60-99	96.6	60-99	97.4	71-99	97.1	73-99
Mean Triage Temperature in °F (range)	98.1	94.5-104.7	98.0	94.1-104	98.1	90.1-106	98.0	94.3-104.4	98.0	93.3-103.4
Mean Prior Admissions (range)	1.2	0-49	1.4	0-48	1.7	0-40	0.6	0-46	0.9	0-42

difficulties in clinical datasets.

3.3.1.2 Clinical Data

The clinical dataset used in this study was previously detailed and is publicly available [48]. This dataset includes 560,486 patient visits at three EDs, collected from March 2014-July 2017. This dataset was preprocessed to include all adult patients who were either discharged or admitted. Data collected about the patients includes disposition, triage evaluation, chief complaint, hospital usage, past medical history, outpatient medications, historical labs and vitals, and imaging/ekg counts. The dataset used in this paper is available online as described in [48].

Early analysis did not show cluster stability when the method was applied to the entire dataset, and so the data was broken down into subsets by patient chief complaint. The five largest chief complaint subsets were analyzed here, but this technique could be applied to any chief complaint subset. The five subsets analyzed here were the only subsets consisting of at least 3% of the total dataset. Statistics relating to the patient chief complaint subpopulations are shown in Table 3.1.

These patients represent a wide variety of adult patients who presented at an ED. ED care is difficult as the patient population seen is highly variable: there are acute, emergent cases, as well as patients who come to the ED for more routine care that could be more appropriate in other health care settings. Therefore, the patients shown in Table 3.1 represent a wide

spectrum from healthy to critically ill. One proxy for severity is the means by which a patient arrives at the ED; a patient arriving via ambulance is more likely to be sick than a patient who walked or drove themselves. Number of prior admissions can be a proxy for long term health of a patient. Patients with chronic illness are more likely to have been hospitalized a larger number of times, while patients without chronic illness are less likely to have been hospitalized a larger number of times.

3.3.2 Data Preprocessing

Each chief complaint dataset was split into 80% training and 20% testing. The training set was further split into 5 training folds by 5-fold cross validation. However, the testing folds produced by this cross validation were not used for testing, as the classification metric used (ARI) requires the same test fold for every model trained. All splits were random with no prior weighting of the datasets. Admission data for the given ED visit and emergency severity index (ESI) were omitted so as to censor the model from the eventual clinical outcome and from direct triage assessment. Within each fold, the categorical data was one-hot encoded. The numerical training data was normalized to have unit range and zero mean. Missing data was mean imputed. The test data was normalized by the same scaling factor that produced the training normalization, and missing data was imputed to the mean of the training data.

3.3.3 Dimensionality Reduction and Clustering

Both datasets were embedded into two dimensions for ease of visualization. Two methods were examined for dimensionality reduction: PCA and UMAP. These methods were chosen for their ability to provide a parametric dimensionality reduction technique. This way, the transformations can be trained and subsequently applied to previously unseen data, allowing for clinical phenotyping of new patients.

UMAP was trained with 2, 15, or 150 neighbors included in the local manifold approximation. These parameters were chosen to represent a spectrum of distance at which structure is considered- a smaller number of neighbors in the approximation emphasizes local structure,

while a larger number of neighbors emphasizes global structure. The minimum embedding distance between points was set to 0, 0.1, or 0.25. These parameters were chosen to allow for samples to “clump" to varying amounts. Smaller distance between points allows more similar points to be stacked together at the expense of losing their relationship to more distant points. All distances were calculated using the Euclidean metric. After training each UMAP model on a given training fold, that model was applied to the unseen test fold.

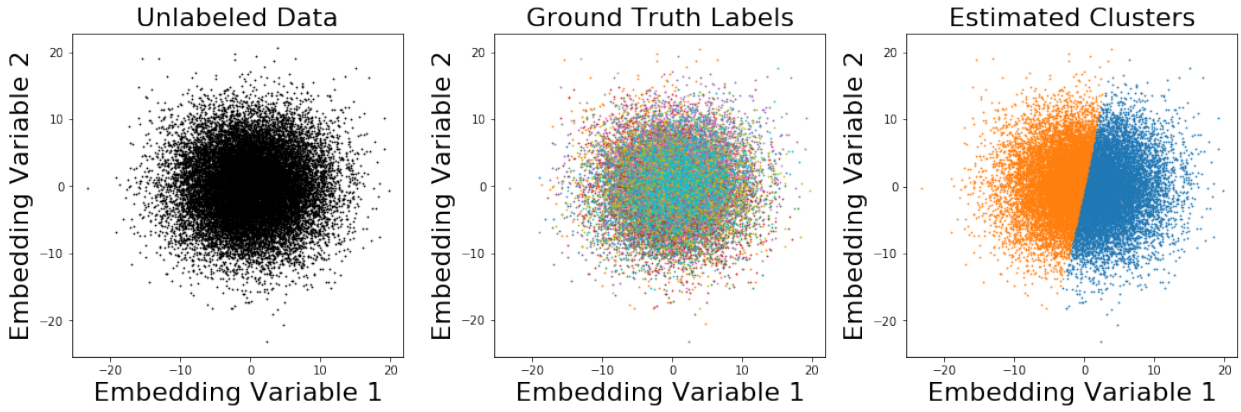
Full covariance GMMs were trained on the training data using an expectation-maximization algorithm. The GMMs were trained with number of clusters n ranging from 2 to 20. These models were then used to predict labels of the testing dataset. The testing data was clustered once for each training fold.

3.3.4 Clustering Analysis

The different clusterings of testing data were analyzed for stability using the ARI. The mean pairwise ARI was computed between each test set labeling produced with a given value of n. For some values of n, the model failed to find n non-null clusters. The final cluster number was selected by choosing the value of n that maximized mean pairwise ARI while still finding at least a mean of $n - 0.5$ clusters. A high ARI indicates that the clustering process is stable among a given dataset, as certain entries are consistently assigned to the same cluster. A low ARI indicates that the clustering process is dominated by noise, and that there is a large amount of variability between different folds.

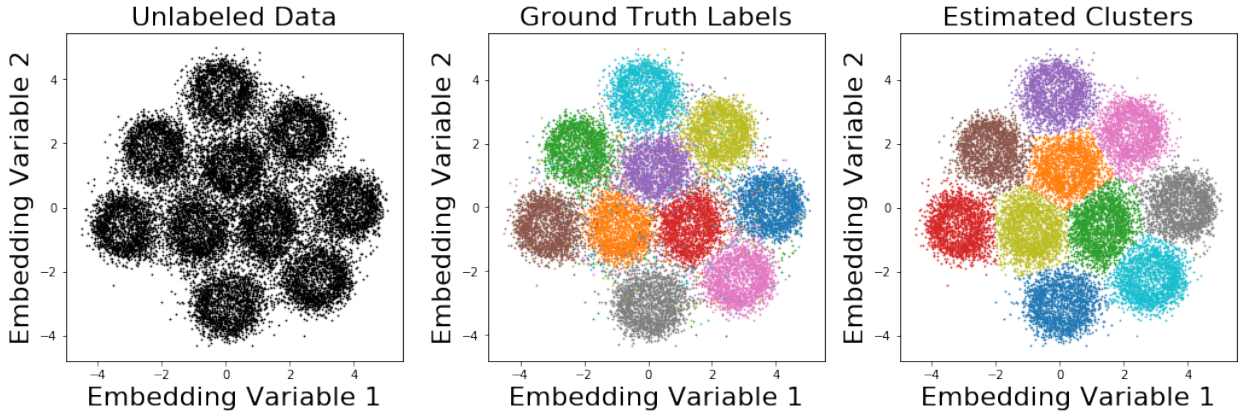
3.3.5 Clinical Cluster Analysis

Clusters may be compared to each other or to a dataset as a whole. In this analysis, comparisons were performed by calculating the difference of the mean of all normalized variables in a cluster with the mean of all normalized variables in the entire chief complaint test set. Training data was not used in analysis. Variables with differences furthest from zero represent the features in the cluster that are most distinct from the rest of the dataset.



(a) (b) (c)

Figure 3.2: PCA Embeddings of synthetic data. Here the embedding and GMMs have been trained on one of the five training splits, and then applied to the test set. This test set application is shown here. In 3.2a, the embedded data is shown without labels. In 3.2b, all data has been labeled with the ground truth cluster identities. In 3.2c, the GMM-predicted clusters are shown.



(a) (b) (c)

Figure 3.3: UMAP Embeddings of synthetic data. Here the embedding and GMMs have been trained on one of the five training splits, and then applied to the test set. This test set application is shown here. In 3.3a, the embedded data is shown without labels. In 3.3b, all data has been labeled with the ground truth cluster identities. In 3.3c, the GMM-predicted clusters are shown.

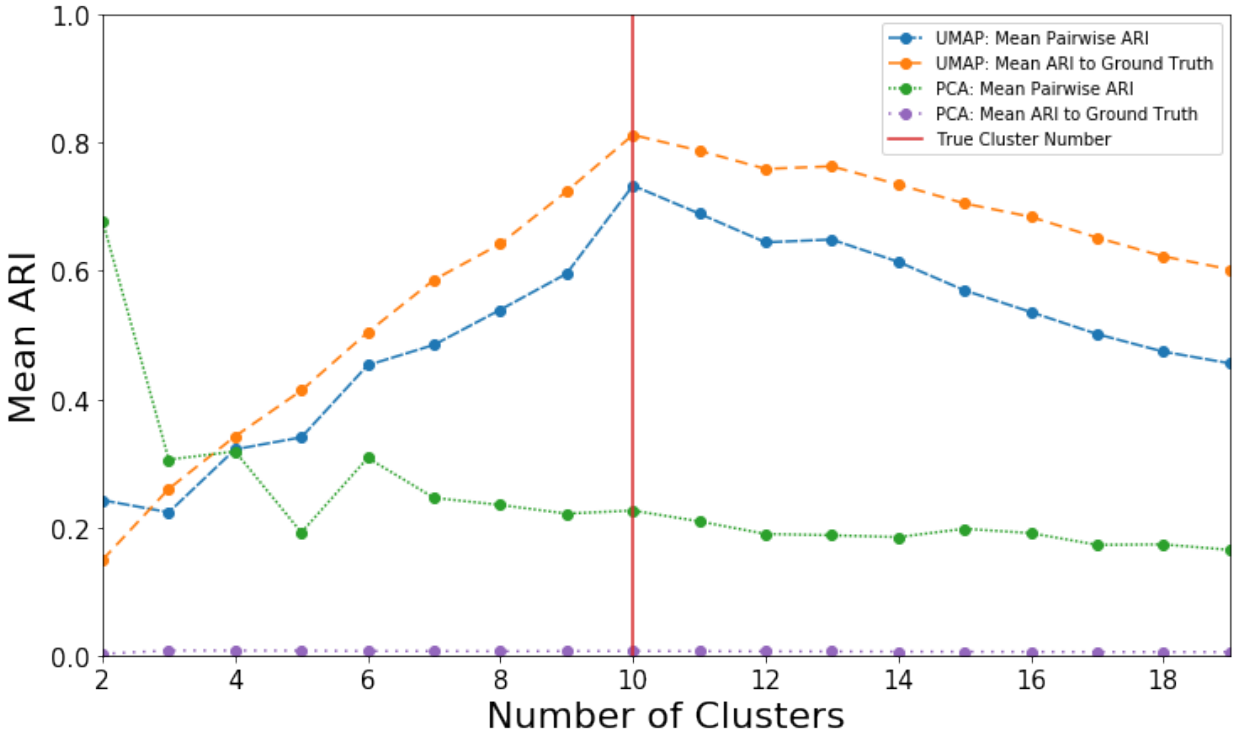


Figure 3.4: Mean pairwise ARIs of clusterings on synthetic data. The solid line denotes the true number of clusters. ARIs are shown both pairwise between different training folds and with respect to the ground truth cluster labeling.

3.4 Results

3.4.1 Synthetic Data

Embeddings of the synthetic data are shown in Figures 3.2 and 3.3. Figure 3.2 shows PCA embeddings. The embeddings used to generate this image were trained on a randomly chosen training split, and applied to the test set. Figure 3.2a shows the embedding without any cluster labels. Figure 3.2b shows the points labeled with their ground truth cluster identity, and Figure 3.2c shows the most stable prediction from GMMs. As can be seen by comparing these figures, the GMM here does not show any strong relationship with the true cluster identities. ARIs of GMMs trained on this data can be seen in Figure 3.4. Although small cluster number shows high pairwise ARI, no choice of cluster number ever provides a high ARI with respect to the ground truth cluster labels. These data suggest that PCA

struggles to reveal true clusters.

We then trained UMAP embeddings using random training splits on synthetic data (Figure 3.3) and applied the mapping to held-out test data (Figure 3.3b). We then used GMMs to discover the most stable cluster predictions (Figure 3.3c). In contrast to the PCA data in Figure 2, we observed significantly improved capture of phenotypes within the synthetic dataset. While the embedding does not perfectly separate the clusters, a clear trend towards separation can be seen. ARIs of the GMMs trained here can be seen in Figure 3.4. The peak ARI is reached when grouping with 10 clusters, which is the true number of clusters present.

3.4.2 Clinical Data

Within our emergency department electronic health record database, the most common chief complaints were abdominal pain (present in 54,315 patient visits, 9.7% of total), chest pain (35,778, 6.4%), shortness of breath (24,652, 4.4%), back pain (20,633, 3.7%) and fall (19,012, 3.4%). All other chief complaints were present in less than 3% of patient visits. Of note, patient visits could have multiple chief complaints, with the average patient visit having 1.13 chief complaints. We sought to implement our embedding and clustering pipeline to visits within each of these categories. We then compared clusters to one another to determine the features driving the phenotype.

3.4.2.1 Shortness of Breath

We first explored hyperparameters for the Shortness of Breath population (Figure 3.5). In this plot, the highest ARI (15 neighbors, 0 min distance, 3 clusters) is invalidated as a best possible option, as the mean fold produced 1.2 clusters, which is below the cutoff of no more than 0.5 below the number of clusters used for training the model. The next highest ARI (150 neighbors, 0.1 min distance, 2 clusters) was chosen for further analysis. This analysis was replicated across chief complaints and summarized in Table 3.2.

The best clustering of shortness of breath was found with two clusters that contain 72.1% (95% CI 67.5-76.7) and 27.9% (95% CI 23.3-32.5) of patients. The larger cluster was slightly

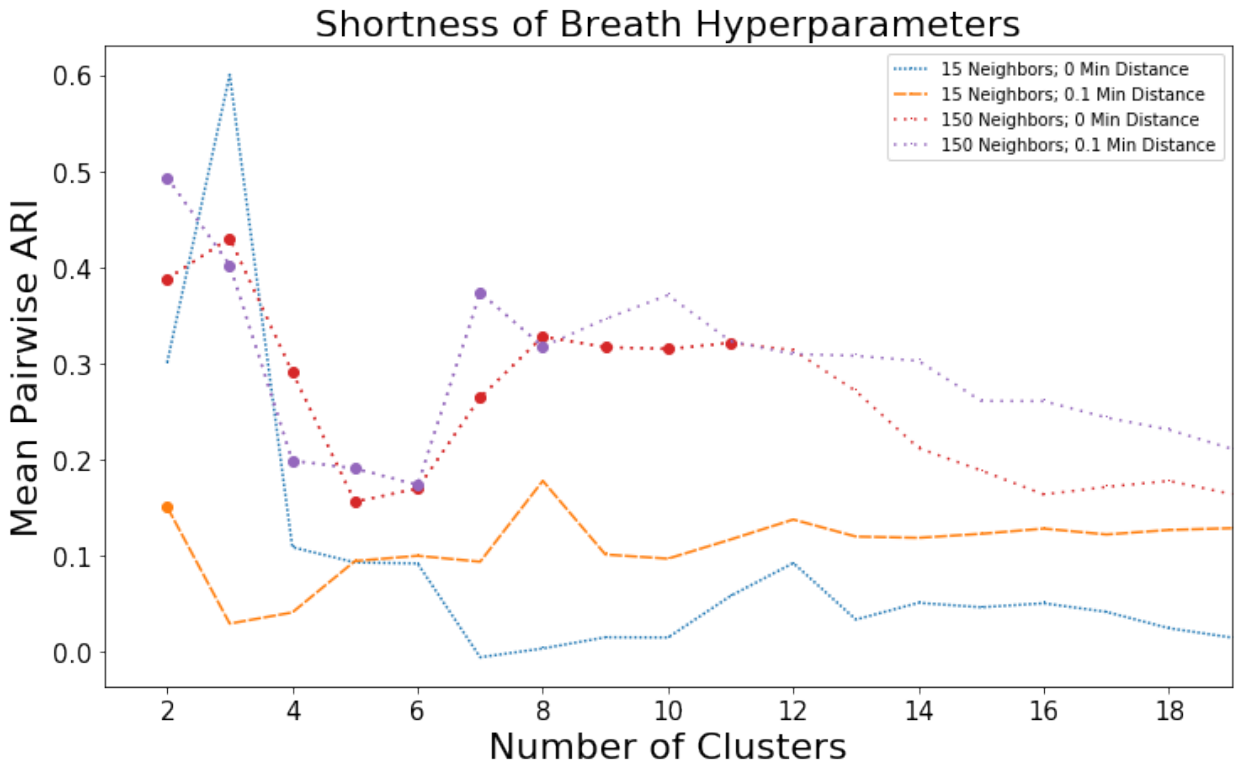
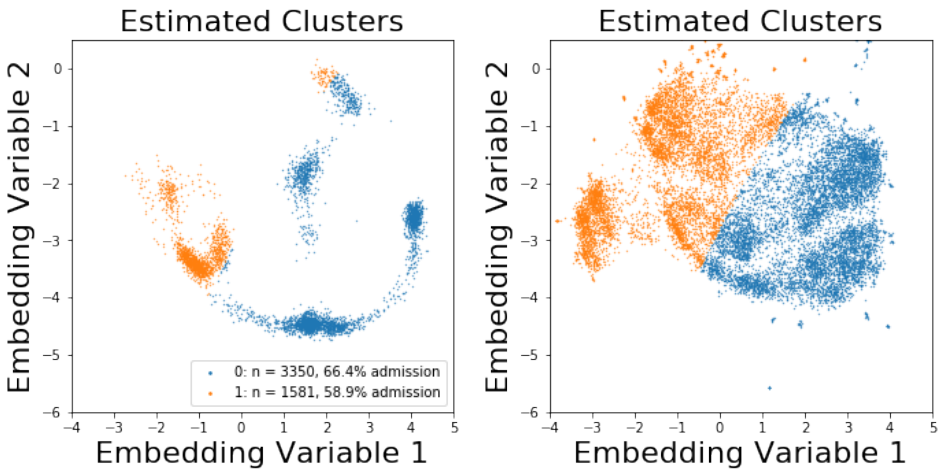


Figure 3.5: Representative plot of hyperparameters. Here, four sets of hyperparameters and the resulting mean pairwise ARI are shown. All ARIs are plotted. The markers indicate hyperparameters where the mean number of clusters produced was no more than 0.5 less than the number of clusters with which the GMM was trained. For instance, the blue peak at 3 clusters was built with a model where although 3 clusters were indicated, the mean number of clusters used was 1.2, indicating that four folds categorized all test data as belonging to a single cluster, while one fold categorized all test data as belonging to two clusters. Therefore, the ARI is elevated through a trivial clustering of only one cluster present.

more likely to be admitted (66.5%, 95% CI 66.3-66.6) while the smaller cluster was less likely to be admitted (57.3%, 95% CI 55.3-59.2). A representative embedding of these clusters are shown in Figure 3.6.

The larger cluster here was much more likely to have been previously visited this hospital system. These patients were more likely to have arrived via ambulance, suggesting higher acuity. These patients were slightly more likely to have urinalysis results positive for blood and leukocytes, and were more likely to have risk factors such as chronic obstructive pulmonary disease or congestive heart failure. On average, these patients had been admitted in



(a) (b)

Figure 3.6: UMAP embeddings of patients with Shortness of Breath. Figure 3.6a shows the training data, while Figure 3.6b shows the application of the model to the test data.

Table 3.2: Best mean pairwise ARIs and associated hyperparameters per chief complaint.

	Best ARI	# of Neighbors	Minimum Distance	# of Clusters
Abdominal Pain	0.353	150	0.0	2
Chest Pain	0.589	15	0.0	6
Shortness of Breath	0.493	150	0.1	2
Back Pain	0.385	150	0.25	4
Falls	0.741	150	0.0	2

this system 2.4 times each previously, with a median of 1 previous admission.

The smaller cluster here was much more likely to have been a first time patient to this ED system. These patients were more likely to have commercial insurance, and to be employed full time. These patients were also more likely to present with additional chief complaints, such as cough or palpitations.

3.4.2.2 Abdominal Pain

The best clustering of abdominal pain was found with two clusters that contain 89.2% (95% CI 83.7-94.6) and 10.8% (95% CI 5.4-16.3) of patients (Figure 3.7). The larger cluster was generally more likely to be admitted (37.6%, 95% CI 36.7-38.5), while the smaller cluster

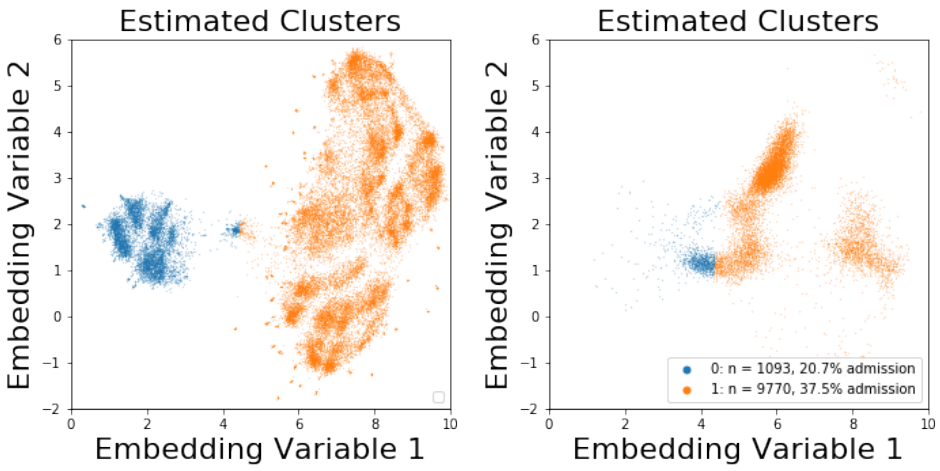
Table 3.3: Selected Patient Characteristics of Abdominal Pain Clusters

	Cluster 0		Cluster 1	
Count	1093	10.1%	9770	89.9%
Disposition- Admit (%)	226	20.7%	3667	37.5%
Gender— Male (%)	355	32.5%	3488	35.7%
Insurance Status— Medicare (%)	539	49.3%	3656	37.4%
Insurance Status— Medicaid (%)	61	5.6%	1753	17.9%
Language— English (%)	719	65.8%	8981	91.9%
Arrival via Ambulance (%)	103	9.4%	2454	25.1%
Mean Age in Years (range)	37.8	18-89	46.8	18-103
Mean Triage Heart Rate (range)	83.7	43-148	86.0	40-185
Mean Triage Systolic BP(range)	129.3	82-232	132.5	66-243
Mean Triage Diastolic BP (range)	79.6	40-128	80.6	28-163
Mean Triage Respiratory Rate (range)	17.5	13-28	17.6	10-40
Mean Triage O_2 Saturation (range)	97.7	91-99	97.5	74-99
Mean Triage Temperature in °F (range)	98.1	96-104.4	98.1	94.6-103.5
Mean Prior Admissions (range)	<0.01	0-3	1.3	0-47

was generally less likely to be admitted (24.2%, 95% CI 19.3-29.0). Selected patient characteristics of these clusters are shown and compared in Table 3.3. The peak mean pairwise ARI found was 0.35, with 2 clusters, 150 neighbors, and no minimum distance between points.

In the smaller cluster, patients tended to have either been previously evaluated in this emergency department system and discharged, or had never been previously evaluated in this system. These patients had lower blood pressure and less hypertension than the other cluster. These patients were less likely to have very low O_2 saturation, and were less likely to be English speakers. Relatively few of these patients arrived at the hospital via ambulance, indicating lower acuity. These patients were less likely to be admitted to the hospital.

The larger cluster tended to have an older population, and more had esophageal disease or hyperlipidemia. Nearly a quarter of these patients arrived at the hospital via an ambulance, indicating higher acuity. Most of these patients had been admitted from this ED system before, and over a third were admitted in this visit.



(a) (b)

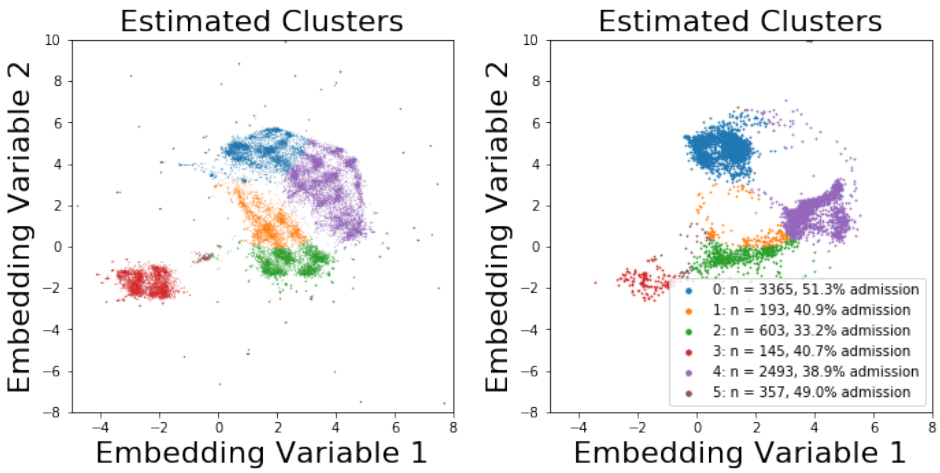
Figure 3.7: UMAP embeddings of patients with Abdominal Pain. Figure 3.7a shows the training data, while Figure 3.7b shows the application of the model to the test data.

3.4.2.3 Chest Pain

The best clustering of chest pain was found with six clusters that contain 54.5% (95% CI 46.4-62.6), 24.3% (95% CI 16.8-31.7), 12.5% (95% CI 7.8-17.2), 5.8% (95% CI 3.8-7.8), and 0.5% (95% CI 0.0-1.0) of patients. Most clusters had similar rates of admission, with overlapping 95% confidence intervals. The peak mean pairwise ARI found was 0.59, with 6 clusters, 15 neighbors, and no minimum distance between points. A representative visualization of these clusters is shown in Figure 3.8.

In the largest cluster (47.0% of the population), the patients were slightly more likely to have been previously admitted. This population was more likely to have hypertension or mood disorders, or to have arrived via ambulance, indicating higher clinical acuity. These patients were slightly more likely to have positive urine protein and leukocytes. These patients had previously been admitted to this hospital an average of 2.8 times each, with all patients in the top quartile having been admitted 3 or more times each. 51.4% of patients in this cluster were admitted, as opposed to 44.8% overall for patients with this chief complaint.

The second largest cluster (34.8% of the population) was notable for feature patients



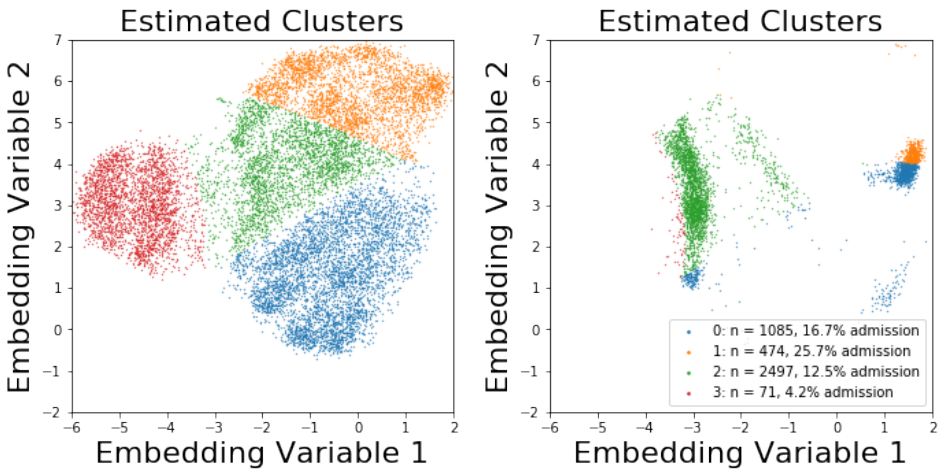
(a) (b)

Figure 3.8: UMAP embeddings of patients with Chest Pain. Figure 3.8a shows the training data, while Figure 3.8b shows the application of the model to the test data.

who were more likely to have never been seen within this hospital system, and very few had ever been admitted. These patients were more likely white or Caucasian, and were less likely to be on Medicaid (28%) or Medicare (14%). These patients had less hypertension and diagnosed mood disorders than the rest of the population. 38.9% of these patients were admitted, as opposed to 44.8% overall. Patients in this cluster were an average of 6 years younger than the patients in the larger cluster.

The third largest cluster here (8.4% of the population) was clustered largely on arrival mechanism, and were more likely to have arrived via car or as a walk-in, and were significantly less likely to have arrived via ambulance, indicating a generally lower acuity. These patients were less likely to have risk factors including male gender, hyperlipedemia, or diagnosed CAD. These patients were relatively unlikely to be admitted, with only a 33.17% admission rate.

One cluster contained 5.0% of the population, and this cluster was more likely to include diabetic patients who had previously been seen in this system. Patients in this cluster were slightly more likely to have alcohol-related disorders or other substance-related disorders. These patients also had slightly higher rates of mood or anxiety disorders, and were more



(a) (b)

Figure 3.9: UMAP embeddings of patients with Back Pain. Figure 3.9a shows the training data, while Figure 3.9b shows the application of the model to the test data.

likely to have been previously discharged. Patients in this cluster were admitted at a rate of 49.0%.

The remaining two clusters each contained less than 3% of the total population. Their primary differences relative to the rest of the population were racial demographics.

3.4.2.4 Back Pain

The best clustering of back pain was found with four clusters that contain 63.3% (95% CI 60.1-66.6), 22.5% (95% CI 20.0-24.9), 12.6% (95% CI 10.9-14.3), and 1.6% (95% CI 1.3-1.9) of patients. Clusters had similar rates of admission, with overlapping 95% confidence intervals. The peak mean pairwise ARI found was 0.39, with 4 clusters, 150 neighbors, and 0.25 minimum distance between points. A representative visualization of these clusters are shown in Figure 3.9.

The largest cluster, consisting of 60.5% of the testing data, primarily features patients that were not previously seen in this ED system. These patients were more likely to be employed full time, and predominantly did not arrive at the hospital via ambulance. These patients were slightly more likely to have male gender, and were less likely to have asthma.

12.5% of these patients were admitted, as opposed to 15.0% of patients with this chief complaint.

The second largest cluster, consisting of 26.3% of the testing data, features patients that were more often previously seen in this ED system. These patients were more often insured via Medicaid, and were more likely to have risk factors such as hypertension. These patients were more likely to have urinalysis positive for leukocytes, blood, or protein. These patients were more likely to have female gender. 16.7% of these patients were admitted, as opposed to 15.0% of the patients overall with this chief complaint.

The next cluster consists of 11.5% of the testing data. Patients in this cluster were much more likely to have been previously admitted, and were more likely to have arrived via ambulance. These patients were generally more hypertensive than the rest of this population, and were more likely to have been diagnosed with mood or anxiety disorders. These patients were insured with Medicare more often than the remainder of the patients in this population.

The smallest cluster consists of 1.7% of the testing data. These patients were predominantly clustered by their racial demographics, and were generally more likely to have arrived at the ED as a walk-in patient. These patients were slightly more likely to have female gender and to co-present with a chief complaint of having suffered a fall. Only 4.2% of these patients were admitted, as opposed to 15.0% overall.

3.4.2.5 Falls

The best clustering of falls was found with two clusters that contain 52.7% (95% CI 50.6-54.8) and 47.3% (95% CI 45.2- 49.4) of patients. Clusters had very similar rates of admission. The peak mean pairwise ARI found was 0.74, with 2 clusters, 150 neighbors, and no minimum distance between points. Three of five representative visualizations of these clusters are shown in Figure 3.10.

The larger cluster consists of 52.0% of the testing data. In this cluster, patients were more likely to have been previously seen in this ED system. These patients were more likely to have arrived via ambulance, and were more likely to have positive urinalysis findings for

protein, leukocytes, or blood. These patients were more likely to be on Medicare. These patients were slightly more likely to have lower blood oxygen saturation. 31.6% of these patients were admitted, as opposed to 28.7% overall.

The smaller cluster consists of 48.0% of the testing data. In this cluster, patients were more likely to have never been seen in this ED system. These patients were more likely to be employed full or part time, and to have commercial insurance. These patients were admitted at a rate of 25.5%.

3.5 Discussion

In this work, we sought to further efforts towards the summary visualization of high-dimensionality EHR data. We focused on a previously published emergency department dataset with the goal of revealing data-driven phenotypes at time of hospital presentation. We first benchmarked our approach using a synthetic dataset and subsequently moved our analysis towards investigating the five most common emergency department chief complaints.

The synthetic dataset here was generated to have similar size and variability to the clinical dataset. As was anticipated, two-dimensional PCA was unable to capture the variability contained within the multiple informative dimensions. We observed very little useable structure within the PCA embedding of the synthetic data (Figure 3.2b). While the ARI is higher than would be expected by chance for 2 clusters, it is unremarkable at the true value of 10 clusters. Furthermore, while the pairwise ARI for the PCA model maintains a value around 0.2, the ARI to the ground truth cluster labels remains at nearly 0. This suggests that even though some patterns are found within the PCA embedding of the data, these patterns do not correlate with the underlying structure of the data.

In contrast, UMAP clearly identifies elements of the high-dimensional structure of this same dataset. Even without clustering the UMAP-embedded data, a strong relationship can be seen between most points of a given cluster (Figure 3.3b). Though visually trivial, mapping the mean pairwise ARI of GMMs trained with different numbers of clusters in Figure 3.4 shows that the peak ARI correlates with this correct number of clusters. Furthermore,

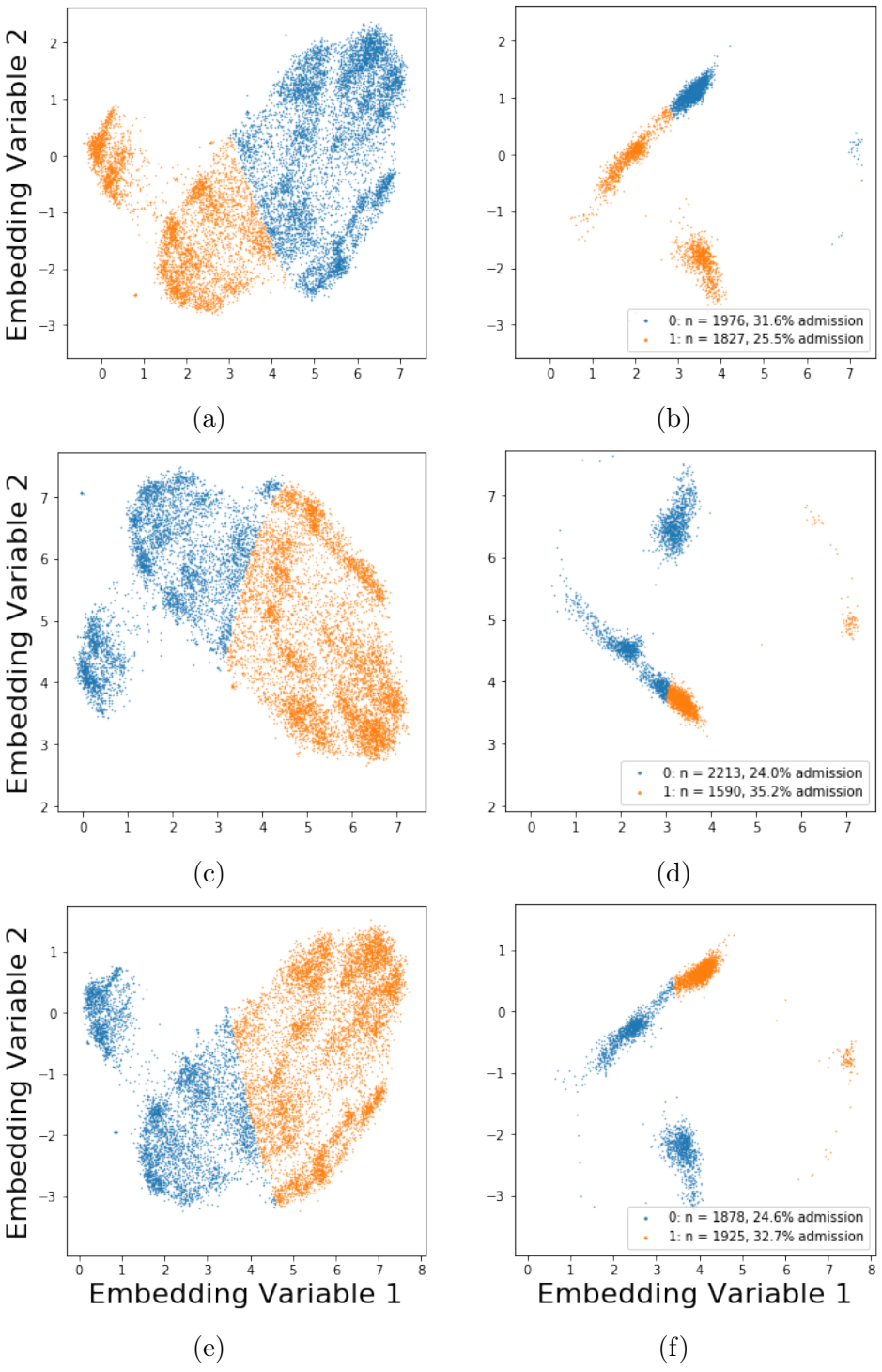


Figure 3.10: Three different folds of UMAP embeddings of patients who suffered Falls. Figures 3.10a, 3.10c, and 3.10e show the training data, while Figures 3.10b, 3.10d, and 3.10f show the application of the model to the test data. In each of these, it can be seen that the embedding follows a similar overall pattern even among different folds. Two additional folds are not shown, but exhibit the same global shape and cluster characteristics.

the pairwise ARI and ARI to ground truth are both high values. This suggests that when using UMAP, the pairwise ARI is useful as a proxy for the ground truth ARI. When adapting this model to datasets without known ground truth, the pairwise ARI should function to ensure that valid cluster patterns are being observed. These data motivate our efforts to investigate a real-world clinical dataset.

3.5.1 Clinical Interpretation

Here, we consider more broadly the clinical implications of the discovered phenotype clusters. In the shortness of breath clusters described above, the characteristics of the two clusters paint two very different clinical images. In the larger cluster, the characteristics present a description of patients who are generally sick: patients who are high utilizers of healthcare with medical comorbitities such as congestive heart failure or chronic obstructive pulmonary disease. On the other hand, the smaller cluster represents first-time utilizers of the health care system who are generally healthier. They are more likely to present with other chief complaints, such as cough. This suggests that this cluster represents more acute causes of shortness of breath, while the other cluster represents more chronic causes of shortness of breath.

The two clusters discovered among patients with abdominal pain generally separate the population into a younger, healthier group and an older, sicker group. The younger group was less likely to arrive via ambulance, and was much less likely to have previously been admitted to this hospital. The older group was much more likely to have underlying health issues. Abdominal pain clusters are dominated by demographics and arrival mechanism.

While the abdominal pain clusters appeared more trivial in nature (old/sick, young/healthy), our workflow suggested optimized chest pain clustering would be produced with 15 neighbors used in the UMAP approximation rather than 150 neighbors. This results in the discovered clusters having a greater reliance on local structure rather than more distant global structure. The clusters here vary more in size, which allows for more precise phenotypes to be observed. For instance, one of the clusters here featured patients with low risk factors for heart disease

who transported themselves to the ED. This group was less likely to need admission to the hospital.

The clusters discovered in patients presenting with back pain fit several different patient populations. The largest cluster presents the image of patients how experience first time lower back pain, but who have good socioeconomic factors as a positive prognostic indicator. These patients were relatively less likely to be admitted to the hospital. The second largest cluster presents a clinical picture of patients who have medical comorbidities and non-musculoskeletal reasons for their back pain. For instance, these patients often had leukocyturia, potentially indicating pyelonephritis (kidney infection) as the cause of their pain, or hematuria, potentially indicating nephrolithiasis (kidney stone) as the cause of their pain. These patients were slightly more likely to be admitted to the hospital than were other patients with this chief complaint.

It is interesting to note the dissimilarity in the general shape of the training data and testing data in the patients with back pain (Figure 3.9). Although the training data exhibited four clusters of relatively similar density, the test data shows unbalanced clusters with much more variable density. This is an aspect of this model where the visual nature of the embeddings can be leveraged. While in the training data, the model easily differentiates between the blue and orange clusters (Figure 3.9a), the test data shows that there are a number of patients that are very similar to each of these clusters, and that these new patients are embedded with a density unlike the training densities.

The clusters discovered in patients presenting after a fall were the clusters with the highest mean pairwise ARI of any clusters within the dataset. As shown in Figure 3.10, the resulting embeddings are very similar on subsequent training folds. These patients were split nearly in half, but split in almost the exact same way each time. In these two clusters, the cluster with the higher admission rate is also the cluster with increased comorbidities. These patients would likely be at increased risk of fall due to their increased age and medical comorbidities. Similarly, falls suffered could have the potential to cause greater harm to the

patient.

In sum, across the five chief complaints studied, we observed a range of phenotypes and underlying demographic and physiologic drivers. In cases of more trivial two-group separations, as observed with abdominal pain, shortness of breath, and falls, our approach reveals elements of patient comorbidities and presentation acuity. Of note, the ARIs for these complaints differ significantly, with falls (0.741) being the most stably captured and abdominal pain (0.353), the least stable. We hypothesize that this metric may be capturing the underlying heterogeneity inherent to the complaint. For example, falls may be mechanical in nature (e.g., slipping) or result from a cardiovascular cause (e.g., syncope, arrythmia) or change in mental status (e.g., transient ischemic attack). The etiologies of abdominal pain are myriad, and more difficult to bin [68] - a minimal differential diagnosis of pain etiology includes a half-dozen organ systems. For this reason, abdominal pain is typically clinically assessed by abdominal quadrant. For complaints of chest and back pain, we were able to optimize for larger number of clusters. Further research is required to reveal interesting elements evolving from the pairwise comparisons of these phenotypes.

3.5.2 Limitations and Future Work

In most clinical data clusters, the visualization of the training data and the testing data appear very different. For instance, in Figure 3.9a, there are well-demarcated clusters with similar appearing sizes and densities. However, in Figure 3.9b, the clusters are of different sizes and appear grossly dissimilar to the clusters in Figure 3.9a. The cluster on the left has nearly disappeared, while the bulk of the top and bottom clusters have become much closer to each other than in the training data. Future work will continue to explore model robustness across variations including the use of alternative distance metrics.

Additionally, other potential clusterings of patients are likely present within these datasets. For instance, in Figure 3.5, it appears that another clustering might be present with 3 clusters or with 7 clusters. Future work should evaluate these different potential clusterings to see if additional clinical characteristics can be achieved from these clusters, and should

evaluate if greater number of clusters with slightly lower stability can be clinically useful.

We attempted to apply this analytic pipeline to the full clinical dataset, without separating into populations by chief complaint. We observed that it was unable to consistently find any given clusters and hypothesize that this is likely due to the large number of potential clusters to which any given patient could belong. Patients with the same core pathology could present with different chief complaints. For example, a myocardial infarction (heart attack) could present as either chest pain, or as syncope. A more generally applicable model should incorporate a method to include multiple possible chief complaints so that similar pathologies are not treated as separate entities.

An interesting extension to this work would be to utilize the clusters here to assess patients for risk of adverse events. The clustering of patients presenting after falls exhibited characteristics that suggest that patients were grouped by their risk of falls. The patients at lower overall risk were more likely to be discharged, while the patients at higher overall risk were more likely to be admitted. This suggests that this technique could be applied to isolate risk factors for adverse outcomes. This information could then help identify similar patients based on cluster membership, and allow health care providers to rapidly and effectively apply interventions to improve patient health and outcomes.

3.6 Conclusion

This paper presents a technique for the two-dimensional visualization of complex emergency department patient data. We show that UMAP and GMMs enable robust cluster identification using a synthetic dataset and then apply these tools to a real-world clinical dataset. We explore the patient phenotypes emerging from varied patient chief complaints, revealing pertinent clinical characteristics of these populations. Among patients with abdominal pain, chest pain, shortness of breath, back pain, and falls, the populations are reliably divided into 2-6 clusters. These clusters group patients based on characteristics such as demographics and triage variables, allowing for clear clinical pictures of the type of patients involved to be seen. We anticipate future medical scenarios where deployment of

this visualization pipeline will enable both rapid, real-time patient triage and retrospective cohort discovery from electronic health records.

4. DYNAMICALLY EXTRACTING PROBLEM LISTS FROM CLINICAL NOTES*

In this chapter, we once again aim to utilize machine learning to rapidly extract useful clinical information. Here, however, we turn to focusing on the intensive care setting. Rather than a tool to rapidly visualize and characterize a new patient with a given chief complaint, here we instead extract the most pertinent features and problems for a patient based on free text notes. This allows for a rapid distillation of the key parts of a longer narrative, aiding clinicians in rapidly understanding the patient's state.

4.1 Introduction

Problem lists are an important component of the electronic health record (EHR) that are intended to present a clear and comprehensive overview of a patient's medical problems. These lists document illnesses, injuries, and other details that may be relevant for providing patient care and are intended to allow clinicians to quickly gain an understanding of the pertinent details necessary to make informed medical decisions and provide patients with personalized care [69, 70]. Despite their potential utility, there are shortcomings with problem lists in practice. One such shortcoming is that problem lists have been shown to suffer from a great deal of clutter [71]. Irrelevant or resolved conditions accumulate over time, leading to a problem list that is overwhelming and difficult for a clinician to quickly understand. This directly impairs the ability of a problem list to serve its original purpose of providing a clear and concise overview of a patient's medical condition.

A challenge that comes with attempting to reduce clutter is that many conditions on the list may be relevant in certain situations, but contribute to clutter in others. For example, if a patient ends up in the intensive care unit (ICU), a care unit for patients with serious medical conditions, then the attending physician likely does not care about the patient's

*This chapter is reprinted with permission from "Dynamically Extracting Outcome-Specific Problem Lists from Clinical Notes with Guided Multi-Headed Attention" by Lovelace, J., Hurley, N. C., Haimovich, A. D., & Mortazavi, B. J., 2020. Machine Learning for Healthcare Conference (pp. 245-270). PMLR. Copyright 2020 by J. Lovelace, N.C. Hurley, A.D. Haimovich & B.J. Mortazavi.

history of joint pain. That information, however, would be important for a primary care physician to follow up on during future visits. In this case, the inclusion of chronic joint pain clutters the list for the attending physician in the ICU, but removing it from the list could decrease the quality of care that the patient receives from his/her primary care physician.

In this work, we address this problem by developing a novel end-to-end framework to extract problems from the textual narrative and then utilize the extracted problems to predict the likelihood of an outcome of interest. Although our framework is generalizeable to any clinical outcome of interest, we focus on ICU readmission and patient mortality in this work to demonstrate its utility. We extract dynamic problem lists by utilizing problem extraction as an intermediate learning objective to develop an interpretable patient representation that is then used to predict the likelihood of the target outcome. By identifying the extracted problems important for the final prediction, we can produce a problem list tailored to a specific outcome of interest.

We demonstrate that this framework is both more interpretable and more performant than the current state-of-the-art work using clinical notes for the prediction of clinical outcomes [72, 73, 74]. Utilizing the intermediate problem list for the final outcome prediction allows clinicians to gain a clearer understanding of the model's reasoning than prior work that only highlighted important sections of the narrative. This is because our framework directly identifies clinically meaningful problems while the prior work requires a great deal of inference and guesswork on the part of the clinician to interpret what clinical signal is being represented by the highlighted text.

For example, prior work predicting the onset of heart disease found that the word "daughter" was predictive of that outcome. The authors stated that the word usually arose in the context of the patient being brought in by their daughter which likely signaled poor health and advanced age [74]. While this makes sense after reviewing a large number of notes, this connection is not immediately obvious and a clinician would not have the time to conduct the necessary investigation to identify such a connection. By instead directly extracting pre-

defined clinical conditions and procedures and using those for the final prediction, we reduce the need for such inference on the part of the physician.

The primary contributions of this work are:

- A novel end-to-end framework for the extraction of clinical problems and the prediction of clinical outcomes that is both more interpretable and performant than models used in prior work.
- An expert evaluation that demonstrates that our problem extraction model exhibits robustness to labeling errors contained in a real world clinical dataset.
- Dynamic problem lists that report the quantitative importance of each extracted problem to an outcome of interest, providing clinicians with a concise overview of a patient's medical state and a clear understanding of the factors responsible for the model's prediction.
- A qualitative expert user study that demonstrates that our dynamic problem lists offer statistically significant improvements over a strong baseline as a clinical decision support tool.

Generalizable Insights about Machine Learning in the Context of Healthcare

A significant body of past work develops predictive models that can not be used in clinically useful settings due to their reliance on billing codes assigned after a patient leaves the hospital [75, 76, 77, 78, 79, 80]. While there may be value in the technical innovations made by such work, research that acknowledges and addresses the constraints of the domain is essential to develop methods that can actually be implemented in practice. We demonstrate that recent methods for automated ICD code assignment are sufficiently performant to extract billing information in real-time for downstream modeling tasks. Although we focus on extracting problem lists for clinical decision support in this work, this finding has broader ramifications for the field. It both enables the real-time implementation of previously impracticable work and paves the way for future work to develop clinically feasible models that utilize dynamically extracted diagnosis and procedure information from clinical text.

4.2 Related Work

There has been a large body of prior work utilizing natural language processing (NLP) techniques to extract information from clinical narratives. Blecker et al. [81] demonstrated that unstructured clinical notes could be used to effectively identify patients with heart failure in real time. Their methods that involved data from clinical notes outperformed those using only structured data, demonstrating the importance of effectively utilizing the rich source of information contained within the clinical narrative.

Prior work has found success predicting ICD code assignment using clinical notes within MIMIC-III and has found that deep learning techniques outperform traditional methods [82, 83, 84, 85]. Mullenbach et al. [83] augmented a convolutional model with a per-label attention mechanism and found that it led to both improved performance and greater interpretability as measured by a qualitative, expert evaluation. Sadoughi et al. [84] later improved upon their model by utilizing multiple convolutions of different widths and then max-pooling across the channels before the attention mechanism.

There has also been work done demonstrating that machine learning models can effectively leverage the unstructured clinical narrative for the prediction of clinical outcomes [77, 74, 72]. Jain et al. [72] augmented long short-term memory networks (LSTMs) with an attention mechanism and applied it to predict clinical outcomes such as mortality and ICU readmission. However, when defining readmission, they treated both ICU readmissions and deaths as positive examples. The clinical work by Krumholz et al. [86] has demonstrated that these are orthogonal outcomes, and thus modeling them jointly as a single outcome does not make sense from a clinical perspective. By treating them as separate outcomes in this work, we are able to independently explore the risk factors for these two distinct outcomes.

Jain et al. [72] also raised some questions about the interpretability of attention in their work with clinical notes, repeating the experiments introduced by Jain and Wallace [87] to evaluate the explanatory capabilities of attention. However, Wiegreffe and Pinter [88] explored some of the problems with their underlying assumptions and experimental setup

and demonstrated that their experiment failed to fully explore their premise, and thus failed to support their claim.

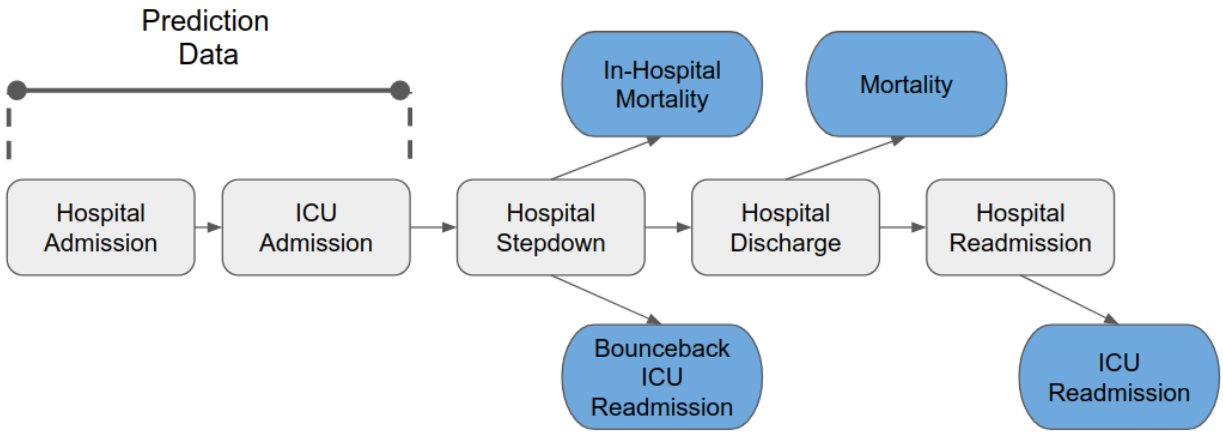


Figure 4.1: Outcomes explored in this work

4.3 Data and cohort

This work is conducted using the free text notes stored in the publicly available MIMIC-III database [2]. The database contains de-identified clinical data for over forty thousand patients who stayed in the critical care units of the Beth Israel Deaconess Medical Center. This information was collected as part of routine clinical care and, as such, is representative of the information that would be available to clinicians in real-time. This makes the dataset well-suited for developing clinical models.

To develop our cohort, we first filter out minors because children have different root causes for adverse medical outcomes than the general populace. We also remove patients who died while in the ICU and filter out ICU stays that are missing information regarding the time of admission or discharge. We then extract all ICU stays where the patient had at least three notes on record before the time of ICU discharge to develop a cohort with a meaningful textual history. This leaves us with $33,311$ unique patients and $45,260$ ICU stays.

For ICU readmission we extract labels for two types of readmissions, bounceback and 30 day readmisssion. Bounceback readmissions occur when a patient is discharged from the ICU and then readmitted to the ICU before being discharged from the hospital. For 30 day readmissions, we simply look at any readmission to the ICU within the 30 days following ICU discharge. For mortality, we also look at two different outcomes, in-hospital mortality and 30-day mortality. Because we use all data available at the time of ICU discharge, in-hospital mortality is constrained to mortality that occurs after ICU discharge but prior to hospital discharge. All the outcomes that we explored in this work are laid out in Figure 4.1. This provides us with a cohort with $3,413$ (7.5%) bounceback readmissions, $5,674$ (12.5%) 30-day readmissions, $3,761$ (8.3%) deaths within 30 days, and $1,898$ (4.2%) in-hospital deaths. For our experiments, we then split our cohort into training, validation, and testing splits following an 80/10/10 split and use 5-fold cross validation. We divide our cohort based on the patient rather than the ICU stay to avoid data leakage when one patient has multiple ICU stays.

We extract all clinical notes associated with a patient's hospital stay up until the time of their discharge from the ICU. The text is then preprocessed by lowercasing the text, normalizing punctuation, and replacing numerical characters and de-identified information with generic tokens. All of the notes for each patient are then concatenated and treated as a continuous sequence of text which is used as the input to all of our models. We truncate or pad all clinical narratives to 8000 tokens. This captures the entire clinical narrative for over 75% of patients and we found that extending the maximum sequence length beyond that point did not lead to any further improvements in performance.

4.4 Methods

In this work, we develop an end-to-end framework to jointly extract problems from the clinical narrative and then use those problems to predict a target outcome of interest. An overview of our framework can be seen in Figure 4.2. We embed the clinical notes using learned word embeddings and then apply a convolutional attention model with a guided

multi-headed attention mechanism to extract problems from the narrative. We then utilize the intermediate problem predictions to predict the target outcome. This differs from standard deep learning models because the features used for our final prediction are clearly mapped to clinically meaningful problems rather than opaque learned features. We also describe the training procedure that we develop to ensure that our problem extraction model maintains a high level of performance, something that is essential for the intermediate features to maintain their clinical significance.

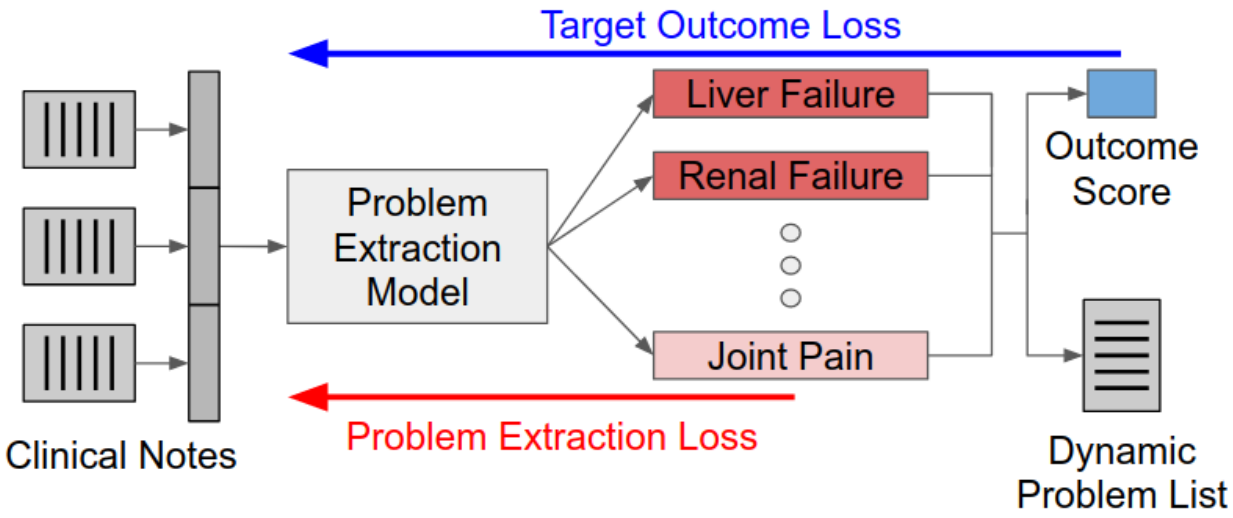


Figure 4.2: Overview of our proposed framework

4.4.1 Embedding techniques

We utilize all notes in the MIMIC-III database associated with subjects who are not in our testing set to train embeddings using the Word2Vec method [89]. This allows for training on a greater selection of notes than if training had been limited to the training set. This training is done using the continuous bag-of-words implementation and it generates embeddings for all words that appear in at least 5 notes in our corpus. We replace out-of-vocabulary words with a randomly initialized UNK token to represent unknown words. Both

100 and 300 dimensional word embeddings were explored and early testing showed that 100 dimensional word embeddings led to better performance.

4.4.2 Target Problems

We experiment with multiple different representations for the intermediate problems in this work. The first representation we explore are the ICD9 codes assigned to all hospital stays in our dataset. These codes are used for billing purposes and represent diagnostic and procedure information for each patient. Although prior work has found that these codes are predictive of adverse outcomes [77, 78, 80], these codes are assigned after a patient has been discharged from the hospital and, as such, directly using these codes as features in a predictive model limits the clinical utility of such a model. By instead learning to dynamically assign these codes within our framework, we can use these codes to predict the outcomes we explore using only the information available at the time of prediction.

However, the large ICD9 label space will likely hinder our frameworks's ability to effectively extract and utilize the codes. To address this, we leverage the heirarchical nature of the ICD9 taxonomy. Full ICD9 codes are represented by character strings up to 6 characters in length where each subsequent character represents a finer grained distinction. We experiment with rolled up ICD9 codes which consist of only the first three characters of each ICD9 code to address the problem of the large label space. The rolled up codes still represent clinically meaningful procedures and conditions while substantially reducing the number of labels.

We also explore using phecodes which were developed to conduct phenome-wide association studies (PheWAS) in EHRs [90]. Prior work demonstrated that phecodes better represent clinically meaningful phenotypes than ICD9 codes [91]. Because of this, phecodes may lead to a more clinically meaningful and predictive intermediate representations than ICD9 diagnosis codes. A mapping from ICD9 codes to phecodes already exists and can be used to extract phecodes from our dataset. Similar to ICD9 codes, we explore both full and rolled up phecodes. For every problem representation in this work, we only use codes that

occur at least 50 times in our training set to reduce label sparsity. After this filtering, there are an average of 1047.4 full ICD diagnosis codes, 331.8 full ICD procedure codes, 695.6 full phecodes codes, 419.6 rolled ICD diagnosis codes, 203.4 rolled ICD procedure codes, and 356.0 rolled phecodes across our 5 folds.

4.4.3 Problem extraction model

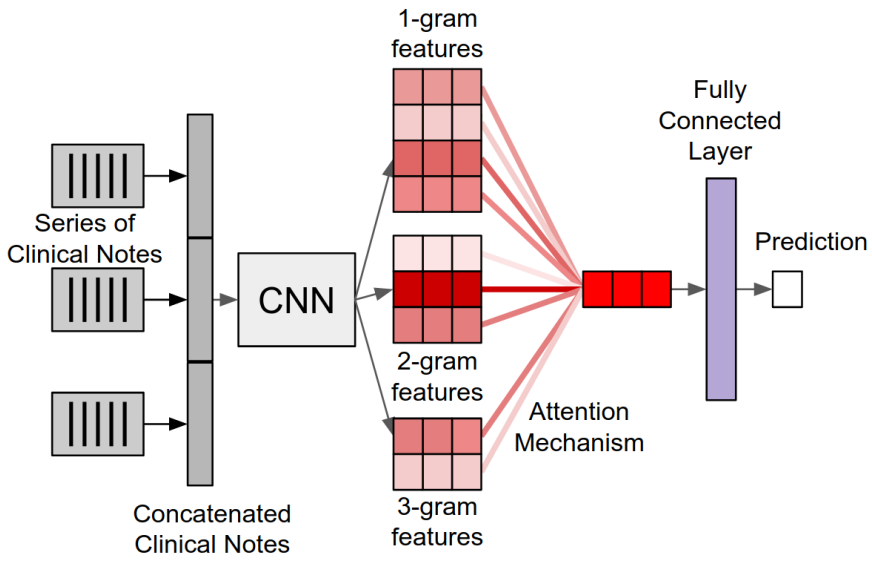


Figure 4.3: Illustration of our problem extraction model with a single attention mechanism shown.

The convolutional attention architecture used in this work is similar to that developed by Mullenbach et al. [83] and Sadoughi et al. [84] for automatic ICD code assignment. The model can be described as follows. We represent the clinical narrative as a sequence of d_e-dimensional dense word embeddings. Those word embeddings are then concatenated to create the matrix $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; ...; \mathbf{x}_N]$ where N is the length of the clinical narrative and $x_n \in \mathbb{R}^{d_e}$ is the word embedding for the n^{th} word in the narrative. We then apply a convolutional neural network (CNN) to the matrix $\mathbf{X}$.

In this work, we use three convolutional filters of width 1, 2, and 3 with output dimen-

sionality d_f. These filters convolve over the textual input with a stride of 1, applying the learned filters to every 1-gram, 2-gram. and 3-gram in the input. In this work, we augment the CNN with a multi-headed attention mechanism where each head is associated with a problem [92]. Unlike the work of Mullenbach et al. [83] and Sadoughi et al. [84], we apply our attention mechanisms over multiple convolutional filters of different lengths. This allows our model to consider variable spans of text while still maintaining the straightforward interpretability of the model introduced by Mullenbach et al. [83].

To apply the attention mechanisms, we learn a query vector, $\mathbf{q}_\ell \in \mathbb{R}^{d_f}$, for each problem ℓ that will be used to calculate the importance of the feature maps across all filters for that problem. We calculate the importance using the dot product of each feature map with the query vector. We let $\mathbf{H} \in \mathbb{R}^{d_f \times (3N)}$ be the concatenated output of our CNN and can then calculate the attention distribution over all of the feature maps simultaneously using the matrix vector product of our final feature map and the query vector as $\boldsymbol{\alpha}_\ell = softmax(\frac{\mathbf{H^T q_\ell}}{\sqrt{d_f}})$ where d_f is used as a scaling factor and $\boldsymbol{\alpha_\ell} \in \mathbb{R}^{\mathbf{3N}}$ contains the score for every position across all the filters. The softmax operation is used so that the score distribution is normalized. We calculate the final representation used for classification for problem ℓ by taking a weighted average of all of the outputs based on their calculated weights given by $\mathbf{v}_\ell = \sum_{i=1}^{3N} \boldsymbol{\alpha_{\ell,i}} \mathbf{h_i}$ where $\mathbf{h}_i$ is the i^{th} feature vector in $\mathbf{H}$ and $\mathbf{v}_\ell$ is the final representation used for predicting the presence of problem ℓ.

Given the representation $\mathbf{v}_\ell$, we calculate the final prediction as $\hat{y}_\ell = \sigma(\mathbf{w^T}_\ell \mathbf{v}_\ell + b_\ell)$ where $\mathbf{w}_\ell$ is a vector of learned weights, $\mathbf{b}_\ell$ is the bias term, and σ is the sigmoid function. We train our problem extraction model by minimizing the binary cross-entropy loss function given by $\mathcal{L}_p = -\sum_{\ell=1}^{L} y_\ell \log(\hat{y}_\ell) + (1 - y_\ell)\log(1 - \hat{y}_\ell)$ where y_ℓ is the ground truth label and $\hat{y}_\ell$ is our model's prediction for problem ℓ.

4.4.4 Outcome classification

In our proposed framework, the feature vector used for the outcome prediction is $\mathbf{s} = [s_0; s_1; ...; s_{L-1}; s_L]$ where $\mathbf{s} \in \mathbb{R}^L$ and s_ℓ is the scalar score for problem ℓ defined by $s_\ell =$

$\mathbf{w}_\ell^T \mathbf{v}_\ell + b_\ell$. We calculate our final prediction using this vector similarly to our intermediate problem prediction as $\hat{y_0} = \sigma(\mathbf{w^T}_o\mathbf{s} + b_o)$. Using the score for each outcome as the features for the final prediction allows for the straightforward interpretation of each feature. This differs from the standard deep learning models used in prior works where the final feature vector used for the prediction is composed of learned features that are not interpretable. We utilize this improvement to explain our model's decision making process and to develop dynamic problem lists.

To optimize the classification objective for our target outcome, we also minimize the binary cross-entropy loss function $\mathcal{L}_o = -(y_o \log(\hat{y}_o) + (1 - y_o)\log(1 - \hat{y}_o))$ where y_o is the ground truth label for our target outcome and $\hat{y}_o$ is our model's prediction for that outcome.

4.4.5 Training procedure

For our intermediate features to be interpretable, it is important for our problem extraction model to maintain a high level of performance. This motivates the development of our training procedure. We define a threshold for the performance of our problem extraction model and train only that component of our framework if the validation performance falls below that threshold. This ensures that we are only training the final classification layer using intermediate representations that effectively represent their corresponding problem. This also prevents our target classification objective from degrading the performance of our problem extraction model as that would harm the interpretability and clinical utility of our framework.

Thus our final loss function $\mathcal{L}$ can be defined as $\mathcal{L} = \begin{cases} \mathcal{L}_o + \mathcal{L}_p & \text{if } val_p \geq threshold_p \\ \mathcal{L}_p & \text{if } val_p < threshold_p \end{cases}$

where val_p is the validation performance and $threshold_p$ is a pre-defined performance threshold. We measure the performance of our problem extraction model by calculating the micro-averaged Area Under the Receiver Operating Curve (AU-ROC) on the validation set and use a threshold of 0.90 for the models trained in this work. We found this training procedure to be necessary to maintain good problem extraction performance for problem configurations

that involved full codes while the configurations with rolled codes were able to maintain performance during joint training. We optimize our final loss function using the Adam optimizer [93]. Our code is made publicly available[2] and we relegate full implementation details to the appendix.

4.5 Experiments and results

4.5.1 Baselines

To evaluate the efficacy of our proposed framework at predicting our target outcomes, we develop three strong baselines based on recent work for clinical outcome prediction using clinical text [73, 74, 72]. The first baseline is the convolutional model developed by Kim [94] for text classification. This model consists of three convolutions of width 1, 2, and 3 which are applied over the clinical narrative and then max-pooled. The three pooled representations are then concatenated and used for the final prediction.

The second baseline is similar to the model used for problem extraction in our proposed framework and is a straightforward extension of the model proposed by Mullenbach et al. [83]. Unlike our problem extraction model, this baseline utilizes a single attention head and directly predicts the outcome of interest. This baseline allows us to not only compare the predictive performance of our model, but to also explore the improved interpretability that our framework provides. For our third baseline, we use a bidirectional LSTM augmented with an additive attention mechanism which was used by [72] in their work predicting clinical outcomes from notes.

4.5.2 Outcome Results

For each outcome in this work, we explore using both full and rolled ICD codes and phecodes as our intermediate problems. To gain insight into the effectiveness of each subset of codes, we also explore using only the rolled ICD diagnosis codes, rolled ICD procedure codes, and rolled phecodes. For every model, we report the mean and standard deviation

[2] `https://github.com/justinlovelace/Dynamic-Problem-Lists`

across the five testing folds for the area under the Receiver Operating Curve (AU-ROC) and the area under the Precision-Recall Curve (AUC-PR) to evaluate the effectiveness of our models. The results for all of the outcomes explored in this work can be found in Table 4.1.

As expected, we find that trying to use the entire set of ICD codes for our intermediate problem representation is relatively ineffective, being outperformed by at least one of our baselines across all outcomes. We also observe that this problem extends to trying to utilize the full set of phecodes. However, we find that our model is very effective when using rolled ICD codes or phecodes. When using rolled codes, we find that our proposed framework outperforms all baselines with multiple different problem configurations across all outcomes and performance metrics.

Somewhat surprisingly, we find that using the individual subsets of codes does not lead to any loss in performance and appears to marginally improve performance. It is possible that the additional information provided by combining diagnostic and procedure codes is offset by difficulties that come from increasing the label space. We find that our framework leads to not only improved clinical utility (which we demonstrate later in this work), but also improved predictive performance.

4.5.3 Problem Extraction Results

For our model to be interpretable, it is important for the problem extraction model to be effective. To explore the performance of our problem extraction model and the effect that the additional learning objective has on that performance, we conduct an additional experiment where we train our problem extraction model independently and compare it with the performance of our intermediate problem extraction model in our framework across all outcomes. We report results for this experiment in Table 4.2.

We observe that our problem extraction method is performant across all of the target outcomes in this work. However, we find that our problem extraction model is consistently more effective when using rolled codes as opposed to full sets of codes. This is understandable as the larger label space and finer grained distinctions between the codes leads to a more

Table 4.1: Outcome Prediction Results

Model	Problem Set	In-Hospital Mortality		30-Day Mortality	
		AU-ROC	AU-PR	AU-ROC	AU-PR
CNN-Max	-	0.852 ± 0.015	0.323 ± 0.048	0.842 ± 0.008	0.430 ± 0.009
Conv-Attn	-	0.865 ± 0.015	0.330 ± 0.038	0.852 ± 0.007	0.415 ± 0.012
LSTM-Attn	-	0.853 ± 0.015	0.308 ± 0.046	0.855 ± 0.008	0.431 ± 0.007
DynPL	F-ICD$_{\text{Diag}}$ & F-ICD$_{\text{Proc}}$	0.823 ± 0.023	0.218 ± 0.036	0.821 ± 0.012	0.352 ± 0.031
DynPL	F-Phe & R-ICD$_{\text{Proc}}$	0.837 ± 0.047	0.252 ± 0.090	0.836 ± 0.013	0.393 ± 0.028
DynPL	R-ICD$_{\text{Diag}}$ & R-ICD$_{\text{Proc}}$	$\mathbf{0.866 \pm 0.011}$	0.322 ± 0.046	$\mathbf{0.857 \pm 0.005}$	$\mathbf{0.438 \pm 0.012}$
DynPL	R-Phe & R-ICD$_{\text{Proc}}$	$\mathbf{0.865 \pm 0.016}$	$\mathbf{0.330 \pm 0.040}$	$\mathbf{0.855 \pm 0.006}$	$\mathbf{0.435 \pm 0.007}$
DynPL	R-ICD$_{\text{Diag}}$	$\underline{\mathbf{0.869 \pm 0.010}}$	$\underline{\mathbf{0.332 \pm 0.037}}$	0.852 ± 0.008	0.424 ± 0.021
DynPL	R-ICD$_{\text{Proc}}$	0.863 ± 0.011	0.329 ± 0.030	$\mathbf{0.855 \pm 0.005}$	$\underline{\mathbf{0.443 \pm 0.011}}$
DynPL	R-Phe	$\mathbf{0.867 \pm 0.014}$	0.327 ± 0.040	$\underline{\mathbf{0.858 \pm 0.007}}$	$\mathbf{0.440 \pm 0.021}$

Model	Problem Set	Bounceback Readmission		30-Day Readmission	
		AU-ROC	AU-PR	AU-ROC	AU-PR
CNN-Max	-	0.661 ± 0.018	0.148 ± 0.016	0.650 ± 0.011	0.212 ± 0.018
Conv-Attn	-	0.707 ± 0.009	0.173 ± 0.018	0.684 ± 0.004	0.235 ± 0.017
LSTM-Attn	-	0.695 ± 0.010	0.154 ± 0.009	0.681 ± 0.008	0.231 ± 0.021
DynPL	F-ICD$_{\text{Diag}}$ & F-ICD$_{\text{Proc}}$	0.667 ± 0.015	0.138 ± 0.018	0.659 ± 0.011	0.213 ± 0.016
DynPL	F-Phe & R-ICD$_{\text{Proc}}$	0.692 ± 0.014	0.154 ± 0.013	0.669 ± 0.006	0.219 ± 0.008
DynPL	R-ICD$_{\text{Diag}}$ & R-ICD$_{\text{Proc}}$	0.703 ± 0.013	0.168 ± 0.021	0.683 ± 0.003	0.234 ± 0.016
DynPL	R-Phe & R-ICD$_{\text{Proc}}$	0.705 ± 0.017	0.168 ± 0.019	$\mathbf{0.687 \pm 0.005}$	0.234 ± 0.011
DynPL	R-ICD$_{\text{Diag}}$	$\underline{\mathbf{0.710 \pm 0.014}}$	0.170 ± 0.011	$\mathbf{0.688 \pm 0.004}$	$\mathbf{0.238 \pm 0.012}$
DynPL	R-ICD$_{\text{Proc}}$	$\mathbf{0.708 \pm 0.012}$	$\underline{\mathbf{0.178 \pm 0.019}}$	$\underline{\mathbf{0.690 \pm 0.006}}$	$\underline{\mathbf{0.239 \pm 0.017}}$
DynPL	R-Phe	$\underline{\mathbf{0.710 \pm 0.013}}$	$\mathbf{0.173 \pm 0.019}$	$\mathbf{0.689 \pm 0.003}$	$\mathbf{0.238 \pm 0.011}$

F=Full Codes, R=Rolled Codes. Bolded values indicate equivalent or superior performance compared to all baselines and the best performance is underlined.

Table 4.2: Problem Extraction Results

Target Outcome	F-ICD$_{\text{Diag}}$ & F-ICD$_{\text{Proc}}$		F-Phe & R-ICD$_{\text{Proc}}$		R-ICD$_{\text{Diag}}$ & R-ICD$_{\text{Proc}}$		R-Phe & R-ICD$_{\text{Proc}}$	
	Micro AU-ROC	Macro AU-ROC	Micro AU-ROC	Macro AU-ROC	Micro AU-ROC	Macro AU-ROC	Micro AU-ROC	Macro AU-ROC
Problem Extraction	**0.946± 0.001**	**0.887± 0.002**	**0.945± 0.001**	**0.877± 0.002**	**0.952± 0.000**	**0.888± 0.002**	**0.952± 0.001**	**0.879± 0.003**
Bounceback Readmission	0.853 ± 0.005	0.753 ± 0.005	0.889 ± 0.003	0.760 ± 0.007	0.905 ± 0.002	0.754 ± 0.009	0.908 ± 0.002	0.744 ± 0.010
30-Day Readmission	0.865 ± 0.022	0.756 ± 0.013	0.891 ± 0.009	0.764 ± 0.006	0.905 ± 0.001	0.748 ± 0.008	0.908 ± 0.002	0.739 ± 0.010
In-Hospital Mortality	0.862 ± 0.022	0.738 ± 0.014	0.887 ± 0.012	0.753 ± 0.021	0.906 ± 0.004	0.754 ± 0.007	0.906 ± 0.003	0.740 ± 0.009
30-Day Mortality	0.847 ± 0.026	0.733 ± 0.021	0.893 ± 0.011	0.757 ± 0.006	0.902 ± 0.002	0.749 ± 0.006	0.902 ± 0.002	0.733 ± 0.007

challenging classification problem. This reduced problem extraction performance when using the full set of codes is likely a contributing factor to the poorer target outcome performance observed when using full sets of codes.

We do observe that the addition of the target outcome objective does degrade performance when compared to a model trained exclusively on problem extraction. This degradation demonstrates the importance of our training procedure to ensure that the intermediate problem extraction remains effective.

4.5.4 Effect of End-to-End Training

We conduct an ablation experiment to evaluate the effect of end-to-end training on our framework's performance by first training our framework only on problem extraction, freezing the problem extraction component, and then fine-tuning the final classification layer to predict the outcome of interest. We report results for this experiment in Table 4.3 and observe a consistent decrease in performance when training the two components separately. This decrease is particularly notable for both mortality outcomes. This is likely because the feature space defined by the problems fail to represent all pertinent information from the notes and training the network end-to-end allows for some adaptation to the final outcome. For example, the frozen problem extraction model would not be incentivized to recognize the severity of problems while such information would be useful when predicting the target outcomes.

4.5.5 Comparison Against Oracle

We conduct an additional experiment to explore the effectiveness of our problem extraction model. In this experiment we train a logistic regression oracle to predict the outcomes directly from the ground truth labels derived from ICD codes. It is important to note that because ICD codes are associated with entire hospital stays in our dataset, this experiment involves using future information compared to the clinically useful application setting of our other models. Not only are ICD codes themselves unavailable at the time of ICU discharge,

Table 4.3: Effect of End-to-End Training

Model	Problem Set	In-Hospital Mortality		30-Day Mortality	
		AU-ROC	AU-PR	AU-ROC	AU-PR
DynPL	R-ICD$_{\text{Diag}}$ & R-ICD$_{\text{Proc}}$	$\mathbf{0.866 \pm 0.011}$	$\mathbf{0.322 \pm 0.046}$	$\mathbf{0.857 \pm 0.005}$	$\mathbf{0.438 \pm 0.012}$
Frozen DynPL	R-ICD$_{\text{Diag}}$ & R-ICD$_{\text{Proc}}$	0.852 ± 0.008	0.254 ± 0.032	0.847 ± 0.006	0.365 ± 0.024
DynPL	R-Phe & R-ICD$_{\text{Proc}}$	$\mathbf{0.865 \pm 0.016}$	$\mathbf{0.330 \pm 0.040}$	$\mathbf{0.855 \pm 0.006}$	$\mathbf{0.435 \pm 0.007}$
Frozen DynPL	R-Phe & R-ICD$_{\text{Proc}}$	0.837 ± 0.017	0.215 ± 0.035	0.834 ± 0.011	0.322 ± 0.032

Model	Problem Set	Bounceback Readmission		30-Day Readmission	
		AU-ROC	AU-PR	AU-ROC	AU-PR
DynPL	R-ICD$_{\text{Diag}}$ & R-ICD$_{\text{Proc}}$	$\mathbf{0.703 \pm 0.013}$	$\mathbf{0.168 \pm 0.021}$	$\mathbf{0.683 \pm 0.003}$	$\mathbf{0.234 \pm 0.016}$
Frozen DynPL	R-ICD$_{\text{Diag}}$ & R-ICD$_{\text{Proc}}$	0.698 ± 0.011	0.161 ± 0.012	0.677 ± 0.004	0.224 ± 0.010
DynPL	R-Phe & R-ICD$_{\text{Proc}}$	$\mathbf{0.705 \pm 0.017}$	$\mathbf{0.168 \pm 0.019}$	$\mathbf{0.687 \pm 0.005}$	$\mathbf{0.234 \pm 0.011}$
Frozen DynPL	R-Phe & R-ICD$_{\text{Proc}}$	0.700 ± 0.008	0.163 ± 0.008	0.680 ± 0.008	0.229 ± 0.017

but the codes could represent medical problems or procedures that arise or occur later in a patient's hospital stay after the patient is discharged from the ICU.

Nevertheless, this experiment can provide some insight into the effectiveness of our problem extraction model and whether it is currently a performance bottleneck. We report results for this logistic regression oracle across two of our problem configurations in Table 4.4. We find that using the ground truth labels leads to notably improved performance compared to our framework for the readmission outcomes, but actually leads to worse performance for most of the mortality outcomes. While the improvement for readmission outcomes can likely be attributed in part to the use of future information, the improvement likely also results from the improved accuracy of the problem labels, suggesting that the efficacy of our problem extraction model is a limiting factor in our framework's performance. However, our framework is not reliant on any particular architecture for problem extraction and this experiment demonstrates that as advances continue to be made on the task of automated ICD coding, our framework will become increasingly viable. The worse performance for mortality outcomes again suggests that the problem space doesn't perfectly represent all of the relevant information contained within the notes and highlights the importance of our

Table 4.4: Comparison Against Oracle

Model	Problem Set	In-Hospital Mortality		30-Day Mortality	
		AU-ROC	AU-PR	AU-ROC	AU-PR
DynPL	R-ICD$_{Diag}$ & R-ICD$_{\text{Proc}}$	0.866 ± 0.011	0.322 ± 0.046	$\mathbf{0.857 \pm 0.005}$	$\mathbf{0.438 \pm 0.012}$
LR Oracle	R-ICD$_{Diag}$ & R-ICD$_{\text{Proc}}$	$\mathbf{0.875 \pm 0.015}$	$\mathbf{0.331 \pm 0.062}$	0.839 ± 0.003	0.404 ± 0.012
DynPL	R-Phe & R-ICD$_{\text{Proc}}$	$\mathbf{0.865 \pm 0.016}$	$\mathbf{0.330 \pm 0.040}$	$\mathbf{0.855 \pm 0.006}$	$\mathbf{0.435 \pm 0.007}$
LR Oracle	R-Phe & R-ICD$_{\text{Proc}}$	0.850 ± 0.015	0.268 ± 0.049	0.818 ± 0.010	0.320 ± 0.039

Model	Problem Set	Bounceback Readmission		30-Day Readmission	
		AU-ROC	AU-PR	AU-ROC	AU-PR
DynPL	R-ICD$_{\text{Diag}}$ & R-ICD$_{\text{Proc}}$	0.703 ± 0.013	0.168 ± 0.021	0.683 ± 0.003	0.234 ± 0.016
LR Oracle	R-ICD$_{\text{Diag}}$ & R-ICD$_{\text{Proc}}$	$\mathbf{0.807 \pm 0.013}$	$\mathbf{0.294 \pm 0.039}$	$\mathbf{0.732 \pm 0.007}$	$\mathbf{0.314 \pm 0.016}$
DynPL	R-Phe & R-ICD$_{\text{Proc}}$	0.705 ± 0.017	0.168 ± 0.019	0.687 ± 0.005	0.234 ± 0.011
LR Oracle	R-Phe & R-ICD$_{\text{Proc}}$	$\mathbf{0.808 \pm 0.013}$	$\mathbf{0.286 \pm 0.034}$	$\mathbf{0.733 \pm 0.007}$	$\mathbf{0.312 \pm 0.013}$

end-to-end training regime which allows for some adaptation to the outcome of interest.

4.5.6 Label Integrity

Although our framework's problem extraction performance provides a straightforward way to validate the effectiveness of our problem extraction model, it is not a perfect method due to the nature of our ground truth labels. A number of past works have demonstrated that ICD codes are an imperfect representation of ground truth phenotypes in actual clinical practice [95, 96, 97, 98, 99, 100, 101, 102, 103]. A common trend observed in work exploring the accuracy of ICD codes is that they have strong specificity but poorer sensitivity. In other words, a patient assigned a given code very likely has the corresponding condition, but there are likely more patients with that condition than only the patients who were assigned that ICD code. Given that our dataset contains information gathered during routine clinical care, the ICD codes we use as ground truth labels in this work likely suffer from the same problem.

Because of this complication, perfect problem extraction performance, as evaluated by using ICD codes as ground truth labels, is actually suboptimal. In such a case, the model would have learned to perfectly replicate the biases and mistakes in the ICD coding process

instead of correctly identifying all of the clinical problems. We hypothesize that if our problem extraction model is effective, then there are likely some 'incorrect' predictions that count against our model in the evaluation above that are actually correct. To evaluate this hypothesis, we conduct an expert evaluation over a limited set of predictions.

Because ICD codes tend to have problems with sensitivity, most of the errors with our ICD labels should be false negatives. To evaluate whether our problem extraction model is correctly recognizing some of the problems missed by the ICD codes, we extract the 50 most confident false positives for one of the models trained in this work and manually evaluate whether the patient actually has the corresponding problem. It is important to note that when conducting the evaluation, we are not necessarily following ICD coding standards. We are instead identifying whether the patient has the corresponding problem to explore challenges with using ICD codes to represent phenotype labels as is being done in this work and has been done in prior work [104]. We report the results for this experiment in Table 4.5.

Table 4.5: Expert Evaluation of 50 False Positives

	Count	Percentage
Correct Prediction	37	74%
Correct Label	13	26%

We observe that our hypothesis was correct and that a large majority (74%) of the false positives that we extracted from our model were actually correct predictions penalized due to label inaccuracies. This demonstrates that our model is already reasonably robust to these label inaccuracies and is successfully extracting problems despite noisy labels. We also observe that the actual false positives are often well grounded in the text. For example, radiologists prioritize sensitivity over specificity when reporting observations, and we found multiple false positives resulting from radiological findings that required clinical correlation.

Although there is a large body of work in ICD code classification in MIMIC [83, 84, 82, 85], we are the first to conduct such an analysis demonstrating the ability of our model to overcome label inconsistencies.

4.6 Interpretability

While we demonstrated that our framework is performant, its primary strength is the simplicity of interpretation that it provides. Tonekaboni et al. [105] surveyed clinicians to identify aspects of explainable modeling that improve clinician's trust in machine learning models. Clinicians identified being able to understand feature importance as a critical aspect of explainability so that they can easily compare the model's decision with their clinical judgement. Clinicians expected to see both global feature importance and patient-specific importance so we explore both of those in this work.

4.6.1 Global Trends

A large body of prior work has explored the interpretability of attention, but that exploration is typically limited to individual predictions [83, 106, 107]. While that is useful, it is also important to gain an understanding of population level trends.

By designing our frameworks such that the value for the final prediction is a linear combination of the extracted problem scores, we can simply extract the weights from the final layer of our model to gain an understanding of which problems are important. We calculated the mean and standard deviation for each problem over the five folds and present the strongest risk factors across all outcomes in Table 4.6. We observe that there are a number of common risk factors between outcomes. We find that the top four risk factors for both readmission tasks were fluid disorders; puncture of vessel; renal failure; and congestive heart failure, not hypertensive. We find that urinary tract infections and pneumonia were both strong factors for mortality as well as the shared readmission risk factors of puncture of vessel and fluid disorders.

We also explored whether there were factors associated with healthy outcomes but found

that even the most negative weights had a small magnitude that was insignificant given their variance. Thus our model appears to recognize a limited number of positive risk factors while the majority of the intermediate problems have little effect on the outcome. This makes it well-suited for producing clutter-free problem lists for clinicians which we explore in the next section.

Table 4.6: Risk Factors for Target Outcomes

30-Day Mortality		In-Hospital Mortality	
Problem	Weight	Problem	Weight
Disorders of fluid, electrolyte, and acid-base balance	0.151 ± 0.057	Disorders of fluid, electrolyte, and acid-base balance	0.189 ± 0.027
Puncture of vessel	0.091 ± 0.035	Urinary tract infection	0.081 ± 0.018
Pneumonia	0.078 ± 0.026	Puncture of vessel	0.079 ± 0.060
Urinary tract infection	0.072 ± 0.040	Renal failure	0.073 ± 0.020
Congestive heart failure; nonhypertensive	0.067 ± 0.016	Pneumonia	0.071 ± 0.018
30-Day Readmission		Bounceback Readmission	
Problem	Weight	Problem	Weight
Disorders of fluid, electrolyte, and acid-base balance	0.110 ± 0.019	Disorders of fluid, electrolyte, and acid-base balance	0.081 ± 0.025
Renal failure	0.081 ± 0.022	Puncture of vessel	0.076 ± 0.028
Puncture of vessel	0.072 ± 0.019	Congestive heart failure; nonhypertensive	0.059 ± 0.014
Congestive heart failure; nonhypertensive	0.069 ± 0.021	Renal failure	0.037 ± 0.014
Other anemias	0.069 ± 0.061	Hypertension	0.037 ± 0.034

4.6.2 Individual Predictions

We construct dynamic problem lists by extracting the 14 strongest problem predictions. We chose to extract 14 problems because the patients in the training fold had an average of 13.8 codes assigned to their hospital stay so 14 problems should provide an adequate summary of the patient's state. We report these problems sorted by their extraction probability and also report the importance of each problem for the final outcome so that clinicians can easily identify what factors are driving the prediction. For the problem importance, we scale the

Table 4.7: Dynamic Problem Lists

Problem	Extraction Probability	Problem Weight	Top Two Spans of Attended Text
High-Risk Bounceback Readmission Patient			
Other operations of abdominal region (includes paracentesis)	0.950	0.16	[to attempt **paracentesis** again today] [suitable for **paracentesis** was marked]
Chronic liver disease and cirrhosis	0.939	0.28	[to attempt **paracentesis** again today] [suitable for **paracentesis** was marked]
Injection or infusion of therapeutic or prophylactic substance	0.838	0.31	[started on **tpn** plan was] [remains on **tpn** at present]
Puncture of vessel	0.797	1.00	[, beir **hugger applied d/t** low temp.;] [reddend alovesta **cream applied id** : tmax]
Disorders of fluid, electrolyte, and acid-base balance	0.732	0.92	[will be **performed lft's elevated** being followed] [pt is **jaundiced , excoriated** perianal area]
Septicemia	0.556	0.15	[support , **sepsis work-up** p-will] [levofloxacin and **flagyl** po skin]
Ascites (non malignant)	0.539	0.12	[to attempt **paracentesis** again today] [suitable for **paracentesis** was marked]
Transfusion of blood and blood components	0.484	0.06	[pt had **egd** this pm] [rec'd # **units** ffp with]
Prophylactic vaccination and inoculation against certain viral diseases	0.460	0.06	[support , **sepsis** work-up p-will] [history of **hepatorenal** failure and]
Chronic ulcer of skin	0.450	0.32	[, beir **hugger applied** d/t low temp.;] [reddend alovesta **cream applied id** : tmax]
Renal failure	0.404	0.39	[s/p now **with renal failure** reason for] [s/p now **with renal failure** reason for]
Peritonitis and retroperitoneal infections	0.360	0.04	[to attempt **paracentesis** again today] [suitable for **paracentesis** was marked]
Other anemias	0.359	0.06	[rec'd n **units** ffp with] [rec'd n **unit** ffp with]
Viral hepatitis	0.349	0.13	[status , **lactulose** prn as] [remains on **lactulose** prn to]
Low-Risk Bounceback Readmission Patient (Truncated)			
Diagnostic procedures on small intestine	0.547	−0.07	[presently another **endoscopy** is scheduled] [had an **endoscopy** which revealed]
Other anemias	0.284	0.06	[nnd unit **prbc** infusing presently] [n unit **prbc** with initial]
Diseases of esophagus	0.252	−0.05	[presently another **endoscopy** is scheduled] [had an **endoscopy** which revealed]
Effects radiation NOS	0.223	0.06	[nnd unit **prbc** infusing presently] [n unit **prbc** with initial]

Table 4.8: Baseline Attention Interpretation

Highly Attended Text	
High-Risk Bounceback Readmission Patient	Low-Risk Bounceback Readmission Patient
[radiology to **attempt paracentesis** again today]	[small amts **ice chips awaiting** nnd endoscopy]
[iv bid **old tap** site from]	[understanding of **discharge instructions** and new]
[planning to **do tap** this evening]	[daughters care **discharge instructions** reviewed with]
[further oozing **needs c-diff spec** pmicu nursing]	[fbleeding noted **discharge instructions** , pt]
[was d/cd **a paracentesis** was attempted]	[ice chips **per team neuro** : a&oxn]
[of ice **chips tpn** infusing as]	[taking medication **discharge planning** complete with]
[overnight mushroom **cath draining loose** brown-green stool]	[scheduled for **this am- ?** nam pt]
[was started **on tpn** plan was]	[of chron's **disease** and lower]
[pt remains **on tpn** at present]	[, denies **sob** rr nn-nn]
[status , **lactulose** prn as]	[, dry **, intact without** reddness or]
[remains on **lactulose** prn to]	[up the **clots** pt transferred]
[re-oriented rec'ing **lactulose** po has]	[chron's disease **and lower gib** , now]
[pt given **lactulose** x n]	[in the <**loc**> **area plan** : repeat]
[on po **lactulose** perl ,]	[given iv **erythromycin** and iv]

problem weights to the range $[-1, 1]$ by dividing by the problem weight with the greatest magnitude to allow for easier interpretation, and we also provide the spans of text attended to by the model to make each problem prediction. To provide a comparison using our baseline convolutional attention model, we extract the 14 spans of text with the greatest attention weights associated with them.

We provide an example of a dynamic problem list for a patient predicted to be at high risk of bounceback readmission in Table 4.7. From looking at the dynamic problem list, we can quickly identify the most important problems driving the risk prediction (puncture of vessel, fluid disorder, renal failure, skin ulcer, intravenous feeding, and liver disease) while understanding that the other problems are insignificant. Reporting the quantitative importance of each problem saves the clinician from having to manually filter through the long list of problems. Furthermore, the extraction probability provides a measure of uncertainty which, along with the attended text, allows clinicians to intelligently verify the model's performance. For example, renal failure is an important risk factor but has a relatively low extraction prob-

ability of 0.404. Upon inspecting the highlighted text, the clinician can clearly observe that the extraction was accurate and the patient is suffering from that condition. It is also worth noting that in this example the problem extraction model was able to successfully recognize that the patient had bed ulcers and a platelet transfusion, both of which are not represented by the ICD labels in the dataset.

By comparison, we provide the baseline visualization from the convolutional attention model for the same patient in Table 4.8. Here, we can only observe much broader trends and there is a large degree of redundancy (e.g. paracentesis and tap refer to the same procedure). We can observe that the patient has severe liver problems from the need for paracentesis and the use of the medication, lactulose. We can also observe that the patient required intravaneous feeding from the references to total parenteral nutrition (TPN). However, there is a significant amount of redundancy and it is not clear how to meaningfully aggregate these observations to actually gain an understanding of what clinical outcomes the model is extracting and how important they are for the final outcome. Furthermore, the overview of the patient is much less comprehensive than that provided by the dynamic problem list, with all of the information extracted by the baseline being concisely aggregated into three codes (Chronic liver disease and cirrhosis, Other operations of abdominal region, and Injection or infusion of therapeutic or prophylactic substance) in the dynamic problem list that quantitatively reports the importance of those conditions.

We compare a dynamic problem list to our baseline for a low-risk bounceback patient in the same tables and find that the benefits are even more pronounced. When examining the baseline visualization we observe that the model is primarily focusing on references to discharge instructions which don't actually convey any clinically meaningful information. Similarly, the other phrases attended to do not seem to convey any important medical information. On the other hand, the dynamic problem list for the low-risk patient still effectively extracts clinical conditions (that the patient had an esophageal disease, was anemic, etc.) and then concludes that the extracted conditions do not warrant concern. This clearly

demonstrates to a clinician that the model is still effectively extracting the patient's clinical condition, but that it judges that condition to be safe. This transparency is important for the clinician to be able to trust that the model is effective.

4.7 Qualitative Expert User Study

While we have argued for the improved utility of our framework compared against recent work within the domain, it is important to verify that claim by conducting a user study with medical experts. For example, it may be possible that while our framework is sound in theory, the problem extraction stage is sufficiently noisy to render the extracted problem lists useless in practice. To examine the utility of our framework, we recruited four medical experts and conducted a user study where our experts evaluated the utility of our dynamic problem list and the baseline interpretation method. Three of our experts are currently practicing physicians while one is an MD-PhD student with one year of medical school remaining. Two of the medical experts are co-authors who were involved in some parts of the development of this work while the other two had no involvement with our work beyond taking part in the user study.

We conducted our user study by randomly sampling 25 ICU stays from the test set of one of our 30 day readmission models. Because of the imbalanced nature of our dataset, we sample 10 stays from the top 5% of predicted risks and sample the other 15 stays from the remaining ICU stays. This ensures that we evaluate our framework for both high risk patients and patients that are representative of the general patient population. We then provided each of our expert reviewers with the clinical notes associated with each patient and instructed them to briefly review them to gain an understanding of the patient's medical condition. We then presented them with our dynamic problem list and the baseline attention extraction along with the predicted readmission risk and the reviewers evaluated both methods independently using the Likert Scale seen in Figure 4.9.

We report the results for this study in Table 4.10 and compute the statistical significance for two comparisons. We examine the relationship between the two interpretation meth-

ods using a two-tailed paired t-test and also explore whether the dynamic problem list is meaningfully better than a neutral rating using a two-tailed one sample t-test. The first comparison allows us to examine whether our method is an improvement over the baseline while the second allows us to evaluate whether the medical expert's judged our method favorably. We observe that every expert found our framework to be more effective than the baseline method and the difference was statistically significant for all but one expert. Additionally, every expert found the problem list to be meaningfully better than a neutral rating by a statistically significant margin. By contrast, two of our experts rated the baseline worse than neutral and none of the experts rated it to be better than neutral by a statistically significant margin. When averaging the scores for each patient across all experts, we find that our method received a rating of 3.66 on average compared to 2.85 for the baseline method, a meaningful improvement over both the baseline ($p < 0.001$) and a neutral rating ($p < 0.001$). These improvements are still significant even when limiting the evaluation to the two external experts to account for potential biases from the experts who were familiar with this work. While a much more stringent evaluation would need to be conducted (such as a randomized controlled trial) before implementing our method in practice, this preliminary qualitative evaluation is promising and more rigorous evaluations are left to future work.

Table 4.9: Likert Scale

The list effectively identifies and presents relevant medical factors for evaluating readmission risk for this patient.				
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	2	3	4	5

Table 4.10: User Study

	Medical Expert 1	2	3	4	Average	Average of External Experts
Convolutional Attention	3.13	2.52	2.52	3.32	2.85	2.92
Dynamic Problem List	**4.08**	**3.48**	**3.52**	**3.56**	**3.66**	**3.52**
DynPL > Conv-Attn	$p < 0.005$	$p < 0.005$	$p < 0.01$	$p = 0.110$	$p < 0.001$	$p < 0.005$
DynPL > Neutral	$p < 0.001$	$p < 0.05$	$p < 0.05$	$p < 0.01$	$p < 0.001$	$p < 0.01$

4.8 Limitations and Future Work

We did not make the problem extraction architecture a large focus of this work and instead used a model representative of the recent state-of-the-art. In the future, we intend to improve upon the problem extraction module in our framework. In particular, we intend to explore whether we can utilize pre-trained language models to improve our problem extraction and downstream performance given their recent success across a wide variety of tasks both outside of and within the clinical domain [108, 109]. In this work, we augmented our problem extraction module with a linear layer for its simplicity of interpretation and found that it led to strong performance. However, incorporating our problem extraction module into a more sophisticated model could potentially lead to meaningful improvements in performance and we intend to pursue this in future work. We would also like to extend this framework to other outcomes of clinical interest such as sepsis or the onset of intubation to evaluate its ability to generalize beyond the outcomes examined in this work.

5. USE OF MACHINE LEARNING MODELS TO PREDICT DEATH AFTER ACUTE MYOCARDIAL INFARCTION*

Turning back to outcome prediction, this chapter compares several machine learning methods in predicting mortality following acute myocardial infarction. Here, we compare several the performance of several models with both expanded and parsimonious variable sets. We compare to a previously published model and report on differences in model calibrations. While accurate predictions are important, so too is it important to understand the confidence of a model's prediction.

5.1 Introduction

An assessment of risk of death after an acute myocardial infarction (AMI) is useful for guiding clinical decisions for patients and for assessing hospital performance [110, 111, 112]. New analytic approaches may enhance risk prediction with existing data beyond traditional statistical approaches. Existing risk prediction models developed in the prediction of AMI outcomes have been limited by lack of inclusion of nonlinear effects and complex interactions among variables in national samples or have only evaluated these effects in small patient groups [113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124]. With advances in computation and analytics, however, it may be possible to create models in large and diverse patient groups, which may improve on traditional models with existing information. Specifically, the application of machine learning techniques has the potential to improve on accuracy in the prediction of in-hospital mortality after AMI [125, 126, 127].

Accordingly, using data collected in the Chest Pain–MI Registry (CP-MI Registry; formerly known as the ACTION Registry) of the National Cardiovascular Data Registry (NCDR),

*Reprinted with permission from "Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction" by Khera, Rohan; Haimovich, Julian; Hurley, Nathan C; McNamara, Robert; Spertus, John A; Desai, Nihar; Rumsfeld, John S; Masoudi, Frederick A; Huang, Chenxi; Normand, Sharon-Lise; Mortazavi, Bobak J; and Krumholz, Harlan M, 2021. *JAMA Cardiology*, 6, 633-641, Copyright© 2021 American Medical Association.

a national clinical quality program from the American College of Cardiology, we assessed whether machine learning techniques, compared with logistic regression, could improve prediction of in-hospital AMI mortality. The CP-MI Registry includes information on more than 1 million AMI hospitalizations at 1,163 hospitals across the US. We used the most contemporary published model for mortality after AMI, which used logistic regression [116, 117], to compare the performance characteristics of our models derived using machine learning.

5.2 Methods

This cohort study used the American College of Cardiology CP-MI Registry to identify all AMI hospitalizations between January 1, 2011, and December 31, 2016. Data analysis was performed from February 1, 2018, to October 22, 2020. The Yale University Institutional Review Board reviewed the study and waived the requirement for informed consent given the deidentified data. The study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline [128].

5.2.1 The CP-MI Registry

The CP-MI Registry collects data from participating hospitals on patients admitted with AMI, including both ST-elevation myocardial infarction (STEMI) and non-STEMI. Data are collected through retrospective medical record review and submitted using a standardized data collection tool. Collected data include patient demographics, presentation information, pre-hospital vital signs, selected laboratory data from the hospital course, procedures, timing of procedures, and select in-hospital outcomes. The NCDR data quality program enhances data completeness and accuracy through audits and feedback [129].

5.2.2 Patient Population

Between January 1, 2011, and December 31, 2016, a total of 993,905 patients with AMI from 1,128 hospitals were included. Similar to the approach used in prior studies [1, 130], patients transferred to another facility for management (n =47,308) or missing information on history of percutaneous coronary intervention, a key risk factor included in the current

standard for predicting mortality outcomes (n =191,195), were excluded (Table 5.1). Those patients excluded had age and sex distribution similar to those of patients included in the analysis but slightly higher rates of STEMI and unadjusted mortality (Table 5.1). After the exclusion of these patients, 755,402 patients remained for modeling. We also constructed a secondary cohort in which patients were not excluded for missing variables and covariates with missingness greater than 5% were excluded as predictors in the model (n =946,597).

5.2.3 Patient Variables and Data Definitions

Patient variables available to a practitioner at the time of presentation were selected for modeling. These variables include patient demographics, medical history, comorbidities, home medications, electrocardiogram findings, and initial medical presentation and laboratory values. The outcome of this study was death from any cause during hospitalization.

The current standard model for AMI mortality built within the NCDR uses 9 variables to predict mortality and was derived from 29 candidate variables using logistic regression by McNamara et al [1]. We included 2 sets of variables to build our machine learning models. First, we included the 29 variables used to derive the current NCDR standard [1]. Second, we used an expanded variable set with all other variables that would be available to a practitioner at the time of hospital presentation with an AMI (Table 5.2). A priori, we included variables that were available in at least 90% of patients, resulting in 8 candidate continuous variables and 48 categorical variables with a missing variable rate of less than 1%. For these variables, we imputed missing values to the mode for categorical variables and median for continuous variables. In sensitivity analyses, we pursued multiple imputation using the multivariate imputation by chained equations method, which derives predicted values of the missing values using a regression-based approach. These analyses were tested in a 5-fold validation exercise to evaluate the robustness of our strategy. Finally, we evaluated models that included patients who had been excluded from the primary analyses because of missing covariates (threshold $\geq$5%), thereby excluding key variables that are a part of the current standard.

Table 5.1: Differences in characteristics of patients excluded vs included in analyses

Characteristic	Excluded Patients (n = 191,195)		Included Patients (n = 755,402)	
	N(%)	Missing	N(%)	Missing
Demographic Characteristics				
Age, mean (SD), y	64 (14)	0	65 (14)	0
Weight, mean (SD), kg	87 (22)	566	87 (22)	1421
Female Sex	66,349 (35)	0	260,200 (34)	0
Race				
White	161,417 (84)	0	640,995 (85)	0
Black	21,612 (11)	0	87,089 (12)	0
Medical History				
History of diabetes	64,869 (34)	392	257,072 (34)	144
History of hypertension	141,071 (74)	257	562,423 (75)	74
History of dyslipidemia	189 (78)	190953	461,269 (61)	127
Current or recent smoker	64,963 (34)	246	253,829 (34)	145
Current dialysis	29 (27)	191089	68,086 (14)	283305
History of MI	5,176 (3)	603	19,055 (3)	244
History of HF	60 (39)	191040	188,297 (25)	175
Prior PCI	57 (40)	191053	94,897 (13)	704
Prior CABG	-	191195	193,179 (26)	0
History of atrial fibrillation	54 (36)	191046	100,897 (13)	393
Prior cerebrovascular disease	38 (25)	191040	62,312 (8)	519
Prior peripheral arterial disease	20,293 (11)	372	91,723 (12)	148
Presentation				
Presentation after cardiac arrest	8,499 (5)	2047	29,458 (4)	2581
In cardiogenic shock	8,180 (4)	384	28,783 (4)	584
In HF	22,569 (12)	306	95,240 (13)	529
Heart rate, mean (SD), beats/min	84 (24)	1089	84 (24)	2216
SBP at presentation, mean (SD), mm Hg	146 (36)	1166	147 (35)	2678
Presentation ECG findings				
STEMI	85,634 (45)	0	292,784 (39)	0
New or presumed new				
ST depressions	16,772 (9)	0	83,555 (11)	0
T-wave inversions	11,078 (6)	0	56,791 (8)	0
Transient ST-segment elevation lasting <20 min	1,920 (1)	0	8,279 (1)	0
Initial laboratory values				
Troponin ratio, mean (SD)	7.8 (8.3)	4088	7.3 (8.1)	12071
Creatinine, mean (SD), mg/dL	1.3 (1.2)	1098	1.3 (1.2)	4404
Creatinine clearance, mean (SD), mL/min	85 (43)	1634	85 (43)	5756
Hemoglobin, mean (SD), g/dL	14 (2)	1141	14 (2)	4426
Outcome				
In-hospital mortality	9432 (4.9)	0	33,468 (4.4)	0

Table 5.2: List of patient variables used in modeling. ∗denotes model variables used in McNamara et al. study [1]

Model Variables			
Demographics	Age∗	Presentation ECG	ST-elevation myocardial infarction∗
	Weight, kg∗		New or presumed new ST-segment depression∗
	BMI kg/m^2		New or presumed new T-wave inversion∗
	Sex∗		Transient ST-segment elevation $<$ 20 minutes∗
	Race		ST elevation
	Hispanic origin		Left bundle branch block
Medical History	History of diabetes mellitus∗		Isolated posterior MI
	Diabetes control	Home Medications	Aspirin
	History of hypertension∗		Clopidogrel
	History of dyslipidemia∗		ACE inhibitor
	Current/recent smoker∗		Angiotensin receptor blocker
	Current dialysis∗		Beta blocker
	Chronic lung disease∗		Statin
	History of MI∗		Non-statin lipid-lowering agent
	History of heart failure∗		Prasugrel
	Prior PCI∗		Warfarin
	Prior CABG∗		Aldosterone blocking agent
	History of atrial fibrillation∗	Initial Laboratory Tests	Initial CKMB collected
	Prior cerebrovascular disease∗		Initial Troponin collected
	Prior peripheral artery disease∗		Initial Creatinine collected
	Prior stroke		Initial Hemoglobin collected
	Prior transient ischemic attack		Lipid panel collected
Presentation	After Cardiac Arrest∗		Initial BNP collected
	In Cardiogenic shock∗		Initial pro-BNP collected
	In heart failure∗		Troponin Ratio∗
	Heart rate, bpm∗		Creatinine mg/dL∗
	SBP, mmHg∗		Creatinine Clearance∗
			Hemoglobin, g/dL∗

5.2.4 Modeling Strategies

We divided the data into an initial 75% subset (April 1, 2011, through September 30, 2015) for model development and the more recent 25% subset (October 1, 2015, through December 31, 2016) for model testing. The model development period was further divided into 2 equal halves (April 1, 2011, to September 30, 2013, and October 1, 2013, to September 30, 2015) to develop level 1 and level 2 models, respectively (Figure 5.1).

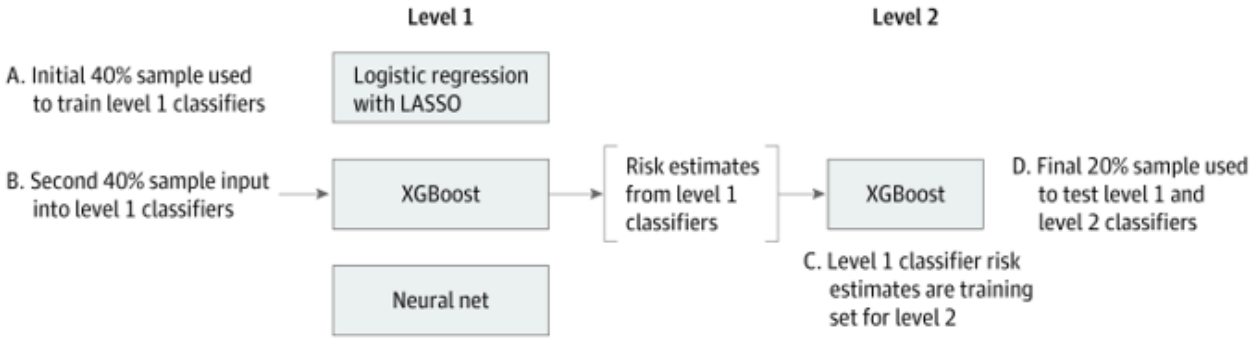


Figure 5.1: The level 1 classifiers consist of 3 independent models each trained on the same initial training sample (sample A), including logistic regression with least absolute shrinkage and selection operator (LASSO), extreme gradient descent boosting (XGBoost), and a neural network. The next training sample (sample B) is then input into the level 1 classifiers, resulting in 3 risk estimates for each observation in sample B, 1 from each level 1 model. These 3 risk estimates are then used to train the level 2 XGBoost classifier (sample C). A final sample (sample D) is input into the level 1 classifiers to obtain risk estimates for input into the level 2 classifier. Performance of the level 1 and level 2 classifiers is assessed using this final training set D.

We compared 3 modeling strategies with logistic regression: (1) gradient descent boosting, (2) a neural network, and (3) a meta-classifier approach that combined logistic regression with least absolute shrinkage and selection operator (LASSO) regularization, a gradient descent boosting, and a neural network. Gradient descent boosting models make predictions using a series of decision trees, representing an interpretable model. Unlike logistic regression, this model can include higher-order interactions and account for complex nonlinear relationships between model variables and outcomes. The method of gradient descent boosting chosen

was extreme gradient boosting, or XGBoost [131]. XGBoost incorporates a measure of how much model accuracy is improved by the addition of a given variable, with a higher gain value implying greater importance in generating a prediction. Neural networks are a type of machine learning technique that, like the human brain, connects layers of nodes (neurons) to model an output. Finally, the meta-classification approach uses an XGBoost model to combine the outputs of 3 supervised learning models, including LASSO, XGBoost, and a neural network (Figure 5.1) [67]. Therefore, the meta-classifier was a level 2 model that was based on the results of prediction models applied directly to patients (level 1 models).

The computational approach is shown in Figure 5.1. The first half of the derivation cohort was used to train 4 methods; logistic regression, LASSO, XGBoost, and a neural network. The second half of the derivation cohort was then used as a training set for the level 2 meta-classifier. We validated the various approaches with the remaining 25% of the sample.

5.2.5 Statistical Analysis

Model discrimination was measured using the area under the receiver operating characteristic curve (AUROC or C statistic) and its 95% CIs.25 In addition, the positive predictive value (or precision) and the sensitivity (recall) across all possible risk thresholds for predicting mortality were plotted using the precision-recall curve. The precision-recall curve, unlike the AUROC, is not affected by the number of true-negative results. In data sets with small event rates and therefore a large expected true-negative rate, such as the one studied here, the precision-recall curve is well suited for comparing different models. For both the C statistic and area under the precision-recall curve, values closer to 1 correspond to more accurate models.

Because the objective of the models is to address prediction at an individual level, we calculated the mean squared prediction error for each model, which represents the mean probability of an inaccurate prediction for a patient. A lower value suggests more accurate prediction. We also calculated the F score, sensitivity, specificity, positive predictive value,

and negative predictive value. In addition, we calculated a Brier score for each model as a measure of model accuracy. The score represents the reliability of the model minus the resolution plus an error term and represents the mean squared error between the observed and predicted risk [132, 133]. Further details are included in the eMethods and Figure 5.2.

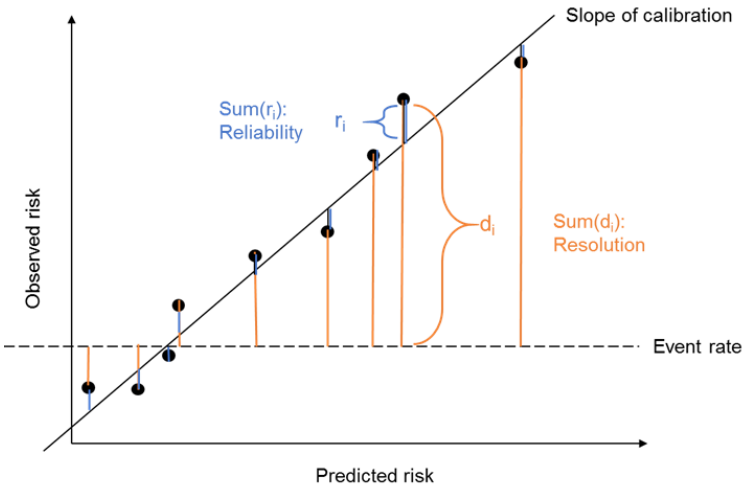


Figure 5.2: Each point represents the predicted versus observed risk at a given decile of risk. Reliability is the sum of the mean-squared error between the deciles of predicted risk and observed risk, and resolution is the mean-squared error between deciles of predicted risk and the event rate of the entire cohort.

Model calibration was measured using (1) the calibration slope, which was calculated as the regression slope of the observed mortality rates across the deciles of predicted mortality rates; (2) the reliability component of the Brier score; and (3) shift tables, in which we classified patients in the validation cohort into prespecified categories of low (<1%), moderate (1%-5%), or high risk (>5%) of death based on logistic regression and one of the machine learning models, creating a 9-way matrix of patients that included risk profiles assigned by the 2 models (low-low, low-moderate, and so on). We then calculated the actual rate of events in these groups, focusing on discordant categories, and compared them against the observed rates of mortality. We conducted sensitivity analyses with risk thresholds set at

less than 1.5%, 1.5% to 3%, and greater than 3%.

All analyses were conducted using open-source Python, version 3.8.0 (Python Software Foundation) and R software, version 3.6 (R Foundation for Statistical Computing). The level of significance was set at a 2-sided P<.05.

5.3 Results

5.3.1 Characteristics of Study Population

A total of 755,402 patients (mean [SD] age, 65[13] years; 495, 202[65.5%] male) were identified during the study period. Among the 755,402 patients in the primary study cohort, the overall in-hospital mortality rate was 4.4%. The derivation cohort consisted of 281,997 patients used to derive the level 1 classifiers, 282,921 to train the meta-classifier model (level 2 model), and the remaining 190,484 patients for the test cohort. Table 5.3 includes characteristics for the derivation and validation cohorts. A total of 562,423 patients (74%) had hypertension, 257,072 (34%) had diabetes, 188,297 (25%) had experienced a prior myocardial infarction, and 94,897 (13%) had a diagnosis of heart failure. In addition, 292,784 (39%) presented with a STEMI, 95,240 (13%) with heart failure, 28,783 (4%) with cardiogenic shock, and 29,458 (4%) after cardiac arrest (Table 5.3).

5.3.2 Model Discrimination

The current NCDR model with 9 variables had good discrimination (AUROC, 0.867) using β coefficients in the original model applied to the data. In models that used the 29-variable set that was used to derive the NCDR standard, machine learning models achieved modest improvements in discrimination over logistic regression using the same data inputs (Table 5.4). The AUROC for all 3 models was numerically higher than logistic regression in both the limited variable set and the expanded variable set, with corresponding improvements in the area under the precision-recall curve (Table 5.4; Figure 5.3). The XGBoost and meta-classifier models achieved a discrimination of 0.898 (95% CI, 0.894-0.902) and 0.899 (95% CI, 0.895-0.903), respectively, applied to the expanded set of variables compared with 0.888

Table 5.3: Baseline characteristics of the derivation and validation cohorts

Characteristic	Derivation Cohort (n = 564,918)	Validation Cohort (n = 190,484)
Demographic Characteristics		
Age, mean (SD), y	65 (14)	65 (13)
Weight, mean (SD), kg	87 (22)	88 (22)
Male Sex	369,455 (65)	125,747 (66)
Race		
White	479,428 (85)	161,567 (85)
Black	65,726 (12)	21,363 (11)
Medical History		
History of diabetes	190,280 (34)	66,792 (35)
History of hypertension	419,803 (74)	142,620 (75)
History of dyslipidemia	344,758 (61)	116,511 (61)
Current or recent smoker	191,638 (34)	62,191 (33)
History of chronic lung disease	67,370 (14)	716 (11)
Current dialysis	14,153 (3)	4,902 (3)
History of MI	140,878 (25)	47,419 (25)
History of HF	70,925 (13)	23,972 (13)
Prior PCI	142,900 (25)	50,279 (26)
Prior CABG	76,462 (14)	24,435 (13)
History of atrial fibrillation	44,164 (8)	18,148 (10)
Prior cerebrovascular disease	68,891 (12)	22,832 (12)
Prior peripheral arterial disease	52,660 (9)	15,167 (8)
Presentation		
Presentation after cardiac arrest	22,368 (4)	7,090 (4)
In cardiogenic shock	22,095 (4)	6,688 (4)
In HF	72,621 (13)	22,619 (12)
Heart rate, mean (SD), beats/min	84 (24)	84 (24)
SBP at presentation, mean (SD), mm Hg	146 (35)	148 (36)
Presentation ECG findings		
STEMI	117,078 (39)	73,136 (38)
New or presumed new		
ST depressions	219,648 (39)	19,261 (10)
T-wave inversions	64,294 (11)	12,918 (7)
Transient ST-segment elevation lasting <20 min	43,873 (8)	1,667 (1)
Initial laboratory values		
Troponin ratio, mean (IQR)	2.5 (0.50-16.3)	3.5 (0.78-20.0)
Creatinine, mean (SD), mg/dL	1.3 (1.2)	1.3 (1.2)
Creatinine clearance, mean (SD), mL/min	85 (43)	85 (42)
Hemoglobin, mean (SD), g/dL	14 (2)	14 (2)

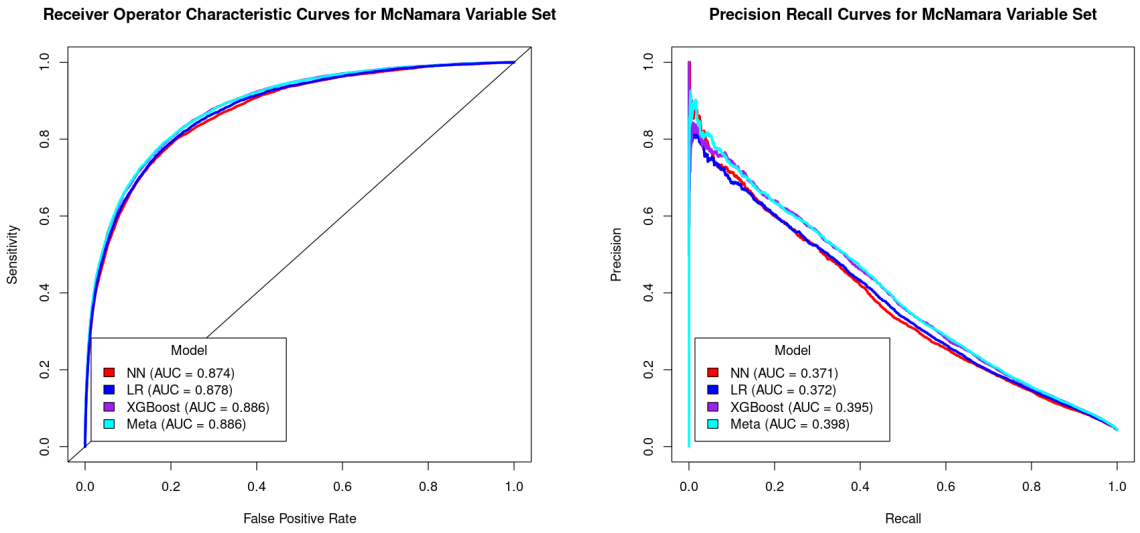
Table 5.4: Performance characteristics of models for predicting in-hospital mortality in acute myocardial infarction

Characteristic	Logistic regression	LASSO	Neural network	XGBoost	Meta-classifier
Variables included in the model of McNamara et al [1]					
Model performance metrics					
AUROC, (95% CI)	0.878 (0.875, 0.881)	0.874 (0.870, 0.879)	0.874 (0.870, 0.878)	0.886 (0.882, 0.890)	0.886 (0.882, 0.890)
Precision-recall AUC	0.372	0.367	0.371	0.395	0.398
F score	0.415	0.408	0.411	0.432	0.432
Sensitivity	0.42 (0.41, 0.43)	0.43 (0.42, 0.45)	0.41 (0.40, 0.42)	0.44 (0.43, 0.45)	0.43 (0.42, 0.44)
Specificity	0.97 (0.97, 0.97)	0.97 (0.97, 0.97)	0.97 (0.97, 0.97)	0.97 (0.97, 0.97)	0.98 (0.97, 0.98)
PPV	0.41 (0.40, 0.42)	0.38 (0.37, 0.39)	0.41 (0.40, 0.42)	0.42 (0.41, 0.43)	0.44 (0.43, 0.45)
NPV	0.97 (0.97, 0.97)	0.97 (0.97, 0.98)	0.97 (0.97, 0.97)	0.98 (0.97, 0.98)	0.97 (0.97, 0.98)
Brier Score					
Reliability, mean (SD), $\times 10^{-6}$	28.4 (9.2)	96.3 (16.5)	224.0 (26.1)	9.5 (3.8)	2.3 (2.1)
Resolution, mean (SD), $\times 10^{-3}$	5.6 (0.1)	5.5 (0.1)	5.4 (0.1)	5.8 (0.1)	5.9 (0.1)
Uncertainty	0.04	0.04	0.04	0.04	0.04
Overall $\times 10^{-2}$	3.52	3.54	3.56	3.49	3.48
Expanded variables included from the CP-MI registry					
Model performance metrics					
AUROC, (95% CI)	0.888 (0.884, 0.892)	0.886 (0.882, 0.890)	0.885 (0.881, 0.889)	0.898 (0.894, 0.902)	0.899 (0.895, 0.903)
Precision-recall AUC	0.421	0.415	0.406	0.451	0.453
F-score	0.436	0.436	0.428	0.458	0.459
Sensitivity	0.47 (0.45, 0.48)	0.42 (0.41, 0.43)	0.43 (0.42, 0.44)	0.45 (0.44, 0.47)	0.43 (0.42, 0.44)
Specificity	0.97 (0.97, 0.97)	0.98 (0.98, 0.98)	0.97 (0.97, 0.98)	0.98 (0.98, 0.98)	0.98 (0.98, 0.98)
PPV	0.41 (0.40, 0.42)	0.45 (0.44, 0.46)	0.43 (0.42, 0.44)	0.46 (0.45, 0.47)	0.49 (0.48, 0.50)
NPV	0.98 (0.98, 0.98)	0.97 (0.97, 0.98)	0.97 (0.97, 0.98)	0.98 (0.98, 0.98)	0.97 (0.97, 0.98)
Brier Score					
Reliability, mean (SD), $\times 10^{-6}$	229.4 (25.6)	40.6 (10.3)	55.7 (11.2)	6.5 (3.5)	4.3 (2.6)
Resolution, mean (SD), $\times 10^{-3}$	6.0 (0.1)	5.9 (0.1)	5.8 (0.1)	6.4 (0.2)	6.5 (0.2)
Uncertainty	0.04	0.04	0.04	0.04	0.04
Overall $\times 10^{-2}$	3.5	3.49	3.5	3.43	3.42

(95% CI, 0.884-0.892) with the logistic regression model. The XGBoost and meta-classifier models had more accurate predictions at an individual level than logistic regression models, with a lower mean squared prediction error across both sets of variables, but this effect was not observed with the neural network (Figure 5.4).

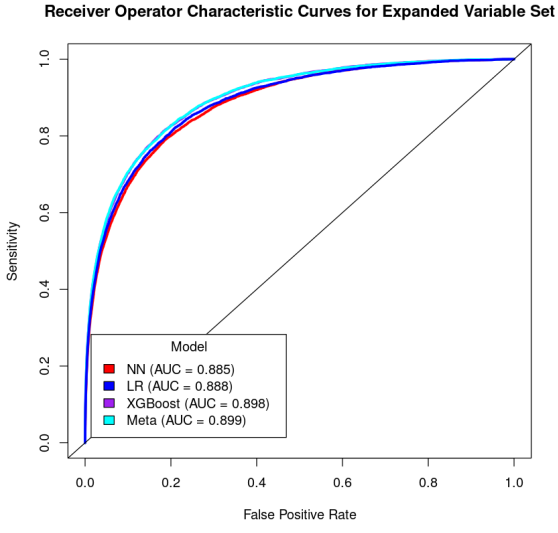
5.3.3 Model Calibration

Of the 3 machine learning models, the XGBoost and the meta-classifier models but not neural network had improvements in calibration slopes compared with logistic regression, when they were applied to a limited or an expanded set of variables (Figures 5.5, 5.6, 5.7). The components and overall Brier score for the different models are included in Table 5.4.

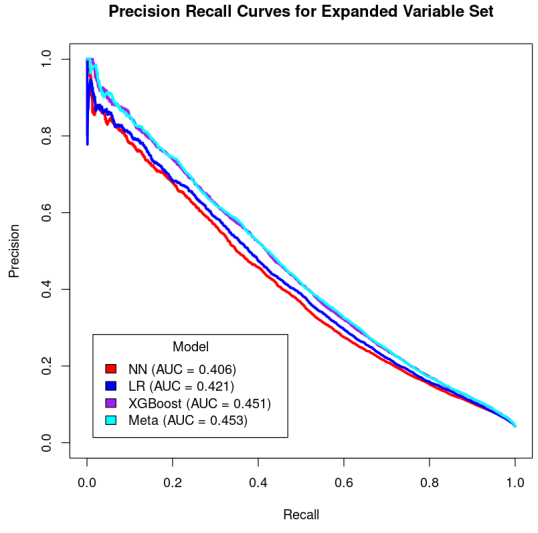


(a)

(b)



(c)



(d)

Figure 5.3: Receiver Operator Characteristic and Precision Recall Curves for each model and each variable set.

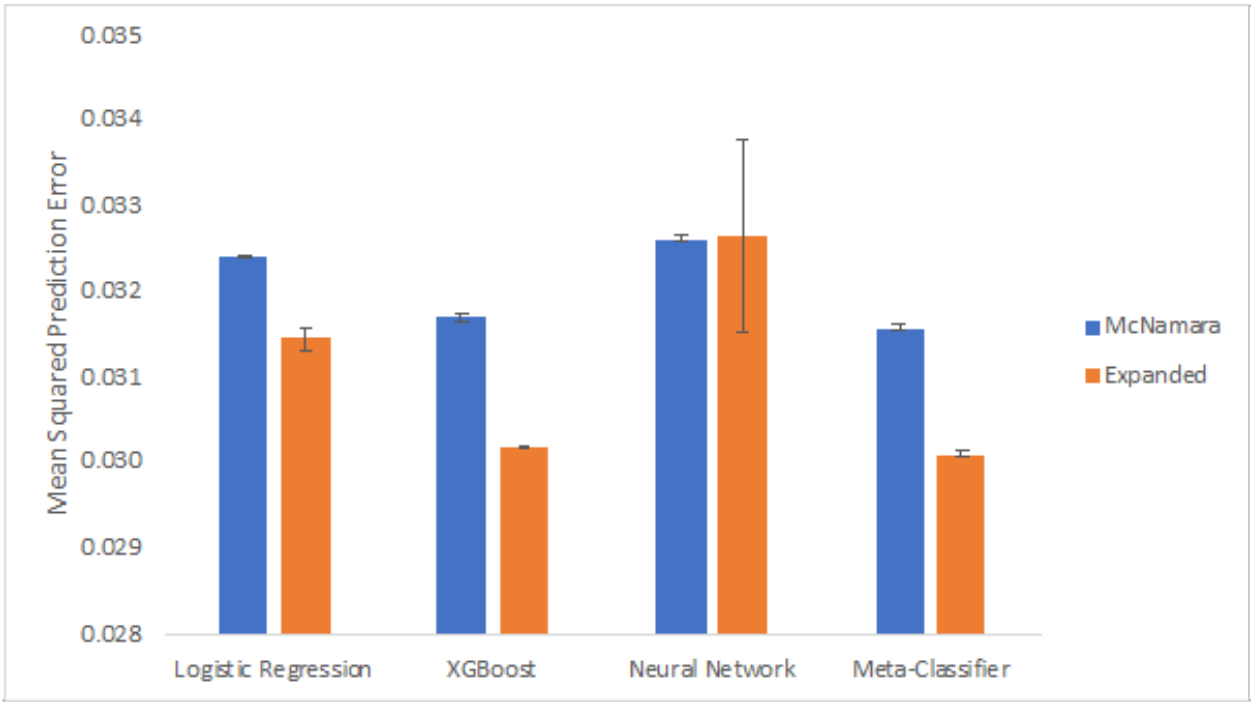


Figure 5.4: Mean Squared Prediction Error of Machine Learning Models Compared With Logistic Regression. The mean squared prediction error for all machine learning models was lower than logistic regression applied to the same set of variables, including the variables used by the current standard [1] and all variables available in the Chest pain-MI registry.

Models with lower values of reliability indicate higher agreement between predicted and observed risk and therefore have better performance. Even with the limited set of model variables, the mean (SD) reliability measure of the meta-classifier ($2.3\ [2.1] \cdot 10^{-6}$) and XGBoost models ($9.5\ [3.8] \cdot 10^{-6}$) but not the neural network ($224.0\ [26.1] \cdot 10^{-6}$) were smaller (and therefore more accurate) compared with the logistic regression model ($28.4\ [9.2] \cdot 10^{-6}$). The machine learning models also had significantly greater resolution (higher range of accurate prediction across the spectrum of risk) than the model based on logistic regression. The highest mean (SD) resolution was found in the meta-classifier ($5.9\ [0.1] \cdot 10^{-3}$) and XGBoost ($5.8\ [0.1] \cdot 10^{-3}$) models followed by the logistic regression model ($5.6\ [0.1] \cdot 10^{-3}$) and the neural network ($5.4\ [0.1] \cdot 10^{-3}$).

All 3 machine models more accurately classified patients in clinically relevant categories of risk. In shift tables, predicted risk across each of the machine learning models (<1%, 1%-5%,

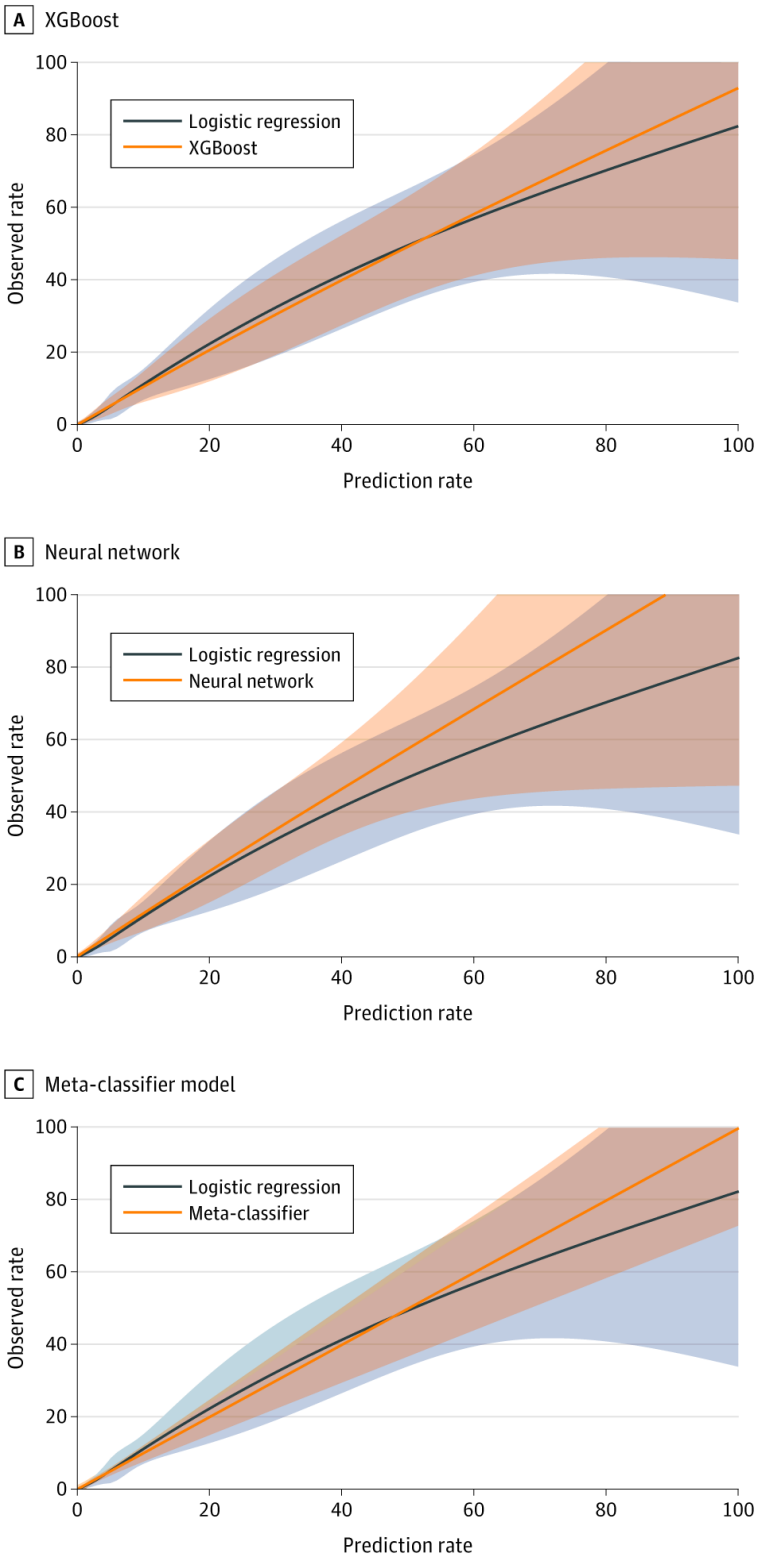


Figure 5.5: Extreme gradient boosting model (XGBoost) (A), neural network (B), and meta-classifier model (C), using the 29-variable input used in the development of the model by McNamara et al. [1]. The shaded areas denote standard error of the calibration.

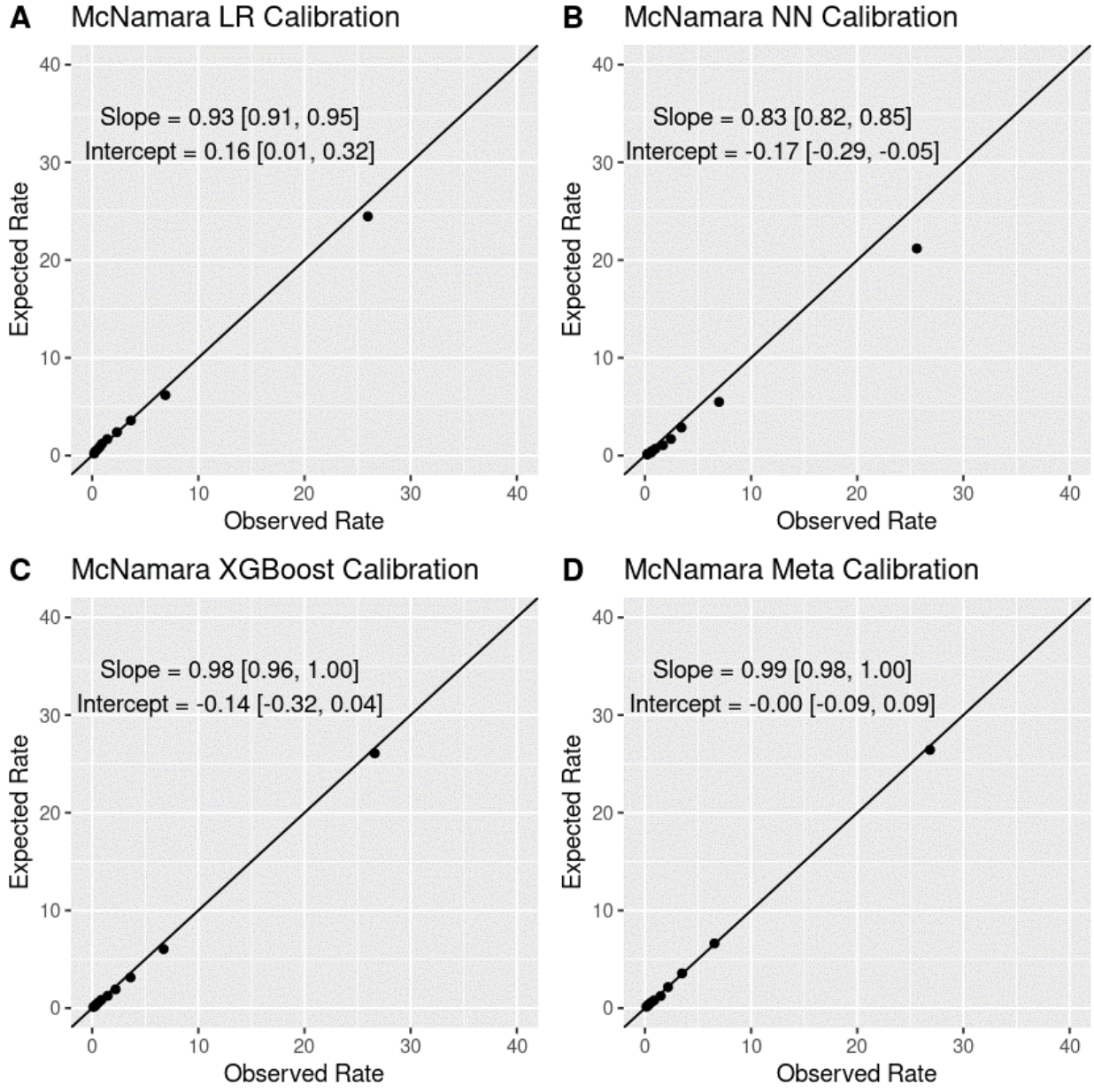


Figure 5.6: Calibration of Models Developed Using Limited Number of Variables Included in the Current Standard [1]. Calibration curves for logistic regression (LR, A), Neural Network (B), XGBoost (C) and Meta-Classifier (D) models for validation cohort predictions. Slope of 1 represents perfect model calibration with values greater than 1 suggesting overestimation of risk and less than 1 suggesting underestimation of risk.

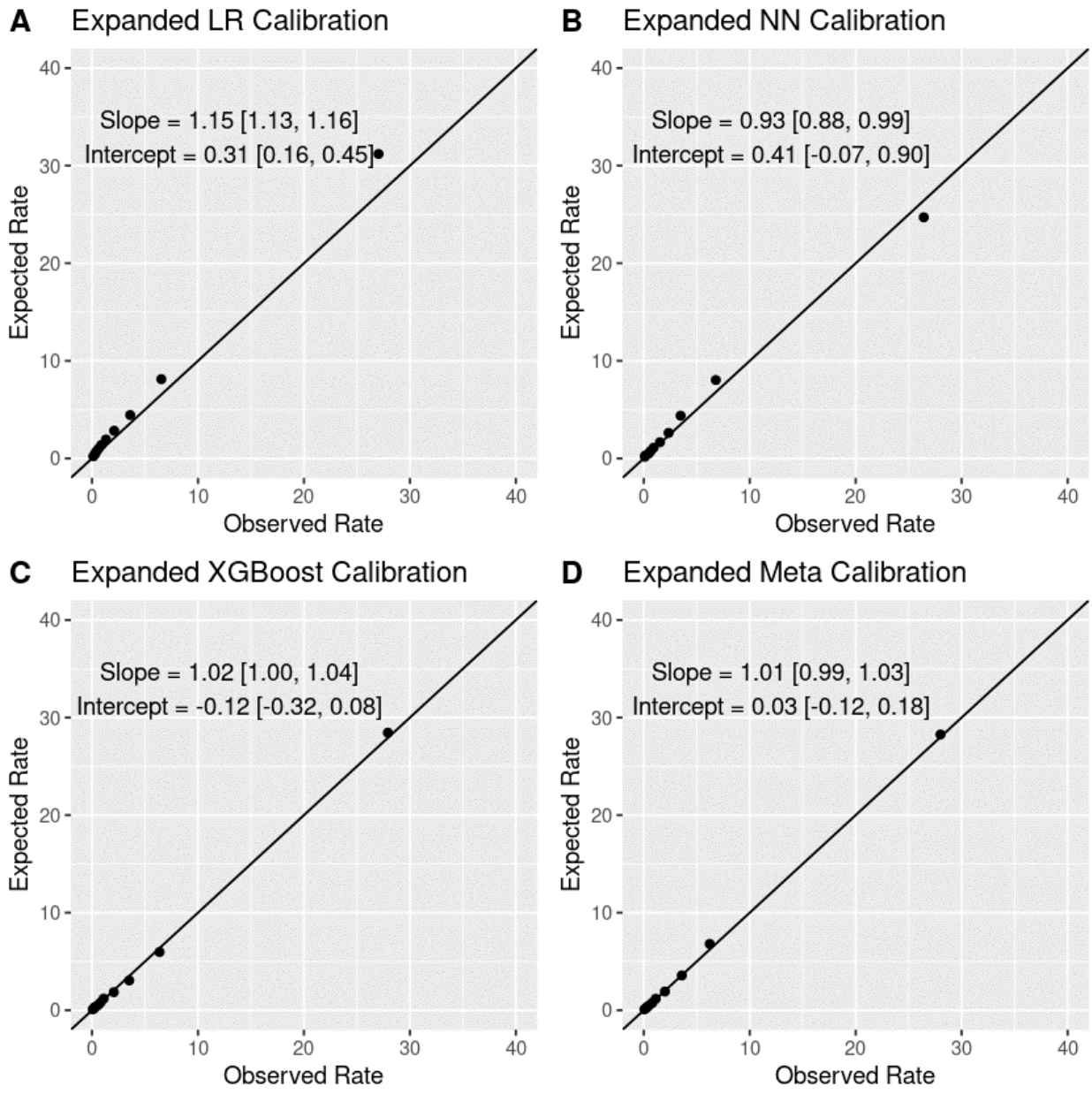


Figure 5.7: Calibration of Models Developed Using Expanded Number of Variables Included in the Chest Pain-MI Registry. Calibration curves for logistic regression (LR, A), Neural Network (B), XGBoost (C) and Meta-Classifier (D) models for validation cohort predictions. Slope of 1 represents perfect model calibration with values greater than 1 suggesting overestimation of risk and less than 1 suggesting underestimation of risk.

and >5%) were individually compared against the predicted risk categories across logistic regression models. In these analyses, individuals with a predicted risk that was discordant between 1 of the machine learning methods and logistic regression was evaluated against the actual rate of observed events in the group. Each of the 3 machine learning models more accurately identified the actual rate of mortality for a group of patients when discordance was found. For example, among patients predicted to be at low risk based on the meta-classifier or XGBoost models and low, moderate, or high risk based on logistic regression, a negligible difference was found in the mortality rate among those also predicted to be at low risk by logistic regression (mortality rate, 0.3%) or moderate or high risk (mortality rate, 0.5%), despite predicted mortality risk of greater than 1% by logistic regression. In contrast, patients who were at low risk based on logistic regression had an observed mortality rate of 2.2% if at moderate or high risk based on the meta-classifier model. A similar pattern was observed for all, compared with logistic regression models applied to the same data.

Notably, 30,836 of 121,839 individuals (25%) deemed to be at moderate or high risk by logistic regression were more appropriately classified as being at low risk by the meta-classifier, consistent with their actual observed rates of mortality after AMI, even with models using the same model inputs. Moreover, 2,951 of 68,645 individuals (4%) who were deemed to be at low risk by logistic regression were reclassified as moderate to high risk (Table 5.5). There was a similar reclassification of risk in the XGBoost model, which reclassified 32,393 medium-high risk individuals (27%) based on logistic regression to low risk, which is more consistent with the observed rates. Furthermore, 3,452 patients (5%) classified as low risk by logistic regression were reclassified as medium-high risk by XGBoost (Table 5.5). The reclassification of low-risk individuals to moderate-high risk was also not consistent with observed events with machine learning models. The models based on expanded variables more accurately categorized patient risk than the limited set of variables, with machine learning models offering additional calibration of risk for the same set of variables. The observations on reclassification were consistent in sensitivity analyses using different risk

Table 5.5: Performance of the XGBoost and meta-classifier models compared with logistic regression

	Expanded LR, No. of patients (% observed mortality)			
Model	<1%	1%-5%	>5%	All
XGBoost vs LR				
Expanded XGBoost				
<1%	65,193 (0.27)	31,971 (0.65)	422 (1.18)	97,586 (0.40)
1%-5%	3,384 (0.95)	44,486 (2.21)	13,155 (3.91)	61,025 (2.51)
>5%	68 (2.94)	2,899 (6.21)	28,906 (20.79)	31,873 (19.42)
All	68,645 (0.30)	79,356 (1.73)	42,483 (15.37)	190,484 (4.26)
Meta-classifier vs LR				
Expanded Meta-classifier				
<1%	65,694 (0.27)	30,661 (0.65)	175 (0.00)	96,530 (0.39)
1%-5%	2,930 (1.06)	45,726 (2.17)	9,033 (3.55)	57,689 (2.33)
>5%	21 (0.00)	2,969 (6.03)	33,275 (18.66)	36,265 (17.61)
All	68,645 (0.30)	79,356 (1.73)	42,483 (15.37)	190,484 (4.26)

Table 5.6: Area under the receiver operator characteristic curve for the 5-fold multiple imputation.

Model	Models Constructed using Limited variables	Models Constructed using Expanded variables
Logistic Regression	0.877	0.888
Neural Network	0.874	0.886
XGBoost	0.885	0.897
Meta-classifier	0.885	0.898

thresholds (<1.5%, 1.5%-3%, and >3%), wherein patients reclassified by XGBoost and meta-classifier but not neural networks had observed event rates consistent with the classified groups (Table 5.6).

The improvements in calibration were consistent across imputation strategies for missing variables, including the mode imputation and 5-fold multiple imputation strategies (Table 5.7). Furthermore, in an additional sensitivity analysis that included most patients by using a smaller number of features, XGBoost achieved an AUROC of 0.899 (95% CI, 0.895-0.904) and meta-classifier achieved an AUROC of 0.901 (95% CI, 0.896-0.905), largely similar to

Table 5.7: Model calibration slopes in patient subgroups.

Group	Logistic regression	Neural network	XGBoost	Metaclassifier
Overall	0.93 [0.91, 0.95]	0.83 [0.82, 0.85]	0.98 [0.96, 1.00]	0.99 [0.98, 1.00]
Age in years				
18-44	0.90 [0.87, 0.93]	0.81 [0.77, 0.84]	0.98 [0.95, 1.00]	0.97 [0.94, 1.00]
45-64	0.93 [0.92, 0.94]	0.83 [0.82, 0.85]	0.97 [0.96, 0.98]	0.98 [0.96, 1.00]
≥65	0.94 [0.91, 0.97]	0.83 [0.81, 0.86]	0.99 [0.96, 1.03]	1.00 [0.99, 1.01]
Sex				
Male	0.94 [0.92, 0.95]	0.84 [0.82, 0.85]	0.98 [0.97, 1.00]	0.99 [0.98, 1.01]
Female	0.92 [0.89, 0.95]	0.82 [0.80, 0.85]	0.97 [0.94, 1.00]	0.97 [0.96, 0.99]
Race/ethnicity				
White	0.93 [0.92, 0.95]	0.83 [0.82, 0.84]	0.98 [0.96, 1.00]	0.99 [0.97, 1.00]
Black	0.95 [0.89, 1.00]	0.86 [0.83, 0.90]	1.00 [0.94, 1.06]	1.01 [0.97, 1.04]

logistic regression (AUROC, 0.890; 95% CI, 0.886-0.895).

5.3.4 Subgroup Analyses

In assessments of subgroups of age, sex, and race, logistic regression models were less well calibrated in patients who were younger and White compared with older (calibration slope, 0.90; 95% CI, 0.87-0.93 in those 18-44 years of age vs 0.94; 95% CI, 0.91-0.97 in ≥65 years of age) and Black patients (calibration slope, 0.93; 95% CI, 0.92-0.95 in White patients vs 0.95; 95% CI, 0.89-1.00 in Black patients). In contrast, the meta-classifier model was well calibrated across patient groups. Of the other models, XGBoost, but not the neural network, was better calibrated in patient subgroups relative to logistic regression.

5.4 Discussion

In this cohort study, in a large national registry of patients with AMI, machine learning models did not substantively improve discrimination of in-hospital mortality compared with models based on logistic regression. However, 2 of these models were associated with improvement in the resolution of risk over logistic regression and with improved classification of patients across risk strata, particularly among those at greatest risk for adverse outcomes. One of these models, XGBoost, is interpretable and represents the collection of individual-

ized decision trees that address complex relationships among variables. The second model, meta-classifier, which aggregated information from multiple machine learning models, also had better model calibration than logistic regression. Despite almost no improvements in discrimination, these models led to reclassification of 1 in every 4 patients deemed moderate or high risk for death with logistic regression as low risk, which was more consistent with their observed event rates. However, machine learning models were not uniformly superior to logistic regression, and a neural network model had worse performance characteristics than a logistic regression model based on the same inputs.

The study builds on prior studies [113, 114, 115, 116, 117, 118, 119, 120, 121, 123] that used machine learning in predicting AMI outcomes. Most of these studies found improved prediction with applications of classification algorithms of varying complexity. However, they were limited by smaller patient groups, with limited generalizability in the absence of standard data collection. In a large national registry with standardized data collection across more than 1000 hospitals, improvements in risk prediction for in-hospital mortality with machine learning models were small and likely do not meet the threshold to be relevant for clinical practice.

However, there are notable aspects of the new models. Without the cost of collecting additional data or a reliance on literature review or expert opinion for variable selection, the models achieved similar model performance characteristics as logistic regression, which is relevant for predictive modeling in clinical areas where disease mechanisms are not well defined. Moreover, 2 of the 3 models were much better calibrated across patient groups based on age, sex, race, and mortality risk and were therefore better suited for risk prediction despite only modest improvement in overall accuracy. Notably, this improvement in predictive range occurred in critical areas by accurately reclassifying individuals at high risk to categories more accurately reflecting their risk. A focus on traditional measures of accuracy underperform in capturing the scale of these improvements because the events are rare and model discrimination is driven by patients not experiencing the mortality event

[134, 135]. In this respect, the Brier score offers a more comprehensive assessment of model performance, combining model discrimination and calibration. The Brier score represents the mean squared difference between the predictions and the observed outcome. A perfect model has a Brier score of 0, and when 2 models are compared, a smaller Brier score indicates better model performance. Both XGBoost and meta-classifier models had scores that were lower than the logistic models by several multiples of the SDs of the score. Given the only marginal improvements in model discrimination, the lower Brier scores reflect the improved calibration noted in the calibration slope and shift tables.

Of note, 1 of the models that performs well is interpretable because it represents a collection of decision trees, thereby ensuring transparency in its application that specifically addresses the concerns with black-box machine learning models. Furthermore, although their development is computationally intensive, their eventual deployment at an individual patient level does not require substantial computational resources. Therefore, the clinical adoption of these models likely depends on whether their gains in prediction accuracy are worth their computationally intensive development and lack of interpretability. Some machine learning models may, therefore, have greater clinical utility in higher-dimensional data where they can uncover complex relationships among variables [136, 137, 138] and of variables with outcomes but only provide limited gains in relatively low-dimension registry data. Furthermore, not all machine learning performed well. The neural network model developed using all available variables in the registry was inferior to the logistic regression based on similar inputs, indicating that not all machine learning models are uniformly superior to traditional methods of risk prediction.

5.4.1 Limitations

This study has limitations. First, although the CP-MI registry captures granular clinical data on patients with AMI, relevant information, such as duration of comorbidities and control of chronic diseases (besides diabetes), was not captured in the registry and is, therefore, not included in the assessment. Furthermore, certain prognostic characteristics of the

patients' general health are not included [139, 140]. Second, although models are based on sound mathematical principles, the study does not identify whether the excess risk identified with the models is modifiable. Third, shift tables judge classification across risk thresholds but may overemphasize small effects around thresholds. However, other calibration metrics also suggest more precise risk estimation by XGBoost and the meta-classifier among patients classified as being at high risk by logistic regression. Fourth, the study was not externally validated. Therefore, although the observations may be generalizable to the data in the NCDR CP-MI Registry, they may not apply to patients not included or hospitals not participating in the registry. However, because the data are collected as a part of routine clinical care at a diverse set of hospitals, other hospitals that collect similar data could likely apply these modeling strategies.

5.5 Conclusions

In a large national registry, machine learning models were not associated with substantive improvement in the discrimination of in-hospital mortality after AMI, limiting their clinical utility. However, compared with logistic regression, the models offered improved resolution of risk for high-risk individuals.

6. USE OF MECHANICAL CIRCULATORY SUPPORT DEVICES AMONG PATIENTS WITH ACUTE MYOCARDIAL INFARCTION COMPLICATED BY CARDIOGENIC SHOCK*

With this chapter, we shift our focus from solely predicting outcomes, to instead investigating reasons for those outcomes. This chapter and the next look at a particular type of intervention for patients with acute myocardial infarction complicated by cardiogenic shock. In this disease state, the damaged heart is unable to adequately supply blood to the body's vital organs. To treat this condition, one approach is to use a device to aid the heart in driving blood. However, evidence supporting the use of these devices is limited. These next chapters aim to use machine learning to elucidate their effectiveness. In this chapter, we first examine trends in the utilization of these devices.

6.1 Introduction

Intra-aortic balloon pumps (IABPs) have been the mainstay of mechanical circulatory support (MCS) for patients with cardiogenic shock in the setting of acute myocardial infarction (AMI) [141]. However, randomized clinical trial (RCT) data [142, 143] and subsequent meta-analyses [144, 145] have reported no clinical benefit from routine IABP use in patients with AMI complicated by cardiogenic shock. Impella devices (intravascular microaxial left ventricular assist devices [LVADs]), which offer greater improvement in hemodynamic parameters compared with IABPs [146], received US marketing clearance in 2008 for providing partial circulatory support for up to 6 hours using an extracorporeal bypass control unit and providing circulatory support during procedures not requiring cardiopulmonary bypass [147]. Studies through 2012 showed a substantial uptake of these devices, from 4.6 per mil-

*This chapter is reprinted with permission from "Use of Mechanical Circulatory Support Devices Among Patients With Acute Myocardial Infarction Complicated by Cardiogenic Shock" by Dhruva, S.S., Ross, J.S., Mortazavi, B.J., Hurley, N.C., Krumholz, H.M., Curtis, J.P., Berkowitz, A.P., Masoudi, F.A., Messenger, J.C., Parzynski, C.S. and Ngufor, C.G., Girotra, S., Amin, A.P., Shah, N.D., Desai, N.R., 2021. JAMA Network Open. Copyright 2021 by Dhruva, S.S. et al. under the CC-BY license.

lion hospital discharges in 2007 to 138 per million discharges in 2012 [148, 149], despite the absence of demonstrated benefits for hard clinical end points in RCTs [146, 150]. National Cardiovascular Data Registry (NCDR) records through September 2013 showed that use of MCS devices other than IABP was clustered around a relatively small number of hospitals but did not increase [149].

Despite the substantial risk of death associated with cardiogenic shock [151] and the relatively high cost of some MCS devices [148, 152], the temporal and contemporary trends in MCS device use have not been examined in terms of detailed demographic and clinical characteristics abstracted from medical records, such as coronary anatomy. Furthermore, previous studies have focused on IABPs and other MCS devices, providing no granularity about other MCS therapies such as extracorporeal membrane oxygenation (ECMO). Understanding changes in use as well as the patients likely to receive MCS devices and the hospitals that are likely to use these devices is particularly important given the recent safety concerns about intravascular microaxial LVADs [152, 153]. In this retrospective cross-sectional study, we collected data from 2 national US registries (of the American College of Cardiology NCDR) to examine trends in the use of MCS devices, providing greater granularity of the clinical characteristics and device type than previous studies, among a large cohort of patients who underwent percutaneous coronary intervention (PCI) for AMI complicated by cardiogenic shock. We also examined hospital-level use variation and factors associated with use.

6.2 Methods

The Human Investigation Committee of the Yale University School of Medicine approved the use of a limited data set from the NCDR for research purposes without requiring informed consent because all of the data were deidentified and maintained centrally by the NCDR. This cross-sectional study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline [128].

6.2.1 Data Sources and Study Population

We linked the NCDR CathPCI and Chest Pain-MI registries, both of which have been described previously [154, 155]. In brief, the CathPCI Registry is a voluntary registry of diagnostic cardiac catheterizations and PCIs performed in the US. More than 1500 hospitals across the US participate in this program and are required to submit data on all PCI procedures. The Chest Pain-MI Registry includes patients with AMI. The CathPCI Registry, version 4.4, identifies whether a patient received an IABP or any other MCS device and the timing of MCS. Version 2.4.2 of the Chest Pain-MI Registry data collection form, released in the third quarter of 2015, includes the type of MCS device.

We identified all patients who underwent PCI for AMI complicated by cardiogenic shock between October 1, 2015, and December 31, 2017, and had available data in both registries. We included individuals in the Chest Pain-MI Registry who had cardiogenic shock at first medical contact or as an in-hospital event or individuals in the CathPCI Registry who had cardiogenic shock within 24 hours prior to the PCI, at the start of the PCI, or as an intra- or postprocedure event. Cardiogenic shock was defined in both registries as systolic blood pressure less than 90 mm Hg and/or cardiac index lower than 2.2 L/min/m2 for at least 30 minutes that was secondary to ventricular dysfunction and/or a requirement for parenteral inotropic or vasopressor therapy or MCS devices to support blood pressure and cardiac index [156]. For patients who underwent multiple PCIs during the hospitalization, we included data from only the initial PCI.

6.2.2 Hemodynamic Support and Covariates

We categorized patients according to the hemodynamic support that they received. The CathPCI Registry details if a patient received an IABP or a different MCS device. The Chest Pain-MI Registry details if a patient received an IABP, intravascular microaxial LVAD, TandemHeart (CardiacAssist Inc), ECMO, LVAD, or other device. The Chest Pain-MI Registry allows documentation of only 1 MCS device per patient. Therefore, by linking the

2 registries, we could identify the MCS devices used (Chest Pain-MI Registry) in combination with IABPs (CathPCI Registry). Patients who did not receive any MCS device composed the medical therapy only group.

Patient-level covariates were patient demographic characteristics, medical history, and clinical presentation. Hospital-level covariates were number of beds, location, type (government, private, or university), presence of teaching program, and mean annual PCI volume. For continuous values with missing values, the mean was imputed. For binary (yes or no) variables, all missing variables were coded as no; for categorical variables, all missing variables were coded as no or other (if a no category did not exist).

6.2.3 Statistical Analysis

We characterized overall MCS device use, including for specific sociodemographic and clinical subgroups (age, sex, race, insurance status, ST-segment elevation MI [STEMI] or non-STEMI, cardiac arrest or no arrest, and transfer status). We examined trends in the use of hemodynamic support by calendar quarter using the Cochrane-Armitage test to determine the significance of changes over time.

We performed multivariable logistic regression to identify independent variables associated with MCS device use compared with medical therapy among all patients with AMI complicated by cardiogenic shock, accounting for clustering by facility (ie, accounting for the possible associations among patients who received care at a given hospital such that the observations were not independent). The model included demographic variables (age, sex, race, and insurance status), comorbidities (previous PCI, previous coronary artery bypass graft [CABG], and peripheral artery disease), clinical presentation variables (cardiac arrest at first medical contact or during hospitalization, STEMI, anterior infarction, left main or proximal left anterior descending coronary artery [LAD] disease, and left ventricular ejection fraction), and hospital variables (number of beds, location, type, teaching program, and mean annual PCI volume). Using the same model, we performed an additional multivariable logistic regression to examine the odds of a patient receiving an intravascular microaxial

LVAD compared with an IABP, restricting the analyses to patients with AMI complicated by cardiogenic shock who received either an IABP or intravascular microaxial LVAD only.

We examined hospital-level variation in MCS device use among hospitals that cared for at least 10 patients with AMI complicated by cardiogenic shock during the study period. We calculated a median odds ratio (OR) by building a generalized linear mixed model with random hospital intercepts. The median OR (always $\geq$1) was derived from the estimate of the variance of the random intercept of the model [157]. Conceptually, the median OR represents the relative odds for 2 identical patients receiving an MCS device at 1 randomly selected hospital vs another randomly selected hospital. A median OR of 1.0 indicates no hospital-level variation, whereas a median OR of 2 indicates that the odds of receiving an MCS device are 2-fold higher in 1 randomly selected hospital vs another hospital. Using the same methods, we calculated a hospital-specific median OR for a patient with AMI complicated by cardiogenic shock to receive an intravascular microaxial LVAD.

We compared hospital characteristics (number of beds, location, type, teaching program, and mean annual PCI volume) by quartiles of MCS device use. We also compared the characteristics of hospitals that used at least 1 intravascular microaxial LVAD vs hospitals that did not. Among hospitals that used at least 1 intravascular microaxial LVAD, we compared the characteristics by tertiles. We used χ^2 test for categorical variables and Kruskal-Wallis test for non-normally distributed continuous variables.

All statistical analyses were 2-sided, with an $\alpha = .05$ for statistical significance. All analyses were conducted in R, version 3.6.0 (R Foundation for Statistical Computing), with packages clubSandwich 0.4.2 [158]; ggplot2, version 3.2.1 [159]; DescTools 0.99.34[160]; and lubridate 1.7.4 [161]. Data were analyzed from October 2018 to August 2020.

6.3 Results

6.3.1 MCS Device Use and Change Over Time

Among 28,304 patients with AMI complicated by cardiogenic shock who received PCI at 928 hospitals during the study period, the mean (SD) age was 65.4 (12.6) years and 18 968 were men (67.0%). Overall, 12,077 patients (42.7%) received an MCS device and 16,227 (57.3%) received medical therapy only during the hospitalization. Of the 12,077 patients who received an MCS device, 1768 (14.6%) received an intravascular microaxial LVAD only, 8471 (70.1%) received an IABP only, 5 (0%) received TandemHeart, 182 (1.5%) received ECMO, 23 (0.2%) received an LVAD, 276 (2.3%) received both an IABP and intravascular microaxial LVAD, 4 (0%) received an IABP and TandemHeart, 138 (1.1%) received an IABP and ECMO, 17 (0.1%) received an IABP and LVAD, and 1193 (9.9%) received another MCS device or a combination of MCS devices (Figure 6.1).

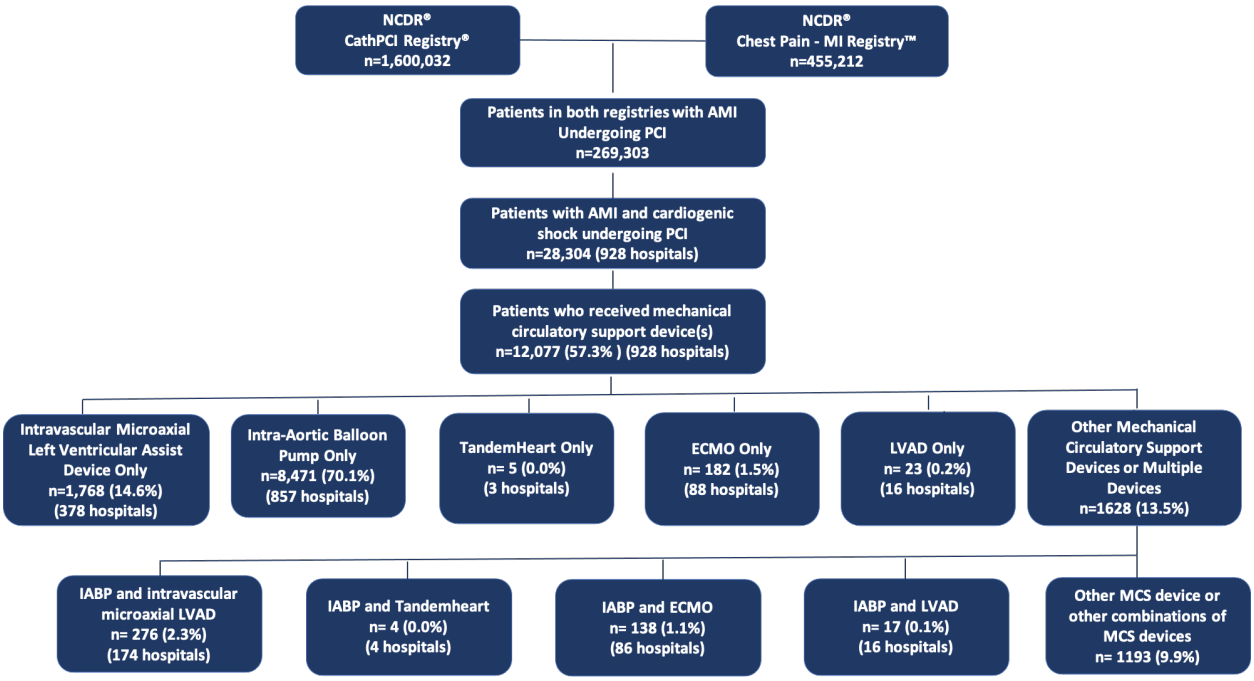


Figure 6.1: Flow Diagram of Patients Undergoing Percutaneous Coronary Intervention for Acute Myocardial Infarction Complicated by Cardiogenic Shock

During the study period, the proportion of patients who used any MCS device remained similar from October through December 2015 to October through December 2017 (from 41.9% to 43.1%; P = .07) (Figure 6.2). A significant increase in the use of intravascular microaxial LVADs (either alone or in combination with IABPs) was found (from 4.1% to 9.8%; P < .001) during this period along with a corresponding decrease in the percentage of patients who received IABPs either alone or in combination with other MCS devices (from 34.8% to 30.0%; P < .001). When limited to patients receiving any MCS, the use of intravascular microaxial LVADs increased from 9.9% to 20.6%, whereas IABP use decreased from 83.1% to 73.2% (Figure 6.2).

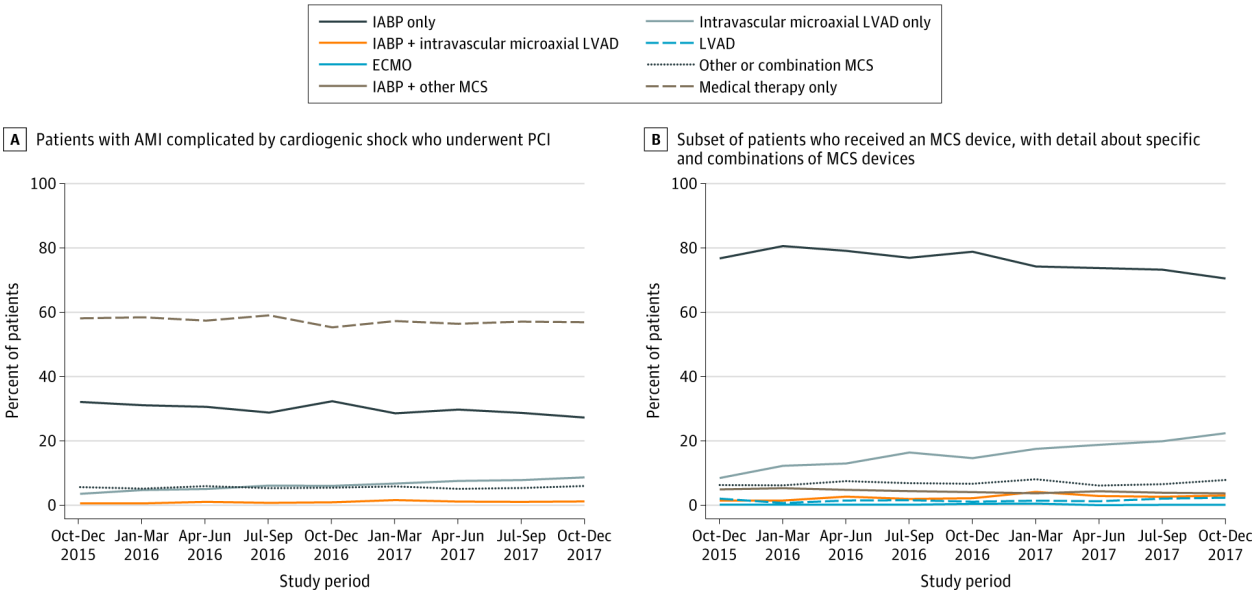


Figure 6.2: Quarterly Use of Mechanical Circulatory Support (MCS) Devices for Patients Who Underwent Percutaneous Coronary Intervention (PCI) for Acute Myocardial Infarction (AMI) Complicated by Cardiogenic Shock From October 2015 to December 2017 at Hospitals Participating in the National Cardiovascular Data Registry CathPCI and Chest Pain-MI Registries

6.3.2 Hospital-Level Variation in MCS Device Use

Of the 928 hospitals included in the study, 521 (56.1%) did not use any intravascular microaxial LVADs for patients with AMI complicated by cardiogenic shock. Among hospitals with at least 10 cases of AMI with cardiogenic shock during the study period, a significant variation in MCS device use was observed (Figure 6.3). The median (interquartile range [IQR]) proportion of patients who received an MCS device at the hospital level was 42% (30%-54%; range 4%-94%). The median (IQR) proportion of patients who received any intravascular microaxial LVAD was 1% (0%-10%; range, 0%-83%).

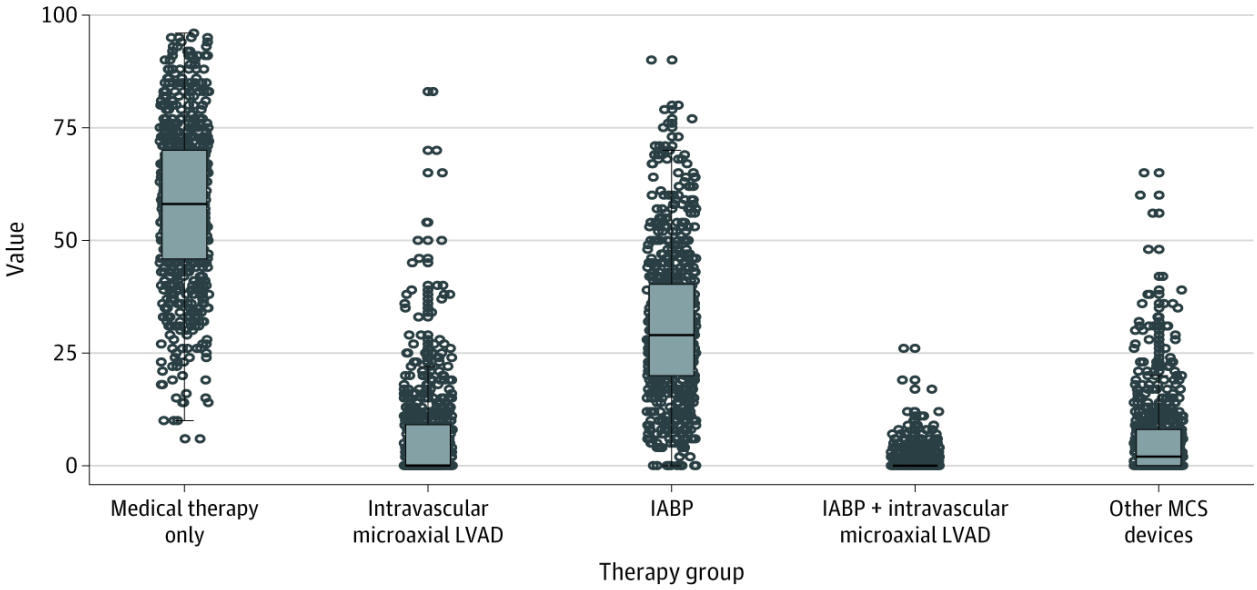


Figure 6.3: Quarterly Use of Mechanical Circulatory Support (MCS) Devices for Patients Who Underwent Percutaneous Coronary Intervention (PCI) for Acute Myocardial Infarction (AMI) Complicated by Cardiogenic Shock From October 2015 to December 2017 at Hospitals Participating in the National Cardiovascular Data Registry CathPCI and Chest Pain-MI Registries

The hospital-specific median OR for use of any MCS device over the study period was 1.79 (95% CI, 1.71-1.86). This OR indicates that the odds of receiving an MCS device were 1.79-fold higher in 1 randomly selected hospital vs another. The hospital-specific median OR

Table 6.1: Hospital characteristics after stratification by quartiles of use of any mechanical circulatory support (MCS) device

Characteristic	Any use of MCS device Quartile 1 (n = 230)	Quartile 2 (n = 233)	Quartile 3 (n = 229)	Quartile 4 (n = 236)	P value
Patients with AMI complicated by cardiogenic shock who underwent PCI and received an MCS device at each hospital, %	<29	29 to <42	≥42 to <55	≥55	NA
Beds, No. (%)					0.005
<200	77 (33.5)	65 (27.9)	47 (20.5)	60 (25.4)	
200-399	100 (43.5)	96 (41.2)	91 (39.7)	93 (39.4)	
400-599	39 (17.0)	37 (15.9)	51 (22.3)	47 (19.9)	
≥600	14 (6.1)	35 (15.0)	40 (17.5)	36 (15.3)	
Location					0.96
Rural	45 (19.6)	40 (17.2)	36 (15.7)	44 (18.6)	
Suburban	79 (34.3)	80 (34.3)	79 (34.5)	81 (34.3)	
Urban	106 (46.1)	113 (48.5)	114 (49.8)	111 (47.0)	
Type					<.001
Government	7 (3.0)	4 (1.7)	3 (1.3)	1 (0.4)	
Private	215 (93.5)	216 (92.7)	195 (85.2)	205 (86.9)	
University	8 (3.5)	13 (5.6)	31 (13.5)	30 (12.7)	
Teaching Program	68 (29.6)	92 (39.5)	104 (45.5)	106 (44.9)	0.001
Annual PCI volume, mean (SD)	482.9 (521.4)	546.0 (458.5)	681.0 (644.6)	584.3 (553.4)	<.001

for use of any intravascular microaxial LVAD only over the study period was 3.33 (95% CI, 3.03-3.63). This OR indicates that the odds of receiving an intravascular microaxial LVAD were 3.33-fold higher in 1 randomly selected hospital vs another.

6.3.3 MCS Device Use by Hospital Characteristics

Among all hospitals that cared for patients with AMI complicated by cardiogenic shock, larger hospitals (≥600 beds) were more likely to be in higher quartiles of MCS device use and smaller ones (≤200 beds) were more likely to be in the lowest quartile of MCS device use (Table 6.1). University hospitals and those with teaching programs were more likely to be in higher quartiles of MCS device use. Hospitals with higher mean annual PCI volumes were more likely to use MCS devices. No significant difference in MCS device use was found across hospitals that were rural, suburban, or urban.

Hospitals that placed at least 1 intravascular microaxial LVAD for patients who under-

Table 6.2: Hospital characteristics after stratification by use of intravascular microaxial left ventricular assist device (LVAD)

Characteristic	No use of intravascular microaxial LVAD (n = 521)	Any (at least 1) use of intravascular microaxial LVAD Tertile 1 (n = 125)	Tertile 2 (n = 135)	Tertile 3 (n = 147)	P value (among tertiles)	P value (no vs any use)
Patients with AMI complicated by cardiogenic shock who underwent PCI and received an intravascular microaxial LVAD at each hospital, %	NA	<7	7 to <15	≥15	NA	NA
Beds, No. (%)					0.5	<.001
<200	179 (34.4)	16 (12.8)	23 (17.0)	31 (21.1)		
200-399	205 (39.3)	54 (43.2)	59 (43.7)	62 (42.2)		
400-599	89 (17.1)	30 (24.0)	24 (17.8)	31 (21.1)		
≥600	48 (9.2)	25 (20.0)	29 (21.5)	23 (15.6)		
Location					0.83	0.003
Rural	105 (20.2)	18 (14.4)	21 (15.6)	21 (14.3)		
Suburban	192 (36.9)	38 (30.4)	38 (28.1)	51 (34.7)		
Urban	224 (43.0)	69 (55.2)	76 (56.3)	75 (51.0)		
Type					0.76	0.17
Government	9 (1.7)	1 (0.8)	3 (2.2)	2 (1.4)		
Private	474 (91.0)	108 (86.4)	120 (88.9)	129 (87.8)		
University	38 (7.3)	16 (12.8)	12 (8.9)	16 (10.9)		
Teaching Program	184 (35.3)	57 (45.6)	69 (51.1)	60 (40.8)	0.22	0.001
Annual PCI volume, mean (SD)	442.1 (498.1)	821.3 (562.7)	753.4 (666.7)	662.8 (471.2)	0.03	<.001

went PCI for AMI complicated by cardiogenic shock were more likely to be large (≥200 beds), be in an urban setting, have a teaching program, and have a higher annual PCI volume (Table 6.2). Across tertiles of hospitals that used intravascular microaxial LVADs, no significant difference was observed in the number of beds, location, type, or presence of teaching program. Hospitals with lower annual PCI volume were more likely to be in the highest tertile of intravascular microaxial LVAD use.

6.3.4 MCS Device Use by Patient Demographic and Clinical Characteristics

In comparing the unadjusted use of intravascular microaxial LVADs only with use of IABPs only within a denominator of all therapies for AMI complicated by cardiogenic shock, men were more likely than women to receive intravascular microaxial LVADs (6.6% vs 5.4%; $P < .01$) and IABPs (30.9% vs 27.9%; $P < .01$) (eFigure 2 in the Supplement). Patients younger than 65 years were more likely to receive intravascular microaxial LVADs than those aged 75 years or older (6.5% vs 5.4%; $P = .002$) (eFigure 3 in the Supplement). Black patients were significantly more likely to receive intravascular microaxial LVADs compared

with patients who were not Black individuals (7.5% vs 6.5%; P = .005) (eFigure 4 in the Supplement). Additional analyses that characterize the use of MCS devices by insurance, type of myocardial infarction, cardiac arrest status, and transfer status are provided in eFigures 5 to 8 in the Supplement.

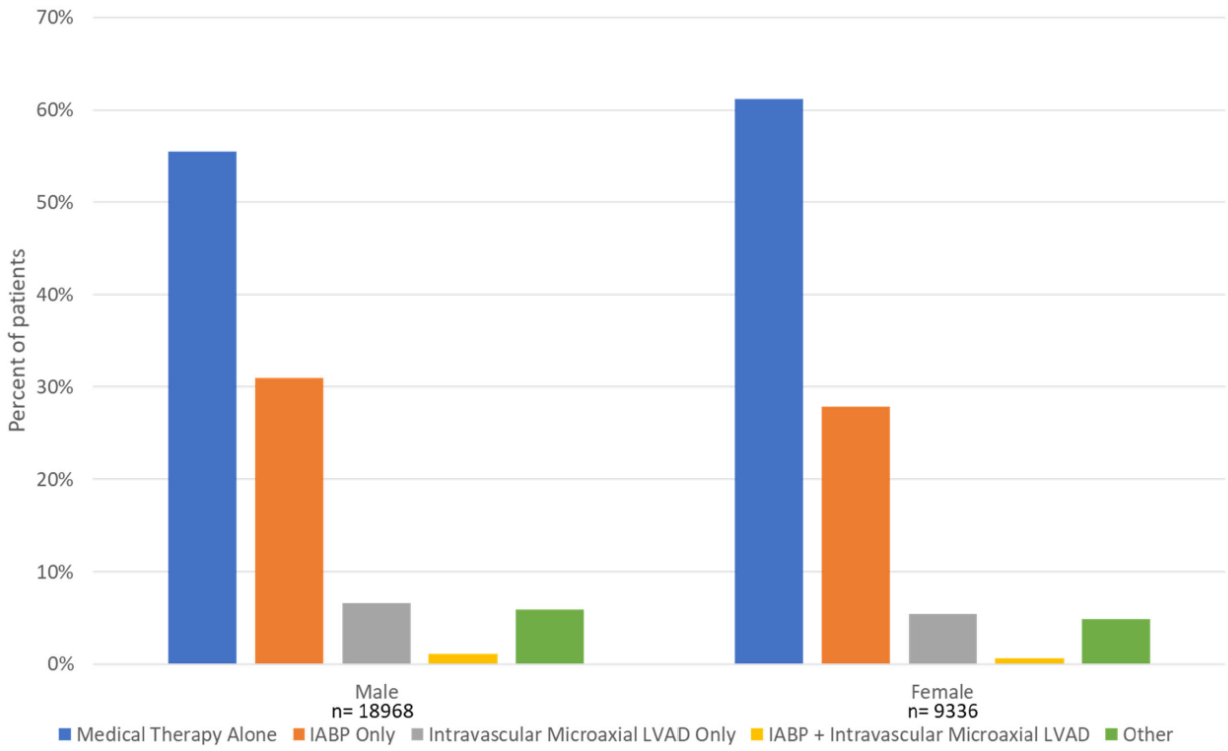


Figure 6.4: Sex Distribution by Therapy of Patients Undergoing Percutaneous Coronary Intervention for Acute Myocardial Infarction Complicated by Cardiogenic Shock from October 1, 2015 – December 31, 2017

6.3.5 Characteristics Associated With MCS Device Use and With Intravascular Microaxial LVAD vs IABP Use

In multivariable regression analysis, female sex (OR, 0.88; 95% CI, 0.83-0.93), the presence of peripheral artery disease (OR, 0.78; 95% CI, 0.71-0.86), and previous CABG (OR, 0.74; 95% CI, 0.67-0.81) were associated with lower odds of receiving any MCS device (Table 6.3). Private insurance (vs no insurance), cardiac arrest at first medical contact or

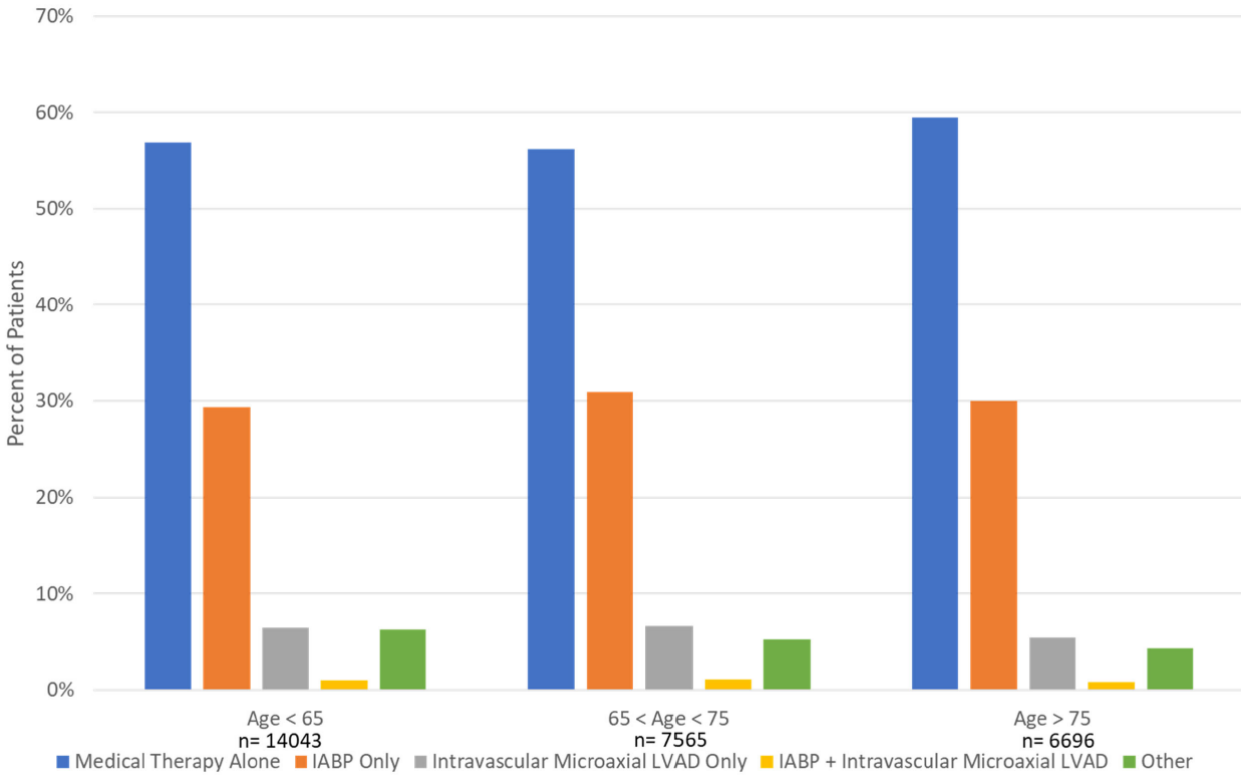


Figure 6.5: Age Distribution by Therapy for Patients Undergoing Percutaneous Coronary Intervention for Acute Myocardial Infarction Complicated by Cardiogenic Shock from October 1, 2015 – December 31, 2017

during hospitalization, STEMI, anterior infarction, and severe left main and/or proximal LAD stenosis were associated with greater MCS device use. Patients treated at private or university hospitals were more likely to receive any MCS device.

In multivariable regression analysis, patients who presented with STEMI (OR, 0.69; 95% CI, 0.60-0.80) and with previous CABG (OR, 0.79; 95% CI, 0.64-0.99) had significantly lower odds of use of intravascular microaxial LVADs only vs use of IABPs only (Table 6.3). Cardiac arrest at first medical contact or during hospitalization (OR, 1.82; 95% CI, 1.58-2.09) and severe left main and/or proximal LAD stenosis (OR, 1.36; 95% CI, 1.20-1.54) were associated with higher odds of intravascular microaxial LVAD use compared with IABP use.

Table 6.3: Patient and hospital characteristics associated with use of any mechanical circulatory support (MCS) device vs medical therapy only and with use of intravascular microaxial left ventricular assist device (LVAD) only vs intra-aortic balloon pump only

	OR (95% CI)	
Variable	Use of any MCS device vs medical therapy only	Use of intravascular microaxial LVAD vs intra-aortic balloon pump
Patient characteristics		
Age	1.00 (1.00-1.00)	1.00 (0.99-1.00)
Female sex	0.88 (0.83-0.93)	0.93 (0.82-1.05)
BMI	1.01 (1.00-1.01)	1.02 (1.01-1.03)
Race		
Other	1 [Reference]	1 [Reference]
White	0.88 (0.77-1.00)	1.04 (0.75-1.46)
Black	0.86 (0.74-1.00)	1.29 (0.86-1.94)
Insurance		
None	1 [Reference]	1 [Reference]
Medicaid	1.06 (0.92-1.22)	0.87 (0.64-1.16)
Medicare	1.11 (0.99-1.24)	0.93 (0.73-1.18)
Private	1.13 (1.03-1.23)	1.00 (0.81-1.24)
Medicaid and Medicare	0.97 (0.84-1.13)	0.91 (0.65-1.28)
Private and Public	1.06 (0.94-1.18)	0.94 (0.74-1.20)
Other	1.07 (0.93-1.24)	1.05 (0.79-1.39)
Medical History		
PAD	0.78 (0.71-0.86)	1.26 (1.05-1.52)
Cardiac arrest	1.70 (1.58-1.83)	1.82 (1.58-2.09)
STEMI	1.19 (1.11-1.28)	0.69 (0.60-0.80)
Anterior Infarct	1.19 (1.12-1.27)	0.94 (0.82-1.07)
Left main or proximal LAD disease	2.21 (2.08-2.35)	1.36 (1.20-1.54)
Previous PCI	1.00 (0.94-1.07)	1.04 (0.92-1.18)
Previous CABG	0.74 (0.67-0.81)	0.79 (0.64-0.99)
LVEF (per 1% increase)	0.96 (0.96-0.97)	0.97 (0.97-0.98)
Hospital characteristics		
Beds, No.		
<200	1 [Reference]	1 [Reference]
200-399	1.09 (0.94-1.27)	10.40 (0.65-1.65)
400-599	1.13 (0.94-1.35)	0.98 (0.61-1.57)
≥600	1.22 (0.99-1.50)	0.81 (0.48-1.39)
Location		
Rural	1 [Reference]	1 [Reference]
Suburban	0.91 (0.77-1.08)	0.86 (0.55-1.37)
Urban	0.89 (0.75-1.05)	1.08 (0.70-1.66)
Type		
Government	1 [Reference]	1 [Reference]
Private	1.33 (0.80-2.21)	0.91 (0.44-1.85)
University	1.84 (1.08-3.12)	0.82 (0.36-1.90)
Teaching Program	1.02 (0.91-1.15)	1.05 (0.79-1.39)
Annual PCI volume (per increase of 1 annual PCI)	1.00 (1.00-1.00)	1.00 (1.00-1.00)

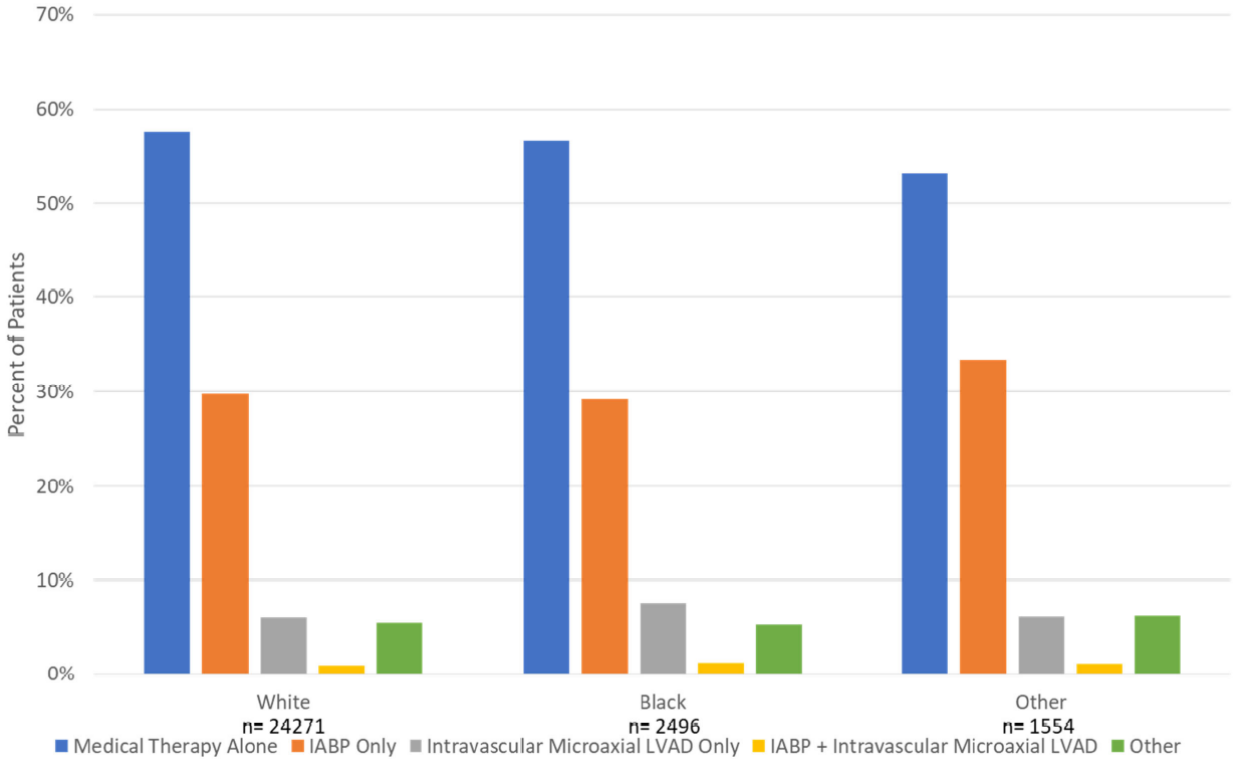


Figure 6.6: Race Distribution by Therapy for Patients Undergoing Percutaneous Coronary Intervention for Acute Myocardial Infarction Complicated by Cardiogenic Shock from October 1, 2015 – December 31, 2017

6.4 Discussion

This large, national cross-sectional study of patients who underwent PCI for AMI complicated by cardiogenic shock showed that, although overall use of MCS devices remained constant between 2015 and 2017, use of intravascular microaxial LVADs increased substantially, whereas use of IABPs decreased. Significant hospital-level variation in MCS device use was observed, with some hospitals not using any MCS devices and some hospitals using only intravascular microaxial LVADs or only IABPs. Other MCS devices remained infrequently used but may be used in combination with or as part of sequential therapy.

Previous studies through 2013 demonstrated a decrease in IABP use [149], which may be attributed to RCTs not demonstrating the clinical benefits of this device [142, 143].

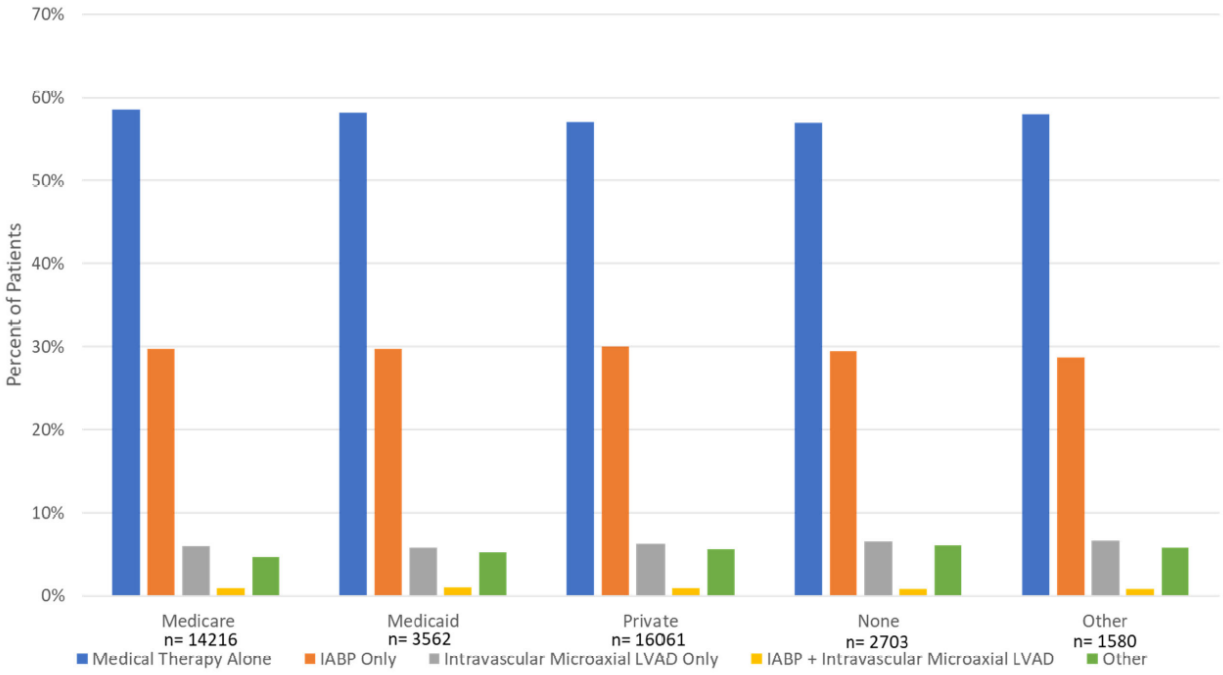


Figure 6.7: Insurance Distribution by Therapy for Patients Undergoing Percutaneous Coronary Intervention for Acute Myocardial Infarction Complicated by Cardiogenic Shock from October 1, 2015 – December 31, 2017

This study extends these past findings to more recent years. Regardless, we found that IABP remains the most commonly used MCS device in patients with AMI complicated by cardiogenic shock; more than 70% of patients who received an MCS device received an IABP. The ongoing use of this device despite its lack of association with improved clinical outcomes may be explained by familiarity with IABPs and because clinical practice guidelines in the US have not recommended against routine IABP use [141]. This finding is in contrast to the European Society of Cardiology clinical practice guidelines published in August 2017 (near the end of the study period), which gave routine IABP use a class III recommendation for patients with cardiogenic shock and STEMI [162].

The increasing use of intravascular microaxial LVADs may be associated with the greater hemodynamic support they provide compared with IABPs to patients with cardiogenic shock [146], who have a high mortality risk. Patients expected to have greater hemodynamic

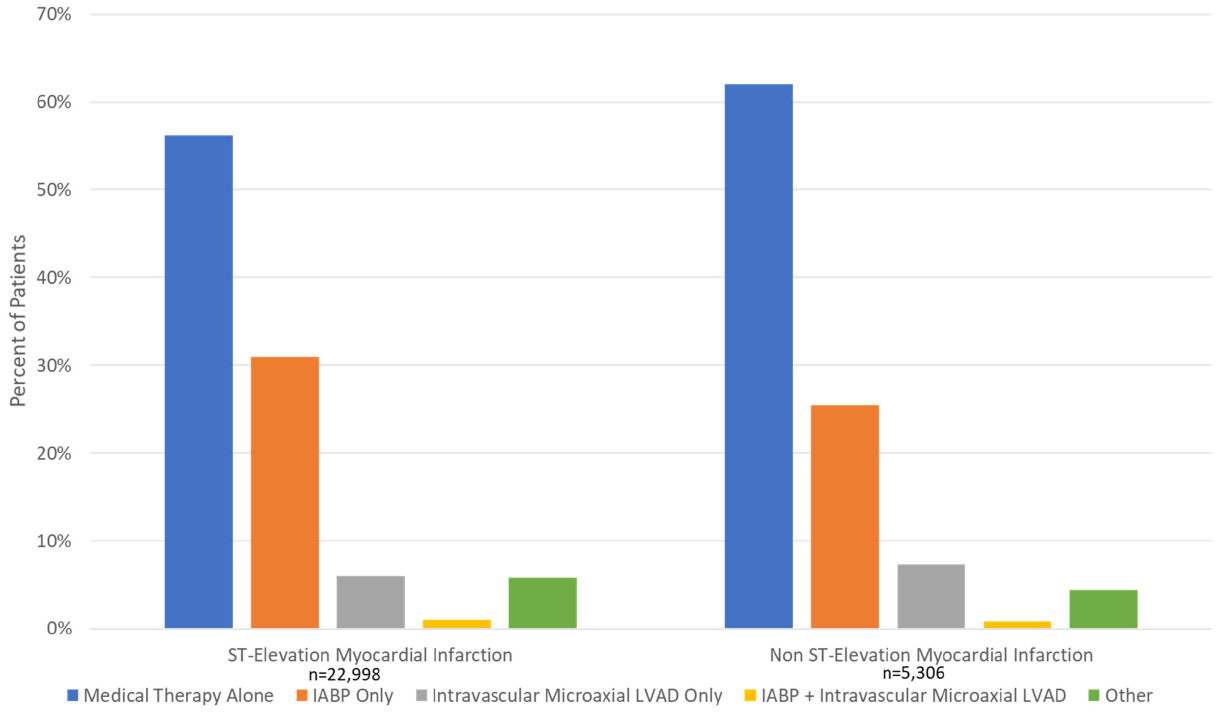


Figure 6.8: Type of Myocardial Infarction by Therapy for Patients Undergoing Percutaneous Coronary Intervention for Acute Myocardial Infarction Complicated by Cardiogenic Shock from October 1, 2015 – December 31, 2017

compromise, including those with cardiac arrest and left main or proximal LAD disease, were more likely to receive intravascular microaxial LVADs. Some groups have recommended intravascular microaxial LVADs for patients with severe cardiogenic shock [163].

However, the significant hospital-level variation in MCS device use and intravascular microaxial LVAD use suggests that no standard of care exists. This lack of consensus is consistent with multiple other studies, including a study of patients with AMI complicated by cardiogenic shock that reported that patient characteristics were not associated with MCS device use [164] and with another study of patients with cardiogenic shock in cardiac intensive care units in which hospital-level variation in MCS device use could not be explained by differences in illness severity [165].

One reason for the substantial variation in hospital use of MCS devices may be the

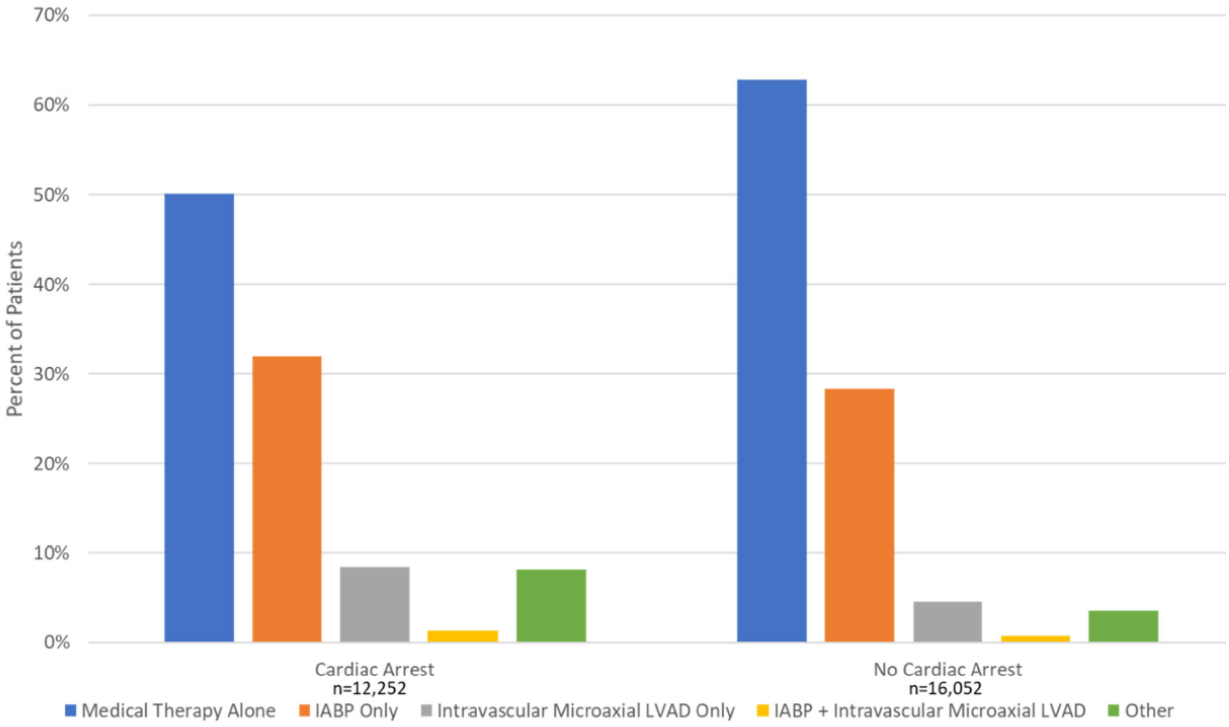


Figure 6.9: Cardiac Arrest Status by Therapy for Patients Undergoing Percutaneous Coronary Intervention for Acute Myocardial Infarction Complicated by Cardiogenic Shock from October 1, 2015 – December 31, 2017

paucity of clinical study data demonstrating the clinical benefit of intravascular microaxial LVAD use among patients with AMI complicated by cardiogenic shock [166]. Existing RCTs do not show the benefits of IABP use in AMI with cardiogenic shock [146, 150], although recent large observational studies have found that intravascular microaxial LVAD use was associated with higher mortality compared with IABP use [152, 153]. Intravascular microaxial LVADs were also significantly more expensive than IABPs, suggesting significant differences in total cost [148, 152, 167]. Another reason for the variation in hospital-level device use could be that patient and device selection for AMI with cardiogenic shock remains uncertain because of the clinical heterogeneity of cardiogenic shock [151]. A recently released classification scheme [168] could help establish the specific cardiogenic shock stages under which different MCS devices should be deployed. A third reason for the hospital-level use varia-

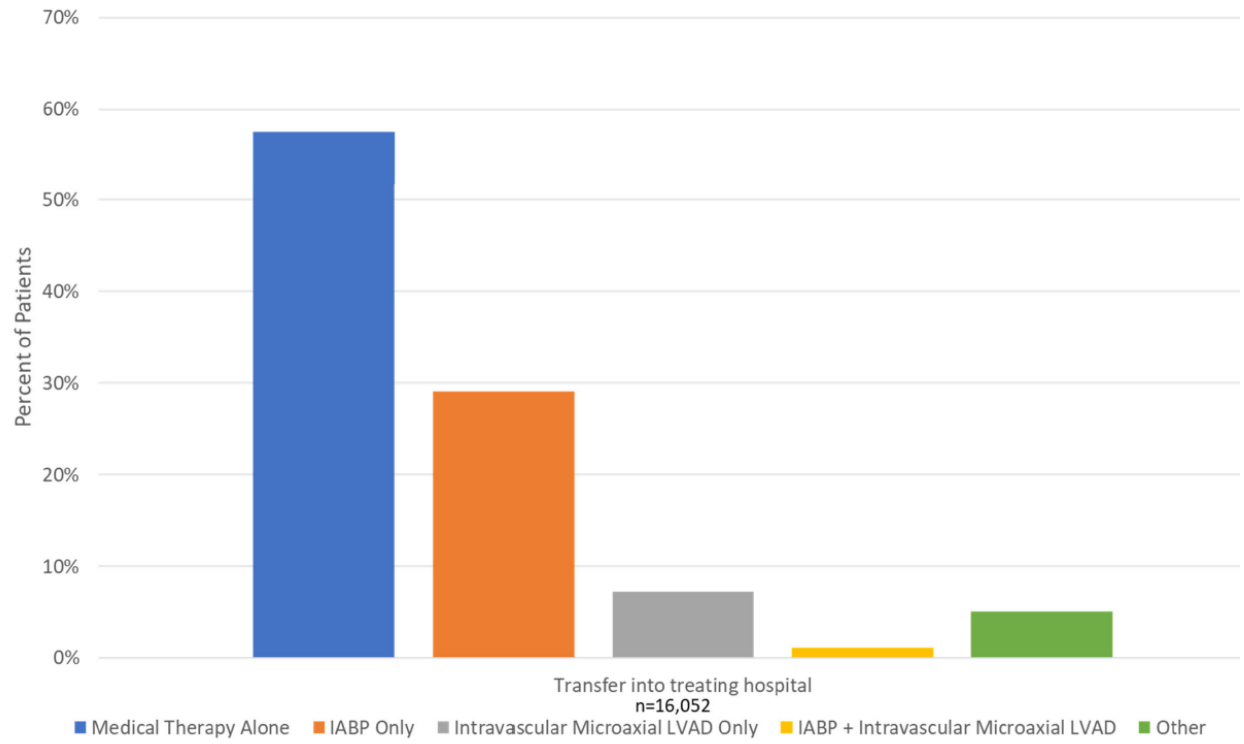


Figure 6.10: Therapies for Transfer Patients Undergoing Percutaneous Coronary Intervention for Acute Myocardial Infarction Complicated by Cardiogenic Shock from October 1, 2015 – December 31, 2017

tion may be that hospitals that have invested in the infrastructure to deploy intravascular microaxial LVADs for the care of patients are more likely to use these devices. Differences in reimbursement for intravascular microaxial LVADs vs IABPs [148] as well as other factors may also be associated with the observed use trends. Additional RCT evidence, which would help guide the selection, use, and timing of MCS devices in patients with AMI complicated by cardiogenic shock, could play a role in reducing hospital-level variation and improving patient outcomes as well as targeting these devices to patients who are most likely to find them beneficial [166, 169].

Among patients with STEMI, we found increased odds of MCS device use but lower use of intravascular microaxial LVADs. Because patients with STEMI in general have more acute, unstable presentations and are more likely to have a cardiac arrest, it is not surprising

that these patients often received MCS devices. However, in the model adjusted for clinical presentation and coronary anatomy, the lower odds of intrasvascular microaxial LVAD vs IABP use broadly highlighted the substantial variation in use trends that seemed to be associated not only with clinical presentation or physiological features but also with discretionary decision-making by physicians and institutions.

A novel finding of this study was that women with AMI complicated by cardiogenic shock were less likely than men to receive any MCS therapy. This finding extends the reports of differences in treatment provided to women with AMI, such as primary PCI [170] and other device-based therapy for cardiovascular disease [171]. These differences may be associated with the smaller vascular anatomy, which cannot accommodate the large bore access needed for MCS devices, and a greater predisposition to bleeding complications in women compared with men [172]. Further research is needed to ascertain the reasons for these sex-based differences.

6.4.1 Limitations

This study has several limitations. First, the presence of cardiogenic shock was based on site documentation. Second, different types of intravascular microaxial LVADs, specifically the Impella 2.5, CP, 5.0, and RP devices (ABIOMED), could not be distinguished. Third, because the Chest Pain-MI Registry allows only a single MCS device to be coded, some patients may have received combinations of devices that were not captured. Fourth, we did not have information on all variables relevant to cardiogenic shock (eg, lactate levels or number of vasopressors used), which may be associated with use of specific MCS devices.

6.5 Conclusions

Among patients who underwent PCI for AMI complicated by cardiogenic shock from October 2015 to December 2017, use of intravascular microaxial LVADs increased, with a corresponding decrease in use of IABPs despite limited clinical trial evidence of improved outcomes associated with device use. Significant hospital-level variation in use of MCS

devices was also found.

7. ASSOCIATION OF USE OF AN INTRAVASCULAR MICROAXIAL LEFT VENTRICULAR ASSIST DEVICE VS INTRA-AORTIC BALLOON PUMP WITH IN-HOSPITAL MORTALITY AND MAJOR BLEEDING AMONG PATIENTS WITH ACUTE MYOCARDIAL INFARCTION COMPLICATED BY CARDIOGENIC SHOCK*

Having established that the use of intravascular microaxial LVADs increased significantly despite a lack of clinical trial evidence to support their benefits, we now look at the association between their usage and major adverse outcomes compared with the usage of intra-aortic balloon pumps. To accomplish this, we used propensity matching to group similar patients who received opposite interventions. We compared the outcomes among those groups, and found that usage of intravascular microaxial LVAD was associated with a significantly higher risk of in-hospital death (45.0%) when compared with IABP (34.1%).

7.1 Introduction

Based on data collected from 1995 to 2013, cardiogenic shock occurs in an estimated 4% to 12% [173, 174, 175] of patients with acute myocardial infarction (AMI) and is associated with substantial morbidity and mortality. Percutaneous coronary intervention (PCI) is the cornerstone of management with a consideration of hemodynamic support with mechanical circulatory support (MCS) devices—most commonly intra-aortic balloon pumps (IABPs) and Impella devices (intravascular microaxial left ventricular assist devices [LVADs]).

Although intravascular microaxial LVADs improve hemodynamic parameters more than IABPs, it is not known whether this translates into improved outcomes among patients. The first intravascular microaxial LVAD received US Food and Drug Administration (FDA)

*Reprinted with permission from "Association of Use of an Intravascular Microaxial Left Ventricular Assist Device vs Intra-aortic Balloon Pump With In-Hospital Mortality and Major Bleeding Among Patients With Acute Myocardial Infarction Complicated by Cardiogenic Shock" by Dhruva, Sanket S; Ross, Joseph S; Mortazavi, Bobak J; Hurley, Nathan C; Krumholz, Harlan M; Curtis, Jeptha P; Berkowitz, Alyssa; Masoudi, Frederick A; Messenger, John C; Parzynski, Craig S; Ngufor, Che; Girotra, Saket; Amin, Amit P; Shah, Nilay D; and Desai, Nihar R, 2021. *JAMA* , 323, 734-745, Copyright© 2020 by American Medical Association.

clearance (in 2008 through the 510[k] regulatory pathway) for temporary support for up to 6 hours during cardiac procedures based on substantial equivalence to previously approved circulatory support devices but without a pivotal trial to demonstrate clinical efficacy compared with a control group [147]. Intravascular microaxial LVADs were later reclassified as higher-risk class III medical devices in 2014, which now require premarket approval. In April 2016, FDA-approved indications for intravascular microaxial LVADs were expanded through premarket approval to include treatment of cardiogenic shock following AMI. This was based, in part, on a randomized clinical trial (RCT) that showed improved hemodynamics as compared with IABP [146], as well as data from a manufacturer-initiated registry demonstrating improved outcomes relative to historical data [176]. Two RCTs that compared intravascular microaxial LVAD and IABP have demonstrated no statistically significant difference in 30-day mortality in AMI complicated by cardiogenic shock [146, 150]. A matched-pair analysis of 474 patients treated with intravascular microaxial LVAD in clinical practice compared with treatment using IABP (from patients in the IABP-SHOCK II trial) similarly showed no statistically significant mortality difference [177]. Despite limited data demonstrating improvements in clinical outcomes relative to IABP, use of intravascular microaxial LVAD has steadily increased over time [148, 149].

Accordingly, this study sought to use the clinical data collected in 2 national registries to examine clinical outcomes associated with intravascular microaxial LVAD and IABP among patients with AMI complicated by cardiogenic shock undergoing PCI.

7.2 Methods

7.2.1 Data Source

For this study, we linked CathPCI and Chest Pain-MI, 2 registries under the American College of Cardiology's National Cardiovascular Data Registry (described previously) [154, 155]. The CathPCI Registry is a national voluntary registry of diagnostic cardiac catheterizations and PCIs. More than 1500 hospitals across the United States participate

and are required to submit data on all PCI procedures. The Chest Pain-MI Registry includes patients with AMI and is used in more than 1000 US hospitals. Both registries capture standardized data elements, including patient demographics, medical history, laboratory data, procedural details, and in-hospital outcomes including mortality and major bleeding. Version 4.4 of the CathPCI Registry includes details of angiographic findings and can identify whether a patient received an IABP or any other MCS device. Version 2.4.2 of the Chest Pain-MI data collection form (released in the third quarter of 2015) includes the type of MCS device. All data submissions must meet prespecified quality standards. The registries include automatic system validation, education and training of staff, reporting of completeness, and random on-site auditing [129]. The human investigation committee of the Yale University School of Medicine approved the use of a limited data set from the registry for research without requiring informed consent.

7.2.2 Study Population

All patients who underwent PCI for AMI complicated by cardiogenic shock between October 1, 2015, and December 31, 2017, were identified. Patients with cardiogenic shock were identified as those in the Chest Pain-MI Registry who had cardiogenic shock at first medical contact, as an in-hospital event, or those defined in the CathPCI Registry who had cardiogenic shock within 24 hours prior to and up to PCI, at the start of PCI, or as an intra- or postprocedure event. Cardiogenic shock is defined in both registries as 1, 2, or all 3 of the following: systolic blood pressure lower than 90 mm Hg, a cardiac index of less than 2.2 L/minute/m2 for at least 30 minutes that is secondary to ventricular dysfunction, or requirement for parenteral inotropic or vasopressor or MCS devices to support blood pressure and cardiac index [156]. For patients who underwent multiple PCIs during the hospitalization, only data from the initial PCI were included.

7.2.3 Registry Linkage

A probabilistic linkage [178] of patients across the 2 registries was performed to include detailed procedural data from the CathPCI Registry and the specific MCS device type from the Chest Pain-MI Registry. Multiple iterations of matching were then performed, with each subsequent match omitting variables that had been previously included. The matching variables were patient sex, date of admission, time of arrival to facility, age at hospital arrival, a unique hospital identifier, discharge date, and whether PCI was performed as documented in the Chest Pain-MI Registry. This match algorithm identified patients with entries in both registries at the same hospital. To identify patients with AMI complicated by cardiogenic shock who were transferred to another hospital for PCI or who had minor missing data elements that may have affected the match, up to 4 variables were allowed to be mismatched, but these variables always included sex and at least 1 date variable to ensure temporal factors limited matches for similar patients at different hospital encounters. The resulting linked CathPCI-Chest Pain-MI registry cohort formed our analytic cohort.

7.2.4 Hemodynamic Support

Patients were categorized based on hemodynamic support: IABP only, intravascular microaxial LVAD only, and other (such as use of a percutaneous extracorporeal ventricular assist system, extracorporeal membrane oxygenation, LVAD, or patients receiving multiple devices during the hospitalization). Patients coded as not receiving any MCS device constituted the medical therapy group.

7.2.5 Outcomes

The primary outcomes were all-cause in-hospital death and in-hospital major bleeding. Death was captured in the Chest Pain-MI Registry. Major bleeding was defined using the Chest Pain-MI Registry as a decline in hemoglobin level of at least 3 g/dL; transfusion of whole blood or packed red blood cells; procedural intervention/surgery at bleeding site to treat the bleeding; or documented or suspected retroperitoneal bleed, gastrointestinal bleed,

genitourinary bleed, or a bleed in a location not specified elsewhere [179].

7.2.6 Covariates

Covariates were obtained from the CathPCI and Chest Pain-MI registries and included patient demographics, medical history, clinical presentation, laboratory values, administered medications, procedural characteristics, and coronary anatomic data. For continuous values with missing values, the mean was imputed. For binary (yes/no) variables, all missing variables were coded as no, and for categorical variables, all missing variables were coded as no or other (if there was not a no category).

Race and ethnicity were included in this study because our goal was to use all available patient information when risk-standardizing through propensity matching. This determination was made by the patient or family member and then entered into the CathPCI Registry. This determination was based on fixed categories, although multiple response options were possible. For race, the categories were white, black/African American, American Indian/Alaskan Native, Asian, and Native Hawaiian/Pacific Islander. For ethnicity, the categories were Hispanic/Latino or not.

7.2.7 Statistical Analysis

First, overall use of hemodynamic support among all patients was characterized. Characteristics of patients receiving intravascular microaxial LVAD vs those receiving IABP were compared using χ^2 tests for categorical variables and 1-way analysis of variance or Kruskal-Wallis tests for continuous variables.

Clinical outcomes of mortality and major bleeding among patients undergoing PCI for AMI complicated by cardiogenic shock were characterized using propensity-matched analyses to compare patients who received either intravascular microaxial LVAD or IABP only. Seventy-five variables were preselected for matching using previously described methods [180]. Among patients who received either an intravascular microaxial LVAD or IABP, a probabilistic model was developed that calculated the log-odds probability of receiving an

Table 7.1: C-statistic for discrimination between intravascular microaxial left ventricular assist device and intra-aortic balloon pump among all hospitals

	Intravascular Microaxial LVAD vs IABP	IABP vs Medical Therapy Only
C-statistic	0.790	0.745

Table 7.2: C-statistic for discrimination between intravascular microaxial left ventricular assist device and intra-aortic balloon pump among all hospitals with at least 1 intra-aortic balloon pump and 1 intravascular microaxial left ventricular assist device

	Intravascular Microaxial LVAD vs IABP	IABP vs Medical Therapy Only
C-statistic	0.778	0.759

intravascular microaxial LVAD. To develop the log-odds probability and to handle higher-dimensional, nonlinear relationships between covariates, a gradient descent–boosted decision tree algorithm was used to develop the propensity model (called extreme gradient boosting) [131]. The hyperparameters of learning rate were set to 0.1, as is common in slow-learning algorithms, and the number of trees and maximum depth of each tree was selected optimally in a 5-fold cross-validation analysis (depth range, 1-10; number of trees range, 50-1000 in increments of 10). The final model used a depth of 3 for each decision tree and 100 decision trees, which optimally maximized the C statistic for discriminating between intravascular microaxial LVAD and IABP (Tables 7.1 and 7.2).

For each patient who received an intravascular microaxial LVAD, we found all IABP patients with a similar propensity for intravascular microaxial LVAD usage (within 0.6 standard deviations, a value that eliminates approximately 90% of the bias in observed confounders) [181] and randomly selected 1 IABP patient for paired matching. This pair was then removed from the cohort, and the process was repeated until all patients were either matched

or could not be matched due to probability differences.

The standardized mean difference of each covariate was calculated in the propensity-matched cohort. Next, outcomes in the cohort were examined and the absolute risk difference (ARD) and associated 95% CIs were calculated. To verify results, a second independent statistician blinded to the results of the initial analysis confirmed results from the gradient descent–boosted decision tree algorithm using standard logistic regression to propensity match patients using 75 variables.

For sensitivity, analyses were repeated stratified by timing of MCS device placement (either before or after initiation of PCI, when these data were available) in patients from hospitals that had placed at least 1 intravascular microaxial LVAD and IABP, therefore demonstrating capability to use both devices, and in patients who were not transferred to a facility.

As a secondary analysis, a comparison of patients receiving IABP vs medical therapy only was made (using the methods previously described) to determine whether outcomes using propensity matching were similar to those observed from the IABP-SHOCK II trial [142].

As an additional step to address potential unmeasured confounding, an instrumental variable analysis was conducted using hospital-level propensity to use intravascular microaxial LVAD during our study period as the instrumental variable. A 2-stage ordinary least-squares regression analysis was conducted. In the first stage, the predicted probability of receiving intravascular microaxial LVAD at the facility-level was calculated after adjustment for covariates included in our propensity-score matching. The F statistic was calculated to determine the strength of the instrumental variable (a value >10 suggested proceeding to the second stage). In the second stage, the predicted probability of receiving intravascular microaxial LVAD (determined during the first stage) was used as the primary predictor, again adjusting for the same covariates, to examine differences in in-hospital clinical outcomes. The instrumental variable analysis was conducted in 2 populations: in the entire cohort of patients

with AMI complicated by cardiogenic shock and in the patients who received intravascular microaxial LVAD only or IABP only. Analyses were conducted in R, with packages XGBoost for gradient descent boosting [131] and pROC for C statistic calculations [182]. The primary analyses examining outcomes of intravascular microaxial LVAD vs IABP were repeated using SAS version 9.4. All statistical analyses were 2-sided (α=.05 for statistical significance).

7.3 Results

7.3.1 Study Cohort

Of the 269,303 patients with AMI receiving PCI between October 1, 2015, and December 31, 2017, and matched across the Chest Pain-MI and CathPCI registries, 28,304 (10.5%) were classified as having cardiogenic shock. The mean (SD) age was 65.0 (12.6) years (Table 7.3). Approximately two-thirds of patients were men and 86% were white. Approximately 25% had been transferred from another acute care hospital, 81% presented with acute ST-segment elevation myocardial infarction, 38.9% had anterior infarct location, and 43.3% had cardiac arrest either at first medical contact or during hospitalization. Among those with cardiogenic shock at first medical contact, the mean (SD) systolic blood pressure was 94.9 (51.4) mm Hg.

Table 7.3: Characteristics of patients undergoing percutaneous coronary intervention for acute myocardial infarction complicated by cardiogenic shock and of propensity-matched patients receiving intravascular microaxial left ventricular assist device vs intra-aortic balloon pump from October 1, 2015, through December 31, 2017

Patient Characteristics	Medical Therapy Alone	Other MCS	Intravascular Microaxial LVAD Only	IABP Only	p-value	Intravascular Microaxial LVAD matched	IABP Matched	SMD
Total Patients	16,227 (57.3)	1838 (6.5)	1768 (6.2)	8471 (29.9)		1680	1680	
Age, mean (SD) years	65.3 (12.8)	63.0 (12.4)	64.2 (12.0)	65.2 (12.4)	<0.001	64.3 (11.9)	64.0 (11.9)	0.03
BMI, mean (SD) kg/m2	28.9 (6.5)	29.3 (6.6)	29.7 (6.3)	28.9 (6.1)	<0.001	29.6 (6.3)	30.0 (6.4)	0.06
Men	10,517 (64.8)	1326 (72.1)	1260 (71.3)	5865 (69.2)	<0.001	1194 (71.1)	1198 (71.3)	0.005
Women	5710 (35.2)	512 (27.9)	508 (28.7)	2606 (30.8)		486 (28.9)	482 (28.7)	0.005
Race								
White	13,915 (85.8)	1552 (84.4)	1465 (82.9)	7194 (84.9)	<0.001	1390 (82.7)	1414 (84.2)	0.04
Black	1 422 (8.8)	167 (9.1)	185 (10.5)	739 (8.7)		178 (10.6)	170 (10.1)	0.02
Asian	491 (3.0)	75 (4.1)	43 (2.4)	334 (3.9)		41 (2.4)	48 (2.9)	0.03
American Indian	178 (1.1)	16 (0.9)	42 (2.4)	89 (1.1)		37 (2.2)	27 (1.6)	0.04
Native Hawaiian/Pacific Islander	42 (0.3)	2 (0.1)	5 (0.3)	34 (0.4)		5 (0.3)	9 (0.5)	0.04
Not recorded	222 (1.4)	34 (1.8)	33 (1.9)	123 (1.5)		N/A	N/A	N/A
Hispanic or Latino Ethnicity	1057 (6.5)	126 (6.9)	116 (6.6)	576 (6.8)	0.82	N/A	N/A	N/A
Insurance								
Medicaid	2072 (12.8)	225 (12.2)	205 (11.6)	1060 (12.5)	0.51	194 (11.5)	206 (12.3)	0.02
Medicare	8330 (51.3)	797 (43.4)	857 (48.5)	4232 (50.0)	<0.001	812 (48.3)	792 (47.1)	0.02
Private	9172 (56.5)	1058 (57.6)	1003 (56.7)	4828 (57.0)	0.79	957 (57.0)	959 (57.1)	0.002
None	1540 (9.5)	189 (10.3)	178 (10.1)	796 (9.4)	0.57	173 (10.3)	189 (11.3)	0.002
Other	916 (5.6)	105 (5.7)	105 (5.9)	454 (5.4)	0.7	101 (6.0)	95 (5.7)	0.02
Clinical History								
Cardiovascular Risk Factors								
Hypertension	11,394 (70.2)	1214 (66.1)	1239 (70.1)	5784 (68.3)	<0.001	1176 (70.0)	1160 (69.0)	0.02
Dyslipidemia	9443 (58.2)	979 (53.3)	1023 (57.9)	4760 (56.2)	<0.001	971 (57.8)	923 (54.9)	0.06
Current/recent smoker	6169 (38.0)	625 (34.0)	548 (31.0)	2800 (33.1)	<0.001	516 (30.7)	545 (32.4)	0.04
Diabetes mellitus	5118 (31.5)	637 (34.7)	646 (36.5)	2857 (33.7)	<0.001	609 (36.3)	621 (37.0)	0.02
Family history of premature CAD	1914 (11.8)	201 (10.9)	191 (10.8)	933 (11.0)	0.19	184 (11.0)	168 (10.0)	0.03
Currently on dialysis	496 (3.1)	37 (2.0)	52 (2.9)	211 (2.5)	0.01	48 (2.9)	61 (3.6)	0.04
Established Coronary Artery Disease								
Prior MI	3570 (22.0)	369 (20.1)	394 (22.3)	1755 (20.7)	0.04	373 (22.2)	376 (22.4)	0.004
Prior PCI	3772 (23.2)	389 (21.2)	423 (23.9)	1861 (22.0)	0.03	398 (23.7)	369 (22.0)	0.04
Prior CABG	1368 (8.4)	128 (7.0)	127 (7.2)	605 (7.1)	0.001	125 (7.4)	124 (7.4)	0.002

Table 7.3 continued from previous page

Patient Characteristics	Medical Therapy Alone	Other MCS	Intravascular Microaxial LVAD Only	IABP Only	p-value	Intravascular Microaxial LVAD matched	IABP Matched	SMD
Established Cardiovascular Disease								
Prior Heart Failure	1939 (11.9)	223 (12.1)	226 (12.8)	963 (11.4)	0.3	212 (12.6)	231 (13.8)	0.02
Heart Failure within 2 weeks	2941 (18.1)	550 (29.9)	603 (34.1)	2177 (25.7)	<0.001	549 (32.7)	563 (33.5)	0.02
NYHA Class I-III	1364 (8.4)	141 (7.7)	182 (10.3)	798 (9.4)	0.002	175 (10.4)	167 (9.9)	0.02
NYHA Class IV	1556 (9.6)	406 (22.1)	419 (23.7)	1372 (16.2)	<0.001	373 (22.2)	395 (23.5)	0.03
Cardiomyopathy or LV dysfunction	1886 (11.6)	350 (19.0)	366 (20.7)	1350 (15.9)	<0.001	333 (19.8)	350 (20.8)	0.02
Atrial fibrillation/flutter	1373 (8.5)	109 (5.9)	144 (8.1)	626 (7.4)	<0.001	138 (8.2)	137 (8.2)	0.002
Established Vascular Disease								
Cerebrovascular disease	1898 (11.7)	176 (9.6)	191 (10.8)	877 (10.4)	0.002	181 (10.8)	176 (10.5)	0.01
Peripheral artery disease	1618 (10.0)	120 (6.5)	172 (9.7)	641 (7.6)	<0.001	161 (9.6)	147 (8.8)	0.03
Chronic lung disease	2503 (15.4)	216 (11.8)	223 (12.6)	1072 (12.7)	<0.001	205 (12.2)	215 (12.8)	0.02
Cancer	1463 (9.0)	103 (5.6)	152 (8.6)	721 (8.5)	<0.001	146 (8.7)	144 (8.6)	0.004
Heart rate, median (IQR)	76 (57-97)	83 (63-106)	87 (68-106)	84 (60-104)	<0.001	83.7 (35.3)	82.6 (36.1)	0.03
Systolic blood pressure, median (IQR)	120 (92-147)	114 (87-140)	116 (90-142)	118 (92-143)	<0.001	112.7 (46.3)	111.4 (47.5)	0.03
Cardiac Arrest within 24 hours of PCI	4441 (27.4)	739 (40.2)	624 (35.3)	2666 (31.5)	<0.001	578 (34.4)	601 (35.8)	0.03
Cardiac arrest at first medical contact	3791 (23.4)	531 (28.9)	449 (25.4)	2060 (24.3)	<0.001	421 (25.1)	455 (27.1)	0.05
Heart failure at first medical contact	2611 (16.1)	476 (25.9)	464 (26.2)	1835 (21.7)	<0.001	425 (25.3)	474 (28.2)	0.07
STEMI pre-PCI	12,938 (79.7)	1552 (84.4)	1383 (78.2)	7148 (84.4)	<0.001	425 (25.3)	474 (28.2)	0.07
Thrombolytics Given	518 (3.2)	60 (3.3)	50 (2.8)	257 (3.03)	0.78	50 (3.0)	62 (3.7)	0.04
Anterior Infarction	5207 (32.1)	903 (49.1)	894 (50.6)	4019 (47.4)	<0.001	846 (50.4)	879 (52.3)	0.01
Left Main and/or Proximal LAD Disease	5434 (33.5)	1092 (59.4)	1106 (62.6)	4630 (54.7)	<0.001	1037 (61.7)	1046 (62.3)	0.04
Multivessel Disease	7886 (48.6)	1127 (61.3)	1171 (66.2)	5376 (63.5)	<0.001	1110 (66.1)	1110 (66.1)	0
Chronic Total Occlusion	502 (3.1)	84 (4.6)	55 (3.1)	317 (3.7)	0.001	51 (3.0)	65 (3.9)	0.05
Renal Failure	1271 (7.8)	128 (7.0)	171 (9.7)	666 (7.9)	0.02	163 (9.7)	167 (9.9)	0.008
Chronic Kidney Disease								
GFR≥60 mL/min	8326 (51.3)	862 (46.9)	821 (46.4)	4102 (48.42)	<0.001	211 (12.6)	225 (13.4)	0.02
45≤GFR<60 mL/min	3276 (20.2)	400 (21.8)	377 (21.3)	1843 (21.8)	0.02	786 (46.8)	752 (44.8)	0.04
30≤GFR<45 mL/min	1798 (11.1)	243 (13.2)	229 (13.0)	1018 (12.0)	0.003	360 (21.4)	364 (21.7)	0.006
GFR< 30 mL/min or Currently on Dialysis	2828 (17.4)	333 (18.1)	341 (19.3)	1509 (17.8)	0.24	159 (9.5)	163 (9.7)	0.008
GFR Data Missing	1617 (10.0)	210 (11.4)	176 (10.0)	885 (10.4)	0.2	164 (9.8)	176 (10.5)	0.02
GFR, median (IQR) mL/min	64.5 (48.5-81.6)	62.2 (47.2-78.0)	61.5 (45.2-77.8)	62.9 (47.4-78.7)	<0.001	59.4 (21)	59.9 (20.6)	0.03
LVEF, mean (SD)	42.6 (13.5)	31.6 (14.7)	29.1 (13.5)	34.6 (14.0)	<0.001	32.0 (8.6)	32.1 (8.6)	0.02
ECG Findings								

Table 7.3 continued from previous page

Patient Characteristics	Medical Therapy Alone	Other MCS	Intravascular Microaxial LVAD Only	IABP Only	p-value	Intravascular Microaxial LVAD matched	IABP Matched	SMD
ST Elevation	12 614 (77.7)	1506 (81.9)	1328 (75.1)	6933 (81.8)	<0.001	1277 (76.0)	1288 (76.7)	0.02
LBBB	160 (1.0)	37 (2.0)	38 (2.1)	97 (1.1)	<0.001	32 (1.9)	23 (1.4)	0.04
Isolated Posterior MI	146 (0.9)	15 (0.8)	14 (0.8)	83 (1.0)	0.82	15 (0.9)	15 (0.9)	0
Initial Labs, median (IQR)								
BNP, pg/mL	245.0 (69.0-703.0)	300.0 (103.2-873.2)	409.0 (121.0-1003.0)	326.0 (100.0-807.0)	<0.001	745.2 (692.0)	757.8 (746.1)	0.02
NT-ProBNP, pg/mL	1796.5 (330.8-6081.2)	1920.0 (353.0-6211.0)	2280.0 (585.0-6630.0)	2030.0 (362.8-6940.2)	0.14	5795.6 (3375.2)	5944.6 (3984.7)	0.04
Hemoglobin, g/dL	13.9 (12.4-15.2)	13.9 (12.3-15.3)	13.9 (12.2-15.4)	13.9 (12.4-15.3)	0.01	13.7 (2.3)	13.8 (2.2)	0.05
Hemoglobin A1c, %	5.9 (5.5-7.1)	6.0 (5.5-7.5)	6.0 (5.6-7.4)	6.0 (5.5-7.4)	0.005	6.8 (1.3)	6.8 (1.4)	0.01
INR	1.1 (1.0-1.2)	1.1 (1.0-1.3)	1.1 (1.0-1.3)	1.1 (1.0-1.2)	<0.001	1.3 (1.0)	1.4 (1.2)	0.03
INR<2	15 663 (96.5)	1759 (95.7)	1668 (94.3)	8108 (95.7)	<0.001	N/A	N/A	N/A
INR≥2	564 (3.5)	79 (4.3)	100 (5.7)	363 (4.3)		N/A	N/A	N/A
Pre-procedure labs, median (IQR)								
CK-MB, ng/mL	6.8 (2.6-31.6)	8.6 (3.0-40.9)	8.8 (3.4-36.3)	7.8 (3.0-37.8)	<0.001	52.0 (59.0)	53.0 (107.9)	0.01
Troponin I, ng/mL	0.42 (0.06-4.5)	0.61 (0.07-6.2)	1.04 (0.10-8.50)	0.59 (0.08-6.80)	<0.001	12.6 (30.0)	12.5 (30.5)	0.002
Troponin T, ng/mL	0.07 (0.01-0.69)	0.15 (0.01-0.99)	0.20 (0.01-0.94)	0.11 (0.01-0.99)	0.004	5.2 (3.8)	5.2 (1.6)	0.02
Creatinine, mg/dL	1.1 (0.9-1.4)	1.2 (1-1.5)	1.2 (1-1.6)	1.2 (0.9-1.5)	<0.001	1.5 (1.1)	1.5 (1.2)	0.04
Hemoglobin, g/dL	13.9 (12.4-15.2)	13.9 (12.3-15.3)	13.9 (12.2-15.4)	13.9 (12.4-15.3)	0.1	13.6 (2.2)	13.7 (2.1)	0.05
Medications								
Aspirin	15 508 (95.6)	1682 (91.5)	1643 (92.9)	7987 (94.3)	<0.001	1564 (93.1)	1561 (92.9)	0.007
Ticagrelor	7146 (44.0)	782 (42.5)	793 (44.9)	3554 (42.0)	0.008	754 (44.9)	756 (45.0)	0.002
Clopidogrel	5964 (36.8)	512 (27.9)	462 (26.1)	2857 (33.7)	<0.001	449 (26.7)	448 (26.7)	0.001
Beta-blocker	7244 (44.6)	447 (24.3)	439 (24.8)	2420 (28.6)	<0.001	411 (24.5)	427 (25.4)	0.02
Prasugrel	1655 (10.2)	151 (8.2)	148 (8.4)	638 (7.5)	<0.001	132 (7.9)	140 (8.3)	0.02
ACE Inhibitor	2279 (14.0)	91 (5.0)	123 (7.0)	627 (7.4)	<0.001	131 (7.8)	117 (7.0)	0.03
Angiotensin Receptor Blocker	444 (2.7)	23 (1.3)	35 (2.0)	132 (1.55)	<0.001	30 (1.8)	35 (2.1)	0.08
PCI Status								
Emergency	12 715 (78.4)	1352 (73.6)	1223 (69.2)	6808 (80.4)	<0.001	1200 (71.4)	1168 (69.5)	0.04
Salvage	860 (5.3)	307 (16.7)	322 (18.2)	878 (10.4)	<0.001	273 (16.3)	297 (17.7)	0.04
Urgent	2511 (15.5)	168 (9.1)	213 (12.0)	747 (8.8)	<0.001	197 (11.7)	205 (12.2)	0.02
Elective	135 (0.8)	11 (0.6)	10 (0.6)	36 (0.4)	0.003	10 (0.6)	10 (0.6)	0
Timing of MCS Placement								
Prior to initiation of PCI	N/A	925 (50.3)	717 (40.6)	2078 (24.5)	<0.001	653 (38.9)	747 (44.5)	0.11
Post initiation of PCI	N/A	939 (51.08)	724 (41.0)	5633 (66.5)	<0.001	720 (42.9)	635 (37.8)	0.1
Transferred into CathPCI Hospital	3966 (24.4)	426 (23.2)	494 (27.9)	2002 (23.6)	0.001	461 (27.4)	451 (26.8)	0.01
Transferred into CP-MI Hospital	4018 (24.6)	428 (23.3)	418 (27.3)	2013 (23.8)	0.2	450 (26.8)	449 (26.7)	0.001

7.3.2 Mechanical Circulatory Support Device Utilization

In this cohort of 28,304 patients with AMI complicated by cardiogenic shock undergoing PCI, 1768 (6.2%) received only an intravascular microaxial LVAD, 8471 (29.9%) received only an IABP, 1838 (6.5%) received other MCS devices or multiple devices, and 16,227 (57.3%) received medical therapy alone and were not treated with MCS (Figure 7.1). Patients receiving intravascular microaxial LVAD were significantly younger than patients receiving IABP (Table 7.3). Patients with intravascular microaxial LVAD were significantly less likely to have acute ST-segment elevation myocardial infarction (78.2%) vs patients with IABP (84.4%; P $<$ .001) but significantly more likely to be transferred into a Chest Pain-MI–reporting facility (patients with intravascular microaxial LVAD [27.3%] vs patients with IABP [23.8%]; P = .02). There was no significant difference in the percentage of patients who experienced cardiac arrest at first medical contact (intravascular microaxial LVAD [25.4%] vs IABP [24.3%]; P = .35).

7.3.3 Outcomes of Intravascular Microaxial LVAD vs IABP

Unadjusted outcomes are provided in Table 7.4. The 1:1 propensity matching algorithm using data from all patients with AMI complicated by cardiogenic shock undergoing PCI yielded 1680 matched pairs, accounting for 95.0% (1680 of 1768) of patients who received an intravascular microaxial LVAD. Standardized mean differences for 74 of 75 (99%) characteristics of the propensity-matched cohorts were below 0.10 (Table 7.3).

In the propensity-matched cohort, use of intravascular microaxial LVAD was associated with a significantly higher risk of in-hospital death (45.0%) when compared with use of IABP (34.1%; ARD, 10.9 percentage points [95% CI, 7.6-14.2]; P $<$ .001; Figure 7.2). These statistically significant differences were consistent, regardless of the timing of device placement, among patients with intravascular microaxial LVAD placement before initiation of PCI (45.5%) compared with patients receiving IABP before initiation of PCI (36.8%; ARD, 8.7 percentage points [95% CI, 3.1-14.4]; P = .003), and among those with intravascular microax-

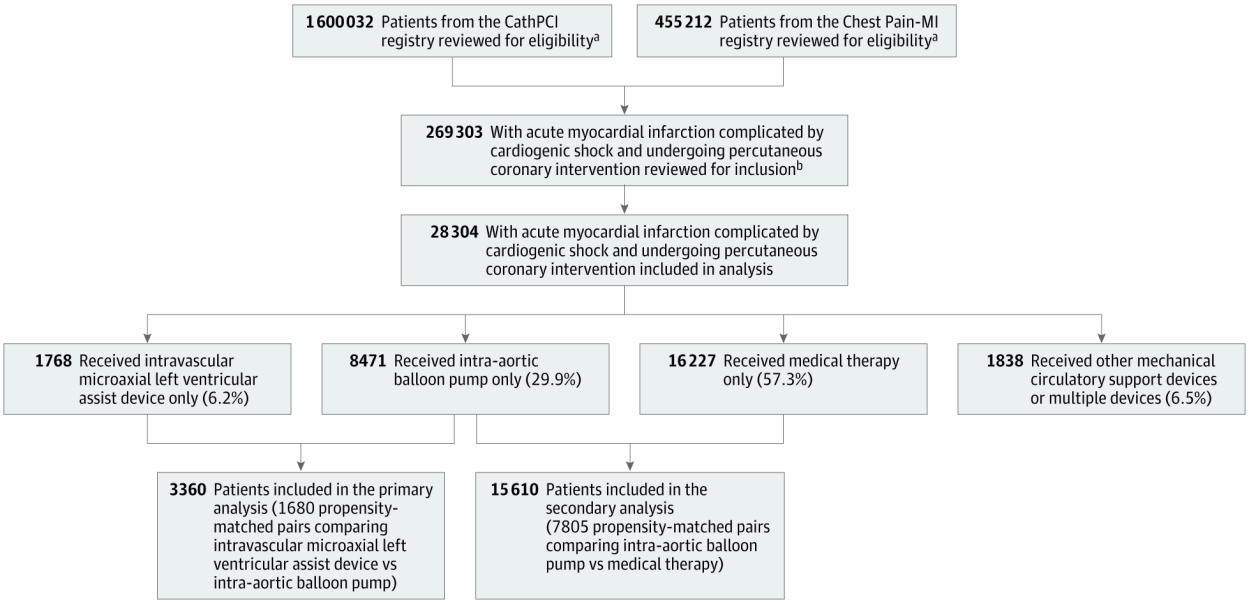


Figure 7.1: Patient Population With Acute Myocardial Infarction Complicated by Cardiogenic Shock Undergoing Percutaneous Coronary Intervention. [a]CathPCI and Chest Pain-MI are registries under the American College of Cardiology's National Cardiovascular Data Registry. PCI indicates percutaneous coronary intervention; MI, myocardial infarction. [b]Patient data were accessed from linked registries.

ial LVAD placement after initiation of PCI (44.0%) compared with IABP after initiation of PCI (32.2%; ARD, 11.8 percentage points [95% CI, 6.6-17.0]; $P < .001$).

Use of an intravascular microaxial LVAD was also associated with a significantly higher risk of in-hospital major bleeding (31.3%) compared with use of IABP (16.0%; ARD, 15.4 percentage points [95% CI, 12.5-18.2]; $P < .001$; Figure 7.2); both access site and nonaccess site bleeding were significantly higher with intravascular microaxial LVAD (Table 7.5).

A secondary 1:1 propensity-matching algorithm using data from the 390 hospitals that used both intravascular microaxial LVAD and IABP included 1570 matched pairs and found consistent results (Figure 7.3). Results were also consistent in an additional 1:1 propensity-matching algorithm among 1201 matched pairs that excluded patients transferred into a treating facility (Figures 7.4 and 7.5). Results were also consistent in an instrumental variable analysis among all patients with AMI complicated by cardiogenic shock and when limited to patients with AMI complicated by cardiogenic shock receiving intravascular microaxial

Table 7.4: Unadjusted Outcomes Among Patients Undergoing Percutaneous Coronary Intervention for Acute Myocardial Infarction Complicated by Cardiogenic Shock from October 1, 2015 – December 31, 2017

	Medical Therapy Alone	Other MCS	Intravascular Microaxial LVAD Only	IABP Only
Number of patients	16,227	1,838	1,768	8,471
Outcomes				
Death	3,241 (20)	798 (43)	801 (45)	2,461 (29)
Major Bleeding	1,688 (10)	552 (30)	556 (31)	1233 (14.6)

Table 7.5: Characteristics of Bleeding Type in Matched Cohort Undergoing PCI and Receiving Intravascular Microaxial Left Ventricular Assist Device or Intra-aortic Balloon Pump for Acute Myocardial Infarction Complicated by Cardiogenic Shock, Among All Hospitals

	Intravascular Microaxial LVAD matched	IABP Matched
Number of patients	1,680	1,680
Overall bleeding	526 (31.3)	268 (16.0)
Access site bleed	188 (11.2)	55 (3.3)
Gastrointestinal bleed	115 (6.8)	73 (4.3)
Genito-urinary bleed	43 (2.6)	12 (0.7)
Retroperitoneal bleed	25 (1.5)	7 (0.4)
Other bleed	210 (12.5)	108 (6.4)
Surgery required for bleed	96 (5.7)	37 (2.2)
Red blood cell / whole blood transfusion	623 (37.1)	315 (18.8)

	Intravascular Microaxial Left Ventricular Assist Device		Intra-aortic Balloon Pump		Absolute Risk Difference (95% CI), %	P Value
	No. of Patients	Patients, %	No. of Patients	Patients, %		
Overall (n = 1680 matched pairs)						
Mortality	756	45.0	573	34.1	10.9 (7.6-14.2)	<.001
Major bleeding	526	31.3	268	16.0	15.4 (12.5-18.2)	<.001
Device placement before initiation of percutaneous coronary intervention (n = 573 matched pairs)						
Mortality	261	45.5	211	36.8	8.7 (3.1-14.4)	.003
Major bleeding	157	27.4	95	16.6	10.8 (6.1-15.6)	<.001
Device placement after initiation of percutaneous coronary intervention (n = 662 matched pairs)						
Mortality	291	44.0	213	32.2	11.8 (6.6-17.0)	<.001
Major bleeding	228	34.4	104	15.7	18.7 (14.2-23.3)	<.001

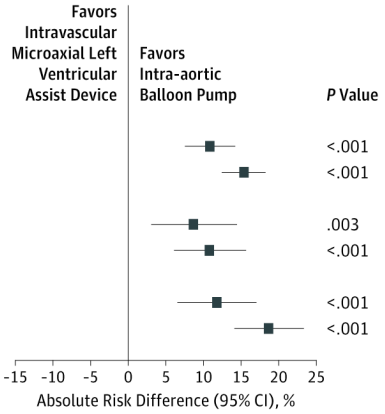


Figure 7.2: In-Hospital Outcomes Among Propensity-Matched Patients With Acute Myocardial Infarction Complicated by Cardiogenic Shock Undergoing Percutaneous Coronary Intervention With Intravascular Microaxial Left Ventricular Assist Device vs Intra-aortic Balloon Pump

LVAD or IABP only.

7.3.4 Outcomes of IABP vs Medical Therapy Alone

The 1:1 propensity-matching algorithm using data from all patients with AMI complicated by cardiogenic shock undergoing PCI yielded 7805 matched pairs. Standardized differences for the characteristics of both propensity-matched cohorts were all below 10%.

In the propensity-matched cohort, IABP use was not associated with lower in-hospital mortality when compared with medical therapy alone; there was a small but statistically significantly higher risk (IABP, 28.6% vs 26.5% for medical therapy alone; ARD, 2.2 percentage points [95% CI, 0.8-3.6]; P = .002). In-hospital major bleeding was significantly higher among patients receiving IABP (14.5% vs 11.0%; ARD, 3.5 percentage points [95% CI, 2.5-4.5]; P < .001) (Figure 7.6).

7.4 Discussion

Among patients with AMI complicated by cardiogenic shock, use of intravascular microaxial LVAD was associated with significantly higher risks of patients experiencing in-hospital mortality and major bleeding compared with use of IABP. These findings were

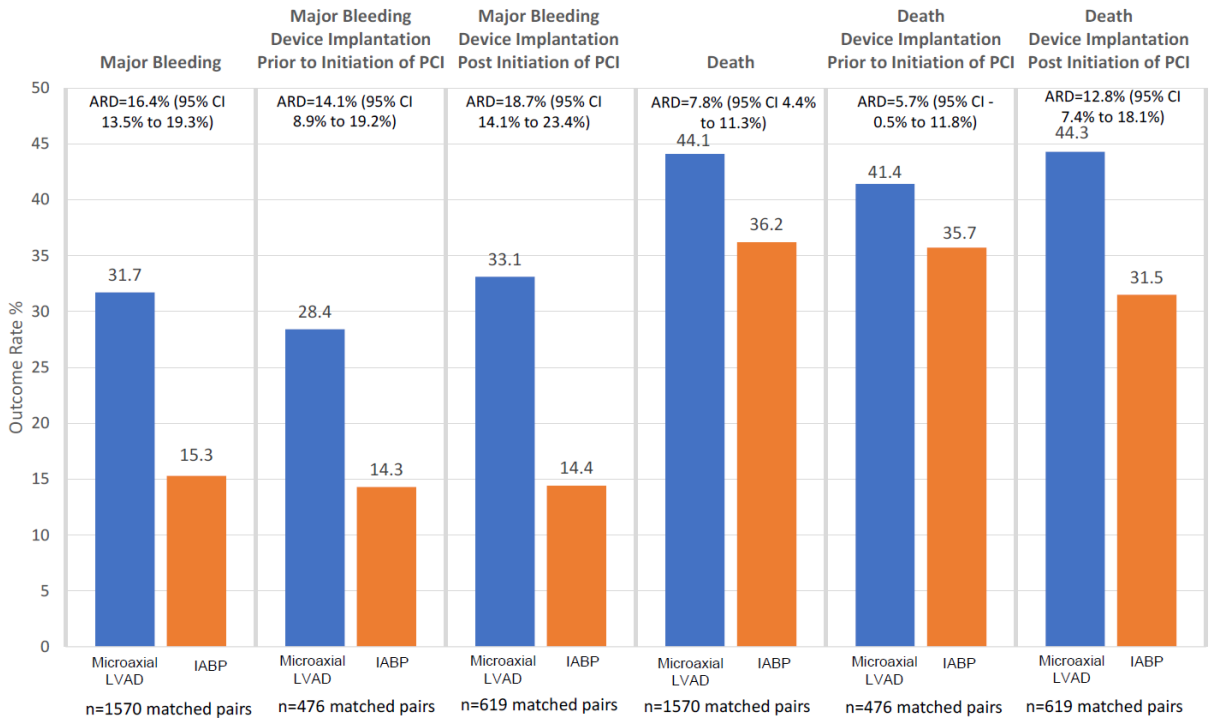


Figure 7.3: In-Hospital Outcomes among Propensity-Matched Patients with Acute Myocardial Infarction Complicated by Cardiogenic Shock Undergoing PCI with Intravascular Microaxial Left Ventricular Assist Device vs Intra-Aortic Balloon Pump, Among All Hospitals with At Least 1 Intra-aortic Balloon Pump and 1 Intravascular Microaxial Left Ventricular Assist Device

consistent for patients regardless of the timing of device placement and transfer status.

The significantly higher risk of in-hospital mortality contrasts with prior RCTs, which failed to show a mortality benefit of intravascular microaxial LVAD but did not show overall harm. There are a number of potential explanations for the findings of this study relative to the previous clinical trials. First, by using national registry data, this study was larger than prior RCTs [146, 150], which cumulatively enrolled only 74 total patients. Second, this study examined clinical experience, rather than device performance among highly selected patients treated by experienced physicians and hospitals in RCTs [146, 150]. The recent experience with Impella RP is instructive: the FDA's May 2019 advisory warning [183] suggests experience with devices as they are used in everyday clinical practice must be closely monitored.

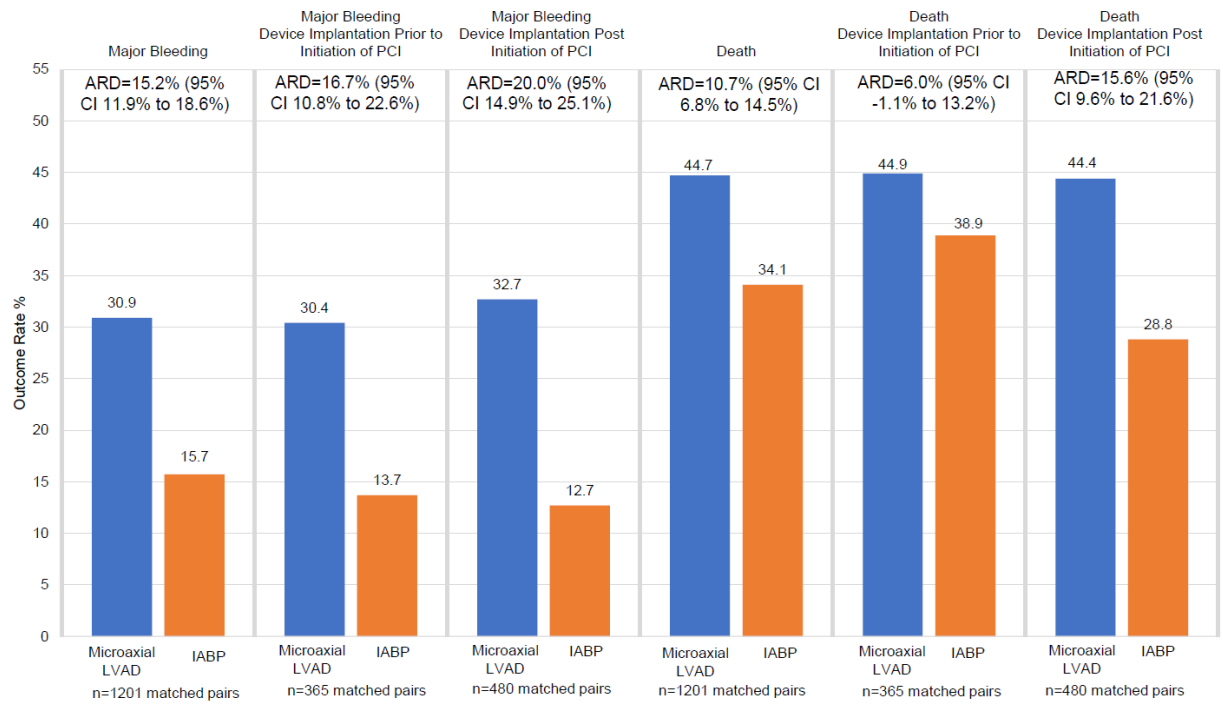


Figure 7.4: In-Hospital Outcomes among Propensity-Matched Patients Who Were Not Transferred to a Treating Facility with Acute Myocardial Infarction Complicated by Cardiogenic Shock Undergoing PCI with Intravascular Microaxial Left Ventricular Assist Device vs Intra-Aortic Balloon Pump (IABP), Among All Hospitals

While a recent matched pair analysis of 237 patients from the IABP-SHOCK II trial and 237 patients receiving intravascular microaxial LVAD in a multinational registry found a point estimate of 2.1% higher mortality among patients receiving intravascular microaxial LVAD compared with IABP that was not statistically significant, that analysis did demonstrate higher risk of severe or life-threatening bleeding in patients receiving intravascular microaxial LVAD [177]. The results showing higher risk of severe or life-threatening bleeding are consistent with the current analysis, which shows higher risk of severe in-hospital major bleeding among patients treated with intravascular microaxial LVAD. These results are also consistent with a large observational study of patients undergoing PCI with MCS, which found that use of intrasvascular microaxial LVAD was associated with higher risk of in-hospital adverse events, including death and major bleeding [152]. Additionally, mortality

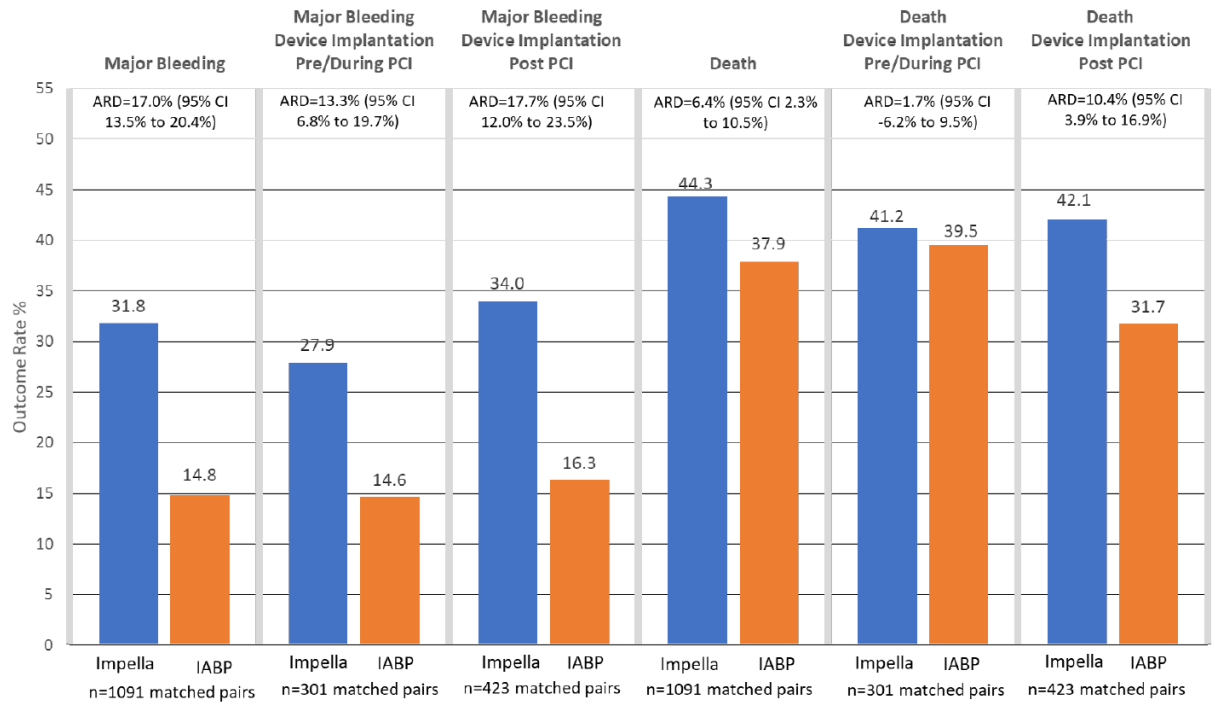


Figure 7.5: In-Hospital Outcomes among Propensity-Matched Patients with Acute Myocardial Infarction Complicated by Cardiogenic Shock Undergoing PCI with Intravascular Microaxial Left Ventricular Assist Device vs Intra-Aortic Balloon Pump (IABP), Among All Hospitals with At Least 1 Intra-aortic Balloon Pump and 1 Intravascular Microaxial Left Ventricular Assist Device

risk when comparing patients receiving IABP vs medical therapy only were consistent with those observed in the IABP-SHOCK II trial [142], the largest RCT of IABP in cardiogenic shock, and support the robustness of this analytic approach. A potential explanation for the increase in mortality may be the increased bleeding with intravascular microaxial LVAD as compared with IABP, which is consistent with prior studies [150, 184, 185]. Bleeding and transfusions are associated with adverse outcomes, including mortality, among patients with AMI [186, 187] and receiving PCI [188].

Taken together, these results highlight the need for additional data to guide the optimal management of AMI complicated by cardiogenic shock in general and the role of MCS devices, in particular. Specifically, robust RCTs and complementary analyses of clinical populations are necessary. The former, including the ongoing DanGer trial of intravascular microaxial

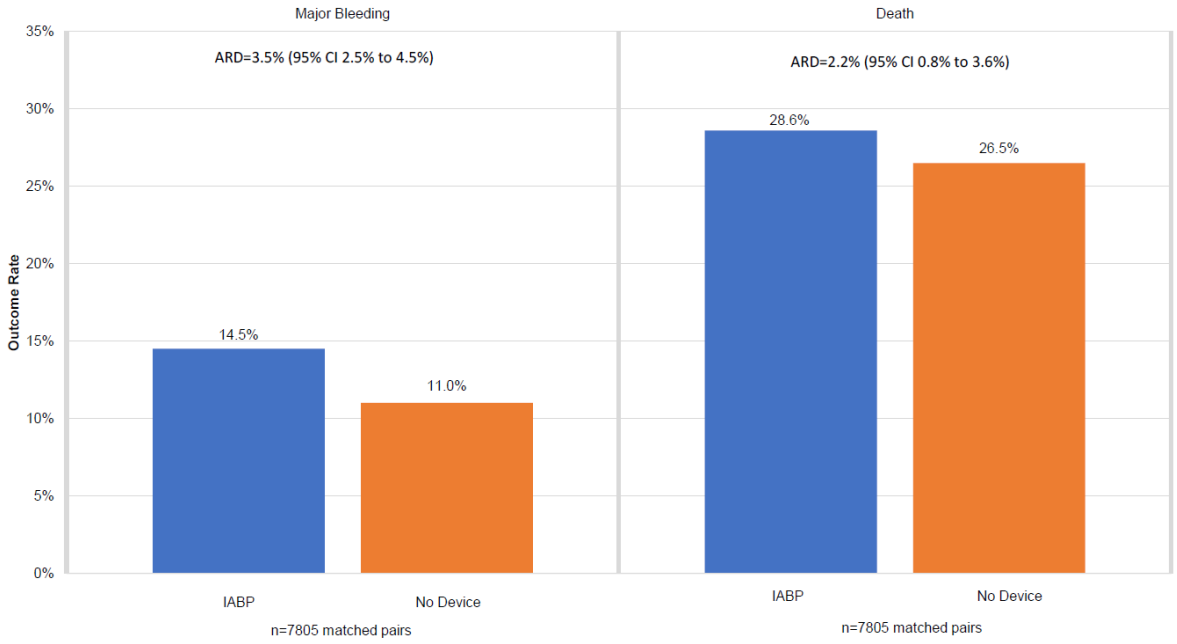


Figure 7.6: In-Hospital Outcomes among Propensity-Matched Patients with Acute Myocardial Infarction Complicated by Cardiogenic Shock Undergoing PCI with Intra-Aortic Balloon Pump vs Medical Therapy Alone

LVAD vs medical therapy in AMI complicated by cardiogenic shock [189], may provide definitive data on efficacy and safety, while the latter may provide important information about device performance in unselected settings and possible off-label indications.

A 2017 American Heart Association scientific statement noted little evidence to guide the timing or selection of patients with cardiogenic shock who are suitable for MCS devices [151]. Furthermore, given that cardiogenic shock is a complex, heterogenous syndrome requiring complex team-based clinical care infrastructure and highly specialized clinicians, a recently released classification scheme [168] may improve the phenotyping of patients with shock to better align therapeutic interventions with the cause and degree of hemodynamic derangement. In this context, this analysis of outcomes among patients receiving MCS devices may inform these efforts.

7.4.1 Limitations

This study has several limitations. First, the presence of cardiogenic shock was based on site documentation. More detailed hemodynamic and clinical data, including the use of vasopressor therapy at the time of MCS device placement, would have enabled a more granular patient profile but are not captured in either the Chest Pain-MI or CathPCI case report forms. However, the registry definition for shock is consistent with many clinical trials, and sites are subject to random audit [129]. Moreover, the event rate among patients included in the propensity-matched analyses suggests these patients had cardiogenic shock.

Second, registry data provide clinical information, such as hemodynamics and laboratory values, at a single time point; cardiogenic shock is an evolving process, and the specific information at the time of decision to use a particular MCS device was not available. Therefore, the possibility cannot be excluded that comparable patient clinical status at presentation might have changed during subsequent hospital course prior to initiation of therapy and have affected the observed outcome differences between intravascular microaxial LVAD and IABP.

Third, there may be residual confounding whereby patients receiving intravascular microaxial LVADs had greater severity of illness than those receiving IABPs. While this study employed a propensity match using detailed demographics, clinical history and presentation, infarct location, coronary anatomy, and clinical laboratory data from a large, national registry, other clinical parameters that may affect or be associated with MCS device selection, such as right heart catheterization measurements, lactate levels, or success of reperfusion were not available. However, approximately 95% of all patients receiving intravascular microaxial LVADs were matched, suggesting that these results may represent the experience of the majority of patients receiving intravascular microaxial LVAD for AMI complicated by cardiogenic shock.

Fourth, different types of intravascular microaxial LVADs, specifically the Impella 2.5, CP, 5.0, and RP, could not be distinguished. While there are differences in the degree of

hemodynamic support provided by these devices, the Impella 5.0 device requires specialized vascular access and is unlikely to be the initial support device used in a patient with AMI complicated by cardiogenic shock. Impella RP was approved in September 2017 and was only available for approximately 3 months of the study period. To further mitigate these concerns, this analysis was limited to patients who received only the intravascular microaxial LVAD or IABP such that patients who had an escalation in their support with device replacement were excluded.

7.5 Conclusions

Among patients undergoing PCI for AMI complicated by cardiogenic shock from 2015 to 2017, use of intravascular microaxial LVAD compared with IABP was associated with higher adjusted risk of in-hospital death and major bleeding complications, although study interpretation is limited by the observational design. Further research may be needed to understand optimal device choice for these patients.

8. A DYNAMIC MODEL TO ESTIMATE EVOLVING RISK OF MAJOR BLEEDING AFTER PCI

This chapter shifts focus slightly away from AMI-CS and more broadly to all patients receiving percutaneous coronary intervention. Here, we look to predict bleeding following that procedure. A key focus in this chapter is examining risk. We do this not as a static one-time estimate, but rather as a more complex estimation that evolves over time. As a patient is treated, many decisions occur and the clinical scenario evolves. This chapter discusses a method to follow that patient through a complex process with multiple rounds of decision making. At each step, new information is learned and the model updates to account for what is available. Through this, the model is able to dynamically adapt to reflect the most current state of the patient, paving the way for improved clinical support.

8.1 Introduction

Bleeding is a common and serious complication associated with percutaneous coronary intervention (PCI). Bleeding is associated with significant morbidity, mortality, and cost [190]. To assess post-PCI bleeding risk, several prediction tools have been developed, including two risk models [191, 192] from the National Cardiovascular Data Registry (NCDR) [193]. By informing clinicians about bleeding risk, these models can aid use of bleeding avoidance strategies (e.g., combination of radial access and bivalirudin), particularly in high-risk patients, thereby reducing rates of major bleeding complications and improving care quality and clinical outcomes [194, 195]. However, rates of bleeding among different demographic groups and among different sites exists [196, 197]. This variation suggests the existence of opportunities for improvement. In addition, many patients at the highest risk for bleeding complications fail to receive guideline based bleeding avoidance strategies [198]. One possible reason for the lack of use of bleeding avoidance strategies is the static nature of the existing risk models.

Current models produce a single estimate of bleeding risk anchored at a single point in time. As treatment decisions are made or unforeseen events occur, these models are unable to adapt and incorporate new information. Development of a dynamic model is needed to allow for estimations to adapt and update throughout an episode of care. Bleeding risk is a dynamic process affected by multiple pre-, intra-, and post-PCI patient and procedural factors throughout the care pathway. As data are gathered and treatment decisions made, risk estimates should account for all of the information currently available, including changes in patient clinical status.

The development of risk models that update across the patient episode of care has the potential to improve our ability to individualize risk prediction. Providing physicians up-to-date feedback may inform optimization of therapeutic strategies, through enhanced decision-support at actionable points across a cardiac catheterization laboratory visit. These models may also improve the understanding of the dynamics and key variables affecting bleeding risk. Such models would represent a transformational change in risk prediction and embrace the principles of a learning health care system [199].

8.2 Methods

8.2.1 Study cohort

This study included all index PCIs in the National Cardiovascular Data Registry (NCDR) CathPCI registry version 4.4 from July 2009 through April 2015 [193, 154]. To examine the improvement of dynamic data over static risk prediction models, we used the same cohort as Mortazavi et al. [200]. We excluded patients during readmissions, who died in the hospital, who had missing data regarding if they had any bleeding events or not, or who underwent coronary artery bypass grafting (CABG) during the index admission, consistent with previous work [192, 200].

The primary outcome was any in-hospital bleeding event within 72 hours after the start of PCI. Bleeding was defined as a hemoglobin drop $\geq$3 g/dL, whole blood or packed red

blood cell transfusion, or intervention/surgery at the bleeding site to reverse/stop or correct bleeding. We further excluded patients with multiple, unknown, or brachial access sites, to evaluate the treatment decision point of radial versus femoral access. We additionally excluded patients with multiple closure methods.

8.2.2 Variables of Interest

This study considered all data available from the CathPCI Registry prior to patient discharge [193]. This included all data used by the existing models [192, 200], as well as additional variables as described below. The full existing NCDR bleeding risk model [192] uses 31 variables: 23 patient characteristics at the time of presentation and 8 characteristics related to coronary anatomy and lesion characterization. The additional data considered here consists of additional laboratory data, past medical history, coronary anatomy (including percent stenosis), stent type, and closure method categories (manual compression, sealant, mechanical, suture, patch, staple, other, or none).

8.2.3 Staged Model Analysis

We first sorted all available CathPCI data into key decision stages of a PCI episode of care. First, we defined three decision points that affect bleeding risk: 1) choice of access site (radial versus femoral); 2) choice of medications (including those administered 24 hours pre-procedure and intra-procedure); and 3) choice of closure device.

Using these three key decision points, we evaluated variables available at three stages: Stage 1) variables available at patient presentation to the catheterization lab; stage 2) variables available after diagnostic coronary angiography; and stage 3) variables related to the PCI procedure. Combining these three decision points and the information available at each of them, six models were designed Figure 8.1. The first included only variables available at presentation. Each subsequent model adds either a decision node or the information that can inform the next decision. The final model included all variables and clinical decisions through the PCI procedure, evaluating remaining bleeding risk for post-procedure care.

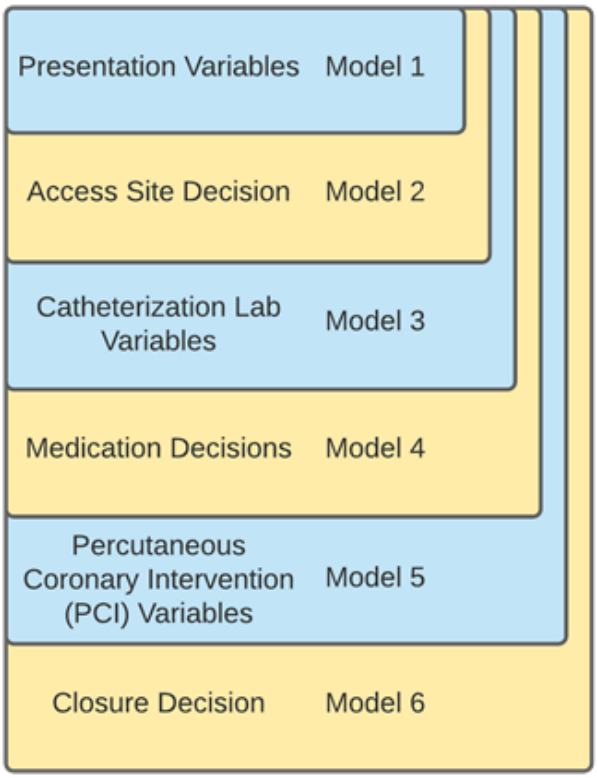


Figure 8.1: Model hierarchy. Each model integrated information of all features from prior models, as well as an added set of features.

8.2.4 Data Preparation

The NCDR goes through a high quality review and adjudication process, ensuring minimal missingness of variables [129]. However, several steps in data preparation were necessary prior to model development. First, situations exist where a parent variable value of "no", indicates that daughter variables would not be captured. Most daughter variables already had a category of missing, unknown, or other. We re-categorized the daughter variables to have a value of No/Not Drawn, and integrated "missing" for the few cases where the parent variable was a Yes/Drawn variable and daughter variable was in fact missing.

Second, medication was categorized as no, yes, blinded, or contraindicated. We recategorized blinded as missing and recategorized contraindicated as no.

Third, missing values were imputed using multiple multivariate feature imputation. Each missing feature was modeled using Bayesian ridge regressors trained in a round-robin fashion. This imputation was performed using the Iterative Imputer package in scikit-learn 0.24.1 [67], based on the multivariate imputation by chained equations (mice) package for R [201]. Following imputation, binary and ordinal variables were set to the nearest allowed value. Multiple imputations were produced by sampling from the regressor models multiple times; each discrete sampling is a new overall sample from the model. This sampling was used to produce five folds of imputations. In this work, we used a re-imputation technique for handling the longitudinal nature of the data, producing multiple imputed datasets at each stage through an episode of care [202]. The imputation models were trained on the training dataset prior to cross validation [203]. The test sets were multiply imputed, but the regressors used for this imputation were trained only on training data.

8.2.5 Training, Testing, and Evaluating

The cohort was divided into an initial 80% cohort (July 1, 2009 through December 31, 2013) for model training and a later 20% cohort (January 1, 2014 through December 31, 2014) for model validation [204, 205]. This temporal split allows for a fair evaluation of the model through testing it on the most distinct possible subset. This approach protects against biasing the result of the overall model by including more recent data. While an external validation set would be preferable, this method of separating data allows for the most realistic testing of the model. As discussed above, five imputation folds were created for each stage. This, in effect, resulted in five trained models for each stage or decision point. The model used was XGBoost, a gradient descent boosted decision tree model [206]. This model is capable of evaluating higher order, non-linear interactions between variables. This is necessary, since bleeding models based upon logistic regression selected the key variables based primarily on statistical tests between the variable and its relationship with incidence of major bleeding [192]. We conducted an internal five-fold cross-validation to tune the hyperparameters of the model for each stage and each fold. The hyperparameters tuned were maximum depth

of each tree (2, 4, 6, or 8), number of tree estimators (100, 500, 1000, or 5000), and learning rate for the model (0.1, 0.15, 0.2, or 0.3). This internal cross-validation was performed using a halving grid search as implemented by the scikit-learn package HalvingGridSearchCV [67]. Optimal performance was found in 27 of the 30 experiments (6 stages * 5 folds) to be given by 1000 estimators with a max depth of 2 and a learning rate of 0.1.

The five imputation folds allow for training and validating the model multiple times, providing both estimates of overall model performance and uncertainty [204]. From this, we calculated the area under the receiver operating characteristics curve (AUROC) for evaluating model discrimination. To better understand the model's positive predictive value across the full range of risk stratification, we also calculated the area under the precision recall curve (AUPRC). Briefly, the precision-recall curve calculates the tradeoff between precision (positive predictive value) and recall (sensitivity) across the full range of thresholds [207]. This number is directly impacted by the number of positive predictions made versus false positive or false negative estimates made, an important factor when considering cases with low event rates such as major bleeding (4.1%, Table 8.1). This model calibration also allows us to calculate the Brier Skill Score, which provides the Brier score for model calibration on an easy to interpret scale of 0-100% for calibration fit. Model performance metrics at each stage are provided in Table 8.2.

8.2.6 Variable Importance

Looking beyond model performance, model interpretation is a key factor in clinical utility [204]. One approach for interpreting models is Shapley Additive exPlanations (SHAP) [208]. SHAP attributes an importance value to each feature of a given sample and allows for an ordering of features from those with the greatest impact on the model output to least impact on the model output. This technique allows for explanations in nonlinear models where a given feature might have different impacts at different values. We use SHAP here to provide an analysis for patients undergoing PCI, therefore providing visual understanding about the impact of factors driving changes in accuracy of the risk prediction model and decisions

Table 8.1: Bleeding model patient characteristics.

	Overall	Training	Validation
	(n=2,868,808)	(n=2,314,446)	(n=554,362)
Demographics			
Age, mean (SD), y	64.6 (12.0)	64.6 (12.0)	64.9 (11.9)
Men	1,960,409 (68.3)	1,577,369 (68.2)	383,040 (69.1)
BMI, mean (SD)	30.0 (6.4)	30.0 (6.4)	30.1 (6.4)
Cardiovascular Comorbidities			
Diabetes	1,057,221 (36.9)	844,928 (36.5)	212,291 (38.3)
Hypertension	2,353,798 (82.1)	1,895,949 (81.9)	457,849 (82.6)
Peripheral Vascular Disease	339,316 (11.8)	274,039 (11.8)	65,278 (11.8)
Chronic Kidney Disease	861,391 (30.0)	705,765 (30.5)	155,626 (28.1)
Previous PCI	1,178,346 (41.1)	948,367 (41.0)	229,978 (41.5)
Previous CABG	510,781 (17.8)	414,560 (17.9)	96,222 (17.4)
PCI Procedural Status			
Elective	1,196,485 (41.7)	992,525 (42.9)	203,961 (36.8)
Urgent	1,152,328 (40.2)	906,226 (39.2)	246,101 (44.4)
Emergent	512,404 (17.9)	409,659 (17.7)	102,744 (18.5)
Salvage	6,440 (0.2)	5,029 (0.2)	1,411 (0.3)
Unknown	1,150 (0.04)	1,007 (0.04)	145 (0.03)
STEMI	468,270 (16.3)	373,792 (16.2)	94,477 (17.0)
Cardiogenic Shock	64,743 (2.3)	51,689 (2.2)	13,055 (2.4)
Cardiac arrest within 24h of PCI	49,008 (1.7)	38,840 (1.7)	10,168 (1.8)
Preprocedure hemoglobin, median (IQR), g/dL	13.7 (12.4-14.9)	13.7 (12.4-14.9)	13.7 (12.4-14.9)
Access Site			
Femoral	2,394,173 (83.5)	1,997,049 (86.3)	397,124 (71.6)
Radial	474,635 (16.5)	317,397 (13.7)	157,238 (28.4)
Medications Used			
Ticlopidine	5,895 (0.2)	5,186 (0.2)	709 (0.1)
Clopidogrel	1,988,178 (69.3)	1,654,031 (71.5)	334,147 (60.3)
Prasugrel	433,079 (15.1)	339,902 (14.7)	93,177 (16.8)
Ticagrelor	167,838 (5.9)	80,318 (3.5)	87,520 (15.8)
Fondaparinux	15,816 (0.6)	14,837 (0.6)	979 (0.2)
Low Molecular Weight Heparin	272,261 (9.5)	224,180 (9.7)	48,081 (8.7)
Unfractionated Heparin	1,528,882 (53.3)	1,197,304 (51.7)	331,578 (59.8)
Bivalirudin	1,695,225 (59.1)	1,375,031 (59.4)	320,194 (57.8)
GP llb/llla (any)	677,865 (23.6)	576,753 (24.9)	101,112 (18.2)
Direct Thrombin Inhibitor	29,512 (1.0)	24,961 (1.1)	4,551 (0.8)
Closure Method			
Manual Compression	965,618 (33.7)	815,354 (35.2)	150,264 (27.1)
Sealant	916,374 (31.9)	745,456 (32.2)	170,918 (30.8)
Mechanical	512,968 (17.9)	358,927 (15.5)	154,041 (27.8)
Suture	264,494 (9.2)	215,420 (9.3)	49,074 (8.9)
Patch	99,690 (3.5)	84,853 (3.7)	14,837 (2.7)
Staple	184 (0.0)	171 (0.0)	13 (0.0)
Other	94,818 (3.3)	82,626 (3.6)	12,192 (2.2)
None/Missing	14,662 (0.5)	11,639 (0.5)	3,023 (0.5)
Post-PCI Major Bleeds	118,327 (4.1)	98,167 (4.2)	20,160 (3.6)

Table 8.2: Comparison of model performances for bleeding prediction.

Model	AUROC	AUPRC	Brier Skill Score	Brier Reliability	Brier Resolution
1	0.812 (0.812-0.812)	0.203 (0.203-0.203)	0.088 (0.088-0.088)	2.6E-4 (2.6E-4-2.6E-4)	3.3E-3 (3.3E-3-3.3E-3)
2	0.817 (0.817-0.817)	0.204 (0.204-0.205)	0.091 (0.091-0.091)	2.0E-4 (1.9E-4-2.0E-4)	3.4E-3 (3.4E-3-3.4E-3)
3	0.825 (0.825-0.825)	0.208 (0.208-0.208)	0.094 (0.094-0.094)	2.1E-4 (2.0E-4-2.1E-4)	3.5E-3 (3.5E-3-3.5E-3)
4	0.832 (0.832-0.832)	0.217 (0.216-0.217)	0.102 (0.102-0.102)	1.4E-4 (1.3E-4-1.4E-4)	3.7E-3 (3.7E-3-3.7E-3)
5	0.844 (0.844-0.845)	0.241 (0.240-0.241)	0.118 (0.118-0.118)	1.1E-4 (1.1E-4-1.2E-4)	4.2E-3 (4.2E-3-4.2E-3)
6	0.845 (0.845-0.845)	0.242 (0.241-0.242)	0.119 (0.119-0.119)	1.0E-4 (1.0E-4-1.1E-4)	4.3E-3 (4.3E-3-4.3E-3)

through the stages outlined in Figure 8.1.

SHAP feature importance plots are shown in Figures 8.2-8.7. In each plot, the variables are sorted in order of decreasing importance. The color of a variable relates to the value of that variable, while location along the x-axis represents how much that variable contributes to the risk of bleeding. Features most strongly driving predictions of high bleeding risk appear on the right with high SHAP values, while features most predictive of low bleeding risk appear at the left. To further understand dynamic risk predictions following a decision, shift tables were generated. These tables are useful for visualizing changing risks of bleeding before and after a decision or for comparing the performance of the initial and final models. To describe these changes, patients were classified into categories of low risk ($<$1%), moderate risk (1%-4%), or high risk ($>$4%) of bleeding. These thresholds were chosen to approximately balance patients between categories across all models [209]. Table 8.3 shows shift tables before and after each decision, while Table 4 shows a shift table from the initial to final model. The earlier model is displayed left to right, while the later model is displayed top to bottom. The top number in each cell represents the number of patients assigned to that risk bin by each model. The bottom number in each cell is the overall bleeding rate of all patients in that cell. NaN represents that no patients were in that combination of bins.

All analyses were conducted in Python version 3.8.6 or R version 4.0.3. Data analysis was performed using scikit-learn 0.24.1 [67] and XGBoost 1.3.3 [206] for gradient descent boosting. SHAP explanations were generated and visualized with SHAP 0.38.1 [208]. Model calibrations generated in R with mgcv 1.8-33 [210] and calibration variances with sandwich

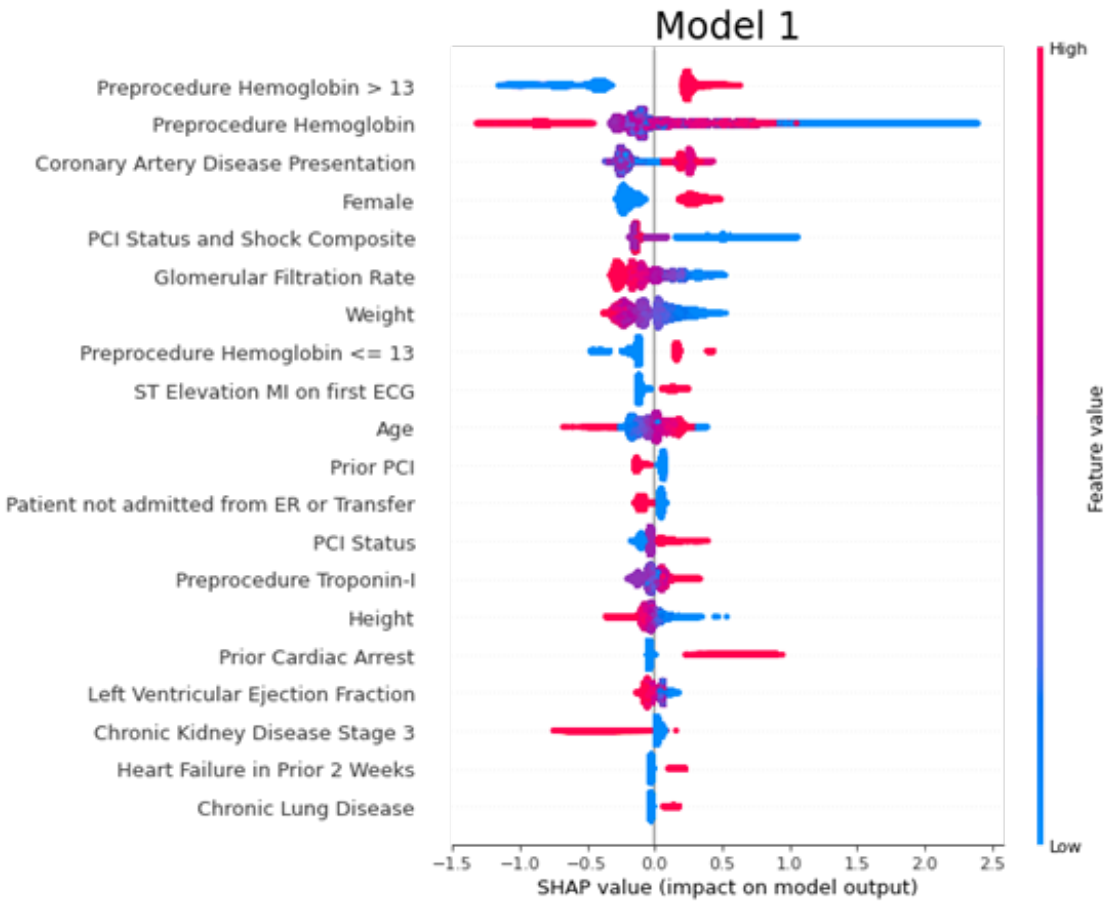


Figure 8.2: SHAP Tree explainer for Model 1

3.0-0 [211]). Source code is available online.

8.3 Results

8.3.1 Patient Cohort and Variables Used

We included 2,868,808 PCIs in the NCDR CathPCI registry; 2,314,446 (80.7%) prior to 2014 for model training and 554,362 (19.3%) after 2014 for validation and model interpretation. The mean (SD) age of patients was 64.6 (12.0) years and 68.3% were male (Table 8.1). Overall, there were 118,327 (4.1%) major bleeding events: 98,167 (4.2%) major bleeding events in the training set and 20,160 (3.6%) in the validation set.

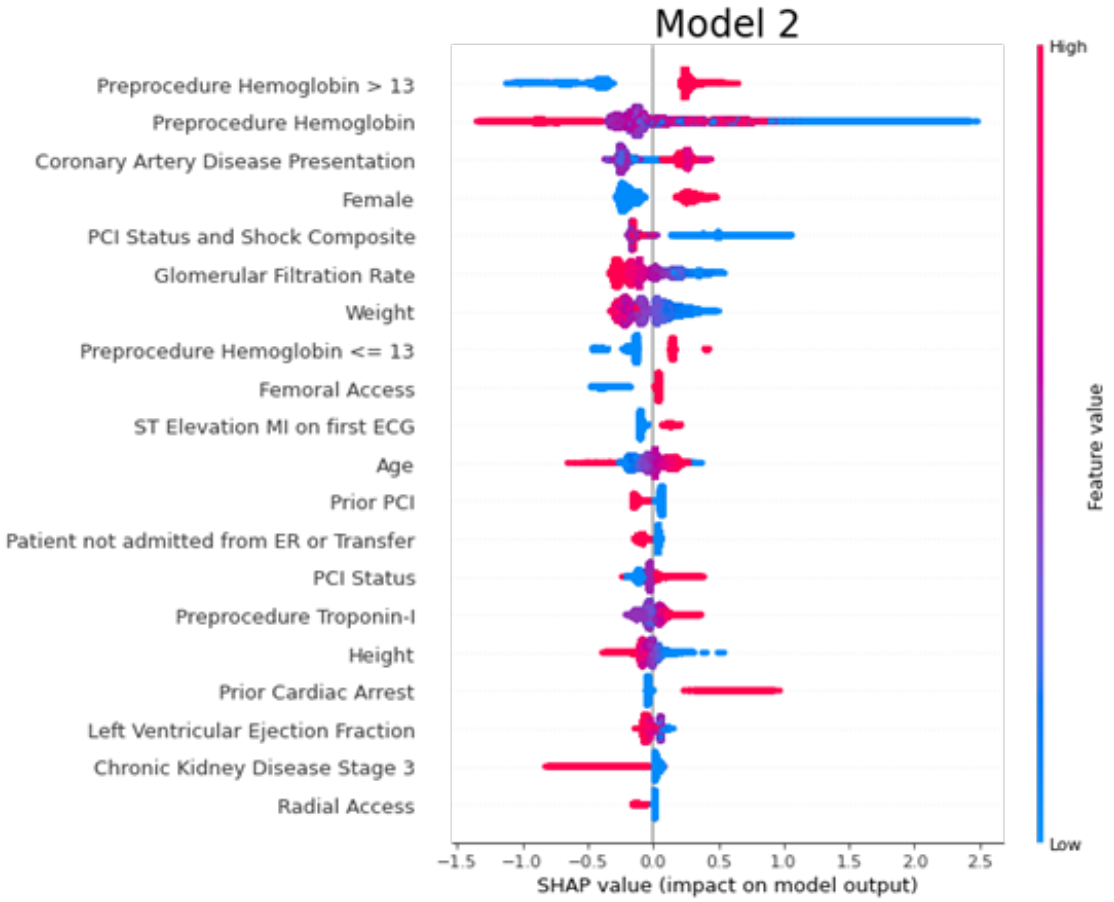


Figure 8.3: SHAP Tree explainer for Model 2

8.3.2 Stage 1: Clinical Presentation (Model 1)

The initial model uses information available to a clinician at the time that a patient presents and predicted bleeding risk with an AUROC of 0.812 and AUPRC of 0.203. The Brier skill score of this model is 0.088, representing the degree that this model improves over a naïve model (higher is better). The Brier reliability is 2.6E-4, representing distance to true probabilities (lower is better), while the resolution is 3.3E-3, representing forecast distances to the mean rate (higher is better). The SHAP plot shows that this model is most strongly driven by preprocedural hemoglobin and coronary artery disease symptoms at presentation (Figure 8.2). The higher a variable is in Figure 8.2, the greater overall importance that variable exhibits for the model. Red values indicate high values for that

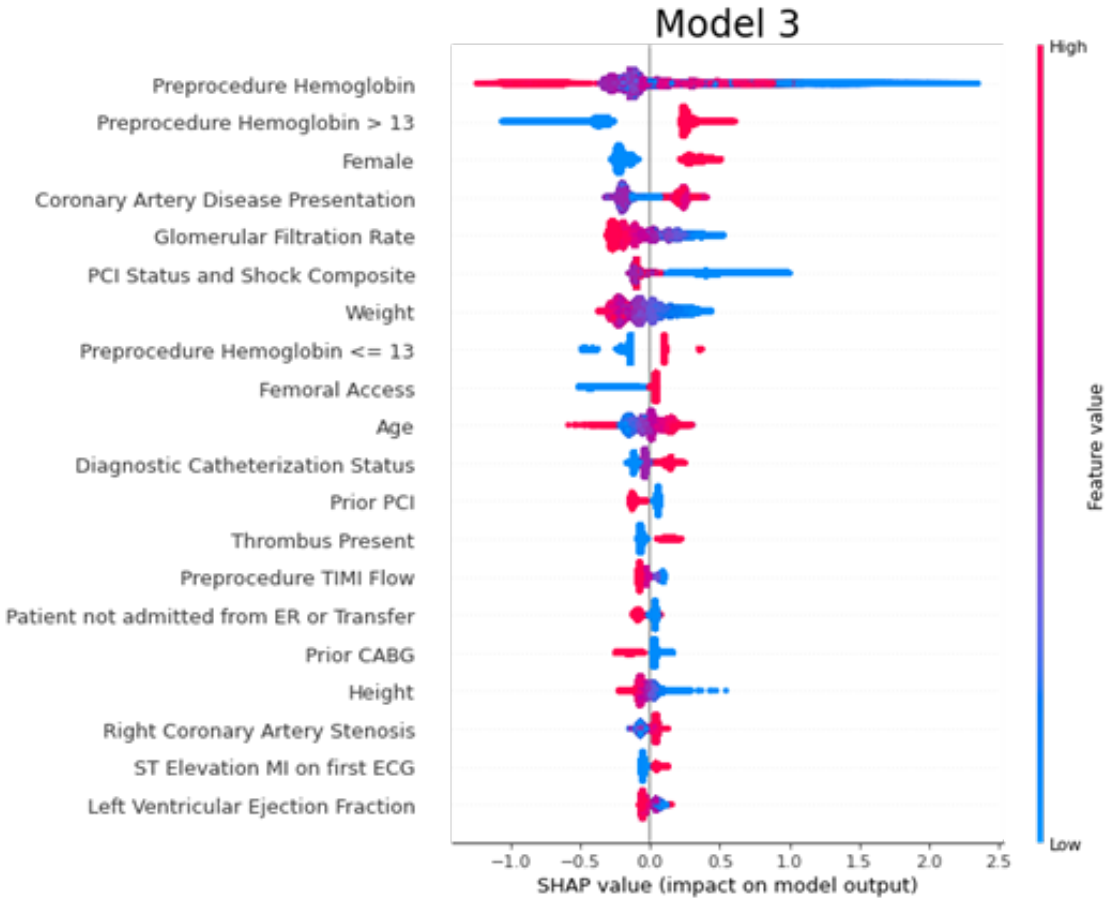


Figure 8.4: SHAP Tree explainer for Model 3

variable (if continuous or ordinal) or "true" (if binary), while blue values indicate the opposite. Points to the right of the axis (positive SHAP values) indicate that a feature of that value increases model estimate of bleeding risk, while points to the left of the axis (negative SHAP values) indicate that a feature of that value decreases the model estimate of bleeding risk. For instance, on the top row of Figure 8.2, a preprocedural hemoglobin greater than 13 is associated with an increased risk of bleeding, while value less than 13 is associated with a decreased risk. The wide range of SHAP values for this variable shows that while the direction of association is constant, the degree to which this feature impacts risk is not constant.

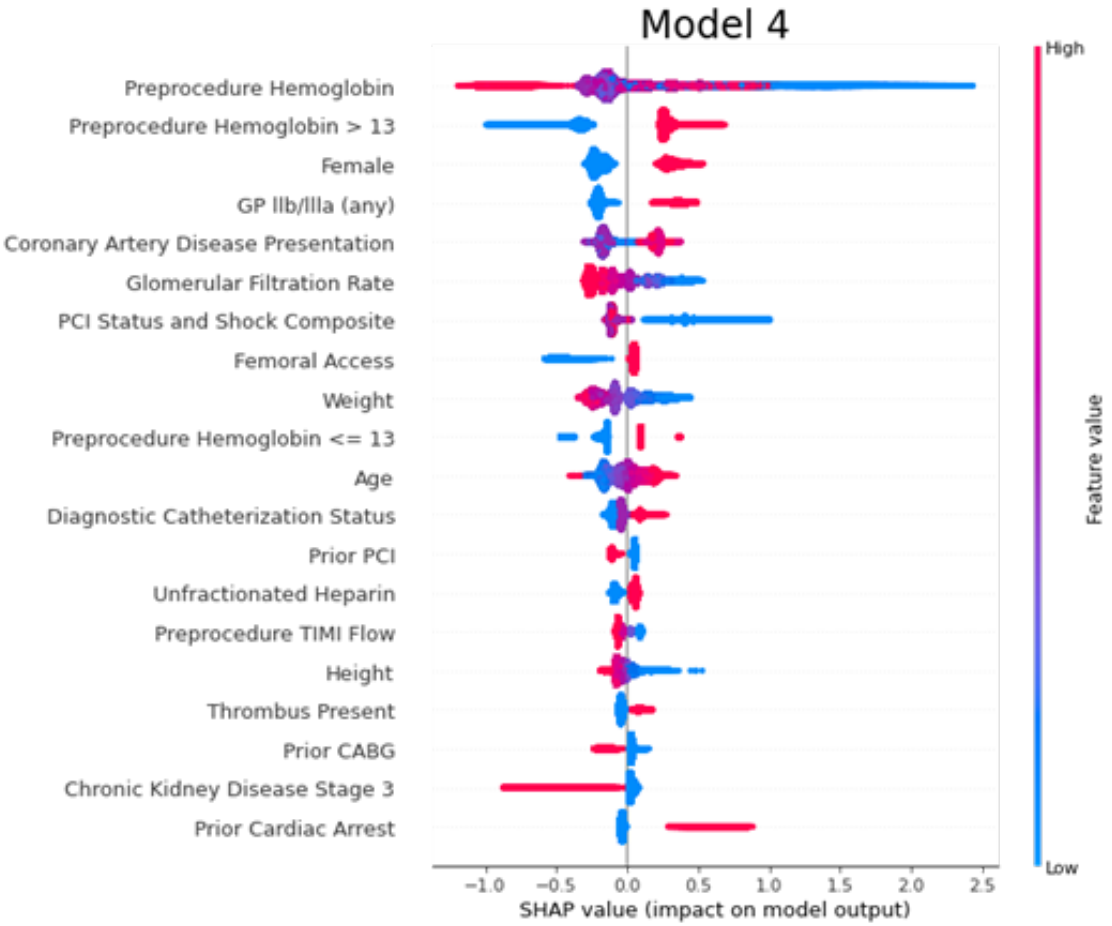


Figure 8.5: SHAP Tree explainer for Model 4

8.3.3 Decision 1: Access Site (Model 2)

The first decision point in this model is the choice of arterial access site: femoral or radial. When accounting for this decision, the AUROC improved to 0.817 and the AUPRC improved to 0.204. The Brier skill score improved to 0.091, the Brier reliability improved to 2.0E-4. The SHAP plot shows that femoral access is the ninth most informative variable in Model 2 (Figure 8.3). Procedures performed via a femoral access site had a slightly increased rate of bleeding, while those performed via a radial access site had a variably decreased rate of bleeding. Procedures performed via femoral access are represented by the narrow red line to the right of the axis. The fact that these procedures have similar SHAP values indicates

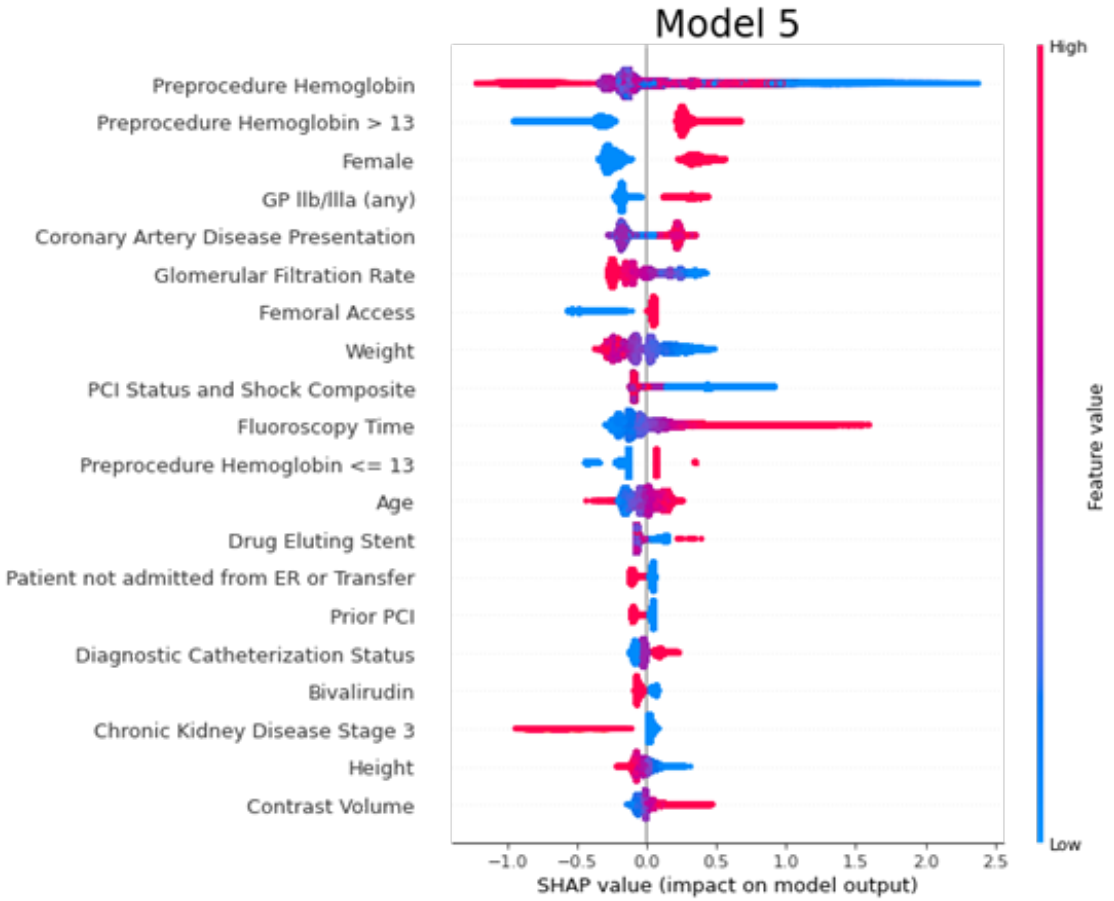


Figure 8.6: SHAP Tree explainer for Model 5

that the model assigns similar risk to procedures with femoral access. In contrast, the blue points to the left of the axis represent procedures performed with radial access, and their elongated shape indicates that the radial access has a variable effect on bleeding risk, with the risk for some procedures being decreased by much more than the risk for others.

Table 8.3a presents a shift table describing patient risk categories (23). Among 123,712 patients classified as low (<1%) risk of bleeding by the clinical presentation model, 9,071 (7.3%) were reclassified as medium (1-4%) risk of bleeding by the model incorporating access site (Table 8.3a). Among 270,485 patients classified as medium risk of bleeding by the initial model, 33,129 (12.2%) were reclassified as low risk by the subsequent model, while 6,465 (2.4%) were reclassified as high (>4%) risk. Among 160,165 patients classified as high risk

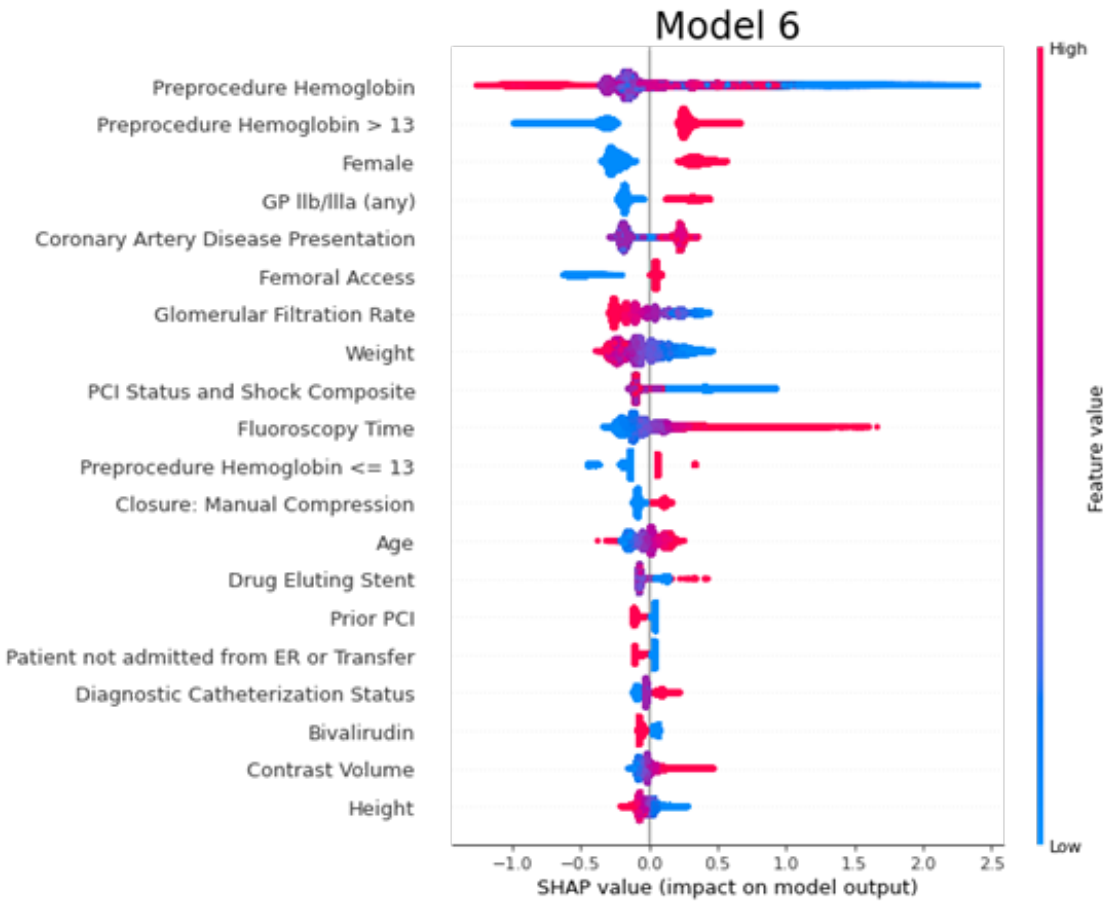


Figure 8.7: SHAP Tree explainer for Model 6

of bleeding by the initial model, 14,582 (9.1%) were reclassified as medium risk.

8.3.4 Stage 2: Cardiac Catheterization Laboratory (Model 3)

The cardiac catheterization laboratory model uses information available after performing a diagnostic cardiac catheterization, but prior to choice of peri-procedural medications and initiation of PCI. In this model, bleeding prediction improved with an AUROC of 0.825 and an AUPRC of 0.208. The Brier skill score improved to 0.094, the Brier reliability worsened to 2.1E-4, and the Brier resolution improved to 3.5E-3. Of features added in this model, the SHAP plot shows that predictions are highly influenced by the presence of thrombi in coronary lesions (13th most informative variable) and pre-procedure Thrombolysis in Myocardial Infarction (TIMI) flow (14th most informative variable) (Figure 8.4).

Table 8.3: Shift tables following each decision point. Top value in each cell is number of patients classified into that risk bin by the two respective models. The bottom value in each cell indicates the actual bleeding rate of all patients within that cell.

Initial vs After Access Site Decision				
	Model 1			
	<1%	1-4%	>4%	All
Model 2	Patients, N Observed Rate	Patients, N Observed Rate	Patients, N Observed Rate	Patients, N Observed Rate
<1%	114,641	33,129		147,770
	0.57%	0.54%	NaN	0.56%
1-4%	9,071	230,891	14,582	254,544
	0.99%	1.69%	2.54%	1.72%
>4%		6,465	145,583	152,048
	NaN	3.11%	10.14%	9.84%
All	123,712	270,485	160,165	554,362
	0.60%	1.58%	9.45%	3.64%

(a) Shift table comparing Model 1 to Model 2.

In Cath Lab Before vs After Medication Decisions				
	Model 3			
	<1%	1-4%	>4%	All
Model 4	Patients, N Observed Rate	Patients, N Observed Rate	Patients, N Observed Rate	Patients, N Observed Rate
<1%	140,103	32,764		172,867
	0.43%	0.64%	NaN	0.47%
1-4%	11,244	207,541	21,612	240,397
	1.20%	1.67%	3.24%	1.79%
>4%		11,448	129,650	141,098
	NaN	4.37%	11.22%	10.67%
All	151,347	251,753	151,262	554,362
	0.49%	1.66%	10.08%	3.64%

(b) Shift table comparing Model 3 to Model 4.

Post PCI Before vs After Closure Decision				
	Model 5			
	<1%	1-4%	>4%	All
Model 6	Patients, N Observed Rate	Patients, N Observed Rate	Patients, N Observed Rate	Patients, N Observed Rate
<1%	178,911	12,095		191,006
	0.43%	0.54%	NaN	0.44%
1-4%	8,561	213,367	7,268	229,196
	0.81%	1.73%	3.56%	1.76%
>4%		5,703	128,457	134,160
	NaN	3.30%	11.76%	11.40%
All	187,472	231,165	135,725	554,362
	0.45%	1.71%	11.32%	3.64%

(c) Shift table comparing Model 4 to Model 5.

8.3.5 Decision 2: Pre-Procedure Medication (Model 4)

The next decision point is the choice of intra-procedural antiplatelets and anticoagulants. Following inclusion of this decision, the model performance increases to an AUROC of 0.832 and an AUPRC of 0.217. The Brier skill score improved to 0.102, the Brier reliability improved to 1.4E-4, and the Brier resolution improved to 3.7E-3. Of features added in this model the SHAP plot shows that use of GP IIb/IIIa inhibitors was strongly associated with an increased risk of bleeding (4th most informative variable), while unfractionated heparin was less strongly associated with an increased risk of bleeding (14th most informative variable) (Figure 8.5).

Among 151,347 patients classified as low risk of bleeding by the cardiac catheterization lab model (Model 3), 11,244 (7.6%) were reclassified as moderate risk by the model incorporating medication choices (Model 4), (Table 8.3b). Among 251,753 patients classified as moderate risk of bleeding by Model 3, 32,764 (13.0%) were reclassified as low risk by Model 4, while 11,448 (4.5%) were reclassified as high risk. Among 151,262 patients classified as high risk of bleeding by Model 3, there were 21,612 (14.3%) patients who were reclassified as moderate risk by Model 4.

8.3.6 Stage 3: PCI (Model 5)

The post-PCI model uses all information through PCI but prior to choice of closure method. This model improves upon the performance of prior models, with an AUROC of 0.844 and AUPRC of 0.241. The Brier skill score improved to 0.118, the Brier reliability improved to 1.1E-4, and the Brier resolution improved to 4.2E-3. The new features introduced in this model that were most associated with increased bleeding risk are proxies of PCI complexity and duration, including fluoroscopy time (10th most informative variable) and contrast volume (20th most important variable) (Figure 8.6). Use of a drug eluting stent was variably associated with risk of bleeding (13th most important variable).

8.3.7 Decision 3: Closure Method (Model 6)

The final decision point is closure. This decision had minimal effect on overall prediction, with AUROC remaining at 0.845 and AUPRC improving slightly to 0.242. The Brier skill score improved slightly to 0.119, the Brier reliability improved to 1.0E-4, and the Brier resolution improved to 4.3E-3. Figure 8.7 shows a SHAP explanatory plot for this model. Manual compression is the closure method most strongly predictive of increased bleeding risk (12th most informative variable).

Among 187,472 patients classified as low risk of bleeding by the post-PCI model (Model 5), 8,561 (4.6%) were reclassified as moderate risk by the model incorporating closure decision (Model 6) (Table 8.3c). Notably, those patients had a bleeding rate of 0.81%, suggesting that this reclassification on average may have been overly pessimistic. Among 231,165 patients classified as moderate risk of bleeding by Model 5, 12,095 (5.2%) were reclassified as low risk by Model 6, while 5,703 (2.5%) were reclassified as high risk. While the patients reclassified as low risk had an appropriately low rate of bleeding in aggregate (0.54%), those patients reclassified to high risk had a moderate aggregated rate of bleeding (3.30%). Among 135,725 patients classified as high risk of bleeding by Model 5, there were 7,268 (5.4%) patients who were reclassified as moderate risk by Model 6.

Total reclassification from the initial model to the final model is shown in Table 4. Among 123,712 patients classified as low risk by the initial model, 14,441 (11.7%) were reclassified as moderate risk, while 723 (0.6%) patients were reclassified as high risk. Notably, the bleeding rate among those patients reclassified to high risk was 12.5%. Among 270,485 patients classified as moderate risk by the initial model, 82,418 (30.5%) were reclassified to low risk by the final model, while 16,577 (6.1%) were reclassified to high risk by the final model. Finally, among 160,165 patients classified as high risk by the initial model, there were 40 (<0.1%) patients reclassified to low risk, and 43,265 (27.0%) patients reclassified to moderate risk.

Table 8.4: Shift table across models. Top value in each cell is number of patients classified into that risk bin by the two respective models. The bottom value in each cell indicates the actual bleeding rate of all patients within that cell.

Initial vs Final Estimate				
	Model 1			
Model 6	<1% Patients, N Observed Rate	1-4% Patients, N Observed Rate	>4% Patients, N Observed Rate	All Patients, N Observed Rate
<1%	108,548	82,418	40	191,006
	0.41%	0.48%	0.00%	0.44%
1-4%	14,441	171,490	43,265	229,196
	1.47%	1.59%	2.50%	1.76%
>4%	723	16,577	116,860	134,160
	12.45%	6.99%	12.02%	11.40%
All	123,712	270,485	160,165	554,362

8.3.8 Case Studies

Because severe bleeding is a relatively rare event, risk for most patients will change by a small amount. However, some patients' risk changes dramatically throughout the course of their treatment. Overall, the median difference of risk from the initial to final prediction is -0.41% (IQR -1.16%, +0.02%). However, the full range of risk changes was much larger, ranging from -44.42% to +83.21%. To understand the factors that drive risk prediction, two patients with illustrative risk profiles were chosen as case studies.

8.3.8.1 Case Study A

A 63-year-old man presented for emergent PCI for STEMI. In the initial model, his risk of bleeding was estimated to be 4.3%. This risk was driven predominantly by the emergent need for PCI, his preprocedural hemoglobin, his coronary artery disease on presentation, his sex, and his weight. The decision was made to use radial access, after which his bleeding risk was estimated to be 2.8%. There were no factors that increased predicted risk during diagnostic coronary angiography, and risk decreased to 2.1%. He received prasugrel and

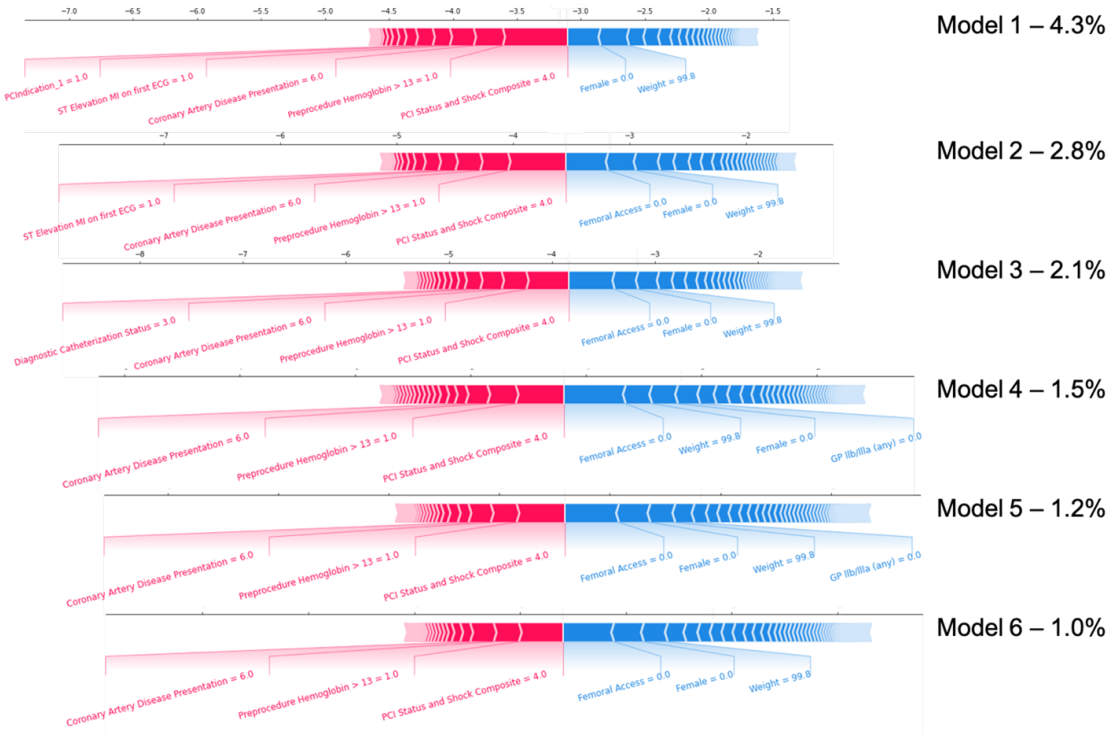


Figure 8.8: SHAP explainer for Case Study A

unfractionated heparin, after which his risk fell to 1.5%. Following successful PCI, his risk further fell to 1.2%. His access site was sutured, and the final bleeding risk was 1.0%. Plots of individual SHAP values for this patient at each model stage are shown in Figure 8.8.

8.3.8.2 Case Study B

A 66-year-old woman presented for elective PCI for stable angina. In the initial model, her risk of bleeding was 0.9%. This risk was driven predominantly by her preprocedural hemoglobin, her sex, the stable nature of her coronary artery disease, and her weight. The decision was made to use femoral access, after which her risk increased to 1.0%. Upon diagnostic coronary angiography, it was discovered that she had significant stenosis present. The model at this stage estimated risk of bleeding to be 1.7%. She received unfractionated heparin and clopidogrel, after which her estimated risk was 1.4%. PCI was notable for high complexity (reflected by a long fluoroscopy time), after which her risk of bleeding increased

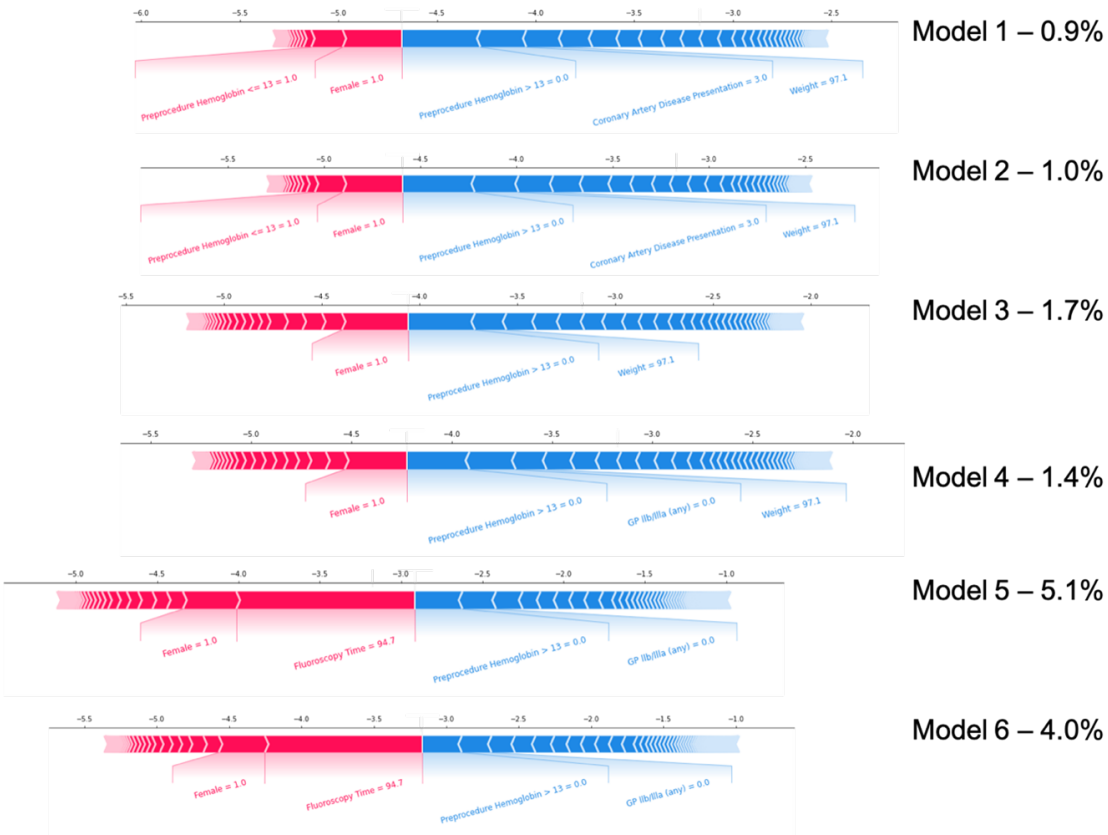


Figure 8.9: SHAP explainer for Case Study B

to 5.1%. Her access site was sutured, and her final risk was 4.0%. Ultimately, this patient experienced a bleed. Plots of individual SHAP values for this patient at each model stage are shown in Figure 8.9.

8.4 Discussion

8.4.1 Limitations and Future Directions

There are several key limitations to our study. First, the real-time data availability are a significant challenge. The NCDR relies upon manual chart abstraction for many variables, such as past medical history variables. The implementation of such a system within an electronic health record environment, where data may be available within near real-time capacity, requires additional investigation of natural language processing techniques, models that handle advanced time-series data, and appropriate evaluation of user interface design

for reducing clinician burden when interacting with such a model.

The second is the timing of the available variables. The registry abstracts some variables as pre- and intra- procedure, so some confounding may exist from reactions to bleeds during the procedure. This nature of the variables could prevent accounting for some intra-procedural variables, such as if there was initially acute closure after PCI requiring additional steps in the procedure.

8.5 Conclusion

By evaluating risk at different stages of patient care, we present a model that provides up-to-date and dynamic bleeding risk estimates. These advanced methods demonstrate evolution in variable importance as clinical decisions are made through course of PCI. These models hold significant potential to provide updating information that informs clinical decision-making that can mitigate bleeding risk.

9. OUTCOMES-DRIVEN CLINICAL PHENOTYPING IN CARDIOGENIC SHOCK USING A MIXTURE OF EXPERTS*

Returning to the cohort discussed in Chapters 6 and 7, we now look at developing a method beyond propensity matching for cohort analysis. As discussed, a key limitation of propensity matching is that while care is taken to group by similarity, at a fundamental level, the propensity score is a scalar used for matching. Here, we apply the deep mixture of experts approach in order to jointly learn phenotypes and predict outcomes.

9.1 Introduction

Cardiogenic shock (CS) is a life threatening condition where the heart is unable to sufficiently supply blood. Despite being an active area of research, in-hospital and 30-day mortality from CS remains near 40% [212, 213]. Assisting the heart with a mechanical circulatory support (MCS) device such as an intra-aortic balloon pump (IABP) or an intravascular microaxial left ventricular assist device (Impella®) is a common strategy to treat CS. However, evidence proving the treatment effect of MCS devices is lacking [212]. Several RCTs have attempted to assess the treatment effectiveness of these devices, but have failed to show difference in mortality outcomes [177]. Despite this lack of evidence, device usage remains high [214].

Given the difficult nature of performing randomized prospective studies with MCS devices, observational studies are an attractive approach for hypothesis generating studies. A recent retrospective study performed propensity score matching and found that Impella® use was associated with a higher risk of in-hospital death in comparison to IABP [153]. However, the study could have been influenced by unmeasured confounders, a drawback of propensity score matching. While propensity matching on a full population may provide some insights, unidentified clinical phenotypes, representing homogeneous subpopulations

*This chapter is reprinted with permission from SUBMISSION PENDING

within the heterogeneous cohort, may actually have an opposite relationship with the outcome [215]. Identifying these phenotypes could help explain treatment effectiveness while appropriately accounting for risk.

A technique is needed that jointly learns phenotypes while also learning patient risk of death. Prior selection of features by clinicians may yield interesting results, but are likely to identify phenotypes that match current clinical understanding, limiting discovery of novel phenotypes. Using a supervised learning technique to model risk, then looking at treatment effectiveness heterogeneity has numerous causal inference implications and challenges [216, 217]. We propose here a two stage model that jointly learns phenotypes and verifies risk stratification through accurate risk models. We propose this model as a deep mixture of experts (MoE) approach [218, 219] to jointly address both risk modeling and clinical phenotyping in observational clinical data. We apply the technique designed by [219] for human activity recognition to the principles of CS using data from [153] and find that we appropriately quantify risk as well as phenotype jointly. We are better able to explore the impact of the use of Impella® and IABP with patients who are better matches than those that do not consider learning risk factors for major adverse events.

9.2 Related Work

Identification of heterogeneity in disease can allow for improved understanding of disease processes and in treatment approaches. Unsupervised and semi-supervised clinical phenotyping approaches are valuable techniques given the vast amount of electronic health data produced and the sparsity of known phenotype labels. Classical unsupervised techniques are one approach for finding clinical phenotypes [220]. Deep learning techniques have expanded the scope of data-driven strategies in clinical phenotyping. Reference [53] used sparse autoencoders followed by t-distributed stochastic neighbor embedding to identify subpopulations in patients with a shared laboratory finding to distinguish very different pathologies. Reference [221] used uniform manifold approximation and projection (UMAP) to extract clusters from patients presenting to an emergency room with given clinical complaints.

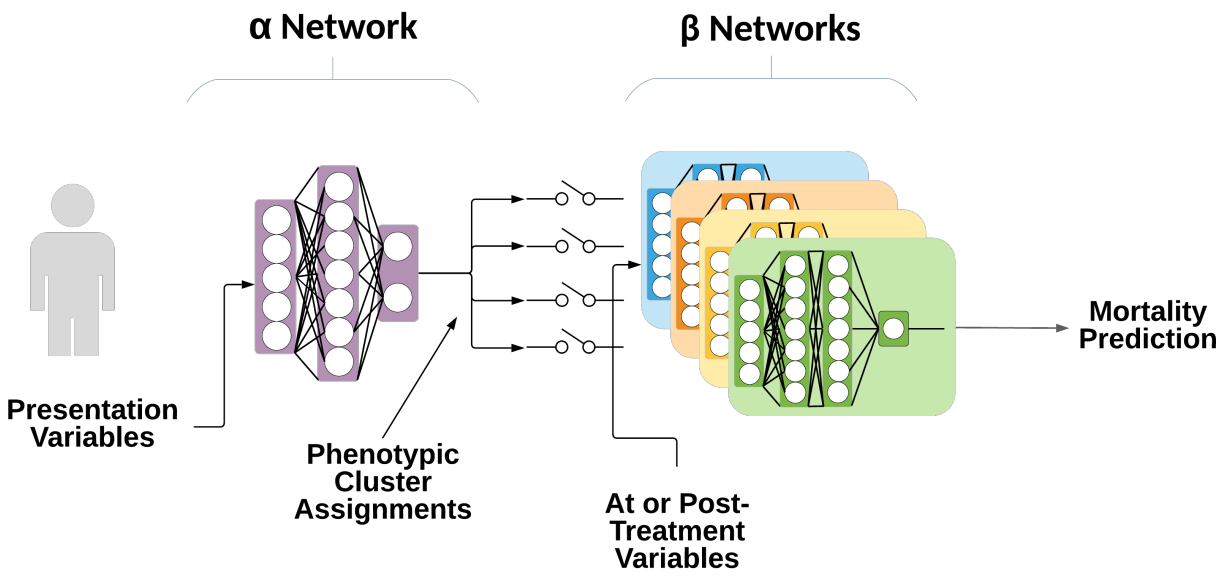


Figure 9.1: Deep MoE model for clustering and predicting clinical outcomes.

Deep learning has also been utilized in causal effect inference [222, 223]. One approach has been to perform causal inference using variational autoencoders and inferring causal structure within the latent space of those encoders [222]. Another approach has been to estimate treatment effect using local similarities within a latent space representation [223].

9.3 Methods

The objective of this work is two-fold: 1) A supervised learning problem where clinical outcomes are predicted from patient features; 2) An unsupervised learning problem of finding phenotypes among the patient population. We assume that treatment decisions are based on presenting characteristics and that outcomes are based on presenting characteristics and treatments. While no direct causal inference is made here, this work lays the groundwork for future causal inference. The objectives are learned from information in three tiers: initial data, data following treatment, and outcomes.

The three tiers of information result in a staged model design. In the first stage, initial information is fed into a fully connected neural network, the α-network, with two outputs. The first output feeds into the second stage, while the second output gives a softmax distribution

of probabilities that the patient belongs to any of n clusters, where n is a searchable hyper-parameter denoting the number of experts in the model. In the second stage, information from the point of treatment onward is concatenated with the output from the first stage and fed into n fully connected neural networks, the β-networks. Each of these networks estimate the likelihood of outcome occurrence which is then weighted by the probability distribution given by the α-network. What follows is our modification to the $\alpha\beta$-network first introduced by Huo et al. for clinical data and objectives, trained as in their work [219]. An illustration of the model is shown in Figure 9.1. Models were implemented in Tensorflow 2.3.2.

There are five different configurations of the overall model: three of different overall size, and two with variant connections between the α-network and the β-network. In the largest network, the α-network and each β-network are composed of 3 fully connected layers with 50 nodes, the mid-sized network features 2 fully connected layers with 24 nodes, and the smallest network features 2 fully connected layers with 24 nodes in the α-network, and 1 fully connected layer with 24 nodes in each β-network. Those with variant connections are broadly constructed in nearly the same way as the largest network. However, in one variant, there is no connection from the α-network's fully connected layers into the β-networks' fully connected layers- the only connection is via the gating function of the α-network. In the other variant, the α-network is concatenated with the second fully connected layer of the β-network rather than the first.

Rather than being used exclusively as a monolithic model, this model is intended to function in two modalities. In the simplest case, this model performs well in the supervised learning task. In the second modality, after being trained on data incorporating all tiers of information, the model can be assign clusters learned during the training from the α-network. This cluster assignment, although learned with gradients incorporating information from outcomes and treatment decisions, only requires information from the earliest tier. By this means, the model can be trained on one set while producing cluster assignments based only on factors present prior to treatment.

Table 9.1: Clinical dataset mortality AUROC values for all models with L2 regularization = 0.01.

	Baseline	2 Experts	4 Experts	5 Experts	6 Experts	8 Experts	10 Experts
XGBoost	**0.880 ± 0.006**						
Logistic Regression	0.839 ± 0.004						
Small Model		**0.843 ± 0.011**	**0.851 ± 0.006**	**0.848 ± 0.009**	**0.851 ± 0.004**	**0.849 ± 0.008**	**0.848 ± 0.005**
Mid-Sized Model		0.841 ± 0.013	0.846 ± 0.011	0.842 ± 0.003	0.843 ± 0.007	0.841 ± 0.008	0.840 ± 0.011
Large Model		0.763 ± 0.098	0.822 ± 0.012	0.814 ± 0.009	0.812 ± 0.006	0.810 ± 0.013	0.818 ± 0.017
Later Hidden Connection		0.626 ± 0.132	0.559 ± 0.108	0.688 ± 0.134	0.701 ± 0.107	0.756 ± 0.097	0.645 ± 0.115
No Hidden Connection		0.618 ± 0.115	0.755 ± 0.040	0.722 ± 0.054	0.777 ± 0.007	0.712 ± 0.071	0.729 ± 0.079

9.3.1 Number of Experts

Much work has been done on very large deep MoEs with huge numbers of experts [218]. Here we limit the number of experts to be searched to 10, aiming to constrain patient populations to at most 10 clusters. This limit was chosen in order to maximize the clinical interpretability of any discovered clusters, but also protects this model from the shrinking batch problem. The objective of this work is not only to generate experts, but to harness them in a way that their utilization and gating can be of use in the clinic.

9.3.2 Baseline Models

For baseline outcome predictions, logistic regression and XGBoost were used. Logistic regression represents a versatile linear classifier (with L2 regularization), while XGBoost is a powerful nonlinear classifier (with hyperparameters: 0.3 learning rate, maximum depth of each tree 6, and number of trees 100).

9.3.3 Metrics

For the supervised task, area under the receiver operating characteristic curve (AUROC) is reported. Similarity of different partitionings into clusters is assessed with the adjusted Rand index (ARI). ARI is used as a way to assess clustering stability between different model folds, where zero indicates that any similarity in the paritionings is likely due to chance and one indicates similar clustering.

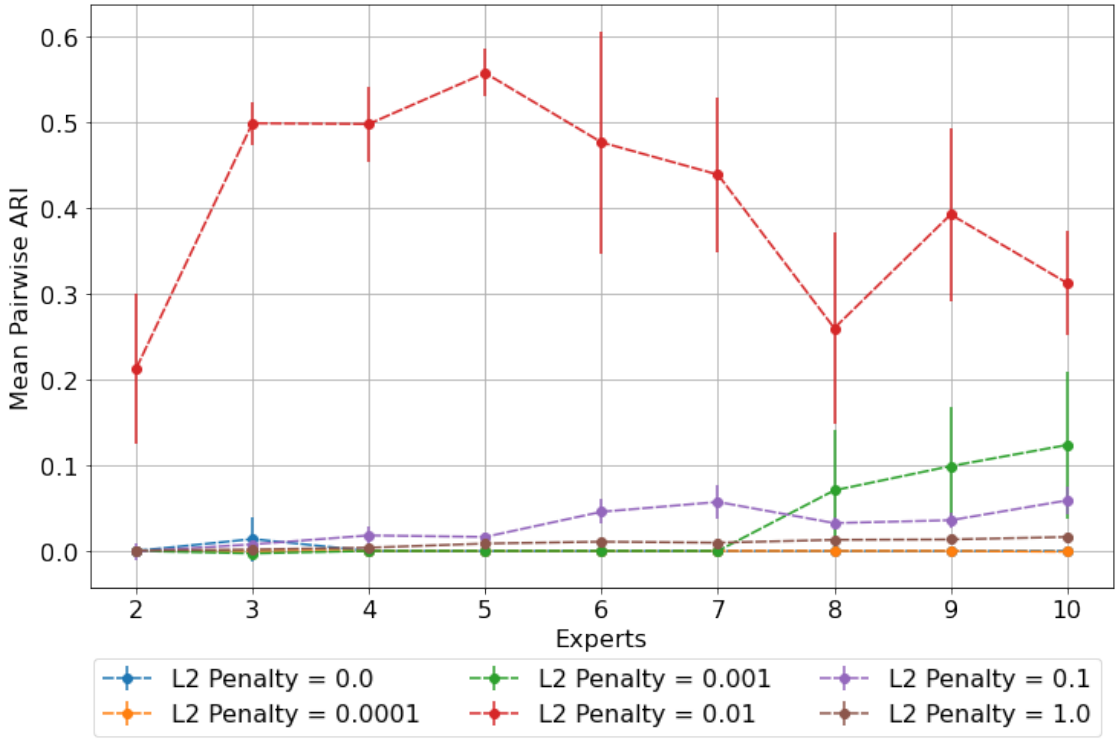


Figure 9.2: Mean pairwise ARI given n experts and various L2 penalties in the clinical dataset. Confidence bars express 95% CI.

9.4 Experiment and Results

The clinical dataset used in this project is derived from two linked American College of Cardiology registry datasets from the National Cardiovascular Data Registry (NCDR) [154]: the Chest Pain-MI registry and the CathPCI registry. These registries include well-curated, independently audited data on patient characteristics, clinical features, and in-hospital outcomes for all patients admitted for an acute myocardial infarction (AMI) in the Chest Pain-MI registry or undergoing percutaneous coronary intervention (PCI) in the Cath-PCI registry from over 1500 participating clinical institutions, with extremely low rates of missing data or outliers due to noise [129]. This study was reviewed and approved by the human investigation committee (TAMU IRB # 2018-0856).

The clinical cohort as first described by [153] included linking episodes of care through the two registries with a probabilistic linkage technique, resulting in 28,304 CS patients. Twelve

presentation features were selected for clustering and 15 additional features were selected for mortality prediction [153]. Five-fold cross validation was used for every experiment. When training neural networks, 20% of the available training data was used as an internal validation split for hyperparameter tuning. All results reported here reflect the mean ± 95% confidence interval over the cross validated results.

The final complete model produces an estimation of the outcome occurring. This estimate is informed by decisions made at treatment time and information learned after treatment, and so should not be used for describing patient phenotypes. However, the multilayered setup of the model allows for similarity groupings made entirely prior to treatment choices. While the model is learned with the insight provided by the later information, it is applicable to patients prior to that point. This allows for similar groups of patients to be found prior to treatment, allowing for group characterization. This is similar to the overall goal of techniques such as propensity matching, but allows for extension of the classification method to subjects lacking treatment and outcome information. Future extensions of this work could be to analyze patient clusters and to produce improved treatment effect estimations for the subpopulations of those clusters, rather than on the population as a whole.

Each model was trained on the clinical dataset with initial and treatment variables as described above. The outcome prediction AUROCs for all models is shown in Table 9.1. The small and mid-sized models each have performance comparable to one of the baselines, while the others show less discrimination. However, the focus of this work is not primarily to predict outcomes, but to use that prediction in classification.

Figure 9.2 shows a plot of pairwise ARI values among each fold of the small model. Following each cross validation fold, the entire dataset was classified using the α-network. (A single monolithic cluster is undesirable; this value was set to 0 for any pairings involving a singular clustering.) Despite the lack of a ground truth label, the best-performing model shows a significantly higher level of inter-fold cluster stability.

In all folds of this highest-scoring model, there was one large cluster, smaller clusters, and

Table 9.2: Clinical cluster characteristics. STEMI = ST Elevation Myocardial Infarction. MD = Multivessel Disease

	C0	**C1**	**C2**	**C3**	**C4**
Number	21,728	2,809	2,053	1,639	75
Age (SD)	68.2 (11.8)	54.9 (9.4)	56.1 (7.5)	52.3 (9.5)	46.0 (7.5)
Male	61.14%	92.49%	81.39%	81.82%	97.33%
Smoker	31.06%	48.74%	55.14%	52.47%	44.00%
STEMI	81.45%	80.99%	71.02%	55.09%	97.33%
MD	56.83%	71.09%	19.48%	48.81%	18.67%
Urgent PCI	11.03%	11.82%	19.82%	30.69%	0.00%
Emergent PCI	78.78%	82.48%	75.35%	64.37%	93.33%
Mortality	30.33%	14.63%	7.45%	7.26%	10.67%
Only IABP	30.3%	35.2%	22.7%	25.1%	22.7%
Only Impella®	6.5%	7.7%	3.4%	4.3%	1.3%
No MCS Device	56.7%	49.1%	68.4%	65.3%	75.3%

Table 9.3: Per-cluster mortality rate. Gray values reflect those groups for which there is not a significant difference between that group and the corresponding group in the total population.

	Entire Group	Only IABP	Only Impella®	None
C0	30.3 (29.7-30.9)	33.9 (32.7-35.0)	49.8 (47.1-52.4)	24.3 (23.5-25.0)
C1	14.6 (13.3-16.0)	14.5 (12.3-16.8)	29.6 (23.6-36.2)	8.6 (7.2-10.2)
C2	7.5 (6.4-8.7)	10.5 (7.9-13.7)	25.7 (16.0-37.6)	4.2 (3.2-5.4)
C3	7.3 (6.1-8.6)	7.8 (5.4-10.8)	22.9 (13.7-34.4)	5.0 (3.7-6.4)
C4	10.7 (4.7-19.9)	11.8 (1.5-36.4)	0.0 (0.0-97.5)	9.1 (3.0-20.0)
All	25.7 (25.2-26.2)	29.0 (28.1-30.0)	45.2 (42.9-47.6)	19.9 (19.3-20.5)

one empty or nearly empty cluster. Cluster features are shown in Table 9.2. Patients in the largest cluster are generally older, more likely to be female, and have a higher mortality rate than the other clusters. Patients in the next cluster were younger, more likely to be male, and had lower mortality rates despite having increased MCS device utilization. Patients in the next two clusters were of similar age, gender, and mortality. However, those in cluster 2 were differentiated by much higher rates of emergent PCI featuring much lower rates of multivessel disease. The last cluster was very small. Patients in Cluster 1 were at higher risk

of receiving IABP in comparison to the overall population, and patients in Cluster 2 were at lower risk of receiving IABP. Patients in Cluster 1 were more likely to receive any type of MCS device, while patients in Clusters 2, 3, and 4 were less likely to receive any type of MCS device.

Mortality rates in each clinical cluster and by MCS device are shown in Table 9.3. This table is shown with 95% confidence intervals as calculated by the Clopper-Pearson method. Among given MCS utilizations, patients in cluster 0 have significantly higher mortality than patients with the same MCS utilization in clusters 1, 2, or 3. In every cluster except cluster 4 the mortality rate among patients receiving only Impella® was higher than the mortality rate among patients receiving only IABP, and patients with IABP had a significantly higher mortality rate than patients without an MCS device.

9.5 Limitations and Future Directions

This work features several aspects that will be expanded in future work. The regularization term here is pivotal to the model's function. One approach would be to extend this with an additional L1 penalty. Another uses a Bayesian model, reformulating the α-network to output a probabilistic distribution of experts. The clustering here loses some nuance of the α-network gating. The gating is a soft gating, with each β-network output being multiplied by its share of the α-network's softmax activation, but the analysis here classifies each subject into a single cluster (via argmax).

9.6 Conclusion

CS is a life-threatening condition with an extremely high mortality rate. Given this high loss of life, it is desirable to understand the heterogeneities of patients with CS. Given the difficulty in collecting randomized data, development of techniques applied to retrospective data is necessary. A technique is needed that jointly learns phenotypes while also learning patient risk. We apply a deep MoE here to find phenotypes among patients with CS. We predict mortality with an AUROC of 0.85 ± 0.01, finding five interpretable clusters. The

largest cluster is generally older, sicker, and at higher risk of mortality. The smaller clusters are younger and at lower risk of mortality. Patients receiving only Impella® had a significantly higher rate of mortality than did patients receiving only IABP, and in three of the clusters patients receiving only IABP had a significantly higher rate of mortality than did patients receiving no MCS device. This method is suitable for jointly performing observational comparative effectiveness and risk modeling.

10. LATENT SPACE ANALYSIS OF SEMI-SUPERVISED LEARNING WITH A DEEP MIXTURE OF EXPERTS

Developing on from the prior chapter, we now turn to a more refined analysis of heterogeneity as exposed by the deep mixture of experts model. Here, we look more closely at the gating provided by the α-network and analyze clustering assignments not as mutually exclusive phenotypes, but rather as complementary archetypes. We examine the latent space to better understand the certainty with which the model makes predictions and assigns groupings.

10.1 Introduction

The practice of medicine consists of diagnosis, prognosis, and treatment [224]. Diagnosis categorizes illness into understood and recognized diseases. Once an underlying disease has been diagnosed, patient prognosis and appropriate treatment can be decided. However, diagnosis is only as good as the understanding of the disease. Heterogeneity inherent in disease processes can limit this understanding [225]. While two patients may appear to have the same disease, one might benefit from an intervention while the other might be harmed by that same intervention. To improve patient care, it is essential to understand this underlying heterogeneity in diseases.

Phenotyping and archetyping represent two similar methods to understand a heterogeneous population and to group it into smaller, more homogeneous subpopulations. Phenotyping involves the identification of discrete clusters within the population and assigning each population member into one of those clusters. Each member belongs to one phenotype and exactly one phenotype. As used here, phenotyping is the clinical application of hard clustering. Archetyping involves the identification of example members throughout the population that are representative of particular set of characteristics. Each member of the population can be described as some combination of the archetypes. A particular member

might more strongly align with one archetype over another, but can share varying degrees of similarities with all archetypes. As used here, archetyping is the clinical application of soft clustering.

By identifying archetypes or phenotypes within a heterogeneous population, subpopulations of increasing homogeneity can be found. Within these subpopulations, personalized prognosis and treatment decisions can be made. By recognizing that a patient belongs to a particular group, an improved understanding of likely outcomes for that patient may be reached. With this improved understanding, personalized decisions can be made, leading to improved outcomes.

While the method described previously (Chapter 9) was able to find phenotypes within a population, that method can be improved. No subject is inherently classified as belonging to a single cluster- but the classifier assigns subjects to whichever phenotype is the most probably. This decision could be made even if a subject's cluster likelihoods were nearly evenly split. How appropriate is it, then, to group that subject with another subject who is purely in that cluster? This inspection reveals the weakness of phenotyping and suggests a need to proceed to archetyping.

Further developing on the earlier work, this work develops an approach for finding more nuanced archetypes within the deep mixture of experts (MoE) model. We inspect the latent space of the α-network and discover that the certainties with which it assigns β-networks can be harnessed for both a more confident prediction of clinical outcome and a deeper understanding of patient archetypes.

10.2 Related Work

10.2.1 Clinical Phenotyping

Clinical identification of a disease is the first step in treating that disease. However, diseases vary and their heterogeneity can greatly complicate the process of diagnosing and treating them [226]. Successfully finding and identifying that heterogeneity can allow for

drastically improved understanding and targeted treatment [225]. Unsupervised and semi-supervised clinical phenotyping approaches are valuable techniques given the vast amount of electronic health data produced and the sparsity of known phenotype labels. Classical unsupervised techniques have been widely studied for finding clinical phenotypes [220].

More recently, deep learning techniques have expanded the scope of data-driven strategies in clinical phenotyping. Beaulieu-Jones et al. [227] used a semi-supervised technique combining a denoising autoencoder and random forest classifier to discover phenotypes of patients with amyotrophic lateral sclerosis. Several techniques have recently been used for finding phenotype clusters using risk profiles. Chapfuwa et al. proposed a novel approach that involved clustering patients in a latent space and differentiated clusters by different time-to-event risks. Their clustering was performed with a joint learning of cluster assignments and time-to-event predictions. Nagpal et al. [229] used a generative mixture model for inferring treatment effect differences in patients treated with opioids and whether or not patients would still be using opioids one year after initial treatment.

10.2.2 Deep Mixture of Experts

The concept of using a deep MoE for supervised learning was first proposed in the early 1990s [230, 231]. In their earliest formulation, these MoEs were each simple feed-forward networks and the gating function a softmax function over those outputs. More recently, the concept has attracted a great deal of attention in applications involving large neural networks for a variety of tasks such as language modeling [218] activity recognition [232, 219], and image recognition [233].

Shazeer et al. [218] used an extremely wide MoE between stacked long short-term memory (LSTM) layers for language recognition and described how sparse gating of their model allowed for decreased overall computational cost. They described several challenges faced in the training of MoE-based models, namely the problem of batch sizes effectively shrinking as gating functions are learned and the challenge of balancing expert contributions. The shrinking batch problem can be addressed by keeping number of experts small, by increasing

batch size in proportion to number of experts, or by taking advantage of data structure via convolutionality. In this work, we keep the number of experts small with the primary goal to improve clinical interpretability, but with the secondary goal of addressing this problem. Shazeer et al. balanced expert contributions through the use of a loss function penalizing unequal expert importances.

10.3 Methods

Building off of the model described in Chapter 9, the joint objective remains in the semi-supervised realm. As a supervised learning problem, the goal is to predict clinical outcomes in a patient population. As an unsupervised learning problem, the goal has advanced to discovering and explaining the interaction of archetypes within the patient population. By coupling these objectives, we formulate a semi-supervised problem where a clustering schema can be learned as part of the supervised problem, but separated from the supervised case for the classification of new patients prior to treatment. In this work information is learned in three tiers: initial data, data following treatment, and outcomes. The initial data contain information that could be useful for identifying how to treat a patient. Once treatment has begun, a second tier of additional information is available, but as this information was influenced by treatment choices, it should not be used for clustering. Finally, after treatment, outcomes occur. A key assumption in this work is that tiers of information are causally blocked; features in one layer may have causal effect on later features, but the temporal setup should not allow later features to have causal effect on earlier features.

10.3.1 Data

The dataset used in this work is the same as that previously described in Section 9.4. The data was separated at random into 80% for model development and training, and 20% for validation. The model was trained using a 5-fold internal cross validation set of the training data, with each fold evaluated independently and used for establishing performance confidence intervals.

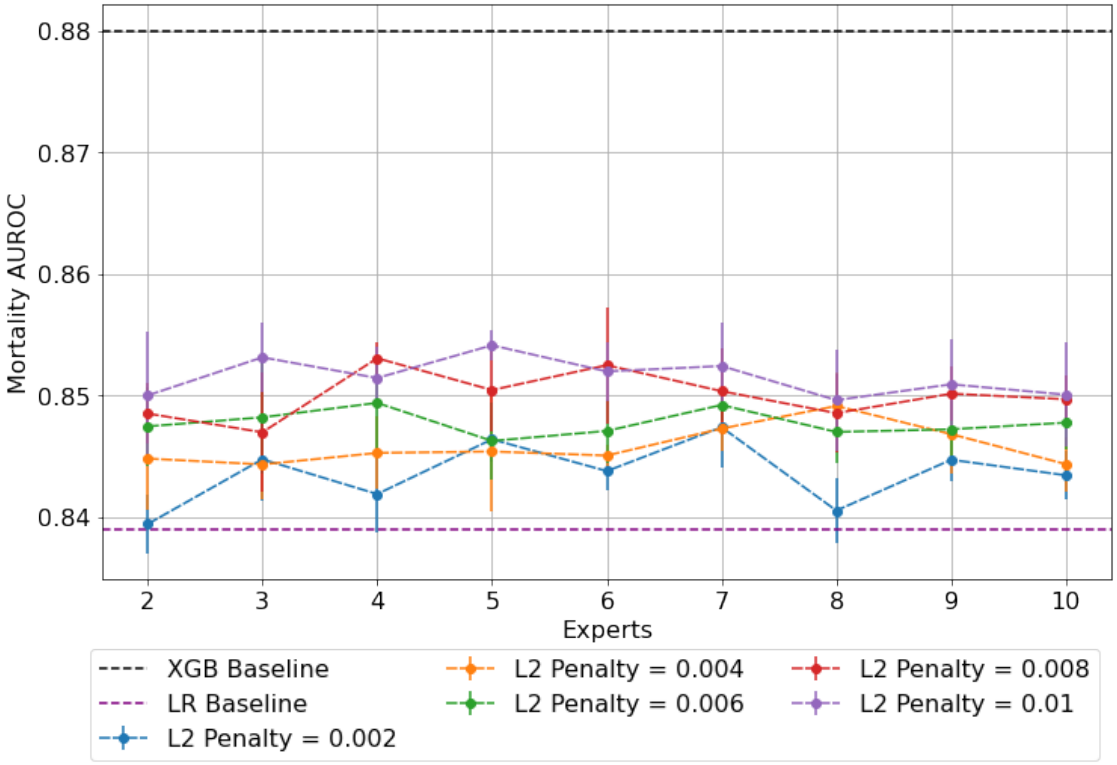


Figure 10.1: AUROC for mortality prediction given n experts and various L2 penalties. Confidence bars express 95% CI. Note truncated axis.

10.3.2 Model

The main MoE model remains unchanged from that presented in Chapter 9. The input to the α-network (first tier of information) is passed through a number of fully connected layers. The final fully connected layer is passed on to each β-network, and is also fed into a gating layer with size equal to number of experts. Each β-network concatenates its input (second tier of information, treatment information and later) to the pass-though output of the α-network, and then this is fed through several fully connected layers. The final layer of each expert outputs an outcome prediction. Each of these predictions are weighted by the α-network's gating layer and summed for the final model prediction.

A chief difficulty in building and training this model lies in the appropriate choice of regularization for the α-network probability outputs. In underregularized setups, this model

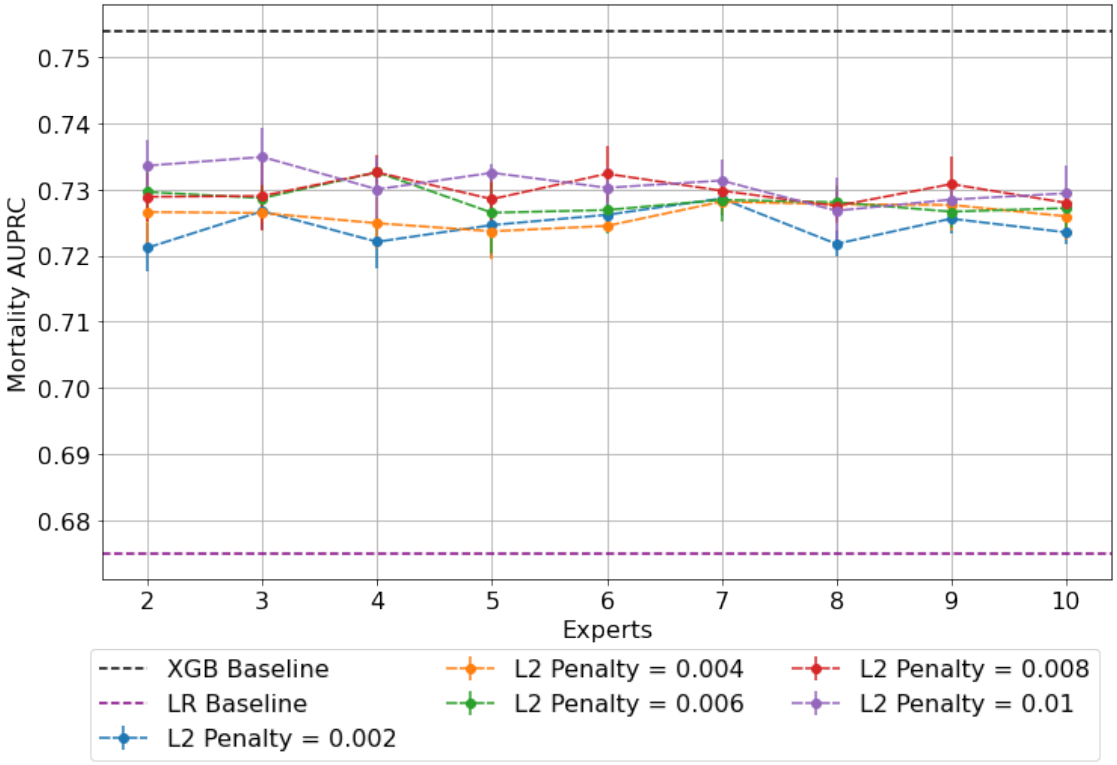


Figure 10.2: AUPRC for mortality prediction given n experts and various L2 penalties. Confidence bars express 95% CI. Note truncated axis. Overall mortality (25.7%) is the lower bound of a naive model.

collapses to use only a single expert, while in overregularized setups the model uses all experts evenly, but does not consistently group any subjects together. L2 regularization penalties were applied to the output of the α-network. These penalties encourage an even utilization of all experts, allowing the overall model to selectively use specific experts for classification. This work elaborates and refines over the regularization shown in Figure 9.2. Given that the 0.01 term there was the only performant model, we searched over regularizations closer to that value and used that search to refine our final results.

The latent space produced by the gating layer is a key point of interest. The softmax output of that layer scales all outputs collectively to sum to 1. This results in many outputs being either very close to 0 or very close to 1, but some of subjects have less extreme weightings assigned to their experts. We observed previously that even though optimal

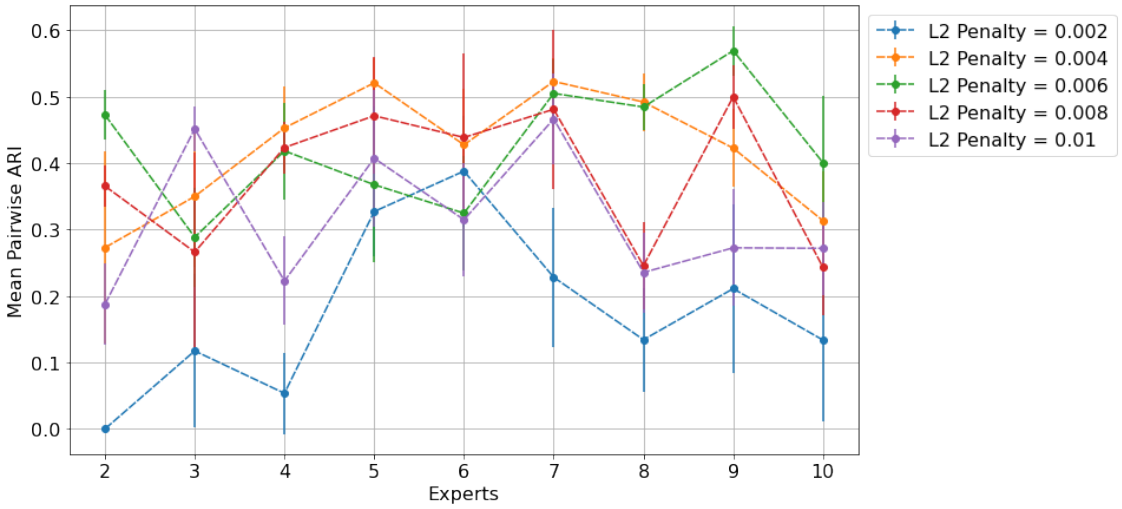


Figure 10.3: Mean pairwise ARI given n experts and various L2 penalties in the clinical dataset. Confidence bars express 95% CI.

Table 10.1: Weight contribution of each cluster. As only three clusters can be visualized at once, this table aids in assessing impact of truncation.

Cluster	5 Experts, L2 Reg = 0.004	3 Experts, L2 Reg = 0.01
A	77.0%	62.9%
B	12.2%	20.0%
C	8.5%	17.2%
D	2.1%	-
E	0.09%	-

performance was achieved with 3-5 experts (Figure 9.2), some number of experts contain very few assignments (Table 9.2). Therefore, for ease of visualization, we display the latent space of the α-network's gating layer using a ternary plot and we truncate beyond three experts. We align the expert assignments so that the expert receiving the greatest overall weight is always shown in the bottom right of these plots (labeled "A"), the expert receiving the second most weight is shown to the top (labeled "B"), and the expert receiving the third most weight is shown to the bottom left (labeled "C"). Points in these plots are colored based on their predicted label using a simple 50% prediction threshold (True/False

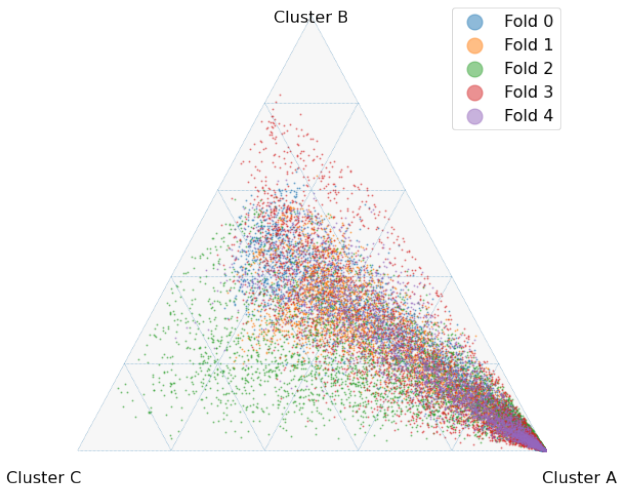


Figure 10.4: Ternary plot of α-network output with 5 Experts and L2=0.004.

Positive/Negative). To assess the model performance and uncertainty as a function of the latent space, heatmaps were generated with local AUROCs.

10.4 Results

The mean pairwise adjusted Rand index for each regularization and expert combination is shown in Figure 10.3. 5 Experts/L2=0.004 and 3 Experts/L2=0.01 were selected for further analysis. 5 Experts/L2=0.004 was selected for having a high ARI with few experts, while 3 Experts/L2=0.01 was selected for having the highest AUROC performance with few experts (Figure 10.1), as well as the highest AUPRC performance overall (Figure 10.2).

10.4.1 5 Experts, L2=0.004

A ternary plot of all folds overlaid on each other is shown in Figure 10.4. Two experts are truncated. However, their combined impact to all predictions is minimal: one expert supplies a total of 2.1% weight of all experts, while the other supplies less than 0.1% of that weight (Table 10.1). Cluster A can be seen to be the most weighty cluster, contributing 77% of all weight.

A representative fold from this model is shown in Figure 10.5. In this figure, true positives

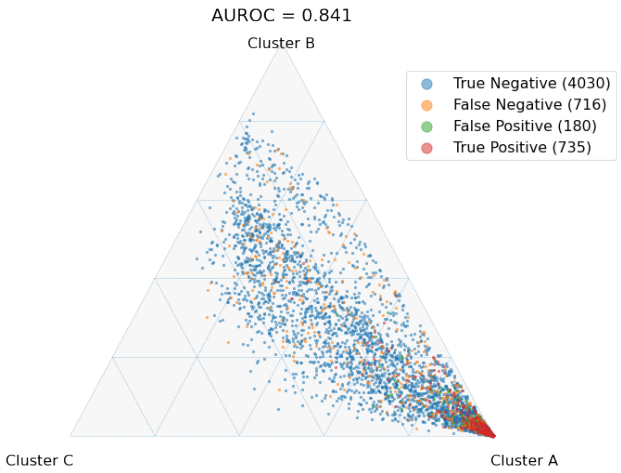


Figure 10.5: Selected single-fold ternary plot with 5 Experts and L2=0.004. Colors indicate prediction and correctness assuming a simple 50% threshold.

are found more often near Cluster A, while subjects more distributed throughout the other clusters exhibit lower overall risk. Local prediction performance is shown in Figures 10.6 and 10.7. Prediction quality can be seen to improve among patients who are more balanced between clusters.

10.4.2 3 Experts, L2=0.01

A ternary plot of all folds overlaid on each other is shown in Figure 10.8. All experts are shown. Cluster A is again the most heavily weighted cluster with 63% of assignments, with Clusters B and C near equal (Table 10.1).

A representative fold from this model is shown in Figure 10.9. In this figure, true positives are one again found more often near Cluster A, while subjects more distributed throughout the other clusters exhibit lower overall risk. Local prediction performance is shown in Figures 10.10 and 10.11. Prediction quality can be seen to improve among patients who are more balanced between clusters.

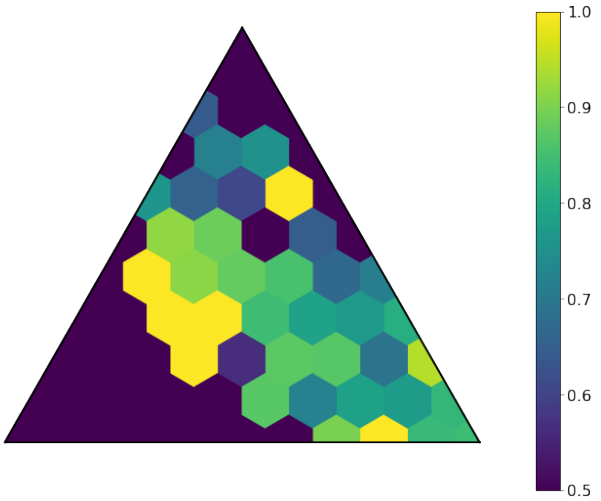


Figure 10.6: Local AUROCs of model with 5 Experts and L2=0.004. Cells with insufficient subjects for scoring are set to 0.5.

10.5 Discussion

In this work, we explore the latent space of a deep MoE classifier to understand how it can aid in assigning archetypes. Rather than assigning to a single most representative cluster, this approach allows for a nuanced balancing and understanding of soft clustering. The outcomes-driven nature of the joint model training does strongly bias patients with the outcome of interest into one group. This allows for the model to separate patients by severity, with patients further from the strongest cluster exhibiting lower clinical risk. At the same time, the overall model performance does well further from this strong cluster membership. This approach shows that the deep MoE is successful and appropriate for separating heterogeneity in a clinical population and for separating those patients by disease severity.

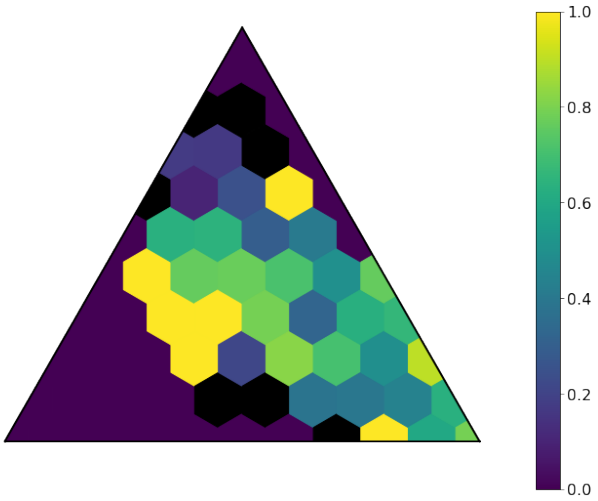


Figure 10.7: Local AUPRCs of model with 5 Experts and L2=0.004. Cells with insufficient subjects for scoring are set to 0.

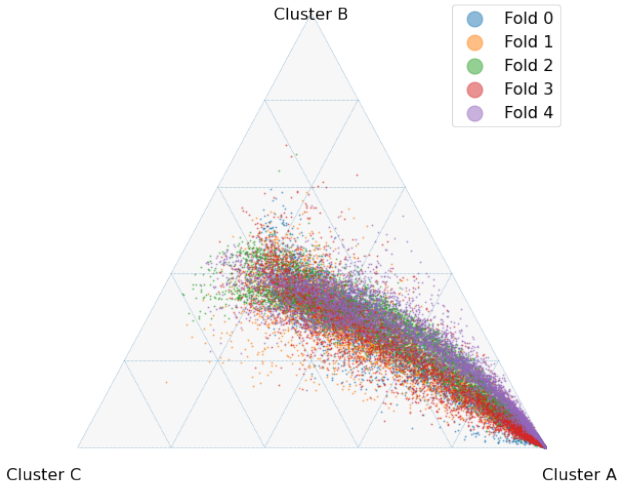


Figure 10.8: Ternary plot of α-network output with 3 Experts and L2=0.01.

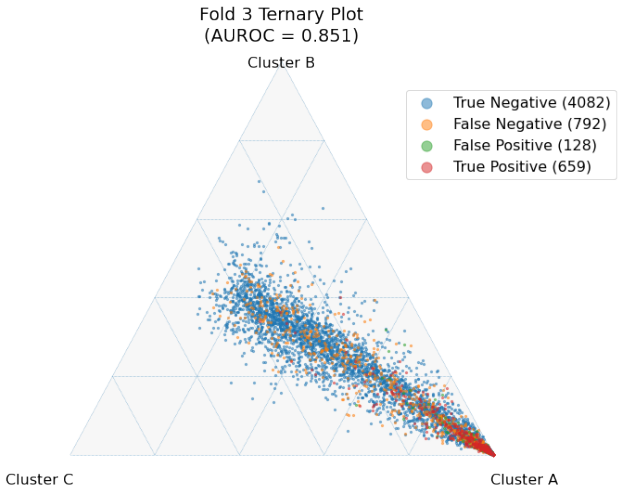


Figure 10.9: Selected single-fold ternary plot with 3 Experts and L2=0.01. Colors indicate prediction and correctness assuming a simple 50% threshold.

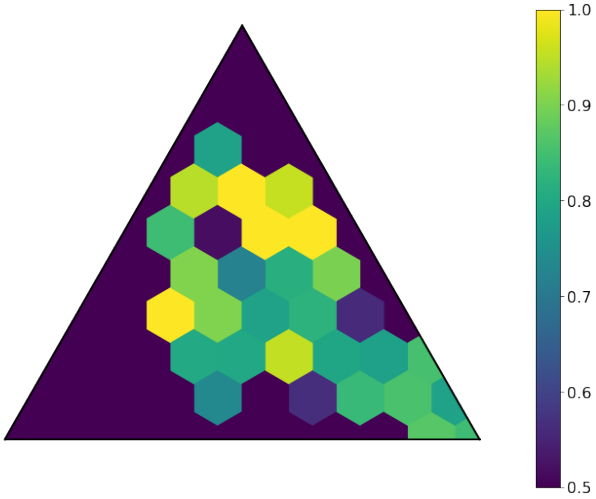


Figure 10.10: Local AUROCs of model with 3 Experts and L2=0.01. Cells with insufficient subjects for scoring are set to 0.5.

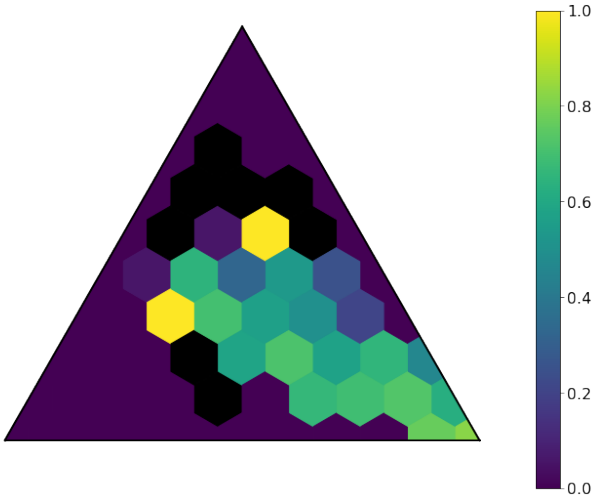


Figure 10.11: Local AUPRCs of model with 3 Experts and L2=0.01. Cells with insufficient subjects for scoring are set to 0.

11. CHALLENGES AND OPPORTUNITIES IN SENSING AND ANALYTICS FOR RISK FACTORS OF CARDIOVASCULAR DISORDERS*

We now look beyond the clinic to applying machine learning in remote settings. The development of remote sensors are outpacing the development of machine learning techniques to analyze and use the rich data that they produce. With the advent of the Internet of things, opportunities for remote sensing for healthcare is boundless. This chapter details ways in which these signals can be harnessed for cardiovascular monitoring.

11.1 Introduction

Cardiovascular diseases are the worldwide leading cause of death [234]. In 2016, cardiovascular diseases accounted for nearly 1 in 3 deaths in the United States. While the range of cardiovascular diseases and treatments can be broad, the Framingham Heart Study teaches us that a number of the risk factors that lead to primary adverse events or secondary recurrent events are often the same or quite similar [235, 236, 237]. Real-time monitoring of these risk factors (i.e. the signs and symptoms associated with cardiovascular disorders) allows for care providers to track patient progress and to rapidly respond to any changes in patient condition. In the hospital, monitoring patients is part of routine clinical practice. Providers are able to monitor cardiac status and basic vitals from anywhere in the hospital at any time. Slight deterioration in health can be observed and interventions put into place before patients suffer worsening harm. However, length of stay in these acute care settings is often quite short [9, 10], representing only a small portion of a patient's life despite the prolonged impact that the decision making in these settings have. Such monitoring is currently deficient in remote settings, where the ability to diagnose new conditions or monitor treatment effectiveness based upon measured changes in vitals and cardiac status that are known to be

*This chapter is reprinted with permission from "A Survey of Challenges and Opportunities in Sensing and Analytics for Risk Factors of Cardiovascular Disorders" by Hurley, N. C., Spatz, E. S., Krumholz, H. M., Jafari, R., & Mortazavi, B. J., 2020. ACM Health. Copyright 2020 by Nathan C. Hurley. et al.

risk factors for primary adverse events or secondary recurrent events is important to prevent future admissions to acute care settings. Monitoring physiologic parameters and symptoms outside of the hospital in ambulatory and/or remote settings can enable better detection and response systems before a person becomes acutely ill and requires hospitalization or after hospitalization to prevent early readmission to the hospital; however, many of the devices today are targeted to healthy people. With the prevalence and ubiquitous nature of remote and wearable sensors, opportunities exist to broaden the applications of sensing and for adapting analytic techniques to enhance diagnosis, monitoring, and treatment of risk factors for primary and secondary prevention of cardiovascular disease. In particular, the ability to capture these measurements is only the first step. Indeed, end-to-end smart health systems are needed that couple the hardware development with advanced analytic techniques to provide both patient and clinical provider necessary confidence in data and risk prediction based upon the measured risk factors.

A challenge in monitoring patients with or at risk for cardiovascular disorders is designing the technology and algorithms to support a variety of conditions and signs/symptoms. While the treatment of cardiovascular disorders such as heart failure [236], coronary artery disease [238], and stroke [237], may differ (the latter, for example, moving from monitoring a potential cardiovascular disorder to neurological treatment), they share a common set of cardiovascular risk factors [239, 240, 241]. The selection of these three disorders highlights their global disease burden, but certainly the Framingham Heart Study teaches that the important risk factors that should be monitored are not limited to tracking only these disorders.

Patients at risk for cardiovascular disorders (or recurrent events due to diagnosed cardiovascular disorders) present a number of challenges for remote monitoring and diagnosis because of complexities within the diseases or trajectory leading to the initial diagnosis. Many of these diseases involve seemingly trivial symptoms that may suddenly change from a minor inconvenience to a debilitating lack of function. A patient with a given disease

may feel well for multiple years, and then suddenly decompensate and require emergent care. Ideally, remote monitoring along with advanced analytics on the captured ambulatory data should be able to track the slow, daily progression of a disease states and alert the patient and healthcare providers to worsening disease before decompensation and patient suffering. However, preliminary studies in remote monitoring have failed at preventing adverse events, such as in preventing repeated hospital admissions in patients diagnosed with HF. For example, the Telemonitoring in Patients with Heart Failure trial (Tele-HF) used patient self-reports of daily changes in symptoms, weight, and a variety of other factors (e.g., medication changes, depression scores, etc.) to identify worsening symptoms in an effort to intervene prior to another acute event, but did not find a statistically significant difference between control and intervention arms [242]. However, an analysis of participant subgroups did find that patient self-reported data could improve prediction of readmission likelihood, showing potential for more advanced analytic techniques to better identify participant risk and to improve estimates in this space [243]. The BEAT-HF trial was designed as a further exploration in automating the capture of the relevant biometric signals, including heart rate, blood pressure, and weight, using remote sensors rather than participant self-report of such data. This study, however, was similarly unable to find a statistically significant difference in control and intervention arms [244], suggesting that further exploration of additional biomedical signals are needed and that the advancements in improved remote and ambulatory monitoring of these key risk factors, alone, is not sufficient to address clinical need. Instead, improved remote and ambulatory sensing likely needs to be coupled with advancements analytic techniques needed to process and interpret data generated by these sensors.

Remote sensing technologies have increased in prevalence and have made personalized health data collection feasible. In human activity recognition (HAR), wearable sensors and inertial measurement units embedded within smartphones and smartwatches have enabled the tracking of detailed motions [245, 246]. Coupled with nearable sensors that capture mo-

tion via video, these sensing systems allow for the tracking of motions of healthy participants [247] to tracking of disease state with custom-built sensors, such as smartshoes [248]. The data provided by these wearable and remote sensors has more recently enabled advanced machine learning techniques to identify more complex patterns of motions, better understanding personalized behavior [249, 250]. Eventually, these techniques have emerged to personalize models of activity recognition to individual users, and this personalized modeling provides the most robust interpretation of activities of daily living per user [251], enabling feedback and the measurement of clinical outcomes [252]. This progression from the development of new sensing modalities to the analytic techniques that detect patterns within the data and finally to personalization in tracking and disease progression modeling is an end-to-end pathway that is required for advanced clinical disorders monitoring for smart health technologies.

The development of new sensors to measure risk factors (e.g. symptoms) of cardiovascular disorders would ideally enable a similar progression for tracking of cardiovascular outcomes. These new sensors would be able to identify conditions that may not be apparent to patients or providers, such as different sounds from the heart, slowly decreasing patterns of activity, or combination of vitals that may appear normal in isolation but may be indicative of risk given a combination of values and certain patient contexts. By identifying dangerous signs before symptoms manifest, earlier interventions can lead to improved health outcomes. A variety of technologies and machine learning techniques to this purpose exist in condition-specific settings [253, 254] to varied success [125, 255, 256]. Understanding the pathologies of the disorders is important in understanding the clinical needs and opportunities that exist in developing new wearable and remote sensors for diagnosis and treatment of a variety of cardiovascular conditions and using advanced analytic techniques that are enabled from the collection of new, comprehensive patient ambulatory risk factor data.

In this survey, we break down the needs and opportunities in monitoring risk factors for the prevention of primary or secondary recurrent adverse events of select cardiovascular dis-

orders into key technological areas that couple remote sensing with analytic developments: 1) we discuss different sensing modalities that have been or that could be applied to tracking cardiac health in remote settings; 2) we consider the opportunities that advanced analytic strategies present with the acquisition of remote sensing data for continuous risk modeling; and 3) we discuss the needs and opportunities for advancements in clinical models using machine learning techniques, including advancements in longitudinal monitoring and interpretability made possible through newer deep learning techniques. As cardiac pathologies manifest, they can also be indirectly observed through physical changes in the body, potentially measured by sensors on or around the body. These changes can be utilized to track patient health, to plan interventions to maximize patient wellness, and to decrease the overall impacts of the disease. One of the oldest technologies used for assessing cardiac health is the stethoscope. In the digital era, the electronic stethoscope is a varied group of technologies that incorporate a microphone in order to automate acoustic diagnose and facilitate remote monitoring [257]. Other technologies, such as photoplethysmography and sphygmomanometry, allow for remote measurement of the characteristics of a heart beat including heart rate and blood pressure [258]. Doppler radar can detect vital signs such as respiratory rate and heart rate [259]. Electrical techniques such as electrocardiography (ECG) or other conduction studies such as Bio-impedance can give insights into the internal physiology of the heart [260].

Sensing systems provide for opportunities to proactively detect and alert patients and physicians to worsening health states. However, to allow for timely and effective interventions as well as to rapidly evaluate the impact of those interventions, development of advanced signal processing and machine learning techniques need to keep pace with the development of raw sensor modalities. This paper presents a survey of state-of-the-art sensing technologies and analytics with respect to monitoring key risk factors for cardiovascular disorders, in order to highlight successes and provide areas for additional growth. Two key ways in which analytics associated with sensing systems can provide support are to develop personalized

models for longitudinal tracking of the risk factor measurements and to develop clinical risk prediction models that monitor disease state trajectories for identifying the onset of a new disease and to track the progression of preexisting disease to avoid recurrent adverse events. Tracking the progression of existing disease is the easier task: once an underlying disease state is known, appropriate monitoring can be put into place and utilized to follow the progression of the disease. Monitoring for the start of new disease is more difficult, as the focus is more general. In either case, sensing and clinical characteristics must be combined for decision support with the aid of machine learning approaches. In this paper, we survey the current state of the art in patient monitoring and analytics for patient risk and care, highlighting needs and opportunities for advancements in the field of smart health with respect to monitoring signs, symptoms, and treatments in patients at risk for diagnosis and adverse events with respect to cardiovascular disorders. We highlight the need to view this technical challenge as an end-to-end smart health solution, requiring both advancements in sensing systems and advancements in analytic techniques to properly analyze and interpret data generated from these systems. The workflow described in this paper towards developing new tools for remote clinical decision support is shown in Figure 11.1.

The rest of this work is organized as follows. Section 11.2 introduces the cardiovascular disorders, their common risk factors, and needs in remote and ambulatory monitoring for these conditions. This section provides a focus for the clinical tasks and describes how the particular case studies generalize to common risk factors and outcomes. Then the paper provides a description of the current state of technologies, remaining needs (technical gaps), and opportunities for technological advancements in end-to-end smart health systems designed for addressing the clinical needs by discussing sensing (Section 11.3), analytics on the sensing systems (Section 11.4), and clinical analytic models on the data generated for patient and provider use (Section 11.5). Finally, Section 11.6 provides a discussion and conclusion.

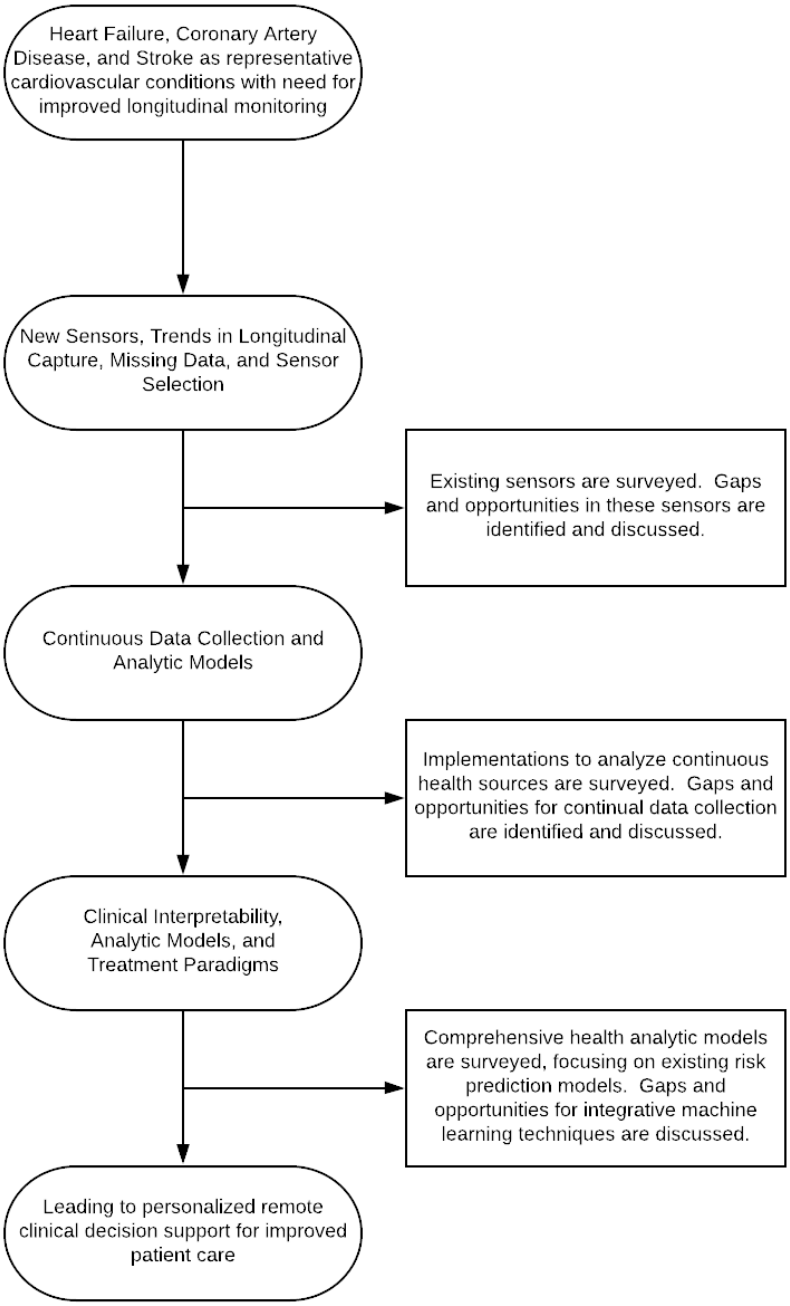


Figure 11.1: Overview of a workflow to developing personalized, remote clinical decision support tools for patients to monitor risk factors of cardiovascular disorders. Needs are shown in three categories: needs in sensor development and data handling, needs in continuous data collection and analysis, and needs in developing comprehensive and personalized analytical models. Addressing these three categories will allow for improved personalized remote clinical decision support for patients and the design of end-to-end smart health systems for clinical modeling.

11.2 Case Studies and Needs

This work considers risk factors associated with primary adverse events and secondary recurrent adverse events associated with the diagnosis and treatment of cardiovascular disorders. A number of the chronic conditions listed may have disparate treatment patterns, however, the underlying risk factors that lead to the initial events have significant overlap. To highlight this, we consider several conditions, namely, heart failure (HF), coronary artery disease (CAD) and acute myocardial infarction (AMI), and stroke. In particular, we include stroke as a condition given the primary risk factors are cardiovascular in nature, even if treatment afterwards may tend to be covered by neurologists. In this section we provide a brief overview of the conditions, their measurable factors, and provide definitions and abbreviations used throughout the manuscript. Table 11.1 provides a list of the key terms and definitions for this section.

HF is typically a chronic condition where the heart is unable to drive blood forward through the body sufficiently or can only do so under damagingly high pressures. HF is a debilitating disease that causes significant global disease burden. In 2016, HF was the most rapidly growing cardiovascular condition in the world [261]. CAD occurs when blood flow through the coronary arteries, the small arteries that provide blood to the heart, becomes impeded. This occurs both gradually as plaque builds up within the coronary arteries and suddenly when a plaque ruptures and clots. The former causes chest pain and exercise intolerance, while the latter, commonly known as an MI, can cause severe pain, loss of consciousness, and death. Each year around 800,000 Americans suffer an AMI, and rapid care following an AMI is a chief predictor for minimizing long term morbidity and mortality [262, 263, 264]. Stroke is any disease impacting the blood vessels to the brain. In particular, acute stroke is a condition that occurs when either a blood vessel in the brain ruptures, or when one of those blood vessels becomes blocked. Stroke manifests with the sudden onset of neurological deficits, some of which may be irreversible. Stroke is the fifth leading cause of death in the United States and is a leading cause of long-term disability [234].

This work considers three primary cardiovascular disorders for the review of gaps and opportunities, though by no means encompasses the entirety of technologies available for monitoring and treating these conditions nor the entirety of conditions to which these technologies could be applied. Instead, these conditions serve as meaningful examples in which technical solutions that monitor and model the known clinical risk factors would be clinically impactful, and demonstrate the similarity in key risk factors despite the potentially divergent care required after the diagnosis of each condition.

11.2.1 Clinical Conditions

HF occurs when one or both halves of the heart are unable to drive blood flow forward at the rate required by the body or can only do so under high pressures. This discussion of pathology will focus primarily on left-sided HF rather than right-sided HF, but the two are often closely associated and technologies for monitoring the two will have a large amount of overlap. The two will also often coexist. HF can result from ineffective heart contractions, from high pressure limiting the effect of heart contractions, or from difficulty in filling the heart. The first two causes lead to HF with reduced ejection fraction (HFrEF), and the last leads to HF with preserved ejection fraction (HFpEF). Ineffective heart contractions can result from muscle damage caused by CAD, by chronic volume overload as seen in mitral regurgitation (MR) or aortic regurgitation (AR), or by a family of cardiac muscle disorders known as cardiomyopathies. High pressure can lead to HF either from aortic stenosis (AS) or from uncontrolled hypertension. In either case, the pressure that the heart works against is so high that the pumping becomes ineffective. Difficulty in filling the heart can be caused by ventricular hypertrophy, cardiomyopathy, fibrosis, disease around the outside of the heart (the pericardium), or by CAD.

Coronary artery disease (CAD) is a family of diseases where blood flow through the small arteries of the heart, the coronary arteries, is restricted. This restriction can be caused by deposits of fatty plaques within the arteries, or by clotting caused by the rupture of one of these plaques. Depending on the extent of the blood flow restriction and the current

oxygen demands of the heart, CAD may cause different symptoms. CAD is represented by a spectrum of conditions that are defined by specific clinical and physiological signs.

Stroke occurs when blood supply in and around the brain is acutely disrupted, and results in acute neurologic defects. Ischemic stroke is a type of stroke where a blockage in cerebral arteries rapidly blocks off blood flow, leading to cell death. Hemorrhagic stroke is a type of stroke where a blood vessel in the brain ruptures, rapidly raising pressure inside the skull and causing cell death. Transient ischemic attacks (TIAs) are similar in cause and presentation to strokes but resolve spontaneously. They are often an indicator of underlying disease and put the patient at increased risk for future TIA or stroke. The neurological pathology goes beyond the scope of this work, but there are several notable cardiovascular impairments that may cause a stroke.

A common key risk factor to all the conditions above is hypertension (HTN). HTN is a condition where a patient's blood pressure is persistently elevated and is often a condition that serves as a modifiable precursor to each of the three cardiovascular disorders discussed [265]. HTN is divided by cause into two categories: primary (or essential) HTN, which has no particular medical cause, and secondary HTN, which is caused by some other medical condition. Primary HTN accounts for roughly 90% of all HTN, while secondary HTN accounts for the remaining 10%. Causes of secondary HTN include renal disease and endocrine diseases that disrupt the body's natural control of blood pressure [266]. Essential HTN is a diagnosis of exclusion and requires ruling out the possibility of any secondary causes. Risk factors for essential HTN include both hereditary and environmental factors [267]. There is a strong association between HTN, obesity, and insulin resistance. HTN is associated with poor diet, excessive alcohol intake, and age. By measuring blood pressure and identifying patients with HTN, we can consider HTN as both a disease state, and potential progression to the other conditions listed in this work, while also similarly considering it a measurable risk factor for those conditions. Because an HTN diagnosis is a modifiable risk factor prevalent in numerous cardiovascular disorders, we highlight it here specifically as a clinical condition

in its own right, but consider the measurement of blood pressure as a key sensing parameter for the rest of this work for both diagnosing HTN and for using blood pressure directly as a risk factor for the other cardiovascular conditions.

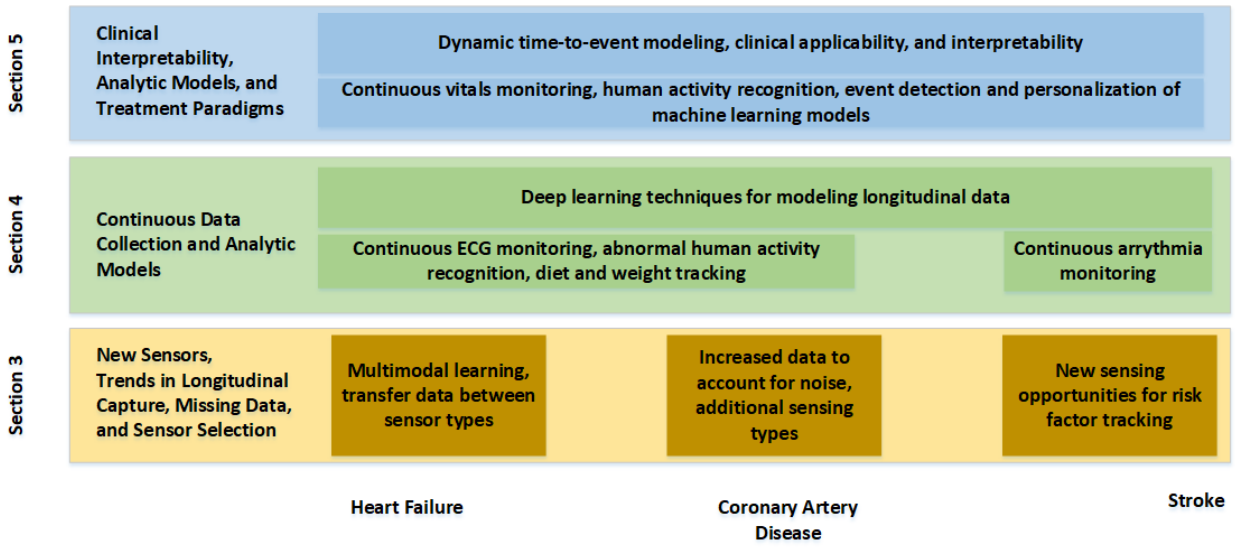


Figure 11.2: Progress from individual building blocks provided by new sensing opportunities to joint, multi-modal analytics, to combined end-to-end modeling for clinical use (y axis) and how they generally relate to each of the three conditions (x axis).

11.2.2 Needs for Monitoring Signs and Symptoms for Cardiovascular Disorders

Figure 11.2 illustrates the three primary needs this survey will discuss: 1) need for sensing technologies that track longitudinal trends of the measures important in identifying risk of cardiovascular disorder despite infrequent, noisy, or missing data measurements; 2) need for new analytic techniques designed in a longitudinal, continual fashion to aid in the development of new risk prediction techniques and in tracking disease progression; and 3) need for personalized and interpretable machine learning techniques, allowing for advancements in clinical decision making. A number of varied signs and symptoms exist for HF, CAD, and stroke. The remainder of this section briefly introduces some common signs and symptoms. Here, a symptom is a change caused by disease that is noticed by and likely an irritant to

the patient, while a sign is a change that the patient may not notice or that may not be concerning to the patient.

In HF, the symptoms result both from insufficient blood flow and from excess fluid buildup. The three main symptoms that are associated with diagnosis of HF and quantification of its severity are dyspnea (shortness of breath) on exertion; sudden, choking dyspnea at night; and difficulty breathing while lying down. In left-sided HF pulmonary vein pressure increases, causing buildup of fluid in the lungs (pulmonary edema) that worsens while lying down. In right-sided HF systemic venous congestion results in fluid buildup in the periphery (peripheral edema) that worsens while upright, resulting in noticeable swelling in the wrists and ankles. HF is difficult to precisely define as it is a clinical syndrome resulting from many different heart conditions, and many variants exist. Therefore, attempts to understand HF and to monitor its progression must focus on identifying the symptoms and identifying cardiac dysfunctions. Symptoms that can be measured include peripheral edema (swelling of ankles, rapid weight gain), decreased activity, and changes in respiratory patterns when lying down versus remaining upright. Changes in blood flow to the kidneys result in decreased urine production during the day, and increased urine production at night. Patients with HF will therefore often get up frequently in the night. These patients will also likely change posture in the night, with patients with advanced HF needing to sleep upright. One of the most used classification schemes for HF is the New York Heart Association (NYHA) Functional Classification [268]. In this classification scheme, classes are separated based on the physical activity that the patient is able to achieve and the discomfort that results from physical activity. Class I is when no symptoms are present, and in Class IV the patient is unable to perform any physical activity without discomfort and symptoms of heart failure are never alleviated. As can be seen, a variety of sensing modalities could be employed to track signs and symptoms of HF, from measurements of peripheral blood flow, respiration rate, exercise capacity, and posture while sleeping. This illustrates the need for new sensors that can measure each of these various symptoms. However, not every sensor

may be worn at all times, due to excessive burden on the user. Therefore, there is a need for new sensing modalities that can track different patterns and trends in captured data, as well as transfer learning techniques that can be adapted to estimate values of sensors that may be malfunctioning or not worn.

If the right set of sensors are selected and are designed to be worn longitudinally, new patterns and trends in signs and symptoms might be detected. In CAD, for example, restrictions in blood flow of the coronary artery may result in a condition called stable angina (SA). The rate at which the restrictions in blood flow occur, however, might change as the disease progresses. At some point, the restriction responsible for SA may rapidly increase, producing a situation where the patient is in emergent need of medical care. The most common way for this progression to occur is for a fatty plaque to rupture, leading to the formation of a clot that blocks blood flow. The first disease after this point is unstable angina (UA). As the restriction increases to a partial occlusion, the patient will experience chest pain that worsens without activity or that is not relieved with rest. Both stable and unstable angina present similarly in a patient. Typically, the patient will have episodes of chest pain that last from 3-10 minutes, but potentially lasting up to 30 minutes. This pain may radiate to the jaw, neck, shoulder, or arm. The patient will likely feel short of breath and may also experience nausea. If the patient takes a medication called nitroglycerine, the pain should resolve within 1-3 minutes. In UA, damage is still reversible, but intervention is emergently necessary to ensure that the disease does not progress. If UA progresses, it will progress to a condition commonly known as a heart attack, or in medical terminology as a myocardial infarction (MI). There are two types of MI: non-ST-elevation MI (NSTEMI), and ST-elevation MI (STEMI). In NSTEMI, some muscle in the heart has begun to die, and therefore at least some of the damage caused is irreversible. In a STEMI, there is a complete blockage of blood flow at some point and a large amount of muscle in the heart has begun to die. NSTEMI and STEMI are distinguished by characteristic findings on ECG; in a STEMI, the ST segment will be elevated above the baseline in some leads, while this elevation is

absent in NSTEMI. The leads showing this change reflect the area of the heart impacted by the MI. This demonstrates the second need, longitudinal monitoring of continuous signals that can identify disease progression, and machine learning techniques that can account for the personal progression and varied rates of this progression.

In order to prevent conditions such as stroke, which are treated by neurologists after the primary adverse event, interventions are necessary in known cardiovascular risk factors, such as HTN, which can lead to stroke in multiple ways. Very high blood pressure raises the risk of hemorrhagic stroke, as blood vessels in the brain may not be able to support higher pressures. Additionally, chronic HTN is the main risk factor associated with ischemic stroke. The diagnosis of HTN requires repeated blood pressure measurements (sustained HTN), as measured by ambulatory blood pressure measurements. Various reasons for blood pressure elevation must be identified, including white coat HTN (when the blood pressure is elevated during a visit to a doctor but normal when measured in home settings), masked HTN (when blood pressure is regularly elevated but detected as normal during a visit to a doctor), and evaluation in changes of blood pressure when sleeping versus when awake (nocturnal nondipping HTN). HTN typically does not manifest with any symptoms, as the body is very good at masking the feeling of this pressure. Although high blood pressure has been colloquially associated with stress, headaches, or dizziness, these symptoms are typically not caused by chronic HTN. The primary sign (and part of the diagnostic criteria) of HTN is an elevated blood pressure. For diagnosis, at least two measurements on two different occasions of blood pressure above 120/80 mmHg are required. More recently, guidelines have suggested measuring blood pressure with an ambulatory blood pressure monitor over a 24-hour period, measuring blood pressure every 15 minutes during the day and every 30 minutes during sleep at night, and using the average values to have a better understanding of a patient's blood pressure [269]. This sustained elevation may result in stiffer arteries, reducing arterial compliance. Additionally, over time, this chronic elevation may result in left ventricular hypertrophy seen on ECG or in changes in the retina. Most patients with

HTN are largely asymptomatic, with the chief clinical sign being that of elevated blood pressure. When symptoms of HTN do manifest, they are largely caused by organ damage that results from chronically elevated blood pressures. Chronically elevated blood pressure can lead to heart damage, as the heart must work harder than normal to produce these elevated pressures. This can lead to HF as the heart gains mass and loses efficiency, or to CAD as the increased mass of the heart requires increased myocardial oxygen supply. Chronically elevated blood pressure can also lead to damage of the arteries. This can lead to atherosclerosis, where plaque buildups can compromise coronary arteries, leading to CAD or cerebral arteries, leading to stroke. Weakening of arterial walls can lead to kidney disease or to retinal disease. Advanced HTN can cause changes to the eye that can be observed visually by a physician. The definition of high blood pressure has undergone changes in recent years, with the SPRINT trial indicating that aggressive treatment of blood pressure to $<120/<80$ mmHg is associated with decreased mortality [270]. The potential measurement of blood pressure from new sensing modalities can enable analytic techniques to identify cases of HTN and evaluate the effectiveness of medication on reducing blood pressure, such as in the SPRINT trial. This illustrates the third need, where machine learning techniques, trained on continual data captured from new sensing modalities (the prior two needs), must provide actionable, interpretable estimations of signs, symptoms, and disease progress, in order to help guide treatment decision making and evaluate treatment effectiveness both prior to a diagnosis of a cardiovascular disorder and in the treatment and evaluation of recovery from an adverse cardiovascular event.

Table 11.2 highlights the available commercial devices that currently suited for tracking a number of the risk factors highlighted for the three cardiovascular disorders. Most devices use light-based sensing for tracking heart rate, pulse oxygenation, and a few have additional sensing capabilities. In the following sections, we explore the state-of-the-art in technology associated with each of the clinical needs, highlighting research advancements beyond the currently available commercial solutions. This survey reviews the technology available, the

gaps that remain in addressing the needs, and highlights opportunities for researchers within the smart health field to design solutions with impact to clinical decision-making problems.

11.3 New Sensors, Trends in Longitudinal Capture, Missing Data, and Sensor Selection

New sensing techniques that capture acute data as well as detecting changes in sensed data over time, are needed to measure the important signs and symptoms that are risk factors for HF, CAD, and stroke. Each condition has a set of similar risk factors as well as unique signs and symptoms that manifest through a variety of changes in the body. For HF, improper blood flow can result in fluid retention (edema) in the lungs or the periphery, as well as causing signs of heart remodeling. Heart remodeling can be evidenced by third and fourth heart sounds (S3 and S4), as well as by a laterally or inferiorly displaced point of maximal impulse (PMI) of the heart on physical exam; the place where the heartbeat can be felt most strongly will migrate down and to the left of the thorax. One way in which improper blood flow can be detected is that the extremities will be cooler than normal.

In CAD, stable and unstable angina will often result in physical pain felt by the patient in an episode that may last up to 30 minutes in the chest that may also radiate to the jaw neck and arm. The patient's heart rate and blood pressure will initially be elevated, although these can potentially decrease in NSTEMI and STEMI as the heart fails to operate optimally. The patient will breathe more quickly and will put more effort into breathing. Additionally, abnormal sounds may be heard with a stethoscope. It is possible for rales, an abnormal lung sound, to be heard at the posterior base of each lung. During chest pain, an ECG will show ST-segment depression, but this will change and progress to ST-segment elevation in STEMIs.

For stroke, this work focuses on the signs and symptoms that might lead to a stroke. Atrial fibrillation (AFib) is a relatively common arrhythmia that increases risk of stroke. AFib results when the atria of the heart beat ineffectively and randomly, causing turbulence within the atria. This turbulent flow allows for clots to form within the atria. If these

clots are dislodged, they may travel through the arteries and become lodged in the brain, causing an ischemic stroke. AFib is classically defined as an "irregularly irregular" beat- the beat is not a typical rhythm (irregular) and additionally has no pattern determining when beats occur (irregularly). This is most often seen as absent P waves on ECG with variably occurring QRS complexes over a noisy baseline. However, this pattern could be detected by many techniques that measure pulse. Chief risk factors that predispose patients to AFib are age, other heart disease, diabetes, and chronic lung disease. HTN can also lead to stroke in multiple ways. Very high blood pressure raises the risk of hemorrhagic stroke, as blood vessels in the brain may not be able to support higher pressures. Chronic HTN is the main risk factor associated with ischemic stroke.

These cardiac conditions present a range of sensing opportunities:

- Acoustic measurement: capture of heart sounds to identify specific classes as well as respiratory effort are important in understanding acute conditions and changes in heart function over time. This also includes respiratory distress when lying down, causing patients diagnosed with HF to need to sleep in a more upright position. *(See Section 11.3.1.1)*

- Electrical measurement: Remote ECG measurements can identify periods of atrial fibrillation and other arrhythmias or help identify progression of CAD during an acute event. *(See Section 11.3.1.2)*

- Heart Beat and Associated Characteristics: Understanding cardiac output, as well as measurement of blood pressure, is an important risk factor that needs periodic measurement. *(See Section 11.3.1.3 and 11.3.1.4)*

- Fluid retention/Weight change: HF often results in lung and peripheral edema that results in swelling and can be measured by cooler temperatures in the periphery and changes in weight. *(See Section 11.3.1.5)*

- Diet, exercise, and pain: In all cases, patient diet (for identifying glucose intolerance, obesity, etc.), patient self-reported pain, fatigue, and general physical activity may be surrogates for worsening conditions. Activity recognition can include posture detection to link with respiratory measurements, and can impact monitoring of glucose intolerance, which can lead to diabetes. *(See Section 11.3.1.6 and 11.3.1.7)*

11.3.1 Existing Technologies and Applications

11.3.1.1 Acoustic Sensing/Vitals

Vital sign monitoring has been explored through a variety of technologies. Each sensor type has been designed to address some of the sensing needs described in the previous section in an effort to replace or replicate tools available in acute care settings for remote environments. The stethoscope is one of the oldest such tools in medicine and is an implementation of acoustic sensing. By hearing and interpreting sounds from the patient, the physician can develop insights into the health of the patient and the functionality of the organs. Recently, digital stethoscopes have been utilized to better capture sounds. Digital stethoscopes provide benefit in allowing soft sounds to be more easily heard, but also allow for recording of sounds for later manual or computational analysis. As physicians have grown more reliant on advanced imaging techniques such as ultrasound, physical exam skill, including skill at auscultation, has decreased [288].

Developing a digital stethoscope involves multiple components requiring heart sound capture, segmentation of the audio signal, and understanding of the cardiac cycle, best paired with an external signal such as ECG or pulse to determine the reference interval as described by Leng et al. [257]. A limitation here is that the time from electrical activity to sound production is not constant in all samples. Direct segmentation techniques involve utilizing Shannon energy to calculate an envelope and to find its peaks, and then use those peaks to reconstruct the cardiac cycle. Following sound segmentation, it is necessary to then classify these sounds. Leng et al. describe various machine learning techniques to classify these sounds, including support vector machines (SVM), artificial neural networks (ANN),

hidden Markov models (HMM), and Gaussian mixture models (GMM), for identifying sounds and identifying next likely sound given the state in the heart beat cycle currently detected. Leng et al. report that these techniques have accuracies near 90% for classifying signals as either normal or as having aortic or mitral valvular lesions [257]. In 2016 a collection of heart sounds was published [289] and this dataset has served as a standardized way to benchmark progress in identifying heart sounds. Work in this dataset was summarized by Clifford et al. in the 2016 PhysioNet Computing in Cardiology Challenge, who reported that several varied techniques reached high performance [290]. Notably, the top three models had completely different approaches but similar performances. Those three models consisted of AdaBoost and a convolutional neural network (CNN), an ensemble of SVMs, or a regularized neural network. Subsequent work has continued to improve on this task with performance improving with more sophisticated ensemble algorithms [291].

Work has also been done to develop low-cost devices that can act as a bridge between a traditional stethoscope and a cell phone [292]. Constructing a cavity with good resonance is necessary in collecting good quality sound transmissions from the stethoscope. In particular, Sinharay et al. have evaluated using different kind of sensors to capture sounds to be transmitted from and to smartphones for analysis.

In addition to detecting abnormal sounds in the cardiac cycle, there has been successful work in eliciting heart pathology from abnormalities within normal heart sounds. The normal cardiac cycle is composed of two sounds, S1 and S2. S2 in turn is caused by the superposition of two separate sounds occurring nearly simultaneously, one from the aortic valve closing and the other from the pulmonic valve closing. Both happen at nearly the same time, typically creating a single sound. However, some heart pathologies can impact the time between these. In a study of pediatric patients, high pressure in the pulmonary vasculature was found to be predicted by certain aortic and pulmonic valve relative intensities [293]. Although this work has not been applied to adult patients, it could theoretically help to elicit information about the pressures at different points within the heart.

In several cases, radar has been utilized instead of direct, on-body measurement for detecting vital signs. Radar is able to detect periodic changes caused by both breathing and the heart, allowing heart rate and respiratory rate to be detected. Vinci et al. described a remote sensor that uses a six-point radar to monitor respiration and heartbeat [294]. It uses a continuous 24 GHz wave and a radiated power of less than 3 microwatts. It captures these values noninvasively in patients at rest. This is notable as it is a sensing modality that does not require attaching sensors to the human body. This is particularly valuable in infants, in adults in severe conditions that cannot have additional attachments placed on the body, and as a modality that improves patient quality of life by limiting on-body sensors. The sensor designed in this paper does not have the limitations of other radar systems that require a wide frequency band to achieve more accurate results. Because of the six-point receiver architecture, this sensor can accurately measure angle and displacement by only measuring phase difference in backscatter patterns. Models regarding the permittivity of the skin allow them to estimate that their signal has 1.52 mm penetration as well as estimates of blanket and clothing impact. As a result, they can estimate where the edge of the torso is to aid in monitoring breathing. This provides an opportunity to noninvasively measure respiration and heart rate. However, it requires known, fixed postures of the individuals. Additionally, it will only work for one patient at a time. While this modality provides activity, displacement, and vitals monitoring in controlled, clinical environments or within specific remote environments (such as in the bedroom while asleep), it does not provide flexibility while moving. There are needs to extend such sensing systems to a variety of environments.

Work by Li et al. explore the use of radar technology for vital sign monitoring [259]. Their system uses a hardware-controlled clutter cancellation system. This allows their radar technology to identify the difference between the person being monitored and background clutter that are likely present in rooms the person would be in. Authors propose taking ka-band radar systems that are meant for motion sensing and modify them for vitals sensing.

Authors discuss existing work, design considerations for advancements, then opportunity to extend this to infant monitoring. The advancements in radar usage have come through the detection of the right frequency band to use. Different frequencies were shown to be able to go through different rubble with and without metal mesh. Authors then discuss the chip-level decisions that need to be made to create CMOS Doppler-based motion detectors. This allows vital sign detection through obstacles which can be important for noninvasive monitoring and for detection of vitals in emergency disaster scenarios. The application, however, is not clear for advanced signal processing of multiple vitals.

11.3.1.2 Electrical Measurements

Remote ECG monitoring has been utilized since the development of the Holter monitor in 1962 [295]. However, recent advances allow for not only recording of remote ECGs, but for real-time analysis and for longer periods. One necessary advancement for increased length of monitoring was the long-term electrode. Traditional wet electrodes are poor choices for long term monitoring due to their inconvenience [296]. Chi et al. surveyed a number of advancements in dry-contact and noncontact electrodes that have been developed [296]. Majumder et al. similarly survey numerous developments in dry electrodes that provide superior remote monitoring performance for long duration ECG monitoring [297].

Remote ECG monitoring has been explored by a number of researchers, primarily to solve the challenges that arise in noisy measurement. One issue that arises in continuous ECG monitoring, as with wearable ECG implementations, is that signals are often hidden by the noise of activity. Li et al. presented an approach for quantifying this noise [298]. While earlier approaches focused on labeling ECGs as either clean or noisy, the approach presented by Li introduced five classifications, each with different amounts of information available to be extracted from the ECG. They defined the noisiest strips as those where artifact obscures signals to the point that there can be no confidence in any interpretation of the ECG. Strips with severe noise were those where some interpretation could be made, but interpretations could be confused as to where the QRS complexes fell or to whether ventricular flutter

rhythms were present. In strips with moderate noise the QRS complex and presence or absence of ventricular flutter rhythms could be assessed, but finer signals such as P or T waves could not be extracted. Minor noise was the label given to strips with some amount of noise, but where P waves and T waves could be extracted. This level of noise allows for the analysis of atrial arrhythmias such as atrial flutter. Finally, clean ECGs were those where no noise was present. The authors produced training data by adding three types of noise to the original clean dataset: baseline wandering, electrode motion, and muscle artifact. They trained an SVM to classify strips based on the amount of noise present and validated this classification scheme on real noisy data. This validation showed good agreement between manually annotated labels and model output labels, with the greatest confusion present where samples had been manually annotated as having minor noise, but the model labeled the samples as having moderate noise. The authors note that a chief limitation of this work was that the model was not trained for or with an arrhythmia database, which substantially lowers its effectiveness on samples with arrhythmias. Additionally, they note that methods based on continuous features rather than discretely extracted features would be likely to show greater performance.

Once identified, several approaches have been implemented in order to account for and to correct motion artifacts. Sriram et al. addressed this problem by utilizing a triaxial accelerometer [299]. ECG signals are usable as a means of continuous biometric security. However, this continuous security is lost when the ECG signal is distorted with motion artifact. This approach shows that supplementing the raw ECG signal with features extracted from acceleration allows for accurate classification of ECG subject identity. They segmented signals to windows containing roughly four heartbeats, averaged those four beats together, and then corrected for baseline abnormalities with linear interpolation of q-minima and a high pass filter in association with the accelerometer features. These features then served to correctly identify users using either a k-nearest neighbors or a Bayesian network classifier.

Several wearable ECG devices have been developed recently. The BioStamp is a wireless

wearable device that received FDA 510(k) for medical use [300]. The BioStamp provides ECG signals that are comparable to a traditional ECG [301]. It also includes accelerometers and gyroscopes, and in a population of 30 healthy adults and was able to provide accurate measures of heart rate, heart rate variability, respiratory rate, activity, and sleep events [302]. Another FDA approved device incorporating ECG monitoring is the iRhythm Zio^{XT} [303] This device is applied to a patient as an adhesive patch, and was found to be more sensitive than a traditional Holter monitor at detecting arrhythmias [304]. This device is able to be worn for up to 14 days.

Another issue that arises with automatic ECG monitoring is that many abnormalities might be troubling in one patient while normal in another. Chen et al. [305] described an approach to train ECG monitoring systems to discover patient-specific abnormalities. This work utilized an accelerometer to reduce the number of false alarms in monitoring systems. Over time, this system learns the normal for a given patient and uses a knowledge of this normal in order to reduce false alarms.

11.3.1.3 Blood Pressure

The American College of Cardiology and the American Heart Association (ACC/AHA) recently released guidelines that suggest ambulatory blood pressure measurements, those taken at home in 15 minute intervals including during sleep, should be captured to better understand a patient's blood pressure and potential cardiovascular risks associated with HTN [267]. The sphygmomanometric and oscillometric techniques are well-established as the predominant means by which blood pressure is typically measured [306]. Both methods involve the inflation of a pressurized cuff, typically around the patient's upper arm and maintained at the level of the heart. The pressure in the cuff is increased to above realistic values of the systolic blood pressure, and then slowly decreased. In the auditory sphygomomanometric method, sounds called Korotkoff sounds can be heard just distal to the cuff as it deflates. The pressure at which these sounds are first heard is the systolic pressure, and the pressure at which these sounds are no longer heard is the diastolic pressure. In the oscillometric

technique, minute variations in pressure as the heart beats against the pressurized cuff are measured and the systolic and diastolic blood pressures are extracted from these variations [307]. Most at-home blood pressure monitoring devices utilize the oscillometric technique, which is well-validated to have performance similar in quality to the sphygmomanometric technique [308]. Recently, cuff-less blood pressure monitoring techniques have been explored in order to record blood pressure.

The most common cuff-less approach thus far is to use photoplethysmography (PPG) and ECG to capture pulse arrival time, pulse transit time (and pulse wave velocity), as surrogates for blood pressure, then use analytic techniques to estimate the systolic and diastolic blood pressure values [309, 310]. If the posture of an individual is known, these techniques are able to measure an estimate of the blood pressure, without disturbing the individual with frequent cuff inflations. However, the ECG and PPG combination can result in error in blood pressure estimation because it does not appropriately account for artifacts that exist between the ECG measurement of a pulse and the PPG capture of the pulse arrival time [311]. In particular, the ECG and PPG combination shortcomings are a direct result of the pre-ejection period of the heart. The pre-ejection period constitutes a time delay between the electrical stimulation of the heart and the actual mechanical expulsion of the blood for each heartbeat [312]. The pre-ejection period can vary under different conditions and is not easy to measure, leading to an unpredictable error in estimating blood pressure when using ECG. Vascular tone can additionally complicate this estimation. Vascular tone can change as patients age or take different medications, and these changes can increase this error [313]. To account for this, researchers have turned towards dual PPG capture [314, 315] over a small portion of the artery to account for pulse transit time, which are better able to locate the artery and avoid capturing blood profusion time into capillaries [316, 317]. Ballistocardiogram approaches look to capture pulse arrival time through the small changes in pressure sensed by the waves in each pulse, providing a method for capturing cuff-less blood pressure whenever participants are still [318, 319, 320]. These approaches all look

to address cuff-less blood pressure when the participant is in a fixed, known position, and provide the opportunity for more frequent ambulatory blood pressure measurement.

More recently, bio-impedance based approaches have also been developed to measure blood pressure in a cuff-less manner [321]. The impedance signals allow the sensors to identify the location of the arteries within the wrist, eliminating errors in blood pressure estimation that are a direct result of the pre-ejection period or the misplacement of light sources that may capture both the pulse transit time in the artery and blood profusion through the capillaries. Estimation of blood pressure characteristics were then made by extracting characteristic features from the multiple bio-impedance channels. This is enhanced by adding other heart beat characteristics, including capturing the inter-beat intervals for heart rate and heart rate variability characteristics [322], as well as respiratory rate [323].

11.3.1.4 Blood Flow

Blood flow is a complex system characterized by pulsatile flow in a dynamic system [324]. While measurements related to arterial blood pressure are often a good proxy for systemic blood flow, different physiologic or pathologic states can alter this relationship [325]. Most notably, isolated vasoconstriction or a thromboembolic event can cause flow along an artery to drop while systemic pressure is relatively unchanged, or atherosclerosis can cause chronically decreased flow to various organs [326]. Ultrasonography can be used to assess blood flow along an artery [327, 328] and can also be used to estimate degree of systemic atherosclerosis [329]. Magnetic resonance imaging (MRI) can also be used to measure blood flow [330].

11.3.1.5 Fluid Retention

While prior studies, such as Tele-HF and Beat-HF, attempted to use weight scales as a surrogate for fluid retention in HF, the measurement of 3 pounds of weight change was not an alert that was able to reduce HF readmissions [242, 243, 244]. A number of attempts to measure peripheral edema and fluid retention have focused on the development of smart

socks that look to measure fluid buildup in the ankles [331, 332]. A stretch sensor measures the expanding duration of the patient's ankle both as edema increases throughout the day and as edema increases over time. The context-awareness allows the device to discard ankle measurements when motion, muscle contractions, or an incorrect posture would interfere with the measurement. This sock was able to reliably determine the participant's posture, and measurements of fluid retention were well correlated, but additional study is needed to determine if this measurement is accurate enough, and whether it can generate alerts early enough to intervene in HF patients. Yao et al. came to similar conclusions of needing further study of their sensor to classify edema [333], as this remains an open area of research.

11.3.1.6 Physical Activity and Posture

Activity, posture, and pain are important measurements in understanding symptom and treatment effectiveness in patients diagnosed with cardiovascular disorders. Measurement of respiratory distress in HF patients requires a measurement of posture, measurement of blood pressure through proxy measures such as pulse transit time require a measurement of posture, as did the smart sock for fluid retention (Section 11.3.1.5). While each sensor can capture posture, smartphones excel at this [334], often coupled with other applications tracking activities of daily living [335, 336]. Recently, smartwatches have shown to accurately detect postures and exercises [337, 338], which is important for patient monitoring, since smartphones are often in the proximity of the user, but often not physically on the user, unlike smartwatches [339]. These can also provide important context to the measurements captured by the other modalities discussed in this section [253].

11.3.1.7 Diet Monitoring and Glucose Intolerance

Thirty million Americans live with diabetes, and another 80 million have pre-diabetes, a condition that left untreated often leads to diabetes [340]. Diabetes occurs when blood sugar is too high due to poor nutrition (e.g., too many refined carbohydrates) and/or inadequate insulin regulation (i.e., insulin resistance). Sustained high levels of blood glucose

can have disastrous long-term health consequences, including cardiovascular diseases. An essential component of clinical interventions for diabetes is monitoring dietary intake, as it can help individuals and health practitioners manage dietary habits and understand how dietary choices affect blood glucose. Various sensing techniques have been explored to capture dietary intake, such as wearable sensors (microphones, accelerometers) to detect eating behaviors such as hand gestures and chewing/swallowing [341], or computer vision techniques to recognize foods from photographs [342]. Using continuous glucose monitors has allowed researchers to develop models of estimated food intake [343], and when coupled with other personal measures, such as gut microbiome data can provide educational information towards treating glucose intolerance at a personalized level [344]. Not only is glucose intolerance, and a diagnosis of diabetes, a key factor that increases risk of cardiovascular disorders, but other parameters, such as salt intake, may impact blood pressure [345]. More recently, authors have shown that detecting glucose excursions, such as hyperglycemia or hypoglycemia is possible from ECG signals [346]. This provides a potentially non-invasive way to track glucose variability while primarily developing sensors for tracking risk factors of a primarily cardiovascular nature.

11.3.2 Gaps

Table 11.3 summarizes the key developments in sensing including remaining gaps in the technologies. As these technology gaps are addressed, richness of the available data will increase. As richness of data increases across the variety of sensors, the potential for noise and missingness increases as well. It is difficult to understand the context in which measurements are captured. Accuracy of posture detection and presence of other noisy attributes impact the potential success of different sensing modalities. It is also unlikely a patient will wear all sensors all the time, as this will provide excessive burden. While a measurement performed on occasion is likely to be a high-quality measurement, continuous and automated measurements introduce a greater deal of variability in the quality of measurements. For instance, a once-a-day measurement is likely to be a measurement where the patient will intention-

ally position themselves appropriately and remain motionless during the measurement. A patient monitoring their blood pressure will likely sit upright with their legs uncrossed, or a technician performing an ECG will ensure that the printed ECG is taken at a point where the patient is motionless, and no artifacts are present. Conversely, more frequent or continuous monitoring must account for noise introduced by motion artifacts as well as from noise introduced from other sub-optimal measuring conditions. As such, a number of challenges remain in capturing the necessary signals:

- Acoustic measurement: Non-wearable sensors are limited by the challenge of identifying a particular patient when multiple people are present. Wearable sensors must account for noise across a variety of motions, environments, and potential sensor misplacement.

- Electrical Measurement: Continuous ECG requires multiple leads to be worn at the same time. Devices such as the Apple watch provide potential for requesting ECG periodically when other sensing modalities dictate when it is necessary [260], but the correlation between these modalities and necessary ECG readings has not been well studied outside of AFib.

- Blood Pressure: Pre-ejection period and vascular tonal changes can impact estimation, resulting in pulse transit time calculations capturing both the arterial pulse as well as profusion into the capillaries. Additionally, misplacement of sensors may alter the accuracy of the readings, impacting performance of analytic models used to estimate blood pressure from data captured by these sensors. Cuff-less blood pressure monitoring must extend to continuous, beat-to-beat measurements without constantly restraining users to fixed, known postures.

- Fluid Retention/Weight Change: Edema measurements have not been clinically validated to show the degree of fluid retention which must generate alerts that can clinically improve outcomes.

- Physical activity and pain: Remote measurement of acute and chronic pain remains an open challenge.

- Glucose Intolerance: Tracking of diet, nutrition, and the direct link to cardiovascular care remains an open-ended problem without the use of invasive glucose monitoring technologies.

11.3.3 Opportunities

An additional source of noise can be introduced by the redundancy of signals that can exist. Different physical phenomena can be measured by different modalities, many of which will produce slightly different readings. Heart rate can be derived from multiple sources: auditorily by stethoscope, electrically by ECG, optically by PPG, and electromagnetically by radar. It stands to reason that these redundant values could be exchanged for each other, but that exchange may not completely be a one-to-one relationship. Transfer learning is an ongoing field of study that seeks to apply existing models to data that was not used in training or was only used minimally in training [347, 348, 349]. Transfer learning could be applied to this problem as a way to apply a single model to patients with disparate data collection modalities.

Missingness in data also increases as richness increases. While binary parameters used in many risk models (e.g. history of HF, current diabetic status, etc.) are easy to collect and even possible to impute, continuous monitoring opens the possibility of more complicated missingness. A battery may fail on a sensor leading to a variable period of missingness. Wearable sensors may introduce missingness secondary to poor compliance or poor utilization. The missingness introduced by gaps in continuous monitoring is more difficult to impute and presents a challenge in building comprehensive models [350, 351]. Deep learning techniques to address missing data have shown promising results, however, simple imputation of time-series signals is currently the best approach [350], leaving the door open to further work to address this at the sensors and anlytics level. A number of opportunities emerge for im-

mediate and impactful research on sensing signs and symptoms of cardiovascular disorders, illustrated in Figure 11.3, and listed below:

- Integration of multiple sensing modalities into a single platform, reducing the number sensors needed to be worn. High impact areas appear to be the wrist (smartwatch) and chest (heart and lung sounds). Analytics that leverage this integration will be discussed further in Section 11.4.

- Using analytic techniques to estimate parameters traditionally captured invasively with non-invasive surrogates (e.g. glucose and hypoglycemia using wearable ECG.

- Integration of machine learning techniques to help identify when longitudinal data capture is necessary, similarly to ECG requests to verify periods of arrhythmias associated with AFib detection with the Apple Watch [260].

- Transfer learning, when coupled with uncertainty quantification techniques, enables improvement of model performance through personalization *(See Section 11.4.1.5)*. However, when accounting for varying sensor types of the different domains, techniques are needed to quantify what domains of data and what quantity of those data are needed to transfer learn. Additionally, knowing which portions of models to re-train in a transfer learning mechanism should be further explored.

11.4 Continuous Data Collection and Analytic Models

Beyond the acute sensing and detection of symptoms related to HF, CAD, and stroke, analytic opportunities arise in the processing of this data longitudinally and continuously. As discussed, the progression of CAD from stable and unstable angina to NSTEMI and STEMI represent longitudinal changes that may have periods of rapid change interspersed. Similarly, untreated HTN can lead to stroke if untreated. Changes in heart remolding in HF may be represented by changes in heart sounds as captured by acoustic sensing. Patients living with HF may experience long term changes in the amount of physical exertion required

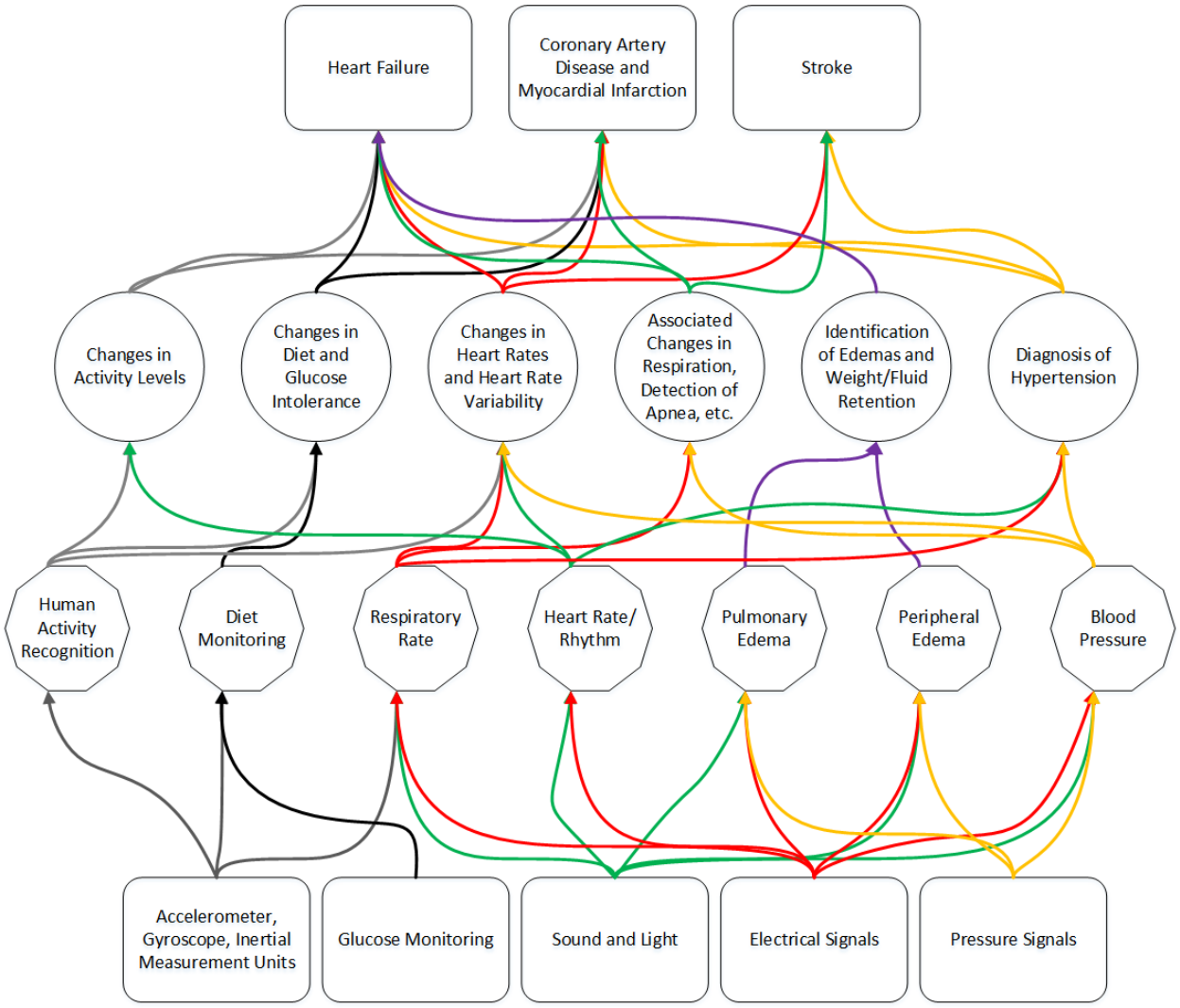


Figure 11.3: Overview of selected sensor categories proceeding to selected signs and symptoms measured and their potential progression to adverse events and diagnoses. The number of crossing connections illustrate the commonality in risk factors that can be sensed in progression to primary adverse events and secondary recurrent adverse events for a variety of cardiovascular conditions. The colors are only illustrative of different pathways in each level and are not meant to be illustrative between subsequent levels.

to perform activities of daily living. These changes may be gradual and unnoticeable to the patient, but may represent worsening condition or recovery.

These cardiac conditions present a range of analytic techniques necessary to capture longitudinal changes in continuously-sensed data:

- Continuous capture of acoustic sensing: Understanding how sounds change over time may allow for the identification of new signals that represent earlier identifiers of worsening conditions or treatment effectiveness. *(See Section 11.4.1.1)*

- Continuous capture of electrical signals: While the detection of arrhythmias may be present in surrogate measures such as heart rate, detection of changes in ST segments of an ECG may allow for early alerts and acute care. *(See Section 11.4.1.2)*

- Continuous capture of vitals signals: Understanding the changes in the variety of vitals signals captured and how they may relate to each other can provide an understanding of improving or worsening risk factors relevant to HF, CAD, and stroke. *(See Section 11.4.1.3)*

- Continuous capture of physical activity: Physical activity and sleep are important functional measures of recovery, and accurate, longitudinal understanding of functional change can be correlated with improved mortality and prevention of adverse events. *(See Section 11.4.1.4)*

- Deep learning techniques for data analysis and modeling: A variety of deep learning techniques have the ability to develop personalized models using continuous, longitudinal data. While long short-term memory networks (LSTM) and general Recurrent Neural Networks (RNN) provide a standardized framework for signals, this section explores modification of existing techniques to work with a wider array of data discussed in this section. *(See Section 11.4.1.5)*

11.4.1 Existing Technologies and Applications

11.4.1.1 Continuous Capture of Acoustic Sensing

A primary application of acoustic sensing is for the assessment of cardiac murmurs [288]. Most auscultative techniques have not been developed for continuous monitoring but are rather focused on individual discrete observations. However, continuous wave Doppler monitoring can be used in fetal monitoring [352] and continuous fetal monitoring has been shown to have superior outcomes relative to intermittent monitoring [353]. There has been some work in extending this technology to continuous adult cardiac auscultation [354]. Mc Loughlin and Mc Loughlin found that continuous auscultation was able to detect impaired ventricle relaxation and lesions of the aortic and mitral valves with higher sensitivity than was available with traditional auscultation alone [354]. However, there is a pronounced absence of further work in continuous cardiac auscultation.

Electronic auscultation is useful for deriving characteristics of other parts of the cardiovascular system than sounds generated specifically by the heart. A carotid bruit is a sound created by turbulent blood in a carotid artery, often caused by narrowing that in turn is produced by atherosclerotic plaques. Knapp et al. looked at the effectiveness of carotid bruit detection by electronic auscultation [355]. Out of 1,371 patients in this study, 84 were found to have carotid bruits by electronic auscultation. These patients were matched with controls who did not have bruits, and both patients from each pair were assessed with duplex ultrasound to determine extent of carotid stenosis. Bruit detection with electronic auscultation and manual annotation was found to have a sensitivity of 88% for stenosis $\geq$ 50%, and a specificity of 58% with duplex ultrasound providing the ground truth.

Work by Palaniappan et al. surveyed machine learning techniques to further analyze lung sounds [356]. They evaluated 59 papers that used signal processing and machine learning techniques on a variety of lung sound problems including normal breath sounds, abnormal breath sounds, and a series of sounds called adventitious lung sounds. This survey high-

lights an important need by evaluating short term sounds, long term sounds, and identifying normal and abnormal sounds across the different time periods. Most works in this survey focused on specific frequencies (between 150 and 2000 Hz, though they found that most work typically worked at 150 Hz), and evaluated machine learning techniques such as k-nearest neighbor, ANNs, HMMs, GMMs, genetic algorithms, SVMs, and fuzzy logic to classifying a variety of lung sounds. They found that by using piezoelectric microphones, contact microphones, and electric microphones, and one commercially available lung sound instrument, they could design electronic stethoscopes that filtered out heart sounds to capture necessary lung sounds. Similarly, one could use the same techniques to filter out the lung sound to capture the heart sounds. Using standard time-domain and frequency-domain signal processing features, algorithms were able to classify lung sounds with between 83-93% accuracy. Rocha et al. published a database of lung sounds that were used in the 2017 ICBHI Scientific Challenge as a challenge for lung sound classification [357]. These sounds consisted of wheezes, crackles, wheezes and crackles, or normal breath sounds. Several groups have achieved good performance on this dataset by applying CNNs to this dataset [358, 359]. There is continuing work in applying RNN and LSTM architecture to this task as well [360, 361]. Work in this domain is largely limited by the large variation in pathological sounds and by a lack of additional publicly available datasets. Given the traditionally subjective nature of sound interpretation, there has also been some disagreement in lung sound nomenclature [362]. In addition to wheezes and crackles, there are many other sounds which should be included in training, including rhonchi, pleural rubs, diminished breath sounds, and differentiation of crackles into either fine or coarse crackles. Another step that could be accomplished in this domain is the replacement of particular sound identification with the identification of the underlying pathology. As additional data is collected and annotated, further developments should be made possible.

11.4.1.2 Continuous Capture of Electrical Signals

In clinical settings, most ECGs are performed as 12-lead ECGs. In these ECGs, there are 10 electrodes attached to the patient and 12 different measurements taken from these electrodes. Each provide a one-dimensional view of the magnitude of the vectors of all electrical impulses in the heart relative to a given axis. Different axes allow for information to be obtained about the functionality about different parts of the heart. Depending on the goals of remote monitoring, remote ECGs will typically only include a subset of these typical views. As a result, methods that can accurately detect essential signals from minimal lead ECGs are necessary.

Work by Jambukia et al. surveyed machine learning techniques to analyze and classify ECG signals [363]. They evaluated 31 papers that used signal processing and machine learning techniques in order to extract clinically significant features from raw ECG signals. Most of the papers evaluated used the MIT-BIH arrhythmia dataset [364] for both training and testing purposes. Two aspects of ECG classification considered were ECG beat classification for individual, isolated beats, and ECG signal classification for interpretation of a longer signal. Some approaches evaluated involve signal feature extraction followed by threshold-based algorithms such as the Pan-Tompkins algorithm. Other approaches utilized various neural network architectures, with the authors finding that of the architectures studied, multilayer perceptron neural networks provided the best performance. Recurrent neural networks, such as the LSTM architecture, were not evaluated in this survey. Deep learning techniques have also been utilized for ECG evaluation. Yildirim showed that a bidirectional LSTM architecture can reliably classify five different rhythms from the MIT-BHI arrhythmia database [365]. This bidirectional LSTM model achieved accuracies greater than other techniques. Additional deep learning techniques that combine CNN and LSTM have been used to detect AFib without explicit feature extraction (such as R peak extraction) [366]. Further deep learning techniques have looked at a variety of processing individual beat anomalies and sequence anomalies [367], though time series presented to CNN models often needs fixed

windows of time to be pre-determined for evaluation. Additionally, some work uses a single lead [368] for detecting arrhythmia, though it is likely at least two leads are currently necessary for other ECG feature extraction.

There is evidence to suggest that patients at risk of cardiac pathology benefit from more continuous remote ECG monitoring. The mHealth Screening to Prevent Strokes (mSToPS) randomized clinical trial is an ongoing trial of 2659 patients investigating the benefit of continuous monitoring for AFib [303]. As reported by Steinhubl et al., the initial phase of the trial discovered that for individuals at risk of AFib, home ECG monitoring was superior to routine care for discovering new incidence of AFib. In the actively monitored group, there was a 3.9% diagnosis of new-onset AFib, vs 0.9% in the control. This resulted in earlier initiation of anticoagulative therapy (a preventative measure for stroke) in these patients. However, this has also resulted in a higher healthcare utilization among these actively monitored patients. This trial is still ongoing- the ultimate clinical impact is still unknown. Clinical outcomes are due to be published in a 3-year follow-up.

11.4.1.3 Continuous Capture of Vitals Sensing

Ultrasonography is a technique that uses ultrasonic sound waves to produce images of tissues beneath the skin. Ultrasonography is valuable for visualizing structures that are unreachable noninvasively. In hospital settings, point-of-care ultrasound has increasingly grown in utilization as mobile ultrasound systems become cheaper and comparable in quality to larger ultrasound systems [369]. Point-of-care ultrasonography is useful as a tool that physicians can bring to the bedside for aid in diagnosis, much like a stethoscope, but deep learning techniques are necessary to evaluate the ultrasound images and classify changes in conditions.

Ultrasonography can also be used to evaluate the fluid status of the lungs. As described in Assaad et al., lung ultrasound is a valuable tool for quickly assessing the health of a patient's lungs [370]. Certain visual findings, such as "B-lines" are highly associated with edema and various pulmonary pathologies. These visual findings also change very rapidly,

reflecting the present disease state more accurately in some cases than measures such as blood oxygen saturation. Lung ultrasonography is also useful in differentiating between cardiogenic and noncardiogenic pulmonary edema; cardiogenic pulmonary edema typically shows more uniform findings and plural effusion (fluid buildup in the tissue surrounding the lungs). Lung ultrasonography is an underutilized technique in medicine and lacks standardization in training and implementation.

Work by Bhuyan et al. explores an exciting possibility of wearable ultrasound for the monitoring of internal function noninvasively [371]. In order to create a small form factor that could be used to measure organ function with wearable, remote ultrasound, they created a small, flexible probe through a flexible PCB integrated circuit. They also used a system that has only one transmit and one receive channel to avoid excess signal degradation. This system has a bandwidth of 10 MHz, power consumption of 6.72 mW per channel, and uses 16 such channels to measure a 5.6 mm x 1.6 mm area. They used classical image processing with ultrasound for their validation. Their system, however, used an attached cable to measure. There is an opportunity to create a remote, continuous version of such a system if a flexible PCB-based wearable ultrasound with necessary battery and wireless transmission capabilities were added, but, computer vision techniques are needed to enhance the analytic component of the wearable ultrasound.

Echocardiography is the practice of using ultrasound in order to visualize the structures of the heart. Echocardiography can take place either as an invasive transesophageal echocardiography (TEE) or as the noninvasive transthoracic echocardiography (TTE). Many different aspects of the heart can be described and quantified via echocardiography [372], including size, function, and mass of various structures of the heart. Measurement of these parameters aids in the diagnosis of HF. For instance, left ventricular mass or poor emptying are markers of HF. Valvular disfunction, such as stenoses or regurgitations can be directly observed. These measurements also aid in assessing cardiac function in CAD, particularly following MI; injured portions of the cardiac wall will often move less than they normally

would.

Various groups have found preliminary success in applying deep learning computer vision techniques to the analysis of ultrasonographic images. The first step in automatic analysis of ultrasonographic images is to recognize the view in question. Østvik et al. describe the use of a CNN to classify TTE images according to the view being presented [373]. This method showed classification high accuracy in distinguishing among seven different TTE views. Additionally, the authors described a technique for extracting 2D slices from 3D images and achieved a mean error of 4°.

Techniques for measuring edema include cuffs that track ankle circumference and measurement of electrical impedance. Weight monitoring is sometimes used as a proxy for tracking edema, as edema co-presents with fluid retention. There has been success in implantable impedance monitors to measure pulmonary edema. Yu et al. found that intrathoracic impedance serves as a predictor for imminent hospitalization due to fluid overload [374]. In a population of 33 patients with HF, a device consisting of a pacemaker and defibrillator was implanted. The device measured the impedance between those two leads. This study found that there was a significant decrease in impedance prior to hospitalization with fluid overload. This decrease began on average two weeks prior to hospitalization and continued through the date of hospitalization.

Impedance monitoring has also been implemented in noninvasive and ambulatory monitoring systems. Weyer et al. describe a system that incorporates both ECG and noninvasive impedance cardiography [375]. This device includes Bluetooth connectivity and a battery that lasts for up to 21 hours. This system could be implemented for long-term monitoring in patients with HF to monitor pulmonary congestion and to potentially allow remote interventions before hospitalization is necessary.

The internal and external jugular veins provide drainage from the head into the heart. The right jugular veins are positioned almost directly above the right atrium, and therefore the pressure within them is very closely tied to the pressure of the right atrium. The external

jugular vein's filling level indicates the pressure within the right atrium and will be distended in cases of right heart failure. Pulsations can be observed with great difficulty in the internal jugular vein. These pulsations provide evidence as to the relative timing and forces involved in right atrial contraction, atrial relaxation, right ventricular contraction, venous filling after the closing of the tricuspid valve, and emptying of the atrium after opening of the tricuspid valve.

As venous pressures are so much lower than arterial pressures, measurement of the jugular venous pulse is much more difficult than the measurement of arterial pulses. However, Amelard et al. were recently able to utilize a technique called PPG Imaging (PPGI) as a viable technique to correctly measure the jugular venous pressure [376]. This technique uses a system located approximately 1.5 meters away from a supine patient. A light shines on the patient and the reflected light is analyzed to identify pulsations. The arterial pulsation from the carotid artery is easier to detect, and the jugular pulsation can be identified as a corresponding inverted pulsation at a location near but lateral to the arterial pulsation. In this study, the ground truth arterial waveform was verified with a PPG measuring device. Pertinent clinical features were consistently able to be extracted from the venous waveform, including the c, v, x, and y waves (corresponding to the contraction of the right ventricle, systolic filling of the right atrium, relaxation of the right atrium, and beginning of the filling of the right ventricle). In about half of subjects, the a wave was also observed (corresponding to the contraction of the right atrium). The ability to regularly monitor and quantify these waveforms could allow for new techniques in monitoring right heart function.

Signals that capture continuous blood pressure, described in Section 11.3, may also be extended to capture a variety of heart rate, heart rate variability, blood pressure, respiratory rate, and changes in these values [323, 377, 378]. Obstructive sleep apnea, a condition in which airwaves are restricted causing the body to wake up from sleep to begin breathing again, increases heart rate, respiratory rate, and blood pressure, keeping patients from falling asleep. This has a direct relationship with blood pressure and nocturnal nondipping HTN,

and treatments for apnea have shown to be correlated to improvements in blood pressure [379]. This means approaches for measuring cuff-less blood pressure cannot be restricted to periodic, ambulatory measurements, but must transition to continuous beat-to-beat measurement and interpretation.

Finally, telemonitoring trials for HF readmission have tracked longitudinal measurements of symptoms, vitals, and patient qualitative reports [242, 243, 244]. In the Tele-HF study, a number of vitals signals and patient reported outcomes generated alerts for interventions if the values were below a specified threshold, or represented a significance drop from the prior day's values. However, the study was unable to find a statistically significant reduction in readmissions in the intervention arm. Ong et al. in the Beat-HF study tried to use some machine learning techniques to further identify risks of adverse events, and while the study was unable to reduce readmissions, the techniques did show some promise in stratifying patients [380], as did further statistical techniques applied to the Tele-HF data [243, 125]. With the addition of more signals captured, and techniques that can better account for varied time-domain aspects of analytics, it is possible that better just-in-time alerts can be generated for preventing future recurrent HF events.

11.4.1.4 Continuous Capture of Physical Activity

For cardiovascular disorders, the detection of activities and postures is important in understanding the other biomarkers captured, providing context for their readings. For example, at nighttime, knowing the posture of the user provides context for dyspnea measurements and heart sound recordings for HF patients. In addition to providing context for the other vitals measurements related to cardiovascular disorders, the change in physical activity performance can show increasing effects of HF symptoms, pain as a result of CAD, and acts as a surrogate for the general well-being of these patients.

Many research-oriented activity recognition platforms focus on the detection of activities of daily living [253, 381, 382, 383] and understanding daily exercise intensities. These sensors are capable of tracking sports movements in the healthy and measure sedentary time in the

elderly, and come in many forms of smartwatches, smartphones, smartwatch-sized sensors [337, 381], embedded within shoes, and most recently within eTextiles [384, 385].

HF participants have had improved outcomes in mortality and readmission when involved in cardiac rehabilitation programs that encourage continuous physical exercise [386]. This physical exercise routine has shown that measurements in improved peak exercise capacity correlate with improved cardiovascular outcomes. Home-based cardiac rehabilitation systems centered on physical activity detection in order to quantify a home-based exercise routine [387]. However, such systems do not yet quantify improvement directly from physical activity measurements. This is necessary since adherence in cardiac rehabilitation programs is often quite low [388].

11.4.1.5 Deep Learning for Personalized and multi-modal models

Deep neural network techniques have enabled the analysis and modeling of the data gathered from these sensing systems for a variety of event detection techniques. RNN and LSTM are deep learning models that are particularly well-suited for developing models for event detection on time-series data, such as segments of ECG signals [389], blood glucose [390], sleep [391], and in general are good for medical diagnostics using time-series data [392]. Often, the key to these techniques is the ability to generate its own features. For example, improvement in processing of ECG signals with deep neural networks rather than other machine learning techniques allows for the automatic identification of arrhythmias [393]. CNNs, when combined with LSTMs allow for robust, automatic feature engineering that improves the classification of signals extracted from wearable sensors.

These models are further improved through two techniques: through personalization and through integrated multiple sensing modalities together in one model. While these deep neural networks are able to extract features that represent important classification properties, person-to-person differences may impact model performance. Therefore, techniques that can train on a user's own data then perform functions on later-captured data can show improved performance [346]. Personalized modeling techniques have been used to detect and warn of

cardiac arrhythmias [394], and risk of recurrent events in HF patients by tracking changes to cardiac biomarkers [395], which presents a modeling opportunity if sensors can be developed to track those biomarkers in remote and ambulatory settings.

While having enough properly-labeled data is a challenge, uncertainty quantification techniques can identify when labeled data is necessary to personalize models for improved performance. Deep neural networks are particularly well suited to this task because of the ability to implement uncertainty quantification techniques on the probabilistic output generated by the models as well as rapidly re-weigh the network through transfer learning or domain adaptation techniques. In HAR tasks, for example, uncertainty quantification techniques that look at the maximum entropy measure from the generated predictions to determine if the model is performing well on existing activities or identifying new users or activity types [396, 219]. Once these periods of uncertainty are found, new data can be captured and transfer learning techniques identify what part of the deep neural network must be modified to account for the new user, new activity, or new sensor type [397, 349, 251].

A number of tools have been developed for assistance in annotating subject-dependent data [398]. An initial challenge in the annotation of data is event detection and segmentation. Adams et al. described a model for event detection and activity segmentation in wearable sensor data streams [399]. They validated their model on several datasets, including one in which events were instances when a user took a puff from a cigarette and activity segmentation was determining whether a user was smoking or not at a given time. This allowed for a system by which smoking event analysis was able to proactively provide feedback to assist in smoking cessation [400]. Labeled data is often collected in artificially constrained environments. However, several groups have focused on developing approaches for collecting annotations in natural environments. Akbari et al. described one such algorithm [401]. This algorithm was demonstrated on a smartwatch, and requested the user to annotate activities whenever uncertainty exceeded a threshold, but limited these requests to prevent overwhelming the user with request fatigue. Similarly, Fallahzadeh et al. describe an algorithm which

uses context in order to determine the optimal time to request annotations from the user [402]. Each of these techniques allows for collection of personalized annotations which could then be used in training of more sophisticated models.

While requiring a large amount of data, deep neural networks are well suited to improving classification tasks and analyzing the data from wearable sensors of various sources. Multiple-sensor fusion approaches have improved a number of modeling tasks, including estimating of blood pressure from ECG and PPG signals [403] and HAR through the use of inertial measurement units that integrate accelerometers, gyroscopes, and magnetometers, with use of multiple sensors across the body [404]. Additionally, multi-modal deep learning has been used to estimate certain biomedical signals with devices primarily meant to capture other signals, such as smartphones for heart rate estimation [405] and in recognition tasks outside of the cardiovascular domain, including stress monitoring [406], highlighting a key gap and opportunity to improve sensing in cardiovascular care.

In a multi-modal setting, data synchronization, sensor selection, and power optimization are important, in both settings with multiple sensors and settings with sensors integrated into a single platform device. When distinct events are to be detected and classified that event detection can be used to synchronize the data streams to ensure all are capturing the same events at the same time [407]. Additionally, sensor selection techniques can identify the key features for specific recognition tasks and reduce the number of sensors needed, optimizing the power [396]. Synchronizing data and optimizing usage of sensors and power is of extreme importance in longitudinal use of these sensors, but are themselves a subject that requires deep analytic technique reviews out of the scope of this cardiovascular disorder survey, given the current lack of integrated sensing platforms for important risk factor monitoring.

11.4.2 Gaps

Deep learning techniques have made the exploration of time series data more fruitful with the development of automatic features that represent longitudinal risk or outcomes. Techniques such as attention-based LSTMs have shown promise in exploring continuous

time-series data to predict clinical mortality, decompensation, and length of stay, which outperform hand-crafted feature extraction, and other deep-learning techniques that do not find focus on specific periods of time, in intensive care unit data [408]. However, these techniques have not been applied to this remote data yet because the integration of these sensing techniques have not yet occurred.

A number of the analytic techniques tied to the use of these new sensing paradigms have focused on the diagnosis of specific anomalies or classification of specific types of sounds or signals captured. What is still needed is the following:

- Integrated sensors that can capture signals frequently, or continuously, over entire study periods.

- Machine learning techniques that can explore multiple windows of time over multiple combinations of available signals in order to quantify trajectories in signals, identifying longitudinal patterns and changes in signal that may be indicative or worsening conditions or treatment effectiveness and recovery, extending beyond anomaly detection and signal classification.

- Data synchronization in multi-modal platforms, identifying how many sensors are needed for an application and minimizing the user burden in wearing them and needing to re-charge them is an important problem that will need to be addressed as these systems are developed for longitudinal use.

- Deep learning techniques listed demonstrate how personalization can improve model performance. However, personalization requires the labeling of data samples for supervised learning techniques. The longitudinal capture of these labels may result in undue burden on the users of the system. Finding a balance between the types and quantity of data needed and the passive collection is of utmost importance for user adherence.

11.4.3 Opportunities

One way in which advanced work in analytics could be incorporated is in better personalized monitoring for risk of CAD. As previously discussed, CAD is a condition often characterized by gradual worsening of chest pain that culminates in a heart attack. Ideally, monitoring systems would be able to follow gradual changes prior to the rupture of a plaque that causes a heart attack. In the early stages of CAD, activity monitoring could be used to assess wellness. By tracking a certain threshold of activity that the patient does not (or cannot) exceed, a monitoring system can estimate the severity of angina. As that threshold begins to decrease, the patient's angina is likely increasing and greater intervention may be warranted. The monitoring system for a patient at risk of CAD should also include an ECG system to watch for changes associated with a heart attack. If any electrical changes concerning for a heart attack begin to appear in the patient's ECG, then emergency services would be required. Earlier interventions are associated with better outcomes, and a monitoring system like this coupled with improved analytics could potentially allow for earlier treatment, leading to less overall damage and better patient outcomes.

Improved analytics could also be implemented to better treat valvular diseases. Unlike the other pathologies discussed here, there are few risk models for predicting future valvular heart disease. However, advanced analytics could be implemented to allow for earlier detection of valvular disease. As discussed above, these abnormalities change the way in which blood flows through the chambers of the heart, producing turbulence that can be detected as sound. The most straightforward evaluation for valvular disease in remote wearable settings would involve electronic stethoscopes continuously monitoring the patient's heart sounds. By learning the normal sound profile of a patient, new changes and murmurs could be quickly identified. After identifying a particular valvular disease and its associated murmur, long-term monitoring with electronic stethoscopes could be used to characterize the severity of the valvular insult; most murmurs initially increase in intensity, but in later disease stages decrease in intensity. Rather than risking false negative screening in physical examinations,

a longitudinal monitoring system could detect the changes along this trajectory to allow for more informed decision making. As a more advanced option in monitoring valvular disease, miniaturized ultrasound probes could be incorporated into a wearable system. These could be used for imaging and analyzing the valvular parameters such as cross section and flow. Additional work into computer vision interpretation of ultrasound images would be necessary in order to automatically process these signals. Vital monitoring can also directly feed into an understanding of valvular disease. In particular, blood pressure can reflect aortic valve lesions. Finally, systems to monitor valvular disease could monitor symptomatic disease progression. As many types of valvular disease may ultimately lead to HF, the opportunities presented above for HF apply here as well. Foremost among them would be activity recognition, where late stage valvular disease can manifest with a loss of stamina in day-to-day activities.

Opportunities:

- Improved Machine Learning Processing of Existing Sensor Modalities: Development of machine learning techniques that can extract meaningful data from non-numerical sources, expanding on the computer vision work done in automatically processing and interpreting ultrasound images.

- Time-Series Machine Learning Models: Development of machine learning techniques that can process longitudinal data and account for multiple channels of data, sampled at different frequencies, and with different segment lengths of importance, are required to develop new risk prediction techniques and alerts based upon continuously captured data.

- Applying attention mechanisms to deep learning techniques to interpret what features are being extracted and better understand the interdependence of the multi-modal learning techniques will enable more rapid selection of key sensors for longitudinal tracking and event detection.

11.5 Clinical Interpretability, Analytic Models, and Treatment Paradigms

Clinical risk prediction models and those that predict adverse events have helped guide medical treatments and improve patient care. These techniques, with machine learning modeling, have the potential to improve clinical care in both the acute care settings [409] and remote care settings. This includes understanding the diagnosis and progression of diseases and the personalized patterns and signals that can be captured by advances made in categories listed in Sections 3 and 4. Some preliminary work has been conducted in clinical trials on HF patients, understanding distinct patient phenotypes within the disease. In one such HF trial, clinically-distinct clusters of patients were found to have different time-to-event predictions and outcome rates [57]. Another relevant clinical trial in HF patients is the Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist (TOPCAT) trial [410, 411]. The purpose of this trial was to determine if a treatment designed specifically for HF patients with preserved ejection fraction could improve outcomes, a patient population where such treatments have not been found to universally treat these patients. This trial was also unsuccessful in showing that HF patients treated with Spironolactone had better outcomes [411]. However, due to some issues with data gathered in certain regions, investigators began taking a closer look at subsets of patients, to determine if specific patients were actually helped by the treatment. The investigators found regional variations lead to different treatment effectiveness in cohorts of participants [412, 410]. This indicates that HF patients diagnosed with preserved ejection fraction may benefit from cluster analysis, looking at personalized differences in outcome rates where different treatments may be helpful for different subsets of patients. These provide for the basis of the following needs:

- Risk prediction models: as illustrated by the TOPCAT findings, these diseases are quite complex and understanding the person-to-person variation allows for specific risk prediction based upon data collected, along with matching techniques that allow for comparison to patients most similar to the individual modeled. *(See Section 11.5.1.1)*

- Dynamic adaptation: models must account for the varied data types potentially collected, the varied rate at which they are collected, how well to link them to data gathered in acute care settings, and be able to update as a disease progression worsens or treatment regimen proves effective, including providing confidence metrics that suggest the collection of additional data, if necessary. *(See Section 11.5.1.2)*

- Time-to-Event Modeling: With longitudinal sensing, methods of survival analysis that adapt to time-varying would allow for updated risk estimates for adverse events both in terms of likelihood of event as well as in estimating the likely time to that event occurrence. *(See Section 11.5.1.3)*

- Multi-task learning: Deep learning techniques are well-suited to estimate risks of multiple, potentially varying adverse events, leveraging the commonality in risk factors associated with the primary adverse events or secondary recurrent events related to the different cardiovascular disorders. *(See Section 11.5.1.4)*

- Interpretable machine learning: as the data size progresses, medical models must be able to explain the driving risk factors in a manner interpretable to clinicians in order to guide treatment decision making. *(See Section 11.5.1.5)*

11.5.1 Existing Technologies and Applications

11.5.1.1 Risk Prediction Models

Much of stroke risk prediction is tied to the risk associated with AFib. In particular, it may be appropriate for patients with AFib to undergo anticoagulation therapy in order to reduce their risk of stroke. Anticoagulation therapy is any therapy that works to reduce the rate at which blood clots form. This type of therapy can be beneficial by preventing thromboembolic stroke. Conversely, this type of therapy can be detrimental by promoting life-threatening bleeds, such as in hemorrhagic stroke. Therefore, implementation of any anticoagulation therapy must be implemented with great care. In addition to models that

predict only stroke risk, the ACC/AHA Pooled Cohort Equations treat stroke and CAD together.

CHA2DS2-VASc is a model that predicts 12-month thromboembolic event rate (including stroke, pulmonary embolism, and peripheral thromboembolism) in patients with AFib who are not undergoing anticoagulation therapy [413]. Creation of this model drew upon the efforts of and improved upon multiple older models in order to apply more broadly and accurately to diverse patient populations. One chief exclusion in this model is that only patients with non-valvular AFib are considered. The parameters considered in this model are presence of HF, HTN, age, diabetic status, history of stroke or other thromboembolic event, history of any vascular disease, and gender. This model aids clinicians in prescribing anticoagulants, which increase the risk of bleeds but decrease the risk of thromboembolic events (including stroke).

The HAS-BLED model was created to predict the risk of bleeding in anticoagulated patients with AFib [414]. The parameters included in this model are HTN, history of liver or kidney dysfunction, history of stroke, history of bleeding, difficulty calibrating oral anticoagulation therapy, use of alcohol, and use of certain drugs that may increase bleeding risk. Recommendations by groups such as the European Society of Cardiology [415] are that CHA2DS2-VASc and HAS-BLED be used in conjunction for informed decision making, and that HAS-BLED alone should not be a reason to withhold anticoagulant therapy.

Other models have been produced to predict general risk of stroke. The MyRisk_Stroke Calculator is a model to predict 10-year risk of stroke [416]. This estimator was built on a prospective dataset where collection began in 1992 and validated with a second dataset with collection beginning in 1998. Follow-up was through the year 2007. In this cohort, the parameters found with an association to stroke risk were age, gender, education status, high blood pressure, smoking status, alcohol consumption, activity levels, anger, depression, and anxiety. Additionally, comorbidities such as renal disease, diabetic status, HF status, CAD, peripheral arterial disease were included as features in this model. The model was created

as a Cox proportional-hazards model and predicts 10-year risk of any type of stroke.

Another stroke risk model is the QStroke score [417]. QStroke was developed to be used for all patients without history of stroke but intended specifically to be used as a supplement or replacement for CHA2D2-VASc in predicting risk associated with AFib. The QStroke model features the following as parameters: age, gender, ethnicity, Townsend deprivation index (an index related to socioeconomic status), smoking status, body mass index, systolic blood pressure, blood lipid levels, and family history of CAD. HTN, diabetic status, AFib status, HF status, CAD, presence of rheumatoid arthritis, renal disease, and valvular disease were also included as pertinent comorbidities. The QStroke model was created as a Cox proportional-hazards model and predicts 10-year risk of any type of stroke.

Many attempts have been made to assess the risks of developing CAD. Among the most current of these are from the ACC/AHA Task Force on Practice Guideline [269]. That work introduced a set of models termed the Pooled Cohort Equation to predict a primary CAD event within 10 years. The predicted risk in this model is based on age, gender, race, blood pressure (systolic and diastolic), diabetic status, smoking status, various cholesterol lab values, and on certain current medications (HTN control, statins, or aspirin). This risk prediction tool was built to predict any type of "hard" atherosclerotic-based disease, and therefore in addition to predicting future CAD is also predicts future stroke. However, it does not distinguish between risks for these two different outcomes and treats them both as a positive outcome.

11.5.1.2 Remote and Dynamic Models

Remote and telemonitoring studies that use telephones and call-centers as the primary source of data have been used to track HF patients, in the hope of reducing heart failure admissions. These systems are intended to track patient symptoms, including impact of medication, weight gain (as a surrogate for edema), and depression, to identify early signs of decompensation aimed at providing interventions that prevent hospital readmissions in HF patients. In Tele-HF, Krumholz et al. found that a self-report telemonitoring system was not

able to reduce readmissions in heart failure patients based upon daily reports of symptoms, medication usage, weight, and depression [243]. Ong et al., in the Beat-HF trial, looked to automate some of the data collection surrounding blood pressure and weight, with machine learning risk models to drive interventions, but found similarly that HF readmissions were not reduced [244]. Anker et al. surveyed meta-analyses and prospective clinical trials that evaluated the efficacy of telemonitoring in patients with HF [418]. They found disagreement between the efficacy of telemonitoring for HF in different types of trials, but stress that the outcomes of telemedicine depend on personalization to the particular patient.

Models that predict risks within varied windows of time have, thus far, been restricted to medical settings. Henry et al. used a rolling model to predict the risk of sepsis in a hospital setting, selecting important features and identifying dynamic risks of sepsis within a single hospital admission [419]. Such dynamic models could be adapted to remote and longitudinal settings, but have not done so yet.

Few models exist, however, that estimate clinical risks of adverse events using remote and sensible data. Cakmak et al. used a smartphone to estimate answers to the Kansas City Cardiomyopathy Questionnaire, which aims to rate health status and severity of symptoms of conditions such as HF [420]. These personalized models that use remote data to estimate clinically-validated instruments used in current clinical models present a significant gap and opportunity for the systems discussed here.

11.5.1.3 Deep Time-to-Event

Survival analysis is an important domain of clinical modeling where data provided to a model estimates the likelihood of an event occurring as a function of time and the measured risk factors. This provides both an estimate of the likelihood of an event occurring but when it will occur, providing better longitudinal analysis of risk. The primary method for this technique has been a Cox Proportional Hazards model, which estimates the likelihood of survival over time with an underlying logistic regression model, which is linear in nature. Recently, deep learning techniques have improved upon the fit of these estimates over time by

allowing for complex, non-linear interactions [421]. Similarly, work by Lee et al. estimate the directly survival using deep neural networks [422], and continued improving upon this work by allowing for a range of dynamic covariates prior to the point of estimation rather than just the last value available for each time series [423]. Lee, Chen, and Ishwaran proposed a new model for survival analysis based upon adaptive boosting, which allows for time-varying covariates that can change at different points in time, allowing for flexibility in captured time-series signals and the features used from them to estimate risk [424]. Ultimately, these models will need to be adapted to the use of remote sensing data to track long-term risk of events based upon daily captured data for individuals wearing remote sensing systems.

11.5.1.4 Multi-task learning and Attention

The use of multiple sensing signals to track common risk factors over an array of differing cardiovascular disorders requires models that are robust to estimating multiple outcomes. Deep learning techniques are well-suited for this multi-task learning, having already demonstrated model performance superiority in clinical settings, such as estimating outcomes of patients in an intensive care unit [425]. Similarly, multi-task learning has proven to have superior performance using clinical time-series data [350] and in estimating in-hospital mortality [426]. The primary principle behind the multi-task learning environment is modifying loss functions to account for how accurate the model is estimating a set of outcomes rather than an individual prediction, with the assumption that the features being extracted from the data sources can estimate risk of each outcome, due to their dependent nature. This presents an opportunity to adapt these models to estimate risks from both in-hospital and remote, sensing data.

While these models are primarily built with CNNs, RNNs, and LSTMs, two key challenges arise when using these techniques: Limiting the input space to signals of the same length sampled at the same duration and Interpretation of their findings. While RNNs and LSTMs can handle varying-length sequences better than CNNs, with padding and masking techniques, they still sample time-series data across multiple channels at fixed time inter-

vals. With larger sequences of varying types of data, models that can adapt to different lengths, such as those used in natural language processing techniques, are needed, namely, transformers. The transformer architecture presents opportunities to further enhance model performance of time-series signals by allowing for additional flexibility to point the deep learning model at which portions of signals to focus on [427] that may not be properly aligned [428]. This modification to the more standard CNN, RNN, and LSTM architecture allows for more accurate modeling of clinical time-series data by leveraging when signals change and when they are invariant [429], and by adapting time-warping techniques [430]. Transformers have an attention property that allow it to attune to specific regions of data. This attention allows for more accurate clinical risk prediction models using time-series data [408]. The attention mechanism also has a property of providing a level of interpretability to deep learning techniques by identifying portions of signals that are deemed more important for feature extraction and model training.

11.5.1.5 Interpretable Machine Learning

A recent push in the machine learning field has been to explain predictions provided by deep learning methods that are generally considered black box techniques. Ribiero et al. developed a technique by which local logistic regression models are able to identify the reasons a particular prediction is made based upon the variables that generated the prediction of that element and similar model elements [431]. This work demonstrates model interpretability, which comes naturally in CNN deep learning models that can visualize data in intermediate models but becomes much more complicated in time-series based models such as LSTMs. Work by Lundberg and Lee looked to develop personalized levels of interpretation that are model agnostic, demonstrating a feature distribution and visualization technique that shows how certain actors matter for each user in a model and how that impacts the overall model performance [208]. Additional interpretation of models for how personal factors impact estimations provide personalization of interpretation Additional machine learning techniques look to automatically cluster patients and explain the phenotype discovery [432], while also

learning to predict multiple outcomes at the same time across different patient types [425], but work on explaining the findings remains in preliminary stages [433].

Interpretability also indicates the confidence in estimations, and understanding what data helps and hurts the predictive accuracy of techniques. In work aimed at improving real-time context and activity detection, Ardywibowo et al. evaluated selected sensors to improve HAR with constraints on the types of sensors and the power those sensors consume [396]. Work by Solis et al. use the idea of uncertainty quantification in order to direct users to gather more data in real-time, diet logging settings [434]. Uncertainty quantification is an emerging field of interpretable machine learning that has the ability to guide confidence in predictions collected as well as suggest additional data that patients and clinicians should consider collecting.

11.5.2 Gaps

Existing models for predicting risk in cardiovascular conditions rely on sparse data that are measured on rare occasions. Many parameters are trivial to measure (age, gender), and many parameters are Boolean values relating to history. In comparison to the data produced continuous monitoring systems, these data are sparse and likely overly-simplistic. There are two chief ways in which the limitations posed by this sparsity of data can be overcome with richer data: existing models can be updated to include richer data sources, and richer data sources can be analyzed for anomaly detection and rare event detection.

The following gaps remain in developing personalized analytic models based upon the remote sensing data gathered:

- Integration of sensing data with acute care data and outcomes for robust risk prediction models.
- The clinical models, to date, do not use complex remote, ambulatory sensing data. The initial development of models that leverage this data is needed.
- Learning key features for predicting adverse events from longitudinal capture without

the presence of ground truth data remains a challenge. Understanding how to extract features that work in accurate risk prediction models using remote sensing data and providing confidence to clinicians on these findings will require collecting vast amounts of data on user's to track for the potential of clinical events. This means integrating the sensing systems with electronic health record models that have the ground truth diagnosis and treatment information for such events.

- Development of dynamic models that are flexible to the types of data collected, the windows of data collected, and the changing in patient condition throughout observation.

- Deep learning-based time-to-event models either do not update when covariates change over time, fixing a longitudinal prediction with data set at a certain point in time, or update model estimations at fixed time-grid intervals. A time-to-event model that is able to adapt to time-varying covariates as they are captured, such as those from the sensing systems described in this work, is needed to update risk estimation.

- Interpretable machine learning to explain the predictions of these complex models, and help guide clinical decision making, including identifying similar patients and explaining potentially new phenotypes that might be discovered.

11.5.3 Opportunities

Existing models to quantify disease state and future disease risk could be improved through the implementation of richer data sources into the mode. The NYHA Functional Classification of HF relies in part on physical activity levels. The levels are subjective, with definitions in part of "no limitation of physical activity" (class I), "slight limitation of physical activity" (class II), "marked limitation of physical activity" (class III), and "unable to carry on any physical activity" (class IV). These classes are inherently subjective, and therefore susceptible to variability between patients with the same underlying disease state. Augmenting this classification with patterns detected from signals such as HAR and effort involved

in activity (such as via heartrate monitoring) would allow for objective measurements from beyond the limited scope of direct patient-physician contact. The increase of objective measurements would likely lead to updates to existing models and better information to aid in making clinical decisions.

Existing models could also be improved by the detection of rare or uncommon events. For instance, when a patient presents with AFib, the duration of the AFib is typically unknown. As discussed above, the CHA2DS2-VASc score can aid physicians in predicting stroke and the appropriateness of implementing oral anticoagulation therapy. However, the parameters which contribute to the CHA2DS2-VASc score are simplistically sparse. Age and gender are (for cardiac risk purposes) nonmodifiable risk factors. Each of the other parameters are positive if the patient has ever had the given event once in their life: HF, HTN, stroke/TIA/thromboembolism, vascular disease, or diabetes. It stands to reason that this model may be improved from richer data, such as the pattern or frequency with which the patient experiences episodes of AFib. Addition of this richer data to the model could potentially result in a model which is better able to discriminate between those at risk of stroke and those at lower risk of stroke, allowing for more appropriate and judicious use of oral anticoagulants.

The emergence of new sensing and internet of things (IoT) technologies creates a need for new models to incorporate new data for better prediction and understanding of disease states. The drastic increase in technology such as smartphones and smartwatches allows for new rich data sources, and also creates a need for the utilization of these data sources. Recently, smartwatches have been adapted to detect conditions such as AFib [435]. Further work should look at implementing these new modalities into longitudinal risk models. For instance, HAR recognition could be implemented as a parameter in monitoring activity tolerance in patients with HF. This could supplement the existing subjective measures of heart failure with newer objective measures.

Ultimately, data from new rich data sources is only valuable so far as it contributes to

improving the quality of patient healthcare. In order for this contribution to take place, models must generate actionable feedback that can be used for informed clinical decision making. Rather than presenting modeling through a black box approach where data is supplied to the model and an answer is returned, it is desirable that the reasoning behind the risk score is understandable. If a model is interpretable, then the factors leading to a given score can be understood and interventions made to address the risk and to improve patient outcomes. Additionally, the greater the interpretability of a model, the more information that the physician and patient are able to have about the overall disease state. As this information is understood by the physician and the patient, it can be used to better inform and guide care. As a result, the following opportunities exist for immediate and impactful machine learning research:

- Machine learning models with cross-sectional and time-series data: Integration of sensing data with acute care data and outcomes for robust risk prediction models.

- Development of dynamic models that are flexible to the types of data collected, the windows of data collected, and the changing in patient condition throughout observation.

- Interpretable machine learning to explain the predictions of these complex models, and help guide clinical decision making, including identifying similar patients and explaining potentially new phenotypes that might be discovered.

- Transfer learning: transfer learning techniques will be able to take developed models and adapt to a variety of signals captured, a variety of patients modeled, or a combination therein, improving the flexibility of any analytic techniques developed to advance the prior three opportunities.

- Deep learning: adaptation of deep learning techniques that have proven successful in natural language processing tasks and computer vision tasks to time-series modeling

based upon the remote sensors discussed in this work provide an opportunity to develop new transformer and attention-based models that are adaptable to various signals of different domains and lengths.

- Adaptive time-to-event models: As deep time-to-event models improve the estimation of risk longitudinally, developing dynamic models that adapt the model structure to data that is newly available through different sensors is needed that account not only for changes in the values of the modeled covariates, but that can adapt to new time-series signals as they become available.

11.6 Discussion and Conclusion

We surveyed the field of sensing technologies and machine learning analytics that exist in the field of remote monitoring for the tracking of risk factors that lead to primary adverse events and secondary recurrent events associated with cardiovascular disorders. Through the evaluation of these sensing modalities and machine learning techniques, we highlighted the potential for addressing three critical areas of need for care in patients monitoring risk factors associated with heart failure, coronary artery disease (and myocardial infarction), and stroke: 1) need for sensing technologies that track longitudinal trends of the cardiovascular disorder despite infrequent, noisy, or missing data measurements; 2) need for new analytic techniques designed in a longitudinal, continual fashion to aid in the development of new risk prediction techniques and in tracking disease progression; and 3) need for personalized and interpretable machine learning techniques, allowing for advancements in clinical decision making. We highlight these needs based upon the current state-of-the-art in smart health technologies and analytics and discuss the ample opportunities that exist in addressing all three needs in the development of smart health technologies and machine learning (primarily deep learning) approaches applied to the field of cardiovascular disorders and care. Whereas the progression of smart health technologies in these needs has demonstrated success in fields such as HAR and physical disorder monitoring, the opportunities for addressing cardiovascular care are

many.

These cardiovascular disorders are often very complex conditions characterized by multiple changes in a patient, many of which are slow and difficult to notice. However, systems could be built to take into account and monitor many different changes in order to track risk factor progress for disease state monitoring and to allow clinical decisions to be made before rapid decompensations. As a disease progresses, regular monitoring of heart sounds could be used in order to track heart remodeling. Instead of noticing these sounds in an acute care visit, computer-aided auscultation though wearable electronic stethoscopes could allow for earlier detection. Quantitative edema tracking would allow for monitoring functional changes within the heart. As pulmonary edema increases, clinicians are able to tell that left heart function is decreasing. Changes in ECG signals may indicate progression of CAD disorders, that may result in additional patient pain, prior to leading to heart attack. As a patient's condition worsens, they may gradually lose the stamina to walk certain distances or to perform a certain amount of activity, demonstrating changes in physical activity capacity, respiratory rate, or sleep quality. These changes may be so gradual that patients may not notice them. Instead, new analytics for progression could instead build activity recognition into the modeling to understand slow changes in baseline function. In this, departures from a patient's baseline level of activity would be significant and could be useful information for guiding clinical care. Similarly, alterations in blood flow may lead to changes in urination habits. As blood flow to the kidneys might be restricted during the day and increased at night due to postural changes, the kidneys will produce more urine at night. Tracking frequency of nocturnal urination could provide more clues as to overall health. Note that this will be sensitive, but not specific for heart failure. Additional measurement of hemodynamic characteristics, such as HTN, may show treatment effectiveness and better guide the improvement of factors that would lead to conditions such as stroke. In total, a comprehensive system to track the progression of cardiovascular disorders should incorporate a body of integrated sensors, capture this data over longitudinal periods of time, and as a result, enable

new advancements in machine learning techniques that can make best use of this data to help guide patient and clinician alike in improving patient care through personalized, dynamic time-to-event modeling.

This survey highlighted the needs in developing smart health applications to treat HF, CAD, and stroke, and the risk factors associated with them. It reviewed the existing technologies, highlighting the current gaps in solutions presented for those needs. Finally, it presented a series of opportunities, including advanced analytic techniques to be developed once new sensing solutions are available that can guide impactful changes in the way patients with cardiovascular disorders are cared for.

Table 11.1: Abbreviations and Definitions of key clinical terms

Abbreviation	Clinical Term	Definition
AFib	Atrial Fibrillation	Cardiac arrythmia that is highly associated with risk of stroke
AR	Aortic Regurgitation	Disease state where the aortic valve fails to completely close during diastole, allowing for backwards flow of blood into the heart
AS	Aortic Stenosis	Disease state where the aortic valve encounters resistance during opening, requiring additional force from the heart to drive blood forward
BP	Blood Pressure	The pressure present in the arterial circulatory system. Blood pressure oscillates from a peak value (systolic blood pressure) to a trough value (diastolic blood pressure) as the heart beats
CAD	Coronary Artery Disease	Disease state characterized by impaired blood flow in the small arteries around the heart (the coronary arteries)
ECG	Electrocardiography/Electrocardiogram	Tracing that shows a measurement of cardiac electrical activity
HAR	Human Activity Recognition	Utilizing sensors to classify a patient's physical activities
HF	Heart Failure	Disease state characterized by an impaired ability of the heart to drive blood forward
HFpEF	HF with preserved ejection fraction	Heart failure state characterized by a normal ejection fraction, often related to defects in ventricular filling during diastole
HFrEF	HF with reduced ejection fraction	Heart failure state characterized by a decreased ejection fraction, often related to impaired ventricular contractility or to pressure overload
HTN	Hypertension	A health condition characterized by chronically increased blood pressure that puts patients at risk of heart disease
(A)MI	(Acute) Myocardial Infarction	Disease state characterized by impaired coronary blood flow leading to some degree of cardiac muscle death. Commonly known as a "heart attack."
MR	Mitral Regurgitation	Disease state where the mitral valve fails to completely close during systole, allowing for backwards flow of blood within the heart
MRI	Magnetic Resonance Imaging	Imaging technique using powerful magnets to image internal structures
NSTEMI	Non-ST-elevation MI	MI characterized by a lack of ST segment elevation on ECG
NYHA	New York Heart Association	Entity that issues guidelines for heart failure classification
PMI	Point of Maximal Impulse	Point on a patient's chest where the movement of the patient's heart can most strongly be felt.
S3	Third heart sound	A heart sound occurring after the normal second heart sound. It is benign in younger patients, but indicative of pathology in older patients
S4	Fourth heart sound	A heart sound occurring prior to the normal first heart sound. It is indicative of pathology
SA	Stable Angina	Chest pain that occurs after a certain amount of exertion caused by restricted blood flow through the coronary arteries
STEMI	ST-elevation MI	MI characterized by an elevation of the ST segment on ECG
TEE	Transesophageal echocardiography	Imaging technique that uses ultrasonography and a probe in the esophagus to evaluate the function of the heart
TIA	Transient Ischemic Attack	Temporary restriction of blood flow to a part of the brain. Similar to a stroke, but resolves within minutes.
TTE	Transthoracic echocardiography	Imaging technique that uses ultrasonography and a probe on the chest to evaluate the function of the heart
UA	Unstable Angina	Chest pain that occurs at rest caused by restricted blood flow through the coronary arteries

Table 11.2: Sample of current commercially-available devices and common cardiovascular parameter monitoring

Product Class	Product	(Optical) Heart Rate	Blood Pressure	Other Measures
Smartwatch	Amazfit Verge[271]	✓	×	ECG
	Apple Watch[272]	✓	×	ECG
	Empatica Watch[273]	✓	×	Galvanic Skin Response & Temperature
	Fitbit[274]	✓	×	
	Garmin Fenix Watch[275]	✓	×	PulseOx & Temperature
	Samsung Galaxy Watch[276]	✓	✓	
	Valencell-associated smartwatches[277]	✓	×	
	Withings (Move ECG, ScanWatch, Steel HR, Pulse HR)[278]	✓	×	PulseOx, ECG
Smart Ring	Oura (Ring)[279]	✓	×	
Headphones	Samsung Galaxy Buds[276]	✓	×	
	Valencell Blood Pressure Kit[280]	✓	✓	
	Valencell-associated earbuds[277]	✓	×	
Chest Strap	Garmin HRM-Series Chest Strap[281]	×	×	ECG
	Polar H-Series Chest Strap[282]	×	×	ECG
	QardioCore Chest Strap[283]	×	×	ECG
Medical Devices	AliveCor ECG[284]	×	×	ECG
	Caretaker[285]	✓	✓	PulseOx
	Finapres Nova[286]	✓	✓	PulseOx
	QardioArm[287]	×	✓	

Table 11.3: Summary of sensing types, analytic possibilities, and the advantages and disadvantages of the technologies

Technology	Sensor	Analytics	Advantages	Disadvantages
Digital stethoscope[288, 292]	Microphone and appropriate casing	Automated segmentation could be applied to segment beats and analyze characteristics of heart sounds	Relatively cheap and accessible technologies	Commercially available models require physician evaluation
Radar vital sign measuring[294, 259]	Radar receiver to monitor patient physiologies	Analyze movement to extract heartrate and respiratory rate	No contact required	Requires knowing patient posture, Limited to a single patient at a time
Electrocardiography[300, 301, 304]	Measurement of cardiac electrical impulses	Automated segmentation required for analysis, pattern recognition required for detection of changes such as arrythmias or ST elevations	Gold standard in cardiac monitoring	Requires lead placement, Susceptible to noise, Difficult to interpret
Cuff-Based BP Monitor[306]	Pressure impulse or sound during cuff deflation	Automated analysis of cuff pressure allows for extraction of systolic and diastolic BPs	Perform well at providing accurate measures of patient BP, Easy to use	Unable to provide continuous blood pressure, Obtrusive
PPG-Based BP Monitor[309, 310]	PPG signal in coordination with ECG signal	Extracts blood pressure by relationship of electrical signal and pulse signal arrivals.	Could be applied to continous monitoring	Requires multiple signals, Poor accuracy due to variable pre-ejection period timing
Bioimpedance-based BP Monitor[321]	Bioimpedance signal	Extracts blood pressure by relationship of bioimpedance along arteries.	Could be applied to continous monitoring	Not commercially available
Ultrasonography[327, 328]	Ultrasound generator sends pulses into patient and analyzes rebounding signals to construct an image	Image analysis required for flow assessment and vessel characterization	Allows direct visualization of blood vessel, allows for real time visualization	Not available in wearable form, Noisy image, quality dependent on skill with device, Requires expert evaluation
Smart Sock [331, 332]	Accelerometer and stretch sensor	Estimates leg edema using learned mapping of inputs	Allows for remote monitoring of extremity edema	Not commercially available
Continuous Glucose Monitor[343]	Subdermal electrode to sample interstitial fluid	Allows for estimation of food intake and for personalized treatment of glucose intolerance	Captures a difficult signal that is closely related to long term health issues	Invasive, Unlikely to reach universal adaptation, Needs to be replaced after 14 days

12. ESTIMATING BEAT-TO-BEAT CUFFLESS BLOOD PRESSURE WITH NEURAL ARCHITECTURE SEARCH

This chapter now examines a particular application of machine learning to remote sensors. As detailed in Chapter 11, smart watches are able to provide a wide range of monitoring. However, there does not yet exist a commercially available watch-based cuffless blood pressure monitoring system. This chapter details analytic approaches for estimating blood pressure from a novel bioimpedance wrist-worn device.

12.1 Introduction

Hypertension, a disease marked by chronically elevated blood pressure, affects one in four adults worldwide, causes an estimated 7.6 million deaths per year, and results in the loss of 92 million disability-adjusted life-years [436]. A cornerstone of lowering cardiovascular risk is the management of hypertension [265, 437, 438, 439, 440], as blood pressure is the most modifiable physiological marker that drives risk for cardiovascular disorders [265]. However, evidence is abundant that measurement of blood pressure outside a clinical visit provides better prognostic information [265], including the identification of masked hypertension [441], white coat hypertension [442], or nocturnal non-dipping hypertension [443]. Studies estimate 10% of the population is unaware of their added risks in these cases [444, 445, 446, 447, 448, 449], supporting the need for ambulatory measurements taken more frequently for each patient. While ambulatory blood pressure monitoring devices better capture blood pressure measurements that indicate hypertension, these devices present a number of serious shortcomings that prevent widespread adoption and use, for example, startling or even awakening asleep subjects and leading to a misdiagnosis of a 'non-dipper' [450]. As a result, studies are hard to repeat and typically limited to measuring ambulatory blood pressure in a single 24-hour period [451, 452, 453]. Studies [454, 455] have shown that additional measurements over a single 48-hour period or two 24-hour measurement periods

provide significant additional prognostic value. Therefore, a solution is needed that allows for continuous and unobtrusive monitoring. More specifically, a cuffless device that enables readings without subject disturbance is needed.

Cuffless blood pressure monitoring devices measure surrogates of blood pressure and utilize regression modeling techniques in order to provide diastolic and systolic blood pressure readings from these surrogates. Devices that use photoplethysmography (PPG) and electrocardiography (ECG) [456], or bioimpedance [457, 458] have recently been developed to measure the pulse transit time (PTT) or pulse wave velocity (PWV), which are known surrogates for blood pressure [459, 321, 460]. These work focus on collecting physical signals and formulating the relationship to the blood pressure, but lack algorithms for further modeling and analysis. Recent work by Ibrahim and Jafari [321] demonstrated that their bioimpedance-based sensor better located arterial sites from which to measure these physiologic surrogates of blood pressure. Using an AdaBoost regression technique and a series of maneuvers on feature selection, they built a window-based personal model that measures diastolic and systolic blood pressure to with respective errors of 2.6 mmHg and 3.4 mmHg, within the ISO standard that requires errors less than 10 mmHg [461]. However, the averaging calculation of features and labels lost many details, and is not a straightforward map from the physical signals to blood pressure. Additionally, this tree-based algorithm ignores the relationship between diastolic and systolic blood pressure, and relies on personal feature extraction thus does not fit future exploration, e.g. a generalized model.

As the dataset from Ibrahim and Jafari [321] is made available to us, we seek to examine this model and dataset, with the intention of developing regression models that provide an extension of a beat-to-beat regression model that provides estimates of blood pressure for each heartbeat, rather than average value over time. Since blood pressure is a function of cardiac output and arterial compliance, a change in arterial compliance impacts how much the PTT varies when blood pressure varies [462, 463]. The beat-to-beat measurements of blood flow relate to both diastolic and systolic blood pressure, but in a non-linear relationship. For

instance, high systolic blood pressure can be associated with a correspondingly high diastolic blood pressure but might also be associated with an unchanged diastolic blood pressure. Therefore, we develop a multitask learning (MTL) network to estimate both diastolic and systolic blood pressure, and hypothesize that an MTL framework that first measures the changes in bioimpedance signals and then uses task-specific layers to correlate those changes to diastolic and systolic blood pressure will be more accurate than developing task-specific models. Additionally, while work by Ibrahim and Jafari [321] identified a number of manual features that correlate PTT and specific bioimpedance features to blood pressure, we seek to both explore a neural architecture that is able to automatically learn features of importance as well as to use a neural architecture search (NAS) technique to identify the appropriate model structure to optimize this task. We train a controller network that produces an optimal number of layers and network depth without requiring background knowledge on the type of data and depth of network necessary to tune in a manual grid-searched fashion.

Generalizable Insights about Machine Learning in the Context of Healthcare

- We implement models able to produce patient blood pressure by noninvasive means from wearable sensor signals.

- The design of the MTL model allows for top-level healthcare knowledge to guide machine learning exploration. Model design encompassing this top-level domain knowledge allows for better models that are able to realize complex relationships present in the underlying physiological processes.

- Application of the NAS algorithm into the physiologically-motivated MTL network allows for further design space exploration and improved performance at the extremes of prediction error. This integration improves upon variation present in underlying physiological assumptions.

12.2 Related Work

12.2.1 Cuffless Blood Pressure

There is substantial interest in developing cuffless blood pressure monitoring devices. One common approach to develop such a device is through the simultaneous use of ECG and PPG signals. Kachuee et al. [464] describe an algorithm for extracting blood pressure from ECG and PPG using both manual feature extraction techniques and vector-based matching on shape and timing, but cannot meet the ISO standard. Zheng et al. [460] used an armband equipped with ECG and PPG to measure PTT by manually deriving the PTT using R-R intervals and manually discarding signals with motion artifacts. Linear and non-linear models based upon Moens-Korteweg were applied and obtained a root mean squared error (RMSE) of 8.7 mmHg for systolic blood pressure over a 24-hour monitoring period. Approaches that use ECG and PPG suffer inaccuracies resulting from the pre-ejection period, which constitutes a time delay between the electrical stimulation of the heart and the actual mechanical expulsion of the blood for each heartbeat [312]. One approach to overcoming the limitations of PPG is to use a dual PPG system. Wang et al. [465] proposed utilizing one PPG on the forearm and one on the wrist coupled with an accelerometer to build a system that can measure PWV while remaining robust to motion artifacts. Nabeel et al. [466] similarly used a dual-PPG system and were able to estimate diastolic blood pressure with an RMSE of 5.26 mmHg, but reported greater difficulties in measuring systolic blood pressure. In addition to PPG and ECG, Chandrasekaran et al. [467] proposed a novel method of using the built-in camera and microphone from two smartphones to estimate blood pressure. The authors used a physical model relating the transit time to the blood pressure, and report accuracies of 90% when implementing their regression models. This smartphone-based cuffless blood pressure estimation method has a high error margin of 6 mmHg, and requires subjects to touch the phone, and therefore is not able to capture nocturnal blood pressure. A chief limitation in all of these works was often the manual extraction of key features with pre-designed models

for linear or non-linear regressions.

Ibrahim and Jafari [321] developed a wrist-worn device to capture bioimpedance signals for continuous blood pressure estimation. By evaluating measurements along the wrist, this method does not suffer from timing errors as a result of the pre-ejection period. To evaluate the method, they recruited ten healthy subjects aged between 18 and 30 years old and collected data from them using the wrist-worn sensors as well as the reference using the Finapres NOVA system. Data was collected after exercising in order to induce physiologic changes in subject blood pressure. The Finapres NOVA device measures the reference diastolic and systolic blood pressure for each heartbeat. This reference beat-by-beat blood pressure is related to the bioimpedance curves measured by the four pairs of sensors throughout the beat. There are 50 features extracted from the bioimpedance signals representing diastolic peak, maximum slope, systolic foot, and inflection point, representing PTT and personal vessel elasticity information. Finally, the extracted features are averaged over 10-beat windows and applied to two separate AdaBoost regression models to estimate window-based diastolic and systolic individually. The results show an average RMSE and correlation coefficient (R) of 2.6 mmHg and 0.77 for the diastolic blood pressure and 3.4 mmHg and 0.86 for the systolic blood pressure.

12.2.2 Neural Architecture Search

Neural Architecture Search (NAS) is an approach for optimizing network architecture to improve performance and has been successfully applied in multiple domains. Chen et al. [468] describe a NAS applied to building shared-hierarchical MTL models for natural language processing tasks. They utilize NAS to determine an optimal way in which to share modules between tasks, allowing similar tasks to share additional modules and dissimilar tasks to share fewer modules. Several studies have shown benefits from utilizing NAS to optimize models in medical domains, such as Weng et al. [469] and Faes et al. [470]. Weng et al. [469] proposed a method for applying NAS on medical image segmentation. Fonseca et al. [471] applied NAS to build a convolutional neural net classifying tissue composition

in patients undergoing mammography, producing a model with performance comparable to several experienced radiologists. Balaprakash et al. [472] developed a reinforcement learning-based NAS to automate a finding an optimal deep learning-based model for predicting cancer using a class of non-representative data. Liu et al. [473] built a human activity recognition system from human motion information captured by distributed infrared sensors, and applied NAS to learn the best mask structure for the recognition task. These approaches all demonstrate applicability and utility of NAS methods to work in multiple domains to improve model architecture, and we employ such a method here.

12.3 Dataset and Data Preprocessing

12.3.1 Dataset

The dataset for this paper was collected from 11 subjects by a wrist bioimpedance sensor [321]. The sensor collects four channels at a sampling rate of 20 kHz and measures the impedance corresponding to blood volume changes within the the ulnar and radial arteries. The data is collected for each subject through a few workout trials with resting periods between. Workout serves to elevate blood pressure and ensure that a range of physiologic values are measured. Noise has been removed from the dataset and 13.47 ± 4.05 minutes of resting measurements remain for all subjects, except for one outlier with only has five minutes of data. When collecting data from the wrist bioimpedance sensor, the participants also wore the Finapres NOVA device to capture beat-to-beat diastolic and systolic blood pressure as the reference blood pressure. We use this dataset in our work.

12.3.2 Data Preprocessing

The raw bioimpedance signals are segmented into windows by heartbeat. The segmented beat-to-beat signals are fed into an LSTM for time-series feature extraction. The heartbeat sequence length of the dataset varies from 5 to 25 thousand, representing heart rates from 48-240 beats per minute. Samples with a sequence length over 20,000 are removed as the majority of these represent erroneously detected heartbeats with beat length far longer than

physiologically likely in the subject demographic. While LSTM models have proven to be functional in extracting features from time-series data, a reasonable limit of at most a few hundred timesteps is often used in practice. Therefore, all sequences are downsampled from 20 kHz to 100 Hz by equally sampling at a ratio of 200:1, giving each heartbeat at most 100 measurements. An illustrative figure and further details are provided in the appendix. We then zero pad the beginning of all samples to make their length 100 for shorter sequences. Although the LSTM layers are used as an alternative to manual feature extraction as described in Ibrahim and Jafari [321], we augment the four available signal channels with their first derivative to increase performance of the LSTM. This implicitly provides additional timing information akin to extracting the PTT or PWV, which are important factors in estimating blood pressure. Finally, we add the timing of each point in the sequence as an extra channel. The final input to the LSTM consists of these nine channels across 100 timesteps.

12.4 MTL for Personalized Blood Pressure Estimation

12.4.1 Model Development

The model development consisted of three key steps: 1) embedding time series data, 2) using a network to extract diastolic and systolic blood pressures from the shared LSTM output, and 3) optimal network architecture search. We used this configuration to first encode shared relationship information between the bioimpedance signals for each heartbeat with both diastolic and systolic blood pressure models, then developed task-specific layers. Both diastolic and systolic blood pressure are functions of cardiac output and arterial compliance but are always linearly related. For instance, variability in this difference is known to be related to the nocturnal dipping status of patients [474]. The MTL framework takes the features extracted though data embedding and uses two task-specific networks to correlate the same features to diastolic and systolic blood pressures separately.

A structure of our general model is represented in figure 12.1. Our model uses a single

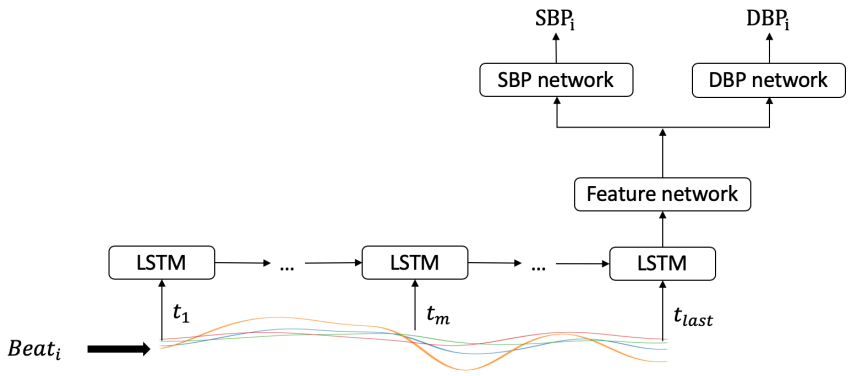


Figure 12.1: Overall network architecture. Each time point along a heartbeat is fed into the LSTM, and the final output of the LSTM is fed into the feature network for calculating blood pressures. As depicted here, the network is structured using an MTL approach following a shared layer. For the baseline single-task models, only one branch of the MTL (systolic (SBP) or diastolic (DBP) blood pressure) is present.

layer LSTM at each time step with 30 units to encode the beat-to-beat time-series signals. The LSTM first step is initialized with random values. LSTM allows for the memorizing of historical information, so we use LSTM to embed time-series signals in order to extract the changing of signals within each beat. A given state h in the LSTM includes the historical bioimpedance information of a beat sequence, and x is the input at each time point within a sequence. The output from the final LSTM cell, which memorizes the bioimpedance information of the entire beat, is utilized in predicting the blood pressure.

After the LSTM, we add a dropout layer with keep probability 0.7. The LSTM model extracts historical information from the time-series data. However, it is essential that the model be able to incorporate the relationship between adjacent signals, consisting of pairs of channels with data from the same artery. This relationship implicitly describes the movement of the arterial pulsation distally along the artery. Notably, the PTT and PWV, hallmarks of earlier work, can only be measured by observing the signal at two separate locations along the same artery, and knowing the distance between those locations. Therefore, we added a shared fully connected layer after the LSTM and dropout layer for feature extraction. The hidden size of this single layer is 30 nodes, which are delivered to the task-specific networks

for estimating diastolic and systolic blood pressures.

The two networks take the same features from the shared layer as inputs. Both networks contain a tunable number of fully connected layers with ReLU activation functions and tunable hidden sizes for each layer. We manually searched the number of layers from two to seven and let the hidden size vary from 20 to 100 as an initial exploration space to prevent overfitting. We observed that the results of blood pressure estimation decrease whenever the number of layers exceeds five or the size of a hidden layer exceeds 50, and then we decided to extensively search from two to five layers with 20 to 50 hidden nodes per layer. Initial exploration outside of this range lead to rapidly decreasing performance metrics. After the multiple task-specific layers, each network output is converted to a single value using a fully connected layer with a linear activation function to estimate the respective blood pressure. The optimal setting varies between subjects. However, three layers with hidden size 30 (prior to the output layer) works well for many of the subjects. When using a different number of layers between the two networks in a given subject, the accuracy suffers as the different network depths fail to converge given the limitations of back propagation. The shared loss function used for blood pressure estimation is:

$$L_{BP} = \sqrt{\sum_i (E_i^S - T_i^S)^2 + (E_i^D - T_i^D)^2} \tag{12.1}$$

where E^D and E^S represent the estimated diastolic and systolic pressures, respectively, and T^S and T^D are their target values respectively. This way, both diastolic and systolic blood pressures are estimated by using different networks but the same features, and when updating the model parameters they control both their own part and the shared network. A comparison between the performances of the MTL-based model and the individual estimation models is provided below.

Using an internal validation set from the training data, we manually searched the number of units in each LSTM cell from 20 to 50, the hidden size of the shared layer from 20 to

50, and the keep probability for the dropout layer from 0.5 to 1. Changing the LSTM cell unit count and the shared layer size did not contribute significantly to performance, while optimal performance was achieved by dropout of 0.7. However, when changing the number of layers and their hidden size for the task-specific layers in the MTL, the result changes for each subject and the best setting varies between subjects. As mentioned earlier, the results decrease whenever the number of task-specific layer is over 5 or the hidden size of them is over 50. Therefore, we hypothesized that a very big network over-computes the features, resulting in a model with excessive bias.

Our current dataset has 11 subjects in total, allowing for a manual search for personalized models. However, searching for optimal architecture and hyperparameters becomes infeasible when expanding to arbitrary future subjects. Therefore, we decided to search model architectures for the best performing architecture when applied to an arbitrary subject in our dataset. Grid searching is an easy approach to this problem, but it has two limitations: 1) it requires manually searching to obtain background knowledge and decide the search range, and 2) when the dataset becomes big, the model needs to run all the possible settings for all the subjects, which is a time-concerning work.

To better search the space of appropriate size and layers of the MTL network, we utilized NAS. NAS uses a recurrent network as a controller which produces an architecture for a child model. The controller network is updated with a reinforcement learning approach that treats the performance of the produced model as its reward. When producing the child model, the output of each step is used as a hyperparameter, and is delivered to the next step for the recurrent network as input. We fix the MTL with five layers and use the recurrent network with ten units as the controller that produces the hidden size for each hidden layer one at a time. The MTL layers are produced as follows:

$$p_i = C(u_{i-1}, p_{i-1}) \tag{12.2}$$

$$R_j = \frac{1}{L_{BP}^j} \tag{12.3}$$

where C represents the recurrent controller network. In each step i, it takes the state u_{i-1} and output p_{i-1} from the previous step $i-1$ to compute the output p_i at the current step. Here p_i refers to the produced hidden size for a given network at a certain MTL layer, and p_{i-1} indicates the hidden size of the opposite network from the previous layer. To update the controller in the reinforcement episode j, we calculate the reward R_j as above to encourage the controller network being trained to obtain lower blood pressure estimation loss L_{BP} as defined in equation 12.1. From grid searching, the best number of layers varies between 2 to 4, and the results decrease rapidly when the number of layers is over 5. Therefore, we searched through adding one layer beyond that when applying NAS in order to give NAS enough space to search and explore. The controller has 10 steps in total including 5 layers and 2 prediction at each layer. The controller network was implemented using the NAScell package in Tensorflow to produce the MTL model, and the produced child network is trained as naive MTL [475].

12.4.2 Experimental Results

We tested both individual task models and the MTL model for each subject. For both models, we used 10-fold cross validation to train each subject. In each fold, 80% of data was randomly picked as training set, 10% as validation set, and 10% as test set. The validation set and test set had no overlap among folds, and the models were re-initialized with random parameters at the beginning of each fold. After 10 folds of training and testing, we took the estimated diastolic and systolic values from all folds and returned to the original time-series order, then calculated RMSE and correlation for evaluation. We then improved over this baseline by utilizing NAS as described above. After finding optimal architecture using NAS to construct the task-specific layers for, the resulting MTL model was trained, validated, and tested in the same way.

Table 12.1 compares the baseline individual task models, our MTL model, and the MTL

model as optimized by NAS. With separate models, diastolic blood pressure estimation gets 3.43 mmHg average MSRE with 0.79 standard deviation and 0.73 average correlation. Systolic blood pressure has 5.25 mmHg average RMSE with 1.87 standard deviation and 0.75 average correlation. All subjects have RMSE within ISO standards for error of 85% of beats with under 10 mmHg absolute error.

Table 12.1: Mean ± Standard Deviation RMSE (mmHg) and R for individual task models, MTL model, and NAS-MTL model for beat-to-beat diastolic and systolic blood pressure estimation (DBP & SBP)

Model	DBP RMSE	SBP RMSE	DBP R	SBP R
Individual task	3.43 ± 0.79	5.25 ± 1.87	0.73 ± 0.10	0.75 ± 0.10
MTL	3.18 ± 0.50	4.53 ± 1.03	0.77 ± 0.09	0.80 ± 0.10
NAS-MTL	$\mathbf{2.91 \pm 0.47}$	$\mathbf{4.46 \pm 0.90}$	$\mathbf{0.89 \pm 0.01}$	$\mathbf{0.92 \pm 0.02}$

Table 12.2: Individual task model beat-to-beat performance per subject for diastolic and systolic blood pressure (DBP & SBP) RMSE (mmHg) and R.

Subject	DBP RMSE	SBP RMSE	DBP R	SBP R
1	2.98	4.08	0.81	0.75
2	3.58	4.65	0.83	0.91
3	2.95	5.64	0.74	0.62
4	3.26	5.21	0.61	0.76
5	2.94	2.66	0.84	0.89
6	4.99	6.61	0.61	0.70
7	3.95	6.00	0.68	0.74
8	4.69	9.88	0.58	0.62
9	2.85	4.17	0.72	0.67
10	2.90	4.46	0.72	0.71
11	2.61	4.39	0.84	0.87
Mean	3.43 ± 0.79	5.25 ± 1.87	0.73 ± 0.10	0.75 ± 0.10

Table 12.2 contains the beat-to-beat results from the baseline model while Table 12.3 shows the results from the MTL model. The baseline models in Table 12.2 are composed

Table 12.3: MTL beat-to-beat performance per subject for diastolic and systolic blood pressure (DBP & SBP) RMSE (mmHg) and R.

Subject	DBP RMSE	SBP RMSE	DBP R	SBP R
1	3.40	4.43	0.74	0.69
2	3.71	4.60	0.81	0.91
3	2.28	3.29	0.85	0.89
4	3.18	5.26	0.65	0.76
5	2.82	2.75	0.85	0.89
6	3.87	5.12	0.79	0.84
7	3.37	5.11	0.78	0.82
8	3.52	6.50	0.79	0.86
9	2.63	3.70	0.77	0.76
10	3.45	4.66	0.56	0.59
11	2.72	4.32	0.83	0.86
Mean	3.18 ± 0.50	4.53 ± 1.03	0.77 ± 0.09	0.80 ± 0.10

of individual separated diastolic or systolic prediction networks. We observe that the MTL model averages RMSE 3.18 mmHg with 0.50 standard deviation for diastolic blood pressure estimation and 4.53 mmHg with 1.03 standard deviation for systolic blood pressure. All the subjects have error with the ISO standard. At the same time, diastolic blood pressure averages 0.77 correlation and systolic blood pressure achieves 0.80, with standard deviation 0.09 and 0.10 respectively.

In order to compare with the previous work of Ibrahim and Jafari [321], we calculated the blood pressure as averaged over a 10-beat window. The previous work, two AdaBoost regression models with manually extracted features, estimated diastolic blood pressure with an RMSE of 2.6 mmHg and correlation of 0.77 and systolic blood pressure with an RMSE of 3.4 mmHg and correlation of 0.86. The MTL model obtained an RMSE of 1.57 mmHg and correlation of 0.93 for diastolic blood pressure and an RMSE of 2.31 mmHg and correlation of 0.94 for systolic blood pressure. **See appendix for more details**. As we primarily focus on beat-to-beat estimation in this paper the remaining results focus solely on beat-to-beat findings.

Figure 12.2 shows a randomly chosen example of estimated blood pressures and their tar-

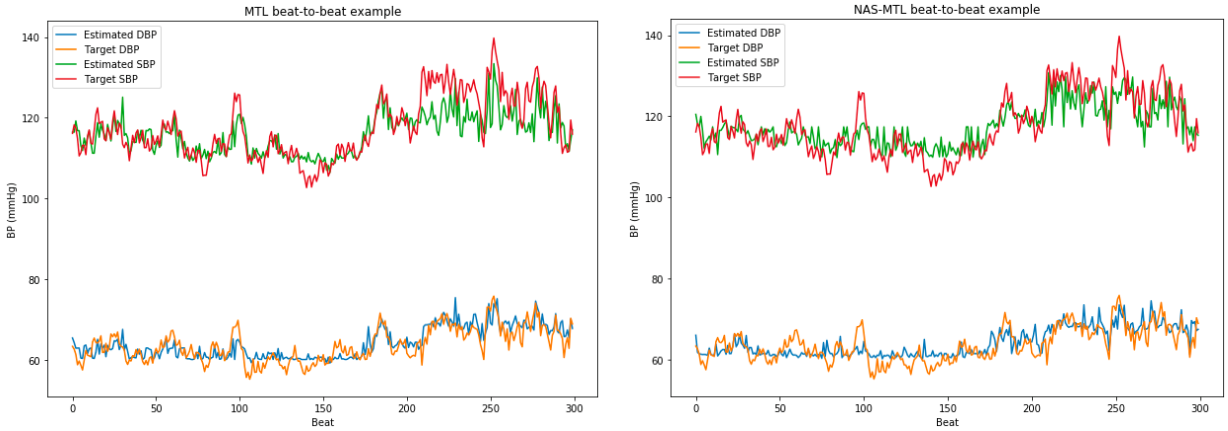


Figure 12.2: The estimated and target blood pressures for a randomly chosen subject as generated by the MTL model (left) and NAS-MTL model (right).

gets from a randomly chosen subject and is representative of all subjects. In these plots, the predicted blood pressures from the MTL and NAS-MTL accurately estimates the changing of blood pressure, even though it does not match every beat perfectly. The impact of Valsalva near beat 200 can be seen by a rise in systolic blood pressure and a corresponding but smaller rise in the diastolic blood pressure. We notice that the MTL model can accurately estimate peaks from targets, e.g. the peak at around the 100th beat and before the 200th. However, the estimated blood pressures are more stable comparing to the targets. The target blood pressures vary more but our estimated blood pressures are smoother. In comparison to the naive MTL, the estimated systolic blood pressure from NAS-MTL is closer to target values after the blood pressure increases near beat 200.

The Bland-Altman plot for beat-to-beat MTL predicted blood pressures for all subjects is shown in Figure 12.3. The top two figures show the estimated and reference (target) blood pressures. The lower two figures include the absolute error for all predictions. The lower figures show the 95% confidence intervals (CIs) for the limits of agreement. To compare with the ISO standard, we also calculated the 85th percentile of errors from the estimated blood pressure. In this model, 85% of predictions have errors within 4.05 mmHg diastolic and within 5.72 mmHg systolic. With this model 98.4% of predictions have diastolic error

less than 10 mmHg and 95.4% of predictions have systolic error less than 10 mmHg.

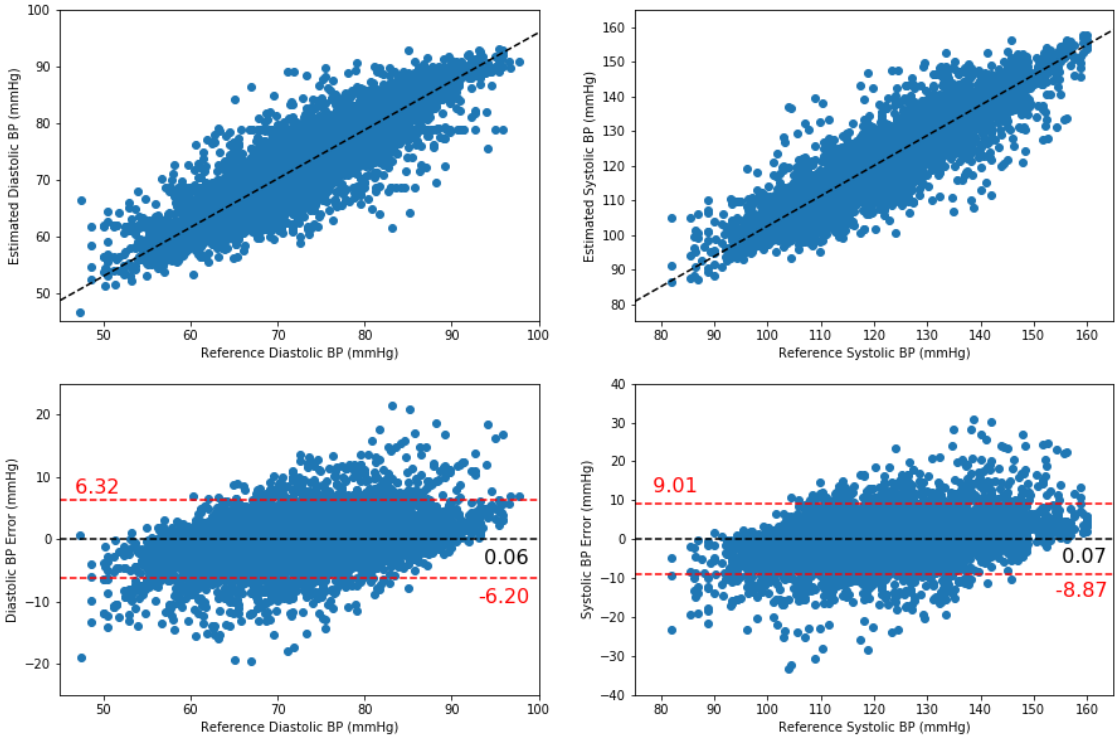


Figure 12.3: Bland Altman plot for MTL beat-to-beat model.

Table 12.4 provides the beat-to-beat results from the MTL with architecture optimized by NAS. Diastolic blood pressure averages RMSE of 2.91 mmHg with 0.47 standard deviation, and systolic blood pressure averages RMSE of 4.46 mmHg with 0.90 standard deviation. All subjects have diastolic RMSE less than 4 mmHg and all systolic RMSE are below 6 mmHg.

The Bland-Altman plot for beat-to-beat NAS-MTL is shown in figure 12.4. The upper figures show that the NAS-MTL model is able to estimate blood pressure in all ranges. The 95% CI for the limit of agreement for diastolic and systolic blood pressure estimation are shown in red. In this model, 85% of predictions have errors within 4.21 mmHg diastolic and within 6.30 mmHg systolic. With this model 99.2% predict with diastolic error less than 10 mmHg, and 95.7% of predictions have systolic error less than 10 mmHg.

Table 12.4: NAS-MTL beat-to-beat performance per subject for diastolic and systolic blood pressure (DBP & SBP) RMSE (mmHg) and R.

Subject	DBP RMSE	SBP RMSE	DBP R	SBP R
1	2.88	4.03	0.91	0.87
2	3.39	5.00	0.92	0.95
3	2.31	3.20	0.92	0.95
4	2.66	4.65	0.87	0.90
5	3.11	3.05	0.90	0.93
6	3.97	5.43	0.88	0.91
7	3.33	5.14	0.89	0.91
8	3.56	5.71	0.89	0.94
9	2.56	3.57	0.88	0.88
10	2.53	3.55	0.90	0.89
11	3.00	4.82	0.88	0.91
Mean	2.91 ± 0.47	4.46 ± 0.90	0.89 ± 0.01	0.92 ± 0.02

12.4.3 Analysis

As shown in figure 12.2, the MTL model is able to accurately estimate the changing of blood pressures and track along with the ground truth blood pressure through its peaks, troughs, and rapid changes. We also notice that the plots from estimated blood pressure are smoother than the plots of measured blood pressure. When comparing table 12.2 and table 12.3, we observe that the MTL model has better performance for both diastolic and systolic blood pressures across all metrics in comparison to individual models. The MTL model decreases the average diastolic RMSE by 0.25 mmHg and the average systolic RMSE by 0.72 mmHg. The lower RMSE and higher correlation indicate that the estimated blood pressure from the MTL model has superior alignment with the targets. More importantly, the performance of the MTL model is more stable across subjects than are the independent models. For some subjects, such as subjects 6, 7, and 8, separate models obtain startlingly worse performance. Therefore, the two outcomes produced by the MTL model and its joint loss function clearly provides some performance boost that is lacking in the individual case.

These results also show a difference in the difficulty of the two tasks. Diastolic blood pressure is typically estimated with a lower error than is the systolic blood pressure. Systolic

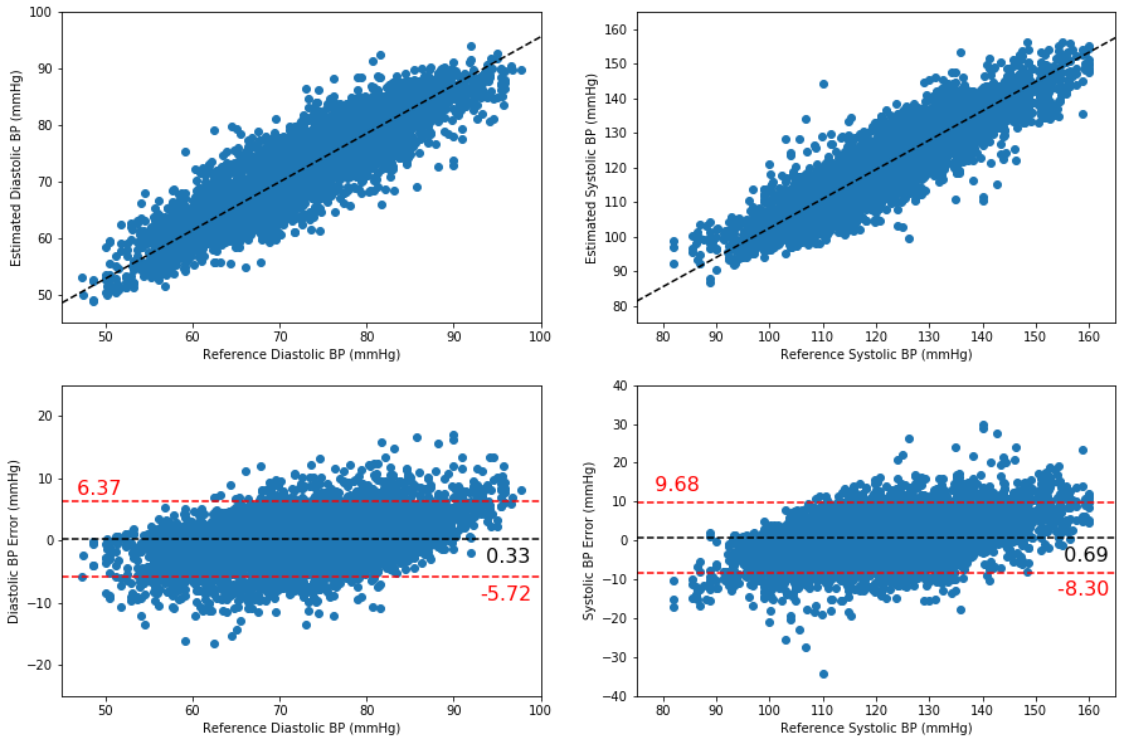


Figure 12.4: Bland Altman plot for NAS-MTL beat-to-beat model.

blood pressure has a wider physiological range than does diastolic and, in our dataset, the systolic blood pressure similarly shows this wider range. However, the high correlation exhibited by our estimations of systolic blood pressure shows that the model follows along with the variations across this wider range.

Applying NAS has variable results, with the optimal model from some subjects providing a much higher performance than for others. For many subjects, it finds architectures that give similar results as to the MTL without NAS, meaning that NAS saves the effort of manually searching and producing models with some good results. However, there are some subjects, namely subjects 5, 6 and 11, which have inferior results to the model produced by manual searching. This likely results from fixing the number of MTL hidden layers to at most five and only searching the hidden size of each layer. It is possible that five layers are overwhelmed for these three subjects, as the manual search found that two layers gives a superior performance for these subjects. Here we only focus on using NAS to search the

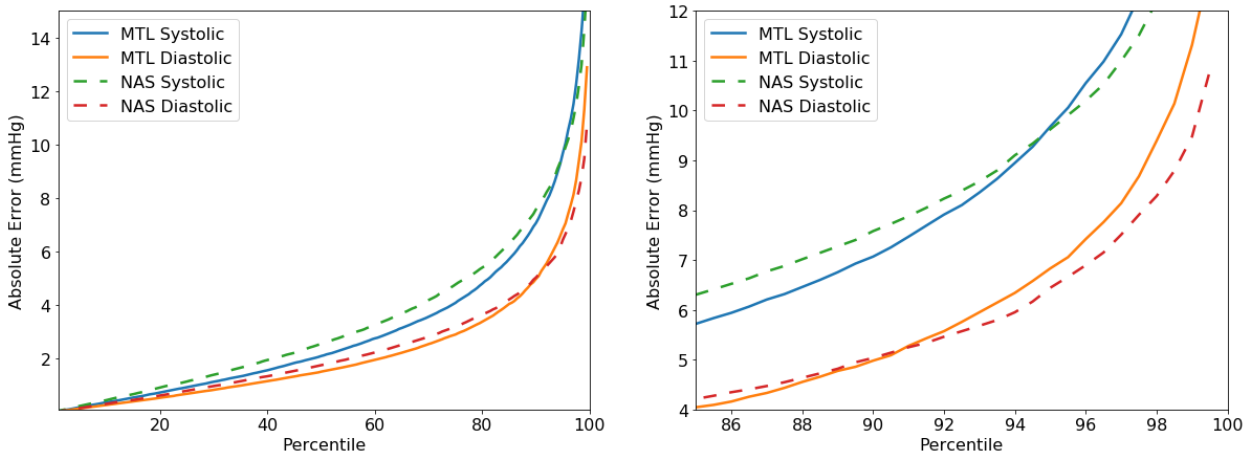


Figure 12.5: Plots comparing absolute error by percentile for MTL and NAS-MTL models. The base MTL model has lower error for the most values, but higher error among its 10% worst predictions of diastolic error and among its 5% worst predictions of systolic error. NAS performs slightly worse on most points, but has smaller error at the extremes, represented by where the plots cross. Both plots show the same data, but the plot on the right is scaled to show the transition between the relative model performance.

hidden size. We plan to extend to the number of layers in our future work.

One unexpected result in analyzing the output from the NAS-searched space was that in several cases, the size of the hidden layer surpassed sizes that were expected and searched manually. In several cases NAS found optimal performance with up to 76 nodes in a hidden layer, which was beyond the range of what we expected to be worth considering in our manual search. This reflects a central limitation of manual grid searching in that the search space can be limited by initial selection. We also found that the NAS-searched architecture has a similar hidden size for all subjects in the first three layers, meaning that all the subjects may have an underlying relation that is found by those three layers, and that the last two layers account for personal variability in blood pressure estimation.

When comparing NAS searched and manually searched MTL model, NAS searched MTL is more accurate for the extreme blood pressure estimation, as shown in figure 12.2, and has less errors over 10 mmHg, as shown in figures 12.3 and 12.4. However, NAS searched MTL acts worse than manually searched MTL for the middle blood pressure. Starting from

90% percentile and 95% percentile of errors, NAS has benefits over manually searching. The tradeoff of where NAS outperforms MTL is more readily seen in figure 12.5. Here we see that the most extreme errors produced by the NAS model are smaller than the errors produced by the MTL model, but that the less extreme errors are greater in NAS model. We conclude that NAS searched architecture has better performance on extreme blood pressure estimation than manually searched MTL but is worse for the middle blood pressure. The reason is that the number of layers for NAS is fixed to be five, and the searched architecture has bigger hidden size than manually searched architecture. Therefore, NAS searched architecture can provide more computation for the more difficult to estimate extreme blood pressures but may cause redundant computation for a middle and easy blood pressure.

12.5 Discussion

Our study shows that an MTL-based regression is able to estimate patient blood pressure successfully with bioimpedance signals. Our proposed beat-to-beat model performs with RMSE of 2.91 mmHg and 4.46 mmHg for diastolic and systolic blood pressure, which has a significant improvement over the ISO standard of 10 mmHg. The high correlation means that our estimation not only has low errors, but also accurately represents the range of physiologic blood pressures in our sample population. The MTL model is more efficient and accurate than separate models, decreasing the RMSE of diastolic blood pressure by 0.52 mmHg and 0.79 mmHg for systolic blood pressure. Diastolic and systolic blood pressure are not linearly related, but are both fundamentally tied to the true arterial blood pressure, and are both estimated from the same bioimpedance signals. The shared LSTM and dense layers in the model are trained by both diastolic and systolic blood pressure, and thus can be balanced during training and allow for the shared network to learn features important to both tasks. This approach can be extended to other applications by predicting other tasks and diseases that are inherently related.

In comparison to traditional tree-based blood pressure estimation algorithms, the deep learning-based MTL model not only improves performance but allows for further optimiza-

tion. We applied NAS on the MTL to optimize model architecture by producing the task-specific networks layer-by-layer. Even though the task-specific networks are separate, the multiple tasks are related and from the same shared layer. Thus, NAS is able to successful produce the hyperparameters layer by layer sequentially from a recurrent network. NAS is encouraged and trained toward better MTL performance, but replaces manual search with an optimizer for more accurate guidance and thus can provide a superior method of searching the hyperparameter space. The MTL model produced by NAS has improved estimation at extreme blood pressures and has lower errors in the extremes,allowing for greater practical utility.

12.5.1 Limitations & Future Directions

A chief limitation in this work is in the nature of the population from which our data was derived. The data was derived from subjects who were healthy individuals aged between 18 and 30 with no evidence of or known cardiovascular disease. Needless to say, these subjects do not represent the full spectrum of patient physiologies that should be incorporated in this modeling approach. Future work has been planned to collect data on patients in a clinical setting, and to better understand the variation present in a more robustly varied population. As more data becomes available, the methods outlined in this paper can be adapted to account for the anticipated increase in diversity. The logical continuation beyond that point is to collect the data of patients with known hypertension and to assess the clinical utility of utilizing this technique for adaptation to direct patient care. From there, this technique could be expanded to acutely ill patients in an ICU setting who may be experiencing rapid changes in blood pressure, and who could benefit from increased continuous noninvasive monitoring.

Another limitation and future direction of NAS is to better understand the searching space in which data measurements that belong to a given subject exist. Our application of NAS is under a restriction of a given number of layers and only searching the hidden size. A reasonable next extension of NAS would be to search over not only the hidden layer width,

but to also search over the optimal number of hidden layers. Further, the time for training a NAS model and the minimal training data are also worth exploring in our future work.

Finally, future work is needed to further explore ways of interpreting this work. What physical signal(s) are important to this model? For instance, while we hypothesize that PTT and PWV are important findings directly related to the physical interpretation and measurement of blood pressure, our current model functions as a black box and does not allow for evaluation which would determine how large of a role derived physical features play in blood pressure determination. Interpreting the mechanism of this model and peering into the black box could allow for more directed sensor measurements and an overall improved system.

12.5.2 Conclusion

In this paper, we propose an MTL based beat-to-beat blood pressure estimation model from cuffless bioimpedance signals. The MTL model achieved 3.18 ± 0.50 mmHg and 4.53 ± 1.03 RMSE with 0.77 ± 0.09 and 0.80 ± 0.10 correlation for diastolic and systolic blood pressure, respectively, which shows benefits of MTL over the individual-task models. Additionally, the MTL model obtained an RMSE of 1.57 mmHg and correlation of 0.93 for diastolic blood pressure and an RMSE of 2.31 mmHg and correlation of 0.94 for systolic blood pressure when comparing against the 10-beat averaged state of the art model. When comparing our manually searched MTL architecture and the results from NAS, we observed that NAS improves over the manually searched model, and the discovered optimal architecture from NAS is beyond the initial expected search parameters from preliminary experiments. In the future, we consider expanding the NAS approach to a bigger searching space, and apply it on reduced training data for a common problem of the time-consuming data collection process in clinic.

13. USING IOT SENSORS OPPORTUNISTICALLY TO ENHANCE HUMAN ACTIVITY RECOGNITION USING A MIXTURE OF DEEP NEURAL NETWORKS

Returning to the deep mixture of experts approach, this penultimate chapter uses environmental sensors in concert with on-body sensors to estimate activity and context. The mixture of experts approach allows for intelligently selecting optimal sensors based on a small subset of on-body sensors. This approach could be expanded for intelligently selecting from among any longitudinal sensors for the most appropriate to use for predictive tasks at any given time.

13.1 Introduction

Personal health monitoring and tracking has become more feasible through ubiquitous, wearable sensors, such as smartwatches and smartphones [476]. Systems built around these devices can track personal activity and query individuals for understanding behavior and health [477, 478]. For example, tracking food intake enables users to maintain a healthier diet [479], tracking exercise can improve recovery after heart attacks [480], and understanding emotion and behavior of individuals allows for better interpersonal conflict resolution [481]. As sensing technologies become more pervasive, the internet of things (IoT) applications can be augmented to enable the internet of medical things (IoMT) through the combination of wearable sensors and sensors in the immediate environment, referred to as nearable sensors [232]; we focus on wearable sensors for personal tracking as a base for all IoMT applications by either extracting direct parameters of human activity recognition (as is necessary for cardiac rehabilitation [480]) or as context for other tracking activities [232]. Smartwatches are powerful IoT devices that allow for accurate activity tracking in a wide variety of scenarios involving constrained problems, such as that as a known workout routine [337]. However, such solutions may present much lower accuracies when relaxed into further unconstrained

environments, potentially requiring help from additional data sources to recover performance [232].

The increasing prevalence of wearable and nearable sensors has increased the quantity and types of data that can be collected on individuals in remote environments [482]. This abundance of sensing results in an abundance of data that can be processed for tracking, processing, decision making, and alerting (if needed). However, this heterogeneous data may be produced at different rates, for different applications, and determining what IoT data is useful to help applications is becoming an equally important challenge to having high accuracy and ease of usability in those applications [482]. Modeling techniques are necessary to intelligently sort through the abundance of available data and identify the key sensors from which to capture data.

In many applications of machine learning, heterogeneous data is divided into smaller homogeneous groups that can be modeled more accurately through clustering. This clustering provides interpretability as each sample can be described more readily by the group to which it belongs. Models can be trained on each specific cluster in order to identify the key features relevant to each cluster rather than the entire heterogeneous dataset. In mixture of experts (MoE) models [483] such group-level clustering and supervised modeling are developed as a single training step, rather than splitting data a priori to then build models. MoEs have successfully served different applications from classification and regression tasks [484, 485] to phenotyping in medical datasets [486].

MoE-based approaches provide an opportunity to address personal tracking systems in IoT environments that have an abundance of sensors providing interaction data. By considering primary sensors to be those on-body (wearable) sensors that track personal health and all other off-body (nearable) sensors as indirect IoT sensors, we develop a deep MoE technique that enhances remote and wearable monitoring applications. This MoE technique opportunistically identifies a limited subset of sensors that can help in human activity recognition (HAR) tasks, boosting performance. This work develops this modeling technique for

improving HAR algorithms through the judicious leverage of IoT data. The contributions of this paper are as follows:

- Creates a multitask learning (MTL) deep mixture of experts technique to improve HAR from wearable sensors by augmenting estimations with data from nearable sensors, which we call the $\alpha\beta$-network.
- Develops an opportunistic MoE sensor selection technique for improving HAR tasks based upon the model performance and accuracy of the activities predicted from the wearable sensors.
- Provides an evaluation of model performance based upon the grouping of IoT sensors, the MoE approach, and the MTL addition to the MoE network.

The rest of this paper is organized as follows. Section 13.2 highlights recent work in IoT sensor selection, highlighting challenges and opportunities in the expanding data availability to enable personal and medical applications (internet of medical things). Section 13.3 discusses our deep MoE approach to enhancing personal tracking and activity recognition with intelligent selection of nearable sensors in an instrumented environment with IoT data, while Section 13.4 highlights case studies in a smart home environment and discusses the results. Section 13.5 discusses future opportunities from those findings, and Section 13.6 concludes this work.

13.2 Related Works

13.2.1 IoT Sensor Selection

The abundance of data generated by IoT applications, while providing accuracy to each individual application, can provide overwhelming amounts of data that make information processing difficult [487]. It is important to understand the context of the specific application

and quality of service needed to modify processing of the data [488]; however, these solutions provide application-specific solutions that are better than random clustering of sensors, focusing on cluster ability and not the end application performance [487].

Shukla, Maiti, and Sahoo discussed this challenge in growing data and the need to intelligently identify the most suited sensors for IoT applications by creating a mapping of sensors to applications in the context of latency and energy usage [489]. While they discuss the challenges, they focus on a latency-driven greedy approach to facilitate data collection. This work identifies an approach to sensor selection that is based in model and application accuracy for evaluation.

Jones et al. investigated a tiered approach to grouping sensors: by performance requirements, by environmental requirements, and by costs associated with application implementation [490].

Yachir et al. discuss methods of aggregating individual sensors into a cohesive system [491]. In this system, they partition sensing devices by general response type in order to improve system latency while retaining performance. Our work similarly looks at sensor response types, but looks at selection in response to performance rather than clustering and quality of service measurements.

Intelligent selection for applications has improved energy efficiency and application performance. However, these solutions are centered on the deployment of sensors for applications rather than processing data after deployment for a variety of applications.

Increasing prevalence of IoT sensors has lead to the problem of too many data streams to incorporate by traditional techniques. Detecting streams with high information yield and excluding streams with low information yield is a necessary task for intelligently utilizing these sensors. One approach to detecting novelty in sensors is to analyze the homoscedasticity and statistical features of incoming data streams [482]. These features have been found to be useful in real time analysis of event detection in IoT monitoring systems. However, these techniques are chiefly beneficial when distinguishing between event or no event, and do not

lend themselves to distinguishing between additional events and events caused by different users within the same environment.

13.2.2 IoT and Health

IoT devices of are particular interest in the context of assisting with user health. Many implementations of IoT device systems have been built with the goal of connecting patients with their healthcare providers [492]. These systems incorporate a wide array of devices including wearable devices to monitor motion or biological signals, and nearable devices to monitor activity at home. Yang et al. investigated some of the issues entailed in integrating the feeds of these various sensors into a single format that helpfully shares information both with the patient and with remote healthcare services [492]. However, while this technology succeeds in integrating specific and handcrafted IoT sensors into a single implementation, many of its sensors are specifically crafted to work with this system. For instance, one sensor is a smart pillbox that is able to calculate how many pills are taken at a given time by measuring the change in weight of the pillbox. This limits the ability of the system in scaling to new sensors in the absence of handcrafting the meaning of those sensors.

IoT devices are of key interest in assisting members of an aging population with performing activities of daily living (ADLs). Zhu et al. presented an overview of sensing systems that can be utilized for ambient assisted living and outlined barriers in sensor platforms for incorporating these systems [493].

Roggen et al. propose a paradigm of opportunistic activity recognition in the context of dynamic IoT environments [494] [495]. They show that opportunistic activity recognition is a superior approach than classical activity recognition given the rise of pervasive IoT technologies. They discuss ways in which this approach could be implemented in highly instrumented environments and provide such an instrumented dataset. Our work provides an implementation of opportunistic activity recognition based on this dataset, but reduces the number of wearable sensors to those that represent those that are currently the most pervasive: smartphones and smartwatches.

Ordóñez and Roggen additionally developed a deep neural network that achieves good activity recognition through fusing multimodal wearable sensors [404]. Their network architecture uses all wearable sensors, and is not expanded to nearable sensors. Our work allows for increased scalability over the model they describe, and incorporates nearable sensors opportunistically.

13.2.3 Mixture of Experts

In conventional MoE approaches, class assignments are completely unobserved and are chosen in the training process in a way that maximizes the log likelihood in each iteration [496]. These techniques can explore individual model performance by showing when a parametric model is incorrect and the impact of data towards model performance [497]. This type of grouping can also be used to interpret class assignments [486], explaining data and samples that provide interpretation for model performance. Therefore, it is important to select a model with enough expressive power that can fully explore the dynamic range of data and classification tasks. These techniques can potentially improve deep learning models, and have seen preliminary improvements [498], providing ample opportunity to improve activity recognition techniques.

Solis et al. used an MoE approach to automatically identify the context in which activities took place [232]. This paper defined a two-level network that pre-defined locations in which certain activities might take place and then used a second level to estimate actions, such as eating, in those places. However, their approach limited the benefits of the MoE approach by supervising the definition of location and the distribution of the two-staged network through the addition of the KL Divergence term in the optimization. In this work, we extend the idea presented by Solis et al. by developing a similar MoE approach that overcomes limitations of the prior approach by using an unsupervised technique to model activity from wearable sensors. and searches for the correct distribution of nearable sensors to augment/enhance the initial estimation of activities.

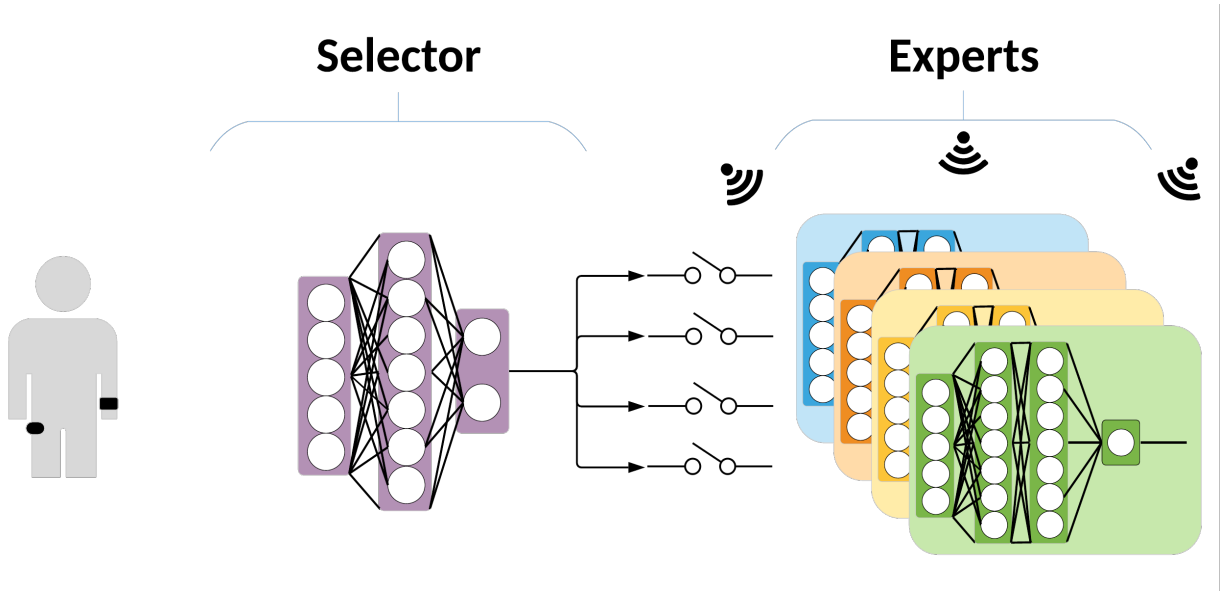


Figure 13.1: Single task deep MoE model with a α-network that always chooses one of several β-networks. Although in the abstract case soft labels of the β-networks could allow for distributions of multiple β-networks to be selected, in this implementation we restrict the α-network to selecting one β-network for every task.

13.3 Methods

To accomplish personal tracking and activity recognition via the efficient utilization of nearable sensors, we implement an MoE model to select the sensors which are most likely to provide high-yield information about user activity. The overall objective of this MoE model is to accurately classify user activity while minimizing the number of sensors used. While user activity can be classified with reasonable accuracy based on wearable sensors, addition of nearable IoT sensors may enhance activity classification. However, not all nearable IoT sensors will necessarily contribute to classification at any given point. For instance, in real-world situations other users may be active within the environment, introducing noise. Additional nearable IoT sensors may in fact cause difficulty in identifying user activity by confounding classification. As the number of nearable IoT sensors grows, the task of incorporating all sensors into a single expanding model might become intractable (or at least impractical). Therefore, depending on particular situation requirements, the goal may be to

provide this classification utilizing a minimal set of nearable sensors.

Our MoE is composed of a two-step set of models, called a $\alpha\beta$-network. The construction of the two-step $\alpha\beta$-network is adaptable to allow the network designer to focus on particular sensing needs and limitations. However, in any implementation, the base design stays the same. The first step of the $\alpha\beta$-network is called the α-network. This α-network takes input from some set of sensors and generates two predictions. The first is a raw prediction of the current user activity. The second is a prediction of which network and set of sensors in the second step is most likely to usefully supply additional information about the user and give a high prediction accuracy. Whatever this α-network's structure, its purpose is to generate an initial prediction and to select the optimal network in the second step for user activity classification.

The second step of the $\alpha\beta$-network is a family of β-networks. Each β-network takes as input different subsets of the overall sensor feeds and makes a prediction as to the user's activity. Each β-network is trained on a different subset of environmental IoT sensors. Although this training must be performed over all groupings of sensors, at runtime not all β-networks will be invoked. Indeed, if the α-network's initial certainty is sufficiently high, no additional β-network needs to be invoked, reducing resource utilization and processing required for prediction.

Here we describe the motivation for the design of the $\alpha\beta$-network, as well as multiple configurations of the $\alpha\beta$-network that allow for differential prioritization of resources and that can be adapted for different problems. We begin by describing the base α-network and the combined $\alpha\beta$-network. We then discuss our training approach, our hyperparameter tuning, and our approach for model evaluation.

13.3.1 β-network: DNN for HAR

Deep neural networks have proven to be effective for Human Activity Recognition (HAR) and have been used extensively in this domain. Convolutional neural networks (CNN) are powerful tools for extracting descriptive features from images or any data with a structure

similar to that of the images. As our base model, we use a 3 layer CNN network developed for HAR [499]. This model serves as our base and the networks that we develop, the α-network and the β-network, are its extensions. Their characteristics are explained below.

13.3.2 $\alpha\beta$-network

Here we develop the $\alpha\beta$-network model and its two components, the α-network and the β-network as a more selective and opportunistic model for HAR. A schematic view of the $\alpha\beta$-network model is presented in Fig. 13.1. Based on the input from a restricted and privileged set of inputs (in this work, wearable sensors), the α-network selects a particular β-network from all available β-networks. This selected β-network will then contribute to the final prediction by drawing on a particular set of inputs (in this work, some subset of nearable sensors). The selection of inputs to both α-network and β-network are determined by the application. In the abstract case, the α-network gives a distribution over different β-networks and then the final output would be the weighted average of the β-networks' outputs using the distribution given by α-network. Each β-network then uses a subset of the sensors to produce activity predictions. One of the objectives of our work is to intelligently select the β-network used in order to minimize the number of sensors that are used, allowing model scalability to larger systems. In order to achieve this minimization given the size of the dataset utilized here, we use hard assignments rather than soft assignments of β-networks. Rather than computing a weighted average where the weight vector is composed of probabilities of different β-networks being chosen, we set a 1 for the most probable β-network and 0 for all others. Therefore, only one β-network would be used for each prediction. This translates to using only one set of senors for each prediction. This is desirable as a smaller number of sensors entails lesser computation and resource expenditure, as discussed previously in Section 13.2.

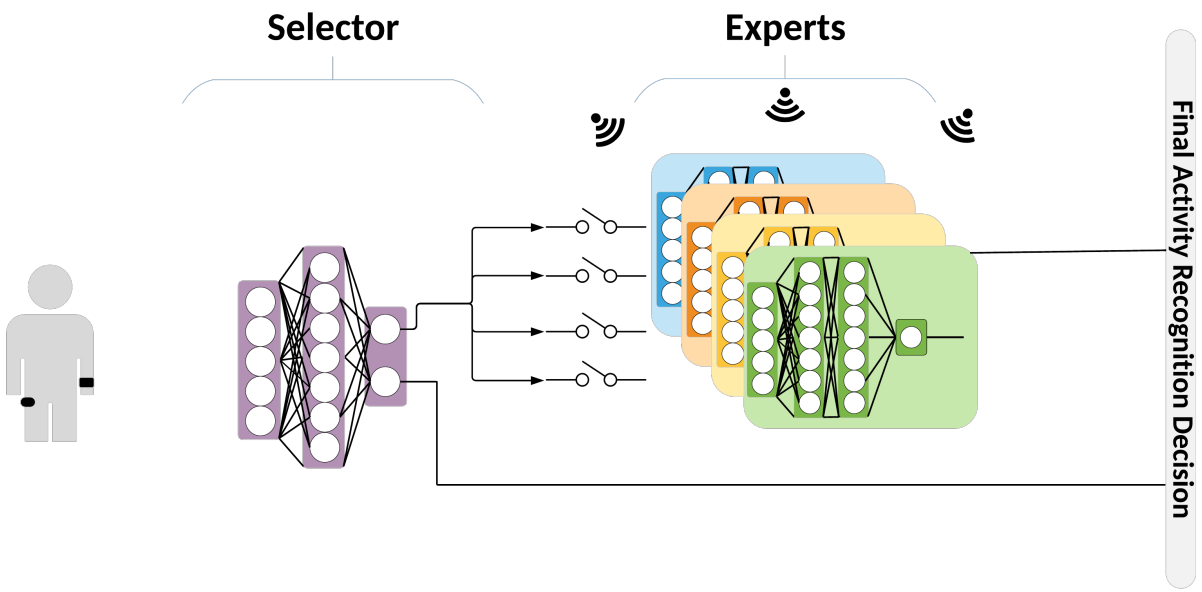


Figure 13.2: Multitask learning deep MoE model. This α-network can optionally select to either return its prediction, or to opportunistically utilize the sensors associated with a particular β-network. While in the abstract case a distribution of multiple β-networks could be selected, in this implementation we restrict the α-network to either implicitly select itself or to explicitly select exactly one β-network.

13.3.3 Baseline Models

In this work, we develop a model that in the first step uses a limited set of signals, and then opportunistically adds an expanded set of signals for a final prediction. In order to establish a baseline, we constructed two models that represent the full spectrum of performance from the minimal to maximal number of sensors. The baseline model shown in Fig. 13.3 consists of an activity recognition network trained only on wearable sensors. At the other extreme, the baseline model shown in Fig. 13.4 consists of wearable sensors and all nearable sensors belonging to a specific β-network. Although that network is tractable for the size of the dataset used here, larger datasets would become intractable in that model, requiring a more selective and opportunistic model to run pragmatically.

13.3.4 $\alpha\beta$-network Extension: Multitask Learning

In the first form of the $\alpha\beta$-network the α-network acts as a model selector and the β-network acts as a predictor. As an extension, we next explored implementing the α-network in a way such that it initially produces some prediction and an estimation of the certainty of that prediction by computing its entropy. Such uncertainty is used to decide whether the α-network and the sensors it is using is sufficient for making a prediction or we need to use an β-network and consequently IoT sensors. In some cases, the sensors in α-network may be sufficient to make a decision, reducing the benefit of utilizing an β-network. Additionally, the original sensors might have detected some activity that is not well-predicted by the β-network. Therefore, we developed the multitask learning framework so that the α-network makes a prediction of the output. If the model certainty in that prediction is below a given threshold, the model will select an additional β-network to augment the initial prediction. In the case that the α-network does not choose a separate β-network, the α-network is implicitly selecting itself for the prediction. Otherwise, it will select some other β-network. This model design is shown in Fig. 13.2.

13.3.5 Hyperparameter Tuning and Pretraining

The training process of deep MoE models can experience instabilities if not handled appropriately [232]. Pretraining can lead to a better modeling performance as the model is able to be guided through the highly non-convex optimization problem of training. It has been seen that in many cases the α-network can collapse into selecting only one β-network, no matter how many we provide. This happens as the initial state of the β-networks can be initialized in a way that a given β-network has a smaller error than others and the α-network gives it more weight. Larger weight causes the network to have a larger share in the output, making the error rate share larger as well. Consequently, this cycle may cause one network to provide a major contribution while others provide trivial contributions to the output. This can hamper the performance by making the α-network selection ineffective and trivial,

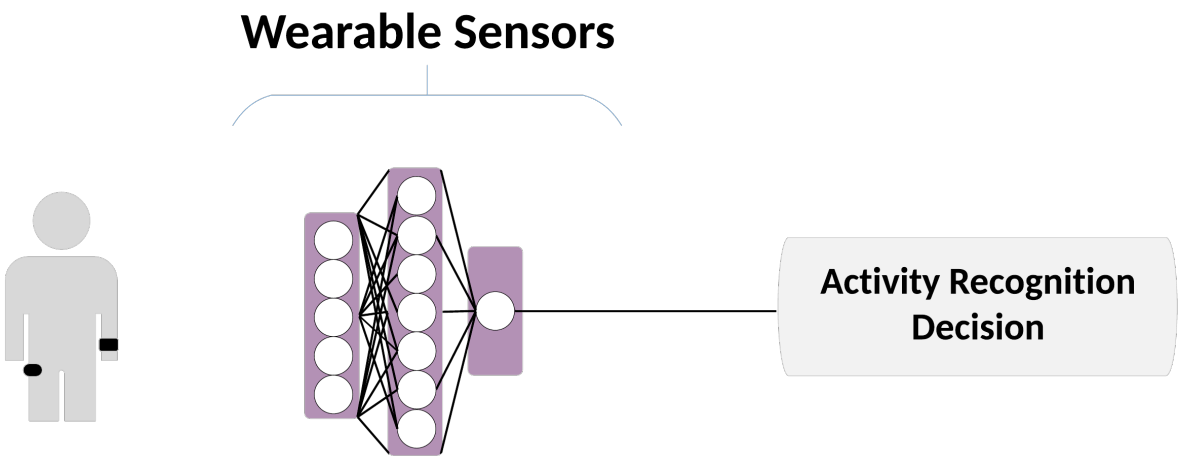


Figure 13.3: Baseline model incorporating only wearable sensors. This is a negative control, where no nearable IoT sensors are ever available to the base model.

hampering the benefit of the MoE approach.

To mitigate the complications of this error, we implemented a pretraining scheme as described in [232]. In pretraining, we train the α-network to output specific β-network assignments so that when the actual training begins our α-network starts from a balanced point so it will be more stable. β-network assignments used differ based on the application. In applications where there is no specific mapping between different β-networks and their usage, methods such as K-means clustering can be used and the α-network can be trained on such assignments. In this application we have used domain knowledge to group and assign β-networks (and equivalently, which sensor sets) to probable labels. We pretrain our α-network based on that assumption in order to rapidly achieve a stable starting point. Pretraining based on unsupervised clustering methods would be appropriate in the absence of such domain knowledge. The details regarding such assignments can be found in Table 13.1.

13.3.6 Training

The training was performed following the EM algorithm [496]. In EM, the lower bound of the log-likelihood is maximized in an iterative approach consisting of two stages, the E-step and the M-step. In the E-step, the parameters of the network computing the expected value of the log-likelihood, the α-network, are optimized while in the M-step, the parameters of the modeling networks, the β-networks, are optimized. Therefore, in the E-step the α-network is optimized while β-networks are frozen and in the M-step vice versa. The EM training alternates iterations of training either α-network network or β-network networks while keeping the other(s) frozen. It should be noted that, while the pre-trained networks are supervised, this final training step is unsupervised in terms of the β-network network assignments and solved using EM.

Multi-Task-Learning Training: The learning of the multitask learning network is slightly different from the training procedure of the basic α-network. In the multitask learning network there are two locations in the network that might output a prediction. The first is by the α-network and the second by the selected β-network. In contrast to the $\alpha\beta$-network where the loss function is only the cross entropy between the β-network output and the target, in multitask learning network the cross entropy between α-network and target is also added. We guide this with domain knowledge, when available. For example, the action of opening a door in our dataset is likely to be sufficiently detected by the door sensors, rather than dishwasher sensors or needing the wearable sensors. In these cases, the sensor set and consequently the β-network that uses that set can be known in advance. This label can be used for guiding the α-network regarding which β-network to choose. This will be similar to the pretraining process detailed above where α-network is trained to choose the β-networks. Here it is done in the while training, rather than before it.

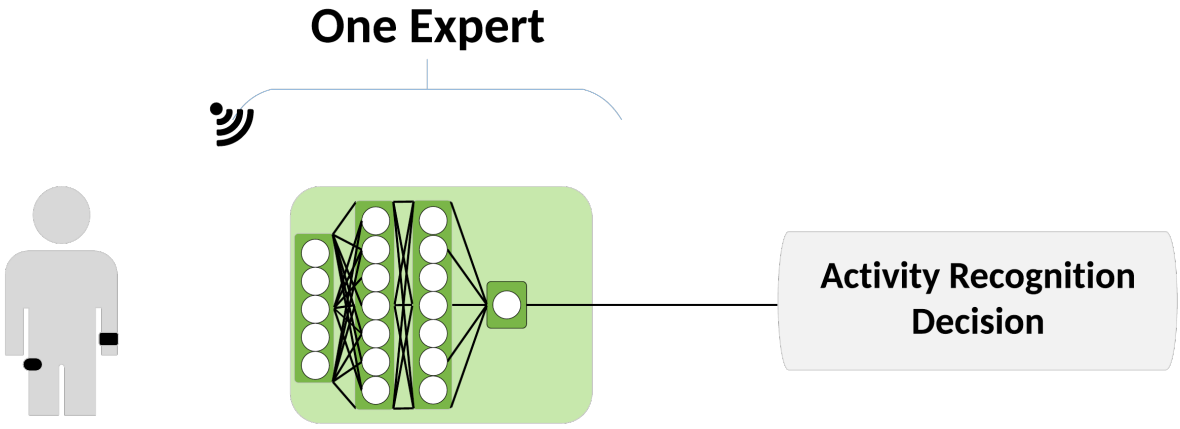


Figure 13.4: Baseline model incorporating a single expert model. This is one β-network from among the MoE, but has access to the signals typically provided to the α-network. Each individual β-network is implemented within its own expert baseline model.

13.4 Evaluation and Results

In this work, we build models which utilize a subset of sensors for a primary α-network. This α-network outputs either an activity prediction or a selection of the most appropriate group of sensors/β-network to provide an activity prediction. This allows for the α-network to select groupings of sensors that are most likely to return a high information yield, while excluding sensors that are unlikely to provide useful information. The α-network here is provided a minimal and privileged set of sensors (all utilized wearable sensors) and uses the information from these sensors to select the most appropriate additional signals, allowing for the system to entirely ignore the signals from irrelevant sensors. This has the desirable effect of opportunistically lowering resource utilization and computational complexity, while maintaining or improving accuracy.

In all implementations discussed here, the α-network makes a hard assignment as to which β-network is most appropriate to further specify the activity. While in the abstract it would be desirable for the α-network instead to output a probability distribution over all β-networks, the size of the dataset utilized here best exhibits the opportunistic gains of

this system by performing hard assignments. This choice of implementation was selected to ensure at each time instance one specific sensor set is used in order to maintain good performance while ensuring low resource utilization. If soft assignments were to be used, all the sensor sets in the dataset here would have a contribution to the output, defeating the purpose of using only one set. Hard assignments ensure that the increased cost of including all sensor streams does not happen.

Here we present the design of the different implementations of our $\alpha\beta$-network. We implement it in several increasingly sophisticated tasks as we develop its robustness. In Section 13.4.1 we detail the overal model setup. In Section 13.4.2 we describe the sensor used from the dataset and the motivation for choosing this dataset. Section 13.4.3 describes the specific network architecture used, and Sections 13.4.4, 13.4.5, and 13.4.6 describe the three case studies explored here. The case study described in Section 13.4.4 is built with a α-network that always chooses one β-network. The case study in Section 13.4.5 introduces artificial noise into the system and allows the α-network to optionally ignore all β-networks and instead to implicitly self-select when its confidence in its own estimation is high. The final case study in Section 13.4.6 expands the sensors used in the base α-network to incorporate multiple wearable sensors, a smartphone in addition to a watch.

13.4.1 Experimental Setup

The first α-network configuration uses a restricted set of wearable sensors in order to select between β-networks that share sensors with the α-network, but otherwise feature sensors that mutually exclusive to each other (Fig. 13.1). The α-network here has access to an wrist-mounted accelerometer and arm-mounted gyroscope that we treat as a smartwatch. This situation is of interest in cases where accessing sensor sets is expensive, but always accessing at least some external information is desirable. The α-network points the model to access one β-network to make the prediction. This configuration differs from the earlier implementation of Solis et al. [232] in the assumptions regarding cost of sensor acquisition and computation. In their configuration all sensors were accessible to all secondary networks

without computational cost as a limiting factor. In this configuration, not all sensors are accessible at the any time and α-network has access to limited wearable sensors. Based on those sensors, only one of several β-networks will be accessed to make the prediction.

The next α-network configuration featured a set of wearable sensors belong to the α-network where initial activity recognition is performed by the α-network, and further activity confirmation if necessary is performed by the identified β-network (Fig. 13.2). Again, the α-network here has access to an arm-mounted accelerometer and gyroscope that we treat as a smartwatch. Here, initial processing is limited to the set of sensors that uniquely belong to the user. This first step can ignore all other sensor feeds (the β-networks). Once a likely class of activity has been identified, the second step can utilize the appropriate β-network to more precisely identify the activity. As before, each β-network is restricted to a mutually exclusive set of sensors. However, here these β-networks are also mutually exclusive to the α-network, and are entirely nearable sensors. This implementation reflects the real-world situation where the wearable sensors are always of interest, but the nearable sensors are only of use in the (relatively) uncommon situation that the user is probably interacting with some instrumented object. In this setting, the decision is made by the first network and the result from the β-networks is used for augmentation rather than making the primary prediction. When the user is unlikely to be interacting with the instrumented object, signals from that object may be safely ignored, allowing for lower resource utilization.

In the final α-network configuration, we repeated the above above while augmenting the α-network. In addition to the arm-mounted accelerometer and gyroscope, we added a hip-mounted accelerometer. We treat this sensor as a smartphone that couples with the smartwatch to provide additional measurements of wearable activity.

In order to assess the $\alpha\beta$-network, we make performance comparisons to the baselines described in Section 13.3.3. These baselines are single β-networks built on different sets of sensors. Such a comparison is made to see how the model would perform if no sensor selection was performed and a single model was built on the sensors. The first baseline

Table 13.1: List of sensors used by β-networks in the case studies. In case studies 1 and 2 the α-network has access to the watch sensors. In case study 3 the α-network has access to watch and phone sensors. β-networks have access to mutually exclusive nearable sensors.

Name	Sensors list
Watch	Wrist Accelerometer and Arm Gyroscope
Phone	Hip Accelerometer
Expert 1	Sensors on Door 1, Door 2, Fridge
Expert 2	Sensors on Drawer 1, Drawer 2, Drawer 3
Expert 3	Dishwasher, Chair, Objects on Table

Table 13.2: Case study 1: Comparison between accuracies and F1 scores of $\alpha\beta$-network and the baselines, a single β-network network. The recognition task was activity recognition in the Opportunity dataset.

Model	Sensors	Accuracy	F1 Score	F1 Gain
β-network	Watch	0.24 (0.03)	0.32 (0.05)	-
β-network	Watch & Expert 1	0.44 (0.08)	0.43 (0.02)	34.4%
β-network	Watch & Expert 2	0.39 (0.06)	0.39 (0.04)	21.9%
β-network	Watch & Expert 3	0.26 (0.01)	0.37 (0.04)	15.6%
$\alpha\beta$-network	Watch & Experts (All)	0.44 (0.05)	0.40 (0.06)	25.0%

is an β-network which using only the restricted set of wearable sensors (Fig. 13.3). This, specifically, is opposed to the case where a α-network uses particular wearable sensors to decide which β-network to select. This baseline would pinpoint the modeling capability of a single β-network built on that restricted sensor set. The other baselines are β-networks built on the sensor sets α-network could to choose from. This baseline was designed to get an understanding of how good any of the β-networks to be chosen can perform assuming they are always chosen when appropriate. This enables us to isolate the effect of the α-network in constructing the composite $\alpha\beta$-network.

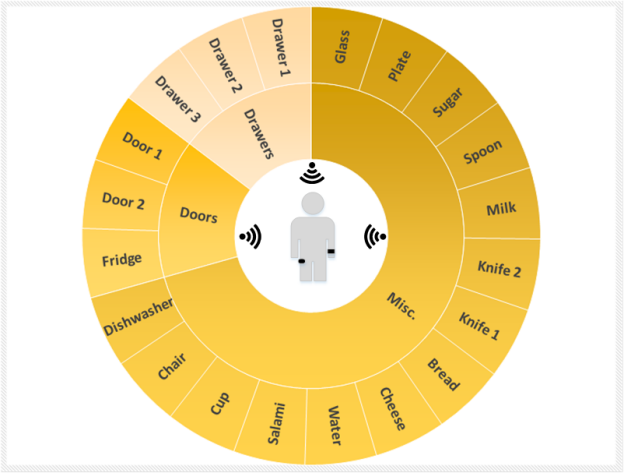


Figure 13.5: Opportunity selected sensors and IoT sensors broken up by category. Notice that the Misc. sensors represent a greater number but that the data provided by all the sensors are roughly evenly distributed.

Table 13.3: Case study 2: Comparison between accuracies and F1 scores of $\alpha\beta$-network and the baselines, a single β-network network, when noise is present in the data. The recognition task was activity recognition in the Opportunity dataset.

Model	Sensors	Accuracy	F1 Score	F1 Gain
β-network	Watch	0.26 (0.03)	0.29 (0.06)	-
β-network	Watch & Expert 1	0.46 (0.04)	0.42 (0.02)	44.8%
β-network	Watch & Expert 2	0.36 (0.09)	0.34 (0.08)	17.2%
β-network	Watch & Expert 3	0.35 (0.08)	0.31 (0.08)	6.9%
$\alpha\beta$-network	Watch & Experts (All)	0.42 (0.04)	0.38 (0.06)	31.0%

13.4.2 Opportunity

The Opportunity dataset [495] is a valuable dataset for human activity recognition gathered by a variety of sensors in a smart-home environment. Items that users could interact with had sensors indicating this interaction, and the users wore inertial measurement units on their upper bodies, hips and legs for activity tracking. This dataset is well-suited for our problem as it entails a great deal of both wearable and nearable sensors in an environment where the user is interacting with various objects. For wearable sensors, we emulated a

smartwatch by using sensors attached to the arm and emulated a smartphone in a pocket by using sensors attached to the hip. The sensors utilized here are shown in Table 13.1 and the instrumented nearable objects are represented in Fig. 13.5. The data collected was labeled in a hierarchical structure, where each activity had different levels of annotations, from higher order understanding of the activity to specific movements each hand is performing (e.g., relaxing vs. opening or closing a specific door). Specifically, the Opportunity dataset has 7 levels of hierarchical labels. Higher level labels describe details such as subject posture while lower level labels describe the hand movements or the objects they are interacting with. We chose activity performed by both hands, such as opening door, as the activity. In total, the dataset contains 18 different activities performed sensed by 72 different sensors.

For data processing, we used non-overlapping windows of 1 second (30 samples) to divide up the time series data. The data had missing values and was mean imputed to fill these missing parts. We used a five-fold cross-validation to evaluate our models with testing accuracy and micro F1 score as performance metrics. All results reported in Tables 13.2-13.5 are presented as mean (standard deviation) of test fold accuracy or F1 score.

13.4.3 Networks Architecture

Following the model developed by [499] we used 3 convolutional layers followed by 3 fully connected layers for both the α-network and β-networks. Note that the only difference between the two networks' architectures was in terms of the number of neurons in the output layer. The implementation of the multitask learning network also differs from the other two in terms of the number of the neurons in the last layer. We implemented all the networks using *PyTorch* library in *Python*. Networks were trained using stochastic gradient descent with an initial learning rate of 0.001 and a momentum of 0.9, which provided the best results in cross-validation.

13.4.4 Case Study 1: Sensor Requesting

In our first case study, we implement the $\alpha\beta$-network to perform well in a situation where the cost of accessing and processing additional sensor feeds is high. In this case, we restrict the α-network to only a specific subset of wearable sensors, specifically, sensors on the wrist which resemble the data collected by a watch which is a ubiquitous means of data collection. Each β-network is composed of the wearable sensors in addition to a category of nearable sensors that is mutually exclusive to the other β-networks. The list of the sensors used for each network can be found in Table 13.1. β-networks, on top of the α-network have access to non-overlapping parts of the data. At each instance, the α-network requests access to and uses only one β-network to perform the prediction. We define each β-network to be all sensors on a specific class of object. In this particular implementation one β-network is composed of all sensors on drawers, another is composed of all sensors on doors that rotate about a vertical axis (i.e. not the dishwasher's door), and the last is all remaining sensors (objects on the table and the dishwasher).

Table 13.2 compares $\alpha\beta$-network with the baselines of single β-networks trained using different sensors in Table 13.1. We find that the baseline network with only wearable sensors performs at an accuracy of 0.24 and F1 score of 0.32, while the $\alpha\beta$-network performs similarly to a single β-network with accuracy of 0.44 and F1 score of 0.40, which represents a 25% gain over the wearable-only baseline. We see that while a single β-network on the watch data can produce similar results to the $\alpha\beta$-network, several of these fail to adequately accommodate for complexities in the data. That yields the worst performance among all as it fails to capture many of the complexities in the data. Note that although the accuracies/F scores presented in the table seem low, these values are deflated due to the utilization of 17 different labels predicted by a limited sensor set. This demonstrates both the degradation of HAR when conducting activities that may or may not be detected by smartwatches in unconstrained environments, and the improvements that can be achieved by augmenting estimations with available nearable sensor data. This result aligns with other published work relating to

Table 13.4: Multitask learning network. As the noise of the dataset increases, the α-network becomes more likely to rely on itself than to utilize separate β-networks. These $\alpha\beta$-networks incorporate signals from the Watch and from all Experts.

Dataset	Accuracy	F1 Score	Implicit Self-Selection Rate
Clean	0.43 (0.03)	0.39 (0.05)	3.62%
Noisy	0.41 (0.06)	0.36 (0.07)	6.15%

Table 13.5: Case study 3: Comparison between accuracies and F1 scores of $\alpha\beta$-network and the baselines, a single β-network network, when implemented with a α-network that includes both a smartwatch and a smartphone.

Model	Sensors	Accuracy	F1 Score	F1 Gain
β-network	Watch & Phone	0.23 (0.02)	0.29 (0.06)	-
β-network	Watch, Phone, & Expert 1	0.48 (0.01)	0.44 (0.01)	51.7%
β-network	Watch, Phone, & Expert 2	0.42 (0.06)	0.39 (0.06)	34.5%
β-network	Watch, Phone, & Expert 3	0.29 (0.06)	0.37 (0.02)	27.6%
$\alpha\beta$-network	Watch, Phone, & Experts (All)	0.38 (0.09)	0.33 (0.12)	13.8%

smartwatch HAR accuracy in unconstrained environments [337].

13.4.5 Case Study 2: Intelligent Sensor Selection

In our second case study, we implement the $\alpha\beta$-network to perform well in a situation where the cost of accessing and processing additional sensor feeds is low, but where the additional feeds come with a large amount of noise. While some signal is contained within these sensors, it is as before desirable to implement the network so that only the most information-rich sensors are included. Here however, we make the assumption that accessing any additional sensors may be too expensive. Therefore, the α-network is structured so that it is able to ignore all other β-networks by implicitly selecting itself, provided that its confidence is high. In this situation, we construct the α-network to receive input from all sensors. The α-network then returns the identity of the β-network with the highest potential

information yield. The list of the sensors the β-networks have access is unchanged and shown in Table 13.1.

In order to implement a noisy environment, in each time window we randomly add noise to various nearable sensors. In this approach, each object is given a chance to have each of three types of noise added. All channels from a given object's sensors (i.e. all acceleration and gyroscope axes of a particular object) are subjected to the same type of noise. The chance of each of these are independent, with the possibility for multiple types of noise to be added. The types of noise are a) zeroing the signal for the entire window, b) adding uniform random noise of the within the amplitude of the signal, and c) adding Gaussian noise with 0 mean and σ equal to the amplitude of the signal. The first type of noise represents a malfunctioning sensor that is inappropriately offline or unable to connect. The second noise represents multiple users interacting within the environment, and activating sensors with signals that are not of interest to the system. The final noise represents random noise from potential communication cross-talk in a highly instrumented environment.

Table 13.3 compares $\alpha\beta$-network with the baselines of single β-networks trained using different sensors in Table 13.1. Here as before, we see that the $\alpha\beta$-network is able to appropriately select from among the various β-networks in order to maintain good accuracy and F1 score, despite the introduced noise. Here we see that the $\alpha\beta$-network reaches as F1 score of 0.38, which represents a 31% gain over the wearable-only baseline. Additionally, Table 13.4 details the self-selection rate in the clean dataset and in the noisy dataset. In the clean dataset, the α-network produces an estimate with high confidence 1% of the time. However, when trained on the dataset with injected noise, the α-network recognizes the decreased reliability of environmental sensors and self-selects 6% of the time. This shows that our model is able to opportunistically recognize and evaluate the quality of its training data. When that data is distorted to the point of no longer being useful, the model intelligently accounts for that lack of reliability.

13.4.6 Case Study 3: Augmented α-network

In our final case study, we replicated the architectures from above, but augmented the α-network by introducing an additional accelerometer to represent a phone in the user's pocket. This duplicates the scenarios from above, but represents the user having an additional wearable IoT device. Both smartphones and smartwatches feature a growing prevalence, and thus this α-network is particularly well-suited as a realistic example of what a user might have access to.

Table 13.5 compares the $\alpha\beta$-network featuring this larger α-network with the baselines of single β-networks trained using different sensors in Table 13.1. The results here show that even though the baseline α-network is a poorer raw predictor than the baseline α-network in earlier studies, this baseline coupled with the β-networks give improved performance. This result underscores a limitation in our design where it appears that the baseline model which always utilizes β-network 1 (accelerometers on doors) outperforms any other system, including the opportunistic $\alpha\beta$-network. As this dataset is limited in environmental size, it stands to reason that certain β-networks have signals containing information that allows for greater overall predictive power than others. Applying this model to a dataset with more β-networks spread out more in space would likely remove this increased performance that this particular β-network has.

13.5 Limitations and Future Work

Although the Opportunity dataset is a high-quality dataset with a great deal of instrumented and labeled data, it is still fundamentally a small enough dataset that training and utilizing a DNN approach on all data within it is not computationally restrictive. Additionally, this dataset is very clean and while we were able to introduce synthetic noise, this is still less structured than true noise from other users would be. However, this work still provides a proof of concept for a system that could be extended to a much larger system with multiple active users introducing noise into the system. Additionally, the dataset used here lacked

a wrist-mounted gyroscope as is contained in many smartwatches. Ideally, the α-network would be constructed of data from a single on-wrist wearable rather than composed of multiple wearables at slightly different locations. Future work may be directed into implementing this framework in larger and noisier datasets and in less constrained environments.

One limitation of this work is the lack of soft selection of β-networks by the α-network. Given the size of this dataset and that the objects belonging to the three β-networks are in such close proximity in a single-user environment, this dataset is inappropriate for implementing soft selection over the β-networks. Future work in a larger dataset could be directed towards utilizing the full potential of this model by allowing for soft β-network selection. This would increase computational overhead, but could be tailored to balance this overhead with expected gains in recognition accuracy.

This work was constructed using domain knowledge to facilitate the construction of the β-networks and to improve the pretraining of the $\alpha\beta$-network. This level of domain knowledge might not always be available when implementing this framework. Future work should be directed to improving and facilitating the process of constructing the β-networks and training the $\alpha\beta$-network without such domain knowledge.

Future work could also be directed towards applying this MoE framework to fields beyond that of HAR. This deep MoE approach could be applied to many situations where there is a large amount of unstructured data with underlying groupings of signals that would be more easily and more correctly predicted with specialized β-networks.

13.6 Conclusion

In this work we present a model for opportunistically utilizing IoT sensors for augmenting HAR. This model is developed with two implementations: one in which a primary α-network selects from among several β-networks in order to classify an activity with a minimal set of nearable sensors, and one in which a primary α-network either provides a classification, or chooses a minimal set of nearable sensors with which to classify the activity. The purpose of this model is to utilize IoT sensors for HAR when there is a significant reason to do so,

but to ignore the sensors when they provide little benefit. This is important as IoT devices continue to become more pervasive as simply merging all IoT features into a single model becomes intractable with increasing sensors. We validate our model using a highly instrumented dataset that provides both wearable and nearable sensors within an environment. In particular, we show that our model is able to change its behavior to recognize unreliable or noisy environmental sensors and can alter its mode of opportunistic MoE consultation depending on the reliability of those sensors.

This work provides a framework for and example of constructing a deep MoE model that opportunistically utilizes additional information when initial model uncertainty is low, or that provides model outcomes without using unnecessary information when model uncertainty is high. This method of opportunistically selecting a particular β-network allows for an overall system that is extensible to environments with large amounts of excess data without needing to directly include every aspect of the environment in individual model outputs. This will allow for HAR systems that are well-suited to operation in large and unbounded environments with improved prediction quality when incorporating all signals is intractable.

14. CONCLUSION

This dissertation presents applications and developments in machine learning for the advancement of healthcare. This dissertation first examines applications of advanced machine learning techniques in clinical settings for outcome prediction. While deep learning techniques represent the cutting edge of research in machine learning, we see that naively applying them on clinical datasets provides small or no benefits in comparison to shallower machine learning techniques such as logistic regression or gradient boosted trees. Along with slim or minimal prediction improvements, neural networks are inherently less interpretable than logistic regression or even gradient boosted approaches.

Logistic regression is trivially interpretable. As seen in Chapter 2, the weights of a logistic regression model are directly translatable to odds ratios. In the analysis of patients testing positive for COVID-19, variables strongly predictive of both admission (Figure 2.4) and mortality (Figure 2.6) are directly interpretable. This highlights one of the most important factors of logistic regression: linear interactions are easily visualizable and interpretable. Cursory inspection of Figures 2.4 and 2.6 immediately shows that age is a driving predictor for those outcomes. Further inspection shows that sex and certain comorbidities similarly contribute to predictions for those outcomes.

However, when judiciously applied, neural networks may aid in interpretation. The deep MoE presented here allows for a structuring such that the underlying model structure allows for patient grouping and interpretation. Rather than exploring traditional dimensionality reduction and clustering techniques such as described in Chapter 3, this model allows for a clustering approach driven by outcomes in a semi-supervised manner. By training multiple experts and a classifier among experts, each expert can be trained to have higher performance on a given subset of patients, while the classifier learns to classify patient subsets. Backpropogation through the entire model allows for the classifier to be driven by outcome prediction, but in a way that allows for future classification prior to knowing outcomes, or

even all inputs.

Neural networks are also increasingly valuable over shallower approaches as data becomes longitudinal. Shallow approaches work well with tabular data. To some extent, longitudinal data can be coerced into a tabular form, but this approach has diminishing returns. This is notably true in natural language processing. Chapter 4 details a novel end-to-end framework for utilizing deep learning to extract problem categories using free-text notes. This approach allows for a translation from longitudinal text data into a tabular problem list. This approach further aided interpretability through the use of an attention mechanism to interpret the relative importance of words in the narrative.

This dissertation next looks at advancing clinical decision support. Propensity score matching is a tool that allows for retrospective estimations of what prospective studies might have uncovered. However, linear techniques for propensity score matching can oversimplify relationships, missing complexities inherent in heterogeneous data. In this work, traditional propensity score matching was first enhanced by use of nonlinear gradient boosted trees instead of the simpler logistic regression to generate propensity scores. This nonlinear approach allows for matching along more complex relationships as higher order interactions may be learned. The deep MoE approach in particular is a novel and suitable tool for this task, allowing for more direct modeling and grouping of heterogeneity within a population. This approach allows for the realization of a more personalized machine learning approach that describes neighborhoods of patients within the deep network latent space, rather than by nearness of a scalar propensity score.

Traditional propensity score matching entails matching one or more cases with one or more controls. A propensity score is created by using logistic regression to estimate the likelihood that a patient will receive an intervention. In the work described in Chapter 7, propensity score matching was performed with XGBoost and validated independently with logistic regression. Nearly all patients receiving a microaxial LVAD (1680 patients out of 1768, Figure 7.1) were successfully matched with a patient receiving an IABP. In this analysis,

use of an intravascular LVAD was associated with a significantly higher risk of in-hospital mortality in comparison with patients receiving IABP (45.0% vs 34.1%, $p<0.001$; Figure 7.2). However, these findings and other findings presented in Chapter 7 were met with concerns that the techniques used may have missed key interactions which would show benefit under certain situations [500, 501, 502]. This limitation was motivating in the following chapters.

Applying a dynamic series of models, Chapter 8 analyzed heterogenous effects in cohorts undergoing PCI. This work highlights the importance of incorporating information in a longitudinal manner, and allows for doing so in a particular setting. The variability in patient risk scores as they proceed through an episode of care highlights the heterogeneity of the population, driving the need for further advanced models that can better describe that heterogeneity among patients and treatment effectiveness at a population level.

Chapter 9 applied the deep MoE approach to discover phenotypic clusters within the population of patients with AMI-CI. This approach aimed to account for heterogeneity that could lead to misleading propensity score matching. This approach allowed for a joint learning of both phenotypes and outcomes, and separation of variables by timing allows for a model that learns with the advantage of outcome information, but that can fairly assign phenotypes earlier in patient presentation. This approach allows for learning a representation more specific and personalized to a local and more homogeneous population from among the wider heterogeneous population.

Finally, this dissertation describes the expansion of the techniques used here to natural environments. A key factor driving the necessity of using deep learning models is longitudinal data: while approaches such as logistic regression and gradient boosted trees perform well on tabular data, they expand poorly to longitudinal data. This is increasingly pertinent as wearable sensors for remote health monitoring are developed and widely adopted. As described in Chapter 11, wearable sensors are not only being developed as medical devices, but are widely proliferating in the consumer sphere. Apple Watches, Galaxy Watches, smartphones, and other consumer devices feature powerful sensors that are able to collect

continuous cardiac signals for upwards of 20 hours per day.

Wrist-worn monitoring devices represent a rich opportunity for growth in remote monitoring. One particular application of this is in using bioimpedance signals to estimate blood pressure. The approach detailed in Chapter 12 characterizes the development of a deep model to infer blood pressure from such a wearable sensor. Future development of this device into a commercially viable device would be a boon for remote health monitoring and for improving diagnosis of various types of masked hypertension.

Additional environmental sensors, while not yet as widely pervasive, are a ripe opportunity in the realm of home health monitoring. In Chapter 13 an approach for using the deep MoE to select subsets of sensors is detailed. This approach assumes that environmental monitoring can grow widely to an unbounded network of longitudinal sensors. The approach described allows for intelligent selection of sensor groups such that the MoE can intelligently infer and utilize context. Increasing breadth of IoT increases need for approaches such as this that can extract and discover biomarkers from the noisy longitudinal sensors.

This dissertation aims to advance and apply machine learning for health care through three main goals. First, it described utilization of advanced machine learning techniques for clinical modeling, predicting harmful outcomes among vulnerable populations (Chapters 2-5). Second, it described advanced machine learning techniques to handle heterogeneity in retrospective analyses, introducing a novel application of a deep MoE for phenotype discovery (Chapters 6-9). Finally, it surveyed needs and opportunities in harnessing remote sensors for medical applications and described two particular instances where useful biomarkers were extracted from longitudinal sensors (Chapters 11-13). Through these goals, this dissertation presents and advances machine learning for healthcare applications both within and beyond the clinic.

REFERENCES

[1] R. L. McNamara, K. F. Kennedy, D. J. Cohen, D. B. Diercks, M. Moscucci, S. Ramee, T. Y. Wang, T. Connolly, and J. A. Spertus, "Predicting in-hospital mortality in patients with acute myocardial infarction," *Journal of the American College of Cardiology*, vol. 68, no. 6, pp. 626–635, 2016.

[2] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[3] B. Moody, G. Moody, M. Villarroel, G. Clifford, and I. Silva, "MIMIC-III waveform database (version 1.0)," 2020. [Online]. Available: https://physionet.org/content/mimic3wdb/1.0/

[4] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[5] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[6] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[7] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[8] D. Cirillo and A. Valencia, "Big data analytics for personalized medicine," *Current opinion in biotechnology*, vol. 58, pp. 161–167, 2019.

[9] S. Wright, D. Verouhis, G. Gamble, K. Swedberg, N. Sharpe, and R. Doughty,

"Factors influencing the length of hospital stay of patients with heart failure," *European Journal of Heart Failure*, vol. 5, no. 2, pp. 201–209, 2003. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/12644013

[10] D. W. Baker, D. Einstadter, S. S. Husak, and R. D. Cebul, "Trends in postdischarge mortality and readmissions: has length of stay declined too far?" *Archives of internal medicine*, vol. 164, no. 5, pp. 538–544, 2004.

[11] N. Genes, S. Violante, C. Cetrangol, L. Rogers, E. E. Schadt, and Y.-F. Y. Chan, "From smartphone to ehr: a case report on integrating patient-generated health data," *NPJ digital medicine*, vol. 1, no. 1, pp. 1–6, 2018.

[12] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *The Lancet infectious diseases*, vol. 20, no. 5, pp. 533–534, 2020.

[13] S. Richardson, J. S. Hirsch, M. Narasimhan, J. M. Crawford, T. McGinn, K. W. Davidson, D. P. Barnaby, L. B. Becker, J. D. Chelico, S. L. Cohen *et al.*, "Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with covid-19 in the new york city area," *Jama*, vol. 323, no. 20, pp. 2052–2059, 2020.

[14] G. Suleyman, R. A. Fadel, K. M. Malette, C. Hammond, H. Abdulla, A. Entz, Z. Demertzis, Z. Hanna, A. Failla, C. Dagher *et al.*, "Clinical characteristics and morbidity associated with coronavirus disease 2019 in a series of patients in metropolitan detroit," *JAMA network open*, vol. 3, no. 6, pp. e2 012 270–e2 012 270, 2020.

[15] G. L. Anesi, S. D. Halpern, and M. K. Delgado, "Covid-19 related hospital admissions in the united states: needs and outcomes," 2020.

[16] E. J. Williamson, A. J. Walker, K. Bhaskaran, S. Bacon, C. Bates, C. E. Morton, H. J. Curtis, A. Mehrkar, D. Evans, P. Inglesby *et al.*, "Opensafely: factors associated with covid-19 death in 17 million patients." *Nature*, 2020.

[17] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. Walker, H. Fu *et al.*, "Estimates of the severity of coronavirus disease 2019: a model-based analysis," *The Lancet infectious diseases*,

vol. 20, no. 6, pp. 669–677, 2020.

[18] G. Onder, G. Rezza, and S. Brusaferro, "Case-fatality rate and characteristics of patients dying in relation to covid-19 in italy," *Jama*, vol. 323, no. 18, pp. 1775–1776, 2020.

[19] F. Chen, W. Sun, S. Sun, Z. Li, Z. Wang, and L. Yu, "Clinical characteristics and risk factors for mortality among inpatients with covid-19 in wuhan, china," *Clinical and translational medicine*, 2020.

[20] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu *et al.*, "Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study," *The lancet*, vol. 395, no. 10229, pp. 1054–1062, 2020.

[21] R. K. Wadhera, P. Wadhera, P. Gaba, J. F. Figueroa, K. E. J. Maddox, R. W. Yeh, and C. Shen, "Variation in covid-19 hospitalizations and deaths across new york city boroughs," *Jama*, vol. 323, no. 21, pp. 2192–2195, 2020.

[22] C. C.-. R. Team, C. C.-. R. Team, C. C.-. R. Team, S. Bialek, V. Bowen, N. Chow, A. Curns, R. Gierke, A. Hall, M. Hughes *et al.*, "Geographic differences in covid-19 cases, deaths, and incidence–united states, february 12–april 7, 2020," *Morbidity and Mortality Weekly Report*, vol. 69, no. 15, pp. 465–471, 2020.

[23] Yale New Haven Health, "Facts and figures," https://www.ynhh.org/ynhhs/about/corporate-overview/system-statistics, 2020, accessed: 2021-4-23.

[24] OHSDI, "OMOP common data model," https://www.ohdsi.org/data-standardization/the-common-data-model/, 2019, accessed: 2021-4-23.

[25] J. McPadden, T. J. Durant, D. R. Bunch, A. Coppi, N. Price, K. Rodgerson, C. J. Torre Jr, W. Byron, A. L. Hsiao, H. M. Krumholz *et al.*, "Health care and precision medicine research: analysis of a scalable data science platform," *Journal of medical Internet research*, vol. 21, no. 4, p. e13043, 2019.

[26] W. L. Schulz, T. J. Durant, C. J. Torre Jr, A. L. Hsiao, and H. M. Krumholz, "Agile

health care analytics: Enabling real-time disease surveillance with a computational health platform," *Journal of Medical Internet Research*, vol. 22, no. 5, p. e18707, 2020.

[27] V. Ogievetsky, J. Heer, and J. Bostock, "D3 data-driven documents," *IEEE Trans. Vis. Comput. Graph*, vol. 17, no. 12, pp. 2301–2309, 2011.

[28] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," *Medical care*, pp. 8–27, 1998.

[29] A. Gasparini, "comorbidity: An r package for computing comorbidity scores," *Journal of Open Source Software*, vol. 3, no. 23, p. 648, 2018.

[30] B. J. Moore, S. White, R. Washington, N. Coenen, and A. Elixhauser, "Identifying increased risk of readmission and in-hospital mortality using hospital administrative data," *Medical care*, vol. 55, no. 7, pp. 698–705, 2017.

[31] C. van Walraven, P. C. Austin, A. Jennings, H. Quan, and A. J. Forster, "A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data," *Medical care*, pp. 626–633, 2009.

[32] R. J. Klein, *Age adjustment using the 2000 projected US population*. Department of Health & Human Services, Centers for Disease Control, 2001, no. 20.

[33] U.S. Census Bureau, "U.s. census bureau quickfacts: Connecticut," https://www.census.gov/quickfacts/CT, 2020, accessed: 2021-4-23.

[34] C. M. Petrilli, S. A. Jones, J. Yang, H. Rajagopalan, L. O'Donnell, Y. Chernyak, K. A. Tobin, R. J. Cerfolio, F. Francois, and L. I. Horwitz, "Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in new york city: prospective cohort study," *Bmj*, vol. 369, 2020.

[35] J.-M. Jin, P. Bai, W. He, F. Wu, X.-F. Liu, D.-M. Han, S. Liu, and J.-K. Yang, "Gender differences in patients with covid-19: focus on severity and mortality," *Frontiers in public health*, vol. 8, p. 152, 2020.

[36] A. Di Castelnuovo, M. Bonaccio, S. Costanzo, A. Gialluisi, A. Antinori, N. Berselli, L. Blandi, R. Bruno, R. Cauda, G. Guaraldi *et al.*, "Common cardiovascular risk

factors and in-hospital mortality in 3,894 patients with covid-19: survival analysis and machine learning-based findings from the multicentre italian corist study," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 30, no. 11, pp. 1899–1913, 2020.

[37] T. Takahashi, M. K. Ellingson, P. Wong, B. Israelow, C. Lucas, J. Klein, J. Silva, T. Mao, J. E. Oh, M. Tokuyama *et al.*, "Sex differences in immune responses that underlie covid-19 disease outcomes," *Nature*, vol. 588, no. 7837, pp. 315–320, 2020.

[38] J. Chen, W. J. Kelley, and D. R. Goldstein, "Role of aging and the immune response to respiratory viral infections: potential implications for covid-19," *The Journal of Immunology*, vol. 205, no. 2, pp. 313–320, 2020.

[39] V. Abedi, O. Olulana, V. Avula, D. Chaudhary, A. Khan, S. Shahjouei, J. Li, and R. Zand, "Racial, economic, and health inequality and covid-19 infection in the united states," *Journal of racial and ethnic health disparities*, pp. 1–11, 2020.

[40] C. P. Gross, U. R. Essien, S. Pasha, J. R. Gross, S.-y. Wang, and M. Nunez-Smith, "Racial and ethnic disparities in population-level covid-19 mortality," *Journal of general internal medicine*, vol. 35, no. 10, pp. 3097–3099, 2020.

[41] E. G. Price-Haywood, J. Burton, D. Fort, and L. Seoane, "Hospitalization and mortality among black patients and white patients with covid-19," *New England Journal of Medicine*, vol. 382, no. 26, pp. 2534–2543, 2020.

[42] K. M. Azar, Z. Shen, R. J. Romanelli, S. H. Lockhart, K. Smits, S. Robinson, S. Brown, and A. R. Pressman, "Disparities in outcomes among covid-19 patients in a large health care system in california: Study estimates the covid-19 infection fatality rate at the us county level." *Health Affairs*, vol. 39, no. 7, pp. 1253–1262, 2020.

[43] D. A. Kass, P. Duggal, and O. Cingolani, "Obesity could shift severe covid-19 disease to younger ages," *Lancet (London, England)*, 2020.

[44] C. Gao, Y. Cai, K. Zhang, L. Zhou, Y. Zhang, X. Zhang, Q. Li, W. Li, S. Yang, X. Zhao *et al.*, "Association of hypertension and antihypertensive treatment with covid-19 mortality: a retrospective observational study," *European heart journal*, vol. 41,

no. 22, pp. 2058–2066, 2020.

[45] M. P. Lin, O. Baker, L. D. Richardson, and J. D. Schuur, "Trends in emergency department visits and admission rates among US acute care hospitals," *JAMA Intern Med*, vol. 178, no. 12, pp. 1708–1710, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30326057

[46] N. Farrohknia, M. Castren, A. Ehrenberg, L. Lind, S. Oredsson, H. Jonsson, K. Asplund, and K. E. Goransson, "Emergency department triage scales and their components: a systematic review of the scientific evidence," *Scand J Trauma Resusc Emerg Med*, vol. 19, no. 1, p. 42, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21718476

[47] P. Tanabe, R. Gimbel, P. R. Yarnold, D. N. Kyriacou, and J. G. Adams, "Reliability and validity of scores on the emergency severity index version 3," *Acad Emerg Med*, vol. 11, no. 1, pp. 59–65, 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/14709429

[48] W. S. Hong, A. D. Haimovich, and R. A. Taylor, "Predicting hospital admission at emergency department triage using machine learning," *PloS one*, vol. 13, no. 7, p. e0201016, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30028888

[49] J. M. Kwon, Y. Lee, Y. Lee, S. Lee, H. Park, and J. Park, "Validation of deep-learning-based triage and acuity score using a large national dataset," *PLoS One*, vol. 13, no. 10, p. e0205836, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30321231

[50] S. Levin, M. Toerper, E. Hamrock, J. S. Hinson, S. Barnes, H. Gardner, A. Dugas, B. Linton, T. Kirsch, and G. Kelen, "Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index," *Ann Emerg Med*, vol. 71, no. 5, pp. 565–574 e2, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28888332

[51] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P.

Sengupta, "Machine learning in cardiovascular medicine: are we there yet?" *Heart*, vol. 104, no. 14, pp. 1156–1164, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29352006

[52] T. Ahmad, L. H. Lund, P. Rao, R. Ghosh, P. Warier, B. Vaccaro, U. Dahlstrom, C. M. O'Connor, G. M. Felker, and N. R. Desai, "Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients," *J Am Heart Assoc*, vol. 7, no. 8, p. e008081, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29650709

[53] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data," *PLoS One*, vol. 8, no. 6, p. e66341, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23826094

[54] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[55] E.-a. D. Amir, K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan, and D. Pe'er, "visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia," *Nature Biotechnology*, vol. 31, p. 545, 2013. [Online]. Available: https://doi.org/10.1038/nbt.2594

[56] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2018.

[57] T. Ahmad, M. J. Pencina, P. J. Schulte, E. O'Brien, D. J. Whellan, I. L. Piña, D. W. Kitzman, K. L. Lee, C. M. O'Connor, and G. M. Felker, "Clinical implications of chronic heart failure phenotypes defined by cluster analysis," *Journal of the American College of Cardiology*, vol. 64, no. 17, pp. 1765–1774, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25443696

[58] C. Seymour, J. Kennedy, S. Wang, Z. Xu, C. Chang, Q. Mi, Y. Vodovotz, G. Clermont, S. Visweswaran, and J. Weiss, "Feasibility of sepsis phenotyping using electronic health record data during initial emergency department care," in *American Journal of Respiratory and Critical Care Medicine*, vol. 197, Amer Thoracic Soc 25 Broadway, 18 FL, New York, NY 10004 USA. Amer Thoracic Soc 25 Broadway, 18 FL, New York, NY 10004 USA, 2018, Conference Proceedings.

[59] B. K. Beaulieu-Jones, C. S. Greene, and A. L. S. C. T. C. Pooled Resource Open-Access, "Semi-supervised learning of the electronic health record for phenotype stratification," *J Biomed Inform*, vol. 64, pp. 168–178, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27744022

[60] J. C. Kirby, P. Speltz, L. V. Rasmussen, M. Basford, O. Gottesman, P. L. Peissig, J. A. Pacheco, G. Tromp, J. Pathak, D. S. Carrell, S. B. Ellis, T. Lingren, W. K. Thompson, G. Savova, J. Haines, D. M. Roden, P. A. Harris, and J. C. Denny, "Phekb: a catalog and workflow for creating electronic phenotype algorithms for transportability," *J Am Med Inform Assoc*, vol. 23, no. 6, pp. 1046–1052, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27026615

[61] Y. Wang, L. Luo, M. T. Freedman, and S. Y. Kung, "Probabilistic principal component subspaces: a hierarchical finite mixture model for data visualization," *IEEE Trans Neural Netw*, vol. 11, no. 3, pp. 625–36, 2000. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18249790

[62] K. Y. Yeung and W. L. Ruzzo, "An empirical study on principal component analysis for clustering gene expression data," *Department of Computer Science and Engineering, University of Washington*, 2000.

[63] K. A. Oetjen, K. E. Lindblad, M. Goswami, G. Gui, P. K. Dagur, C. Lai, L. W. Dillon, J. P. McCoy, and C. S. Hourigan, "Human bone marrow assessment by single-cell rna sequencing, mass cytometry, and flow cytometry," *JCI Insight*, vol. 3, no. 23, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30518681

[64] E. Becht, L. McInnes, J. Healy, C. A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using umap," *Nature Biotechnology*, vol. 37, no. 1, pp. 38–+, 2019. [Online]. Available: <GotoISI>://WOS:000454804600017

[65] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[66] K. Y. Yeung and W. L. Ruzzo, "Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[68] S. L. Cartwright and M. P. Knudson, "Evaluation of acute abdominal pain in adults." *American family physician*, vol. 77, no. 7, 2008.

[69] A. W. Group, "Problem list guidance in the ehr," *Journal of AHIMA*, vol. 82, no. 9, pp. 52–58, Sep. 2011.

[70] C. Holmes, "The problem list beyond meaningful use: Part i: The problems with problem lists," *Journal of AHIMA*, vol. 82, no. 2, pp. 30–33, Feb. 2011.

[71] C. Holmes, M. Brown, D. S. Hilaire, and A. Wright, "Healthcare provider attitudes towards the problem list in an electronic health record: a mixed-methods qualitative study," *BMC medical informatics and decision making*, vol. 12, pp. 127–127, Nov. 2012, 23140312[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/23140312

[72] S. Jain, R. Mohammadi, and B. C. Wallace, "An analysis of attention over clinical notes for predictive tasks," *CoRR*, vol. abs/1904.03244, 2019. [Online]. Available: http://arxiv.org/abs/1904.03244

[73] S. Khadanga, K. Aggarwal, S. Joty, and J. Srivastava, "Using clinical notes with time series data for icu management," 2019.

[74] J. Liu, Z. Zhang, and N. Razavian, "Deep ehr: Chronic disease prediction using medical notes," in *Proceedings of the 3rd Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., vol. 85. Palo Alto, California: PMLR, 17–18 Aug 2018, pp. 440–464. [Online]. Available: http://proceedings.mlr.press/v85/liu18b.html

[75] H. M. Krumholz, Y. Wang, J. A. Mattera, Y. Wang, L. F. Han, M. J. Ingber, S. Roman, and S.-L. T. Normand, "An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction," *Circulation*, vol. 113, no. 13, pp. 1683–1692, 2006. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.105.611186

[76] D. He, S. C. Mathews, A. N. Kalloo, and S. Hutfless, "Mining high-dimensional administrative claims data to predict early hospital readmissions," *Journal of the American Medical Informatics Association : JAMIA*, vol. 21, no. 2, pp. 272–279, 2014, 24076748[pmid]. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24076748

[77] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, "Unfolding physiological state: Mortality modelling in intensive care units," *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining*, vol. 2014, pp. 75–84, Aug. 2014, 25289175[pmid]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25289175

[78] A. Pakbin, P. Rafi, N. Hurley, W. Schulz, M. Harlan Krumholz, and J. Bobak Mortazavi, "Prediction of icu readmissions using data at patient discharge," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2018, pp. 4932–4935.

[79] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers

for medication recommendation," in *Proceedings of IJCAI*, 2019, pp. 5953–5959.

[80] S. Barbieri, J. Kemp, O. Perez-Concha, S. Kotwal, M. Gallagher, A. Ritchie, and L. Jorm, "Benchmarking deep learning architectures for predicting readmission to the icu and describing patients-at-risk," *Scientific Reports*, vol. 10, no. 1, p. 1111, 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-58053-z

[81] S. Blecker, S. D. Katz, L. I. Horwitz, G. Kuperman, H. Park, A. Gold, and D. Sontag, "Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data," *JAMA Cardiology*, vol. 1, no. 9, pp. 1014–1020, 12 2016. [Online]. Available: https://doi.org/10.1001/jamacardio.2016.3236

[82] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, "Multi-label classification of patient notes: Case study on icd code assignment," 2018. [Online]. Available: https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16881

[83] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, and J. Eisenstein, "Explainable prediction of medical codes from clinical text," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 1101–1111. [Online]. Available: http://aclweb.org/anthology/N18-1100

[84] N. Sadoughi, G. P. Finley, J. Fone, V. Murali, M. Korenevski, S. Baryshnikov, N. Axtmann, M. Miller, and D. Suendermann-Oeft, "Medical code prediction with multi-view convolution and description-regularized label-dependent attention," 11 2018.

[85] K. Xu, M. Lam, J. Pang, X. Gao, C. Band, P. Mathur, F. Papay, A. K. Khanna, J. B. Cywinski, K. Maheshwari, P. Xie, and E. P. Xing, "Multimodal machine learning for automated ICD coding," *CoRR*, vol. abs/1810.13348, 2018. [Online]. Available: http://arxiv.org/abs/1810.13348

[86] H. M. Krumholz, Z. Lin, P. S. Keenan, J. Chen, J. S. Ross, E. E. Drye, S. M. Bernheim, Y. Wang, E. H. Bradley, L. F. Han, and S.-L. T. Normand, "Relationship

Between Hospital Readmission and Mortality Rates for Patients Hospitalized With Acute Myocardial Infarction, Heart Failure, or Pneumonia," *JAMA*, vol. 309, no. 6, pp. 587–593, 02 2013. [Online]. Available: https://doi.org/10.1001/jama.2013.333

[87] S. Jain and B. C. Wallace, "Attention is not explanation," *CoRR*, vol. abs/1902.10186, 2019. [Online]. Available: http://arxiv.org/abs/1902.10186

[88] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," *CoRR*, vol. abs/1908.04626, 2019. [Online]. Available: http://arxiv.org/abs/1908.04626

[89] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, vol. abs/1310.4546, 2013. [Online]. Available: http://arxiv.org/abs/1310.4546

[90] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford, "PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations," *Bioinformatics*, vol. 26, no. 9, pp. 1205–1210, 03 2010. [Online]. Available: https://doi.org/10.1093/bioinformatics/btq126

[91] W.-Q. Wei, L. A. Bastarache, R. J. Carroll, J. E. Marlo, T. J. Osterman, E. R. Gamazon, N. J. Cox, D. M. Roden, and J. C. Denny, "Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record," *PLOS ONE*, vol. 12, no. 7, p. e0175508, Jul. 2017. [Online]. Available: https://doi.org/10.1371%2Fjournal.pone.0175508

[92] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[93] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[94] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Association for Computational Linguistics, 2014, pp. 1746–1751. [Online]. Available: http://aclweb.org/anthology/D14-1181

[95] T. E. Chang, J. H. Lichtman, L. B. Goldstein, and M. G. George, "Accuracy of icd-9-cm codes by hospital characteristics and stroke severity: Paul coverdell national acute stroke program," *Journal of the American Heart Association*, vol. 5, no. 6, p. e003056, May 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27247334

[96] C. Benesch, D. M. Witter, A. L. Wilder, P. W. Duncan, G. P. Samsa, and D. B. Matchar, "Inaccuracy of the international classification of diseases (icd-9-cm) in identifying the diagnosis of ischemic cerebrovascular disease," *Neurology*, vol. 49, no. 3, pp. 660–664, 1997. [Online]. Available: https://n.neurology.org/content/49/3/660

[97] E. Birman-Deych, A. D. Waterman, Y. Yan, D. S. Nilasena, M. J. Radford, and B. F. Gage, "Accuracy of icd-9-cm codes for identifying cardiovascular and stroke risk factors," *Medical Care*, vol. 43, no. 5, pp. 480–485, 2005. [Online]. Available: http://www.jstor.org/stable/3768402

[98] H. Ellekjær, J. Holmen, O. Krüger, and A. Terent, "Identification of incident stroke in norway," *Stroke*, vol. 30, no. 1, pp. 56–60, 1999. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/01.STR.30.1.56

[99] E. S. Fisher, F. S. Whaley, W. M. Krushat, D. J. Malenka, C. Fleming, J. A. Baron, and D. C. Hsia, "The accuracy of medicare's hospital claims data: progress has been made, but problems remain." *American Journal of Public Health*, vol. 82, no. 2, pp. 243–248, 1992, pMID: 1739155. [Online]. Available: https://doi.org/10.2105/AJPH.82.2.243

[100] S. R. Heckbert, C. Kooperberg, M. M. Safford, B. M. Psaty, J. Hsia, A. McTiernan, J. M. Gaziano, W. H. Frishman, and J. D. Curb, "Comparison of Self-Report, Hospital Discharge Codes, and Adjudication of Cardiovascular Events in the Women's Health Initiative," *American Journal of Epidemiology*, vol. 160, no. 12, pp. 1152–1158, 12 2004. [Online]. Available: https://doi.org/10.1093/aje/kwh314

[101] S. A. Jones, R. F. Gottesman, E. Shahar, L. Wruck, and W. D. Rosamond, "Validity of hospital discharge diagnosis codes for stroke," *Stroke*, vol. 45, no. 11, pp. 3219–3225, 2014. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.114.006316

[102] H. Kumamaru, S. E. Judd, J. R. Curtis, R. Ramachandran, N. C. Hardy, J. D. Rhodes, M. M. Safford, B. M. Kissela, G. Howard, J. J. Jalbert, T. G. Brott, and S. Setoguchi, "Validity of claims-based stroke algorithms in contemporary medicare data," *Circulation: Cardiovascular Quality and Outcomes*, vol. 7, no. 4, pp. 611–619, 2014. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/CIRCOUTCOMES.113.000743

[103] K. Lakshminarayan, J. C. Larson, B. Virnig, C. Fuller, N. B. Allen, M. Limacher, W. C. Winkelmayer, M. M. Safford, and D. R. Burwen, "Comparison of medicare claims versus physician adjudication for identifying stroke outcomes in the women’s health initiative," *Stroke*, vol. 45, no. 3, pp. 815–821, 2014. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/STROKEAHA.113.003408

[104] V. Rodriguez and A. Perotte, "Phenotype inference with semi-supervised mixed membership models," 2018.

[105] S. Tonekaboni, S. Joshi, M. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," 05 2019.

[106] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *CoRR*, vol. abs/1502.03044, 2015. [Online]. Available: http://arxiv.org/abs/1502.03044

[107] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *CoRR*, vol. abs/1808.00033, 2018. [Online]. Available: http://arxiv.org/abs/1808.00033

[108] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[109] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. [Online]. Available: https://www.aclweb.org/anthology/W19-1909

[110] K. A. Fox, O. H. Dabbous, R. J. Goldberg, K. S. Pieper, K. A. Eagle, F. Van de Werf, Á. Avezum, S. G. Goodman, M. D. Flather, F. A. Anderson *et al.*, "Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (grace)," *bmj*, vol. 333, no. 7578, p. 1091, 2006.

[111] C. B. Granger, R. J. Goldberg, O. Dabbous, K. S. Pieper, K. A. Eagle, C. P. Cannon, F. Van de Werf, A. Avezum, S. G. Goodman, M. D. Flather *et al.*, "Predictors of hospital mortality in the global registry of acute coronary events," *Archives of internal medicine*, vol. 163, no. 19, pp. 2345–2353, 2003.

[112] E. M. Antman, M. Cohen, P. J. Bernink, C. H. McCabe, T. Horacek, G. Papuchis, B. Mautner, R. Corbalan, D. Radley, and E. Braunwald, "The timi risk score for unstable angina/non–st elevation mi: a method for prognostication and therapeutic decision making," *Jama*, vol. 284, no. 7, pp. 835–842, 2000.

[113] A. D. Souza and H. S. Migon, "Bayesian binary regression model: an application to in-hospital death after ami prediction," *Pesquisa Operacional*, vol. 24, no. 2, pp. 253–267, 2004.

[114] M. Zoni-Berisso, D. Molini, S. Viani, G. S. Mela, and L. Delfino, "Noninvasive prediction of sudden death and sustained ventricular tachycardia after acute myocardial

infarction using a neural network algorithm," *Italian Heart Journal*, vol. 2, pp. 612–620, 2001.

[115] X. Li, H. Liu, J. Yang, G. Xie, M. Xu, and Y. Yang, "Using machine learning models to predict in-hospital mortality for st-elevation myocardial infarction patients." *Studies in health technology and informatics*, vol. 245, pp. 476–480, 2017.

[116] M. D. Samad, A. Ulloa, G. J. Wehner, L. Jing, D. Hartzel, C. W. Good, B. A. Williams, C. M. Haggerty, and B. K. Fornwalt, "Predicting survival from large echocardiography and electronic health record datasets: optimization with machine learning," *JACC: Cardiovascular Imaging*, vol. 12, no. 4, pp. 681–689, 2019.

[117] I. Yosefian, E. Mosa Farkhani, and M. R. Baneshi, "Application of random forest survival models to increase generalizability of decision trees: a case study in acute myocardial infarction," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[118] P. D. Myers, B. M. Scirica, and C. M. Stultz, "Machine learning improves risk stratification after acute coronary syndrome," *Scientific reports*, vol. 7, no. 1, pp. 1–12, 2017.

[119] H. Mansoor, I. Y. Elgendy, R. Segal, A. A. Bavry, and J. Bian, "Risk prediction model for in-hospital mortality in women with st-elevation myocardial infarction: a machine learning approach," *Heart & Lung*, vol. 46, no. 6, pp. 405–411, 2017.

[120] P. C. Austin, "A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting ami mortality," *Statistics in medicine*, vol. 26, no. 15, pp. 2937–2957, 2007.

[121] R. Shouval, A. Hadanny, N. Shlomo, Z. Iakobishvili, R. Unger, D. Zahger, R. Alcalai, S. Atar, S. Gottlieb, S. Matetzky *et al.*, "Machine learning for prediction of 30-day mortality after st elevation myocardial infraction: An acute coronary syndrome israeli survey data mining study," *International journal of cardiology*, vol. 246, pp. 7–13, 2017.

[122] R. Bigi, A. Mafrici, P. Colombo, D. Gregori, E. Corrada, A. Alberti, A. De Biase,

P. S. Orrego, C. Fiorentini, and S. Klugmann, "Relation of terminal qrs distortion to left ventricular functional recovery and remodeling in acute myocardial infarction treated with primary angioplasty," *The American journal of cardiology*, vol. 96, no. 9, pp. 1233–1236, 2005.

[123] R. Bigi, D. Gregori, L. Cortigiani, A. Desideri, F. A. Chiarotto, and G. M. Toffolo, "Artificial neural networks and robust bayesian classifiers for risk stratification following uncomplicated myocardial infarction," *International journal of cardiology*, vol. 101, no. 3, pp. 481–487, 2005.

[124] D. Zhang, X. Song, S. Lv, D. Li, S. Yan, and M. Zhang, "Predicting coronary no-reflow in patients with acute st-segment elevation myocardial infarction using bayesian approaches," *Coronary artery disease*, vol. 25, no. 7, p. 582, 2014.

[125] B. J. Mortazavi, N. S. Downing, E. M. Bucholz, K. Dharmarajan, A. Manhapra, S.-X. Li, S. N. Negahban, and H. M. Krumholz, "Analysis of machine learning techniques for heart failure readmissions," *Circulation: Cardiovascular Quality and Outcomes*, vol. 9, no. 6, pp. 629–640, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28263938

[126] L. Breiman *et al.*, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Statistical science*, vol. 16, no. 3, pp. 199–231, 2001.

[127] R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger, and A. Nagler, "Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in sct," *Bone marrow transplantation*, vol. 49, no. 3, pp. 332–337, 2014.

[128] E. Von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke, "The strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines for reporting observational studies," *Annals of internal medicine*, vol. 147, no. 8, pp. 573–577, 2007.

[129] J. C. Messenger, K. K. Ho, C. H. Young, L. E. Slattery, J. C. Draoui, J. P. Curtis, G. J. Dehmer, F. L. Grover, M. J. Mirro, M. R. Reynolds *et al.*, "The national cardiovascular

data registry (ncdr) data quality brief: the ncdr data quality program in 2012," *Journal of the American College of Cardiology*, vol. 60, no. 16, pp. 1484–1488, 2012.

[130] E. D. Peterson, D. Dai, E. R. DeLong, J. M. Brennan, M. Singh, S. V. Rao, R. E. Shaw, M. T. Roe, K. K. Ho, L. W. Klein *et al.*, "Contemporary mortality risk prediction for percutaneous coronary intervention: results from 588,398 procedures in the national cardiovascular data registry," *Journal of the American College of Cardiology*, vol. 55, no. 18, pp. 1923–1932, 2010.

[131] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, "Xgboost: Extreme gradient boosting (r package version 0.6. 4.1)[computer software]," 2018.

[132] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.

[133] S. Siegert, "Specsverification: forecast verification routines for ensemble forecasts of weather and climate. r package version 0.5–2. 2017."

[134] C. G. Walsh, K. Sharman, and G. Hripcsak, "Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk," *Journal of biomedical informatics*, vol. 76, pp. 9–18, 2017.

[135] D. E. Leisman, "Rare events in the icu: an emerging challenge in classification and prediction," *Read Online: Critical Care Medicine; Society of Critical Care Medicine*, vol. 46, no. 3, pp. 418–424, 2018.

[136] C. D. Galloway, A. V. Valys, J. B. Shreibati, D. L. Treiman, F. L. Petterson, V. P. Gundotra, D. E. Albert, Z. I. Attia, R. E. Carter, S. J. Asirvatham *et al.*, "Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram," *JAMA cardiology*, vol. 4, no. 5, pp. 428–436, 2019.

[137] Z. I. Attia, P. A. Noseworthy, F. Lopez-Jimenez, S. J. Asirvatham, A. J. Deshmukh, B. J. Gersh, R. E. Carter, X. Yao, A. A. Rabinstein, B. J. Erickson *et al.*, "An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction," *The*

Lancet, vol. 394, no. 10201, pp. 861–867, 2019.

[138] Z. I. Attia, S. Kapa, F. Lopez-Jimenez, P. M. McKie, D. J. Ladewig, G. Satam, P. A. Pellikka, M. Enriquez-Sarano, P. A. Noseworthy, T. M. Munger *et al.*, "Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram," *Nature medicine*, vol. 25, no. 1, pp. 70–74, 2019.

[139] D. S. Kazi and K. Bibbins-Domingo, "Accurately predicting cardiovascular risk–and acting on it," *Annals of internal medicine*, vol. 172, no. 1, pp. 61–62, 2020.

[140] J. A. Dodson, A. M. Hajduk, M. Geda, H. M. Krumholz, T. E. Murphy, S. Tsang, M. E. Tinetti, M. G. Nanna, R. McNamara, T. M. Gill *et al.*, "Predicting 6-month mortality for older adults hospitalized with acute myocardial infarction: a cohort study," *Annals of internal medicine*, vol. 172, no. 1, pp. 12–21, 2020.

[141] P. T. O'gara, F. G. Kushner, D. D. Ascheim, D. E. Casey, M. K. Chung, J. A. De Lemos, S. M. Ettinger, J. C. Fang, F. M. Fesmire, B. A. Franklin *et al.*, "2013 accf/aha guideline for the management of st-elevation myocardial infarction: a report of the american college of cardiology foundation/american heart association task force on practice guidelines," *Journal of the American college of cardiology*, vol. 61, no. 4, pp. e78–e140, 2013.

[142] H. Thiele, U. Zeymer, F.-J. Neumann, M. Ferenc, H.-G. Olbrich, J. Hausleiter, G. Richardt, M. Hennersdorf, K. Empen, G. Fuernau *et al.*, "Intraaortic balloon support for myocardial infarction with cardiogenic shock," *New England Journal of Medicine*, vol. 367, no. 14, pp. 1287–1296, 2012.

[143] H. Thiele, U. Zeymer, F.-J. Neumann, M. Ferenc, H.-G. Olbrich, J. Hausleiter, A. de Waha, G. Richardt, M. Hennersdorf, K. Empen *et al.*, "Intra-aortic balloon counterpulsation in acute myocardial infarction complicated by cardiogenic shock (iabp-shock ii): final 12 month results of a randomised, open-label trial," *The Lancet*, vol. 382, no. 9905, pp. 1638–1645, 2013.

[144] S. Unverzagt, M. Buerke, A. de Waha, J. Haerting, D. Pietzner, M. Seyfarth, H. Thiele,

K. Werdan, U. Zeymer, and R. Prondzinsky, "Intra-aortic balloon pump counterpulsation (iabp) for myocardial infarction complicated by cardiogenic shock," *Cochrane Database of Systematic Reviews*, no. 3, 2015.

[145] Y. Ahmad, S. Sen, M. J. Shun-Shin, J. Ouyang, J. A. Finegold, R. K. Al-Lamee, J. E. Davies, G. D. Cole, and D. P. Francis, "Intra-aortic balloon pump therapy for acute myocardial infarction: a meta-analysis," *JAMA internal medicine*, vol. 175, no. 6, pp. 931–939, 2015.

[146] M. Seyfarth, D. Sibbing, I. Bauer, G. Fröhlich, L. Bott-Flügel, R. Byrne, J. Dirschinger, A. Kastrati, and A. Schömig, "A randomized clinical trial to evaluate the safety and efficacy of a percutaneous left ventricular assist device versus intra-aortic balloon pumping for treatment of cardiogenic shock caused by myocardial infarction," *Journal of the American College of Cardiology*, vol. 52, no. 19, pp. 1584–1588, 2008.

[147] V. K. Rathi, A. S. Kesselheim, and J. S. Ross, "The us food and drug administration 515 program initiative: addressing the evidence gap for widely used, high-risk cardiovascular devices?" *JAMA cardiology*, vol. 1, no. 2, pp. 117–118, 2016.

[148] R. Khera, P. Cram, X. Lu, A. Vyas, A. Gerke, G. E. Rosenthal, P. A. Horwitz, and S. Girotra, "Trends in the use of percutaneous ventricular assist devices: analysis of national inpatient sample data, 2007 through 2012," *JAMA internal medicine*, vol. 175, no. 6, pp. 941–950, 2015.

[149] A. Sandhu, L. A. McCoy, S. I. Negi, I. Hameed, P. Atri, S. J. Al'Aref, J. Curtis, E. McNulty, H. V. Anderson, A. Shroff *et al.*, "Use of mechanical circulatory support in patients undergoing percutaneous coronary intervention: insights from the national cardiovascular data registry," *Circulation*, vol. 132, no. 13, pp. 1243–1251, 2015.

[150] D. M. Ouweneel, E. Eriksen, K. D. Sjauw, I. M. van Dongen, A. Hirsch, E. J. Packer, M. M. Vis, J. J. Wykrzykowska, K. T. Koch, J. Baan *et al.*, "Percutaneous mechanical circulatory support versus intra-aortic balloon pump in cardiogenic shock after acute myocardial infarction," *Journal of the American College of Cardiology*, vol. 69, no. 3,

pp. 278–287, 2017.

[151] S. Van Diepen, J. N. Katz, N. M. Albert, T. D. Henry, A. K. Jacobs, N. K. Kapur, A. Kilic, V. Menon, E. M. Ohman, N. K. Sweitzer *et al.*, "Contemporary management of cardiogenic shock: a scientific statement from the american heart association," *Circulation*, vol. 136, no. 16, pp. e232–e268, 2017.

[152] A. P. Amin, J. A. Spertus, J. P. Curtis, N. Desai, F. A. Masoudi, R. G. Bach, C. McNeely, F. Al-Badarin, J. A. House, H. Kulkarni *et al.*, "The evolving landscape of impella use in the united states among patients undergoing percutaneous coronary intervention with mechanical circulatory support," *Circulation*, vol. 141, no. 4, pp. 273–284, 2020.

[153] S. S. Dhruva, J. S. Ross, B. J. Mortazavi, N. C. Hurley, H. M. Krumholz, J. P. Curtis, A. Berkowitz, F. A. Masoudi, J. C. Messenger, C. S. Parzynski *et al.*, "Association of use of an intravascular microaxial left ventricular assist device vs intra-aortic balloon pump with in-hospital mortality and major bleeding among patients with acute myocardial infarction complicated by cardiogenic shock," *Jama*, vol. 323, no. 8, pp. 734–745, 2020.

[154] R. G. Brindis, S. Fitzgerald, H. V. Anderson, R. E. Shaw, W. S. Weintraub, and J. F. Williams, "The american college of cardiology-national cardiovascular data registry™(acc-ncdr™): building a national clinical data repository," *Journal of the American College of Cardiology*, vol. 37, no. 8, pp. 2240–2245, 2001.

[155] E. D. Peterson, M. T. Roe, J. S. Rumsfeld, R. E. Shaw, R. G. Brindis, G. C. Fonarow, and C. P. Cannon, "A call to action (acute coronary treatment and intervention outcomes network) a national effort to promote timely clinical feedback and support continuous quality improvement for acute myocardial infarction," *Circulation: Cardiovascular Quality and Outcomes*, vol. 2, no. 5, pp. 491–499, 2009.

[156] C. P. Cannon, A. Battler, R. G. Brindis, J. L. Cox, S. G. Ellis, N. R. Every, J. T. Flaherty, R. A. Harrington, H. M. Krumholz, M. L. Simoons *et al.*, "American college of cardiology key data elements and definitions for measuring the clinical management

and outcomes of patients with acute coronary syndromes: a report of the american college of cardiology task force on clinical data standards (acute coronary syndromes writing committee) endorsed by the american association of cardiovascular and pulmonary rehabilitation, american college of emergency physicians, american heart association, cardiac society of australia & new zealand, national heart foundation of australia, society for cardiac angiography and interventions, and the taiwan society of cardiology," *Journal of the American College of Cardiology*, vol. 38, no. 7, pp. 2114–2130, 2001.

[157] K. Larsen and J. Merlo, "Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression," *American journal of epidemiology*, vol. 161, no. 1, pp. 81–88, 2005.

[158] J. E. Pustejovsky and E. Tipton, "Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models," *Journal of Business & Economic Statistics*, vol. 36, no. 4, pp. 672–683, 2018.

[159] H. Wickham and W. Chang, "ggplot2," *Computer software]. Retrieved from http://ggplot2. org*, 2012.

[160] A. Signorell, K. Aho, N. Anderegg, T. Aragon, A. Arppe, A. Baddeley, B. Bolker, F. Caeiro, S. Champely, D. Chessel *et al.*, "Desctools: tools for descriptive statistics. 2019," *R package version 0.99*, vol. 24, 2019.

[161] G. Grolemund, H. Wickham *et al.*, "Dates and times made easy with lubridate," *Journal of statistical software*, vol. 40, no. 3, pp. 1–25, 2011.

[162] B. Ibanez, S. James, S. Agewall, M. J. Antunes, C. Bucciarelli-Ducci, H. Bueno, A. L. Caforio, F. Crea, J. A. Goudevenos, S. Halvorsen *et al.*, "2017 esc guidelines for the management of acute myocardial infarction in patients presenting with st-segment elevation: The task force for the management of acute myocardial infarction in patients presenting with st-segment elevation of the european society of cardiology (esc)," *European heart journal*, vol. 39, no. 2, pp. 119–177, 2018.

[163] T. M. Atkinson, E. M. Ohman, W. W. O'Neill, T. Rab, J. E. Cigarroa, and I. S. C.

of the American College of Cardiology, "A practical approach to mechanical circulatory support in patients undergoing percutaneous coronary intervention: an interventional perspective," *JACC: Cardiovascular Interventions*, vol. 9, no. 9, pp. 871–883, 2016.

[164] J. B. Strom, Y. Zhao, C. Shen, M. Chung, D. S. Pinto, J. J. Popma, D. J. Cohen, and R. W. Yeh, "Hospital variation in the utilization of short-term nondurable mechanical circulatory support in myocardial infarction complicated by cardiogenic shock," *Circulation: Cardiovascular Interventions*, vol. 12, no. 1, p. e007270, 2019.

[165] D. D. Berg, C. F. Barnett, B. B. Kenigsberg, A. Papolos, C. L. Alviar, V. M. Baird-Zars, G. W. Barsness, E. A. Bohula, J. Brennan, J. A. Burke *et al.*, "Clinical practice patterns in temporary mechanical circulatory support for shock in the critical care cardiology trials network (ccctn) registry," *Circulation: Heart Failure*, vol. 12, no. 11, p. e006635, 2019.

[166] H. Thiele, S. Desch, and A. Freund, "Microaxial left ventricular assist devices: In search of an appropriate indication," *Jama*, vol. 323, no. 8, pp. 716–718, 2020.

[167] J. B. Strom, Y. Zhao, C. Shen, M. Chung, D. S. Pinto, J. J. Popma, and R. W. Yeh, "National trends, predictors of use, and in-hospital outcomes in mechanical circulatory support for cardiogenic shock." *EuroIntervention: journal of EuroPCR in collaboration with the Working Group on Interventional Cardiology of the European Society of Cardiology*, vol. 13, no. 18, pp. e2152–e2159, 2018.

[168] D. A. Baran, C. L. Grines, S. Bailey, D. Burkhoff, S. A. Hall, T. D. Henry, S. M. Hollenberg, N. K. Kapur, W. O'Neill, J. P. Ornato *et al.*, "Scai clinical expert consensus statement on the classification of cardiogenic shock: this document was endorsed by the american college of cardiology (acc), the american heart association (aha), the society of critical care medicine (sccm), and the society of thoracic surgeons (sts) in april 2019," *Catheterization and Cardiovascular Interventions*, vol. 94, no. 1, pp. 29–37, 2019.

[169] C. S. Rihal, S. S. Naidu, M. M. Givertz, W. Y. Szeto, J. A. Burke, N. K. Kapur,

M. Kern, K. N. Garratt, J. A. Goldstein, V. Dimas *et al.*, "2015 scai/acc/hfsa/sts clinical expert consensus statement on the use of percutaneous mechanical circulatory support devices in cardiovascular care: endorsed by the american heart assocation, the cardiological society of india, and sociedad latino americana de cardiologia intervencion; affirmation of value by the canadian association of interventional cardiology-association canadienne de cardiologie d'intervention," *Journal of the American College of Cardiology*, vol. 65, no. 19, pp. e7–e26, 2015.

[170] G. D'Onofrio, B. Safdar, J. H. Lichtman, K. M. Strait, R. P. Dreyer, M. Geda, J. A. Spertus, and H. M. Krumholz, "Sex differences in reperfusion in young patients with st-segment–elevation myocardial infarction: results from the virgo study," *Circulation*, vol. 131, no. 15, pp. 1324–1332, 2015.

[171] A. F. Hernandez, G. C. Fonarow, L. Liang, S. M. Al-Khatib, L. H. Curtis, K. A. LaBresh, C. W. Yancy, N. M. Albert, and E. D. Peterson, "Sex and racial differences in the use of implantable cardioverter-defibrillators among patients hospitalized with heart failure," *Jama*, vol. 298, no. 13, pp. 1525–1532, 2007.

[172] B. Ahmed and H. L. Dauerman, "Women, bleeding, and coronary intervention," *Circulation*, vol. 127, no. 5, pp. 641–649, 2013.

[173] B. Redfors, O. Angerås, T. Råmunddal, C. Dworeck, I. Haraldsson, D. Ioanes, P. Petursson, B. Libungan, J. Odenstedt, J. Stewart *et al.*, "17-year trends in incidence and prognosis of cardiogenic shock in patients with acute myocardial infarction in western sweden," *International journal of cardiology*, vol. 185, pp. 256–262, 2015.

[174] R. V. Jeger, D. Radovanovic, P. R. Hunziker, M. E. Pfisterer, J.-C. Stauffer, P. Erne, and P. Urban, "Ten-year trends in the incidence and treatment of cardiogenic shock," *Annals of internal medicine*, vol. 149, no. 9, pp. 618–626, 2008.

[175] D. Kolte, S. Khera, W. S. Aronow, M. Mujib, C. Palaniswamy, S. Sule, D. Jain, W. Gotsis, A. Ahmed, W. H. Frishman *et al.*, "Trends in incidence, management, and outcomes of cardiogenic shock complicating st-elevation myocardial infarction in the

u nited s tates," *Journal of the American Heart Association*, vol. 3, no. 1, p. e000590, 2014.

[176] US Food and Drug Administration, "Summary of safety and effectiveness data (ssed): Impella ventricular support systems," https://www.accessdata.fda.gov/cdrh_docs/pdf14/P140003S004B.pdf, 2016, accessed: 2021-4-23.

[177] B. Schrage, K. Ibrahim, T. Loehn, N. Werner, J.-M. Sinning, F. Pappalardo, M. Pieri, C. Skurk, A. Lauten, U. Landmesser *et al.*, "Impella support for acute myocardial infarction complicated by cardiogenic shock: Matched-pair iabp-shock ii trial 30-day mortality analysis," *Circulation*, vol. 139, no. 10, pp. 1249–1258, 2019.

[178] J. M. Brennan, E. D. Peterson, J. C. Messenger, J. S. Rumsfeld, W. S. Weintraub, K. J. Anstrom, E. L. Eisenstein, S. Milford-Beland, M. V. Grau-Sepulveda, M. E. Booth *et al.*, "Linking the national cardiovascular data registry cathpci registry with medicare claims data: validation of a longitudinal cohort of elderly patients undergoing cardiac catheterization," *Circulation: Cardiovascular Quality and Outcomes*, vol. 5, no. 1, pp. 134–140, 2012.

[179] American College of Cardiology National Cardiovascular Data Registry, "What each registry collects," https://cvquality.acc.org/NCDR-Home/Data-Collection/What-Each-Registry-Collects, accessed: 2021-4-23.

[180] L. Mauri, T. S. Silbaugh, P. Garg, R. E. Wolf, K. Zelevinsky, A. Lovett, M. R. Varma, Z. Zhou, and S.-L. T. Normand, "Drug-eluting or bare-metal stents for acute myocardial infarction," *New England Journal of Medicine*, vol. 359, no. 13, pp. 1330–1342, 2008.

[181] X. S. Gu and P. R. Rosenbaum, "Comparison of multivariate matching methods: Structures, distances, and algorithms," *Journal of Computational and Graphical Statistics*, vol. 2, no. 4, pp. 405–420, 1993.

[182] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "proc: an open-source package for r and s+ to analyze and compare roc curves," *BMC Bioinformatics*, vol. 12, no. 1, p. 77, Mar. 2011. [Online]. Available:

https://doi.org/10.1186/1471-2105-12-77

[183] US Food and Drug Administration, "Update: increased rate of mortality in patients receiving abiomed impella rp system–letter to health care providers," https://www.fda.gov/medical-devices/letters-health-care-providers/update-increased-rate-mortality-patients-receiving-abiomed-impella-rp-system-letter-health-care, 2019, accessed: 2021-4-23.

[184] B. Wernly, C. Seelmaier, D. Leistner, B. E. Stähli, I. Pretsch, M. Lichtenauer, C. Jung, U. C. Hoppe, U. Landmesser, H. Thiele *et al.*, "Mechanical circulatory support with impella versus intra-aortic balloon pump or medical treatment in cardiogenic shock–a critical appraisal of current data," *Clinical Research in Cardiology*, vol. 108, no. 11, pp. 1249–1257, 2019.

[185] B. Alushi, A. Douedari, G. Froehlig, W. Knie, T. H. Wurster, D. M. Leistner, B.-E. Staehli, H.-C. Mochmann, B. Pieske, U. Landmesser *et al.*, "Impella versus iabp in acute myocardial infarction complicated by cardiogenic shock," *Open Heart*, vol. 6, no. 1, p. e000987, 2019.

[186] J. W. Eikelboom, S. R. Mehta, S. S. Anand, C. Xie, K. Fox, S. Yusuf *et al.*, "Adverse impact of bleeding on prognosis in patients with acute coronary syndromes," *Circulation*, vol. 114, no. 8, pp. 774–782, 2006.

[187] S. Chatterjee, J. Wetterslev, A. Sharma, E. Lichstein, and D. Mukherjee, "Association of blood transfusion with increased mortality in myocardial infarction: a meta-analysis and diversity-adjusted study sequential analysis," *JAMA internal medicine*, vol. 173, no. 2, pp. 132–139, 2013.

[188] T. D. Kinnaird, E. Stabile, G. S. Mintz, C. W. Lee, D. A. Canos, N. Gevorkian, E. E. Pinnow, K. M. Kent, A. D. Pichard, L. F. Satler *et al.*, "Incidence, predictors, and prognostic implications of bleeding and blood transfusion following percutaneous coronary interventions," *The American journal of cardiology*, vol. 92, no. 8, pp. 930–935, 2003.

[189] N. J. Udesen, J. E. Møller, M. G. Lindholm, H. Eiskjær, A. Schäfer, N. Werner, L. Holmvang, C. J. Terkelsen, L. O. Jensen, A. Junker *et al.*, "Rationale and design of danger shock: Danish-german cardiogenic shock trial," *American heart journal*, vol. 214, pp. 60–68, 2019.

[190] A. K. Chhatriwalla, A. P. Amin, K. F. Kennedy, J. A. House, D. J. Cohen, S. V. Rao, J. C. Messenger, S. P. Marso, f. t. National Cardiovascular Data Registry *et al.*, "Association between bleeding events and in-hospital mortality after percutaneous coronary intervention," *Jama*, vol. 309, no. 10, pp. 1022–1029, 2013.

[191] S. K. Mehta, A. D. Frutkin, J. B. Lindsey, J. A. House, J. A. Spertus, S. V. Rao, F.-S. Ou, M. T. Roe, E. D. Peterson, and S. P. Marso, "Bleeding in patients undergoing percutaneous coronary intervention: the development of a clinical risk algorithm from the national cardiovascular data registry," *Circulation: Cardiovascular Interventions*, vol. 2, no. 3, pp. 222–229, 2009.

[192] S. V. Rao, L. A. McCoy, J. A. Spertus, R. J. Krone, M. Singh, S. Fitzgerald, and E. D. Peterson, "An updated bleeding model to predict the risk of post-procedure bleeding among patients undergoing percutaneous coronary intervention: a report using an expanded bleeding definition from the national cardiovascular data registry cathpci registry," *JACC: Cardiovascular Interventions*, vol. 6, no. 9, pp. 897–904, 2013.

[193] I. Moussa, A. Hermann, J. C. Messenger, G. J. Dehmer, W. D. Weaver, J. S. Rumsfeld, and F. A. Masoudi, "The ncdr cathpci registry: a us national perspective on care and outcomes for percutaneous coronary intervention," *Heart*, vol. 99, no. 5, pp. 297–303, 2013.

[194] D. V. Baklanov, S. Kim, S. P. Marso, S. Subherwal, and S. V. Rao, "Comparison of bivalirudin and radial access across a spectrum of preprocedural risk of bleeding in percutaneous coronary intervention: analysis from the national cardiovascular data registry," *Circulation: Cardiovascular Interventions*, vol. 6, no. 4, pp. 347–353, 2013.

[195] S. Subherwal, E. D. Peterson, D. Dai, L. Thomas, J. C. Messenger, Y. Xian, R. G.

Brindis, D. N. Feldman, S. Senter, and L. W. Klein, "Temporal trends in and factors associated with bleeding complications among patients undergoing percutaneous coronary intervention: a report from the national cardiovascular data cathpci registry," *Journal of the American College of Cardiology*, vol. 59, no. 21, pp. 1861–1869, 2012.

[196] S. L. Daugherty, L. E. Thompson, S. Kim, S. V. Rao, S. Subherwal, T. T. Tsai, J. C. Messenger, and F. A. Masoudi, "Patterns of use and comparative effectiveness of bleeding avoidance strategies in men and women following percutaneous coronary interventions: an observational study from the national cardiovascular data registry," *Journal of the American College of Cardiology*, vol. 61, no. 20, pp. 2070–2078, 2013.

[197] A. N. Vora, E. D. Peterson, L. A. McCoy, K. N. Garratt, M. A. Kutcher, S. P. Marso, M. T. Roe, J. C. Messenger, and S. V. Rao, "The impact of bleeding avoidance strategies on hospital-level variation in bleeding rates following percutaneous coronary intervention: insights from the national cardiovascular data registry cathpci registry," *JACC: Cardiovascular Interventions*, vol. 9, no. 8, pp. 771–779, 2016.

[198] S. P. Marso, A. P. Amin, J. A. House, K. F. Kennedy, J. A. Spertus, S. V. Rao, D. J. Cohen, J. C. Messenger, J. S. Rumsfeld, and N. C. D. Registry, "Association between use of bleeding avoidance strategies and risk of periprocedural bleeding among patients undergoing percutaneous coronary intervention," *Jama*, vol. 303, no. 21, pp. 2156–2164, 2010.

[199] G. A. Aarons, A. E. Green, L. A. Palinkas, S. Self-Brown, D. J. Whitaker, J. R. Lutzker, J. F. Silovsky, D. B. Hecht, and M. J. Chaffin, "Dynamic adaptation process to implement an evidence-based child maltreatment intervention," *Implementation Science*, vol. 7, no. 1, pp. 1–9, 2012.

[200] B. J. Mortazavi, E. M. Bucholz, N. R. Desai, C. Huang, J. P. Curtis, F. A. Masoudi, R. E. Shaw, S. N. Negahban, and H. M. Krumholz, "Comparison of machine learning methods with national cardiovascular data registry models for prediction of risk of bleeding after percutaneous coronary intervention," *JAMA network open*, vol. 2, no. 7,

pp. e196 835–e196 835, 2019.

[201] S. v. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in r," *Journal of statistical software*, pp. 1–68, 2010.

[202] X. Kavelaars, S. Van Buuren, and J. Van Ginkel, "Multiple imputation in data that grow over time: A comparison of three strategies," *arXiv preprint arXiv:1904.04185*, 2019.

[203] B. C. Jaeger, N. J. Tierney, and N. R. Simon, "When to impute? imputation before and during cross-validation," *arXiv preprint arXiv:2010.00718*, 2020.

[204] L. M. Stevens, B. J. Mortazavi, R. C. Deo, L. Curtis, and D. P. Kao, "Recommendations for reporting machine learning analyses in clinical research," *Circulation: Cardiovascular Quality and Outcomes*, vol. 13, no. 10, p. e006556, 2020.

[205] B. Nestor, M. B. McDermott, W. Boag, G. Berner, T. Naumann, M. C. Hughes, A. Goldenberg, and M. Ghassemi, "Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks," in *Machine Learning for Healthcare Conference*. PMLR, 2019, pp. 381–405.

[206] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[207] E. Pinker, "Reporting accuracy of rare event classifiers," *NPJ digital medicine*, vol. 1, no. 1, pp. 1–2, 2018.

[208] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[209] R. Khera, J. Haimovich, N. C. Hurley, R. McNamara, J. A. Spertus, N. Desai, J. S. Rumsfeld, F. A. Masoudi, C. Huang, and S.-L. Normand, "Use of machine learning

models to predict death after acute myocardial infarction," *JAMA cardiology*, 2021.

[210] S. N. Wood, *Generalized additive models: an introduction with R*. CRC press, 2017.

[211] A. Zeileis, S. Köll, and N. Graham, "Various versatile variances: An object-oriented implementation of clustered covariances in r," *Journal of Statistical Software*, vol. 95, no. 1, pp. 1–36, 2020.

[212] H. Thiele, U. Zeymer, N. Thelemann, F.-J. Neumann, J. Hausleiter, M. Abdel-Wahab, R. Meyer-Saraei, G. Fuernau, I. Eitel, R. Hambrecht *et al.*, "Intraaortic balloon pump in cardiogenic shock complicating acute myocardial infarction: long-term 6-year outcome of the randomized iabp-shock ii trial," *Circulation*, vol. 139, no. 3, pp. 395–403, 2019.

[213] R. Khera, E. A. Secemsky, Y. Wang, N. R. Desai, H. M. Krumholz, T. M. Maddox, K. A. Shunk, S. S. Virani, D. L. Bhatt, J. Curtis *et al.*, "Revascularization practices and outcomes in patients with multivessel coronary artery disease who presented with acute myocardial infarction and cardiogenic shock in the us, 2009-2018," *JAMA Internal Medicine*, vol. 180, no. 10, pp. 1317–1327, 2020.

[214] S. S. Dhruva, J. S. Ross, B. J. Mortazavi, N. C. Hurley, H. M. Krumholz, J. P. Curtis, A. P. Berkowitz, F. A. Masoudi, J. C. Messenger, C. S. Parzynski *et al.*, "Use of mechanical circulatory support devices among patients with acute myocardial infarction complicated by cardiogenic shock," *JAMA network open*, vol. 4, no. 2, pp. e2 037 748–e2 037 748, 2021.

[215] G. B. Holt, "Potential simpson's paradox in multicenter study of intraperitoneal chemotherapy for ovarian cancer," *Journal of Clinical Oncology*, vol. 34, no. 9, pp. 1016–1016, 2016.

[216] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *arXiv preprint arXiv:2002.02770*, 2020.

[217] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–37, 2020.

[218] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[219] Z. Huo, A. PakBin, X. Chen, N. Hurley, Y. Yuan, X. Qian, Z. Wang, S. Huang, and B. Mortazavi, "Uncertainty quantification for deep context-aware mobile activity recognition and unknown context discovery," *arXiv preprint arXiv:2003.01753*, vol. 108, pp. 3894–3904, 26–28 Aug 2020. [Online]. Available: http://proceedings.mlr.press/v108/huo20a.html

[220] R. Xu and D. C. Wunsch, "Clustering algorithms in biomedical research: a review," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 120–154, 2010.

[221] N. C. Hurley, A. D. Haimovich, R. A. Taylor, and B. J. Mortazavi, "Visualization of emergency department clinical data for interpretable patient phenotyping," *arXiv preprint arXiv:1907.11039*, 2019.

[222] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," *arXiv preprint arXiv:1705.08821*, 2017.

[223] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang, "Representation learning for treatment effect estimation from observational data," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[224] P. Croft, D. G. Altman, J. J. Deeks, K. M. Dunn, A. D. Hay, H. Hemingway, L. LeResche, G. Peat, P. Perel, S. E. Petersen *et al.*, "The science of clinical practice: disease diagnosis or patient prognosis? evidence about "what is likely to happen" should shape clinical practice," *BMC medicine*, vol. 13, no. 1, pp. 1–8, 2015.

[225] D. M. Mannino, "COPD: epidemiology, prevalence, morbidity and mortality, and disease heterogeneity," *Chest*, vol. 121, no. 5, pp. 121S–126S, 2002.

[226] E. Feczko and D. A. Fair, "Methods and challenges for assessing heterogeneity," *Biological psychiatry*, vol. 88, no. 1, pp. 9–17, 2020.

[227] B. K. Beaulieu-Jones, C. S. Greene *et al.*, "Semi-supervised learning of the electronic health record for phenotype stratification," *Journal of biomedical informatics*, vol. 64, pp. 168–178, 2016.

[228] P. Chapfuwa, C. Li, N. Mehta, L. Carin, and R. Henao, "Survival cluster analysis," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020, pp. 60–68.

[229] C. Nagpal, D. Wei, B. Vinzamuri, M. Shekhar, S. E. Berger, S. Das, and K. R. Varshney, "Interpretable subgroup discovery in treatment effect estimation with application to opioid prescribing guidelines," in *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020, pp. 19–29.

[230] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[231] S. Nowlan and G. E. Hinton, "Evaluation of adaptive mixtures of competing experts," *Advances in neural information processing systems*, vol. 3, pp. 774–780, 1990.

[232] R. Solis, A. Pakbin, A. Akbari, B. J. Mortazavi, and R. Jafari, "A human-centered wearable sensing platform with intelligent automated data annotation capabilities," in *Proceedings of the International Conference on Internet of Things Design and Implementation*. ACM, 2019, pp. 255–260.

[233] X. Wang, F. Yu, L. Dunlap, Y.-A. Ma, R. Wang, A. Mirhoseini, T. Darrell, and J. E. Gonzalez, "Deep mixture of experts via shallow embedding," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 552–562.

[234] E. J. Benjamin, P. Muntner, and M. S. Bittencourt, "Heart disease and stroke statistics-2019 update: A report from the american heart association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30700139

[235] R. B. Schnabel, X. Yin, P. Gona, M. G. Larson, A. S. Beiser, D. D. McManus, C. Newton-Cheh, S. A. Lubitz, J. W. Magnani, P. T. Ellinor *et al.*, "50 year trends in

atrial fibrillation prevalence, incidence, risk factors, and mortality in the framingham heart study: a cohort study," *The Lancet*, vol. 386, no. 9989, pp. 154–162, 2015.

[236] D. S. Lee, P. Gona, R. S. Vasan, M. G. Larson, E. J. Benjamin, T. J. Wang, J. V. Tu, and D. Levy, "Relation of disease etiology and risk factors to heart failure with preserved or reduced ejection fraction: insights from the national heart, lung, and blood institute's framingham heart study," *Circulation*, vol. 119, no. 24, p. 3070, 2009.

[237] J. R. Romero, S. R. Preis, A. Beiser, C. DeCarli, A. Viswanathan, S. Martinez-Ramirez, C. S. Kase, P. A. Wolf, and S. Seshadri, "Risk factors, stroke prevention treatments, and prevalence of cerebral microbleeds in the framingham heart study," *Stroke*, vol. 45, no. 5, pp. 1492–1494, 2014.

[238] M. Galderisi, M. S. Lauer, and D. Levy, "Echocardiographic determinants of clinical outcome in subjects with coronary artery disease (the framingham heart study)," *The American journal of cardiology*, vol. 70, no. 11, pp. 971–976, 1992.

[239] D. M. Lloyd-Jones, M. G. Larson, E. P. Leip, A. Beiser, R. B. D'Agostino, W. B. Kannel, J. M. Murabito, R. S. Vasan, E. J. Benjamin, and D. Levy, "Lifetime risk for developing congestive heart failure: the framingham heart study," *Circulation*, vol. 106, no. 24, pp. 3068–3072, 2002.

[240] G. F. Mitchell, S.-J. Hwang, R. S. Vasan, M. G. Larson, M. J. Pencina, N. M. Hamburg, J. A. Vita, D. Levy, and E. J. Benjamin, "Arterial stiffness and cardiovascular events: the framingham heart study," *Circulation*, vol. 121, no. 4, p. 505, 2010.

[241] C. J. O'Donnell and R. Elosua, "Cardiovascular risk factors. insights from framingham heart study," *Revista Española de Cardiología (English Edition)*, vol. 61, no. 3, pp. 299–310, 2008.

[242] S. I. Chaudhry, J. A. Mattera, J. P. Curtis, J. A. Spertus, J. Herrin, Z. Lin, C. O. Phillips, B. V. Hodshon, L. S. Cooper, and H. M. Krumholz, "Telemonitoring in patients with heart failure," *New England Journal of Medicine*, vol. 363, no. 24, pp. 2301–2309, 2010. [Online]. Available: https:

//www.ncbi.nlm.nih.gov/pubmed/21080835

[243] H. M. Krumholz, S. I. Chaudhry, J. A. Spertus, J. A. Mattera, B. Hodshon, and J. Herrin, "Do non-clinical factors improve prediction of readmission risk?: results from the tele-hf study," *JACC: Heart Failure*, vol. 4, no. 1, pp. 12–20, 2016.

[244] M. K. Ong, P. S. Romano, S. Edgington, H. U. Aronow, A. D. Auerbach, J. T. Black, T. De Marco, J. J. Escarce, L. S. Evangelista, and B. Hanna, "Effectiveness of remote patient monitoring after discharge of hospitalized patients with heart failure: the better effectiveness after transition–heart failure (beat-hf) randomized clinical trial," *JAMA internal medicine*, vol. 176, no. 3, pp. 310–318, 2016.

[245] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012.

[246] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," in *23th International conference on architecture of computing systems 2010*, VDE. VDE, 2010, Conference Proceedings, pp. 1–10.

[247] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4405–4425, 2017.

[248] B. Eskofier, S. Lee, M. Baron, A. Simon, C. Martindale, H. Gaßner, and J. Klucken, "An overview of smart shoes in the internet of health things: gait and mobility assessment in health promotion and disease monitoring," *Applied Sciences*, vol. 7, no. 10, p. 986, 2017.

[249] E. Hoque and J. Stankovic, "Aalo: Activity recognition in smart homes using active learning in the presence of overlapped activities," in *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. IEEE, 2012, Conference Proceedings, pp. 139–146.

[250] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[251] S. A. Rokni, M. Nourollahi, and H. Ghasemzadeh, "Personalized human activity recognition using convolutional neural networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[252] B. H. Dobkin and C. Martinez, "Wearable sensors to monitor, enable feedback, and measure outcomes of activity and practice," *Current neurology and neuroscience reports*, vol. 18, no. 12, p. 87, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/30293160

[253] B. Mortazavi, M. Pourhomayoun, H. Ghasemzadeh, R. Jafari, C. K. Roberts, and M. Sarrafzadeh, "Context-aware data processing to enhance quality of measurements in wireless health systems: An application to met calculation of exergaming actions," *IEEE Internet of Things Journal*, vol. 2, no. 1, pp. 84–93, 2014.

[254] I. Fox, L. Ang, M. Jaiswal, R. Pop-Busui, and J. Wiens, "Contextual motifs: Increasing the utility of motifs using contextual data," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, Conference Proceedings, pp. 155–164.

[255] ——, "Deep multi-output forecasting: Learning to accurately predict blood glucose trajectories," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, Conference Proceedings, pp. 1387–1395.

[256] S. Hijazi, A. Page, B. Kantarci, and T. Soyata, "Machine learning in cardiac health monitoring and decision support," *Computer*, vol. 49, no. 11, pp. 38–48, 2016.

[257] S. Leng, R. San Tan, K. T. C. Chai, C. Wang, D. Ghista, and L. Zhong, "The electronic stethoscope," *Biomedical engineering online*, vol. 14, no. 1, p. 66, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26159433https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4496820/pdf/12938_2015_Article_56.pdf

[258] Z. Zhang, "Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction," *IEEE transactions on biomedical engineering*, vol. 62, no. 8, pp. 1902–1910, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26186747

[259] C. Li, J. Cummings, J. Lam, E. Graves, and W. Wu, "Radar remote monitoring of vital signs," *IEEE Microwave Magazine*, vol. 10, no. 1, pp. 47–56, 2009.

[260] J. Wasserlauf, C. You, R. Patel, A. Valys, D. Albert, and R. Passman, "Smartwatch performance for the detection and quantification of atrial fibrillation," *Circulation: Arrhythmia and Electrophysiology*, vol. 12, no. 6, p. e006834, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31113234

[261] B. Ziaeian and G. C. Fonarow, "Epidemiology and aetiology of heart failure," *Nature Reviews Cardiology*, vol. 13, no. 6, p. 368, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26935038

[262] R. L. McNamara, Y. Wang, J. Herrin, J. P. Curtis, E. H. Bradley, D. J. Magid, E. D. Peterson, M. Blaney, P. D. Frederick, and H. M. Krumholz, "Effect of door-to-balloon time on mortality in patients with st-segment elevation myocardial infarction," *Journal of the American College of Cardiology*, vol. 47, no. 11, pp. 2180–2186, 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16750682

[263] H. M. Krumholz, J. Herrin, L. E. Miller, E. E. Drye, S. M. Ling, L. F. Han, M. T. Rapp, E. H. Bradley, B. K. Nallamothu, and W. Nsa, "Improvements in door-to-balloon time in the united states, 2005 to 2010," *Circulation*, vol. 124, no. 9, pp. 1038–1045, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21859971

[264] B. K. Nallamothu, S.-L. T. Normand, Y. Wang, T. P. Hofer, J. E. Brush Jr, J. C. Messenger, E. H. Bradley, J. S. Rumsfeld, and H. M. Krumholz, "Relation between door-to-balloon times and mortality after primary percutaneous coronary intervention over time: a retrospective study," *The Lancet*, vol. 385, no. 9973, pp. 1114–1122, 2015.

[265] M. H. Olsen, S. Y. Angell, S. Asma, P. Boutouyrie, D. Burger, J. A. Chirinos,

A. Damasceno, C. Delles, A.-P. Gimenez-Roqueplo, and D. Hering, "A call to action and a lifecourse strategy to address the global burden of raised blood pressure on current and future generations: the lancet commission on hypertension," *The Lancet*, vol. 388, no. 10060, pp. 2665–2712, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27671667

[266] E. Onusko, "Diagnosing secondary hypertension," *American family physician*, vol. 67, no. 1, pp. 67–74, 2003. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/12537168

[267] D. M. Reboussin, N. B. Allen, M. E. Griswold, E. Guallar, Y. Hong, D. T. Lackland, E. P. R. Miller, T. Polonsky, A. M. Thompson-Paul, and S. Vupputuri, "Systematic review for the 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: a report of the american college of cardiology/american heart association task force on clinical practice guidelines," *Journal of the American College of Cardiology*, vol. 71, no. 19, pp. 2176–2198, 2018.

[268] R. Levin, M. Dolgin, C. Fox, and R. Gorlin, "The criteria committee of the new york heart association: Nomenclature and criteria for diagnosis of diseases of the heart and great vessels," *LWW Handbooks*, vol. 9, p. 344, 1994.

[269] D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'Agostino, R. Gibbons, P. Greenland, D. T. Lackland, D. Levy, and C. J. O'Donnell, "2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines," *Journal of the American College of Cardiology*, vol. 63, no. 25 Part B, pp. 2935–2959, 2014.

[270] S. R. Group, "A randomized trial of intensive versus standard blood-pressure control," *New England Journal of Medicine*, vol. 373, no. 22, pp. 2103–2116, 2015.

[271] "Amazfit verge." [Online]. Available: https://en.amazfit.com/verge.html

[272] A. Support, "Your heart rate. what it means, and where on apple watch you'll find

it." [Online]. Available: https://support.apple.com/en-us/HT204666

[273] Empatica, "Real-time physiological signals | e4 eda/gsr sensor." [Online]. Available: https://www.empatica.com/research/e4

[274] F. Help, "How do i track my heart rate with my fitbit device?" [Online]. Available: https://help.fitbit.com/articles/en_US/Help_article/1565

[275] Garmin and G. L. subsidiaries, "Garmin fenix® 6 | multisport fitness watch." [Online]. Available: https://buy.garmin.com/en-US/US/p/641449

[276] S. E. America, "Samsung heart rate sensor." [Online]. Available: https://www.samsung.com/us/heartratesensor/

[277] Valencell, "Valencell | customers." [Online]. Available: https://valencell.com/customers/

[278] Withings, "Fitness trackers and hybrid smartwatches by withings." [Online]. Available: https://www.withings.com/us/en/watches

[279] O. Ring, "Get to know oura." [Online]. Available: https://ouraring.com/get-to-know-oura

[280] Valencell, "Blood pressure blood pressure." [Online]. Available: https://valencell.com/bloodpressure/

[281] Garmin and G. L. subsidiaries, "Garmin | heart rate monitors." [Online]. Available: https://buy.garmin.com/en-US/US/c14662-p1.html

[282] P. USA, "Polar usa." [Online]. Available: https://www.polar.com/us-en/products/compare

[283] Qardio, "Qardiocore." [Online]. Available: https://store.getqardio.com/products/qardiocore

[284] AliveCor, "Alivecor." [Online]. Available: https://www.alivecor.com/kardiamobile6l/

[285] Caretaker Medical, "Medical papers – caretaker medical." [Online]. Available: https://www.caretakermedical.net/medical-papers/

[286] Finapres Medical Systems, "Finapres medical systems | products - finapres® nova."

[Online]. Available: http://www.finapres.com/Products/Finapres-NOVA

[287] Qardio, "Irregular heart beat detection." [Online]. Available: http://support.getqardio.com/hc/en-us/articles/203579482

[288] R. D. Conn and J. H. O'Keefe, "Cardiac physical diagnosis in the digital age: an important but increasingly neglected skill (from stethoscopes to microchips)," *The American journal of cardiology*, vol. 104, no. 4, pp. 590–595, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19660617https://www.ajconline.org/article/S0002-9149(09)00957-6/fulltext

[289] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. Johnson *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiological Measurement*, vol. 37, no. 12, p. 2181, 2016.

[290] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark, "Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016," in *2016 Computing in Cardiology Conference (CinC)*. IEEE, 2016, pp. 609–612.

[291] M. E. Chowdhury, A. Khandakar, K. Alzoubi, S. Mansoor, A. M Tahir, M. B. I. Reaz, and N. Al-Emadi, "Real-time smart-digital stethoscope system for heart diseases monitoring," *Sensors*, vol. 19, no. 12, p. 2781, 2019.

[292] A. Sinharay, D. Ghosh, P. Deshpande, S. Alam, R. Banerjee, and A. Pal, "Smartphone based digital stethoscope for connected health–a direct acoustic coupling technique," in *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2016, Conference Proceedings, pp. 193–198.

[293] M. Elgendi, P. Bobhate, S. Jain, J. Rutledge, J. Y. Coe, R. Zemp, D. Schuurmans, and I. Adatia, "Time-domain analysis of heart sound intensity in children with and without pulmonary artery hypertension: a pilot study using a digital stethoscope," *Pulmonary circulation*, vol. 4, no. 4, pp. 685–695, 2014.

[294] G. Vinci, S. Lindner, F. Barbon, S. Mann, M. Hofmann, A. Duda, R. Weigel, and A. Koelpin, "Six-port radar sensor for remote respiration rate and heartbeat vital-sign monitoring," *IEEE Transactions on Microwave Theory and Techniques*, vol. 61, no. 5, pp. 2093–2100, 2013.

[295] B. D. Mar, "The history of clinical holter monitoring," *Annals of Noninvasive Electrocardiology*, vol. 10, no. 2, pp. 226–230, 2005.

[296] Y. M. Chi, T.-P. Jung, and G. Cauwenberghs, "Dry-contact and noncontact biopotential electrodes: Methodological review," *IEEE reviews in biomedical engineering*, vol. 3, pp. 106–119, 2010.

[297] S. Majumder, L. Chen, O. Marinov, C.-H. Chen, T. Mondal, and M. J. Deen, "Noncontact wearable wireless ecg systems for long-term monitoring," *IEEE reviews in biomedical engineering*, vol. 11, pp. 306–321, 2018.

[298] Q. Li, C. Rajagopalan, and G. D. Clifford, "A machine learning approach to multi-level ecg signal quality classification," *Computer methods and programs in biomedicine*, vol. 117, no. 3, pp. 435–447, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25306242

[299] J. C. Sriram, M. Shin, T. Choudhury, and D. Kotz, "Activity-aware ecg-based patient authentication for remote health monitoring," in *Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, 2009, Conference Proceedings, pp. 297–304.

[300] M. Kang, E. Park, B. H. Cho, and K.-S. Lee, "Recent patient health monitoring platforms incorporating internet of things-enabled smart devices," *International neurourology journal*, vol. 22, no. Suppl 2, p. S76, 2018.

[301] M. M. Kabir, E. A. Perez-Alday, J. Thomas, G. Sedaghat, and L. G. Tereshchenko, "Optimal configuration of adhesive ecg patches suitable for long-term monitoring of a vectorcardiogram," *Journal of electrocardiology*, vol. 50, no. 3, pp. 342–348, 2017.

[302] E. Sen-Gupta, D. E. Wright, J. W. Caccese, J. A. Wright Jr, E. Jortberg, V. Bhatkar,

M. Ceruolo, R. Ghaffari, D. L. Clason, J. P. Maynard *et al.*, "A pivotal study to validate the performance of a novel wearable sensor and system for biometric monitoring in clinical and remote environments," *Digital Biomarkers*, vol. 3, no. 1, pp. 1–13, 2019.

[303] S. R. Steinhubl, J. Waalen, A. M. Edwards, L. M. Ariniello, R. R. Mehta, G. S. Ebner, C. Carter, K. Baca-Motes, E. Felicione, and T. Sarich, "Effect of a home-based wearable continuous ecg monitoring patch on detection of undiagnosed atrial fibrillation: the mstops randomized clinical trial," *Jama*, vol. 320, no. 2, pp. 146–155, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29998336

[304] P. M. Barrett, R. Komatireddy, S. Haaser, S. Topol, J. Sheard, J. Encinas, A. J. Fought, and E. J. Topol, "Comparison of 24-hour holter monitoring with 14-day novel adhesive patch electrocardiographic monitoring," *The American journal of medicine*, vol. 127, no. 1, pp. 95–e11, 2014.

[305] J. Chen, H. Peng, and A. Razi, "Remote ecg monitoring kit to predict patient-specific heart abnormalities," *Journal of Systemics, Cybernetics and Informatics*, vol. 15, no. 4, pp. 82–89, 2017.

[306] E. O'brien, B. Waeber, G. Parati, J. Staessen, and M. G. Myers, "Blood pressure measuring devices: recommendations of the european society of hypertension," *Bmj*, vol. 322, no. 7285, pp. 531–536, 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11230071

[307] C. F. Babbs, "Oscillometric measurement of systolic and diastolic blood pressures validated in a physiologic mathematical model," *Biomedical engineering online*, vol. 11, no. 1, p. 56, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22913792

[308] J. Hodgkinson, J. Mant, U. Martin, B. Guo, F. Hobbs, J. Deeks, C. Heneghan, N. Roberts, and R. McManus, "Relative effectiveness of clinic and home blood pressure monitoring compared with ambulatory blood pressure monitoring in diagnosis of hypertension: systematic review," *Bmj*, vol. 342, p. d3621, 2011. [Online]. Available:

https://www.ncbi.nlm.nih.gov/pubmed/21705406

[309] T. Ma and Y.-T. Zhang, "A correlation study on the variabilities in pulse transit time, blood pressure, and heart rate recorded simultaneously from healthy subjects," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2006, Conference Proceedings, pp. 996–999.

[310] S. S. Thomas, V. Nathan, C. Zong, E. Akinbola, A. L. P. Aroul, L. Philipose, K. Soundarapandian, X. Shi, and R. Jafari, "Biowatch-a wrist watch based signal acquisition system for physiological signals including blood pressure," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, Conference Proceedings, pp. 2286–2289.

[311] F. C. Bennis, C. van Pul, J. J. van den Bogaart, P. Andriessen, B. W. Kramer, and T. Delhaas, "Artifacts in pulse transit time measurements using standard patient monitoring equipment," *PloS one*, vol. 14, no. 6, p. e0218784, 2019.

[312] L. Peter, N. Noury, and M. Cerny, "A review of methods for non-invasive and continuous blood pressure monitoring: Pulse transit time method is promising?" *Irbm*, vol. 35, no. 5, pp. 271–282, 2014.

[313] R. Payne, C. Symeonides, D. Webb, and S. Maxwell, "Pulse transit time measured from the ecg: an unreliable marker of beat-to-beat blood pressure," *Journal of Applied Physiology*, vol. 100, no. 1, pp. 136–141, 2006.

[314] W.-X. Dai, Y.-T. Zhang, J. Liu, X.-R. Ding, and N. Zhao, "Dual-modality arterial pulse monitoring system for continuous blood pressure measurement," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, Conference Proceedings, pp. 5773–5776.

[315] Z. Trujillo, V. Nathan, G. L. Coté, and R. Jafari, "Design and parametric analysis of a wearable dual-photoplethysmograph based system for pulse wave velocity detection," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2017, Conference Proceedings, pp. 1–4.

[316] T. Huynh, R. Jafari, and W.-Y. Chung, "An accurate bioimpedance measurement system for blood pressure monitoring," *Sensors*, vol. 18, no. 7, p. 2095, 2018.

[317] B. Ibrahim, A. Akbari, and R. Jafari, "A novel method for pulse transit time estimation using wrist bio-impedance sensing based on a regression model," in *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2017, Conference Proceedings, pp. 1–4.

[318] O. T. Inan, A. Q. Javaid, S. Dowling, H. Ashouri, M. Etemadi, J. A. Heller, S. Roy, and L. Klein, "Using ballistocardiography to monitor left ventricular function in heart failure patients," *Journal of Cardiac Failure*, vol. 22, no. 8, p. S45, 2016.

[319] C.-S. Kim, S. L. Ober, M. S. McMurtry, B. A. Finegan, O. T. Inan, R. Mukkamala, and J.-O. Hahn, "Ballistocardiogram: Mechanism and potential for unobtrusive cardiovascular health monitoring," *Scientific reports*, vol. 6, p. 31297, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27503664

[320] C.-S. Kim, A. M. Carek, O. T. Inan, R. Mukkamala, and J.-O. Hahn, "Ballistocardiogram-based approach to cuffless blood pressure monitoring: proof of concept and potential challenges," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 11, pp. 2384–2391, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29993523

[321] B. Ibrahim and R. Jafari, "Cuffless blood pressure monitoring from an array of wrist bio-impedance sensors using subject-specific regression models: Proof of concept," *IEEE transactions on biomedical circuits and systems*, vol. 13, no. 6, pp. 1723–1735, 2019.

[322] A. Aygun, H. Ghasemzadeh, and R. Jafari, "Robust interbeat interval and heart rate variability estimation method from various morphological features using wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, 2019.

[323] K. Sel, J. Zhao, B. Ibrahim, and R. Jafari, "Measurement of chest physiological signals using wirelessly coupled bio-impedance patches," in *2019 41st Annual International*

Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019, pp. 376–381.

[324] D. N. Ku, "Blood flow in arteries," *Annual Review of Fluid Mechanics*, vol. 29, no. 1, pp. 399–434, 1997. [Online]. Available: https://doi.org/10.1146/annurev.fluid.29.1.399

[325] A. C. Burton and S. Yamada, "Relation between blood pressure and flow in the human forearm," *Journal of applied physiology*, vol. 4, no. 5, pp. 329–339, 1951.

[326] N. M. Hamburg and M. A. Creager, "Pathophysiology of intermittent claudication in peripheral artery disease," *Circulation Journal*, pp. CJ–16, 2017.

[327] R. W. Gill, "Measurement of blood flow by ultrasound: accuracy and sources of error," *Ultrasound in Medicine and Biology*, vol. 11, no. 4, pp. 625–641, 1985.

[328] N. A. Martin, C. Doberstein, C. Zane, M. J. Caron, K. Thomas, and D. P. Becker, "Posttraumatic cerebral arterial spasm: transcranial doppler ultrasound, cerebral blood flow, and angiographic findings," *Journal of neurosurgery*, vol. 77, no. 4, pp. 575–583, 1992.

[329] E. de Groot, G. K. Hovingh, A. Wiegman, P. Duriez, A. J. Smit, J.-C. Fruchart, and J. J. Kastelein, "Measurement of arterial wall thickness as a surrogate marker for atherosclerosis," *Circulation*, vol. 109, no. 23_suppl_1, pp. III–33, 2004.

[330] G. Nayler, D. Firmin, D. Longmore *et al.*, "Blood flow imaging by cine magnetic resonance," *J Comput Assist Tomogr*, vol. 10, no. 5, pp. 715–722, 1986.

[331] R. Fallahzadeh, M. Pedram, and H. Ghasemzadeh, "Smartsock: A wearable platform for context-aware assessment of ankle edema," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, Conference Proceedings, pp. 6302–6306.

[332] R. Fallahzadeh, M. Pedram, R. Saeedi, B. Sadeghi, M. Ong, and H. Ghasemzadeh, "Smart-cuff: A wearable bio-sensing platform with activity-sensitive information quality assessment for monitoring ankle edema," in *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. IEEE,

2015, Conference Proceedings, pp. 57–62.

[333] J. Yao, E. M. Weaver, B. D. Langley, S. M. George, and S. R. Hardin, "Monitoring peripheral edema of heart failure patients at home: Device, algorithm, and clinic study," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, Conference Proceedings, pp. 4074–4077.

[334] O. Yürür, C. H. Liu, and W. Moreno, "Light-weight online unsupervised posture detection by smartphone accelerometer," *IEEE Internet of Things Journal*, vol. 2, no. 4, pp. 329–339, 2015.

[335] K. Ouchi and M. Doi, "Smartphone-based monitoring system for activities of daily living for elderly people and their relatives etc," in *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, 2013, Conference Proceedings, pp. 103–106.

[336] I. Pires, N. Garcia, N. Pombo, and F. Flórez-Revuelta, "From data acquisition to data fusion: a comprehensive review and a roadmap for the identification of activities of daily living using mobile devices," *Sensors*, vol. 16, no. 2, p. 184, 2016.

[337] B. J. Mortazavi, M. Pourhomayoun, G. Alsheikh, N. Alshurafa, S. I. Lee, and M. Sarrafzadeh, "Determining the single best axis for exercise repetition recognition and counting on smartwatches," in *2014 11th International Conference on Wearable and Implantable Body Sensor Networks*, IEEE. IEEE, 2014, Conference Proceedings, pp. 33–38.

[338] S. Sen, V. Subbaraju, A. Misra, R. K. Balan, and Y. Lee, "The case for smartwatch-based diet monitoring," in *2015 IEEE international conference on pervasive computing and communication workshops (PerCom workshops)*. IEEE, 2015, Conference Proceedings, pp. 585–590.

[339] K. Van Laerhoven, M. Borazio, and J. H. Burdinski, "Wear is your mobile? investigating phone carrying and use habits with a wearable device," *Frontiers in ICT*, vol. 2,

p. 10, 2015.

[340] C. for Disease Control, Prevention *et al.*, "National diabetes statistics report, 2017: Estimates of diabetes and its burden in the united states. atlanta, ga: Centers for disease control and prevention; 2017," 2019.

[341] H. Kalantarian, N. Alshurafa, T. Le, and M. Sarrafzadeh, "Monitoring eating habits using a piezoelectric sensor-based necklace," *Computers in biology and medicine*, vol. 58, pp. 46–55, 2015.

[342] S. Fang, Z. Shao, D. A. Kerr, C. J. Boushey, and F. Zhu, "An end-to-end image-based automatic food energy estimation technique based on learned energy distribution images: Protocol and methodology," *Nutrients*, vol. 11, no. 4, p. 877, 2019.

[343] Z. Huo, B. J. Mortazavi, T. Chaspari, N. Deutz, L. Ruebush, and R. Gutierrez-Osuna, "Predicting the meal macronutrient composition from continuous glucose monitors," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019, pp. 1–4.

[344] D. Zeevi, T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan *et al.*, "Personalized nutrition by prediction of glycemic responses," *Cell*, vol. 163, no. 5, pp. 1079–1094, 2015.

[345] F. Elijovich, M. H. Weinberger, C. A. Anderson, L. J. Appel, M. Bursztyn, N. R. Cook, R. A. Dart, C. H. Newton-Cheh, F. M. Sacks, and C. L. Laffer, "Salt sensitivity of blood pressure: a scientific statement from the american heart association," *Hypertension*, vol. 68, no. 3, pp. e7–e46, 2016.

[346] M. Porumb, S. Stranges, A. Pescapè, and L. Pecchia, "Precision medicine and artificial intelligence: A pilot study on deep learning for hypoglycemic events detection based on ecg," *Scientific Reports*, vol. 10, no. 1, pp. 1–16, 2020.

[347] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016, Conference Proceedings.

[348] S. A. Rokni and H. Ghasemzadeh, "Plug-n-learn: automatic learning of computational algorithms in human-centered internet-of-things applications," in *Proceedings of the 53rd Annual Design Automation Conference*. ACM, 2016, Conference Proceedings, p. 139.

[349] ——, "Autonomous training of activity recognition algorithms in mobile sensors: A transfer learning approach in context-invariant views," *IEEE Transactions on Mobile Computing*, vol. 17, no. 8, pp. 1764–1777, 2018.

[350] Z. C. Lipton, D. C. Kale, and R. Wetzel, "Modeling missing data in clinical time series with rnns," *arXiv preprint arXiv:1606.04130*, 2016.

[351] Y. Liu, Z. Li, Z. Liu, and K. Wu, "Real-time arm skeleton tracking and gesture inference tolerant to missing wearable sensors," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2019, Conference Proceedings, pp. 287–299.

[352] S. B. Thacker, D. Stroup, M.-h. Chang, and S. L. Henderson, "Continuous electronic heart rate monitoring for fetal assessment during labor," *Cochrane database of systematic reviews*, no. 2, 2001.

[353] Z. Alfirevic, G. M. Gyte, A. Cuthbert, and D. Devane, "Continuous cardiotocography (ctg) as a form of electronic fetal monitoring (efm) for fetal assessment during labour," *Cochrane database of systematic reviews*, no. 2, 2017.

[354] M. J. Mc Loughlin and S. Mc Loughlin, "Cardiac auscultation: Preliminary findings of a pilot study using continuous wave doppler and comparison with classic auscultation," *International journal of cardiology*, vol. 167, no. 2, pp. 590–591, 2013.

[355] A. Knapp, V. Cetrullo, B. A. Sillars, N. Lenzo, W. A. Davis, and T. M. Davis, "Carotid artery ultrasonographic assessment in patients from the fremantle diabetes study phase ii with carotid bruits detected by electronic auscultation," *Diabetes technology & therapeutics*, vol. 16, no. 9, pp. 604–610, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24988112

[356] R. Palaniappan, K. Sundaraj, and N. U. Ahamed, "Machine learning in lung sound analysis: a systematic review," *Biocybernetics and Biomedical Engineering*, vol. 33, no. 3, pp. 129–135, 2013.

[357] B. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques *et al.*, "A respiratory sound database for the development of automated classification," in *International Conference on Biomedical and Health Informatics*. Springer, 2017, pp. 33–37.

[358] K. Kochetov, E. Putin, S. Azizov, I. Skorobogatov, and A. Filchenkov, "Wheeze detection using convolutional neural networks," in *EPIA Conference on Artificial Intelligence*. Springer, 2017, pp. 162–173.

[359] D. Perna, "Convolutional neural networks learning from respiratory data," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 2109–2113.

[360] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. Shalyto, "Noise masking recurrent neural network for respiratory sound classification," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 208–217.

[361] D. Perna and A. Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2019, pp. 50–55.

[362] H. Pasterkamp, P. L. Brand, M. Everard, L. Garcia-Marcos, H. Melbye, and K. N. Priftis, "Towards the standardisation of lung sound nomenclature," *European Respiratory Journal*, vol. 47, no. 3, pp. 724–732, 2016.

[363] S. H. Jambukia, V. K. Dabhi, and H. B. Prajapati, "Classification of ecg signals using machine learning techniques: A survey," in *2015 International Conference on Advances in Computer Engineering and Applications*. IEEE, 2015, Conference Proceedings, pp. 714–721.

[364] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database,"

IEEE Engineering in Medicine and Biology Magazine, vol. 20, no. 3, pp. 45–50, 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11446209

[365] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional lstm network model for ecg signal classification," *Computers in biology and medicine*, vol. 96, pp. 189–202, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29614430

[366] R. S. Andersen, A. Peimankar, and S. Puthusserypady, "A deep learning approach for real-time detection of atrial fibrillation," *Expert Systems with Applications*, vol. 115, pp. 465–473, 2019.

[367] U. B. Baloglu, M. Talo, O. Yildirim, R. San Tan, and U. R. Acharya, "Classification of myocardial infarction with multi-lead ecg signals and deep cnn," *Pattern Recognition Letters*, vol. 122, pp. 23–30, 2019.

[368] S. M. Mathews, C. Kambhamettu, and K. E. Barner, "A novel application of deep learning for single-lead ecg classification," *Computers in biology and medicine*, vol. 99, pp. 53–62, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29886261

[369] C. L. Moore and J. A. Copel, "Point-of-care ultrasonography," *New England Journal of Medicine*, vol. 364, no. 8, pp. 749–757, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21345104

[370] S. Assaad, W. B. Kratzert, B. Shelley, M. B. Friedman, and A. Perrino Jr, "Assessment of pulmonary edema: principles and practice," *Journal of cardiothoracic and vascular anesthesia*, vol. 32, no. 2, pp. 901–914, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29174750

[371] A. Bhuyan, J. W. Choe, B. C. Lee, P. Cristman, Ö. Oralkan, and B. T. Khuri-Yakub, "Miniaturized, wearable, ultrasound probe for on-demand ultrasound screening," in *2011 IEEE International Ultrasonics Symposium*. IEEE, 2011, Conference Proceedings, pp. 1060–1063.

[372] R. M. Lang, L. P. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. A. Flachskampf, E. Foster, S. A. Goldstein, and T. Kuznetsova, "Recommendations for

cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015.

[373] A. Østvik, E. Smistad, S. A. Aase, B. O. Haugen, and L. Lovstakken, "Real-time standard view classification in transthoracic echocardiography using convolutional neural networks," *Ultrasound in medicine & biology*, vol. 45, no. 2, pp. 374–384, 2019.

[374] C.-M. Yu, L. Wang, E. Chau, R. H.-W. Chan, S.-L. Kong, M.-O. Tang, J. Christensen, R. W. Stadler, and C.-P. Lau, "Intrathoracic impedance monitoring in patients with heart failure: correlation with fluid status and feasibility of early warning preceding hospitalization," *Circulation*, vol. 112, no. 6, pp. 841–848, 2005. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16061743

[375] S. Weyer, T. Menden, L. Leicht, S. Leonhardt, and T. Wartzek, "Development of a wearable multi-frequency impedance cardiography device," *Journal of medical engineering & technology*, vol. 39, no. 2, pp. 131–137, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25559781

[376] R. Amelard, R. L. Hughson, D. K. Greaves, K. J. Pfisterer, J. Leung, D. A. Clausi, and A. Wong, "Non-contact hemodynamic imaging reveals the jugular venous pulse waveform," *Scientific reports*, vol. 7, p. 40150, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28065933

[377] S. Dash, K. H. Shelley, D. G. Silverman, and K. H. Chon, "Estimation of respiratory rate from ecg, photoplethysmogram, and piezoelectric pulse transducer signals: a comparative study of time–frequency methods," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 5, pp. 1099–1107, 2010.

[378] M. Bolanos, H. Nazeran, and E. Haltiwanger, "Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals," in *2006 International Conference of the IEEE Engineering in Medicine*

and Biology Society. IEEE, 2006, pp. 4289–4294.

[379] M. Sánchez-de-la Torre, A. Khalyfa, A. Sánchez-de-la Torre, M. Martinez-Alonso, M. Á. Martinez-García, A. Barceló, P. Lloberes, F. Campos-Rodriguez, F. Capote, and M. J. Diaz-de Atauri, "Precision medicine in patients with resistant hypertension and obstructive sleep apnea: blood pressure response to continuous positive airway pressure treatment," *Journal of the American College of Cardiology*, vol. 66, no. 9, pp. 1023–1032, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26314530

[380] M. Pourhomayoun, N. Alshurafa, B. Mortazavi, H. Ghasemzadeh, K. Sideris, B. Sadeghi, M. Ong, L. Evangelista, P. Romano, and A. Auerbach, "Multiple model analytics for adverse event prediction in remote health monitoring systems," in *2014 IEEE Healthcare Innovation Conference (HIC)*. IEEE, 2014, Conference Proceedings, pp. 106–110.

[381] B. Mortazavi, S. Nyamathi, S. I. Lee, T. Wilkerson, H. Ghasemzadeh, and M. Sarrafzadeh, "Near-realistic mobile exergames with wireless wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 2, pp. 449–456, 2013.

[382] T. R. Bennett, C. Savaglio, D. Lu, H. Massey, X. Wang, J. Wu, and R. Jafari, "Motion-synthesis toolset (most): a toolset for human motion data synthesis and validation," in *Proceedings of the 4th ACM MobiHoc workshop on Pervasive wireless healthcare*. ACM, 2014, Conference Proceedings, pp. 25–30.

[383] V. Nathan, S. Paul, T. Prioleau, L. Niu, B. J. Mortazavi, S. A. Cambone, A. Veeraraghavan, A. Sabharwal, and R. Jafari, "A survey on smart homes for aging in place: Toward solutions to the specific needs of the elderly," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 111–119, 2018.

[384] Y.-t. Zhang, C. C. Poon, C.-h. Chan, M. W. Tsang, and K.-f. Wu, "A health-shirt using e-textile materials for the continuous and cuffless monitoring of arterial blood pressure," in *2006 3rd IEEE/EMBS International Summer School on Medical Devices*

and Biosensors. IEEE, 2006, Conference Proceedings, pp. 86–89.

[385] M. Abtahi, J. V. Gyllinsky, B. Paesang, S. Barlow, M. Constant, N. Gomes, O. Tully, S. E. D'Andrea, and K. Mankodiya, "Magicsox: An e-textile iot system to quantify gait abnormalities," *Smart Health*, vol. 5, pp. 4–14, 2018.

[386] B. S. Heran, J. M. Chen, S. Ebrahim, T. Moxham, N. Oldridge, K. Rees, D. R. Thompson, and R. S. Taylor, "Exercise-based cardiac rehabilitation for coronary heart disease," *Cochrane database of systematic reviews*, no. 7, p. CD001800, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21735386

[387] R. Maddison, J. C. Rawstorn, A. Rolleston, R. Whittaker, R. Stewart, J. Benatar, I. Warren, Y. Jiang, and N. Gant, "The remote exercise monitoring trial for exercise-based cardiac rehabilitation (remote-cr): a randomised controlled trial protocol," *BMC Public Health*, vol. 14, no. 1, p. 1236, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25432467

[388] K. N. Karmali, P. Davies, F. Taylor, A. Beswick, N. Martin, and S. Ebrahim, "Promoting patient uptake and adherence in cardiac rehabilitation," *Cochrane Database of Systematic Reviews*, no. 6, p. CD007131, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24963623

[389] S. Saadatnejad, M. Oveisi, and M. Hashemi, "Lstm-based ecg classification for continuous monitoring on personal wearable devices," *IEEE journal of biomedical and health informatics*, 2019.

[390] S. Mirshekarian, R. Bunescu, C. Marling, and F. Schwartz, "Using lstms to learn physiological models of blood glucose behavior," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 2887–2891.

[391] Y. Zhang, Z. Yang, K. Lan, X. Liu, Z. Zhang, P. Li, D. Cao, J. Zheng, and J. Pan, "Sleep stage classification using bidirectional lstm in wearable multi-sensor systems," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops*

(INFOCOM WKSHPS). IEEE, 2019, pp. 443–448.

[392] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[393] S. L. Oh, E. Y. Ng, R. San Tan, and U. R. Acharya, "Automated diagnosis of arrhythmia using combination of cnn and lstm techniques with variable length heart beats," *Computers in biology and medicine*, vol. 102, pp. 278–287, 2018.

[394] S. Kiranyaz, T. Ince, and M. Gabbouj, "Personalized monitoring and advance warning system for cardiac arrhythmias," *Scientific reports*, vol. 7, no. 1, pp. 1–8, 2017.

[395] N. van Boven, L. C. Battes, K. M. Akkerhuis, D. Rizopoulos, K. Caliskan, S. S. Anroedh, W. Yassi, O. C. Manintveld, J.-H. Cornel, A. A. Constantinescu *et al.*, "Toward personalized risk assessment in patients with chronic heart failure: detailed temporal patterns of nt-probnp, troponin t, and crp in the bio-shift study," *American heart journal*, vol. 196, pp. 36–48, 2018.

[396] R. Ardywibowo, G. Zhao, Z. Wang, B. Mortazavi, S. Huang, and X. Qian, "Adaptive activity monitoring with uncertainty quantification in switching gaussian process models," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, Conference Proceedings, pp. 266–275.

[397] A. Akbari and R. Jafari, "Personalizing activity recognition models with quantifying different types of uncertainty using wearable sensors," *IEEE Transactions on Biomedical Engineering*, 2020.

[398] K. D. Feuz and D. J. Cook, "Real-time annotation tool (rat)," in *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[399] R. Adams, N. Saleheen, E. Thomaz, A. Parate, S. Kumar, and B. Marlin, "Hierarchical span-based conditional random fields for labeling and segmenting events in wearable sensor data streams," in *International conference on machine learning*, 2016, pp. 334–343.

[400] R. S. Sadasivam, E. M. Borglund, R. Adams, B. M. Marlin, and T. K. Houston,

"Impact of a collective intelligence tailored messaging system on smoking cessation: the perspect randomized experiment," *Journal of medical Internet research*, vol. 18, no. 11, p. e285, 2016.

[401] A. Akbari, R. S. Castilla, R. Jafari, and B. J. Mortazavi, "Using intelligent personal annotations to improve human activity recognition for movements in natural environments," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[402] R. Fallahzadeh, S. Aminikhanghahi, A. N. Gibson, and D. J. Cook, "Toward personalized and context-aware prompting for smartphone-based intervention," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 6010–6013.

[403] F. Miao, Z. Liu, J. Liu, B. Wen, and Y. Li, "Multi-sensor fusion approach for cuffless blood pressure measurement," *IEEE journal of biomedical and health informatics*, 2019.

[404] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[405] R. Mohamed and M. Youssef, "Heartsense: Ubiquitous accurate multi-modal fusion-based heart rate estimation using smartphones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–18, 2017.

[406] J. H. Lee, H. Gamper, I. Tashev, S. Dong, S. Ma, J. Remaley, J. D. Holbery, and S. H. Yoon, "Stress monitoring using multimodal bio-sensing headset," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2020, pp. 1–7.

[407] D. Bannach, O. Amft, and P. Lukowicz, "Automatic event-based synchronization of multimodal data streams from wearable and ambient sensors," in *European Conference on Smart Sensing and Context*. Springer, 2009, pp. 135–148.

[408] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Thirty-Second AAAI Conference on*

Artificial Intelligence, 2018, Conference Proceedings.

[409] J. A. Rymer and S. V. Rao, "Enhancement of risk prediction with machine learning: Rise of the machines," *JAMA network open*, vol. 2, no. 7, pp. e196 823–e196 823, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31290985

[410] S. de Denus, E. O'Meara, A. S. Desai, B. Claggett, E. F. Lewis, G. Leclair, M. Jutras, J. Lavoie, S. D. Solomon, and B. Pitt, "Spironolactone metabolites in topcat–new insights into regional variation," *New England Journal of Medicine*, vol. 376, no. 17, pp. 1690–1692, 2017.

[411] B. Pitt, M. A. Pfeffer, S. F. Assmann, R. Boineau, I. S. Anand, B. Claggett, N. Clausell, A. S. Desai, R. Diaz, J. L. Fleg *et al.*, "Spironolactone for heart failure with preserved ejection fraction," *New England Journal of Medicine*, vol. 370, no. 15, pp. 1383–1392, 2014.

[412] M. A. Pfeffer, B. Claggett, S. F. Assmann, R. Boineau, I. S. Anand, N. Clausell, A. S. Desai, R. Diaz, J. L. Fleg, and I. Gordeev, "Regional variation in patients and outcomes in the treatment of preserved cardiac function heart failure with an aldosterone antagonist (topcat) trial," *Circulation*, vol. 131, no. 1, pp. 34–42, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25406305

[413] G. Y. Lip, R. Nieuwlaat, R. Pisters, D. A. Lane, and H. J. Crijns, "Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation," *Chest*, vol. 137, no. 2, pp. 263–272, 2010.

[414] R. Pisters, D. A. Lane, R. Nieuwlaat, C. B. De Vos, H. J. Crijns, and G. Y. Lip, "A novel user-friendly score (has-bled) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the euro heart survey," *Chest*, vol. 138, no. 5, pp. 1093–1100, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20299623

[415] A. F. Members, A. J. Camm, G. Y. Lip, R. De Caterina, I. Savelieva, D. Atar, S. H. Hohnloser, G. Hindricks, P. Kirchhof, and E. C. f. P. Guidelines, "2012 focused update

of the esc guidelines for the management of atrial fibrillation: an update of the 2010 esc guidelines for the management of atrial fibrillation developed with the special contribution of the european heart rhythm association," *European heart journal*, vol. 33, no. 21, pp. 2719–2747, 2012.

[416] L. Nobel, N. E. Mayo, J. Hanley, L. Nadeau, and S. S. Daskalopoulou, "Myrisk_stroke calculator: a personalized stroke risk assessment tool for the general population," *Journal of Clinical Neurology*, vol. 10, no. 1, pp. 1–9, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/24465256

[417] J. Hippisley-Cox, C. Coupland, and P. Brindle, "Derivation and validation of qstroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study," *Bmj*, vol. 346, 2013.

[418] S. D. Anker, F. Koehler, and W. T. Abraham, "Telemedicine and remote management of patients with heart failure," *The Lancet*, vol. 378, no. 9792, pp. 731–739, 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/21856487

[419] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, "A targeted real-time early warning score (trewscore) for septic shock," *Science translational medicine*, vol. 7, no. 299, pp. 299ra122–299ra122, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26246167

[420] A. S. Cakmak, E. Reinertsen, H. A. Taylor, A. J. Shah, and G. D. Clifford, "Personalized heart failure severity estimates using passive smartphone data," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1569–1574.

[421] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, p. 24, 2018.

[422] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[423] C. Lee, J. Yoon, and M. Van Der Schaar, "Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data," *IEEE Transactions on Biomedical Engineering*, 2019.

[424] I. H. Lee D, Chen N, "Boosted nonparametric hazards with time-dependent covariates," *SSRN 2906586*, 2017.

[425] H. Suresh, J. J. Gong, and J. V. Guttag, "Learning tasks for multitask learning: Heterogenous patient populations in the icu," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, Conference Proceedings, pp. 802–810.

[426] R. Yu, Y. Zheng, R. Zhang, Y. Jiang, and C. C. Poon, "Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients," *IEEE journal of biomedical and health informatics*, 2019.

[427] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[428] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," *arXiv preprint arXiv:1906.00295*, 2019.

[429] J. Oh, J. Wang, and J. Wiens, "Learning to exploit invariances in clinical time-series data using sequence transformer networks," *arXiv preprint arXiv:1808.06725*, 2018.

[430] S. Lohit, Q. Wang, and P. Turaga, "Temporal transformer networks: Joint learning of invariant and discriminative time warping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 426–12 435.

[431] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, Conference Proceedings, pp. 1135–1144.

[432] J. Henderson, H. He, B. A. Malin, J. C. Denny, A. N. Kho, J. Ghosh, and J. C. Ho,

"Phenotyping through semi-supervised tensor factorization (psst)," in *AMIA Annual Symposium Proceedings*, vol. 2018. American Medical Informatics Association, 2018, Conference Proceedings, p. 564.

[433] A. H. Gee, D. Garcia-Olano, J. Ghosh, and D. Paydarfar, "Explaining deep classification of time-series data with learned prototypes," *arXiv preprint arXiv:1904.08935*, 2019.

[434] R. Solis Castilla, A. Pakbin, A. Akbari, B. J. Mortazavi, and R. Jafari, "A human-centered wearable sensing platform with intelligent automated data annotation capabilities," in *Proceedings of the International Conference on Internet of Things Design and Implementation*. ACM, 2019, Conference Proceedings, pp. 255–260.

[435] J. M. Bumgarner, C. T. Lambert, A. A. Hussein, D. J. Cantillon, B. Baranowski, K. Wolski, B. D. Lindsay, O. M. Wazni, and K. G. Tarakji, "Smartwatch algorithm for automated detection of atrial fibrillation," *Journal of the American College of Cardiology*, vol. 71, no. 21, pp. 2381–2388, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29535065

[436] A. S. Go, D. Mozaffarian, V. L. Roger, E. J. Benjamin, J. D. Berry, M. J. Blaha, S. Dai, E. S. Ford, C. S. Fox, and S. Franco, "Executive summary: heart disease and stroke statistics–2014 update: a report from the american heart association," *Circulation*, vol. 129, no. 3, pp. 399–410, 2014.

[437] P. Sosner, M. Gayda, O. Dupuy, M. Garzon, C. Lemasson, V. Gremeaux, J. Lalongé, M. Gonzales, D. Hayami, and M. Juneau, "Ambulatory blood pressure reduction following high-intensity interval exercise performed in water or dryland condition," *Journal of the American Society of Hypertension*, vol. 10, no. 5, pp. 420–428, 2016.

[438] M. Sipola-Leppänen, R. Karvonen, M. Tikanmäki, H.-M. Matinolli, S. Martikainen, A.-K. Pesonen, K. Räikkönen, M.-R. Järvelin, P. Hovi, and J. G. Eriksson, "Ambulatory blood pressure and its variability in adults born preterm," *Hypertension*, vol. 65, no. 3, pp. 615–621, 2015.

[439] G. Torres, M. Sánchez-de-la Torre, M. Martínez-Alonso, S. Gomez, O. Sacristan, J. Cabau, and F. Barbe, "Use of ambulatory blood pressure monitoring for the screening of obstructive sleep apnea," *The Journal of Clinical Hypertension*, vol. 17, no. 10, pp. 802–809, 2015.

[440] J. A. Staessen, L. Thijs, R. Fagard, E. T. O'brien, D. Clement, P. W. De Leeuw, G. Mancia, C. Nachev, P. Palatini, and G. Parati, "Predicting cardiovascular risk using conventional vs ambulatory blood pressure in older patients with systolic hypertension," *Jama*, vol. 282, no. 6, pp. 539–546, 1999.

[441] E. O'Brien, R. Asmar, L. Beilin, Y. Imai, G. Mancia, T. Mengden, M. Myers, P. Padfield, P. Palatini, and G. Parati, "Practice guidelines of the european society of hypertension for clinic, ambulatory and self blood pressure measurement," *Journal of hypertension*, vol. 23, no. 4, pp. 697–701, 2005.

[442] T. G. Pickering, G. D. James, C. Boddie, G. A. Harshfield, S. Blank, and J. H. Laragh, "How common is white coat hypertension?" *Jama*, vol. 259, no. 2, pp. 225–228, 1988.

[443] T. G. Pickering, G. A. Harshfield, H. D. Kleinert, S. Blank, and J. H. Laragh, "Blood pressure during normal daily activities, sleep, and exercise: comparison of values in normal and hypertensive subjects," *Jama*, vol. 247, no. 7, pp. 992–996, 1982.

[444] R. Sega, R. Facchetti, M. Bombelli, G. Cesana, G. Corrao, G. Grassi, and G. Mancia, "Prognostic value of ambulatory and home blood pressures compared with office blood pressure in the general population: follow-up results from the pressioni arteriose monitorate e loro associazioni (pamela) study," *Circulation*, vol. 111, no. 14, pp. 1777–1783, 2005.

[445] P. Palatini, M. Winnicki, M. Santonastaso, L. Mos, D. Longo, V. Zaetta, M. D. Follo, T. Biasion, and A. C. Pessina, "Prevalence and clinical significance of isolated ambulatory hypertension in young subjects screened for stage 1 hypertension," *Hypertension*, vol. 44, no. 2, pp. 170–174, 2004.

[446] T. Ohkubo, A. Hozawa, J. Yamaguchi, M. Kikuya, K. Ohmori, M. Michimata, M. Mat-

subara, J. Hashimoto, H. Hoshi, and T. Araki, "Prognostic significance of the nocturnal decline in blood pressure in individuals with and without high 24-h blood pressure: the ohasama study," *Journal of hypertension*, vol. 20, no. 11, pp. 2183–2189, 2002.

[447] J. Mallion, N. Genes, L. Vaur, P. Clerson, B. Vaisse, G. Bobrie, and G. Chatellier, "Blood pressure levels, risk factors and antihypertensive treatments: lessons from the sheaf study," *Journal of human hypertension*, vol. 15, no. 12, pp. 841–848, 2001.

[448] E. O'Brien, "Dippers and non-dippers," *Lancet*, vol. 2, p. 397, 1988.

[449] P. Verdecchia, F. Angeli, C. Borgioni, R. Gattobigio, and G. Reboldi, "Ambulatory blood pressure and cardiovascular outcome in relation to perceived sleep deprivation," *Hypertension*, vol. 49, no. 4, pp. 777–783, 2007.

[450] E. Dolan, A. Stanton, L. Thijs, K. Hinedi, N. Atkins, S. McClory, E. D. Hond, P. McCormack, J. A. Staessen, and E. O'Brien, "Superiority of ambulatory over clinic blood pressure measurement in predicting mortality: the dublin outcome study," *Hypertension*, vol. 46, no. 1, pp. 156–161, 2005.

[451] J. R. Banegas, L. M. Ruilope, A. de la Sierra, E. Vinyoles, M. Gorostidi, J. J. de la Cruz, G. Ruiz-Hurtado, J. Segura, F. Rodríguez-Artalejo, and B. Williams, "Relationship between clinic and ambulatory blood-pressure measurements and mortality," *New England Journal of Medicine*, vol. 378, no. 16, pp. 1509–1520, 2018.

[452] J. E. Schwartz, M. M. Burg, D. Shimbo, J. E. Broderick, A. A. Stone, J. Ishikawa, R. Sloan, T. Yurgel, S. Grossman, and T. G. Pickering, "Clinic blood pressure underestimates ambulatory blood pressure in an untreated employer-based us populationclinical perspective: Results from the masked hypertension study," *Circulation*, vol. 134, no. 23, pp. 1794–1807, 2016.

[453] M. R. Irvin, J. N. Booth III, M. Sims, A. P. Bress, M. Abdalla, D. Shimbo, D. A. Calhoun, and P. Muntner, "The association of nocturnal hypertension and nondipping blood pressure with treatment-resistant hypertension: The jackson heart study," *The Journal of Clinical Hypertension*, vol. 20, no. 3, pp. 438–446, 2018.

[454] C. Cuspidi, V. Giudici, F. Negri, and C. Sala, "Nocturnal nondipping and left ventricular hypertrophy in hypertension: an updated review," *Expert review of cardiovascular therapy*, vol. 8, no. 6, pp. 781–792, 2010.

[455] G. Mancia and P. Verdecchia, "Clinical value of ambulatory blood pressure: evidence and limits," *Circulation research*, vol. 116, no. 6, pp. 1034–1045, 2015.

[456] S. S. Thomas, V. Nathan, C. Zong, K. Soundarapandian, X. Shi, and R. Jafari, "Biowatch: A noninvasive wrist-based blood pressure monitor that incorporates training techniques for posture and subject variability," *IEEE journal of biomedical and health informatics*, vol. 20, no. 5, pp. 1291–1300, 2016.

[457] B. Ibrahim, J. McMurray, and R. Jafari, "A wrist-worn strap with an array of electrodes for robust physiological sensing," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 4313–4317.

[458] B. Ibrahim and R. Jafari, "Continuous blood pressure monitoring using wrist-worn bio-impedance sensors with wet electrodes," in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2018, pp. 1–4.

[459] N. Luo, W. Dai, C. Li, Z. Zhou, L. Lu, C. C. Poon, S.-C. Chen, Y. Zhang, and N. Zhao, "Flexible piezoresistive sensor patch enabling ultralow power cuffless blood pressure measurement," *Advanced Functional Materials*, vol. 26, no. 8, pp. 1178–1187, 2016.

[460] Y.-L. Zheng, B. P. Yan, Y.-T. Zhang, and C. C. Poon, "An armband wearable device for overnight and cuff-less blood pressure measurement," *IEEE transactions on biomedical engineering*, vol. 61, no. 7, pp. 2179–2186, 2014.

[461] G. S. Stergiou, B. Alpert, S. Mieke, R. Asmar, N. Atkins, S. Eckert, G. Frick, B. Friedman, T. Graßl, T. Ichikawa *et al.*, "A universal standard for the validation of blood pressure measuring devices: Association for the advancement of medical instrumentation/european society of hypertension/international organization for standardization

(AAMI/ESH/ISO) collaboration statement," *Hypertension*, vol. 71, no. 3, pp. 368–374, 2018.

[462] L. Ghazi, M. M. Safford, Y. Khodneva, W. T. O'Neal, E. Z. Soliman, and S. P. Glasser, "Gender, race, age, and regional differences in the association of pulse pressure with atrial fibrillation: the reasons for geographic and racial differences in stroke study," *Journal of the American Society of Hypertension*, vol. 10, no. 8, pp. 625–632, 2016.

[463] G. E. McVeigh, C. W. Bratteli, D. J. Morgan, C. M. Alinder, S. P. Glasser, S. M. Finkelstein, and J. N. Cohn, "Age-related abnormalities in arterial compliance identified by pressure pulse contour analysis: aging and arterial compliance," *Hypertension*, vol. 33, no. 6, pp. 1392–1398, 1999.

[464] M. Kachuee, M. M. Kiani, H. Mohammadzade, and M. Shabany, "Cuffless blood pressure estimation algorithms for continuous health-care monitoring," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 4, pp. 859–869, 2016.

[465] Y. Wang, Z. Liu, and S. Ma, "Cuff-less blood pressure measurement from dual-channel photoplethysmographic signals via peripheral pulse transit time with singular spectrum analysis," *Physiological measurement*, vol. 39, no. 2, p. 025010, 2018.

[466] P. Nabeel, S. Karthik, J. Joseph, and M. Sivaprakasam, "Arterial blood pressure estimation from local pulse wave velocity using dual-element photoplethysmograph probe," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 6, pp. 1399–1408, 2018.

[467] V. Chandrasekaran, R. Dantu, S. Jonnada, S. Thiyagaraja, and K. P. Subbu, "Cuffless differential blood pressure estimation using smart phones," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 1080–1089, 2012.

[468] J. Chen, K. Chen, X. Chen, X. Qiu, and X. Huang, "Exploring shared structures and hierarchies for multiple nlp tasks," *arXiv preprint arXiv:1808.07658*, 2018.

[469] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "Nas-unet: Neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, pp. 44 247–44 257, 2019.

[470] L. Faes, S. K. Wagner, D. J. Fu, X. Liu, E. Korot, J. R. Ledsam, T. Back, R. Chopra, N. Pontikos, and C. Kern, "Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study," *The Lancet Digital Health*, vol. 1, no. 5, pp. e232–e242, 2019.

[471] P. Fonseca, J. Mendoza, J. Wainer, J. Ferrer, J. Pinto, J. Guerrero, and B. Castaneda, "Automatic breast density classification using a convolutional neural network architecture search procedure," in *Medical Imaging 2015: Computer-Aided Diagnosis*, vol. 9414. International Society for Optics and Photonics, Conference Proceedings, p. 941428.

[472] P. Balaprakash, R. Egele, M. Salim, S. Wild, V. Vishwanath, F. Xia, T. Brettin, and R. Stevens, "Scalable reinforcement-learning-based neural architecture search for cancer deep learning research," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, Conference Proceedings, pp. 1–33.

[473] G. Liu, R. Ma, and Q. Hao, "A reinforcement learning based design of compressive sensing systems for human activity recognition," in *2018 IEEE SENSORS*. IEEE, Conference Proceedings, pp. 1–4.

[474] P. Jerrard-Dunne, A. Mahmud, and J. Feely, "Circadian blood pressure variation: relationship between dipper status and measures of arterial stiffness," *Journal of hypertension*, vol. 25, no. 6, pp. 1233–1239, 2007.

[475] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[476] E. Jovanov, "Preliminary analysis of the use of smartwatches for longitudinal health monitoring," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 865–868.

[477] V. Ahanathapillai, J. D. Amor, Z. Goodwin, and C. J. James, "Preliminary study on activity monitoring using an android smart-watch," *Healthcare technology letters*,

vol. 2, no. 1, pp. 34–39, 2015.

[478] E. Årsand, M. Muzny, M. Bradway, J. Muzik, and G. Hartvigsen, "Performance of the first combined smartwatch and smartphone diabetes diary application study," *Journal of diabetes science and technology*, vol. 9, no. 3, pp. 556–563, 2015.

[479] G. Schiboni and O. Amft, "Automatic dietary monitoring using wearable accessories," in *Seamless Healthcare Monitoring*. Springer, 2018, pp. 369–412.

[480] J. C. Rawstorn, N. Gant, A. Meads, I. Warren, and R. Maddison, "Remotely delivered exercise-based cardiac rehabilitation: design and content development of a novel mhealth platform," *JMIR mHealth and uHealth*, vol. 4, no. 2, p. e57, 2016.

[481] A. C. Timmons, T. Chaspari, S. C. Han, L. Perrone, S. S. Narayanan, and G. Margolin, "Using multimodal wearable technology to detect conflict among couples," *Computer*, no. 3, pp. 50–59, 2017.

[482] H. M. Raafat, M. S. Hossain, E. Essa, S. Elmougy, A. S. Tolba, G. Muhammad, and A. Ghoneim, "Fog intelligence for real-time iot sensor data analytics," *IEEE Access*, vol. 5, pp. 24 062–24 069, 2017.

[483] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural computation*, vol. 6, no. 2, pp. 181–214, 1994.

[484] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE transactions on neural networks and learning systems*, vol. 23, no. 8, pp. 1177–1193, 2012.

[485] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," *arXiv preprint arXiv:1706.06905*, 2017.

[486] M. Courbariaux, C. Ambroise, C. Dalmasso, M. Szafranski, M. Consortium *et al.*, "A mixture model with logistic weights for disease subtyping with integrated genome association study," 2018.

[487] K. Nithya, M. V. Prathap, and K. R. Babu, "Cluster oriented sensor selection for context-aware internet of things applications," in *International Conference on Intelli-*

gent Data Communication Technologies and Internet of Things. Springer, 2018, pp. 981–988.

[488] S.-Y. Yu, C.-S. Shih, J. Y.-J. Hsu, Z. Huang, and K.-J. Lin, "Qos oriented sensor selection in iot system," in *2014 IEEE International Conference on Internet of Things (iThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom)*. IEEE, 2014, pp. 201–206.

[489] J. Shukla, P. Maiti, and B. Sahoo, "Low latency and energy efficient sensor selection for iot services," in *2018 Technologies for Smart-City Energy Security and Power (ICSESP)*. IEEE, 2018, pp. 1–5.

[490] P. M. Jones, Q. Lonne, P. Talaia, G. J. Leighton, G. G. Botte, S. Mutnuri, and L. Williams, "A straightforward route to sensor selection for iot systems," *Research-Technology Management*, vol. 61, no. 5, pp. 41–50, 2018.

[491] A. Yachir, Y. Amirat, A. Chibani, and N. Badache, "Event-aware framework for dynamic services discovery and selection in the context of ambient intelligence and internet of things," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 1, pp. 85–102, 2015.

[492] G. Yang, L. Xie, M. Mäntysalo, X. Zhou, Z. Pang, L. Da Xu, S. Kao-Walter, Q. Chen, and L.-R. Zheng, "A health-iot platform based on the integration of intelligent packaging, unobtrusive bio-sensor, and intelligent medicine box," *IEEE transactions on industrial informatics*, vol. 10, no. 4, pp. 2180–2191, 2014.

[493] N. Zhu, T. Diethe, M. Camplani, L. Tao, A. Burrows, N. Twomey, D. Kaleshi, M. Mirmehdi, P. Flach, and I. Craddock, "Bridging e-health and the internet of things: The sphere project," *IEEE Intelligent Systems*, vol. 30, no. 4, pp. 39–46, 2015.

[494] D. Roggen, G. Troester, P. Lukowicz, A. Ferscha, J. d. R. Millán, and R. Chavarriaga, "Opportunistic human activity and context recognition," *Computer*, vol. 46, no. 2, pp. 36–45, 2012.

[495] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukow-

icz, D. Bannach, G. Pirkl, A. Ferscha *et al.*, "Collecting complex activity datasets in highly rich networked sensor environments," in *Networked Sensing Systems (INSS), 2010 Seventh International Conference on.* IEEE, 2010, pp. 233–240.

[496] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[497] F. G. Cozman, I. Cohen, and M. C. Cirelo, "Semi-supervised learning of mixture models," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003, pp. 99–106.

[498] S. E. Chazan, S. Gannot, and J. Goldberger, "Training strategies for deep latent models and applications to speech presence probability estimation," in *International Conference on Latent Variable Analysis and Signal Separation.* Springer, 2018, pp. 319–328.

[499] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition." in *Ijcai*, vol. 15, 2015, pp. 3995–4001.

[500] H. B. Ravn, O. K. L. Helgestad, and J. E. Møller, "Intravascular Microaxial Left Ventricular Assist Device vs Intra-aortic Balloon Pump for Cardiogenic Shock," *JAMA*, vol. 324, no. 3, pp. 302–303, 07 2020. [Online]. Available: https://doi.org/10.1001/jama.2020.7557

[501] J. A. Rizzo and H. Dove, "Intravascular Microaxial Left Ventricular Assist Device vs Intra-aortic Balloon Pump for Cardiogenic Shock," *JAMA*, vol. 324, no. 3, pp. 303–303, 07 2020. [Online]. Available: https://doi.org/10.1001/jama.2020.7551

[502] S. S. Dhruva, B. J. Mortazavi, and N. R. Desai, "Intravascular Microaxial Left Ventricular Assist Device vs Intra-aortic Balloon Pump for Cardiogenic Shock–Reply," *JAMA*, vol. 324, no. 3, pp. 303–304, 07 2020. [Online]. Available: https://doi.org/10.1001/jama.2020.7560