UCLA
UCLA Electronic Theses and Dissertations

Title
Towards Fair and Interpretable AI Healthcare predictive Models: from wearable sensors to causal graphs

Permalink
https://escholarship.org/uc/item/1bg486jd

Author
Zhang, Wenhao

Publication Date
2023

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
Los Angeles

Towards Fair and Interpretable AI Healthcare predictive Models: from wearable sensors to causal graphs

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Wenhao Zhang

2023

© Copyright by

Wenhao Zhang

2023

ABSTRACT OF THE DISSERTATION

Towards Fair and Interpretable AI Healthcare predictive Models: from wearable sensors to causal graphs

by

Wenhao Zhang

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2023

Professor Majid Sarrafzadeh, Chair

The rapid expansion of data in the healthcare sector has highlighted the need for powerful and user-friendly artificial intelligence (AI) techniques in the medical field. Although AI toolkits have transformed various areas such as image recognition and natural language processing, their integration into healthcare has been relatively slow. Patient records contain data from a variety of sources, including electronic health records, medical imaging, wearable and ambient biosensors, lab results, and genomics, with the aim of capturing the intricacies of patient health conditions. However, the complex, diverse, and high-dimensional nature of medical datasets creates unique challenges for data analysis and limits the effectiveness and practicality of existing solutions. Additionally, ethical and legal concerns regarding the introduction of medical AI models exist, such as the potential model bias against some minority groups in society, lack of interpretability of some AI algorithms, data privacy problems. Hence, further research is necessary on the development and deployment of medical AI models. In this thesis, we mainly focuses on using machine learning and causal inference to solve applied research problems based on healthcare data for fair and trustworthy medical AI models.

The first part of our work involves utilizing machine learning models and statistical toolkits to construct predictive risk models from a patented remote patient monitoring system. The models are based on a comprehensive set of features that are derived from wearable sensors and bluetooth

beacons. These features provide a clear storyline of the daily activities of the frail population in rehabilitation settings. Additionally, we suggest a deep transfer learning framework to classify arrhythmia heartbeat. The proposed method involves fine-tuning a general-purpose image classifier, ResNet-18, with the MIT-BIH arrhythmia dataset. We managed to train the proposed arrhythmia classifier in accordance with the AAMI EC57 standard to ensure that there was no data leakage during model development. The next aspect of my work in healthcare analytics focuses on imbalanced learning where classes are not equally represented in the medical dataset. This issue can be challenging for machine learning classifiers, often leading to biased predictions favoring the majority class and low accuracy for the minority class. To address this issue, we have introduced a new approach that utilizes a weighted oversampling technique and ensemble boosting method to enhance the accuracy of minority data while maintaining accuracy for the majority class.

The second part of our work mainly focuses on using causal inference to develop fair and interpretable machine learning models. By incorporating causality, the model's interpretability and performance can be improved. Causal relationships are often represented in directed acyclic graphs (DAGs) known as causal graphs, which allow researchers to identify the causes of the outcome variables and eliminate irrelevant factors during modeling through visual inspection. In this thesis, we developed a causal discovery algorithm that identifies causal relationships in healthcare datasets with high dimensionality. The proposed algorithm treats causal discovery as a continuous constrained optimization problem with a polynomial constraint. The optimization objective function evaluates the fit of the data to the estimated causal graph, while the constraint ensures that there is no cycle in the estimated graph.

Another aspect of this thesis involves building a causal model to estimate the conversion rate (CVR) in e-commerce recommender systems. This task is particularly challenging in industrial settings due to two major issues: user self-selection leading to selection bias, and data sparsity resulting from rare click events. Our work addresses these challenges by leveraging inverse propensity weight techniques to adjust for selection bias in the final estimation. Additionally, our methods are based on the multi-task learning framework, which can mitigate the impact of data sparsity.

The dissertation of Wenhao Zhang is approved.

Ramin Ramezani

Arash Naeim

Ali Mosleh

Cho-Jui Hsieh

Majid Sarrafzadeh, Committee Chair

University of California, Los Angeles

2023

For my beloved family, friends, and mentors,

who believed in me every step of the way

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am grateful to my Ph.D. advisors, Dr. Ramin Ramezani and Dr. Arash Naeim, for their invaluable guidance, unbelievable support, and exceptional patience throughout my doctoral studies. Their commitment to thorough research, critical mindset, and extensive knowledge have profoundly impacted both my academic pursuits and daily life experiences.

I would like to offer my special thanks to center for smart health (CSH) and the department of Computer Science at UCLA for generously funding me with PhD fellowship and graduate student researcher position . Their support has enabled me to pursue and contribute to my proposed work effectively.

I express my gratitude to my committee members, Dr. Majid Sarrafzadeh, Dr. Ramin Ramezani Dr. Arash Naeim, Dr. Ali Mosleh, Dr. Cho-Jui Hsieh, for their invaluable and constructive input on my dissertation. Additionally, I extend my thanks to Prof. John (Junghoo) Cho for his valuable perspectives regarding the conceptualization of causal discovery as an continuous constraint optimization problem.

I extend my heartfelt appreciation to my esteemed colleagues at the Center for Smart Health and The B. John Garrick Institute for the Risk Sciences, as well as the friends I have had the pleasure of collaborating with. A special thanks to Minh Cao, Bryan P Bendnarski, Dr. Akash Deep Singh, Wentian Bao, Dr. Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, Vishwa Karia, Dr. Karlton Wong, Dr. David Elashoff, and Dr. Pamela Roberts, Dr. Ke Yang, Dr. Yining Dong, Dr. Yantao Zhu, Dr. Tengfei Wang, Dr. Wadie Chalgham, Dr. Xue Yang, Dr. Zhuo Xie, Dr. John Shen. It has been an incredible privilege to collaborate with such brilliant minds. Throughout the past few years, I have thoroughly enjoyed exchanging ideas and engaging in discussions with all of you.

Lastly, I am profoundly grateful to my parents and my wife for their unconditional love and support. Their constant presence and encouragement have served as a guiding light during both the highs and lows of my journey. Without them, this achievement would not have been possible.

VITA

2009–2013 B.S. (Electrical Engineering), Harbin Engineering University.

2013–2015 M.S. (Electrical Engineering), University of Southern California.

2016–2017 M.S. (Computer Science), University of Southern California.

2017–2023 Ph.D. candidate (Computer Science), University of Southern California.

PUBLICATIONS

Temporal convolutional networks and data rebalancing for clinical length of stay and mortality prediction. Scientific Reports 12.1 (2022): 21247.

Range of Motion Sensors for Monitoring Recovery of Total Knee Arthroplasty. 2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks (BSN). IEEE, 2022.

ECG heartbeat classification using deep transfer learning with convolutional neural network and STFT technique. arXiv preprint arXiv:2206.14200 (2022).

Causal Inference in Medicine and in Health Policy: A Summary. HANDBOOK ON COMPUTER LEARNING AND INTELLIGENCE: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation. 2022. 263-302.

Physical Activity Behavior of Patients at a Skilled Nursing Facility: Longitudinal Cohort Study. JMIR mHealth and uHealth 10.5 (2022): e23887.

The Derivation of an ICD-10-based Trauma-related Mortality Model Utilizing Machine Learning. The Journal of Trauma and Acute Care Surgery (2021).

Large-scale causal approaches to debiasing post-click conversion rate estimation with multi-task learning. Proceedings of The Web Conference 2020. 2020.

Gensample: A genetic algorithm for oversampling in imbalanced datasets. arXiv preprint arXiv:1910.10806 (2019).

WOTBoost: Weighted oversampling technique in boosting for imbalanced learning. 2019 IEEE International Conference on Big Data (Big Data). IEEE, 2019

A combination of indoor localization and wearable sensor–based physical activity recognition to assess older patients undergoing subacute rehabilitation: Baseline study results. JMIR mHealth and uHealth 7.7 (2019): e14090.

Using Smart Watch Sensing in At-Risk Populations (SARP) in a Sub-Acute Rehabilitation Center. Archives of Physical Medicine and Rehabilitation 99.12 (2018): e217.

CHAPTER 1

Introduction: healthcare and machine learning

The rapid expansion of data in the healthcare sector has highlighted the need for powerful and user-friendly artificial intelligence (AI) techniques in the medical field. Although AI toolkits have transformed various areas such as image recognition and natural language processing, their integration into healthcare has been relatively slow. Patient records contain data from a variety of sources, including electronic health records, medical imaging, wearable and ambient biosensors, lab results, and genomics, with the aim of capturing the intricacies of patient health conditions. However, the complex, diverse, and high-dimensional nature of medical datasets creates unique challenges for data analysis and limits the effectiveness and practicality of existing solutions. Additionally, ethical and legal concerns regarding the introduction of medical AI models exist, such as the potential model bias against some minority groups in society, lack of interpretability of some AI algorithms, data privacy problems. Hence, further research is necessary on the development and deployment of medical AI models. In this thesis, we mainly focuses on using machine learning and causal inference to solve applied research problems based on healthcare data for fair and trustworthy medical AI models.

1.1 Background and motivation

The advancement in remote patient monitoring systems coupled with AI and machine learning models have provided healthcare professionals with the ability to construct AI predictive risk models using extensive patient data. These medical AI models enable physicians to receive early warnings about patients whose health conditions are deteriorating. As a result, timely medication

and interventions can be administered, leading to reduced hospitalization costs and improved health outcomes [10, 11, 12, 13, 14]. Nonetheless, the raw data obtained from healthcare analysis often requires preprocessing and transformation before it can be effectively analyzed. Machine learning practitioners are faced with the task of handling challenges such as imbalanced data [15, 16], missing data [17], feature selection [18], and other related issues in order to make the data suitable for analysis. The first part of my thesis centers around the aforementioned problems in healthcare data analysis.

During the analysis of healthcare data, healthcare practitioners often pose causal questions which cannot be addressed solely by correlation-based methods. For instance, research questions arise in healthcare data analysis, such as "What is the effectiveness of a specific medication in treating a particular disease?" or "Why do certain statistically significant features fail to predict the outcome?" [19, 20]. These causal questions share a common focus on cause-effect relationships [21, 5]. David Hume, an eighteenth-century philosopher, defined causation in terms of counterfactuals: A is considered a cause of B if, 1) B is consistently observed to follow A, and 2) if A had not occurred, B would not have existed [22]. Judea Pearl, a Turing Award recipient and a pioneer in causal inference research, asserts that a causal learner and human-like AI must acquire three levels of cognitive abilities: seeing, doing, and imagining. Pearl describes these levels as the "ladder of causation" [23]. AI agents become more human-like as they ascend this ladder: starting with passive observation (seeing) of correlations and associations in the data, progressing to active manipulation and intervention (doing) within the systems and models under consideration, and ultimately reaching the highest level of the ladder, which involves reasoning about an unseen world through retrospective analysis (imagining).

I begin my work on causality research by firstly examining the application of causal inference in recommender systems. This field primarily focuses on understanding the cause-and-effect relationships in interventions. For instance, questions like "Will users actually purchase the recommended items on an e-commerce website?" or "Would users click on recommended ads?" are at the core of this inquiry. Consequently, many of the research questions in recommender systems revolve around causal-effect relations, and the mathematical models and methods employed can be extended to

the healthcare research domain [24]. Additionally, I focus on investigating algorithms that can autonomously uncover causal-effect relationships within medical datasets. The identification of such causal relations can be highly valuable in addressing inquiries like "Which factors contribute to the progression of patients' diseases?" This, in turn, can lead to significant enhancements in model interpretability [21, 25]. Moreover, employing causal inference in healthcare data analysis offers the advantage of selecting features that possess causal-effect relationships with predictive outcomes [26, 27].

1.2 Objectives and Main Contributions

The first part of our work involves utilizing machine learning models and statistical toolkits to construct predictive risk models for a patented remote patient monitoring system. The models are based on a comprehensive set of features that are derived from wearable sensors and bluetooth beacons. These features provide a clear storyline of the daily activities of the frail population in rehabilitation settings. Additionally, we suggest a deep transfer learning framework to classify arrhythmia heartbeat. The proposed method involves fine-tuning a general-purpose image classifier, ResNet-18, with the MIT-BIH arrhythmia dataset. We took care to train the proposed arrhythmia classifier in accordance with the AAMI EC57 standard to ensure that there was no data leakage during model development. The next aspect of my work in healthcare analytics focuses on imbalanced learning where classes are not equally represented in the medical dataset. This issue can be challenging for machine learning classifiers, often leading to biased predictions favoring the majority class and low accuracy for the minority class. To address this issue, we have introduced a new approach that utilizes a weighted oversampling technique and ensemble boosting method to enhance the accuracy of minority data while maintaining accuracy for the majority class.

The second part of our work mainly focus on using causal inference to develop fair and interpretable machine learning models. By incorporating causality, the model's interpretability and performance can be improved. Causal relationships are often represented in directed acyclic graphs (DAGs) known as causal graphs, which allow researchers to identify the causes of the

outcome variables and eliminate irrelevant factors during modeling through visual inspection. In this thesis, we developed a causal discovery algorithm that identifies causal relationships in healthcare datasets with high dimensionality. The proposed algorithm treats causal discovery as a continuous constrained optimization problem with a polynomial constraint. The optimization objective function evaluates the fit of the data to the estimated causal graph, while the constraint ensures that there is no cycle in the estimated graph.

Another aspect of this thesis involves building a causal model to estimate the conversion rate (CVR) in e-commerce recommender systems. This task is particularly challenging in industrial settings due to two major issues: user self-selection leading to selection bias, and data sparsity resulting from rare click events. Our work addresses these challenges by leveraging inverse propensity weight techniques to adjust for selection bias in the final estimation. Additionally, our methods are based on the multi-task learning framework, which can mitigate the impact of data sparsity.

1.3 Thesis outline

Chapter 1,2, and 3 provide background knowledge and literature review of our research:

- Chapter 1: **Introduction: healthcare and machine learning**.
  I provide a brief description of the objectives, motivations and contributions.

- Chapter 2: **Background: data analytics in healthcare**.
  I discuss the background to the patented remote patient monitoring system–Sensing at Risk Population–used in this chapter. I also provide a literature reivew on imbalanced learning and arrhythmia classification using deep transfer learning with electrocardiogram dataset.

- Chapter 3: **Background: healthcare analysis with causal inference**. In this chapter, I provide a brief overview of the background information related to causal inference. This includes an explanation of the concept of causality, the fundamental principles of structural causal models, causal graphs, and intervention using do-calculus. Additionally, I explore the

issues of spurious correlation and confounding, including the Simpson paradox. Finally, I discuss the methods for discovering causal relationships from data.

- Chapter 4: **Healthcare data analytics in remote patient monitoring**. This chapter showcases how the Sensing At-Risk Population system, which is a patented remote patient health monitoring system [28], can provide a deeper insight into the health conditions of patients by using wearable technology with sophisticated physical activity tracking algorithms that are specifically designed for geriatric patients. This study has a twofold aim. Firstly, to examine the ability of a combination of physical activity and indoor location features, extracted at baseline, on a cohort of 154 rehabilitation-dwelling patients to discriminate between subacute care patients who are re-admitted to the hospital versus the patients who are able to stay in a community setting. Secondly, to observe longitudinal changes of sensor-based physical activity and indoor localization features of patients receiving rehabilitation at a skilled nursing facility and investigate if the sensor-based longitudinal changes can complement patients' changes captured by therapist assessments over the course of rehabilitation in the skilled nursing facility.

- Chapter 5: **Imbalanced learning in healthcare analytic**. This chapter introduces a new approach called WOT-Boost, which combines a Weighted Oversampling Technique with an ensemble Boosting method. The aim of this method is to enhance the accuracy of minority data classification without compromising the accuracy of the majority class.

- Chapter 6: **Arrythmia classification using deep transfer learning using with eletrocardiogram datasets**. In this chapter, a new deep transfer learning framework is introduced, which is designed for classification tasks with limited training data. The proposed approach involves fine-tuning a general-purpose image classifier ResNet-18 with the MIT-BIH arrhythmia dataset while adhering to the AAMI EC57 standard. The study also inv estigates several existing deep learning models that have failed to avoid data leakage as per the AAMI guidelines. Furthermore, the impact of various data split methods on model performance is also examined and compared.

- Chapter 7: **Causal models to debiasing post-click conversion rate estimation with multi-task learning**. In this chapter, I propose two principled, efficient and highly effective CVR estimators for industrial CVR estimation. The proposed models approach the CVR estimation from a causal perspective and account for the causes of missing not at random. In addition, the proposed methods are based on the multi-task learning framework and mitigate the data sparsity issue. Extensive experiments on industrial-level datasets show that the proposed methods outperform the state-of-the-art CVR models

- Chapter 8: **Causal discovery in high dimension and curse of high dimensionality**. This chapter presents a novel approach to identifying causal relations in high dimensional space by formulating causal discovery as a continuous constraint problem with a polynomial constraint. By utilizing this method, deep learning frameworks like Tensorflow and Pytorch can efficiently solve the problem.

  Finally, I provide a summary of the thesis's main conclusions and propose future research directions for further advancing this study.

- Chapter 9: **Conclusion and future work**.
  In the concluding chapter, I provide a summary of the significant findings and insights obtained from the research. Additionally, I highlight the limitations of the study and suggest future research directions that could help further advance the practical application of the research in real-world scenarios.

CHAPTER 2

Background: data analytics in healthcare

This chapter provides a literature review of concepts involved in various healthcare data analytics projects addressed throughout this thesis. Firstly, it discusses the development of a patented Remote Patient Monitoring (RPM) system known as Sensing At-Risk Population (SARP). The data analyses presented in Chapter 4 is based on data collected from SARP system. Next, the chapter highlights the issue of imbalanced learning. It then presents a comprehensive review arrhythmia classification using electrocardiogram signals thereby reemphasizing the paramount importance of RPMs and related healthcare data analytics challenges.

2.1 Remote patient monitoring

According to the most recent census statistics, by 2050, the population aged 65 years and older is projected to double in size to 83.7 million in the United States [29]. With the increase of this geriatric population, health care utilization will increase dramatically, with a concomitant demand for rehabilitation and in-home care after hospitalization [30]. Finding the best way to support patients during rehabilitation, both at facilities and in home, without compromising patient safety is considered to be a significant challenge. The importance of patient safety and rehabilitation has highlighted the need for constant vigilance and fostered methodologies by which patients can be remotely monitored [30, 31, 32, 33, 34, 35, 36].

Numerous studies have investigated the effectiveness of remote patient health monitoring, some suggesting the potential for such technologies to reduce the overall re-admission cost [37]. With the advent of wearable devices in recent years, remote health monitoring has evolved and

drawn attention, mainly by utilizing physical activity trackers. It is widely assumed that a physical activity regimen implies behavioral patterns that can affect health outcomes. Hence, tracking these patterns and leveraging them may allow the prediction of harmful outcomes, such as falls, in a timely manner. Moreover, tracking individuals' personalized behavioral patterns may allow for the creation of actionable messages to patients and caregivers to improve patient health and outcomes [38]. The purpose of this study was to investigate the physical activity and indoor localization features obtained from our remote patient monitoring system, Sensing At-Risk Population (SARP) [30, 39, 40, 41, 28]. This study reports on SARP sensor–based markers for rehabilitation screening within a geriatric population, exploring if SARP can be used to prospectively distinguish between at-risk patients in a subacute rehabilitation environment.

2.1.1 Sensing At-Risk Population System Overview

Details of the system architecture with proximity-based sensors (beacons) and a Bluetooth-enabled smartwatch as its main components can be found in the study by Moatamed et al [30] and the patent application by Ramezani et al [28]. Building models for physical activity tracking and indoor localization was based on data collected using (1) commercially available Sony SmartWatch 3 with built-in EM7180 $2g$ triaxial accelerometer, 420 mA battery, and BCM43340 Bluetooth module and (2) proximity beacons (MCU ARM Cortex-M4 32-bit processor with floating-point unit). To build the activity tracking and indoor localization models of SARP system, patients were consented on admission to a subacute care rehabilitation center in Los Angeles.

2.1.2 Bluetooth Low Energy Beacons and Indoor Localization

Beacons broadcast their presence to Bluetooth-enabled devices. Utilizing the beacons' Received Signal Strength Indicator (RSSI) values using smartwatches, the SARP system calculates the proximity of the watch to each beacon, thereby inferring the indoor location of the patient wearing that watch. BLE beacons (bluetooth low-energy sensors) have become popular in gathering contextual awareness because of durability and low cost. When used in health care, however, validating

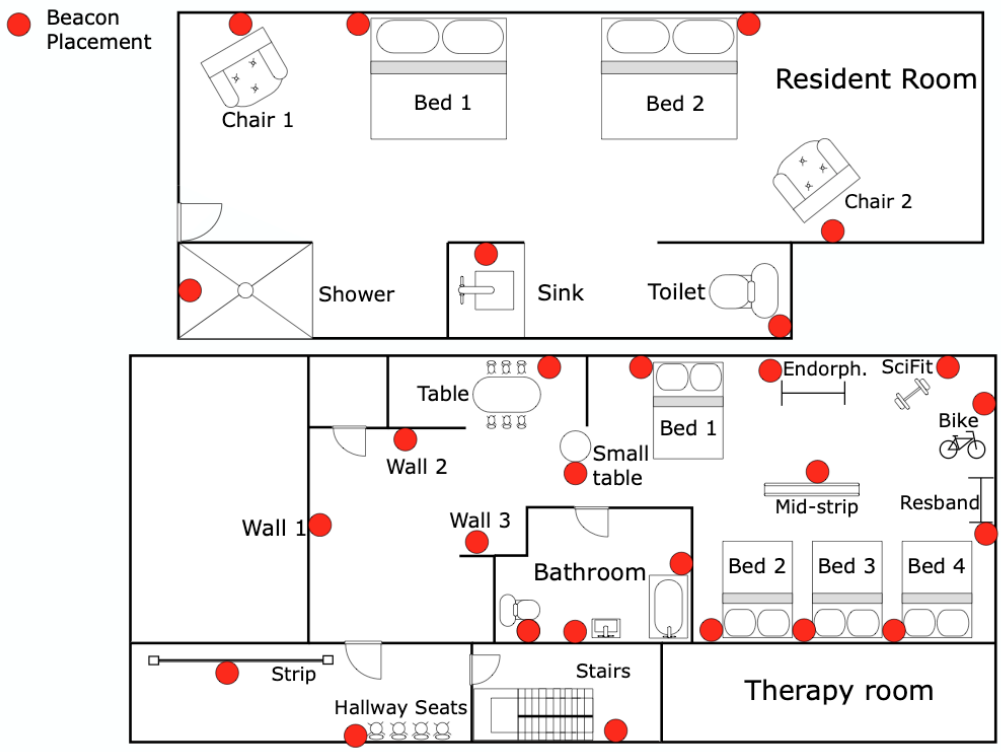


Figure 2.1: Subacute rehabilitation facility map: resident room on top and therapy room at the bottom with locations of mounted beacons shown in red.

reliability and accuracy of their location information is paramount. Beacons are highly susceptible to diffraction, multipath propagation, angle-of-arrival, lack of line-of-sight, and absorption by the human body. In this project, because locations of interest were within close proximity, we considered RSSI values ranged between −50 dBm to −100 dBm. The average RSSI within the line-of-sight, measured by the watch at 1 feet distance, was −66 dBm. To achieve the best accuracy with respect to locations of interest, shown in Figure 2.1, considering beacons hardware specification was crucial. Beacon's antenna configuration and the proximity of locations heavily influence the accuracy of indoor localization. Hence, to achieve a high indoor localization accuracy, it was essential to refine beacon placements iteratively. Moreover, in the rehabilitation facility shown in Figure 2.1, we empirically learned to set the transmission power to −12 dBm and the transmission interval to 250 ms. In studies by Bouchard et al [39, 40, 41], we proposed a few methods and considerations that can help enhance the indoor localization accuracies. A summary of the ground

truth testing executed at the rehabilitation facility shown in Figure 2.1, with an overall accuracy $> 80\%$, can be found in a study by Moatamed et al [30].

2.1.3 Accelerometer Data Processing and Physical Activity Parameters

To infer physical activity of patients in this study, 3-axis raw acceleration signal sampled at 16 Hz was extracted, and the signal magnitude (SM) was initially calculated according to Figure 2.2, equation (1), where *acc* indicates acceleration force around each axis in g units including gravity ($1g = 9.81m/s^2$). The range of the acquired signal is $\pm 2g$. Batches of 160 samples (window size of 10 seconds) were fed to a fifth order Butterworth band-pass filter with cut-off frequencies of 0.5 and 8 Hz. The filtering limited the signal to highlight the frequencies that are most representative of human motion while eliminating the direct current component. Various window sizes ranging from 4 to 12.8 seconds with different overlapping implementations have been used in different studies [42]. These characteristics are normally chosen empirically based on feature extraction, activity labeling, and other annotation factors. In this study, a window size of 10 seconds was used with a 1-second overlap [30]. After preprocessing the accelerometer data, the next step was to infer human activity (positioning) and to later translate the positioning into a quantifiable metric. However, quantifying the physical activity can be deemed challenging and will be discussed after a brief description of physical activity classification.

A decade has passed since the advent of commercially available low-cost, light-weight accelerometers. The enthusiasm about their potential in extracting physical patterns to usually, but not exclusively, improve health outcomes has led researchers to master the techniques of activity recognition [42, 43]. Some researchers have even tried to infer activity intensities and predict energy consumption by comparing accelerometer patterns with measured metabolic equivalents [44, 45, 46]. Despite significant and impressive outcomes, the triumph is mostly based on analyzing small cohorts, or often a homogeneous group of people, with similar age or health conditions. Training and testing datasets in most studies are normally collated from people following a certain protocol, whereas in real life, human movements are intertwined, that is, the sequence of movements does not always form a same pattern. As such, the performance of various activity

recognition algorithms/approaches applied to real-world scenarios should be taken with a grain of salt [42, 43, 45, 46, 47]. The following factors are influential in any human activity tracking algorithm: (1) diversity of human movement habits; (2) variety of human disabilities needing different assistive devices, yielding distinct movement patterns; (3) deficiencies of machine learning algorithms in building one-size-fits-all model; and (4) limitations to distinguish particular motions due to accelerometer placement, for instance, classifying sitting still and laying down with sensor on wrist versus waist [42, 47]. To reduce the negative effect of the mentioned factors, this study uses a combination of classifications in 3 steps according to algorithm shown in Figure 2.3. Time and frequency domain characteristics of the signal (main, median, variance, skewness, kurtosis, peak frequency, and peak power) were used as features. SARP initially categorizes activities broadly into walking and stationary.

Walking embodies active status, and when stationary, the classifier separates brisk (active) and idle (nonactive) movements and later classifies postures into sedentary, standing, and laying down. Both Tables 2.1 and 2.2 depict the summary of physical activity (positioning) classifiers' 10-fold cross-validation results built on 50 patients over approximately 22 hours of collated data at subacute care rehabilitation center in Los Angeles. The algorithms were later validated and refined over the course of 6 months of ground truth testing at the same skilled nursing facility.

2.1.4 Step Counts Versus Raw Accelerometer Assessment

The next stage was to find a way to quantify the difference between different activity status. Step counting is a common way that has long been used to quantify the ambulatory physical activity. However, similar to activity recognition approaches explained earlier, the accuracy of step counters is often the subject of debate among researchers. Comprehensive studies with contradictory results on the accuracy of pedometers and wearable accelerometers can be found in the studies by Crouter et al [48], Mammen et al [49], and Case et al [50]. What is rather clear in using step counters/pedometers is their efficacy in quantifying ambulatory activities and not stationary. For step counters to be more accurate, a user is required to satisfy a minimum walking speed that is often mentioned in the literature as 67 m/min or even higher [51, 52]. Therefore, step counters are less likely to

Number	Equation	Summary
(1)	$SM = \sqrt{acc_x^2 + acc_y^2 + acc_z^2}$	Signal Magnitude
(2)	$MAD = \frac{1}{N}\sum_{i=1}^{N=160} \lvert x_i - x_{ave} \rvert$	MAD of accelerometer magnitude signal
(3)	$d = v_0 t + \frac{1}{2} at^2 = \frac{1}{2}\ 0.02\ (10^2) = 1$	Hand displacement in 10 seconds when threshold on MAD = 0.02 m/s^2
(4)	$Time_{spent}(T\%)$	Time spent in *walking*, *sitting*, *standing*, *laying*, or in *locations of interest* (Table 3) divided by *uptime*
(5)	$energy_{intesity}(E)$	Energy spent in *walking, sitting, standing, laying,* or in *locations of interest* divided by their corresponding *time spent*. In addition to energy intensity spent at each location, we calculated the total energy intensity in *resident* room and *therapy* room. Energy intensity in resident room, (loc_{res_E}) is $\frac{\sum_i loc_{i_E}}{total\_time\_in\_res}$, where $loc_i \in$ resident room. Energy intensity for therapy room was similarly calculated
(6)	$energy_{percentage}(E\%)$	Total energy spent in *walking, sitting, standing, laying* or in *locations of interest* divided by *total energy* spent in that day.

Figure 2.2: Equations. MAD: mean absolute deviation.

produce accurate assessment for less mobile geriatric population. Besides, they are deemed even less effective in quantifying activities in stationary positions. Most studies assess the accuracy of step counters by asking users to walk on a treadmill, which neglects scenarios in which users are stationary, yet pedometers accumulate step counts because of movements in hand. To account for any movements (stationary and ambulatory), this study calculates mean absolute deviation (MAD) of accelerometer magnitude signal using equation (2), Figure 2.2. MAD calculates the statistical dispersion of acceleration from the mean and its unit is meter per second squared,

where x_i is the SM in each 10-second window, and the x_{ave} is the average of accelerometer magnitude for 160 samples (10-second epoch×16 Hz). MAD of accelerometer magnitude represents the average magnitude of acceleration within an interval (in this case, 10 seconds) and is proportionate

```
input: Features extracted from 10s Filtered Accelerometer Signal
       (acc_x, acc_y, acc_z, SM)
begin
    Classify Stationary/Walking
    if Walking then
        Classify: Assistive/Non-Assistive Devices
    else
        /* Stationary case                          */
        Determine: Active / Non-Active
        if Active then
            Classify:
            Sitting/Standing/Laying Down
        else
            Classify:
            Sitting/Standing/Laying Down
```

Figure 2.3: Hierarchical Activity Recognition Pseudo Code.

to force applied to the watch by patient since $f = ma$. This value multiplied into displacement will produce relative work and energy. Take into account that calculating displacement from acceleration, however, is not very accurate because it is the result of accelerometer's double integration, that is, any acceleration jitter accumulates and yields big drifts in displacement. Calculating force, however, is accurate and proportionate to energy; hence, the term energy has been used in this study to quantify human activity movements.

Another way of quantifying activity is to integrate each acceleration channel to produce kinetic energy using $e = \frac{1}{2}mv^2$. This way, however, requires more calculations compared with MAD; for the actual speed, each channel should be considered separately so that the direction of acceleration and deceleration that are removed in SM will be taken into account.

It is worth highlighting that by using a smartwatch accelerometer, it is only possible to calculate the force, proportionate to energy, that is spent on the watch. Hence, if a patient is carrying a weight on the watch-worn hand, the energy expenditure of the patient will not change with regard to the watch.

Active/nonactive is determined in this study using an empirical threshold of 0.02 m/s^2 (2 cm/s^2) over the MAD value. As explained earlier, calculating displacement from the accelerometer

is not highly reliable. However, for illustrative purposes, assume the initial speed of hand movement in each window of 10 seconds is zero. Using equation (3) shown in Figure 2.2, the value 0.02 indicates that a patient's hand displacement has been 1 m in 10 seconds. In case of equal or greater shifts, the patient is considered active, otherwise, idle (nonactive).

Figure 2.4 shows 10-second examples of acceleration SM of a person. It illustrates the difference in walking, active and nonactive stationary positions.

Table 2.1: Online watch classifier.

Class	TP[a] rate	FP[b] rate	Precision	Recall	F-measure	ROC[c] area
Stationary	0.992	0.015	0.977	0.992	0.984	0.954
Walking	0.985	0.008	0.995	0.985	0.990	0.992
Weighted average	0.988	0.011	0.988	0.988	0.974	0.929

[a]TP: true positive.

[b]FP: false positive.

[c]ROC: receiver operating characteristic.

Table 2.2: Activity recognition: positioning.

Position	Accuracy	Precision	Recall	F-measure
Stand	91	0.94	0.91	0.92
Sit	93.7	0.87	0.93	0.90
Lay	90.8	0.97	0.90	0.94
Walk	95.1	0.92	0.95	0.94

2.2 Imbalanced learning in healthcare data analytics

Learning from imbalanced datasets can be very challenging as the classes are not equally represented in the datasets [53]. There might not be enough examples for a learner to form a legit hypothesis that can well model the under-represented classes. Hence, the classification results are often biased towards the majority classes. The curse of imbalanced learning is prevalent in real-world

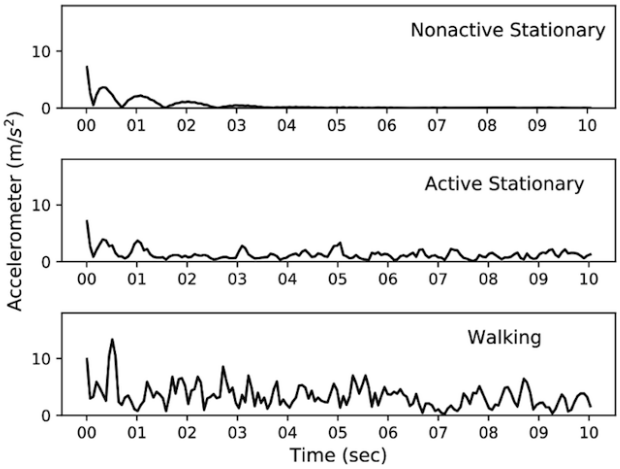


Figure 2.4: Magnitude of accelerometer signal after filtering (direct current component removed before filtering).

applications. In medical research, models are usually trained to give predictions on a dichotomous outcome based on a series of observable features [54]. For example, learning from a cancer dataset which mostly contains non-cancer data samples is perceived to be difficult. Other practical applications with more severely skewed datasets are fraudulent telephone calls [55], detection of oil spills in satellite images [56], detection of network intrusions [57], and information retrieval and filtering tasks [58]. In these scenarios, the imbalance ratio of majority class to minority class can go up to 100,000 [59]. Even though class imbalance issue can exist in multi-class applications, we only focus on the binary class scenario in this paper as it is feasible to reduce a multi-class classification problem into a series of binary classification problems [60].

There have been ongoing efforts in this research domain finding ways to better tackle the imbalanced learning problem. Most of the state-of-the-art research methodologies are fallen into two major categories: 1) Data level approach, or 2) Algorithm level approach [61, 62].

2.2.1 Data level approach

On the data level, skewed datasets can be balanced by either 1) oversampling the minority class data examples, 2) under-sampling the majority class data examples.

2.2.1.1 Oversampling

It aims to overcome the class imbalance by artificially creating new data from the under-represented class. However, simply duplicating the minority class samples would potentially cause overfitting. One of the most widely used techniques is SMOTE. The SMOTE algorithm generates synthetic data examples for minority class by randomly placing the newly created data instances between minority class data points and their neighbors [59]. This technique not only can better model the minority classes by introducing a bias towards the minority instances but also has a lower chance of overfitting. This is due to SMOTE forcing the learners to create larger and less specific decision regions. Based on SMOTE, Hui Han et al. propose the Borderline-SMOTE, which only synthesizes the minorities on the decision borderline [63]. The Borderline-SMOTE classifies minority classes into "safe type" and "dangerous type". The "safe type" is located in the homogeneous regions where the majority of data examples belong to the same class. On the other hand, the "dangerous type" data points are outliers and most likely lie within the decision regions of the opposite class. The intention behind Borderline-SMOTE is to give more weights to the "dangerous type" minority class as it is deemed to be more difficult to learn [64]. Haibo He et al. adopt the same philosophy and proposed ADASYN algorithm, which uses a weighted distribution for different minority class data. The weights are assigned to minority data examples based on the level of difficulty in learning. In other words, harder data examples have more weights thus higher chance of getting more synthesized data. Prior to generating synthetic data, ADASYN inspects the K nearest neighbors for each minority class data example, and counts the number of neighbors from the majority class, Δ_i. Next, the difficulty of learning can be calculated as a ratio of Δ_i/K [65]. ADASYN assigns higher weights on the difficult minority samples. On the contrary, Safe-Level-SMOTE gives more priority to safer minority instances and has a better accuracy performance than SMOTE and Borderline-SMOTE [66]. Karia et al. propose a genetic algorithm, GenSample, for oversampling in imbalanced Datasets. GenSample accounts for the difficulty in learning minority examples when synthesizing, along with the performance improvement achieved by oversampling [67].

2.2.1.2 Undersampling

This technique approaches the imbalanced learning by removing a certain number of data examples from the majority class while keeping the original minority data points untouched. Random undersampling is the most common method in this category [62]. Elhassan AT et al. combine the undersampling algorithm with Tomek Link (T-Link) to create a balanced dataset [54, 68]. However, the undersampling method may suffer severe information loss. In this paper, we mainly focus on the oversampling technique and its variants [69].

2.2.2 Algorithm level approach

On the algorithm level, there are typically three mainstream approaches: a) Improved algorithms, b) cost-sensitive learning, and c) ensemble method [62, 70].

2.2.2.1 Improved algorithms

This approach generally attempts to tailor the classification algorithms to directly learn from the skewed dataset by shifting the decision boundary in favor of the minority class. Tasadduq Imam et al. propose z-SVM to counter the inherent bias in datasets by introducing a weight parameter, z, for minority class to correct the decision boundary during model fitting [71]. Other modified SVM classifiers have also been reported, such as GSVM_RU and BSVM [72, 73]. One special form of an improved algorithm for imbalanced datasets is one-class learning. This method aims to generalize the hypothesis on a training dataset which only contains the target class [74, 75].

2.2.2.2 Cost-sensitive learning

This technique penalizes the misclassifications of different classes with varying costs. Specifically, it assigns more costs to the misclassification of the target class. Hence, the false negative would be penalized more than the false positives [76, 77]. In cost-sensitive learning, a cost weight distribution is predefined in favor of the target classes.

2.2.2.3 Ensemble method

Ensemble method trains a series of weak learners in a fixed number of iterations. A weak learner is a classifier whose accuracy is just barely above chance. At each round, a weak learner is created and a weak hypothesis is generalized. The predictive outcome is produced by aggregating all these weak hypotheses using a weighted voting method [78]. For example, AdaBoost.M2 algorithm calculates the pseudo-loss of each weak hypothesis during boosting. The pseudo-loss is computed over all data examples with respect to the incorrect classifications. The weight distribution is computed using the pseudo-loss (see Algorithm 1). The weight distribution is updated with respect to pseudo loss at the current iteration and will be carried over to the next round of boosting. Hence, the learners in the next iteration will concentrate on the data examples which are hard to learn [79]. Since Adaboost is apt to learn from a imbalanced dataset, several works are based on this boosting framework [80, 81, 82]. SMOTEBoost is proposed to combine the merits of SMOTE and Boosting methods by adding a SMOTE procedure at the beginning of each round of boosting. SMOTEBoost aims to improve the true positives without sacrificing the accuracy of majority class. RUSBoost alleviates class imbalanced by introducing random undersampling technique into a standard boosting procedure. Compared with SMOTEBoost, RUSBoost is a faster and simpler alternative to SMOTEBoost [81]. Ashutosh Kumar et al. proposed RUSTBoost algorithm which adds a redundancy-driven modified Tomek-Link based undersampling procedure before RUSBoost [83]. The Tomek-Link pairs are the pairs of closest data points from different classes. However, all the mentioned boosting algorithms treat the data examples equally. Krystyna Napierala et al. highlighted that the various types of minority data examples (e.g., safe, borderline, rare, and outlier) have unequal influence on the outcome of classification. As such, the algorithms should be designed to focus on the examples which are not easy to learn[64]. DataBoost-IM is reported to discriminate different types of data examples beforehand and adjust the weight distribution accordingly during boosting [82].

2.3 Arrhythmia classification using deep transfer learning with electrocardiogram dataset

Electrocardiogram (ECG) is a simple non-invasive measure to identify heart-related issues such as irregular heartbeats known as arrhythmias. Even though sometimes being observed in healthy people, arrhythmias can develop life-threatening cardiac diseases. Manual inspection on ECG signals to identify arrhythmia can be time consuming and error-prone [84, 85]. Due to the capability of learning complex representation, there have been major developments in utilizing deep learning methods for automatic ECG-based arrhythmia diagnosis [84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96].

In the literature, the ECG analysis generally consists of the following steps: 1) ECG signal preprocessing and noise attenuation, 2) heartbeat segmentation, 3) feature extraction, and 4) learning/classification [85].

The first three steps have been widely studied and discussed in the literature [92]. In this section, we only review a selection of these methods due to page limit. For example, Omid Sayadi et al. proposed a modified extended Kalman filter structure which can be used not only for denoising the ECG signal, but also for compression [92]. C Li et al. presented a heartbeat segmentation algorithm based on wavelet transforms (WT's), which can detect QRS complex (see Figure 2.5) from high P or T waves even with the existence of serious noise or drift [93]. As for feature extraction, Chun-Cheng Lin et al. proposed an automatic heartbeat classification system for arrhythmia classification based on normalized RR intervals (i.e., interval between two successive R waves) and morphological features derived from wavelet transform and linear prediction modeling [97].

Machine learning models are widely used for arrhythmia classification in the literature [85, 86, 88, 90, 91, 96, 97, 98]. Mi Hye Song et al. proposed a support vector machine-based classifier with reduced features derived by linear discriminant analysis [88]. Inspired by the success of Hidden Markov Model (HMM) in modeling speech waveforms for automatic speech recognition, D A Coast et al. applied HMM method in ECG arrhythmia analysis. The model can combine the temporal information and statistical knowledge of the ECG signal in one single parametric model [98]. Awni Y. Hannun et al. proposed an end-to-end deep learning approach which directly takes raw

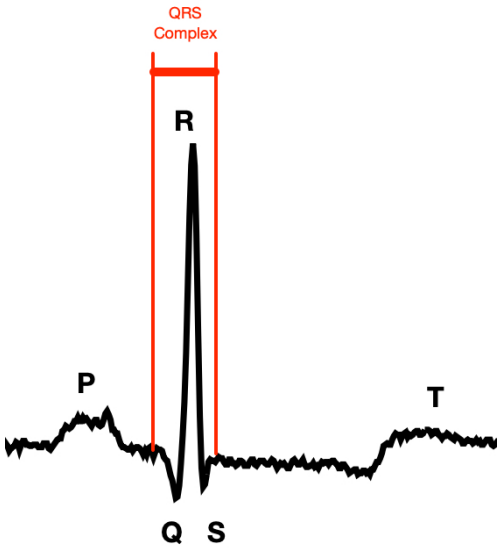


Figure 2.5: A typical heartbeat ECG signal contains P, Q, R, S, and T waves. A QRS complex is a combination of Q,R,S waves.

ECG signal as input and produces classifications without feature engineering or feature selection [91]. Mousavi, Sajad et al. proposed an automatic ECG-based heartbeat classification approach by utilizing a sequence-to-sequence deep learning method to automatically extract temporal and statistical features of the ECG signals [99].

Our work in Chapter 6.1 differs from the studies in 2-fold: 1) it leverages the Short-term Fourier Transform (STFT) to convert 1D ECG signal into 2D time-frequency domain data. Therefore, it is feasible to apply pre-trained 2D Convolution Neural Network in arrhythmia analysis; 2) it is evaluated using MIT-BIH dataset with "inter-patient" training/testing split paradigm detailed in [100].

2.4 Conclusion

In conclusion, this chapter presented a comprehensive literature review of healthcare data analytics projects, with a focus on remote patient monitoring. The chapter started by introducing the Sensing At-Risk Population (SARP) system, a patented remote patient monitoring platform that serves as

the data collection and processing platform for the subsequent data analysis work. The importance of remote patient monitoring in the context of an aging population and the increasing demand for rehabilitation and in-home care was highlighted. The chapter further discussed the challenges of imbalanced learning encountered during the data analysis task and presented a review of the classification of arrhythmias using electrocardiogram signals, which is an integral part of the remote patient monitoring system.

The study emphasized the potential of remote patient monitoring in reducing re-admission costs and improving patient outcomes. It discussed the evolution of remote health monitoring with the advent of wearable devices and the utilization of physical activity trackers to track behavioral patterns that can affect health outcomes.

Overall, this chapter provided a thorough overview of remote patient monitoring, the SARP system, and the challenges and techniques involved in analyzing healthcare data for improved patient care and outcomes. The findings and insights from this chapter lay the foundation for the subsequent chapters that delve into the analysis of data collected from the SARP system and the classification of arrhythmias. These contribute to the advancement of healthcare data analytics and remote patient monitoring.

CHAPTER 3

Background: healthcare analysis with causal inference

3.1 What is causality and why it matters

In this section, we introduce the concept of causality and why causal reasoning is of paramount importance in analyzing data that arise in various domains such as healthcare or social sciences.

3.1.1 What is causality?

The truism of “correlation does not imply causation” is well known and generally acknowledged [101, 102]. The question is how to define "causality".

3.1.1.1 David Hume’s definition

The eighteenth-century philosopher, David Hume, defined causation in the language of the counterfactual: A is a cause of B, if:

1. B is always observed to follow A, and
2. A had not been, B never had existed [103].

In the former case, A is a sufficient causation of B, and A is a necessary causation of B in the latter case [5]. When both conditions are satisfied, we can safely say that A causes B (necessary-and-sufficient causation). For example, sunrise causes rooster’ crow. This cause-effect relation cannot be described the other way around. If the rooster is sick, the sunrise still occurs. Rooster’s crow is not a necessary causation of sunrise. Hence, rooster’s crow is an effect rather the cause of

sunrise.

3.1.1.2 Causality in medical research

In medical research, the logical description of causality is considered too rigorous and occasionally not applicable. For example, smoking does not always lead to lung cancer. Causality in medical literature is often expressed in probabilistic terms [104, 105]. A type of chemotherapy treatment might increase the likelihood of survival of a patient diagnosed with cancer, but does not guarantee it. Therefore, we express our beliefs about the uncertainty about the real world in the language of probability.

One main reason of probabilistic thinking is that we can easily quantify our beliefs in numeric values and build probabilistic models to explain the cause given our observation. In clinical diagnosis, the doctors often seek the most plausible hypothesis (disease) that explain the evidence (symptom). Assume that a doctor observes a certain symptom S, and he or she has two explanations for this symptom, disease A or disease B. If this doctor can quantify his or her belief into conditional probabilities, i.e., $Prob(disease\ A|Symptom\ S)$ and $Prob(disease\ B|Symptom\ S)$ (the likelihood of disease A or B may occur given the symptom S is observed). Then the doctor can choose the explanation that has larger value of conditional probability.

The dilemma: potential outcome framework When we study the causal-effect of a new treatment, we are interested in how the disease responses when we *intervene* upon it. For example, a patient is likely to recover from the cancer when receiving a new type of chemotherapy treatment. To measure the causal effect on this particular patient, we shall compare the outcome of treatment to the outcome of no treatment. However, it is not possible to observe the two *potential outcomes* of the same patient at once. Because this comparison is done using two parallel universes that we imagine: 1) a universe where the patient is treated with the new chemotherapy, and 2) the other where she is not treated. There is always one universe missing. This dilemma is known as the "fundamental problem of causal inference".

3.1.2 Why causal inference?

A data science task can be deemed as making sense of the data or to test a hypothesis about it. The conclusions inferred from data can greatly guide us to make informative decisions. Big data has enabled us to carry out countless prediction tasks in conjunction with machine learning. However, there exist a large gap between highly accurate predictions and decision making. For example, an interesting study [1] reports that there is a "surprisingly powerful correlation" ($\rho = 0.79,\ p < 0.0001$) between the chocolate consumption and the number of Nobel Laureates in a country (Figure 3.1). The policy makers might hesitate to promote chocolate consumption as a way of obtaining more Nobel prizes. The developed western countries where people eat more chocolate are more likely to have better education systems and chocolate consumption has no direct impact on the number of Nobel Laureates. As such, intervening on chocolate cannot possibly lead us to desired outcome. In Section 3.3, we will explore more examples of the spurious correlations explained by confounders (In statistics, confounder is a variable that impacts a dependent variable as well as an independent variable at the same time, causing a spurious correlation [106]) and how to use causal inference to gauge the real causal effect between variables under such circumstances.

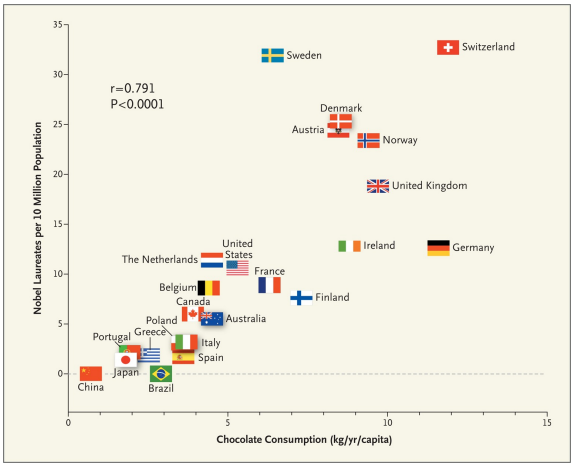


Figure 3.1: A spurious correlation between the chocolate consumption and the number of Nobel Laureates by countries [1].

In predictive tasks, understanding of causality, mechanisms through which an outcome is produced, will yield more descriptive models with better performance. In machine learning, for

instance, one underlying assumption, generally, is that training and testing datasets have identical or at least comparable characteristics/distributions. This assumption is often violated in real practice. For example, an activity recognition model built on a training cohort of highly active participants might perform poorly if it is applied over a cohort of bedridden elderly patients. In this example, variables *age* and *mobility* are the causes that explain the difference between two datasets. Therefore, understanding causality between various features and outcomes is an integral part of a robust and generalized machine learning model. Often, most statistical models rely upon pure correlations perform well under static conditions where the characteristics of the dataset are invariant. Once the context changes, such correlations may no longer exist. On the contrary, relying on the causal relations between variables can produce models less prone to change with context. In Section 3.4, we will discuss on the external validity and transportability of machine learning models.

In many real-world data analytics, in addition to relying solely on statistical relations amongst data elements, it is essential for machine learning practitioners to ask questions surrounding "causal intervention" and "counterfactual reasoning". Questions such as "what would Y be if I do X ?" (causal intervention) or "would the outcome change if I had acted differently?" (counterfactual). Suppose that one wants to find the effects of wine consumption on heart disease. We certainly live in a world in which we cannot run randomized controlled trials asking people to drink wine, say, 1 glass every night, and force them to comply with it for a decade to find the effect of wine on heart disease. In such scenarios we normally resort to observational studies that may eventually highlight associations between wine consumption and reduced risk of heart disease. The simple reaction would be to intervene and to promote wine consumption. However, causal reasoning suggests thinking twice whether wine-reduced heart disease is a causal-effect relation or the association is confounded by other factors, say, people who drink wine are have more money and can buy better quality food, or have better quality of life in general. Counterfactual reasoning, on the other hand, often answers questions in retrospective such as "Would the outcome change if I had acted differently?". Imagine a doctor who is treating a patient with kidney stones. The doctor is left with two choices, conventional treatment that includes open surgical procedures and a new treatment that only involves making a small puncture in the kidney. Each treatment may result

in certain complications, raising the following questions: “Would the outcome be different if the other treatment had been given to this patient.” The “if” statement is a counterfactual condition, a scenario that never happened. The doctor cannot go back in time, give a different treatment to same patient under the same exact condition. So it behooves the doctor to think about counterfactual questions in advance. Counterfactual reasoning enables us to contemplate alternative options in decision-making to possibly avoid undesired outcomes. By understating causality, we will be able to answer questions related to intervention or counterfactuals, concepts we aim to cover in the following sections.

3.1.2.1 Randomized Controlled Trials

Due to the mentioned dilemma, a unit-level of causal effect cannot be directly observed as potential outcomes for an individual subject cannot be observed in a single universe. Randomized controlled trials (RCT) enable us to gauge the population-level causal effect by comparing the outcomes of two groups under different treatments, while other factors are kept identical. Then, the population-level causal effect can be expressed as *average causal effect*(ACE) in mathematical terms. For instance,

$$ACE = |Prob(Recovery|Treatment = Chemotherapy) - Prob(Recovery|Treatment = Placebo)|, \quad (3.1)$$

where ACE is also referred as *average treatment effect* (ATE).

In a randomized controlled trial (RCT), treatment and placebo are assigned randomly to groups that have the same characteristics (e.g., demographic factors). The mechanism is to "disassociate variables of interest (e.g., treatment, outcome) from other factors that would otherwise affect them both"[5].

Another factor that might greatly bias our estimation of causal effect is a century-old problem of "finding confounders" [107, 108, 109]. Randomized controlled trial was firstly introduced by James Lind in 1747 to identify treatment for scurvy, and then popularized by Ronald A. Fisher in

the early 20^{th} century. It is currently well acknowledged and considered as the golden standard to identify the true causal effect without distortions introduced by confounding. However, randomized controlled trials are not always feasible in clinical studies due to ethical or practical concerns. For example, in a smoking-cancer medical study, researchers have to conduct randomized controlled trials to investigate if in fact smoking leads to cancer. Utilizing such trials, researchers should randomly assign participants to an experiment group where people are required to smoke and a control group where smoking is not allowed. This study design will ensure that smoking behavior is the only variable that differs between the groups, and no other variables (i.e., confounders) will bias the results. On the contrary, an observational study where we merely follow and observe the outcomes of smokers and non-smokers will be highly susceptible to confounders and can reach misleading conclusions. Therefore, the better study design would be to choose RCTs, however, it is perceived as highly unethical to ask participants to smoke in a clinical trial. Typically, randomized controlled trials are often designed and performed in a laboratory setting where researchers have full control over the experiment. In many real-world studies, data are collected from observations when researchers cannot intervene/randomize the independent variables. This highlights the need for a different toolkit to perform causal reasoning in such scenarios. In Section 3.3, we will discuss how to gauge the true causal effect from observational studies that might be contaminated with confounders.

3.2 Preliminaries: structural causal models, causal graphs, and intervention with do-calculus

In this section, we primarily introduce 3 fundamental components of causal reasoning: structural causal model (SCM), directed acyclic graphs (DAG), and intervention with do-calculus.

3.2.1 Structural Causal Model

The structural causal model (SCM), $\mathcal{M}$, is proposed by Pearl et al. [6, 110] to formally describe the interactions of variables in a system. A SCM is a 4-tuple $\mathcal{M} =< U, V, F, P(u) >$ where

1. U is a set of background variables, exogenous, that are determined by factors outside the model.

2. $V = \{V_1, ..., V_n\}$ is a set of endogenous variables that are determined by variables within the model.

3. F is a set of functions $\{f_1, ..., f_n\}$ where each f_i is a mapping from $U_i \cup PA_i$ to V_i, where $U_i \subseteq U$ and PA_i (PA is short for "parents") is a set of causes of V_i. In other words, f_i assigns a value to the corresponding $V_i \in V$, $v_i \leftarrow f_i(pa_i, v_i)$, for $i = 1, ..., n$.

4. $P(u)$ is a probability function defined over the domain of U.

Note that there are two sets of variables in a SCM, namely, exogenous variables,U, and endogenous variables, V. Exogenous variables are determined outside of the model and are not explained (or caused) by any variables inside our model. Therefore, we generally assume certain probability distributions $P(u)$ to describe the external factors. The values of endogenous variables, on the other hand, are assigned by both exogenous variables and endogenous variables. These causal mappings are captured by a set of non-parametric functions $F = \{f_1, ..., f_n\}$.f_i can be a linear or non-linear function to interpret all sorts of causal relations. The value assignments of endogenous variables are also referred to as data generation process (DGP) where nature assigns the values of endogenous variables.

Let us consider a toy example: in a smoking-lung cancer study, we can observe and measure the treatment variable *smoking (S)* and the outcome variable *lung cancer (L)*. Suppose these two factors are endogenous variables. There might be some unmeasured factors that interact with the existing endogenous variables, e.g., *genotype (G)*. Then the SCM can be instantiated as,

$$U = \{G\}, \quad V = \{S, L\}, \quad F = \{f_S, f_L\} \tag{3.2}$$

$$f_S : S \leftarrow f_S(G) \tag{3.3}$$

$$f_L : L \leftarrow f_L(S, G) \tag{3.4}$$

This SCM model describes that both *genotype* and *smoking* are direct causes of *lung cancer*. Certain genotype is responsible for nicotine dependence hence explains the smoking behavior [111, 112]. However, no variable in this model explains variable *genotype* and G is an exogenous variable.

3.2.2 Directed Acyclic Graph

Every SCM is associated with a directed acyclic graph (DAG). The vertices in the graph are variables under study and causal mechanisms and processes are edges in DAG. For instance, if variable X is the direct cause of variable Y, then there is a directed edge from X to Y in the graph. The previous SCM model can be visualized as follows:

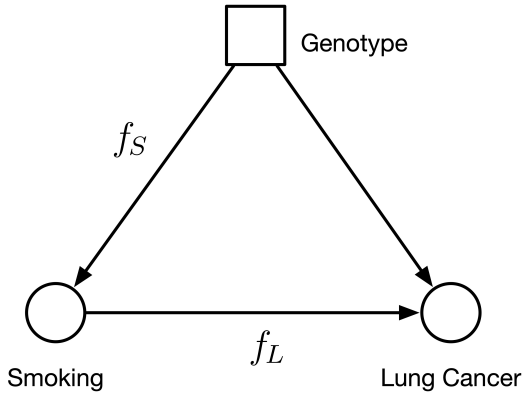


Figure 3.2: Graphical representation of SCM model in Section 3.2.1. Square node denotes the exogenous variable and round nodes denote the endogenous variable. The directed edge represents the causal mechanism.

Note that the graphical representation encodes the causal relations in Equation 3.3-3.4 via a rather intuitive way. In the next section, we will show the strengths of the graphical representation when we need to study the independent relations among variables.

3.2.3 Intervention with do-calculus $do(\cdot)$

Do-calculus was developed by Pearl [113, 114] to gauge the effects of causal interventions [113, 114]. In the example of smoking-lung cancer, the likelihood of getting lung cancer in case of smoking, can be expressed in this conditional probability, $Prob(L = 1|do(S = 1))$, which describes the cause-effect identified in a randomized controlled trial. L and S are Bernoulli random variables which only take two values: 0 and 1. $L = 1$ denotes the fact of getting lung cancer, and $L = 0$ represents the observation of no lung cancer. $S = 1$ means that the observation of smoking whereas $S = 0$ says no smoking is observed. In other words, this post-intervention distribution represents the probability of getting lung cancer $(L = 1)$ when we intervene upon the data generation process by deliberately forcing participant to smoke, i.e., $do(S = 1)$. Post-intervention probability distribution refers to probability terms that contain do-notation, $do(\cdot)$. This post-intervention distribution is different from the conditional probability in an observational study: $Prob(L = 1|S = 1)$, which only represents the likelihood of outcome $(L = 1)$ when we observe that someone smokes. This conditional probability in observational study does not entail the true causal effect as it might be biased by confounders. We will discuss the confounding issue in the next section.

To recall, randomized controlled trials might be impractical or even unethical to conduct at times. For example, we cannot force participants to smoke in a randomized controlled experiment in order to find the cause-effect of an intervention $(do(S = 1))$. Do-calculus suggests us to raise the following question instead: is it possible to estimate the post-intervention $Prob(L|do(S))$ from observational study. If we can express $Prob(L|do(S))$ in terms of the conditional probability $Prob(L|S)$ estimated in the observational study, then we can gauge the causal-effect without performing randomized controlled trials.

3.2.3.1 do-calculus algebra

Here we introduce the algebraic procedure of do calculus that allows us to bridge the gap of probability estimation between observational study and randomized controlled trials. The goal of do-calculus is to reduce the post-intervention distribution that contains the $do(\cdot)$ operator into a set

of probability distributions of $do(\cdot)$ free. The complete mathematical proof of do-calculus can be seen in [115, 6].

Rule 1. *Prob(Y|do(X),Z,W)=Prob(Y|do(X),Z) when we observe the variable W is irrelevant to Y (possibly conditional on the other variable Z), then the probability distribution of Y will not change.*

Rule 2. *Prob(Y|do(X),Z)=Prob(Y|X,Z) if Z is a set of variables blocking all "back-door" paths from X to Y, then* $Prob(Y|do(X), Z)$ *is equivalent to* $Prob(Y|X, Z)$*. Backdoor path will be explained shortly.*

Rule 3. *Prob(Y|do(X))=Prob(Y) we can remove do(X) from Prob(Y|do(X)) in any case where there are no causal paths from X to Y. If it is not feasible to express the post-intervention distribution, Prob(L|do(S)), in terms of do-notation-free conditional probabilities (e.g., Prob(L|S)) using the aforementioned rules, then randomized controlled trials are necessary to gauge the true causality.*

3.2.3.2 Backdoor path and d-separation

In rule 2, the backdoor path refers to any path between cause and effect with an arrow pointing into cause in a directed acyclic graph (or a causal graph). For example, the backdoor path between smoking and lung cancer in Figure 3.2 is "smoking $\leftarrow$ genotype $\rightarrow$ lung cancer". How do we know if a backdoor path is blocked or not?

Judea Pearl in his book [5] introduced the concept of d-separation that tell us how to block the backdoor path. Please refer to [116] for the complete mathematical proof.

a) In a chain junction, A $\rightarrow$ B $\rightarrow$ C, conditioning on B prevents information about A from getting to C or vice versa.

b) In a fork or confounding junction, A $\leftarrow$ B $\rightarrow$ C, conditioning on B prevents information about A from getting to C or vice versa.

c) In a collider, A $\rightarrow$ B $\leftarrow$ C, exactly the opposite rules hold. The path between A and C is blocked when not conditioning on B. If we condition on B, then the path is unblocked. Bear in mind if this path is blocked A and C would be considered independent of each other.

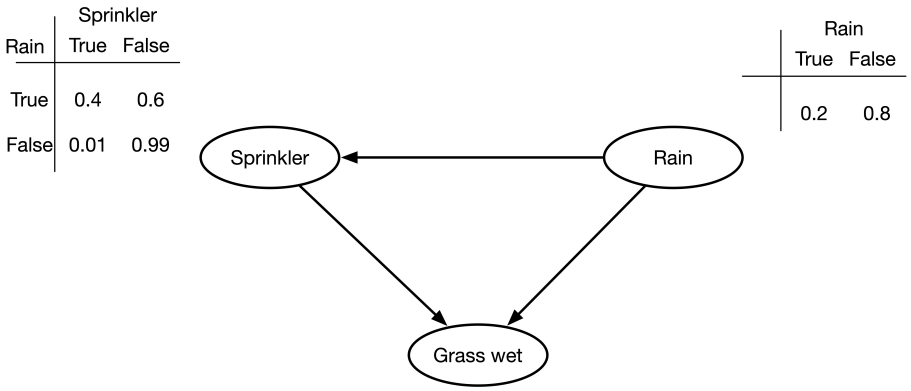
In Figures 3.2, conditioning on genotype will block the backdoor path between smoking and lung cancer. Here, conditioning on genotype means that we only consider a specific genotype in our analysis. Blocking the backdoor between the cause and effect actually prevents the spurious correlation between them in an observational study. Please refer to the next section for more details on confounding bias.

3.2.4 What is the difference between $Prob(Y = y|X = x)$ and $Prob(Y = y|do(X = x))$?

In [6], Pearl et al explains the difference between the two distributions as follows, "In notation, we distinguish between cases where a variable X takes a value x naturally and cases where we fix $X = x$ by denoting the latter $do(X = x)$. So $Prob(Y = y|X = x)$ is the probability that $Y = y$ conditional on finding $X = x$, while $Prob(Y = y|do(X = x))$ is the probability that $Y = y$ when we intervene to make $X = x$. In the distributional terminology, $Prob(Y = y|X = x)$ reflects the population distribution of Y among individuals whose X value is x. On the other hand, $Prob(Y = y|do(X = x))$ represents the population distribution of Y if everyone in the population had their X value fixed at x". This can be better understood with a thought experiment. Imagine that we study the association of barometer readings and weather conditions. We can express this association in terms of $Prob(Barometer|Weather)$ or $Prob(Weather|Barometer)$. Notice that correlations can be defined in both directions. However, causal relations are generally uni-directional. $Prob(Weather = rainy|Barometer = low)$ represents the probability of weather being rainy when seeing the barometer reading is low. $Prob(Weather = rainy|do(Barometer = low))$ describes the likelihood of weather being rainy after we manually set the barometer reading to low. Our common sense tells us that manually setting the barometer low would not affect the weather condition, hence, this post-intervention probability should be zero, whereas $Prob(Weather = rainy|Barometer = low)$ might not be zero.

3.2.5 From Bayesian networks to Structural Causal Models.

Some readers may raise the question: "what is the connection between structural causal models and Bayesian networks, which also aims to interpret causality from the data using DAGs?". Firstly, Bayesian network (also know as belief networks) was introduced by Pearl [117] in 1985 as his early attempt into causal inference. A classic example of Bayesian network is shown in Figure 3.3. The nodes in Bayesian networks represent the variables of interests, and the edges between linked variables denote their dependencies, and the strength of such dependencies are quantified by conditional probabilities. The directed edges in this simple network (Figure 3.3) encodes the following causal assumptions: 1) *Grass wet* is true if the *Sprinkler* is true or *Rain* is true. 2) *Rain* is the direct cause of *Sprinkler* as the latter is usually off in a rainy day to preserve the water usage. We can use this probabilistic model to reason the likelihood of a cause given an effect is observed, e.g., the likelihood of a rainy day if we observe the sprinkler is on is $Prob(Rain = True|Sprinkler = True) = 0.4$ as shown in the conditional probability tables in Figure 3.3.



Rain	Sprinkler: True	Sprinkler: False
True	0.4	0.6
False	0.01	0.99

Rain: True	Rain: False
0.2	0.8

Sprinkler	Rain	Grass wet: True	Grass wet: False
False	False	0.0	1.0
False	True	0.8	0.2
True	False	0.9	0.1
True	True	0.99	0.01

Figure 3.3: A simple example of Bayesian network with conditional probability tables.

However, "*a Bayesian network is literally nothing more than a compact representation of a*

huge probability table. The arrows mean only that the probabilities of child nodes are related to the values of parent nodes by a certain formula (the conditional probability tables)" [5]. On the contrary, the arrows in the structural causal models describe the underlying data generation process between linked variables (i.e., cause and effect) using a function mapping instead of conditional probability tables. If that we construct a SCM on the same example, the DAG remains unchanged, but the interpretations of the edges are different. For example, the edge of "$Rain \rightarrow Sprinkler$" indicates the function $Sprinkler \leftarrow f(Rain)$, which dictates how the effect ($Sprinkler$) would respond if we wiggle the cause ($Rain$). Note that $Sprinkler$ is the effect and $Rain$ is the cause, and the absence of the arrow "$Sprinkler \leftarrow Rain$" in the DAG says there is no such function, $Rain \leftarrow f(Sprinkler)$. Consider we would like to answer an interventional question, "What is likelihood of a rainy if we manually turn on the sprinkler, $Prob(Rain = True|do(Sprinkler = True))$?". It is natural to choose SCMs for such questions: since we know that $Sprinkler$ is not the direct cause of $Rain$ according to the causal graph, the rule 3 of do-calculus algebra (Section 3.2.3.1) permits us to reduce $Prob(Rain = True|do(Sprinkler = True))$ to $Prob(Rain = True)$. That is the status of $Sprinkler$ has no impact on $Rain$. However, a Bayesian network is not equipped to answer such interventional and counterfactual questions. The conditional probability $Prob(Sprinkler = True|Rain = True) = 0.4$ only says the association between $Sprinkler$ and $Rain$ exists. Therefore, the ability to emulate interventions is one of the advantages of SCMs over Bayesian networks [5, 118].

However, Bayesian networks is an integral part of the development of causal inference framework as it is an early attempt to marry causality to graphical models. All the probabilistic properties (e.g., local Marko property) of Bayesian network are also valid in SCMs [5, 116, 118]. Meanwhile, Bayesian networks also impact causal discovery research, which focuses on the identification of causal structures from data through computational algorithms [119].

3.3 Simpson paradox and confounding variables

Spurious correlations introduced by confounder

The famous phrase "correlation does not imply causation" suggests that the observed correlation between variables A and B does not automatically entail causation between A and B. Spurious correlations between two variables may be explained by a confounder. For example, considering the following case (Figure 3.4) where a spurious correlation between yellow fingernails and lung cancer is observed. One cannot simply claim that people who have yellow fingernails have higher risk of lung cancer as neither is the cause of the other. Confounding is a causal concept and cannot be expressed in terms of statistical correlation [120, 121].

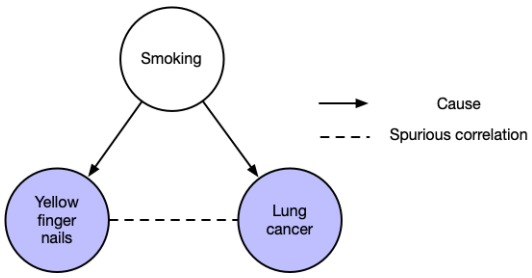


Figure 3.4: Smoking is a common cause and confounder for yellow finger nails and lung cancer. A spurious correlation may be observed between two groups who have yellow fingernails and lung cancer because of the third variable smoking

Another interesting study [1] reported there is a "surprisingly powerful correlation" ($rho = 0.79,\ p < 0.0001$) between the chocolate consumption and the number of Nobel Laureates in a country. It is hard to believe there is any direct causal relation between these two variables in this study. This correlation might be again introduced by a confounder (e.g., advanced educational system in developed countries) that is not included in this study.

Pearl argues that [6]: "*one cannot substantiate causal claims from association alone, even at the population level. Behind every causal conclusion there must lie some causal assumptions that is not testable in an observational study*".

3.3.0.1 Simpson paradox example: kidney stone

Confounder (a causal concept) may not only introduce spurious correlations but can also generate misleading results. Table 3.1 shows a real-life medical study [122] that compares the effectiveness of two treatments for kidney stones. Treatment A includes all open surgical procedures while treatment B is percutaneous nephrolithotomy (which involves making only a small puncture(s) in the kidney). Both treatments are assigned to 2 groups with the same size (i.e., 350 patients). The fraction numbers indicate the number of success cases over the total size of the group.

If we consider the overall effectiveness of two treatments, treatment A (success rate$= \frac{273}{350} = 0.78$) is inferior to treatment B (success rate$= \frac{289}{350} = 0.83$). At this moment, we may think treatment B has higher chance of cure. However, if we compare the treatment by the size of the kidney stones, we discover that treatment A is clearly better in both groups, patients with small stones and patients with large stones. Why is the trend at the population level is reversed when we analyze treatments in sub-populations?

Table 3.1: Kidney stone treatment. The fraction numbers indicate the number of success cases over the total size of the group. Treatment B is more effective than treatment A at overall population level. But the trend is reversed in sub-populations.

	Treatment A	Treatment B
Small Stones ($\frac{357}{700} = 0.51$)	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Large Stones ($\frac{343}{700} = 0.49$)	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
Overall	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$

If we inspect the table with more caution, we realize that treatments are assigned by the severity, i.e., people with large stones are more likely to be treated with method A while most of those with small stones are assigned with method B. Therefore, severity (the size of the stone) is a confounder that affects both the recovery and treatment as shown in Fig.3.5.

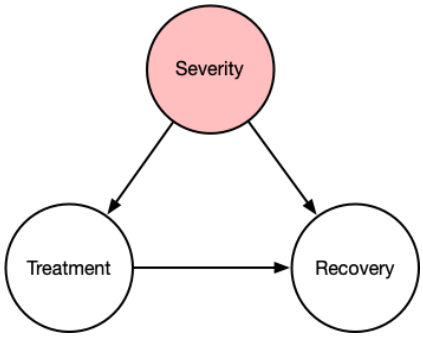


Figure 3.5: Observational study that has a confounder, *severity*

Ideally, we are interested in the pure causal relation of "treatment X $\rightarrow$ recovery" without any other unwanted effects from exogenous variables (e.g., the confounder severity). We de-confound the causal relation of "treatment X $\rightarrow$ recovery" by intervening on variable $Treatment$ and forcing its value to be either A or B. By fixing the treatment, we can remove the effect coming from variable $Severity$ to variable $Treatment$. Note that the causal edge of "$Severity \rightarrow Treatment$" is absent in the mutilated graphical model shown in Figure 3.5. Since Severity does not affects the $Treatment$ and $Recovery$ at the same time after the intervention, it is no longer a confounder. Intuitively, we are interested in understanding if we use treatment A on all patients, what will be the recovery rate, $Prob(Recovery|do(Treatment = A))$. Similarly, what is the recovery rate, $Prob(Recovery|do(Treatment = B))$, if we use treatment B only. If the former has larger value, then treatment A is more effective; otherwise, treatment B has higher chance of cure. The notation $do(X = x)$ is a do-expression which fixes the value of $X = x$. Note that the probability $Prob(Recovery|do(Treatment))$ marginalizes away the effect of severity by $Prob(Recovery|do(Treatment)) = Prob(Recovery|do(Treatment)$,$Severity = treatmentA) + Prob(Recovery|do(Treatment)$,$Severity = treatmentB)$. Essentially, we are computing the causal effects of "treatment A$\rightarrow$ recovery" and "treatment B$\rightarrow$ recovery":

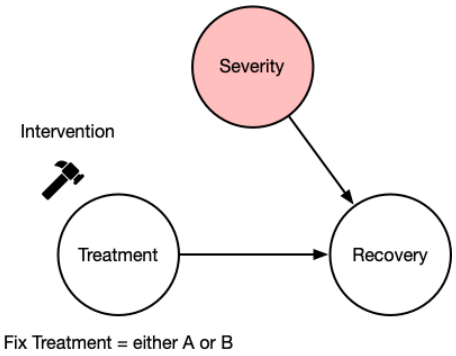


Figure 3.6: We simulate the intervention in the form of a mutilated graphical mode. The causal effect $Prob(Recovery|do(Treatment))$ is equal to the conditional probability $Prob(Recovery|Treatment)$ in this mutilated graphical model.

$$Prob(R = 1|do(T = A)) \tag{3.5}$$

$$= \sum_{s \in \{small, large\}} Prob(R = 1, S = s|do(T = A)) \quad (\textit{ law of total probability}) \tag{3.6}$$

$$= \sum_{s \in \{small, large\}} Prob(R = 1|S = s, do(T = A))Prob(S = s|do(T = A)) \tag{3.7}$$

$$= \sum_{s \in \{small, large\}} Prob(R = 1|S = s, do(T = A))Prob(S = s)^{1} \tag{3.8}$$

$$= \sum_{s \in \{small, large\}} Prob(R = 1|S = s, T = A)Prob(S = s)^{2} \tag{3.9}$$

$$= Prob(R = 1|S = small, T = A)Prob(S = small) + Prob(R = 1|S = large, T = A)Prob(S = large) \tag{3.10}$$

$$= 0.93 \times 0.51 + 0.73 \times 0.49 \tag{3.11}$$

$$= 0.832 \tag{3.12}$$

Similarly, we can compute,

[1] rule #3 in *do-calculus*, see Section 3.2.3
[2] rule #2 in *do-calculus*, see Section 3.2.3

$$Prob(R = 1|do(T = B)) \tag{3.13}$$

$$= \sum_{s \in \{small, large\}} Prob(R = 1, S = s|do(T = B)) \quad (\textit{ law of total probability}) \tag{3.14}$$

$$= \sum_{s \in \{small, large\}} Prob(R = 1|S = s, do(T = B))Prob(S = s|do(T = B)) \tag{3.15}$$

$$= \sum_{s \in \{small, large\}} Prob(R = 1|S = s, do(T = B))Prob(S = s)^{1} \tag{3.16}$$

$$= \sum_{s \in \{small, large\}} Prob(R = 1|S = s, T = B)Prob(S = s)^{2} \tag{3.17}$$

$$= Prob(R = 1|S = small, T = B)Prob(S = small) + Prob(R = 1|S = large, T = B)Prob(S = large) \tag{3.18}$$

$$= 0.87 \times 0.51 + 0.69 \times 0.49 \tag{3.19}$$

$$= 0.782 \tag{3.20}$$

Now we know the causal effects of $Prob(Recovery|do(Treatment = A)) = 0.832$ and $Prob(Recovery|do(Treatment = B)) = 0.782$. Treatment A is clearly more effective than Treatment B. The results also align with our common sense that open surgery (treatment A) is expected to be more effective. A more informative interpretation of the results is that the difference of the two causal effects denotes the fraction of the population that would recover if everyone is assigned with treatment A compared to the other procedure. Recall that we have the opposite conclusion if we read the "effectiveness" at population level in Table 3.1.

3.3.1 How to estimate the causal effect using intervention?

The "interventionist" interpretation of causal effect is often described as the magnitude by which outcome Y is changed given a unit change in treatment T. For example, if we are interested in the effectiveness of a medication in the population, we would set up an experimental study as follows: 1)

[1] rule #3 in *do-calculus*, see Section 3.2.3

[2] rule #2 in *do-calculus*, see Section 3.2.3

We administer the drug uniformly to the entire population, $do(T = 1)$, and compare the recovery rate $Prob(Y = 1|do(T = 1))$ to what we obtain under the opposite context $Prob(Y = 1|do(T = 0))$, where we keep everyone from using the drug in a parallel universe, $do(T = 0)$. Mathematically, we estimate the differenceknown as ACE (defined in section 3.1.2.1),

$$ACE = Prob(Y = 1|do(T = 1)) - Prob(Y = 1|do(T = 0)) \tag{3.21}$$

"*A more informal interpretation of ACE here is that it is simply the difference in the fraction of the population that would recover if everyone took the drug compared to when no one takes the drug*" [21]. The question is how to estimate the intervention distribution with the do operator, $Prob(Y|do(T))$. We can utilize the following theorem to do so,

Theorem 1. ***The causal effect rule*** *Given a graph G in which a set of variables PA are designated as the parents of X, the causal effect of X on Y is given by*

$$Prob(Y = y|do(X = x)) = \sum_z Prob(Y = y|X = x, PA = z)Prob(PA = z) \tag{3.22}$$

If we multiply and divide the right hand side by the probability $Prob(X = x|PA = z)$, we get a more convenient form:

$$Prob(y|do(x)) = \sum_z \frac{Prob(X = x, Y = y, PA = z)}{Prob(X = x, PA = z)} \tag{3.23}$$

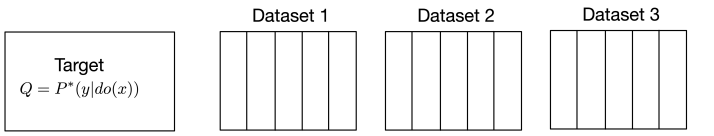
Now the computation of $Prob(Y|do(T))$ is reduced to the estimation of joint probability distributions $Prob(X, Y, PA)$ and $Prob(X, PA)$, which can be directly computed from the corresponding observational dataset. Please refer to [123] for details on probability distribution estimation.

3.4 External validity and transportability of machine learning models.

In the era of big data, we diligently and consistently collect heterogeneous data from various studies. For example, data collected from different experimentational conditions, underlying population,

locations, or even different sampling procedures. In short, the collected data are messy, and rarely serves our inferential goal. Our data analysis should account for these factors. "*The process of translating the results of a study from one setting to another is fundamental to science. In fact, scientific progress would grind to a halt were it not for the ability to generalize results from laboratory experiments to the real world*" [5]. We initially need to take a better look at heterogenous datasets.

3.4.1 How to describe the characteristics of heterogeneous datasets?



d1	**Population**	Los Angeles	New York	Texas
d2	**Obs./Exp.**	Experimental	Observational	Experimental
d3	**Sampling**	Selection on age	Selection on sex	-
d4	**Measured**	X_1, X_2, X_3, Y	X_1, X_2, Y	X_2, Y

Figure 3.7: Heterogeneous datasets can vary on the dimensions $(d1, d2, d3, d4)$ shown above. Suppose we are interested in the causal effect $X \to Y$ in a study carried out in Texas, and we have the same causal effect studied in Los Angeles and New York. This table exemplifies the potential differences between the datasets [2]

.

Big data empowers us to conduct a wide spectrum of studies and to investigate the analytical results. We normally incline to incorporate or transfer such results to a new study. It naturally raises the question: under what circumstances can we transfer the existing knowledge to new studies that are under different conditions. Before we come up with "licenses" of algorithms that permits such transfer, it is crucial to understand how the new dataset in the target study differs from the ones in the existing studies.

Bareinboim [124] summarizes the differences of heterogeneous datasets over the four dimensions

shown in Figure 3.7 [124] that are certainly not enough to enumerate all possibilities in real practice, but more dimensions can be added in future research. In Figure 3.7,

- d1) datasets may vary on the study population;
- d2) datasets may vary in study design. For instance, the study in Los Angeles is an experimental study under laboratory setting, while the study in New York is an observational study in real world;
- d3) datasets may vary in collection process. For instance, dataset 1 may suffer from selection bias on variable age; for example, if the subjects recruited in study 1 are contacted only using landlines, the millennials probably are excluded in the study as they prefer mobile phones;
- d4) Studies might also take measurements on different sets of variables.

3.4.2 Selection bias

Selection bias is caused by preferential exclusion of data samples [125]. It is a major obstacle in validating statistical results, and it can hardly be detected in either experimental or observational studies.

3.4.2.1 COVID-19 example

During the COVID-19 pandemic crisis, a statewide study reported that 21.2% of New York City residents have been infected with COVID-19 [126]. The study tested 3,000 New York residents statewide at grocery and big-box stores for antibodies that are to indicate whether someone has had the virus. Cassie Kozyrkov [127] argues the study might be contaminated with selection bias. The hypothesis notes that the cohort in the study is largely skewed towards the group of people who are high risk-takers/less cautious and have had the virus. The large portion of the overall population may include people who are risk-averse and cautiously stay home; these people were excluded from the research samples. Therefore, the reported toll (i.e., 21.2%) is likely to be inflated. Here, we try to investigate a more generic approach to spot-on selection bias issues.

Causal inference requires us to make certain plausible assumptions when we analyze data. Data sets are not always complete, that is, it does not always tell the whole story. The aresults of the analyses (e.g., spurious correlation) from data alone can be often very misleading. You may recall the example of smokers who may develop lung cancer and have yellow fingernails. If we find this association (lung cancer and yellow fingernails) to be strong, we may come to the false conclusion that one causes the other.

Back to our COVID-19 story, what causal assumptions can we make in the antibody testing study? Consider Figure 3.8 in which each of the following assumptions represent an edge.

i) We know that the antibody will be discovered if we do the related test.

ii) People who have had COVID-19 and survived would generate an antibody for that virus.

iii) Risk taking people are more likely to go outdoor and participate in the testing study.

iv) In order to highlight the difference between the sample cohort and the overall population in the graph, Bareinboim [124, 125] proposed a hypothetical variable S (standing for "selection"). The variable bounded in the square in Fig. 8 stands for the characteristics by which the two populations differ. You can also think of S as the inclusion/exclusion criteria. Variable S have incoming edges from variables "risk taking" and "carried virus". This means that the sample cohort and overall population differ in these two aspects.

Now we have encoded our assumptions into a transparent and straightforward diagram. You may wonder why we go through all the hassle to make this graphical diagram? What value does it add to our identification of selection bias or even debiasing procedure?

Let us begin with the question of how it helps us to find the common pattern/principle of identifying selection bias?

3.4.2.2 Identify selection bias with causal graph

First, a couple of quick tips in identifying selection bias: 1) find any collider variable on the backdoor path between the cause variable and the effect variable, 2) selection bias occurs when

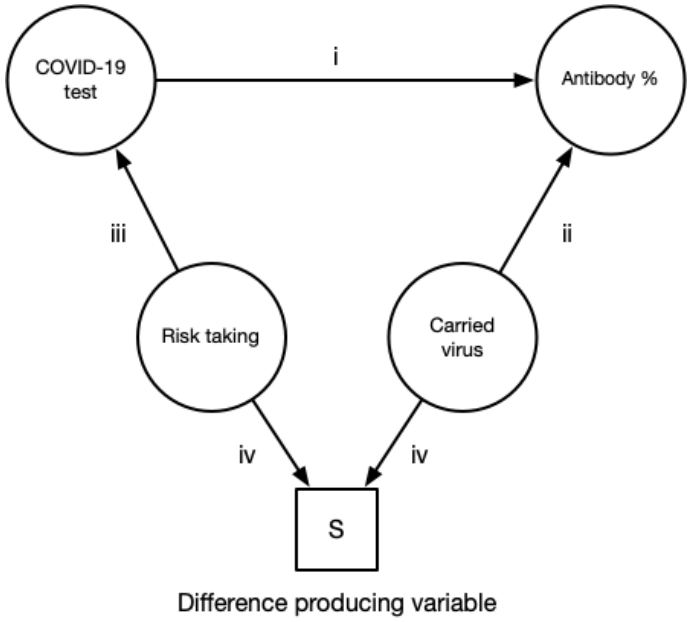


Figure 3.8: Graphical model that illustrates the selection bias scenario. Variable S (squared shape) is a difference producing variable, which is a hypothetical variable that points to the characteristic by which the two populations differ.

your data collection process is conditioning on these collider variables [125]. Note that these tips are only sufficient but not necessary conditions in finding selection bias. By conditioning we mean considering only some of the possibilities a variable can take and not all.

The backdoor path in tip 1) refers to any path between cause variable (COVID testing) and effect variable (antibody %) with an arrow pointing into cause variable ([5] , page 158). In our example, the only backdoor path is "COVID-19 test $\leftarrow$ risk taking $\rightarrow$ S $\leftarrow$ carried virus $\rightarrow$ antibody %". Spurious correlation will be removed if we block every backdoor path.

We observe that there are 3 basic components on the backdoor path: 1) a fork "COVID-19 test $\leftarrow$ high risk $\rightarrow$ S"; 2) a collider "high risk $\rightarrow$ S $\leftarrow$ carried"; 3) another fork "S $\leftarrow$ carried virus $\rightarrow$ antibody %". Now we notice that this backdoor path is blocked if we do not condition on variable S in the collider (rule c). If we condition on variable S (e.g., set S={high risk, have had virus}), the backdoor path will be opened up and spurious correlation is introduced in your antibody testing results.

Now we come to a realization that if we condition on collider on the backdoor path between cause $\rightarrow$ effect and open up the backdoor path, we will encounter the selection bias issue. With this conclusion, we can quickly identify if our study has a selection bias issue given any causal graph. The procedures of this identification can also be automated when graph is complex. So, we offload this judgement to algorithms and computers. Hopefully, you are convinced at this point that using graphical model is a more generic and automated way of identifying selection bias.

3.4.2.3 Unbiased estimation with do-calculus

Now we are interested in the estimation of cause effect between "COVID-19 test $\rightarrow$ Antibody %", $P(Antibody|do(test))$. In other words, the causal effect represents the likelihood of antibody discovery if we test everyone in the population. Our readers may wonder at this point the do-calculus makes sense, but how would we compute and remove the do-operator? Recall the algebric of do-calculus introduced in Section 3.2.3. Let us compute $P(Antibody|do(test))$ as follows,

$$P(Antibody|do(test)) \tag{3.24}$$

$$= P(Antibody|do(test), \{\})^{1} \tag{3.25}$$

$$= P(Antibody|test, \{\})^{2} \tag{3.26}$$

$$= P(Antibody|test) \tag{3.27}$$

$$= \sum_{\substack{i \in \{high, low\} \\ j \in \{true, false\}}} P(Antibody, risk = i, virus = j|test)^{3} \tag{3.28}$$

$$= \sum_{\substack{i \in \{high, low\} \\ j \in \{true, false\}}} P(Antibody|test, risk = i, virus = j) P(risk = i, virus = j|test) \tag{3.29}$$

$$\tag{3.30}$$

[1] Condition on nothing, an empty set.

[2] The backdoor path in Figure 3.8 is blocked naturally if we condition on nothing, {}. According to rule b) of do-alculus, we can safely remove the *do* notation.

[3] Law of total probability.

The last step of the equation show four probability terms measured in the study required to have an unbiased estimation. If we assume close-world (risk={high, low}, virus={true, false}), it means we need to measure every stratified group. Now we have seen that do-calculus can help us identify what pieces we need in order to recover the unbiased estimation.

Model transportability with data fusion.

Transferring the learned knowledge across different studies is crucial to scientific endeavors. Consider a new treatment that has shown effectiveness for a disease in an experimental/laboratory setting. We are interested in extrapolating the effectiveness of this treatment in a real-world setting. Assume that the characteristics of the cohort in the lab setting is different from the overall population in real-world, e.g., age, income, etc. Direct transfer of existing findings into new setting will result in biased inference/estimation of the effectiveness of the drug. Certainly, we can recruit a new cohort that is representative of the overall population and study whether the effectiveness of this treatment is consistent. However, if we could use the laboratory findings to infer our estimation goal in the real-world, this would reduce cost of repetitive data collection and model development. Let us consider a toy example of how causal reasoning could help with data fusion.

3.4.2.4 A toy example of data fusion with causal reasoning

Imagine we developed a new treatment for a disease, and we estimated the effectiveness of the new drug for each age group in a randomized controlled experimental setting. Let $P(Recovery|$ $do(Treatment), Age)$ denote the drug effect at each age group. We wish to generalize the lab results to a different population. Assume that the study cohort in lab setting and the target population are different and the differences are explained by a variable S as shown in Figure 3.8. Meanwhile, we assume these causal effects of each specific age group are invariant across populations. We are interested in gauging the drug effect in the target population ($S = s^*$), and the query can be expressed as $P(Recovery = True|do(Treatment = True), S = s^*)$. Then the query can be solved as follows:

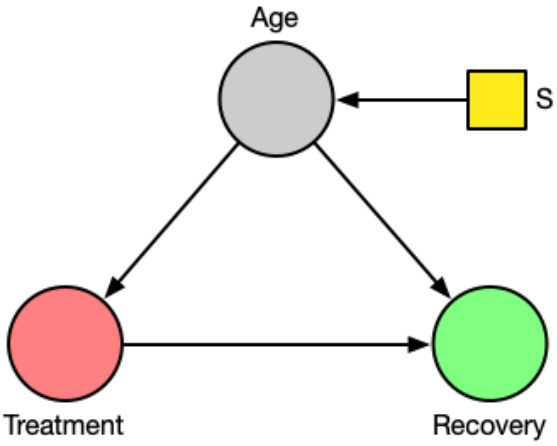


Figure 3.9: An toy example shows how to transfer existing inference in a lab setting to another population where the difference is the age, denoted by a hypothetical difference variable S (in yellow)

$$Query = P(Recovery|do(Treatment), S = s^*) \tag{3.31}$$

$$= \sum_{age} P(Recovery|do(Treatment), S = s^*, Age = age) \times$$

$$P(Age = age|do(Treatment), S = s^*) \tag{3.32}$$

$$= \sum_{age} P^*(Recovery|do(Treatment), Age = age)P^*(Age = age) \tag{3.33}$$

In the last step of the equation, $P^*(Recovery|do(Treatment), Age = age)$ is the effect we discovered through various experimental studies, and it is invariant across study cohorts. Hence, $P^*(Recovery|do(Treatment), Age = age) = P(Recovery|do(Treatment), Age = age)$. $P^*(Age = age)$ is the age distribution in the new population, and is a shorthanded for $P(Age = age|S = s^*)$. We realize that to answer the query, we just need to compute the summation of the experimental findings weighted by the age distribution in the target population.

For example, we assume that we have discovered the effectiveness of the new treatment on different age groups through some experimental studies. The effectiveness is expressed as follows,

$$P(Recovery|do(Treatment), Age < 10) = 0.1 \tag{3.34}$$

$$P(Recovery|do(Treatment), Age = 10 \sim 20) = 0.2 \tag{3.35}$$

$$P(Recovery|do(Treatment), Age = 20 \sim 30) = 0.3 \tag{3.36}$$

$$P(Recovery|do(Treatment), Age = 30 \sim 40) = 0.4 \tag{3.37}$$

$$P(Recovery|do(Treatment), Age = 40 \sim 50) = 0.5 \tag{3.38}$$

The age distribution in our target population is as follows,

- group1 : Age < 10 $P^*(Age < 10) = 1/10$
- group2 : Age = 10 $\sim$ 20 $P^*(10 \leq Age < 20) = 2/10$
- group3 : Age = 20 $\sim$ 30 $P^*(20 \leq Age < 30) = 4/10$
- group4 : Age = 30 $\sim$ 40 $P^*(30 \leq Age < 40) = 2/10$
- group5 : Age = 40 $\sim$ 50 $P^*(40 \leq Age < 50) = 1/10$

According to Equation 3.31-3.33, the effectiveness in the target population should be computed as,

$$P(recovery|do(treatment), S = s^*) = \sum_{age} P(recovery|do(treatment), age)P^*(age) \tag{3.39}$$

$$= 0.1 * \frac{1}{10} + 0.2 * \frac{2}{10} + 0.3 * \frac{4}{10} + 0.4 * \frac{2}{10} + 0.5 * \frac{1}{10} \tag{3.40}$$

$$= 0.03 \tag{3.41}$$

3.5 Learn from missing data using causal inference.

Introduction

Missing data occurs when the collected values are incomplete for certain observed variables. Missingness might be introduced in a study for various reasons: for examples, due to sensors that stop working because the run out of battery; Data collection is done improperly by researchers; Respondents refuse to answer some survey questions that may reveal their private information (e.g., income, disability). Missing data issue is inevitable in some scenarios. In a smoking-cancer medical study, it is perceived highly unethical to ask participants to smoke in order to test the hypothesis of smoking leading to lung cancer.

Typically, building machine learning predictors or statistical models with missing data may expose ourselves to the following risks: a) the partially observed data may bias our inference models, and the study outcomes may largely deviate from the true value [128], b) the reduced sample size may lose the statistical power to provide any informative insights [129], c) this technical impediment might also cause severe predictive performance degradation as most of the machine learning models assume datasets are complete when making inferences.

Extensive research endeavors have been dedicated to this issue. List-wise deletion or mean value substitutions are commonly used in dealing with missing data because of their simplicity. However, these naive methods fail to account for the relationships between the missing data and the observed data. Thus, the interpolations usually deviate from the real values by large. Rubin et al. introduce the concept of missing data mechanism which is widely adopted in the literature [130]. This mechanism classifies missing data into tree categories:

- Missing completely at random (**MCAR**): the observed data are randomly drawn from the complete data. In other words, missingness is unrelated to other variables or itself. For example, in Figure 3.10, job performance ratings is a partially observed variable, and variable IQ is complete without any missingness. The MCAR column shows that the missing ratings are independent of IQ values and itself.

	Job performance ratings			
IQ	Complete	MCAR	MAR	MNAR
78	9	—	—	9
84	13	13	—	13
84	10	—	—	10
85	8	8	—	—
87	7	7	—	—
91	7	7	7	—
92	9	9	9	9
94	9	9	9	9
94	11	11	11	11
96	7	—	7	—
99	7	7	7	—
105	10	10	10	10
105	11	11	11	11
106	15	15	15	15
108	10	10	10	10
112	10	—	10	10
113	12	12	12	12
115	14	14	14	14
118	16	16	16	16
134	12	—	12	12

Figure 3.10: Example source [3]. Job performance ratings is a partially observed variable, and variable IQ is a completely observed variable without any missingness. The second column shows the complete ratings. The $3^{rd}/4^{th}/5^{th}$ columns show the observed ratings under MCAR/MAR/MNAR conditions, respectively.

- Missing at random (**MAR**): the missing values of the partially observed variable dependents on other measured variables. For example, in Fig. 3.10, the MAR column shows that the missing ratings are associated with low IQs.

- Missing not at random (**MNAR**): MNAR includes scenarios when data are neither MCAR nor MNAR. For example, in Fig. 3.10, the MNAR column shows that the missing ratings are associated with itself. i.e., low job performance ratings ($ratings < 9$) are missing.

Pearl et al. demonstrate that theoretical performance guarantee (e.g. convergence and unbiasedness) exists for inference with data that are MCAR and MAR [4, 131]. In other words, we can still have bias-free estimation with even missing data. For example, in Figure 3.10, assume Y is the random variable that represents the job performance ratings. The expectation of job performance ratings under complete, MCAR, MNAR columns are $\mathbb{E}^{Complete}[Y] = 10.35$, $\mathbb{E}^{MCAR}[Y] = 10.60$,

$\mathbb{E}^{MNAR}[Y] = 11.40$, respectively. It can be easily verified that bias of $Bias_{MCAR} = |\mathbb{E}^{Complete}[Y] - \mathbb{E}^{MCAR}[Y]| = 0.25$ is less than $Bias_{MNAR} = |\mathbb{E}^{Complete}[Y] - \mathbb{E}^{MNAR}[Y]| = 1.05$. As the size of the dataset grows, $\mathbb{E}^{MCAR}[Y]$ will converge to the real expectation value $\mathbb{E}^{Complete}[Y]$, i.e., $Bias_{MCAR} = 0$. However, since the MNAR mechanism dictates that low ratings ($Y < 9$) are inherently missing from our observations, we cannot have an bias-free estimation, regardless of the sample size, if we make no assumptions of the missing mechanism. We can notice that the observed data are governed by the missing mechanism (or data generation process). Therefore, missing data issue is inherently a causal inference problem [131, 132]. Details of the causal perspective to missing data can be referred to Section 3.5.

Most statistical techniques proposed in the literature for handling missing data assume that data are MCAR or MAR [133, 134, 135, 136]. For example, expectation maximum likelihood algorithm is generally considered superior to other conventional methods (e.g. list-wise or pair wise deletion) when the data are MCAR or MAR [3]. Moreover, it provides theoretical guarantee (e.g. unbiasedness or convergence) [137] under MCAR or MAR assumptions. However, when it comes to MNAR, estimations with the conventional statistical techniques will mostly be biased. Mohan et al. report that we can achieve unbiased estimation in MNAR scenario under certain constraints using causal methods[4, 131, 132]. These work still leave certain problems unsolved with MNAR case. The limitations will be presented and discussed in details in section 3.5. Therefore, we can focus our future research direction towards solving this unexplored issues.

Causal Perspective

In this section, we briefly explore and discuss the causal approaches proposed by Karthika Mohan and Judea Pearl in dealing with missing data[4, 131, 132]. Firstly, we introduce the concept of causal graph representation for missing data – missing graph(s) (m-graph(s) for short). Then we introduce the definition of recoverability, a formal definition of unbiased estimation with missing data. Next we discuss under what conditions we can achieve recoverability. At last, we identify the unsolved problems with data that are MNAR.

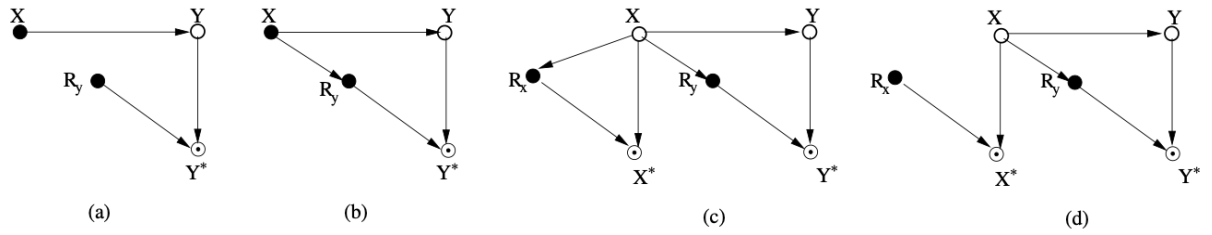


Figure 3.11: m-graphs for data that are: (a) MCAR, (b) MAR, (c) & (d) MNAR; Hollow and solid circles denote partially and fully observed variables respectively [4]

Preliminary on missing graphs

Let $G(\mathbb{V}, E)$ be the causal graph (a DAG) where $\mathbb{V} = V \cup U \cup V^* \cup \mathbb{R}$. V denotes the set of observable nodes, which represents observed variables in our data. These observable nodes can be further grouped into fully observable nodes, V^{obs}, and partially observable nodes, V^{mis}. Hence, $V = V^{obs} \cup V^{mis}$. V^{obs} denotes the set of variables that have complete values whereas V^{mis} denotes the set of variables that are missing at least one data record. Each partially observed variable $v_i \in V^{mis}$ has two auxiliary variables R_{v_i} and V_i^*, where V_i^* is a proxy variable that is actually observed, and R_{v_i} denotes the causal mechanism responsible for missingness of V_i^*,

$$v_i^* = f(r_{v_i}, v_i) = \begin{cases} v_i, & \text{if } r_{v_i} = 0 \\ missing, & \text{if } r_{v_i} = 1 \end{cases} \tag{3.42}$$

In missing graphs, R_{v_i} can be deemed as a switch that dictates the missingness of its proxy variable, V_i^*. For example, in Figure 3.11 (a), X is a fully observable node which has no auxiliary variables. Partially observable node Y is associated with Y^* and R_{Y_i}. Y^* is the proxy variable that we actually observe on Y, and R_y masks the values of Y by its underlying missingness mechanism (e.g., MCAR, MAR, MNAR). E is the set of edges in m-graphs, and U is the set of unobserved variables (latent variables). For example, in a recommender system, the purchase interest of a user can not be measured nor observed. Hence it is an unobserved variable. The toy example in Figure 3.11 is not involved with any latent variable.

We can cast the classification of missingness mechanisms (e.g., MCAR, MAR, MNAR) onto

m-graphs as depicted in Figure 3.11.

- Figure 3.11 a) shows the MCAR case where $R_y \perp\!\!\!\perp X$ [1]. The on-off status of R_y is solely determined by coin-toss. Bear in mind that the absence of an edge between two vertices in causal graph is a strong constraint which represents there is no relation between them. The criterion of deciding if a m-graph represents MCAR is $\mathbb{R} \perp\!\!\!\perp (V^{obs} \cup V^{mis} \cup U)$.
- Figure 3.11 b) shows the MAR case where $R_y \perp\!\!\!\perp Y|X$ [1]. R_y depends on the fully observed variable X. The criterion of deciding if a m-graph represents MAR is $\mathbb{R} \perp\!\!\!\perp (V^{mis} \cup U)|V^{obs}$.
- Figure 3.11 c)&d) show the MNAR cases where neither of the aforemetioned criterions holds.

It is a clear advantage that we can directly read the missingness mechanism from the m-graphs using *d-seperation* [2] without conducting any statistical test.

Recoverability

Before we can discuss under what conditions we can achieve bias-free estimation, we shall firstly introduce the definition of recoverability.

Definition 1. ***Recoverability*** *[4]. Given a m-graph G, and a target query relation Q defined on the variables in V, Q is said to be recoverable in G if there exists an algorithm that produces a consistent estimate of Q for every dataset D such that P(D) is 1) compatible with G and 2) strictly positive over complete cases, i.e.,* $P(V^{obs}, V^{mis}, \mathbb{R} = 0) > 0$.

The definition in the original paper [4] may be a bit obscure at first. To my understanding, a query (e.g. what is the value of joint probability, $Prob(X, Y)$ in Figure 3.11) is recoverable if there exists an algorithm that computes $Prob(X, Y)$ with the observed values of X, Y^*. Then this algorithm is referred as a "consistent estimator" which gives unbiased inference. We will see some examples in this section later.

[1] This (conditional) independence is directly read off from causal graphs using *d-separation* technique

[2] Watch this video on d-separation: `https://www.youtube.com/watch?v=yDs_q6jKHb0`

Corollary 1. *[4]. A query relation Q is recoverable in G if and only if Q can be expressed in terms of the probability P(O) where $O = R, V^*, V^{obs}$ is the set of observable variables in G. In other words, for any two models M_1 and M_2 inducing distribution P^{M_1} and P^{M_2} respectively, if $P^{M_1}(O) = P^{M_2}(O) > 0$ then $Q^{M_1} = Q^{M_2}$.*

Recoverability when data are MCAR

Example 3.5.1. *[4]. Let X be the treatment and Y be the outcome as depicted in the m-graph in Figure 3.11 a). Let it be the case that we accidentally delete the values of Y for a handful of samples, hence $Y \in V_m$. Can we recover P(X,Y)?*

Yes, P(X,Y) under MCAR is recoverable. We know that $R_y \perp\!\!\!\perp (X, Y)$ [1] holds in Figure 3.11 a). Thus, $P(X, Y) = P(X, Y|R_y) = P(X, Y|R_y = 0)$. When $R_y = 0$, we can safely replace Y with Y^* as $P(X, Y) = P(X, Y^*|R_y = 0)$. Note that $P(X, Y)$ has been expressed in terms of the probability ($P(X, Y^*|R_y = 0)$) we can compute using observational data. Hence, we can recover P(X,Y) with no bias.

Recoverability when data are MAR

Example 3.5.2. *[4]. Let X be the treatment and Y be the outcome as depicted in the m-graph in Figure 3.11 b). Let it be the case that some patients who underwent treatment are not likely to report the outcome, hence $X \in R_y$. Can we recover P(X,Y)?*

Yes, P(X,Y) under MAR is recoverable. We know that $R_y \perp\!\!\!\perp Y|X$ [1] holds in Figure 3.11 b). Thus, $P(X, Y) = P(Y|X)P(X) = P(Y|X, R_y)P(X) = P(Y|X, R_y = 0)P(X)$. When $R_y = 0$, we can safely replace Y with Y^* as $P(X, Y) = P(Y*|X, R_y = 0)P(X)$. Note that $P(X, Y)$ has been expressed in terms of the probability ($P(Y*|X, R_y = 0)P(X)$) we can compute using observational data. Hence, we can recover P(X,Y) with no bias.

Recoverability when data are MNAR

Example 3.5.3. *[4]. Figure 3.11 d). depicts a study where (i) some units who underwent treatment (X=1) did not report the outcome Y, and (ii) we accidentally deleted the values of treatment for a handful of cases. Thus we have missing values for both X and Y which renders the dataset MNAR. Can we recover P(X,Y)?*

Yes, P(X,Y) in d) is recoverable. We know that $X \perp\!\!\!\perp R_x$ *and* $(R_y \cup R_y) \perp\!\!\!\perp Y|X$ [1] *holds in Figure 3.11 d). Thus,* $P(X,Y) = P(Y|X)P(X) = P(Y|X,R_y)P(X) = P(Y^*|X^*, R_y = 0, R_x = 0)P(X^*|R_x = 0)$. *Note that* $P(X,Y)$ *has been expressed in terms of the probability (*$P(Y^*|X^*, R_y = 0, R_x = 0)P(X^*|R_x = 0)$*) we can compute using observational data. Hence, we can recover P(X,Y) with no bias.*

In the original paper [4], $P(X,Y)$ is not recoverable in Figure 3.11. c). Mohan et al. provide a theorem (see Theorem 1 in [4]) which states the sufficient condition for recoverability.

Testability

In Figure 3.11 b), we assume that the missing mechanism R_y is the causal effect of X, hence the arrow pointing from X to R_y. The question naturally arises: "is our assumption/model compatible with our data?" Mohan et al. propose an approach to testify the plausibility of missing graphs from the observed dataset [138].

Testability of Conditional Independence (CI) in m-graphs

Definition 2. *[138] Let* $X \cup Y \cup Z \subseteq V_o \cup V_m \cup R$ *and* $X \cap Y \cap Z \neq \emptyset$. $X \perp\!\!\!\perp Y|Z$ *is testable if there exists a dataset* D *governed by a distribution* $P(V_o, V^*, R)$ *such that* $X \perp\!\!\!\perp Y|Z$ *is refuted in all underlying distributions* $P(V_o, V_m, R)$ *compatible with the distribution* $P(V_o, V^*, R)$.

In other words, if the CIs can be expressed in terms of observable variables exclusively, then these CIs are deemed testable.

Theorem 2. *Let $X, Y, Z \subset V_o \cup V_m \cup \mathbb{R}$ and $X \cap Y \cap Z = \emptyset$. The conditional independence statement S: $X \perp\!\!\!\perp Y | Z$ is directly testable if all the following conditions hold:*

1. $Y \nsubseteq (R_{X_m} \cup R_{Z_m})$. *In words, Y should contain at least one element that is not in $R_{X_m} \cup R_{Z_m}$*
2. $R_{X_m} \subseteq X \cup Y \cup Z$. *In words, the missingness mechanisms of all partially observed variables in X are contained in $X \cup Y \cup Z$*
3. $R_{Z_m} \cup R_{Y_m} \subseteq Z \cup Y$. *In words, the missingness mechanisms of all partially observed variables in Y and Z are contained in $Y \cup Z$*

Testability of CIs comprising of only substantive variables

As for the CIs that only includes substantive variables (e.g., Figure 3.11 (b)), it is fairly easy to see $X \perp\!\!\!\perp Y$ is testable when $X, Y \in V_o$.

Missing data from causal perspective

Given an incomplete dataset, our first step is to postulate a model based on causal assumptions of the underlying data generation process. Secondly, we need to determine whether the data rejects the postulated model by identifiable testable implications of that model. Last step is to determine from the postulated model if any method exists that produces consistent estimates of the queries of interests.

3.6 Augmented machine learning with causal inference in recommender systems.

Despite the rising popularity of causal inference research, the route from machine learning to artificial general intelligence still remains uncharted. Strong artificial intelligence aims to generate artificial agents with the same level of intelligence as human beings. The capability of thinking causally is integral for achieving this goal [139]. In Chapter 7, I make my initial endeavor to enhance

deep learning models in recommenders systems by incorporating causal inference techniques. Numerous research inquiries in recommender systems center around assessing the impact of specific interventions or recommendations, essentially involving cause-and-effect relationships. Hence, recommender systems serve as an ideal foundation for exploring causal inference in intervention research.

3.6.1 Selection bias in recommender systems

Selection bias is a widely-recognized issue in recommender systems [140, 141, 142]. For example, music stream services usually suggest genres that have positive user feedbacks (e.g., favorite, share, and buy, etc.), and selectively ignore the ones that are rarely exposed to users [143]. In this section, we study the selection bias that exists in the post-click conversion rate (CVR) estimation.

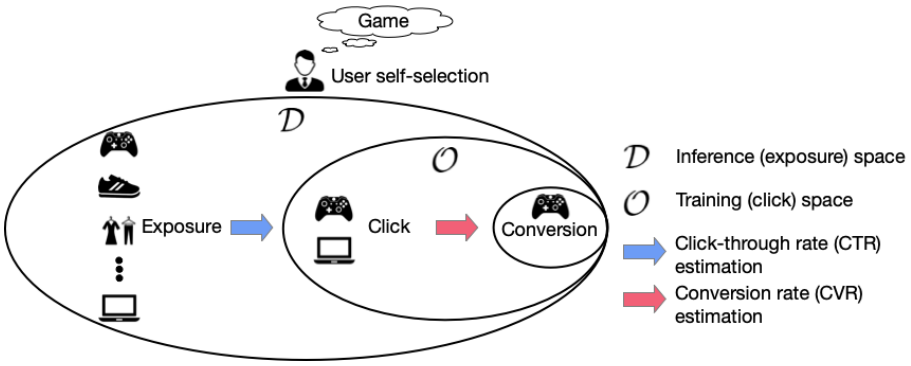


Figure 3.12: Illustration of the selection bias issue in conventional conversion rate (CVR) estimation. The training space of conventional CVR models is the click space $\mathcal{O}$, whereas the inference space is the entire exposure space $\mathcal{D}$. The discrepancy of data distribution between $\mathcal{O}$ and $\mathcal{D}$ leads to selection bias in conventional CVR models.

Problem formulation

Post-click conversion rate (CVR) estimation is a critical task in e-commerce recommender systems [144, 145]. A typical e-commerce transaction has the following sequential events: "exposure $\rightarrow$ click $\rightarrow$ conversion" [141]. Post-click conversion rate indicates the probability of transitions from

click to conversion. Typically, when training CVR models, we only include the items that customers clicked on as we are unaware of the conversion feedback of the items that are not clicked by customers [146]. Bear in mind, not clicking on an item does not necessarily indicate the customer is not interested in purchasing it. Customers may unconsciously skip certain items that might be interesting to them. Figure 3.12 reveals that the exposure space $\mathcal{D}$ is a super set of the click space $\mathcal{O}$. Selection bias occurs when conventional CVR models are trained in the click space, and the predictions are made in the entire exposure space (see Figure 3.12) [141]. Intuitively, data in the click space is drawn from the entire exposure space and is biased by the user self-selection. Therefore, the data distribution in the click space is systematically different from the one in the exposure space. This inherent discrepancy leads to data that is missing not at random (MNAR), and selection bias in the conventional CVR models [142, 137, 147, 148].

We identify two practical issues that make CVR estimation quite challenging in industrial-level recommender systems:

- **Selection bias**: The systematic difference of data distributions between training space $\mathcal{O}$ (i.e., all user self-selected items) and inference space $\mathcal{D}$ (i.e., all exposed items) biases conventional CVR models [149, 146, 150]. This bias usually causes the degradation of model performance.
- **Data sparsity**: In the CVR estimation task, it refers to the fact that item clicks are rare events (we have a CTR of 5.2% in the production dataset and 4% in the public dataset). Conventional CVR models are typically trained only using data in the click space. Therefore, the number of training samples may not be sufficient for the large parameter space. In our experiments, the numbers are 0.6 billion samples vs. 5.3 billion parameters in production dataset, and 4.3 million samples vs. 2.6 billion parameters in public dataset (see Chapter 7) [151, 152].

3.6.2 A causal perspective to unbiased CVR estimation

Recall that selection bias in CVR estimation comes from the fact that models are trained over the click space $\mathcal{O}$, whereas the predictions are made over the exposure space $\mathcal{D}$ (See Figure 3.12). Ideally, we can remove the selection bias by building our CVR estimators using a dataset where the

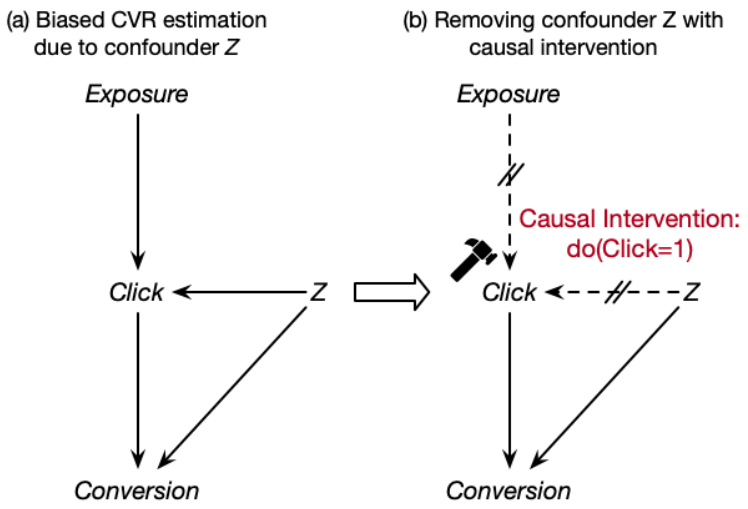


Figure 3.13: This causal graph formulate CVR estimation as a causal problem. [5] In (a), Z is a confounder that affects both clicks and purchases, and it biases the inference. In (b), we apply intervention on click events ($do(Click) = 1$). Once users are "forced" to click on each exposed item, Z has no control over user click behaviors. Note the absence of the arrow from Z to *Click*. Hence, we have successfully removed the confounder Z, and the selection bias [6, 5, 7, 8, 9].

conversion labels of all the items are known. In the language of causal inference, it is equivalent to training CVR estimators on a "do dataset", where causal intervention is applied on click event during the data generation process. Specifically, users are "forced" to click on every item in the exposure space $\mathcal{D}$ and further make their purchase decisions. Note that the training space is the same as the inference space in the "do dataset". Hence, the selection bias is eliminated. Intuitively, we can also understand how causal intervention removes the bias in Figure 3.13. Z denotes the self-selection factors that affect both click events and conversion events. For example, Z can be the purchase interest or price discount that customers consider in online-shopping. In causal inference, we refer Z as "confounder(s)" that biases the CVR inference [7]. Once the causal intervention is applied on the click event (i.e., users are forced to click on all exposed items), Z has no control over user click behaviors. It means that we have successfully removed the confounder Z which biases our CVR estimation [6, 5, 7, 8, 9].

Apparently, this "do dataset" generated in this imaginary intervention experiment cannot be

obtained in reality. Now the challenge is how to train our CVR estimators on the observed dataset $\mathcal{O}$ as if we do on the "do dataset". In Chapter 7, we will discuss two estimators that can achieve unbiased CVR prediction with the data that are MNAR.

3.6.3 Related works

In this section, we review several existing works that attempt to tackle the selection bias issue in recommender systems. Meanwhile, we summarize how our methods are different from prior works.

Ma *et al.* [141] proposed the Entire Space Multi-task Model (ESMM) to remedy selection bias and data sparsity issues in the conversion rate (CVR) estimation. ESMM is trained in the entire exposure space, and it formulates CVR task as two auxiliary tasks, i.e., click-through rate (CTR) and click-through & conversion rate (CTCVR) estimations. However, we argue that ESMM is biased. The details of our argument are presented in Section 7.2.2 in Chapter 7.

Causal inference offers a way to adapt for the data generation process when we attempt to restore the information from MNAR data [147]. Schnabel *et al*. [140] proposed an IPW-based estimator for training and evaluating recommender systems from biased data. IPW-based models may still be biased if the propensities are not accurately estimated. Wang *et al.* [153] proposed a doubly robust (DR) joint learning approach for estimating item ratings that are MNAR. Doubly robust estimator combines the IPW-based methods with an imputation model that estimates the prediction error for the missing data. When the propensities are not accurately learned, DR estimator can still enjoy unbiasedness as long as its imputation model is accurate. However, the existing DR-based methods are not devised for CVR estimation, hence fail to account for the severe data sparsity issue that widely exists in the CVR estimation. In addition, such a joint learning approach is not efficient in industrial setting (see Figure 7.3).

3.7 Causal discovery methods in high dimensional space

3.7.1 Dimension reduction by feature selection

Hao et al. proposed 3-phase causal discovery algorithm framework, Causal Discovery on High Dimension (CDHD), to solve the high dimensional problem by using conventional feature selection algorithms to reduce the size of the feature space before identifying the causal relations [154]. In the first phase, a heuristic-based feature selection algorithm, Max-Relevance and Min-Redundancy (MRMR), is employed to select most relevant features with respect to the target variable, resulting in search space reduction [155]. In the second phase, a constraint-based causal discovery method is utilized to discover the causal skeleton. In the third phase, a causal direction learning algorithm, Information Geometric Causal Inference (IGCI), is incorporated to orient the edge directions in the learned causal graphs [156].

3.7.2 Fast PC algorithm for high dimensional causal discovery

It is well acknowledged that the PC algorithm does not scale well with high dimensional dataset as its runtime is exponential to the number of variables [157, 158, 159, 160, 161]. Meanwhile, the inferred graph from the PC algorithm is variable order-dependent; that is, the resulting graph will change if the order of input features changes [162]. Recall that the PC algorithm generate causal structure skeleton by removing edges if the connected pair of nodes are conditionally independent. In practice, however, we do not have the perfect knowledge of these conditional independence relations. The conditional independence is obtained by statistical conditional independence tests at some predefined significance level, α [162]. For example, the standard Pearson chi-square test is generally performed under i.i.d. assumption (i.e., independent, identically distributed) [163]. Since the causal structure is updated dynamically after each edge removal, the resulting skeleton might be different from the true structure if these statistical tests return false independence relations.

The fast PC algorithm, based on [162], is a parallelized approach that groups conditional independence tests at each level, ensuring they are not correlated. These subtasks are then distributed

across different CPU cores for simultaneous execution. The results obtained from each core are later integrated to form the final outcome. To enable easy combination of subgraphs, this algorithm requires that the subtasks be independent. The parallel-PC algorithm has been recognized as a fast and memory-efficient procedure for learning causal structures. It has been evaluated on both synthesized datasets and real-world datasets [164]. Furthermore, the proposed algorithm is also order-independent, allowing for flexibility in the arrangement of variables. Readers can find the implementation of the algorithm in the R package pcalg, available at `https://cran.r-project.org/web/packages/pcalg/index.html`.

3.7.3 Association rule mining for causal discovery

[160] propose an association rule mining based algorithm for causal discovery. The associate rule mining generally utilizes machine learning models to identify "if-then" association patterns in the data. Meanwhile, association rule mining has already been demonstrate being efficient method for relation discovery. Note that the associations are generally in the scope of correlation interpretation and should be carefully investigated before treated as casual relations [165]. An associate rule generally consists of an antecedent ("if") and a consequent ("then"), both of which are a list of items.

There are generally three metrics in assessing the strength of the association: 1) support, 2) confidence, and 3) lift. The first two metrics are described with more details in [165]. We present the concept of lift to introduce the underlying concept of using association rule mining in causal discovery.

Lift is a metric that calculates the ratio of confidence to support. Specifically, it is defined as,

$$Lift(X \rightarrow Y) = \frac{(\text{Transactions containing both } X \text{ and } Y)/(\text{Transactions containing } X)}{\text{Fraction of transactions containing } Y}$$

The interpretation of $lift(X \rightarrow Y)$ literally means the lift or raise in confidence of observing Y if we have seen X over the probability of observing Y without any knowledge of X. This interpretation has the similar philosophy of average treatment effect (ATE) in a randomized control

trials in gauging the causal effect [160, 166]. $lift(X \rightarrow Y)$ can be seen as the causal effect under the assumption that no confounder exists that impact X and Y at the same time.

The proposed algorithm in [160] identifies the causal relations as follows: firstly, it identifies all the irrelevant variables with respect to the target outcome. This is done by calculating the lift (odds ratio) of an independent variable X to the target outcome Y. The proposed algorithm assumes that an input variable is irrelevant if the odds ratio is significantly lower than 1. Next, all the irrelevant variables are excluded from the candidates feature set for the causal discovery. To check if the remaining features are indeed casual features, the proposed algorithm calculates the confidence interval of the lift through sampling a fraction dataset that contains the antecedent and consequent. If the lower bound of the lift is significantly larger than 1. Then the algorithm assumes that the associate rule is a causal relation. Meanwhile, the proposed algorithm also leverage the anti-monotone property of association rule mining to accelerate the computation. In particular, a rule is redundant if it is implied by a more generalized rule. For example, if rule "college graduate $\rightarrow$ high salary" holds, then we know that both male college graduates and female college graduates enjoy high salaries. It is therefore redundant to have the rules "male college graduate $\rightarrow$ high salary" and "female college graduate $\rightarrow$ high salary". Therefore, this anti-monotone property can prune the search space in the high dimension settings.

3.7.4 Formulating the causal discovery problem as a continuous constrained optimization problem

Zheng et al. introduced an innovative framework that addresses the causal structure learning problem by formulating it as a continuous optimization problem, which can be efficiently solved using gradient-based solvers [167]. Their key contribution lies in the utilization of a continuous equality constraint that ensures acyclicity in the causal structure. The advantage of this continuous constraint is its compatibility with existing toolkits for constraint optimization problems, enabling efficient computation of the optimal solution.

However, the approach presented in [167] relies on the assumption of a linear Structural Equation

Model (SEM), which may be overly restrictive for many real-world datasets. To overcome this limitation, Yue Yu et al. proposed DAG-GNN, an architecture based on graph neural networks that captures more complex data characteristics [168]. DAG-GNN is capable of handling both discrete input variables and vector-valued variables.

Furthermore, it is worth mentioning that the equality constraint proposed in [167] still incurs computational expenses. In Chapter 8.1 of our work, we aim to enhance this constraint and propose improvements upon it.

3.8 Conclusions

In this chapter, we explored the strengths of causal reasoning when facing problems such as confounding bias, model transportability, and learning from missing data. We present some examples to demonstrate pure data-driven or correlation-based statistical analysis may generate misleading conclusions. We argued the need to consider causality in our models to support critical clinical decision-making. Machine learning has been widely employed in various healthcare applications with recently increased efforts on how to augment machine learning models with causality to improve interpretability [169, 170] and predictive fairness [24, 171] and to avoid bias [172, 173]. The model interpretability can be enhanced through the identification of cause-effect relation between the model input and outcome. We can observe how the model outcome responds to interventions upon inputs. For example, powerful machine learning models can be built for early detection of type 2 diabetes mellitus using a collection of features such as age, weight, HDL cholesterol, and triglycerides [174]. However, healthcare practitioners are not content with mere predictions – they are also interested in the variables upon which the intervention will help reduce the risk of the disease effectively. Understanding causality is crucial to answer such questions. We also showed how causality can address confounding bias and selection bias in data analyses. Literature shows that causal inference can be adopted in deep learning modeling to reduce selection bias in recommender systems [24, 171]. Model fairness aims to protect the benefit of people in the minority groups or historically disadvantageous groups from the discriminative decisions produced

by AI. Causal inference can also ensure model fairness against such social discriminations [172]. In addition to the attempts and progressed made in this field, there are many low-hanging fruits in combining causal inference with machine learning methods. We hope this brief introduction of causal inference can inspire more interested readers in this research area.

CHAPTER 4

Healthcare data analytics in remote patient monitoring

4.1 Introduction

In recent years, the field of healthcare has made significant progress in integrating wireless technology into traditional care models. The widespread availability of devices like wearable sensors has enabled researchers to gather large amounts of data and apply it to various aspects of healthcare. One important objective of using wearable sensors is to study and analyze human activity and functional patterns in order to predict harmful outcomes such as falls. These sensors can also track individual movements to identify personalized behavioral patterns and establish standardized measures for frailty, well-being, and independence. Many wearable devices, such as activity trackers and smartwatches, come equipped with affordable embedded sensors that provide users with health statistics. Additionally, Bluetooth low-energy sensors called BLE beacons have gained popularity among researchers in the field of ambient intelligence. These beacons, known for their low cost and durability, are useful for collecting indoor localization data, which is an important component of recognizing human activity. In studies conducted by Moatamed et al. and in a patent application by Ramezani et al., a comprehensive framework called Sensing At-Risk Population was introduced. This framework combines the classification of human movements using a 3-axial accelerometer with the extraction of indoor localization using BLE beacons.

The purpose of this chapter is threefold:

- To assess the effectiveness of combining physical activity and indoor location features, extracted at baseline, in distinguishing between subacute care patients who are readmitted to the hospital and those who are able to remain in a community setting. This assessment was

conducted on a cohort of 154 patients residing in a rehabilitation facility.

- To examine the longitudinal changes in sensor-based physical activity and indoor localization features of patients receiving rehabilitation at a skilled nursing facility. By tracking these changes over time, we can gain insights into the progress and development of these patients.
- To investigate whether the changes detected by sensors over time can complement the assessments made by therapists during the course of rehabilitation at the skilled nursing facility. By comparing the two sources of information, we can better understand the overall progress and recovery of the patients.

4.2 Methods

4.2.1 Overview

From June 2016 to November 2017, we recruited patients after admission to a subacute rehabilitation center in Los Angeles. We performed a cross-sectional baseline study of this cohort to better understand data features collected by the SARP system. We investigated the prevalence of physical activity tracking features and indoor localization features at baseline for both outcome groups (hospital vs long-term care). Moreover, we assessed their efficacy in determining the outcome (hospital vs long-term care).

4.2.2 Participants

Participants aged older than 60 years were recruited from a subacute rehabilitation facility in Los Angeles. The study cohort contains patients who had been admitted to a subacute rehabilitation center for 21 days. After this period, patients were either re-admitted to hospital (H) or stayed in community (C; either at home or long-term care). The inclusion criteria were broad, allowing any patient to participate as long as they were aged older than 60 years, English speaking, and able to consent with the exclusion criteria including movement disorders or paralysis of the upper or

lower extremity. The diversity of cohort included patients who were a postsurgical, poststroke, and postclinical decompensation because of medical illnesses. Eligible participants signed a consent form approved by the University of California, Los Angeles, Institutional Review Board.

4.2.3 Study Design

Patients were given a smartwatch by a clinical coordinator every morning at 9 am. Patients were asked to wear their watches at all times until the coordinator collected the watch at around 6 pm every day. Watch batteries were expected to last longer than the protocol period (>9 hours). Patients normally stayed in the resident room (bedroom) and were scheduled for an hour of daily exercise and activity in the therapy room. Beacons were mounted at locations of interest (Table 4.2.3), shown with color dots in Figure 4.1 within bedroom and therapy room. Take into account that despite imposing an identical protocol for all patients, daily collected data from each individual may differ. This is primarily because of patients not complying with the protocol at all times, losing interest during the day, feeling uncomfortable, and getting concerned about their privacy. Therefore, to provide a situation in which a fair comparison among patients can be enforced, we determined analysis inclusion criteria.

Table 4.1: Locations of interest. For sensor-based feature assessment throughout the paper, shower, toilet, and sink are considered as bathroom; walls 1, 2, and 3 as wall; beds 1 to 4 inside the therapy room and beds 1 and 2 inside the resident room as beds.

Location	Sublocations
Resident room	Bed, chair, shower, toilet
Therapy room	Bed, resband, bike, endorphine, strip, table, small table, hallway, seats, wall, hallway doors, sink, bath

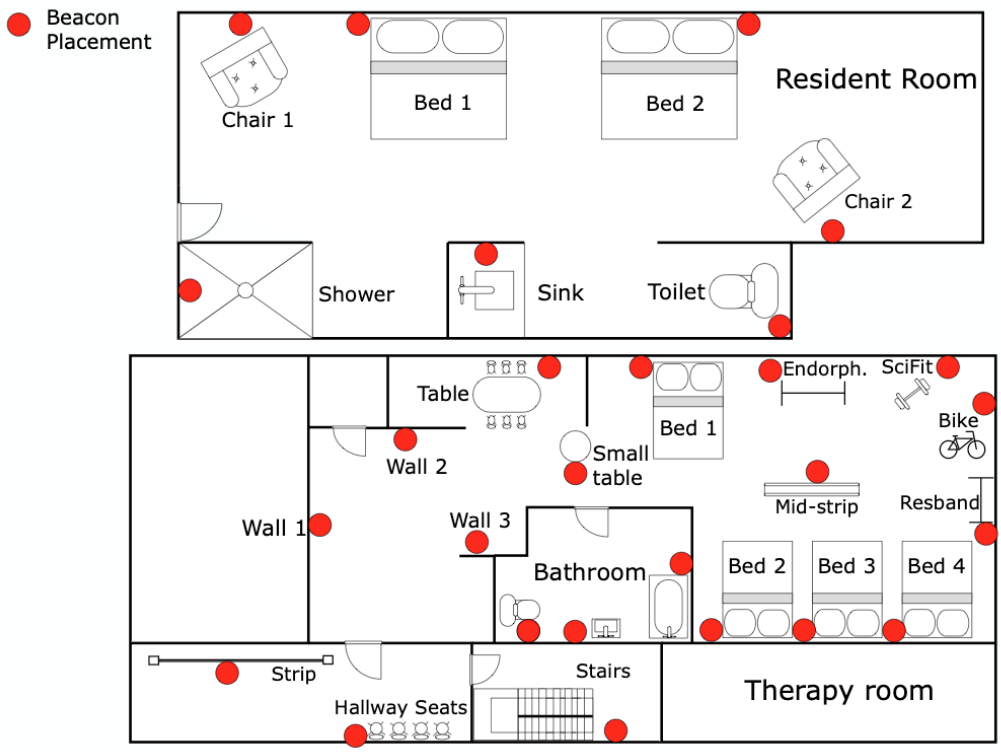


Figure 4.1: Subacute rehabilitation facility map: resident room on top and therapy room at the bottom with locations of mounted beacons shown in red.

4.2.4 Analysis Inclusion Criteria

4.2.4.1 Baseline predictive task

For the baseline analysis, we included study participants who satisfied the following constraints: (1) patients with 4 hours or more of watch wear time data in at least 1 day within the first 3 days of admission (defined as baseline); and (2) having 15 min or more of therapy room wear time in that particular baseline day. In case both inclusion criteria were satisfied on more than 1 day, the earliest day was selected as baseline. The reason for choosing 4 hours or more wear time was to set a standard minimum; given the health of this population whom mostly recently discharged from the hospital, we anticipated variability in watch usage. To have a minimum standard, we agreed that patients needed to wear the watch more than 50% of the available hours per day (in this study, 8 hours).

4.2.4.2 Longitudinal data analysis

The analysis inclusion criteria of longitudinal study were defined to ensure all patients satisfy a minimum amount of daily sensor data and collected PT and OT assessments. Analysis criteria include patients with the following data: (1) ≥ 3 days of watch data; (2) each day ≥ 4 hours of watch wear time; and (3) ≥ 3 sessions of PT or OT or a combination of both PT and OT. Cohort data were agglomerated for analyses according to the consort diagram shown in Figure 4.2.

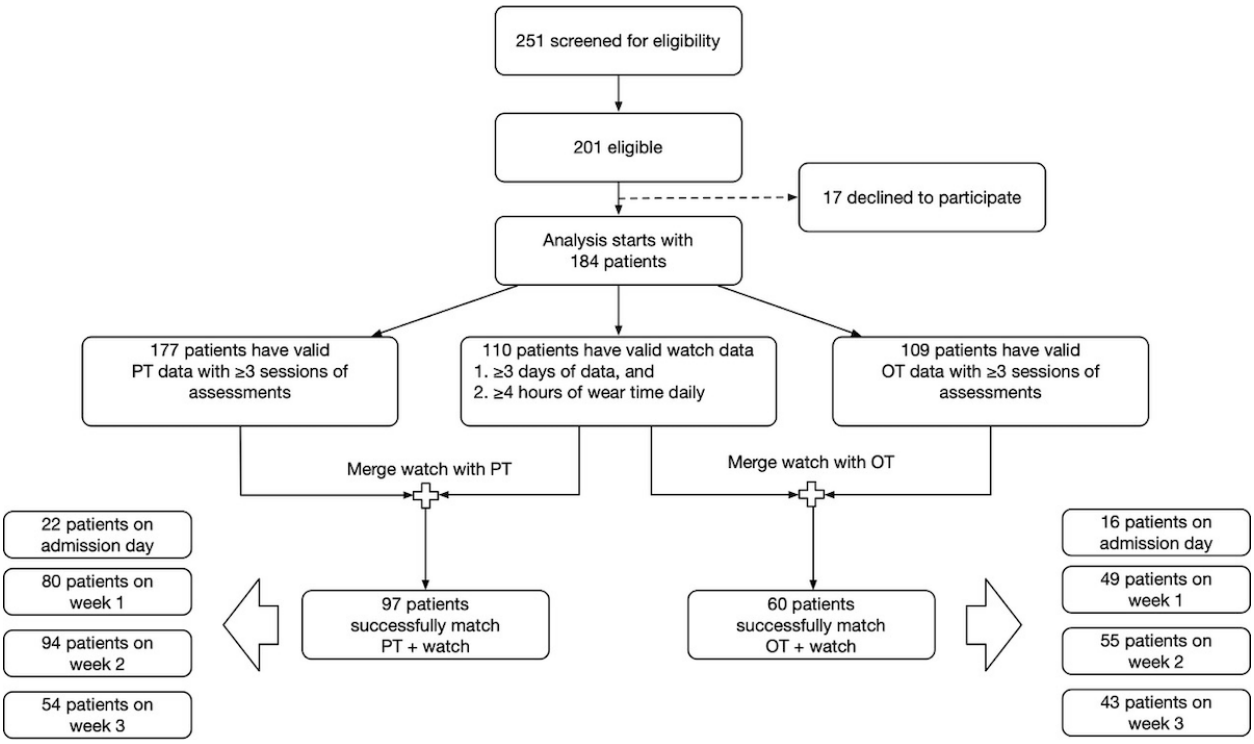


Figure 4.2: Diagram describing the analysis cohort. OT: occupational therapy; PT: physical therapy.

The hours when the watch was not worn were excluded from the study; therefore, hours may not be consecutive.

4.2.5 Measures

4.2.5.1 Demographic and Clinical Characteristics

We collected the demographic characteristics of patients such as age, race, gender, and ethnicity. We also translated the clinical coordinator's assessments including usage of assistive devices and their type, measures of activity of daily living (ADL), pain (yes/no), and number of active diagnosis (more or less than 10). We investigated the significance of such characteristics in distinguishing the outcome (community vs hospital).

4.2.5.2 Sensor-Based Parameters

Sensor-based features are combination of 3 groups of parameters that are achieved by harnessing smartwatch and BLE beacons. The features are based on (1) activity recognition such as sitting time and standing time; (2) indoor localization, for example, time in bed, time in bathroom, or therapy room; and (3) row acceleration quantification, MAD (energy; see section Sensing At-Risk Population System Overview). By combining these attributes, we achieved features such as sitting time in bed or energy spent in walking or in bed.

To perform a fair comparison among patients with different watch wear time, we normalized features: time spent (minutes) in a certain physical activity or location was divided by uptime (the total watch wear time in a day in minutes) to yield normalized time features. Uptime is an essential factor in providing fair comparison

We investigated the significance of sensor-based features with respect to the outcomes: hospital versus community. All measurements are at baseline, that is, the day that satisfies inclusion criteria from 9 am to 6 pm. We calculated "time spent in percentage", "energy intensity (E)", and "energy spent in percentages", as shown in equations (4), (5), and (6) in Figure 2.2.

To recap, for each individual, time-related features such as sitting time were divided by uptime. Energy-related features such as walking were divided by: (1) the uptime, yielding energy intensity and (2) their total daily value, producing the energy percentage.

4.2.5.3 Clinical Features

Clinical assessments in this study are 2-fold: physical therapy (PT) and occupational therapy (OT). PT and OT metrics included functional activities such as bed mobility (includes rolling, moving between supine and sitting, scooting in supine, scooting on the edge of the bed), gait (movement patterns that make up walking and associated interpretations), transfers (moving body from one surface to another without walking), hygiene, toileting, and lower body dressing. Those activities were scored based on the functional levels (1 to 6), from independent to completely dependent [175]. A comprehensive collection of PT and OT key metrics were performed every week; hence, patients were expected to have ≥ 3 PT or OT assessments within 21 days. In this study, a subset of clinical features was chosen; these features were common in more than 65% (n=72) of patients' PT and OT visits. The most common PT functional activities, performed by more than 65% of the cohort, are as follows: gait distance (in feet), transfer activity, and bed mobility, including movement from supine to sit. Common OT functional activities are comprised of lower body dressing, toileting activity, hygiene, and overall ability to tolerate daily activities (activity tolerance).

4.2.6 Statistical Analysis

4.2.6.1 Baseline predictive task

We explored the capability of baseline sensor-based and demographic features to distinguish between subacute rehabilitation patients based on their outcomes (i.e., re-admitted to hospital (H) vs staying in the community (C) either long-term care or home). Chi-squared tests were used to compare categorical demographic variables between outcome groups. We compared quantitative demographic variables and sensor-based metrics (physical activity derived from watch accelerometer and indoor localization inferred from BLE beacons RSSI) between groups using the Kruskal-Wallis test. Cohen's d was used to summarize the effect size and illustrate the discriminatory power of each feature. Commonly, 0.2, 0.5, and 0.8 are Cohen's d cut-off values indicating small, medium, and large effect size, respectively. Spearman rho was used to measure correlations between physical activity and location-based features.

4.2.6.2 Longitudinal data analysis

Visualization of prior analysis was generated to unveil any longitudinal patterns. The time trends of sensor-based features appeared to be approximately linear; hence, we decided to use linear models for longitudinal analysis.

Descriptive statistics (medians and IQR) were computed for clinical assessments (i.e., PT and OT) at each session. Generalized linear mixed effect model was used to understand the longitudinal relationships between the clinical measures and the sensor-based features [176, 177, 178]. Due to the frequency difference in which sensor and clinical assessments were collected, we merged a day of clinical assessment data with its corresponding day or closest day containing the sensor data (SD 3 days).

Three models, each with different sets of sensor-based features, were constructed for each clinical outcome. Model 1 included overall energy intensity as covariate. Model 2 considered energy intensity at resident room and energy intensity at therapy area as covariates. Additionally, sensor-based activity parameters (e.g., energy intensity of sitting) were used in model 3. Linear time indicates the number of weeks since the enrollment day. Interaction effects of sensor features with time were also included.

4.2.6.3 Predictive Models of Outcome

We investigated the capability of features at baseline to triage and predict patients who were re-admitted to the hospital or who stayed in community. We built random forest models (maximum depth=2, random state=40, and class_weight=balanced), with hospital patients as positive group. We used single or combination of features with highest statistical significance in distinguishing outcomes according to Kruskal-Wallis tests. Model generation and evaluating performance characteristics (3-fold cross-validation) including sensitivity, specificity, accuracy, and area under the curve (AUC) estimation were performed using Python Programming Language libraries Pandas (version 0.21.0) and Numpy (version 1.14.5), Scipy (version 1.0.0), and Scikit-learn (version 0.19.1) [179, 180, 181].

4.3 Results: baseline prediction task

4.3.1 Demographic and Clinical Characteristics

From 184 consented subjects, 30 were excluded because of not satisfying the analysis inclusion criteria. A total of 154 patients were included in this study in which 145 (94.2%) of subjects discharged home/community (C), and 9 (5.8%) re-admitted to hospital (H) at the end of their rehabilitation process. Table 4.2 presents detailed sociodemographic and clinical characteristics of this cohort, such as age, gender, race-ethnicity, presence of pain, number of active diagnoses, usage of assistive devices, and ADL. Table 4.2 indicates the mean (SD) and number of patients included for every particular parameter. Among the clinical assessments, Table 4.2 shows that ADL Toilet is significant in determining the outcome (P=.007) with 65% of the cohort in need of extensive assistance and 35% limited assistance.

Table 4.2: Sociodemographic and clinical characteristics of the cohort of 154 patients.

Parameter	Community	Hospital	Community vs hospital (P value)
Subjects, n(%)	145 (94.2)	9 (5.8)	_ [a]
Age (years), mean (SD)	82.16 (9.55)	84.22 (13.87)	.24
Gender, n (%)			.56
Female	104 (71.7)	4 (44.4)	
Male	41 (28.3)	5 (55.6)	
Race/ethnicity, n(%)			>.99
Asian	5 (3.4)	0 (0.0)	
Black/African American	14 (9.7)	0 (0.0)	
Hispanic/Latino	4 (2.7)	0 (0.0)	
Native/hawaiian Pacific Islander	3 (2.1)	0 (0.0)	
White	119 (82.1)	9 (100)	

Table 4.2 continued from previous page

Parameter	Community	Hospital	Community vs hospital (P value)
Pain present, n (%)			.92
No	44 (31.7)	1 (14.3)	
Yes	95 (68.3)	6 (85.7)	
Active diagnoses, n (%)			>.99
<10	22 (15.2)	1 (11.1)	
>= 10	123 (84.8)	8 (88.9)	
ADL [b] transfer, n (%)			.77
Limited assistance	65 (45.1)	2 (22.2)	
Extensive assistance	79 (54.9)	7 (77.8)	
ADL dress, n (%)			.96
Limited assistance	32 (22.2)	1 (11.1)	
Extensive assistance	112 (77.8)	8 (88.9)	
ADL eat, n (%)			.91
Independent	128 (88.9)	7 (77.8)	
Supervision	4 (2.8)	0 (0.0)	
Limited assistance	9 (6.2)	1 (11.1)	
Extensive assistance	3 (2.1)	1 (11.1)	
ADL toilet, n (%) [c]			.007
Limited assistance	50 (34.7)	1 (11.1)	
Extensive assistance	94 (65.3)	7 (77.8)	
Total dependence	0 (0.0)	1 (11.1)	
ADL walk room, n (%)			.73
Limited assistance	73 (50.7)	2 (22.2)	
Extensive assistance	59 (41.0)	5 (55.6)	
Activity did not occur	12 (8.3)	2 (22.2)	

Table 4.2 continued from previous page

Parameter	Community	Hospital	Community vs hospital (P value)
ADL walk hall, n (%)			.88
Limited assistance	73 (50.7)	2 (22.2)	
Extensive assistance	64 (44.4)	6 (66.7)	
Activity occurred only once or twice	2 (1.4)	0 (0.0)	
Activity did not occur	5 (3.5)	1 (11.1)	
ADL walk on unit, n (%)			.85
Supvervision	1 (0.7)	0 (0.0)	
Limited assistance	71 (49.3)	2 (22.2)	
Extensive assistance	72 (50.0)	7 (77.8)	
ADL hygiene, n (%)			.84
Supervision			
Limited assistance			
Extensive assistance			
ADL bed, n (%)			.61
Supervision	1 (0.7)	0 (0.0)	
Limited assistance	83 (57.6)	2 (22.2)	
Extensive assistance	60 (41.7)	7 (77.8)	
Urinary continence, n (%)			.09
Always continent	117 (81.2)	4 (44.4)	
Occasionally incontinent	4 (2.8)	0 (0.0)	
Frequently incontinent	8 (5.6)	2 (22.2)	
Always incontinent	7 (4.8)	3 (33.3)	
Not rated	8 (5.6)	0 (0.0)	
Bowel continence, n (%)			.08
Always continent	128 (88.9)	5 (55.6)	

Table 4.2 continued from previous page

Parameter	Community	Hospital	Community vs hospital (P value)
Occasionally incontinent	3 (2.1)	0 (0.0)	
Frequently incontinent	7 (4.8)	1 (11.1)	
Always incontinent	6 (4.2)	3 (33.3)	
Assistive devices, n (%)			.97
Walker	1 (0.7)	0 (0.0)	
Wheelchair	5 (4.0)	1 (14.3)	
Walker and wheelchair	123 (94.6)	6 (85.7)	
Cane and wheelchair	1 (0.7)	0 (0.0)	

[a] Not applicable.

[b] ADL: activity daily living.

[c] Parameters with $P < .05$.

4.3.2 Energy Intensity Features Assessment

Amongst sensory-based features shown in Figure 2.2, equations (4-6), energy intensity features are the ratio of the total energy spent in a particular activity or location to their corresponding time spent. Taking into account, indoor localization capability of SARP system enabled us to calculate the energy spent at each location of interest, sum of which was broadly categorized into (1) energy intensity in resident room and (2) energy intensity in therapy room. According to Table 4.3, energy features that best discriminated community and hospital patients were energy intensity in resident room (P<.001, d=1.21), resident_bed (P<.001, d=1.23), resident_bath (P=.004, d=1.18), and total energy intensity (P=.003, d=0.87). Features such as energy intensity of laying down (P=.02), and therapy_bathroom (P=.02), despite statistical significance, have low effect sizes (d=0.418 and d=0.17, respectively). Moreover, with P<.001 and d=1.25, energy intensity in resident room has high discriminatory power with respect to outcome.

Figure 4.3 depicts the energy intensity distributions between 2 groups in resident and therapy rooms. It shows that energy intensity in therapy room in both groups has similar mean value (line within the box); therefore, a clear distinction cannot be made within 2 groups based on that feature. However, the mean value of community group in resident room is clearly higher than in hospital patients.

Kernel density estimation (KDE) distributions are shown in Figure 4.4 (subplots A and D). The figure attests to the distinction of energy intensity in resident room among community and hospital patients (subplot A). However, the KDE of energy intensity in therapy room (subplot D) does not indicate the same discriminatory power. Figure 4.4 (subplot B) indicates that energy intensity of most patients in therapy room is higher compared with resident room for both outcome groups because most patients fall below the identity line. Points shown on the identity line represent patients with same therapy and resident intensities. According to subplot (C), the center core of the contour plot (representing most patients) in community group is almost circular contrary to hospital patients. This indicates that the ratio of resident to therapy intensity is closer to one (ie, same activity intensities). On the contrary, more oval shape of the contour core in hospital group can imply that most patients are persistently more active during therapy sessions while being less active in their resident room. The increase in energy levels can be seen clearly in Figure 4.5. The figure depicts the ratio of energy intensity in therapy room to resident room. Most patients in hospital outcome group, demarcated by red line, fall around number 2. In other words, therapy room energy intensity is twice the resident room for most patients in hospital group. However, 50 patients in community group (blue histogram) have the ratio close to 1, that is, the same intensity in both therapy and resident room. A more detailed scenario of both groups within the therapy room can be found in Figure 4.6 and Table 4.4.

Average time spent and energy intensity at each therapy location stratified by groups are shown in Figure 4.6. It is clear that hospital group spent no time at stairs, scifit, table, and endorphine. The 5 most intensive activities were small table, stairs, scifit, table, and bike. Small table and table are places where patient normally carried out hand pedaling exercises. Table 4.4 further highlights the details of therapy room location/facility usage in each group. More than 70% of participants

from both groups had used bed and bathroom in therapy room, with bathroom 's P<.05 (Table 4.3). However, it is worth mentioning that the effect size of bathroom energy intensity is small: 0.17 (cut-off regions: 0.2 small, 0.5 medium, and 0.8 large). Furthermore, Figure 4.6 reveals that both groups' intensities at bed and bathroom were less than 60 per min. In a study by Razjouyan et al [36], a cutoff point of 90 is suggested to differentiate between light and moderate-to-vigorous activities.

Figure 4.7 illustrates Spearman correlations among features. According to annotations explained in the Features section, E indicates energy intensity, E% denotes energy percentage, and T% shows the percentage of time spent. Circles, contrary to ovals, correspond to low correlation, whereas lines imply the highest correlation. Darker spectrum on either side (red or blue) represents higher correlation; red implies positive, whereas blue is indicative of negative correlation. It is clear from the figure that laying down is negatively correlated with the rest of the features. Bath and bed in resident room are understandably correlated strongly with energy spent in resident room because almost all activities happened in those 2 locations, and patients hardly used the chair. Bed, bath, resband, small table, bike, and scifit are strongly correlated with energy spent in therapy room. It is clear that being active is highly correlated with overall energy intensity. Resident room energy intensity is strongly correlated with overall energy intensity.

Table 4.3: Sensor-based (activity and indoor localization) features: assessment according to outcomes. **C** denotes "Community" group and **H** denotes "Hospital" group.

Feature	**Community, mean (SD)**	**Hospital, mean (SD)**	**P value**	**Effect size** [1]	**Frequency**	
					C	**H**
Energy % parameters						
Active [2]	2.37 (3.84)	1.00 (1.29)	.001	1.24	145	9
Walking	2.37 (3.84)	1.00 (1.29)	.08	0.50	145	9
Standing [2]	59.70 (8.70)	57.92 (6.39)	.002	1.24	145	9
Sitting [2]	17.83 (9.69)	13.33 (8.90)	.02	0.86	145	9
Laying down [2]	20.10 (6.43)	27.73 (9.94)	.04	0.54	145	9

Table 4.3 continued from previous page

Feature	Community, mean (SD)	Hospital, mean (SD)	P value	Effect size [1]	Frequency	
					C	H
Energy intensity parameters						
Total energy [2]	52.61 (18.23)	35.85 (16.53)	.003	0.87	145	9
Active	11.94 (18.27)	6.05 (8.02)	.30	0.42	145	9
Walking	450.47 (253.08)	366.45 (218.66)	.44	0.34	145	9
Standing	85.93 (26.92)	82.27 (36.12)	.32	0.11	145	9
Sitting	184.33 (97.58)	156.19 (104.74)	.31	0.28	145	9
Laying down [2]	26.23 (8.68)	19.54 (7.35)	.02	0.418	145	9
Energy intensity therapy room						
Energy therapy room	70.75 (43.11)	68.49 (63.56)	.36	0.04	145	9
Bathroom [2]	74.84 (49.02)	62.35 (83.54)	.02	0.17	114	8
Strip	57.84 (42.33)	13.03 (8.30)	.06	1.43	88	2
Bed	60.22 (40.27)	39.09 (7.15)	.27	0.72	97	4
Resband	61.06 (43.10)	75.73 (85.49)	.57	0.20	100	6
Bike	91.80 (76.82)	120.58 (38.41)	.31	0.43	36	2
Scifit	98.39 (55.04)	0.0 (0.0)	_[3]	_	14	0
Endor	41.38 (6.74)	0.0 (0.0)	-	-	3	0
Midstrip	56.46 (48.92)	65.46 (24.53)	.38	0.22	45	3
Small table	61.07 (40.37)	148.47 (138.78)	.53	.71	57	3
Table	93.49 (66.75)	0.0 (0.0)	-	-	56	0
Hallway seats	42.58 (43.13)	32.52 (7.89)	.87	0.32	43	3
Stairs	133.48 (128.07)	0.0 (0.0)	-	-	8	0
Wall	57.07 (28.49)	25.61 (0.0)	.17	-	73	1
Energy intensity resident room						
Energy resident room [2]	43.32 (17.44)	26.99 (6.05)	<.001	1.25	145	9
Bed [2]	43.93 (19.01)	25.76 (4.37)	<.001	1.23	144	9
Bathroom [2]	55.89 (27.95)	32.50 (9.30)	.004	1.18	141	9
Chair	42.45 (20.61)	0.0 (0.0)	-	-	5	0

Table 4.3 continued from previous page

Feature	Community, mean (SD)	Hospital, mean (SD)	P value	Effect size [1]	Frequency	
					C	H
Time % parameters						
Active [2]	12.92 (6.52)	6.94 (4.01)	.001	1.10	145	9
Walking	0.35 (0.51)	0.15 (0.27)	.09	0.44	145	9
Standing [2]	44.22 (7.94)	32.68 (7.30)	<.001	1.51	145	9
Sitting [2]	8.60 (8.36)	6.16 (7.36)	.04	0.31	145	9
Laying down [2]	46.83 (9.83)	60.99 (11.11)	<.001	1.35	145	9
Time spent % therapy room						
Bathroom	0.03 (0.04)	0.06 (0.08)	.16	0.27	114	8
Strip	0.01 (0.03)	0.005 (0.002)	.62	0.48	88	2
Bed	0.62 (0.19)	0.55 (0.23)	.64	0.43	97	4
Resband [2]	0.02 (0.02)	0.05 (0.03)	.03	0.74	100	6
Bike	0.03 (0.03)	0.01 (0.002)	.51	0.80	36	2
Scifit	0.03 (0.02)	0.0 (0.0)	-	-	14	0
Endor	0.009 (0.01)	0.0 (0.0)	-	-	3	0
Midstrip	0.02 (0.02)	0.02 (0.02)	.31	0.49	45	3
Small table[2]	0.02 (0.03)	0.04 (0.0p2)	.04	0.50	57	3
Table	0.06 (0.05)	0.0 (0.0)	-	-	56	0
Hallway seats	0.006 (0.004)	0.01 (0.16)	.64	0.78	43	3
Stairs	0.02 (0.04)	0.0 (0.0)	-	-	8	0
Wall	0.01 (0.02)	0.01 (0.0)	.98	-	73	1
Time spent % resident room						
Bed	0.62 (0.19)	0.55 (0.23)	.16	0.12	144	9
Bathroom	0.21 (0.17)	0.25 (0.20)	.92	0.52	141	9
Chair	0.007 (0.03)	0.0 (0.0)	-	-	5	0

[1]Effect sizes have been calculated as Cohen d
[2]Parameters with $P < .05$.
[3]Not applicable.

4.3.3 Energy Percentage Features Assessment

Energy percentage feature, as mentioned in Figure 2.2, is the percentage of energy spent in walking, sitting, standing, laying, or energy spent in locations of interest divided by total energy spent in that day. According to Table 4.3, community patients are more active (P=.001, d=1.24) than patients re-admitted to the hospital. Meanwhile, energy percentage of standing (P=.002, d=1.24) and sitting (P=.02, d=0.86) of the community group is higher than those in hospital group. Other than walking, all energy percentage parameters were shown significant in distinguishing between both groups. Walking is not significant in distinguishing the outcome: Energy (%) in walking (P=.08, d=0.50) and energy intensity during walking (P=.44, d=0.34).

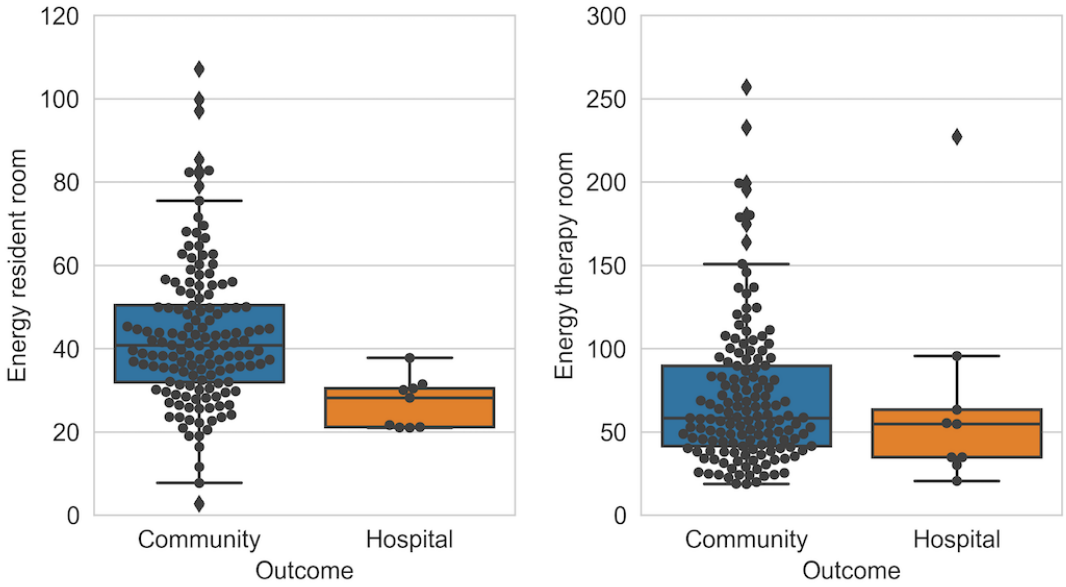


Figure 4.3: Energy intensity distribution.

4.3.4 Time Features Assessment

According to Table 4.3, standing time (%) has the strongest discriminatory power (P<.001, d=1.51) among all watch-derived parameters. Community group has higher time percentage in laying down

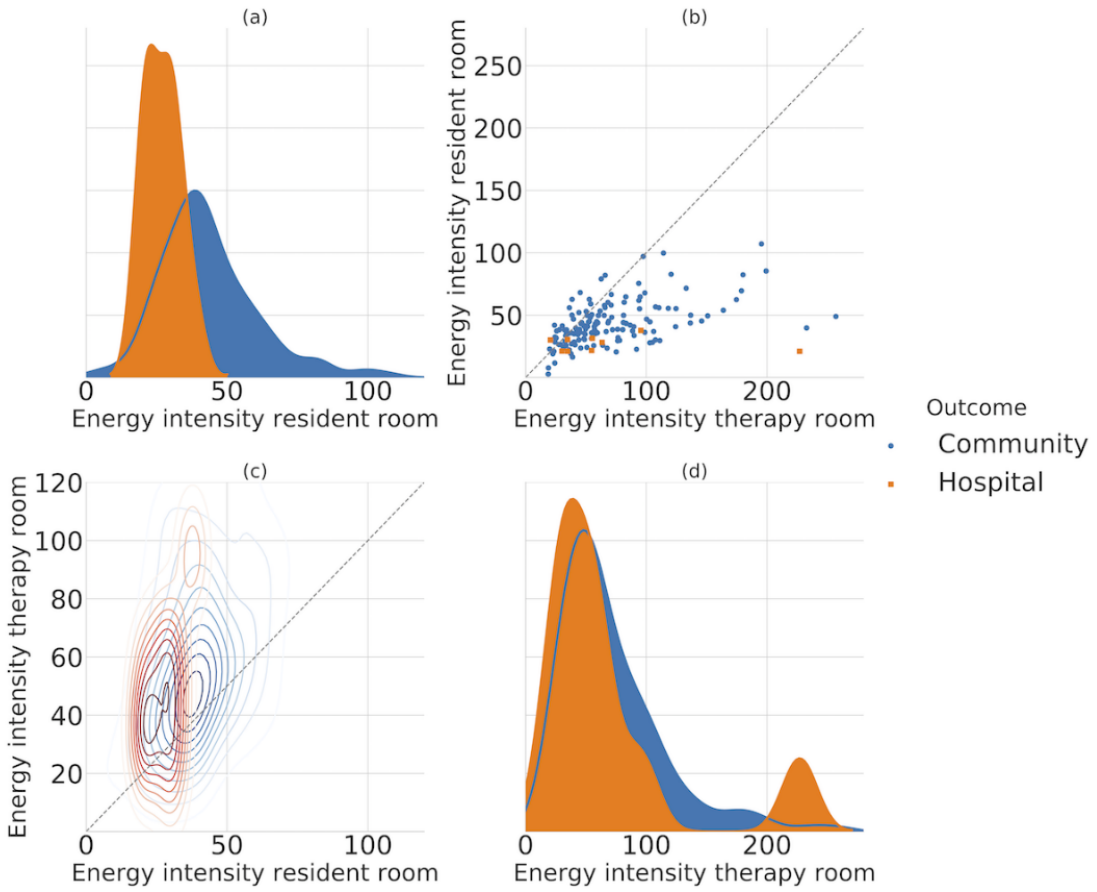


Figure 4.4: Gauging energy intensity in community versus hospital.

(P<.001, d=1.35) and active state (P=.001, d=1.24) compared with hospital group. Despite statistical significance of sitting time (%), its effect size is between small and medium (P=.04, d=0.31). Walking time was quite negligible (<1% of time for both groups with P=.09, d=0.44), whereas overall active state, which captures walking and stationary active periods, was highly significant (P=.001, d=1.10). As shown in the table, none of the time (%) parameters in resident room have the ability to discriminate between the 2 outcome groups.

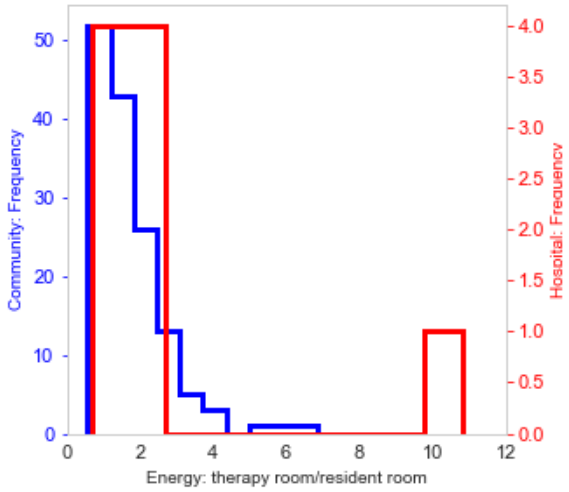


Figure 4.5: Distribution of patients spending energy in therapy room compared with resident room. X-axis indicates the ratio of energy in therapy to resident room.

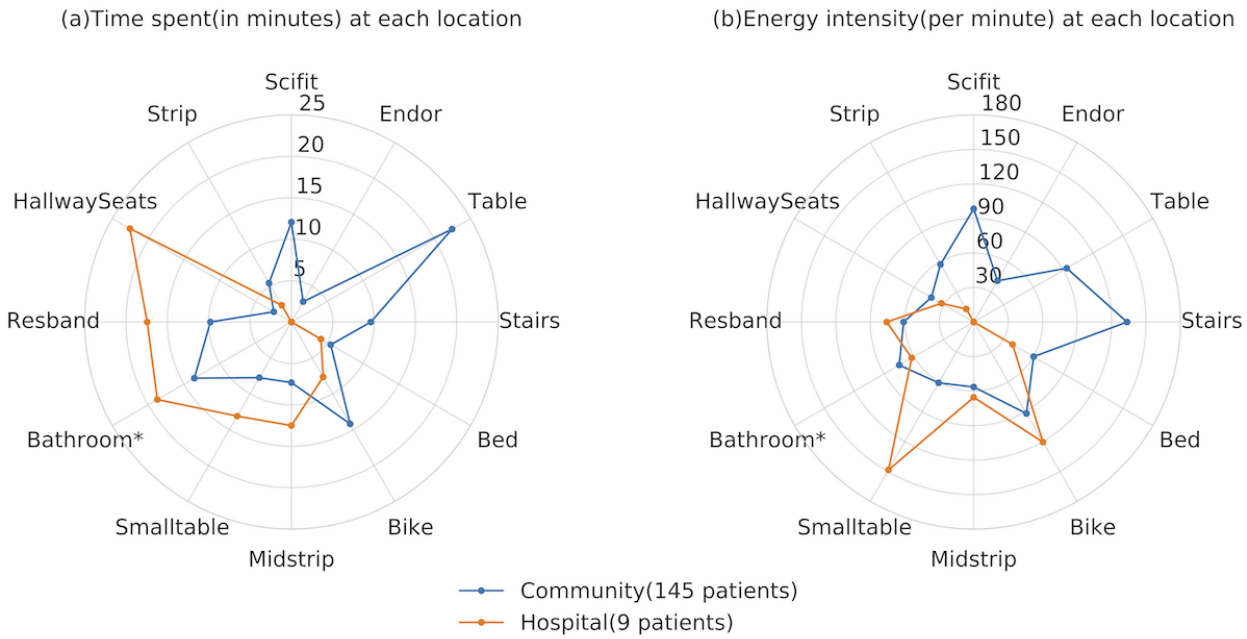


Figure 4.6: Time and energy intensity details of therapy room.

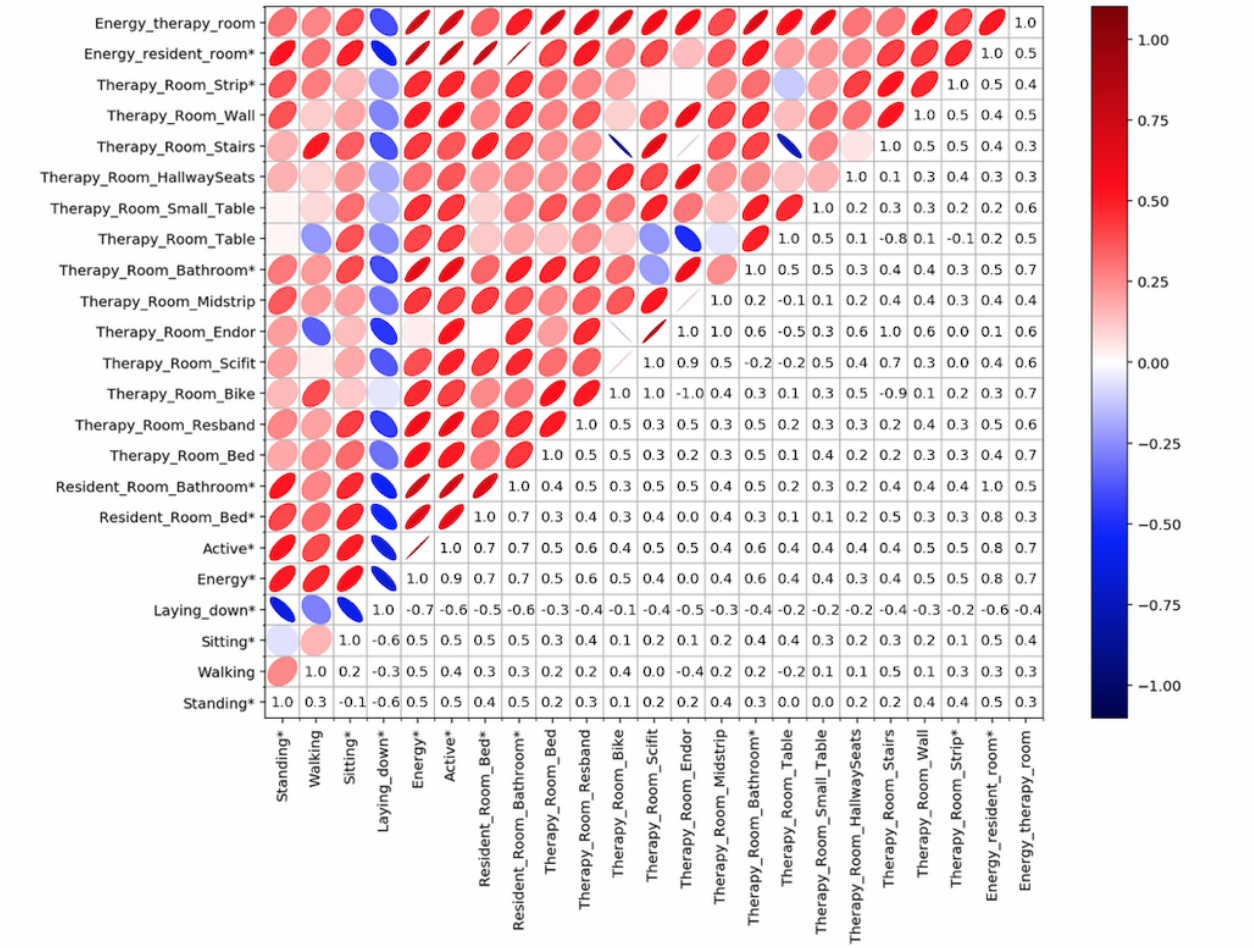


Figure 4.7: Correlations among sensor-based features. Asterisk indicates parameters with $P < .05$.

4.3.5 Performance of Predictive Models at Baseline

Random forest models were built based on the most statistically significant features. In reviewing Table 4.2, the top 3 most influential features in distinguishing the outcomes were % standing time (P<.001, d=1.51), % laying down time (P<.001, d=1.35), and resident room energy intensity (P<.001, d=1.25). Results of 3-fold cross-validation models with their corresponding AUC score are presented in Table 4.5. Take into account that the sensitivity (recall) presented in the table is not the weighted average and reflects only recall of minority (H) group. Specificity indicates the true negative rate when negative group is comprised most patients returning to community setting (C)

Table 4.4: Frequency of therapy room location/facility usage by group.

Location/facility	**Frequency of facility usage**	
	Community, n (%)	Hospital, n (%)
Scifit	14 (9.6)	0 (0.0)
Endor	3 (2.1)	0 (0.0)
Table	56 (38.6)	0 (0.0)
Stairs	8 (5.5)	0 (0.0)
Bed	118 (81.4)	7 (77.8)
Bike	36 (24.8)	2 (22.2)
Midstrip	45 (31.0)	3 (33.3)
Small table	57 (39.3)	3 (33.3)
Bathroom[a]	114 (78.6)	9 (100.0)
Resband	100 (69.0)	7 (77.8)
Hallway seat	43 (29.7)	3 (33.3)
Strip	88 (60.7)	3 (33.3)

[a]Parameters with $P < .05$.

after the rehabilitation period.

4.4 Results: longitudinal data analysis

4.4.1 Demographic and Clinical Characteristics

From 184 consented patients, 110 (60%) met the watch wearing time protocol with mean age of 79.4 (SD 5.9) years. Moreover, 97 (88%) patients were included in PT-watch paired analysis and 60 (54%) in OT with watch analytics. Most participants were female (n=79, 72%) and of White race or ethnicity (n=84, 76%). Additionally, 62% (n=69) of the patients had pain, 99% (n=109) of them needed some level of assistance with functional mobility activities (transfer activity), and 75% (n=83) needed assistive devices for walking. Table 4.6 presents detailed sociodemographic and

Table 4.5: Predictive models: 3-fold cross-validation (community, n=48; hospital, n=3).

Features	Sensitivity mean (SD)[a]	Specificity mean (SD)[a]	Accuracy mean (SD)[a]	AUC[b] mean (SD)
Standing time (%)	22.2 (31.4)	74.4 (15.3)	71.4 (12.9)	0.62 (0.06)
Standing time (%), laying down time (%)	11.1 (15.7)	91.0 (0.9)	86.4 (1.5)	0.70 (0.10)
Standing time (%), laying down time (%), resident room energy intensity (%)	44.4 (41.6)	87.6 (4.3)	85.1 (5.5)	0.85 (0.09)
Resident room energy intensity	77.7 (15.7)	74.5 (8.5)	74.7 (7.3)	0.84 (0.10)

[a]Mean (SD) reported for the validation datasets based on a 3-fold cross-validation. Mean and SD are calculated across all 3 folds.

[b]AUC: area under the curve.

clinical characteristics of the 110 patients. ADL parameters and their significance in determining the outcome are presented based on initial assessments, at the time of admission, or within one day.

Table 4.6: Sociodemographic and clinical characteristics (initial assessment) of the cohort of 110 patients.

Parameter	Community	Hospital	Community vs hospital (P value)
Subjects, n(%)	105 (95.5)	5 (4.5)	_ [a]
Age (years), mean (SD)	78.0 (5.7)	84.1 (6.8)	.03
Gender, n (%)			>.99
Female	76 (72.4)	3 (60)	
Male	29 (27.6)	2 (40)	
Race/ethnicity, n(%)			>.99
Asian	5 (4.8)	0 (0)	

Table 4.6 continued from previous page

Parameter	Community	Hospital	Community vs hospital (P value)
Black/African American	12 (11.4)	1 (20)	
Hispanic/Latino	2 (1.9)	0 (0)	
Native/hawaiian Pacific Islander	2 (1.9)	0 (0)	
White	84 (80)	4 (80)	
Pain present, n (%)			.95
No	29 (30)	2 (50)	
Yes	67 (70)	2 (50)	
Active diagnoses, n (%)			.86
<10	22 (21)	0 (0)	
>= 10	83 (79)	5 (100)	
Transfer, n (%) [b]			.87
Supervision	1 (1)	0 (0)	
Limited assistance	57 (55)	1 (20)	
Extensive assistance	46 (44)	4 (80)	
Dressing, lower body, n (%)			.93
Independent	1 (1)	0 (0)	
Limited assistance	28 (27)	0 (0)	
Extensive assistance	75 (72)	5 (100)	
Eating, n (%)			.93
Independent	90 (90)	4 (80)	
Supervision	4 (4)	1 (20)	
Limited assistance	4 (4)	1 (0)	
Extensive assistance	2 (2)	1 (0)	
Toileting, n (%)			.70
Independent	1 (1)	0 (0)	

Table 4.6 continued from previous page

Parameter	Community	Hospital	Community vs hospital (P value)
Limited assistance	45 (43)	0 (0)	
Extensive assistance	58 (56)	5 (100)	
Walk room, n (%)			.91
Supervision	1 (1)	0 (0)	
Limited assistance	61 (59)	1 (20)	
Extensive assistance	34 (32)	3 (60)	
Activity did not occur	8 (8)	1 (20)	
Walk hall, n (%)			.92
Supervision	1 (1)	0 (0)	
Limited assistance	62 (60)	1 (20)	
Extensive assistance	35 (33)	4 (80)	
Activity occurred only once or twice	1 (1)	0 (0)	
Activity did not occur	5 (5)	0 (0)	
Walk on unit, n (%)			.78
Supvervision	1 (1)	0 (0)	
Limited assistance	62 (60)	1 (20)	
Extensive assistance	41 (39)	4 (80)	
Hygiene, n (%)			.84
Independent	1 (1)	0 (0)	
Limited assistance	59 (57)	2 (40)	
Extensive assistance	44 (42)	3 (60)	
Bed mobility, n (%)			.96
Supervision	1 (1)	0 (0)	
Limited assistance	68 (65)	2 (40)	
Extensive assistance	35 (34)	3 (60)	

Table 4.6 continued from previous page

Parameter	Community	Hospital	Community vs hospital (P value)
Urinary continence, n (%) [b]			.002
Always continent	85 (82)	1 (20)	
Occasionally incontinent	3 (3)	0 (0)	
Frequently incontinent	7 (6)	1 (20)	
Always incontinent	4 (4)	3 (60)	
Not rated	5 (5)	0 (0)	
Bowel continence, n (%) [b]			.006
Always continent	91 (87)	2 (40)	
Occasionally incontinent	3 (3)	0 (0)	
Frequently incontinent	5 (5)	0 (0)	
Always incontinent	5 (5)	3 (60)	
Assistive devices, n (%)			>.99
Wheelchair	3 (4)	0 (0)	
Walker and wheelchair	75 (95)	4 (100)	
Cane and wheelchair	1 (1)	0 (0.0)	

[a] Not applicable.

[b]Parameters with $P < .05$.

4.4.2 Longitudinal Analysis of All Features (Sensor and Clinical Measurements)

The community group spent higher overall energy intensity and energy intensity at the resident room compared to the hospital group, as seen in Figures 4.8 (a) and (b). However, energy intensity during therapy sessions tends to have similar values between two groups, especially toward the end of the rehabilitation period, as seen in Figure 4.8 (c).

The descriptive statistics of clinical parameters are summarized in Table 4.7. It shows that "gait

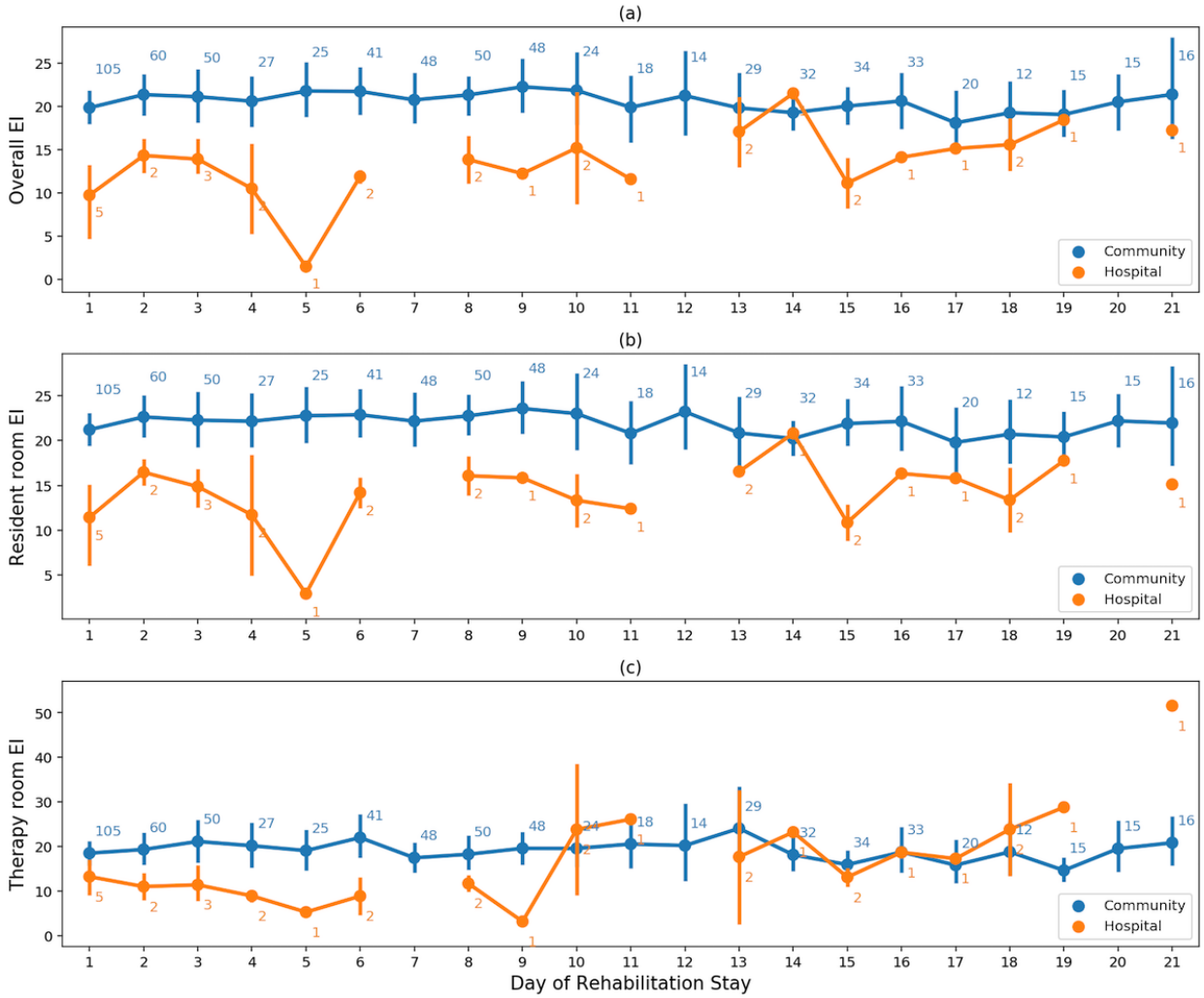


Figure 4.8: Energy intensity averaged per days in 21 days.

distance feet" increases over time (median and IQR after the first week), and "activity tolerance" increases (IQR after first week and median after second week). The table indicates no clear improvements in other clinical-based measures gauged by PT and OT functional levels within 3 weeks.

Table 4.7: Descriptive statistics of all measures.

Measures	Admission day			Week 1			Week 2			Week 3		
	N	Median	IQR	N	Median	IQR	N	Median	IQR	N	Median	IQR
Sensor features												
Overall EI [a]	110	17.97	13.00 ∼23.74	110	18.88	13.76∼25.17	83	19.30	14.97∼25.05	57	18.43	15.10∼23.37
Resident room EI	110	19.41	14.90∼24.74	110	19.94	15.58∼25.85	83	20.65	16.12∼25.66	57	19.45	15.69∼24.39
Therapy room EI	110	15.09	9.02∼25.36	110	15.29	9.83∼25.01	83	17.19	10.30∼24.34	57	14.96	11.20∼20.50
Occupational therapy features												
Dressing, lower body	16	3.00	2.75∼3.00	39	4.00	3.00∼4.00	40	4.00	4.00∼4.00	31	4.00	4.00∼4.00
Toileting general	16	4.00	2.75∼4.00	37	4.00	3.00∼4.00	40	4.00	4.00∼4.00	29	4.00	4.00∼4.00
Activity tolerance general (min)	11	8.00	5.00∼9.00	34	15.00	10.00∼15.00	37	15.00	15.00∼20.00	29	20.00	15.00∼20.00
Hygiene grooming	4	4	4.00∼4.00	15	4.00	4.00∼4.00	19	4.00	4.00∼4.00	15	4.00	4.00∼4.00
Physical therapy features												
Transfer general	20	4.00	3.75∼4.00	72	4.00	4.00∼4.00	86	4.00	4.00 ∼4.00	50	4.00	4.00∼4.00
Gait distance, feet	20	40.00	18.75∼50.00	70	100.00	71.25∼150.00	80	150.00	100.00∼200.00	44	150.00	97.50∼200.00
Gait assistive device	21	2.00	1.00∼2.00	60	2.00	2.00∼2.00	69	2.00	2.00∼2.00	38	2.00	2.00∼2.00
Gait level surface	18	4.00	4.00∼4.00	61	4.00	4.00∼4.00	71	4.00	4.00∼4.00	40	4.00	4.00∼4.00
Bed mobility supine sit	21	4.00	3.00∼4.00	72	4.00	4.00∼4.00	84	4.00	4.00∼4.00	49	4.00	4.00∼4.00

[a]EI: energy intensity.

4.4.3 Longitudinal Association Between Clinical Measures and Sensor-Based Features

The associations of repeated PT, OT, and sensor-based measurements are modeled through three generalized linear mixed models. On PT and sensor associations, according to Table 4.8, the results of model 1 revealed that gait distance feet (β=.28; SE=0.06; P<.001), gait level surface β=.17; SE=0.04; P<.001, and bed mobility including supine to sit (β=.26; SE=0.05; P<.001) improved over time. Higher overall energy intensity indicates a higher score of transfer activity (β=.22; SE=0.08; P=.03).

In model 2, energy intensity at the therapy room was positively associated with transfer activity (β=.19; SE=0.08; P=.02). In addition, gait distance feet (β=.28; SE=0.05; P<.001), gait level surface (β=.17; SE=0.04; P<.001) and bed mobility including supine to sit (β=.26; SE=0.05; P<.001) improved every week.

In model 3, sitting energy intensity showed positive association with transfer activity (β=.16; SE=0.07; P=.02). Meanwhile, according to model 3, participants showed weekly improvements in gait distance (measured in feet; β=.27; SE=0.06; P<.001), gait level surface (β=.16; SE=0.05;

Table 4.8: Generalized linear mixed model association between physical therapy and occupational therapy assessments with sensor-based features

Models	Gait distance feet		Transfer general		Gait level surfaces		Bed mobility supine sit		Dressing lower body		Toileting general		Activity tolerance general	
	Estimate β	SE	Estimate β	SE	Estimate β	SE	Estimate β	SE	Estimate β	SE	Estimate β	SE	Estimate β	SE
Model 1														
Intercept	-.01	0.09	-.01	0.09	.02	0.11	.01	0.09	<.01	0.01	.01	0.13	<.01	0.10
Time (weeks)	.28	0.06[a]	.08	0.05	.17	0.04[a]	.26	0.05[a]	.30	0.07a	.16	0.05[b]	.59	0.06[a]
Overall EI [c]	.14	0.08	.22	0.08[b]	.11	0.08	.18	0.08[b]	.19	0.09[b]	.23	0.09[b]	-.08	0.08
Time × overall EI	.01	0.06	-.05	0.05	-.07	0.05	-.09	0.05	-.09	0.07	-.04	0.06	-.01	0.07
Model 2														
Intercept	<-0.1	0.08	-0.2	0.09	.01	0.10	.01	0.09	<-.01	0.10	.01	0.13	<.01	0.10
Time (weeks)	.28	0.05[a]	.08	0.05	.17	0.04[a]	.26	0.05[a]	.29	0.07[a]	.15	0.05[b]	.59	0.06[a]
Resident room EI	.16	0.10	.06	0.09	.02	0.10	.14	0.09	.07	0.10	.14	0.10	.04	0.29
Therapy room EI	-.05	0.08	.19	0.08[b]	.10	0.08	.07	0.07	.16	0.10	.15	0.08	-.02	0.24
Resident room EI × time	.07	0.07	-0.4	0.07	.01	0.06	-.08	0.06	-.07	0.09	-.06	0.07	-.02	0.12
Therapy room EI × time	-.08	0.07	.02	0.07	-.10	0.06	-.01	0.06	.02	0.09	.05	0.08	-.01	0.10
Model 3														
Intercept	-.01	0.08	-.01	0.09	.02	0.11	.01	0.09	-.01	0.11	.02	0.14	<.01	0.10
Time (weeks)	.27	0.06[a]	.06	0.05	.16	0.05[a]	.26	0.05[a]	.32	0.07[a]	.18	0.05[a]	.59	0.06[a]
Sitting EI	.03	0.07	.16	0.07[b]	.03	0.06	<.01	0.06	.13	0.09	.09	0.07	.10	0.07
Standing EI	-.01	0.09	.06	0.08	.07	0.07	-.03	0.08	.07	0.11	.03	0.08	-.03	0.09
Laying down EI	.13	0.09	.06	0.09	.06	0.08	.14	0.08	.03	0.11	.10	0.11	-.14	0.09
Sitting EI × time	.03	0.06	-.04	0.05	-.01	0.05	-.02	0.05	-.15	0.08	-.13	0.06[b]	-.13	0.07
Standing EI × time	.08	0.07	.11	0.07	.02	0.06	.04	0.06	-.05	0.10	-.07	0.07	.04	0.09
Laying dowm EI × time	-.01	0.08	-.13	0.07	-.09	0.06	-.09	0.07	.11	0.11	.15	0.09	-.10	0.08

[a] $P < .001$.

[b] $P < .05$.

[c] EI: energy intensity.

P<.001), and bed mobility including supine to sit (β=.26; SE=0.05; P<.001).

On OT and sensor associations, Table 4.8 shows that lower body dressing, toileting activity, and activity tolerance in general improved every week in all three models. The higher value of overall energy intensity in model 1 implied a higher functional score of lower body dressing (β=.19; SE=0.09; P=.03) and toileting activity (β=.23; SE=0.09; P=.01).

4.4.4 Longitudinal Analyses of Location Occurrences Between 2 Outcome Categories of Patients

The occurrence of a location is equal to the number of times a patient spends more than 40 continuous seconds within that specific location. In other words, if the smartwatch receives Bluetooth low energy signal of a beacon corresponding a location for 40 seconds, the occurrence of that location increases by one unit. Figure 4.9 (a and b) shows total occurrences of patients in various nursing facility locations (daily) normalized by the number of patients in each category. Darker colors indicate higher frequency of patients visiting a particular location. In short, patients in outcome category "home" traveled within the facility (resident and therapy area) much more frequently than patients eventually admitted to a longer-term care or the "hospital" group. Additionally, no patient in the hospital category used upper body exercise (SciFit), Endorphin, and stair equipment in the therapy area.

4.5 Discussion

To our knowledge, this is the first study that has combined indoor localization and accelerometer-based physical activity recognition to assess older patients. A subset of indoor location and physical activity features were found to be highly correlated with the outcomes (community vs hospital re-admission) at baseline. In this section, we discuss the significant highlights of the result.

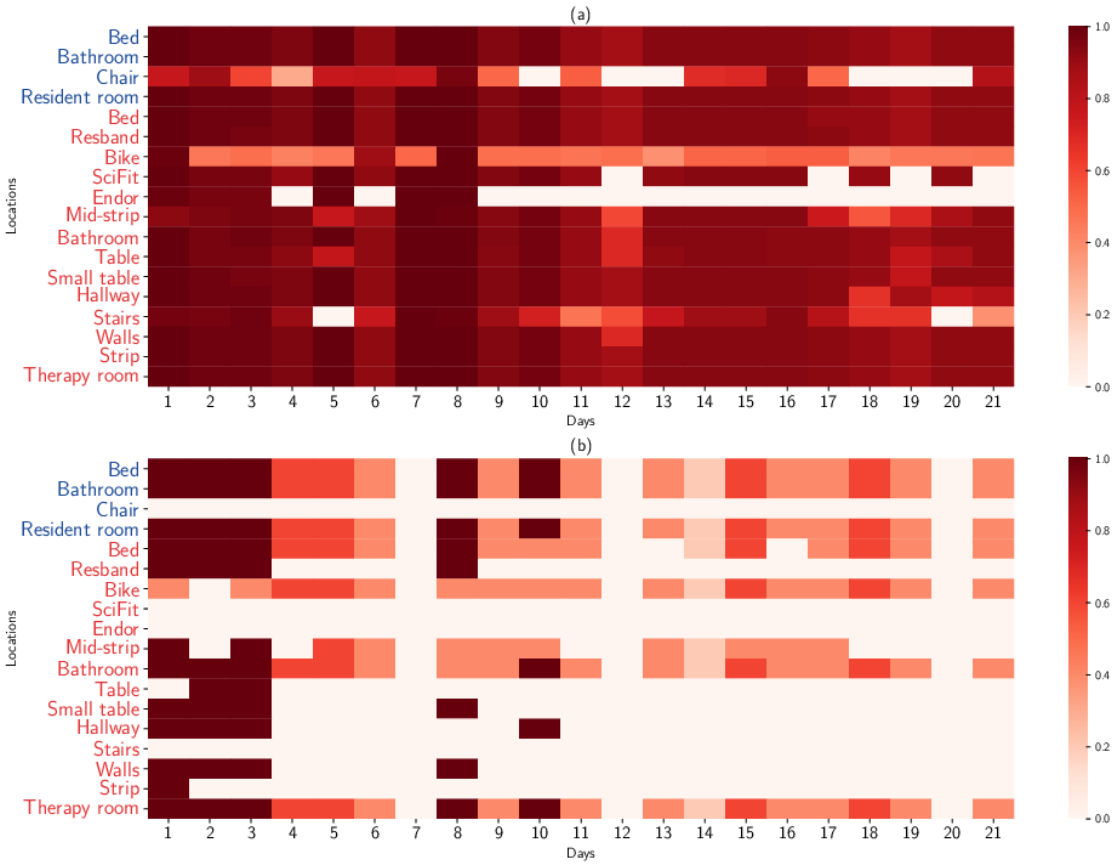


Figure 4.9: Normalized observation counts per patient by location within 21 days; (a): 105 patients in the "community" group; (b): 5 patients in the "hospital" group

4.5.1 Steps Versus Raw Acceleration Signal

Interestingly, walking, a known distinctive parameter in assessing physical functional performance in certain older populations [36], did not yield significance in this study. In populations that are frail, similar to that in subacute rehabilitation, only a negligible amount of time is spent walking (<1% of daily activity). This suggests that in these populations, steps counters may not necessarily be the best way to quantify active state [51, 52]. It would be best to prepare for the stark reality that geriatric population may not be active enough to assess their well-being or infer their independence only based on step counts or by monitoring their walking. A combination of activity features that includes both wearable sensor and stationary beacons that provide corresponding indoor localizations could be a stronger indicator of their general well-being and/or frailty. Moreover, the use of raw acceleration signals to quantify energy intensity allows us to capture even small movements, the movements that may not trigger step counters but still indicate some level of activity. Let us consider an example in which we considered energy spent rather than steps: compared with community, hospital patients show higher percentage of energy while laying down (P=.39, d=0.54). They also spent more overall time (%) in that position (60.99) compared with 46.83 for community patients. However, energy intensity of community patients is higher than hospital patients (26.23 vs 19.54). This indicates that community patients have been more active while lying down. Being more active while lying down may be the result of turning in bed; hence, this feature may denote higher ability to move in community patients. In this scenario, as discussed earlier, step counters will not produce reliable results to quantify patients' activity levels.

4.5.2 Activity With Therapist Versus Resident Time Alone

One interesting aspect of this study was to investigate the activity while a patient is with a physical therapist versus activity during the other hours of the day. It did not appear that a clear distinction could be made between different outcome groups based on therapy room energy intensity. This could be because all patients during therapy sessions are engaged by the therapist in similar physical activities following set protocols. However, the energy intensity of resident room was distinctive

within outcome groups.

4.5.3 Value of Indoor Localization Data

To assess the value of indoor localization in activity tracking, it would be best to highlight some of the scenarios: according to Table 4.2, among clinical characteristic assessment, ADL toilet (P=.007) was the most significant feature in determining the outcome. This feature corresponds to the watch-derived feature energy intensity in resident bathroom. With P=.004 and effect size of d=1.18, energy intensity in resident room (achieved from indoor localization) hence confirms the clinical finding and can be considered in the absence of ADL evaluations. In other words, ADL variant, a highly significant clinical feature, can be replicated using combination of indoor localization and activity/energy derivations.

Both group energy intensities at bed and bath were less than 60 per min. In the study by Razjouyan et al [36], authors use a cutoff point of 90 to differentiate between light and moderate-to-vigorous activities. On the basis of that, given the intensity in both bathroom and bed for either of the groups, we can conclude that patients performed light activities in those locations.

None of the patients in hospital outcome group used therapy room toilet/bathroom. It is likely that those patients were not capable enough to perform such exercises or even not advised by clinicians/nurses to do so to prevent injury. Either way, the lack of performing an activity, in this case, information extracted from indoor localization data, could be an early indication of which group a patient belongs to; it could also potentially be used to identify adverse outcomes and proactively address to prevent a negative outcome.

4.5.4 Predictive Analysis: Statistically Significant Features

P value as statistical significance or strength of evidence index has long been a subject of debate [182, 20]. It is very crucial to know that the P value is not a definite test; increasing more attributes significantly correlated with the outcome variable in a predictive model does not necessarily yield higher predictability. Although statistical significance index and its effect size provide a standard

exploratory data analysis and perhaps a good informal heuristic for choosing attributes of a prediction model, machine learning practice has more freedom from model assumptions. This study shows that the addition of significant variants did not increase predictive power and the model with only energy intensity in resident room produced the highest recall of minority class (hospital outcome) and overall AUC (0.84).

Considering only the prediction results, we can infer that location data add value to our system. It is apparent that energy intensity in resident room is the most decisive feature in predicting the outcome.

4.5.5 Activity With Therapist Versus Resident Time Alone and the Value of Indoor Localization

One of the principal findings of this study is that the energy intensity spent in therapy sessions, unlike in resident room, tend to have similar values in both outcome groups, more significantly toward the end of the rehabilitation period (Figure 4.8). Perhaps the therapists in both patient groups are encouraged to complete their therapy activities and are part of an individually designed therapeutic program that aimed to improve functional activity. Moreover, energy intensity spent in the resident room is very similar to overall energy intensity in that patients generally spend most of their time in the resident room. Resident room activity levels are likely to be crucial in determining the outcome of patients, even at early stages of their rehabilitation. Further understanding of the therapeutic skills learned during therapeutic intervention and carryover into the resident room warrants further study.

Based on Table 4.7, the PT and OT features investigated in this study all improved over time along with the sensor-based feature, energy intensity. However, improvements are more distinguishable between admission day and weeks 1 and 2. On week 3, the mean value for sensor-based features such as overall energy intensity declines. Similarly, OT and PT features show less change compared to week 1 and admission day. One possible reason could be the drop in sample size after week 2 as patients are likely to be discharged earlier. Note that despite the steady PT and

OT functional scores in later times, the interquartile range decreases over time, which indicates less variations in functional levels. This could mean that residents achieved their functional goals or plateaued in functional progression. Other aspects that limit a resident's functional ability need to be examined to determine if nonmotor parameters are limiting a resident's progress. Cognition, vision, and psychological factors are some of the areas that may limit functional progression.

Table 4.7 also shows that except the "gait distance in feet," the improvement of features was not evident after the 2^{nd} and 3^{rd} week. Further exploration of therapy treatment intensity or type of intervention is warranted. Significant improvements in "gait distance in feet" suggest the importance of this feature in clinical assessment. The rest of the gait measures showed they were less likely to change over time. Dynamic gait parameters and their relation to mobility in daily activities need more investigation.

4.5.6 Sensor-Based Features and Changes in Clinical Assessments

The captured sensor-based longitudinal changes such as lying down, sitting, and overall energy intensity reflect changes in PT and OT features (Table 4.8). This finding confirms the benefit of remote patient monitoring systems as adjunct tools to further reveal patients' daily story lines. Such systems can bear valuable information in further understanding the type and intensity of therapy interventions that impact overall functional outcome. Brisk features remained surprisingly unchanged over time when patients were expected to become less sedentary during recovery of functional abilities, at least partially. Average sedentary time among all patients was more than 99.8% and remained unchanged. In other words, the cohort was walking less than 0.2% of the time, measured objectively by the SARP wrist-worn sensor. This finding strongly suggests that focusing on sedentary features among elderly patients is beneficial, confirming the studies in [183, 184, 185], contrary to the emphasis many patient monitoring systems place on using activity trackers to count steps [52, 186]. This study shows the importance of translating all movements into measurements such as energy, or energy intensity, rather than solely relying on steps. This may shed light on the type of intervention needed for improving the mobility of the elderly resident population.

4.6 Limitations and Future Research

Activity classification can best be obtained using a series of motion sensors placed on various parts of the body. Thus, a wide range of activities can be captured as most body motions are detected. However, to simplify the activity detection, using single motion sensors is quite popular. Placing an accelerometer on the hip has been one of the most popular methods because it captures almost all human motions; however, it underestimates the arm ergometry, as it cannot fully extract the arm movements [187]. Wrist-worn accelerometers are popular because of their ease of use, water resistance in most brands, and capturing a comprehensive set of activities. However, interpreting their data for certain sedentary activities such as sitting, standing, and laying is rather challenging, in that, hand movements are very similar in those scenarios. Although ambulation detection is evident in most cases, error rates of classification increase when using assistive devices, walking in very low speed, carrying a weight with the hand that is not wearing the watch, or doing activities involving hand and feet movement together such as sweeping [42, 187, 188].

Patients' compliance with wearing a smartwatch was the main challenge of this study, and we expect it to be a generic obstacle in similar studies that aim to harness wearable technology for patients. Moreover, if the target population is less familiar with new forms of technology such as wearable devices, the compliance issue might become even more crucial. In this study, we recruited 184 patients, of which 30 patients were excluded for not satisfying the analysis inclusion criteria (watch wear time constraint). Our baseline analyses revealed that 50% of patients removed their watches before the study coordinator collects them at the end of the 8 hours.

Dealing with medical datasets is rather challenging in that the datasets predominantly consist of normal cases in addition to minority abnormal instances that deem to be more interesting [189]. Many attempts have been made to overcome the obstacle of the normal and abnormal samples known as imbalanced datasets. There exist approaches to improve the performance of predictive models by oversampling and/or undersampling the dominant and abnormal instances [59, 15, 65]. In our study cohort, the 2 outcome categories are not equally represented, making the dataset imbalanced. In the future, we aim to further investigate the use of oversampling and undersampling

of our dataset as methods that perhaps are not very conventional in the medical field but can possibly improve the predictability of our models.

4.7 Conclusions

Despite the evolution of eHealth and mobile health (mHealth) and the emerging role of wearable and mobile technology in new platforms of health care, there are anecdotal claims that wearable technology may not precisely quantify patients' health [190]. In this study, we showed that wearable technology, equipped with refined physical activity tracking algorithms, in our case, tailored for geriatrics, can result in a better understanding of patients and hopefully pave the way in developing intervention alerts and approaches. We discussed how SARP features provide a clearer storyline of daily activity patterns by merging indoor localization with physical activities. The SARP system can be incorporated into mHealth technology platforms and can provide a more objective assessment of the frail population.

CHAPTER 5

Imbalanced learning in healthcare analytics

5.1 Introduction

Learning from imbalanced datasets can be very challenging as the classes are not equally represented in the datasets [53]. There might not be enough examples for a learner to form a hypothesis that can well model the under-represented classes. Hence, the classification results are often biased towards the majority classes. The curse of imbalanced learning is prevalent in real-world applications as discussed in Section 2.2.

To address this, we propose WOTBoost, a method that combines Weighted Oversampling Technique and ensemble Boosting. WOTBoost synthesizes minority data to balance the dataset and identifies difficult minority data. By generating more synthesized data near difficult-to-learn minority samples, WOTBoost can potentially improve classification accuracy without compromising the majority class. Even though class imbalance issue can exist in multi-class applications, we only focus on the binary class scenario in this paper as it is feasible to reduce a multi-class classification problem into a series of binary classification problems [60].

The contributions in this paper are as follows:

- We identify the minority class data examples which are harder to learn at each round of boosting and generate more synthetic data for this kind.
- We test our proposed algorithm extensively on 18 public accessible datasets and compared the results with the most commonly used algorithms. To our knowledge, this might be first work to carry out such a comprehensive comparison study in ensemble method combined

with oversampling approach.

- We inspect the various distributions of 18 datasets and discussed why WOTBoost performs better on certain datasets.

5.2 WOTBoost: Weighted Oversampling Technique in Boosting

In this section, we propose the WOTBoost algorithm which combines a weighted oversampling algorithm with the standard boosting procedure. The Weighted Oversampling Technique populates synthetic data based on the weights that are associated to each minority data. In other words, higher weighted minority data samples are synthesized more. This algorithm is an ensemble method and creates a series of classifiers in an arbitrary number of iterations. The boosting procedure will be elaborated with details in Algorithm 1 and 2: a) We introduce a weighted oversampling step at the beginning of each iteration of boosting. b) We adjust the weighted oversampling strategy using the updated weights (i.e., D_t at line 8 in Algorithm 1) associated with the minority during each round of boosting [53]. The boosting algorithm gives more weights to the data samples which were misclassified in the previous round. Hence, WOTBoost can be designed to generate more synthetic data examples for the minority data which were misclassified in the previous iterations. Meanwhile, boosting technique would also add more weights to misclassified majority class data, and force the learner to focus on these data as well. Therefore, we combine the merits of weighted oversampling technique and AdaBoost.M2 together. The goal is to improve the discriminative power of the classifier on difficult minority examples without sacrificing the accuracy of the majority class data instances.

Algorithm 1 presents the details of the boosting procedure, which is a modified version of AdaBoost.M2 [79]. It takes a training dataset D_{Tr} with m data samples, $D_{Tr} = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$. x_i is the *ith* feature vector in n-dimensional space, and $y_i \in Y = \{majority, minority\}$ is the true label associated with x_i. $\hat{y}_i$ is the predicted label. We initialize a mislabel distribution, B, which contains all the misclassified data instances (i.e., $\hat{y}_i \neq y_i$). In addition, we also initialize a

Algorithm 1: Boosting with weighted oversampling

Input: Training dataset D_{Tr} with m samples $\{x_i, y_i\}, i = 1, 2, ..., m$, where x_i is an instance in the n dimensional feature space, X, and y_i $\in Y = \{majority, minority\}$ is the label associated with x_i;

Let $B = \{(i, \hat{y}_i) : i = 1, ..., m, \hat{y}_i \neq y_i\}$;

T specifies the number iterations in boosting procedure;

Initialize: $D_1(i, \hat{y}_i) = \frac{1}{m}$, $i = 1, 2, ..., m$

for t=1,2,3,... T **do**

Create N synthetic examples from minority class with the weight distribution D_t using **Algorithm 2**;

Fit a weak learner using the temporary training dataset which is a combination of original data and synthetic data;

Calculate a weak hypothesis $h_t : X \times Y \rightarrow [0, 1]$;

Compute the pseudo-loss of h_t: $\varepsilon_t = \frac{1}{2} \sum_{(i,\hat{y}_i) \in B} D_t(i, \hat{y}_i)(1 - h_t(x_i, y_i) + h_t(x_i, \hat{y}_i))$;

Let $\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$;

Update the weight distribution D_{t+1}: $\tilde{D}_{t+1}(i, \hat{y}_i) = D_t(i, \hat{y}_i)\beta_t^{\frac{1}{2} \times (1 - h_t(x_i, y_i) + h_t(x_i, \hat{y}_i))}$;

Normalize D_{t+1} : $D_{t+1}(i, \hat{y}_i) = \frac{\tilde{D}_{t+1}(i, \hat{y}_i)}{Z_t}$, where Z_t is a normalization constant such that $\sum_{i \in m} D_{t+1}(i, \hat{y}_i) = 1$

end

Output: $h_{final}(x) = \text{argmax}_{\hat{y}_i \in Y} \sum_{t=1}^{T} (log \frac{1}{\beta_t}) h_t(x, \hat{y}_i)$

weight distribution for the training data by assigning equal weights over all samples. During each round of boosting (step 1 - step 9), a weak learner is built on a training dataset which is the output of a weight oversampling procedure. The weak learner formulates a weak hypothesis which is just slightly better than random guessing, hence the name [79]. But this is good enough as the final output will aggregate all the weak hypotheses using weighted voting. As for error estimation, the pseudo loss of a weak hypothesis is calculated as specified at step 5. Instead of using ordinary training loss, pseudo loss is adopted to force the ensemble method to focus on mislabeled data. More

Algorithm 2: Dynamic weighted oversampling procedure

Input: N is the number of synthetic data examples from minority class;

D_t is the weight distribution passed at *line 2* in **Algorithm 1**

Calculate the number of synthetic data examples for each minority class instance:

$g_i = N \times \frac{D_t(i,\hat{y_i})}{\sum\limits_{j \in minority} D_t(j,\hat{y_i})}$;

For each minority class instance, $x(i)$, in original training dataset, generate g_i synthetic data examples using the following rules:

for 1,2,3,...,g_i **do**:

(I) Randomly choose a minority class example, $x_{nn}(i)$, from the k nearest neighbors of $x(i)$, which is a n-dimensional feature vector.

(II) Calculate the difference vector $\delta = x_{nn}(i) - x(i)$.

(III) Create a synthetic data example using the following equation:

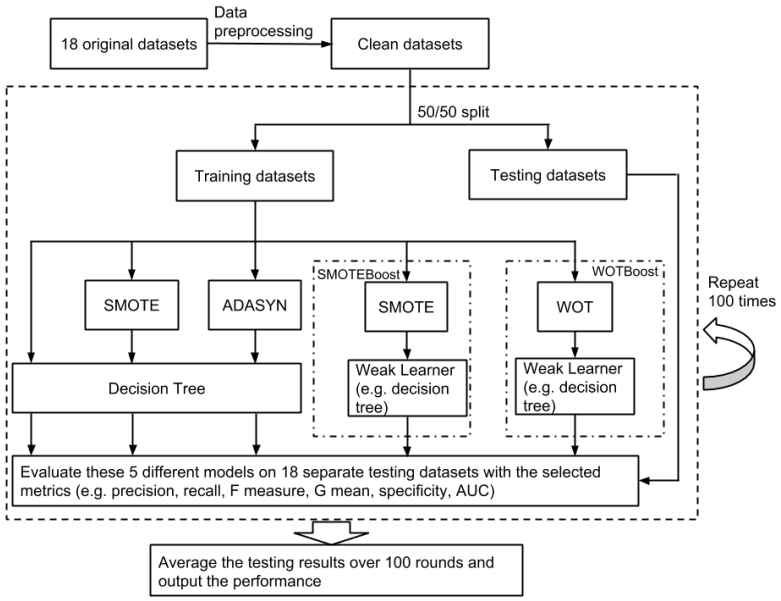
$$x_{syn}(i) = x(i) + \delta \times \lambda$$

where $\lambda \in [0, 1]$. **Output:** A temporary training dataset combining the original data with synthetic data

justification for using pseudo loss can be found in [79, 191]. Once the pseudo loss is computed, the weight distribution, D_t, is updated accordingly and normalized at step 5 - step 8.

Algorithm 2 demonstrates the weighted oversampling procedure. The inputs to oversampling technique are the weight distribution, D_t, and an arbitrary number of synthetic data samples, N. It uses the weight distribution as the oversampling strategy to decide how to synthesize for each minority data samples, as it is described at step 1 in Algorithm 2. As mentioned previously, the ensemble method would assign more weights to misclassified data. Therefore, this oversampling strategy facilitates the classifier to learn a broader representation of mislabeled data by placing more similar data samples around them.

Figure 5.1: Overview of the comparison study



5.3 Experimentation

In this section, we conduct a comprehensive comparison study of WOTBoost algorithm with decision tree, SMOTE + decision tree, ADASYN + decision tree, and SMOTEBoost. Figure 5.1 shows how the models are built and assessed.

5.3.1 Dataset overview

We evaluate these 5 models extensively using 18 imbalanced datasets which are publicly accessible. The imbalanced ratio (i.e., counts of majority class samples to counts of minority class samples) of these datasets vary from 1.7 to 42. Since some testing imbalanced datasets have more than 2 classes, and we are only interested in the binary class problem in this paper, we pre-processed these datasets and modified them into a binary class datasets following the rules in the literature [59, 63, 66, 65, 80, 83]. Meanwhile, only numeric attributes are included when processing datasets. The details of data cleaning can be referred to the prior works [59, 80, 65]. The characteristics of these datasets are summarized in Table 5.1.

Table 5.1: Characteristics of 18 testing datasets

Dataset	Instances	Attributes	Outcome Frequency	Imbalanced Ratio	No. of safe minority	No. of unsafe minority	unsafe minority%
Pima Indian Diabetes [192]	768	9	Maj: 506 Min:268	1.9	86	182	67.9%
Abalone [193]	4177	8	Maj:689 Min:42	6.4	5	37	88.1%
Vowel Recognition [193]	990	14	Maj:900 Min:90	10.0	89	1	1.1%
Mammography [194]	11183	7	Maj: 10923 Min: 260	42	107	153	58.8%
Ionosphere [193]	351	35	Maj: 225 Min 126	1.8	57	69	54.8%
Vehicle [193]	846	19	Maj: 647 Min:199	3.3	154	45	22.6%
Phoneme [195]	5404	6	Ma j: 3818 Min:1580	2.4	980	606	38.2%
Haberman [193]	306	4	Maj: 225 Min:81	2.8	8	73	90.1%
Wisconsin [193]	569	31	Maj: 357 Min: 212	1.7	175	37	17.5%
Blood Transfusion [196]	748	5	Maj: 570 Min: 178	3.2	23	83	87.1%
PC1 [197]	1484	9	Maj: 1032 Min: 77	13.4	8	69	89.6%
Heart [193]	294	14	Maj: 188 Min: 106	1.8	17	89	84.0%
Segment [193]	2310	20	Ma j: 1980 Min: 330	6.0	246	84	25.5%
Yeast [193]	1484	9	Ma j: 1240 Min: 244	5.1	95	149	61.1%
Oil	937	50	Maj: 896 Min: 41	21.9	0	41	100.0%
Adult [193]	48842	7	Maj: 37155 Min: 11687	3.2	873	10814	92.5%
Satimage [193]	6430	37	Maj: 5805 Min: 625	9.3	328	297	47.5%
Forest cover [198]	581012	11	Maj: 35754 Min: 2747	13.0	2079	668	24.3%

5.3.2 Experiment setup

We compare the WOTBoost algorithm with naive decision tree classifier, decision tree classifier after SMOTE, decision tree classifier after ADASYN, and SMOTEBoost. Figure 5.1 shows that the clean datasets are split evenly into training and testing during each iteration [65]. As a control group, a naive decision tree model learned directly from the imbalanced training dataset. SMOTE and ADASYN algorithms are used separately to balance the training dataset before inputting it to decision tree classifiers. SMOTEBoost and WOTBoost take in imbalanced training datasets and synthesize new data samples for the minority at each round of boosting. Both of them use decision tree as the weak learner [80]. Models are evaluated on a separate testing dataset. The evaluating metrics used in this study are precision, recall, F1 measure, G mean, specificity, area under ROC. The final performance assessments are averaged over 100 such runs, and they are summarized in Table 5.3. During each testing run, we oversample the training dataset in a way that both minority class and majority class are equally represented in all models [65]. For SMOTE, ADASYN, SMOTEBoost, and WOTBoost, we set the number of nearest neighbors to be 5.

5.3.3 Metrics

Overall accuracy is typically chosen to evaluate the predictive power of machine learning classifiers provided with a balanced dataset. As for imbalanced datasets, overall accuracy is no longer an effective metric. For example, in the information retrieval and filtering domain by Lewis and Catlette (1994), only 0.2% are interesting cases [56]. A dummy classifier that always gives predictions of majority class would easily achieve an overall accuracy of 99.8%. However, this predictive model is uninformative as we are more interested in classifying the minority class. Common alternatives to overall accuracy in assessing imbalanced learning models are F measures, G mean, and Area Under the Curve (AUC) for Receiver Operating Characteristic (ROC) [199]. By convention, majority class is regarded as negative class and minority class as positive class [59, 200]. Table II shows a confusion matrix that is typically used to visualize and assess the performance of predictive models. Based on this confusion matrix, the evaluation metrics used in this paper are mathematically formulated as follows:

Table 5.2: Confusion matrix of a binary classification problem

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

$$F1\ measure = 2 \cdot \frac{precision \times recall}{precision + recall} \tag{5.3}$$

$$
\begin{aligned}
G\,mean &= \sqrt{Positive\ Accuracy \times Negative\ Accuracy} \\
&= \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}
\end{aligned}
\tag{5.4}
$$

Table 5.3: Evaluation metrics and performance comparison

Dataset	Methods[a]	OA[b]	Precision[b]	Recall[b]	F_measure[b]	G_mean[b]	ROC AUC[b]
Pima Indian Diabetes	DT	0.71 ± 0.02	**0.61 ± 0.04**	0.54 ± 0.05	0.57 ± 0.03	0.66 ± 0.02	0.67±0.02
	S	0.67±0.02	0.55±0.03	0.54±0.04	0.54±0.02	0.63±0.02	0.64±0.02
	A	0.68 ± 0.02	0.56±0.04	0.58±0.05	0.57±0.03	0.66±0.03	0.66±0.02
	SM	0.66±0.02	0.52±0.02	**0.86±0.04**	0.64±0.02	0.68±0.02	0.70±0.01
	WOT	**0.73±0.02**	0.60±0.03	0.78±0.05	**0.68±0.02**	**0.74±0.02**	**0.74±0.02**
Abalone	DT	0.93 ± 0.01	0.46±0.12	**0.46±0.10**	**0.46±0.08**	**0.66±0.08**	**0.71±0.04**
	S	0.88±0.02	0.24±0.07	0.38±0.11	0.29±0.07	0.59±0.08	0.65±0.05
	A	0.88±0.02	0.24±0.06	0.42±0.11	0.31±0.07	0.62±0.09	0.66±0.05
	SM	0.84±0.06	0.19±0.04	**0.46±0.12**	0.27±0.05	0.63±0.05	0.66±0.05
	WOT	**0.94±0.01**	**0.55±0.33**	0.34±0.11	0.42±0.13	0.58± 0.18	0.66 ±0.05
Vowel Recognition	DT	0.97±0.00	**0.90±0.06**	0.79±0.06	0.84±0.04	0.88±0.03	0.89±0.03
	S	0.96±0.00	0.85±0.06	0.74±0.06	0.80±0.04	0.86±0.03	0.87±0.03
	A	0.97±0.00	0.88±0.05	0.79±0.07	0.83±0.04	0.88±0.03	0.89±0.03
	SM	**0.98±0.00**	0.83±0.05	0.96±0.04	0.89±0.03	0.97±0.02	0.97±0.02
	WOT	**0.98±0.01**	0.87±0.10	**0.98±0.01**	**0.93±0.07**	**0.98±0.02**	**0.98±0.01**
Ionosphere	DT	0.86±0.02	0.83±0.06	0.73±0.06	0.77±0.04	0.82±0.03	0.83±0.03
	S	0.85±0.03	0.75±0.05	0.81±0.06	0.78±0.04	0.84±0.03	0.84±0.03
	A	0.88±0.03	0.84±0.05	0.80±0.06	0.82±0.04	0.86±0.03	0.86±0.03
	SM	**0.91±0.02**	0.89±0.06	**0.85±0.04**	**0.87±0.03**	**0.90±0.02**	**0.90±0.02**
	WOT	**0.91±0.02**	**0.92±0.05**	0.79±0.04	0.85±0.03	0.87±0.02	0.88±0.02
Vehicle	DT	0.94±0.01	**0.85±0.04**	0.88±0.04	0.87±0.03	0.92±0.02	0.92±0.02
	S	0.90±0.01	0.75±0.04	0.88±0.05	0.81±0.03	0.89±0.02	0.89±0.02
	A	0.92±0.01	0.81±0.04	0.87±0.04	0.84±0.02	0.90±0.02	0.90±0.02
	SM	**0.95±0.00**	0.84±0.03	**0.97±0.02**	**0.90±0.02**	**0.96±0.01**	**0.96±0.01**
	WOT	0.89±0.10	0.70±0.15	**0.97±0.03**	0.81±0.11	0.92±0.07	0.92±0.06
Phoneme	DT	**0.86±0.00**	**0.75±0.01**	0.74±0.01	0.75±0.01	0.82±0.00	0.82±0.00

Table 5.3 continued from previous page

Dataset	Methods[a]	OA[b]	Precision[b]	Recall[b]	F_measure[b]	G_mean[b]	ROC AUC[b]
	S	**0.86±0.00**	0.74±0.01	0.78±0.01	**0.76±0.01**	**0.83±0.01**	**0.83±0.00**
	A	0.83±0.00	0.68±0.01	0.78±0.01	0.73±0.01	0.82±0.00	0.82±0.00
	SM	0.77±0.00	0.57±0.01	0.86±0.01	0.69±0.01	0.80±0.00	0.80±0.00
	WOT	0.52±0.06	0.38±0.03	**0.99±0.01**	0.54±0.03	0.57±0.07	0.66±0.04
Haberman	DT	**0.67±0.03**	0.38±0.06	0.25±0.08	0.30±0.05	0.46±0.05	0.54±0.03
	S	0.65±0.03	**0.40±0.05**	0.39±0.08	0.39±0.05	**0.64±0.04**	0.57±0.03
	A	0.60±0.03	0.37±0.05	0.52±0.08	0.43±0.05	0.58±0.05	0.58±0.04
	SM	0.48±0.06	0.34±0.03	**0.84±0.07**	**0.48±0.03**	0.53±0.10	**0.59±0.02**
	WOT	0.54±0.05	0.35±0.07	0.70±0.12	0.47±0.05	0.57±0.05	**0.59±0.03**
Wisconsin	DT	0.95±0.01	0.93±0.03	0.93±0.01	0.93±0.01	0.95±0.01	0.95±0.01
	S	0.92±0.01	0.89±0.03	0.90 ±0.03	0.89±0.02	0.91±0.01	0.91±0.01
	A	0.95±0.01	0.93±0.03	0.94±0.03	0.94±0.02	0.95±0.01	0.95±0.01
	SM	**0.98±0.01**	**0.99±0.00**	**0.95±0.01**	**0.97±0.01**	**0.97±0.01**	**0.97±0.01**
	WOT	0.97±0.01	0.97±0.03	**0.95±0.02**	0.96±0.02	0.96±0.01	0.96±0.01
Blood Transfusion	DT	**0.72±0.01**	**0.39±0.06**	0.28±0.08	0.32±0.06	0.49±0.07	0.57±0.04
	S	0.71±0.01	**0.39±0.05**	0.39±0.07	0.39±0.05	0.56±0.05	0.60±0.03
	A	0.70±0.01	0.38±0.05	0.42±0.08	0.40±0.06	0.57±0.07	0.60±0.04
	SM	0.44±0.03	0.29±0.03	**0.93±0.10**	**0.45±0.03**	0.52±0.04	0.61±0.04
	WOT	0.68±0.03	0.38±0.16	0.52±0.12	0.44 ±0.09	**0.61±0.14**	**0.62±0.05**
PC1	DT	0.90±0.01	0.25±0.05	0.27±0.05	0.26±0.04	0.50±0.04	0.61±0.02
	S	0.87±0.02	0.22±0.04	0.38±0.05	0.27±0.04	0.58±0.03	0.64±0.02
	A	0.87±0.02	0.26±0.04	**0.51±0.06**	**0.35±0.03**	**0.68±0.03**	**0.71±0.02**
	SM	0.82±0.05	0.16±0.02	0.41±0.04	0.23±0.03	0.59±0.07	0.63±0.02
	WOT	**0.91±0.03**	**0.34±0.03**	0.30±0.07	0.32±0.03	0.53±0.02	0.63±0.02
Heart	DT	0.77±0.03	**0.68±0.06**	0.63±0.08	0.65±0.05	0.73±0.04	0.74±0.03
	S	0.76±0.03	0.67±0.06	0.57±0.07	0.62±0.04	0.70±0.04	0.71±0.03
	A	**0.79±0.03**	**0.68±0.05**	**0.75±0.06**	**0.71±0.04**	**0.78±0.03**	**0.78±0.03**
	SM	0.70±0.03	0.55±0.05	**0.75±0.06**	0.63±0.03	0.71±0.03	0.71±0.03
	WOT	0.74±0.03	0.60±0.06	0.74±0.06	0.66±0.04	0.74±0.03	0.74±0.03
Segment	DT	**0.96±0.00**	**0.88±0.04**	0.88±0.03	**0.88±0.02**	**0.93±0.02**	**0.93±0.01**
	S	**0.96±0.00**	0.87±0.03	0.85±0.03	0.86±0.02	0.91±0.01	0.91±0.01
	A	**0.96±0.00**	**0.88±0.03**	0.87±0.03	0.87±0.02	0.92±0.01	0.92±0.01

Table 5.3 continued from previous page

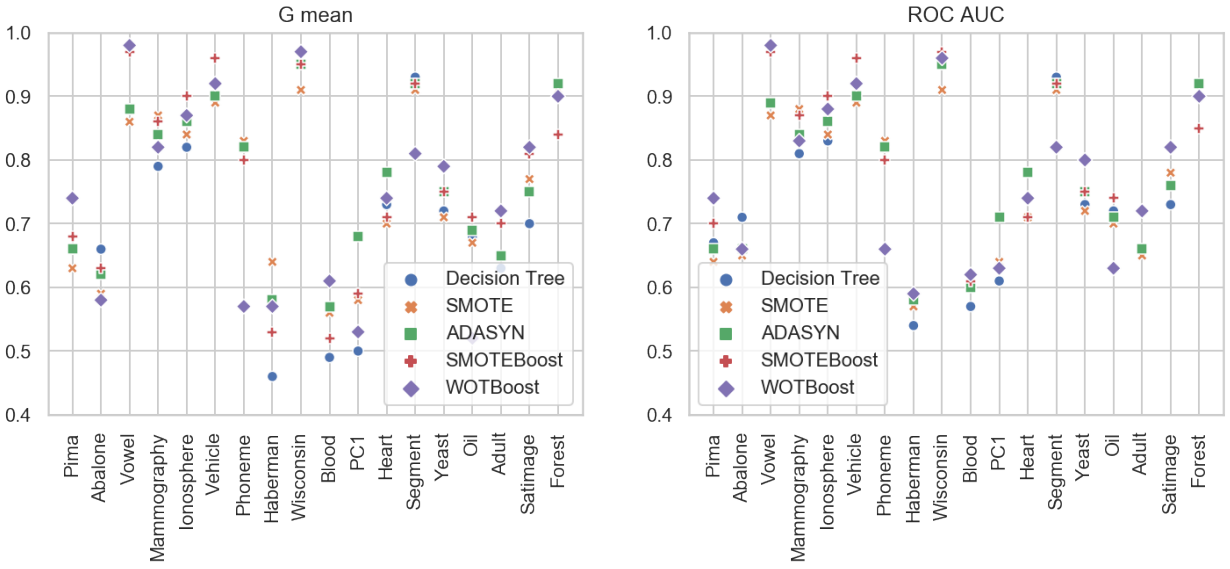
Dataset	Methods[a]	OA[b]	Precision[b]	Recall[b]	F_measure[b]	G_mean[b]	ROC AUC[b]
	SM	0.95±0.00	0.80±0.03	0.87±0.02	0.83±0.02	0.92±0.01	0.92±0.01
	WOT	0.72±0.09	0.34±0.08	**0.97±0.07**	0.51±0.08	0.81±0.06	0.82±0.05
	DT	0.83±0.01	0.46±0.03	0.59±0.05	0.51±0.03	0.72±0.03	0.73±0.02
	S	0.81±0.01	0.41±0.04	0.60±0.04	0.49±0.03	0.71±0.02	0.72±0.02
Yeast	A	0.82±0.01	0.43±0.03	0.66±0.05	0.52±0.02	0.75±0.03	0.75±0.02
	SM	0.70±0.02	0.32±0.02	**0.82±0.03**	0.46±0.02	0.75±0.01	0.75±0.01
	WOT	**0.84±0.02**	**0.50±0.05**	0.73±0.05	**0.59±0.03**	**0.79±0.02**	**0.80±0.02**
	DT	0.93±0.01	0.35±0.11	0.48±0.13	0.41±0.11	0.68±0.12	0.72±0.06
	S	0.91±0.01	0.26±0.07	0.48±0.11	0.33±0.07	0.67±0.08	0.70±0.05
Oil	A	0.89±0.01	0.22±0.09	**0.52±0.11**	0.31±0.08	0.69±0.09	0.71±0.05
	SM	0.94±0.01	0.41±0.07	**0.52±0.10**	**0.46±0.06**	**0.71±0.06**	**0.74±0.05**
	WOT	**0.95±0.02**	**0.47±0.13**	0.28±0.16	0.35±0.09	0.52±0.12	0.63±0.07
	DT	0.75±0.00	0.48±0.00	0.47±0.00	0.48±0.00	0.63±0.00	0.66±0.00
	S	0.70±0.00	0.41±0.00	0.56±0.00	0.48±0.00	0.65±0.00	0.65±0.00
Adult	A	0.71±0.00	0.42±0.00	0.57±0.00	0.48±0.00	0.65±0.00	0.66±0.00
	SM	**0.81±0.00**	**0.62±0.01**	0.55±0.01	**0.58±0.00**	0.70±0.01	**0.72±0.00**
	WOT	0.75±0.02	0.48±0.03	**0.67±0.05**	0.56±0.01	**0.72±0.01**	**0.72±0.01**
	DT	**0.91 ± 0.00**	**0.53±0.02**	0.51±0.03	0.52±0.02	0.70±0.02	0.73±0.01
	S	0.90±0.00	0.51±0.02	0.63±0.02	0.56±0.01	0.77±0.01	0.78±0.01
Satimage	A	0.89±0.00	0.45±0.03	0.60±0.03	0.52±0.02	0.75±0.01	0.76±0.01
	SM	0.90±0.00	0.49±0.02	0.72±0.02	**0.58±0.01**	0.81±0.00	**0.82±0.01**
	WOT	0.88±0.01	0.42±0.03	**0.75±0.03**	0.54±0.02	**0.82±0.01**	**0.82±0.01**
	DT	**0.97±0.00**	**0.81±0.01**	0.82±0.01	**0.82±0.01**	0.90±0.00	0.90±0.00
	S	**0.97±0.00**	0.78±0.01	0.85±0.01	0.81±0.00	0.91±0.00	**0.92 ±0.00**
Forest cover	A	**0.97±0.00**	0.79±0.01	0.86±0.01	**0.82±0.01**	**0.92±0.00**	**0.92±0.00**
	SM	0.96±0.00	0.73±0.01	0.72±0.01	0.72±0.00	0.84±0.00	0.85±0.00
	WOT	0.91±0.02	0.43±0.05	**0.88±0.02**	0.58±0.05	0.90±0.01	0.90±0.01

[a] DT=Decision Tree, S=SMOTE, A=ADASYN, SM=SMOTEBoost, WOT=WOTBoost.

[b] Values are rounded to 2 decimal places

We highlight the best model and its performance in boldface for each dataset in Table 5.3. Figure 5.2 presents the performance comparison of these 5 models on G mean and AUC score in 18 datasets. To assess the effectiveness of the proposed algorithm on these imbalanced datasets, we count the cases when WOTBoost algorithm outperforms or matches other models on each metric. The results presented in Table 5.4 show that WOTBoost algorithm has the most winning times on G mean (6 times) and AUC (7 times). As defined in Equation 5.4 in the metric section, G mean is the square root of the product between positive accuracy (i.e., recall or sensitivity) and negative accuracy (i.e., specificity). Meanwhile, area under the ROC curve, or AUC, is typically used for model selection, and it examines the true positive rate and false positive rate at various thresholds. Hence, both evaluation metrics consider the accuracy of both classes. Therefore, we argue that WOTBoost indeed improves the learning on the minority class while keeping the accuracy of the majority class.

Figure 5.2: Performance comparison of G mean and AUC score on 18 datasets



In Table 5.3, we observe that WOTBoost has the best G mean and AUC score on Pima Indian Diabetes whereas SMOTEBoost is the winner on Ionoshphere with the same assessments. Considering these two datasets have similar global imbalanced ratio, it naturally raises the question: are there any other factors that are influential in the classification performance? To understand the reasons why WOTBoost performs better on certain datasets, we investigate the local characteristics

Table 5.4: Summary of effectiveness of WOTBoost algorithm on 18 datasets

Winning counts	Precision	Recall	F_measure	G_mean	AUC
Decision Tree	9	0	2	2	2
SMOTE	2	1	1	3	3
ADASYN	2	3	3	3	3
SMOTEBoost	3	10	8	4	6
WOTBoost	6	8	4	6	7

of the minority class in these datasets. We use t-SNE to visualize the distribution of these two datasets as shown in Figure 5.3. t-SNE algorithm allows us to visualize high dimensional datasets by projecting it into a two-dimensional panel. Figure 5.3 indicates there are more overlapping between two classes in Pima Indian Diabetes, whereas more "safe" minority class samples in Ionosphere. It is likely that WOTBoost is able to learn better when there are more difficult minority data examples. Figure 5.4 demonstrates the distribution of Pima Indian Diabetes before and after applying WOTBoost. We highlight one of the regions where minority data samples are difficult to learn. WOTBoost algorithm is able to populate synthetic data for these minority data samples.

Figure 5.3: (a) Distribution of Pima Indian Diabetes dataset. (b) Distribution of Ionosphere dataset

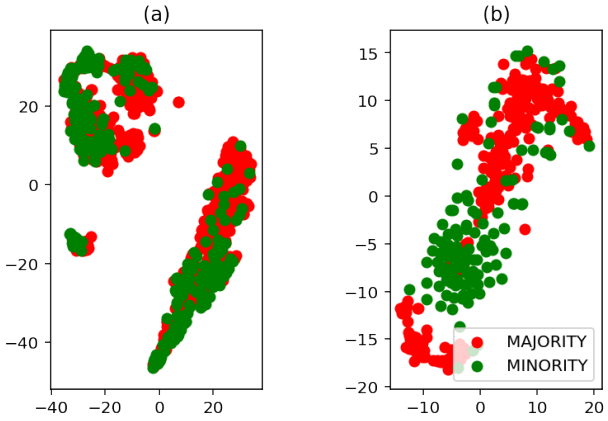
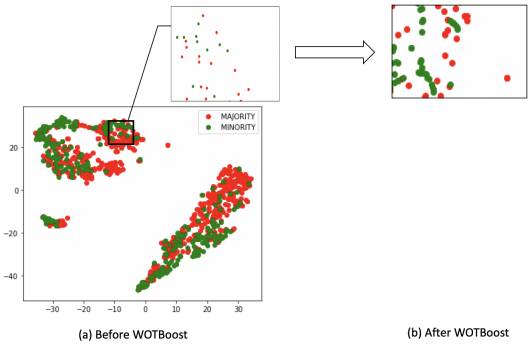


Table 5.1 shows the number of safe/unsafe minority samples of 18 dataset. We consider a minority class sample to be safe if its 5 nearest neighbors contain at most 1 majority class sample; Otherwise, it is labeled as an unsafe minority [63, 64]. Unsafe minority percentage is computed by

$$unsafe\ minority\% = \frac{counts\ of\ unsafe\ minority}{counts\ of\ minority}$$

Figure 5.4: Pima Indian Diabetes distribution before and after applying WOTBoost



We observe that the unsafe minority percentages are around 50% or higher in most of the datasets where WOTBoost has the best G-mean or AUC shown in Table 5.3. For example, Adult, Haberman, Blood Transfusion, Pima Indian Diabetes, and Satimage have 92.5%, 90.1%, 87.1%, 67.9%, 47.5% unsafe minority among the total minority class samples, respectively. Meanwhile, the global imbalanced ratios of these datasets are from 1.9 to 10.0. Hence, WOTBoost might be a good candidate to tackle imbalanced datasets with large proportion of unsafe minority samples and relatively high between-class imbalance ratios.

5.4 Conclusion

In this paper, we propose the WOTBoost algorithm to better learn from imbalanced datasets. The goal is to improve the performance of classification on minority class without sacrificing the accuracy of the majority class. We carry out a comprehensive comparison between WOTBoost algorithm and 4 other classification models. Results indicate that WOTBoost has the best G mean and AUC scores in 6 out of 18 datasets. WOTBoost shows more balanced performance, such as in G mean, than other classification models compared to particularly SMOTEBoost. Even though WOTBoost is not a cure-all method to the imbalanced learning problem, it is likely to produce promising results for datasets that contain a large portion of unsafe minority samples and maybe relatively high global

imbalanced ratios. We hope that our contribution to this research domain would provide more insights and directions.

In addition, our study demonstrates that having the prior knowledge of the minority class distribution could facilitate the learning performance of the classifiers [53, 82, 65, 64, 63, 66]. Further investigating on the data-driven sampling may produce interesting findings in this domain.

CHAPTER 6

Electrocardiogram heartbeat classification using deep transfer learning with Convolutional Neural Network and STFT technique

6.1 Introduction

Electrocardiogram (ECG) is a simple non-invasive measure to identify heart-related issues such as irregular heartbeats known as arrhythmias. While artificial intelligence and machine learning is being utilized in a wide range of healthcare related applications and datasets, many arrhythmia classifiers using deep learning methods have been proposed in recent years. Deep learning methods generally require a large amount of training data. While well-annotated ECG datasets for arrhythmia detection are limited [91], resorting to transfer learning techniques in which a pre-trained image classifier is used can be warranted. There have been recent attempts in using transfer learning framework with MIT-BIH dataset to develop arrhythmia diagnosis models [84].

In this paper, we propose a deep transfer learning framework that is aimed to perform classification on a small size training dataset. The proposed method is to fine-tune a general-purpose image classifier ResNet-18 with MIT-BIH arrhythmia dataset in accordance with the AAMI EC57 standard. This paper further investigates many existing deep learning models that have failed to avoid data leakage against AAMI recommendations. We compare how different data split methods impact the model performance. This comparison study implies that future work in arrhythmia classification should follow the AAMI EC57 standard when using any including MIT-BIH arrhythmia dataset.

The main contributions of this paper are summarized as follows:

- Propose an end-to-end ECG classification framework that can leverage the learning power of existing pre-trained 2D CNN models.
- Demonstrate how the choice of samples in MIT-BIH dataset significantly impacts the deep learning model performance.
- Highlight the unreliable and biased model evaluation in current literature on ECG classification with deep learning methods.

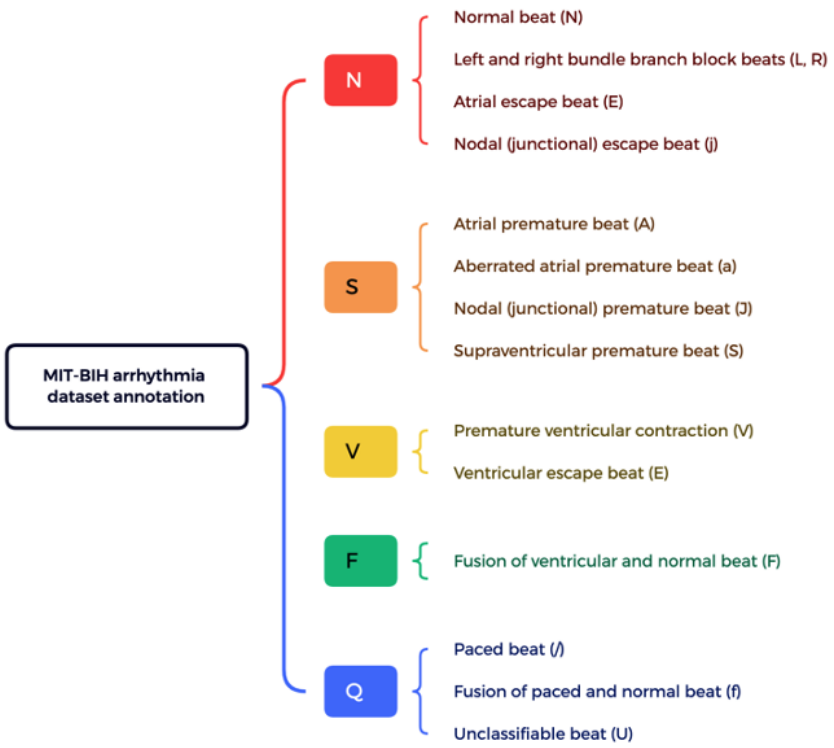


Figure 6.1: Heartbeat annotations in MIT-BIH dataset according to AAMI EC 57. The consolidated classes are N, S, V, F, Q.

6.2 Dataset

Similar to the majority of the arrhythmia analysis studies, this study develops the model on MIT-BIH Arrhythmia dataset [89]. The dataset includes 48 half-hour excerpts of two-channel ambulatory

ECG recordings collected from 47 patients at 360 Hz. The dataset was annotated at heartbeat level by two or more cardiologists independently. 14 original heartbeat types are consolidated into 5 groups according to AAMI recommendation, shown in Figure 6.1.

Set	**Heartbeat types (AAMI EC57 standard)**					
	N	**S**	**V**	**F**	**Q**	**Total**
Full MIT-BIH set	90,631	2,781	7,236	803	8,043	109,494
Intra-patient split						
Training (80% split)	72,471	2,223	5,789	642	6,431	87,756
Testing (20% split)	18,118	556	1,447	161	1,608	21,890
Inter-patient split						
Training (DS1 in [100])	45,866	944	3,788	415	8	51,021
Testing (DS2 in [100])	44,259	1,837	3,221	338	7	49,712

Table 6.1: Heartbeat distribution by classes of the raw data, intra-patient split, and inter-patient

There are two ways to split the dataset into training and testing sets, inter-patient paradigm versus intra-patient paradigm. The intra-patient paradigm creates the training/testing dataset by randomly choosing heartbeat samples. In this paradigm, the heart- beat samples from the same patient might exist in both training and testing dataset. Therefore, the testing data might influence the model training. We argue that this data split paradigm can result in unreliable results. Models developed under intra-patient split paradigm should be reconsidered and re-evaluated for clinical decision making; overall, intra-patient approach should be highly discouraged [87, 100]. On the other hand, in inter-patient paradigm the training/testing datasets are created from different patients [100, 201]. Hence, inter-patient split avoids the information leakage issue existed in its counterpart method. In the inter-patient split, the MIT-BIH dataset is divided into two datasets (DS1 and DS2),

identified by the patient IDs: DS1 = 101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, 230 and DS2 = 100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, 234 proposed by Chazal et al. [100]. DS1 is used for model training (training set) and DS2 is used for model evaluation (testing set). Patient 102, 104, 107, 217 are excluded in inter-patient split.

6.3 Methodology

In this section, we describe the proposed work in two steps: 1) building an end-to-end transfer learning framework with STFT and ResNet18, and 2) investigating how the inter-patient split and intra-patient split of MIT-BIH dataset impact the performance of a series of models presented in the literature [84, 86, 96].

6.3.1 Preprocessing

The raw MIT-BIH data firstly goes through a high pass filter (> 0.5 Hz) to remove baseline constant signal. Moving average is applied to remove base drift. Chebyshev type I 4^{th}-order filter and bandwidth 6-18 Hz coupled with Shannon energy filters are used to find the R peak. Then the ECG recordings are segmented into a set of heartbeats with the length of 1.2 RR interval (the time between two consecutive peaks).

In our proposed approach, we use pretrained 2D CNN models (ResNet18) which requires the input data to be in the format of 2D images. Therefore, Short-Term Fourier Transform (STFT) is used to obtain 2D time-frequency spectrograms of the digitized 1D ECG recordings for capturing the frequency variations [100, 202]. The 2D time-frequency spectrograms for each point in the signal is computed by [202],

$$STFTx[n] = X(x, \omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{6.1}$$

Where $x[n]$ is the signal which is sampled at 360 Hz and $w[n-m]$ is the moving window (e.g.,

Hanning window or Guassian window). We suggest using Hanning window with size 512. The resulting 2D spectrograms are in dimension of 224 $\times$ 224. The STFT transformation is performed using Python library *librosa* (version 0.9.1).

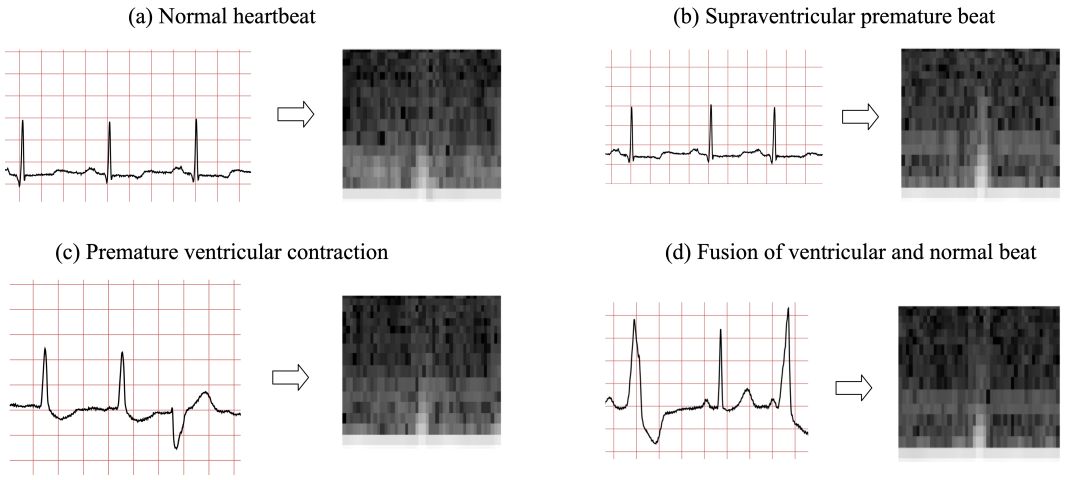


Figure 6.2: ECG grey-scaled spectrograms of the 4 class in MIT BIH dataset.

The class distribution in MIT-BIH dataset is highly imbalanced (the majority class N takes 89% of the entire dataset, see Table 6.1. In this study, we only consider the heartbeat type N, S, V, F, and exclude class Q due to limited samples (n=15). Over- sampling and under-sampling techniques are explored in this study to construct equalized class representation. Note that data sampling methods are only applied on training dataset DS1. We applied oversampling after STFT is performed. Image rotation, flip, and adding Gaussian noise are used to create the artificial data samples.

6.3.2 Arrhythmia Classifier using Transfer Learning

Inspired by the application of using pretrained CNN classifiers (e.g., ResNet18, Res- Net50, etc.) to build predictive models in lung CT scans, we explored the feasibility of using such classifiers in Arrhythmia classification [203]. ResNet18 is used to classify ECG recordings into 4 classes listed in Figure 6.2. The dimension of the input data is adjusted to 224 × 224 × 1 for ResNet 18. A fully

connected layer at the end of ResNet18 is adjusted to predict 4 classes. To classify Arrhythmia, the pretrained ResNet18 network is fine-tuned using the preprocessed DS1 dataset from inter-patient split paradigm. Then the retrained ResNet18 is evaluated using DS2.

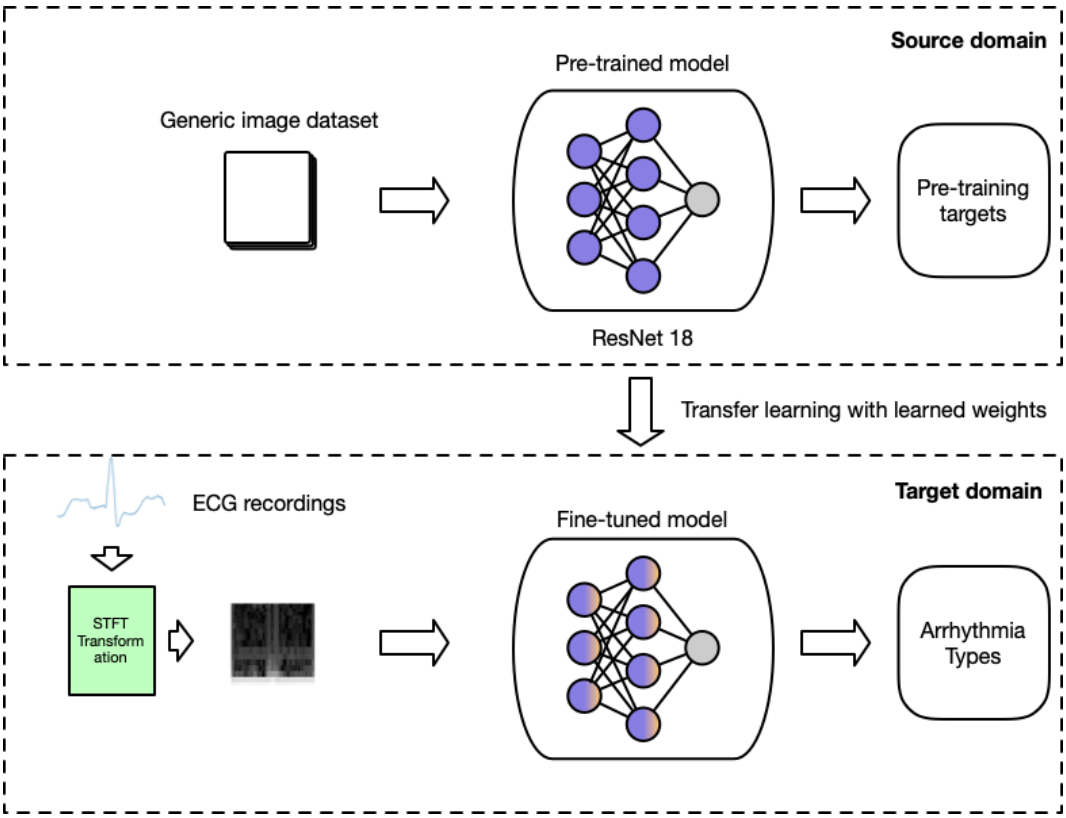


Figure 6.3: Visualization of transfer learning in this work. The pretrained models are developed on generic image dataset. There are wide choices of existing pretrained models such as ResNet18 and VGGs. The pretrained models are then fine-tuned with task-specific data, i.e., 2D ECG data in time-frequency domains transformed from 1D ECG waveform recordings. We suggest using pretrained ResNet 18 for classification.

The training parameters opted for the transfer learning-based model are: i) using Adam optimizer, ii) batch size is 500, iii) model is trained up to 20 epochs, iv) learning rate is .0001. The metrics used for evaluation are precision, recall, and accuracy.

6.3.3 Investigation of how the choice of Intra-patient split versus Inter-patient split paradigm impact model performance

In this section, we investigate how the inter-patient split and intra-patient split methods impact the performance of several state-of -the-art deep learning models in the literature [84, 86]. Eduardo Luz et al. studied the impact of these two data split paradigms [87] on machine learning models such as SVM and shallow neural networks. However, that study was published in 2011 and did not include deep learning models. In our study, we implemented two state-of-the-art deep learning models: Convolutional Neural Networks and its variant described in [84, 86]. Kachuee et al. only evaluates their arrhythmia classifier on intra-patient split [84]. However, we re-implemented[1] these deep learning models and tested the models using both intra-patient and inter-patient split paradigms.

6.4 Results

The performance of the proposed transfer learning framework model is presented in Table 6.2. We also report the comparison study of existing models in the literature and show the performance difference between inter-patient paradigm and intra-patient paradigm in Table 6.2 and Table 6.3, respectively.

It is worth noting that authors in [84] tried to mitigate the imbalance problem even in the "test set" which is contrary to machine learning practice; mitigation occurs during training. Also, in [84] the number of test samples in F class is equal to the entire class which may explain the high precision and recall.

Table 6.2 shows the proposed ResNet18 model with data augmentation achieves best overall accuracy, best recall in normal (N) class, best precisions in arrhythmia (S, V, F) classes. It is worth mentioning that the first model [84] in Table 6.2 was only tested using intra-patient split paradigm in the original paper. In addition, the results from our re-implementation in study [86] are not similar to the reported numbers. We notice that the evaluation procedure is based on intra-patient split even

[1] Our implementation can be found at `https://github.com/wenhaoz-fengcai/ECG-Arrhythmia-Detection-DL`.

Table 6.2: Performance comparison of deep learning models with inter-patient split paradigm. The metrics reported are overall accuracy, precision (Pre), and recall (Rec). Note that the first model was not tested using inter-patient split paradigm in the original paper. The results obtained here are from our re-implementations. The best scores are bold-faced in each column.

Work	**Accuracy (%)**	**Arrhythmia types**				
		N (n=44,259)	**S (n=1,837)**	**V (n=3,221)**	**F (n=338)**	**Q (n=7)**
		Pre/Rec	Pre/Rec	Pre/Rec	Pre/Rec	Pre/Rec
Kachuee [84]	81.2	94.4/84.5	0.0/0.0	30.9/**92.4**	1.0/**1.3**	0.0/0.0
Romdhane [86]	62.1	95.6/64.0	0.0/0.0	12.7/79.3	0.0/0.0	0.0/0.0
Proposed method	**90.8**	**95.3/95.1**	**13.0/9.0**	**68.2**/88.4	**1.3**/0.3	N/A

Table 6.3: Performance of deep learning model re-implementations with intra-patient split paradigm. The reported metric are the overall accuracy, precision (Pre), and recall (Rec) of our implementation. The numbers in paratheses are results reported in the literature.

Work	**Accuracy (%)**	**Arrhythmia types**				
		N (n=18,118)	**S (n=556)**	**V (n=1,447)**	**F (n=161)**	**Q (n=1,608)**
		Pre/Rec	Pre/Rec	Pre/Rec	Pre/Rec	Pre/Rec
Kachuee[84]	93.1	98.4/94.3	38.1/82.6	96.6/82.3	26.6/93.8	98.3/92.6
(reported in literature)	(93.4)	(84.3/97.0)	(98.9/89.0)	(95.0/96.0)	100.0/86.0	100.0/98.0
Romdhane[86] [2]	82.7	82.8/99.9	0.0/0.0	42.9/0.4	0.0/0.0	0.0/0.0

though the authors claimed that their model was tested using inter-patient split method [86]. Table 3 presents the model testing results from our implementation as well as the reported performance in the literature. There is a significant performance drop once deep learning models are trained using inter-patient split instead of intra-patient split for deep learning models described in [84, 86].

Huang et al. only evaluated their model on heartbeat type N, S, V, hence the testing results under F and Q are not available.

6.5 Discussion and Conclusion

We proposed an end-to-end ECG classification framework using 2D CNN classifiers. By transforming the 1D ECG waveforms into 2D frequency-time spectrogram using Short-Term Fourier Transform, the proposed framework provides the opportunity of integrating the general purpose pre-trained 2D CNN models (e.g., VGG-16, Efficient Net, etc) for arrhythmia detection. The proposed method achieves better overall accu- racy compared with deep learning models described in [84, 86].

Our second contribution is to demonstrate how the choice of samples in MIT-BIH dataset significantly impacts the deep learning model performance. We re-implemented two deep learning models for arrhythmia detection in [84, 86], and then tested these models following the AAMI recommendation using the inter-patient data split. We observe that the model evaluation using intra-patient split generates better results compared with the testing results using inter-patient paradigm. However, we argue that the testing set of intra-patient paradigm is susceptible to contamination and is highly likely to have included samples from the same patients appeared in the training set. Therefore, the intra-patient split paradigm is more likely to generate inflated and biased results compared with inter-patient split paradigm.

In addition, there is a lack of consistency regarding the usage of MIT-BIH dataset for arrhythmia classification. For example, study [96] only includes data samples from 14 patients out of 47 in the MIT-BIH dataset. Meanwhile, classifiers in studies [88, 96, 201, 204] only predict a subset of heartbeat types. In [96], only normal beat (NOR), left bundle branch block beat (LBB), right bundle branch block beat (RBB), pre-mature ventricular contraction beat (PVC), atrial premature contraction beat (APC) are in- cluded in the analysis. Moreover, there is a lack of standard reporting in the arrhythmia classification literature. For example, [84, 86] evaluate their models using precision, recall, and overall accuracy, whereas [87, 88] reported their models performance using

specificity and sensitivity.

With the intention of building robust and unbiased arrhythmia classifiers, we highly suggest that practitioners follow the correct practice of splitting the training and testing data to avoid any possible information leakage. Moreover, we are calling for more transparency of data pre-processing and model development, along with a standard of model evaluation. In such a manner, the research community can reproduce and verify the results.

CHAPTER 7

Large-scale Causal Approaches to Debiasing Post-click CVR Estimation with Multi-task Learning

7.1 Introduction

This chapter represents our initial endeavor to enhance deep learning models by incorporating causal inference. Numerous research questions within the realm of recommender systems revolve around estimating the impact of specific interventions or recommendations, essentially exploring cause-and-effect relationships. Consequently, recommender systems serve as a promising foundation for delving into causal inference research regarding interventions.

Post-click conversion rate (CVR) estimation is a critical task in e-commerce recommender systems. This task is deemed quite challenging under industrial setting with two major issues: 1) selection bias caused by user self-selection, and 2) data sparsity due to the rare click events. These two issues have been well discussed in Section 3.6. A successful conversion typically has the following sequential events: "exposure $\rightarrow$ click $\rightarrow$ conversion". Conventional CVR estimators are trained in the click space, but inference is done in the entire exposure space. They fail to account for the causes of the missing data and treat them as missing at random. Hence, their estimations are highly likely to deviate from the real values by large. In addition, the data sparsity issue can also handicap many industrial CVR estimators which usually have large parameter spaces.

In this chapter, we propose two principled, efficient and highly effective CVR estimators for industrial CVR estimation, namely, Multi-IPW and Multi-DR. The proposed models approach the CVR estimation from a causal perspective and account for the causes of missing not at random. In addition, our methods are based on the multi-task learning framework and mitigate the data sparsity

issue. Extensive experiments on industrial-level datasets show that our methods outperform the state-of-the-art CVR models.

We have reviewed several related works in Section 3.6. Our approach differs from those methods in three aspects: 1) The problems are different. We developed our methods for CVR estimation in e-commerce system, while they focus on the rating prediction [205]. 2) The challenges are different. we design our models to address the selection bias and data sparsity issues, while they only consider the former (ESMM considers both). 3) The methods are different. we integrate multi-task framework with causal approaches. Specifically, We co-train propensity model, imputation model and prediction model simultaneously with deep neural networks, while they train these modules separately or alternatively, and usually with models such as linear regression or matrix factorization [206, 207, 208, 209]. We will further justify our design in Section 7.2 and report the performance improvement in Section 7.4.

To simplify the debiasing task of CVR estimation, we assume the exposure space is the entire item space we are interested in (see Figure 3.12) [141]. Such a relaxation is also made based on the postulation that most items are exposed at least once. Table 7.1 shows that our dataset contains 81.5 million items and 11.5 billion exposures, i.e., each item is exposed, on average, about 150 times.

The main contributions of this paper are summarized as follows:

- To the best of our knowledge, this is the first paper that combines IPW-based and DR-based methods with multi-task learning. From a causal perspective, we aim to tackle the well-recognized issues (i.e., selection bias and data sparsity) in CVR estimation in concert.
- We highlight that the state-of-the-art CVR model, ESMM [141], is biased. Different from existing works, our methods adjust for MNAR data, and deal with the selection bias in a principled way. Meanwhile, we give mathematical proofs that the proposed methods are theoretically unbiased. The empirical study shows our approaches outperform ESMM and several state-of-the-art causal models, and demonstrates the efficiency of our methods in real industrial setting.

7.2 Causal CVR Estimators with multi-task learning

7.2.1 Preliminary

Let $\mathcal{U} = (u_1, u_2, ..., u_N)$ be a set of N users and $\mathcal{I} = (i_1, i_2, ..., i_M)$ be a set of M items, $\mathcal{D} = \mathcal{U} \times \mathcal{I}$ be the user-item pairs, $\mathbf{R} \in \mathbb{R}^{N \times M}$ be the true conversion label matrix where each entry $r_{u,i} \in \{0, 1\}$, and $\hat{\mathbf{R}} \in \mathbb{R}^{N \times M}$ be the predicted conversion score matrix where each entry $\hat{r}_{u,i} \in [0, 1]$. Then, the *Prediction inaccuracy* $\mathcal{P}$ over all user-item pairs can be formulated as follows,

$$\mathcal{P} = \mathcal{P}(\mathbf{R}, \hat{\mathbf{R}}) = \frac{1}{|\mathcal{D}|} \sum_{(u,i) \in \mathcal{D}} e(r_{u,i}, \hat{r}_{u,i}), \tag{7.1}$$

where $e(r_{u,i}, \hat{r}_{u,i}) = -r_{u,i} \log(\hat{r}_{u,i}) - (1 - r_{u,i}) \log(1 - \hat{r}_{u,i})$.

Let $\mathbf{O} \in \{0, 1\}^{\mathcal{U} \times \mathcal{I}}$ be the *indicator matrix* where each entry $o_{u,i}$ is an observation indicator: $o_{u,i} = 1$ if a user u clicks on item i, $o_{u,i} = 0$ otherwise. Since the clicks are subjective to certain unobserved factors (e.g., users latent interests), such user self-selection process generates MNAR data [130, 3]. Naive CVR estimators are trained only in the click space $\mathcal{O} = \{(u, i) | o_{u,i} = 1, (u, i) \in \mathcal{D}\}$. Let $\mathbf{R}^{\text{obs}}$ and $\mathbf{R}^{\text{mis}}$ be the set of conversion labels that are present and absent in $\mathcal{D}$. We evaluate these naive CVR models by averaging the cross-entropy loss over the observed data [140, 153],

$$\begin{aligned} \mathcal{E}^{\text{Naive}} &= \mathcal{E}(\mathbf{R}^{\text{obs}}, \hat{\mathbf{R}}) \\ &= \frac{1}{|\mathcal{O}|} \sum_{(u,i) \in \mathcal{O}} e(r_{u,i}, \hat{r}_{u,i}) \\ &= \frac{1}{|\mathcal{O}|} \sum_{(u,i) \in \mathcal{D}} o_{u,i} e(r_{u,i}, \hat{r}_{u,i}), \end{aligned} \tag{7.2}$$

where $|\mathcal{O}| = \sum_{(u,i) \in \mathcal{D}} o_{u,i}$.

We say a CVR estimator $\mathcal{M}$ is *unbiased* when the expectation of the estimated prediction inaccuracy over $\mathbf{O}$ equals to the prediction inaccuracy $\mathcal{P}$, i.e., $\text{Bias}^{\mathcal{M}} = |\mathbb{E}_{\mathbf{O}}[\mathcal{E}^{\mathcal{M}}] - \mathcal{P}| = 0$, otherwise it is *biased*. If data is MNAR, $|\mathbb{E}_{\mathbf{O}}[\mathcal{E}_{\mathcal{M}}] - \mathcal{P}| \gg 0$ [4].

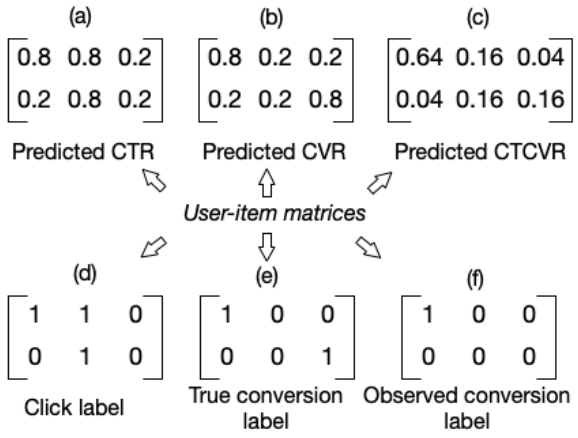


Figure 7.1: A toy example that demonstrates ESMM is biased.

7.2.2 Is ESMM an Unbiased CVR Estimator?

In this section, we demonstrate that ESMM, the state-of-the-art CVR estimator in practice, is essentially biased, though the author claim in the paper that the model eliminates the selection bias [141]. We formulate the estimation bias of ESMM, and prove it is not theoretically unbiased by giving a counter example.

Let $e_{u,i}^{\text{CTR}}, e_{u,i}^{\text{CVR}}, e_{u,i}^{\text{CTCVR}}, (u,i) \in \mathcal{D}$, be the cross-entropy losses of CTR, CVR, and CTCVR tasks. Then we have,

$$\begin{aligned}
&\text{Bias}^{\text{ESMM}} = |\mathbb{E}_{\mathbf{O}}[\mathcal{E}^{\text{ESMM}}] - \mathcal{P}| \\
=&\left| \mathbb{E}_{\mathbf{O}} \left[\frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} (e_{u,i}^{\text{CTR}} + e_{u,i}^{\text{CTCVR}}) \right] - \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} e_{u,i}^{\text{CVR}} \right| \\
=&\left| \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} (e_{u,i}^{\text{CTR}} + e_{u,i}^{\text{CTCVR}}) - \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} e_{u,i}^{\text{CVR}} \right| \\
=&\frac{1}{|\mathcal{D}|} \left| \sum_{(u,i)\in\mathcal{D}} (e_{u,i}^{\text{CTR}} + e_{u,i}^{\text{CTCVR}} - e_{u,i}^{\text{CVR}}) \right|.
\end{aligned} \tag{7.3}$$

We can easily verify that $\text{Bias}^{\text{ESMM}} > 0$ using the counter example in Figure 7.1. Note that to be theoretically unbiased, ESMM should satisfy $|\mathbb{E}_{\mathcal{D}}[\mathcal{E}^{\text{ESMM}}] - \mathcal{P}| = 0, \forall \mathcal{D}$. Therefore, we conclude that ESMM cannot ensure unbiased CVR estimation.

7.2.3 Multi-task Learning Module

To address the data sparsity issue, we adopt the philosophy of multi-task learning and introduce an auxiliary CTR task [210]. The multi-task learning module exploits the typical sequential events in e-commerce recommender system, i.e., "exposure $\rightarrow$ click $\rightarrow$ conversion", and chains the main CVR task with the auxiliary CTR task. The amount of training data in CTR task is generally larger than that in CVR task by $1 \sim 2$ order of magnitudes (see Table 7.1), thus CTR task trains the large volume of model parameters more sufficiently. Besides, the feature representation learned in the CTR task is shared with the CVR task, which makes the CVR model benefit from the extra information via parameter sharing. Hence, the data sparsity issue is remedied [211, 141, 212].

Meanwhile, multi-task learning is also perceived as being cost-effective in training phase [141]. Specifically, multi-task learning co-trains multiple tasks simultaneously as if they were one task. This mechanism can potentially reduce storage space for saving duplicate copies of embedding matrix. In addition, the parallel training mechanism generally reduces the training time by large. The Multi-IPW and Multi-DR models inherit aforementioned merits by incorporating a multi-task learning module.

7.2.4 Multi-task Inverse Propensity Weighting CVR Estimator

Let the marginal probability $P(o_{u,i} = 1)$ denote the propensity score, $p_{u,i}$, of observing an entry in $\mathbf{R}$. In practice, the real $p_{u,i}$ can not be obtained directly. Instead, we estimate the real propensity with $\hat{p}_{u,i}$. The IPW-based estimator uses $\hat{p}_{u,i}$ to inversely weight prediction loss [147, 213, 140, 214],

$$\mathcal{E}^{\text{IPW}} = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i} e(r_{u,i}, \hat{r}_{u,i})}{\hat{p}_{u,i}}. \tag{7.4}$$

Typically, $\hat{p}_{u,i}$ is learned via an independent logistic regression model [215]. In Figure 7.2, we propose the Multi-IPW model which leverages the multi-task learning framework to simultaneously learn the propensity score (i.e., CTR in Multi-IPW) with CVR. Hence, the loss function of Multi-

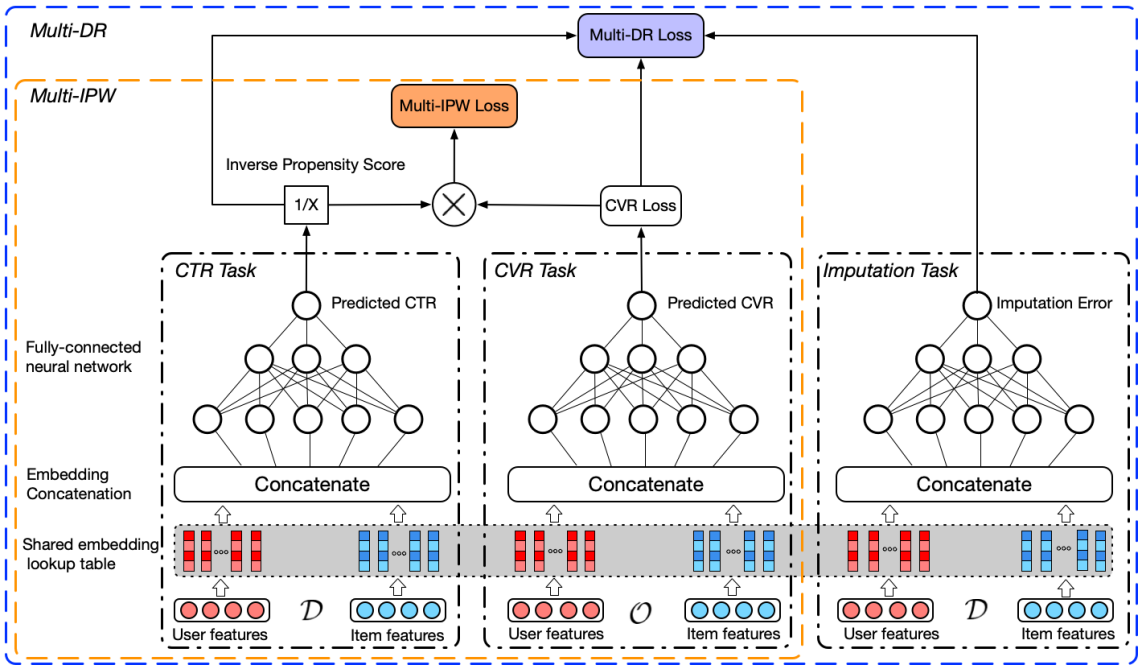


Figure 7.2: Multi-Inverse Propensity Weighting estimator and Multi-Doubly Robust estimator. The Multi-DR estimator augments Multi-IPW with an imputation model. We use predicted CTR as propensity scores in the Multi-IPW estimator. In the multi-task learning module, the CTR task, CVR task, and Imputation task are chained together via parameter sharing.

IPW estimator can be written as follows,

$$\begin{aligned}&\mathcal{E}^{\text{Multi-IPW}}(\mathcal{X}_{\mathcal{O}}; \theta_{\text{CTR}}, \theta_{\text{CVR}}, \Phi) \\ &= \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \frac{o_{u,i} e(r_{u,i}, \hat{r}_{u,i}(\vec{x}_{u,i}; \theta_{\text{CVR}}, \Phi))}{\hat{p}_{u,i}(\vec{x}_{u,i}; \theta_{\text{CTR}}, \Phi)},\end{aligned} \tag{7.5}$$

where Φ represents the shared embedding parameters. θ_{CVR} and θ_{CTR} are neural network parameters of CVR task and CTR task, respectively. $e(r_{u,i}, \hat{r}_{u,i})$, parameterized by θ_{CVR} and Φ, is the cross entropy loss of true CVR label $r_{u,i}$ and predicted CVR score $\hat{r}_{u,i}$. We use the predicted CTR score $\hat{p}_{u,i}$, parameterized by θ_{CTR} and Φ, as propensities. $\mathcal{D}$ denotes all the data in the exposure space. $\mathcal{X}_{\mathcal{O}}$ is the input feature vectors in $\mathcal{O}$.

we formally derive the bias of Multi-IPW and prove it is unbiased given the propensities are accurately estimated.

Theorem 3. *Given the true propensities* $\mathbf{P}$ *and the true conversion label matrix* $\mathbf{R}$*, the Multi-IPW*

CVR estimator gives unbiased CVR prediction when estimated propensity scores are accurate $\hat{p}_{u,i} = p_{u,i}$,

$$|\mathbb{E}_{\mathbf{O}}[\mathcal{E}^{\textit{Multi-IPW}}] - \mathcal{P}| = 0. \tag{7.6}$$

Proof.

$$\begin{aligned}
&|\mathbb{E}_{\mathbf{O}}[\mathcal{E}^{\text{Multi-IPW}}] - \mathcal{P} \\
=&\left|\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \mathbb{E}_{\mathbf{O}}\left[\frac{o_{u,i}e(r_{u,i},\hat{r}_{u,i})}{\hat{p}_{u,i}}\right] - \mathcal{P}\right| \\
=&\left|\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} \frac{p_{u,i}e(r_{u,i},\hat{r}_{u,i})}{\hat{p}_{u,i}} - \mathcal{P}\right| \\
=&\left|\frac{1}{|\mathcal{D}|}\sum_{(u,i)\in\mathcal{D}} e(r_{u,i},\hat{r}_{u,i}) - \mathcal{P}\right| = 0.
\end{aligned} \tag{7.7}$$

□

Multi-IPW estimator inherits the merits of multi-task learning: 1) better CVR prediction due to parameter sharing, and 2) reduced training time and parameter storage. These are clear advantages over conventional IPW-based estimators.

Algorithm 3: Multi-Inverse Propensity Weighting

Input: Observed conversion labels $\mathcal{R}^{obs}$ and user-item features $\mathcal{X}_{\mathcal{O}} = \{\vec{x}_{u,i}\}^{\mathcal{O}}, u, i \in \mathcal{O}$

while stopping criteria is not satisfied **do**

  Sample a batch $\mathcal{B}$ of user-item pairs $\{\vec{x}_{u,i}\}_B$ from $\mathcal{O}$

  Co-train CTR task and CVR task

    Update θ_{CTR}, Φ by descending along the gradients

    $\nabla_{\theta_{CTR}}\mathcal{E}^{Multi-IPW}, \nabla_{\Phi}\mathcal{E}^{Multi-IPW}$

    Update θ_{CVR}, Φ by descending along the gradients

    $\nabla_{\theta_{CVR}}\mathcal{E}^{Multi-IPW}, \nabla_{\Phi}\mathcal{E}^{Multi-IPW}$

end

7.2.5 Multi-task Doubly Robust CVR Estimator

The IPW-based models are unbiased contingent on accurately estimated propensities (i.e., $\hat{p}_{u,i} = p_{u,i}$). In practice, this condition is too restricted. To address this issue, doubly robust estimator is introduced by previous works [216, 153, 217, 218].

Wang *et al.* [153] proposed a joint learning approach for training a doubly robust estimator, and introduced two models: 1) a prediction model $\hat{r}_{u,i} = f_\theta(\vec{x}_{u,i})$, and 2) an imputation model $\hat{e}_{u,i} = g_\phi(\vec{x}_{u,i})$. The prediction model, parameterized by θ, aims to predict the ratings, and its performance is evaluated by $e_{u,i} = e(r_{u,i}, \hat{r}_{u,i}), (u,i) \in \mathcal{D}$. The imputation model, parameterized by ϕ, aims to estimate the prediction error $e_{u,i}$ with $\hat{e}_{u,i}$. Its performance is assessed by $\delta_{u,i} = e_{u,i} - \hat{e}_{u,i}$. The feature vector $\vec{x}_{u,i}$ encodes all the information about the user u and the item i, $(u,i) \in \mathcal{D}$. Then, we can formulate the loss of doubly robust estimator as,

$$\mathcal{E}^{\text{DR}} = \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \left(\hat{e}_{u,i} + \frac{o_{u,i}\delta_{u,i}}{\hat{p}_{u,i}} \right), \tag{7.8}$$

Similarly, we propose the Multi-DR estimator which augments Multi-IPW estimator by including an imputation model estimating the prediction error $e_{u,i}$. Multi-DR optimizes the following loss,

$$\begin{aligned} &\mathcal{E}^{\text{Multi-DR}}(\mathcal{X}; \theta_{\text{CTR}}, \theta_{\text{CVR}}, \theta_{\text{Imp}}, \Phi) \\ &= \frac{1}{|\mathcal{D}|} \sum_{(u,i)\in\mathcal{D}} \left(\hat{e}_{u,i}(\vec{x}_{u,i}; \theta_{\text{Imp}}, \Phi) + \frac{o_{u,i}\delta_{u,i}(\vec{x}_{u,i}; \theta_{\text{CVR}}, \theta_{\text{Imp}}, \Phi)}{\hat{p}_{u,i}(\vec{x}_{u,i}; \theta_{\text{CTR}}, \Phi)} \right), \end{aligned} \tag{7.9}$$

where Φ represents the shared embedding parameters among CTR task, CVR task, and imputation task. θ_{CTR}, θ_{CVR}, θ_{Imp} are neural network parameters of CTR task, CVR task, imputation task, respectively. $\hat{p}_{u,i}$ is the propensity (i.e., predicted CTR score) given by CTR task. Estimated prediction error $\hat{e}$, parameterized by θ_{Imp} and Φ, is given by imputation task. $\delta_{u,i} = e_{u,i} - \hat{e}_{u,i}$ is the error deviation.

We can formally derive the bias of Multi-DR and prove it is unbiased if either true propensity scores or true prediction errors are accurately estimated (i.e., $\Delta_{u,i} = 0$ or $\delta_{u,i} = 0$).

Theorem 4. *Given the true propensities* $\mathbf{P}$ *and the true conversion label matrix* $\mathbf{R}$*, the Multi-DR CVR estimator gives unbiased CVR prediction when either estimated propensity scores are accurate*

$\Delta_{u,i} = \frac{p_{u,i} - \hat{p}_{u,i}}{\hat{p}_{u,i}} = 0$ *or the estimated prediction errors are accurate* $\delta_{u,i} = e_{u,i} - \hat{e}_{u,i} = 0$*,*

$$|\mathbb{E}_{\mathbf{O}}[\mathcal{E}^{\textit{Multi-DR}}] - \mathcal{P}| = 0. \tag{7.10}$$

Proof.

$$\begin{aligned}
&|\mathbb{E}_{\mathbf{O}}[\mathcal{E}^{\text{Multi-DR}}] - \mathcal{P}| \\
&= \frac{1}{|\mathcal{D}|}\left| \sum_{(u,i)\in\mathcal{D}} \left(\hat{e}_{u,i} + \mathbb{E}_{\mathbf{O}}\left[\frac{o_{u,i}\delta_{u,i}}{\hat{p}_{u,i}} \right] \right) - \mathcal{P} \right| \\
&= \frac{1}{|\mathcal{D}|}\left| \sum_{(u,i)\in\mathcal{D}} \left(\hat{e}_{u,i} + \frac{p_{u,i}\delta_{u,i}}{\hat{p}_{u,i}} \right) - \mathcal{P} \right| \\
&= \frac{1}{|\mathcal{D}|}\left| \sum_{(u,i)\in\mathcal{D}} \frac{(p_{u,i} - \hat{p}_{u,i})\delta_{u,i}}{\hat{p}_{u,i}} \right| \\
&= \frac{1}{|\mathcal{D}|}\left| \sum_{(u,i)\in\mathcal{D}} \Delta_{u,i}\delta_{u,i} \right| = 0.
\end{aligned} \tag{7.11}$$

□

Algorithm 4: Multi-Doubly Robust

Input: Observed conversion labels $\mathcal{R}^{obs}$ and user-item features $\mathcal{X}_{\mathcal{D}} = \{\vec{x}_{u,i}\}^{\mathcal{D}}, u, i \in \mathcal{D}$

while stopping criteria is not satisfied **do**

Sample a batch $\mathcal{B}$ of user-item pairs $\{\vec{x}_{u,i}\}_B$ from $\mathcal{D}$

Co-train CTR task, CVR task, and Imputation task

Update θ_{CTR}, Φ by descending along the gradients $\nabla_{\theta_{CTR}}\mathcal{E}^{Multi-DR}, \nabla_{\Phi}\mathcal{E}^{Multi-DR}$

Update θ_{CVR}, Φ by descending along the gradients $\nabla_{\theta_{CVR}}\mathcal{E}^{Multi-DR}, \nabla_{\Phi}\mathcal{E}^{Multi-DR}$

Update θ_{Imp}, Φ by descending along the gradients $\nabla_{\theta_{Imp}}\mathcal{E}^{Multi-DR}, \nabla_{\Phi}\mathcal{E}^{Multi-DR}$

end

7.3 Experimentation

In this section, we evaluate the performance of the proposed models with a public dataset and a large-scale production dataset collected from Mobile Taobao, the leading e-commerce platform in China. The experiments are intended to answer the following questions:

- **Q1**: Do our proposed approaches outperform other state-of-art CVR estimation methods?
- **Q2**: Are our proposed models more efficient in industrial setting than other baseline models?
- **Q3**: How is the performance of our proposed models affected by hyper-parameters?

7.3.1 Datasets

Ali-CCP [1][141]

Alibaba Click and Conversion Prediction (Ali-CCP) dataset is collected from real-world traffic logs of the recommender systems in the Taobao platform. See the statistics in Table 7.1.

Production sets

This industrial production dataset is collected from the Mobile Taobao e-commerce platform. It contains 3-week transactional data (see Table 7.1). Our production dataset includes 109 features, which are primarily categorized into: 1) user features, 2) item features, and 3) combination features. We further divide this dataset into 4 subsets: Set A, Set B, Set C, and Set D, which contain the first two days (5%), the first five days (20%), the first twelve days (50%), and the 3-week of data (100%), respectively. We use the data of the last day in each set as testing set and the remaining data as training set.

Table 7.1: Statistics of experimental datasets

Dataset	# Exposure	# Click	# Conversion	# user	# item
Ali-CCP	84M	3.4M	18k	0.4M	4.3M
Set A	1.1B	54.5M	0.6M	-	22.5M
Set B	2.7B	0.2B	1.9M	-	39.1M
Set C	6.0B	0.4B	4.3M	-	62.6M
Set D	11.5B	0.6B	8.3M	-	81.5M

[1] `https://tianchi.aliyun.com/dataset/dataDetail?dataId=408`

7.3.2 Baseline models

We compare Multi-IPW model and Multi-DR model with the following baselines. Note that some baselines are causal estimators which we modify to predict the unbiased CVR, and others models are existing non-causal estimators designed for CVR predictions.

7.3.2.1 Non-causal estimators

- **Base** is a naive post-click CVR model, which is a Multi-layer Perceptron (See the CVR task in Figure 7.2). Note that this is essentially an independent MLP model which takes the feature embeddings as input and predicts the CVR. The base model is trained in the click space.

- **Oversampling** [219, 220] deals with the class imbalance issue by duplicating the minority data samples (conversion=1) in training set with an oversampling rate k. In our experiment, we set $k = 5$. The oversampling model is trained in the click space.

- **ESMM** [141] utilizes multi-task learning methods and reduces the CVR estimation into two auxiliary tasks, i.e., CTR task and CTCVR task. ESMM is trained in the entire exposure space, and deemed as the state-of-the-art CVR estimation model in real practice.

- **Naive Imputation** takes all the unclicked data as negative samples. Hence, it is trained in the entire exposure space.

7.3.2.2 Causal estimators

- **Naive IPW**[140] is a naive IPW estimator. Note that it is not specifically designed for CVR estimation task as CVR prediction has its intrinsic issues. For example, it cannot deal with the data sparsity issue that inherently exists in CVR task.

- **Joint Learning DR** [153] is devised to learn from ratings that are missing not at random. In this experiment, we tailor Joint Learning DR for the CVR estimation. Similarly, Joint learning DR handles data sparsity issue poorly.

- **Heuristic DR** is designed as a baseline for Multi-DR. It assumes that the unclicked items are negative samples with probability $1-\eta$, where η is smoothing rate and it denotes the probability of having a positive label. In the experiments, we explore η in $\{0.0005, 0.001, 0.002, 0.005, 0.01\}$ and report the best performance.

7.3.3 Metrics

In CVR prediction task, ROC AUC is a widely used metric [221]. One interpretation of AUC in the context of ranking system is that it denotes the probability of ranking a random positive sample higher than a negative sample. Meanwhile, we also adopt Group AUC (GAUC) [222]. GAUC extends AUC by calculating the weighted average of AUC grouped by page views or users,

$$\text{GAUC} = \frac{\sum_{i \in U} w_i \times \text{AUC}_i}{\sum_{i \in U} w_i}, \tag{7.12}$$

where w_i is exposures. GAUC is commonly recognized as a more indicative metric in real practice [222]. In the public dataset, models are only assessed with AUC as the dataset is missing the information for computing GAUC.

7.3.4 Unbiased Evaluation

In this work, we use CTCVR-AUC/GAUC to evaluate the unbiasedness of CVR estimators [141]. We need to point out that testing with an unbiased dataset or randomization is generally a golden standard for unbiasedness assessment [223, 224]. However, the unbiased training/testing dataset for CVR estimation is rather unobtainable in real practice.The real-world system can randomly expose items to users and generate unbiased evaluation sets for CTR estimation. But they cannot force users to randomly click on items to generate unbiased data for CVR estimations. This limitation may be further investigated in the future work.

7.3.5 Experiments setup

7.3.5.1 Ali-CCP experiment

The experiment setup on Ali-CCP mostly follows the prior work [141]. We set the dimension of all embedding vectors to be 18. The architecture of all these multi-layer perceptrons (MLP) in multi-task learning module are identical as $512 \times 256 \times 128 \times 32 \times 2$. The optimizer is Adam with a learning rate $lr = 0.0002$, and batch size is set to $|batch| = 1024$.

7.3.5.2 Production set experiment

In production set experiment, we vary the dimensions of feature embedding vectors according to each feature's real size in order to minimize the memory usage. In order to have a fair comparison study, all the models in this experiment share $|batch| = 10000$, MLP architecture $1024 \times 512 \times 256 \times 128 \times 32 \times 2$, adam optimizer with learning rate $lr = 0.0005$. We also added $l2$ normalization to imputation model in Multi-DR, and the coefficient is $v = 0.0001$.

7.4 Results and Discussion

In this section, we evaluate the proposed models and answer the questions raised in Section 7.4.

7.4.1 Model Assessments (Q1)

In this section, we report the experiment results in Table 7.2, 7.3. Multi-IPW and Multi-DR are clear winners over other baselines across all experiments. Meanwhile, we have the following observations:

- In production dataset, Multi-IPW and Multi-DR consistently outperform Joint Learning DR [153]. We reason that the performance improvement benefits from multi-task learning module, which remedies the data sparsity issue.

Table 7.2: Results of comparison study on Production datasets. The best scores are bold-faced in each column. Note that this table has two sections, AUC scores and GAUC scores. The rows that contain the models proposed in this paper are highlighted in color grey.

Model	Set A (1.1B)		Set B (2.7B)		Set C (6.0B)		Set D (11.5B)	
	CVR	CTCVR	CVR	CTCVR	CVR	CTCVR	CVR	CTCVR
AUC score								
Base	78.24	73.12	78.67	73.86	79.62	74.70	81.66	76.28
Oversampling[219]	78.63	73.53	78.72	74.09	79.69	74.82	81.77	76.30
ESMM[141]	79.29	73.86	79.74	74.33	80.11	74.97	82.17	76.55
Naive Imputation	78.12	73.21	78.44	73.50	79.32	73.81	81.56	76.39
Naive IPW[140]	79.23	73.82	79.73	74.34	80.14	74.92	82.13	76.45
Heuristic DR	78.45	73.45	78.84	73.99	79.52	74.18	81.74	76.40
Joint Learning DR[153]	79.09	73.67	79.53	74.51	80.01	74.90	82.09	76.61
Multi-IPW	79.51	73.99	**79.85**	74.81	80.21	75.01	82.57	76.89
Multi-DR	**79.72**	**74.45**	79.80	**74.91**	**80.50**	**75.39**	**82.72**	**77.23**
GAUC score								
Base	-	59.69	-	60.16	-	60.58	-	61.27
Oversampling[219]	-	60.17	-	60.28	-	60.59	-	61.30
ESMM[141]	-	60.53	-	60.90	-	61.13	-	61.76
Naive Imputation	-	60.14	-	60.39	-	60.56	-	61.39
Naive IPW[140]	-	60.51	-	60.95	-	61.09	-	61.77
Heuristic DR	-	60.01	-	60.30	-	60.65	-	61.35
Joint Learning DR[153]	-	60.43	-	60.83	-	60.97	-	61.67
Multi-IPW	-	60.70	-	**61.09**	-	61.25	-	61.98
Multi-DR	-	**60.90**	-	60.99	-	**61.52**	-	**62.28**

- We notice that Multi-DR mostly has better performance than Multi-IPW. Recall that Multi-DR augments Multi-IPW by introducing an imputation model. Provided that $0 \leq \hat{e} \leq 2e$, the tail bound of Multi-DR is proven to be lower than that of Multi-IPW for any learned

Table 7.3: Results of comparison study on Public dataset: Ali-CCP. Experiments are repeated 10 times and mean $\pm$ 1 std of AUC scores are reported below. The best scores are bold-faced in each column. The rows that contain the models proposed in this paper are highlighted in color grey.

Model	CVR AUC	CTCVR AUC
Base	66.00 ± 0.37	62.07 ± 0.45
Oversampling [219]	67.18 ± 0.32	63.05 ± 0.48
ESMM-NS [141]	68.25 ± 0.44	64.44 ± 0.62
ESMM [141]	68.56 ± 0.37	65.32 ± 0.49
Multi-IPW	69.21 ± 0.42	65.30 ± 0.50
Multi-DR	$\mathbf{69.29 \pm 0.31}$	$\mathbf{65.43 \pm 0.34}$

propensity score $\hat{p}_{u,i}, (u, i) \in \mathcal{O}$ [153]. Therefore, Multi-DR is expected to perform better than Multi-IPW when the imputation model is well-trained.

- We observe that Multi-IPW/Naive IPW estimator are superior to Base in all experiments. Compared with the Base, both IPW-based models introduce estimated propensities to correct the selection bias. Recall that Theorem 1. ensures the CVR estimators are unbiased if the propensities are accurately estimated. While in practice the estimated propensities may deviate from the real values, this control experiment attests to "enough accuracy" of the propensity model.

The experiment results demonstrate that Multi-IPW and Multi-DR counter the selection bias and data sparsity issues in CVR estimation in a principled and highly effective way. In the next subsection, we will discuss other strengths of the proposed methods.

7.4.2 Computational efficiency (Q2)

In this section, we study the computational efficiency of the proposed models against the baselines under industrial setting. We summarize the records of training time and parameter space size of each model in Figure 7.3, and the cluster configuration in Table 7.4.

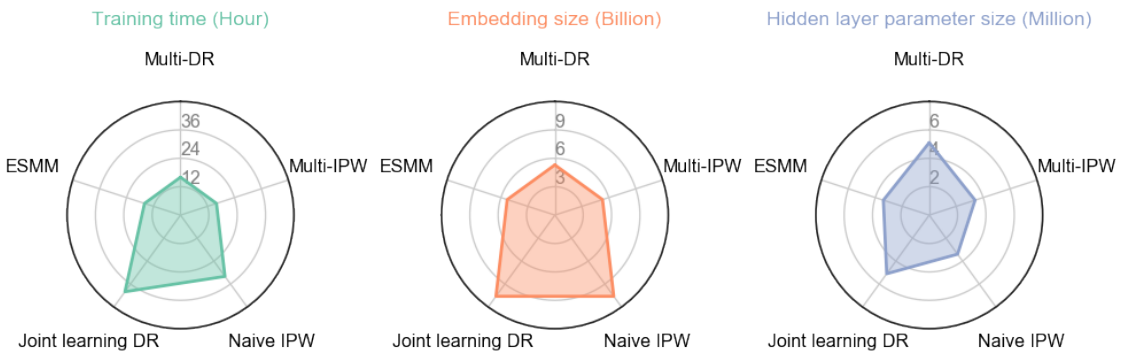


Figure 7.3: Computational cost of Multi-IPW and Multi-DR. The left subplot reveals the hours needed to complete one epoch of training. The middle subplot shows the size of embedding parameters of each model. The right subplot shows the size of hidden layer parameters of each model. Note that the proposed models achieve the best prediction performance, while have the lowest computational cost.

We observe that Multi-IPW and Multi-DR require less or equivalent training time compared with other baselines. Recall that multi-task learning method co-trains multiple tasks simultaneously as if they were one task. We can expect the training time being greatly shortened. Meanwhile, our methods are also economical in memory usage due to parameter sharing in the multi-task learning module.

Table 7.4: Distributed cluster configuration

Cluster configuration	Parameter Server	Worker
# instances	4	100
# CPU	28 cores	440 cores
# GPU[2]	-	25 cards
MEMORY (GB)	40	1000

[2]GPU specs: Tesla P100-PCIE-16G

7.4.3 Hyper-parameters in model implementation

IPW bound τ is a hyper-parameter introduced in our model implementation to handle high variance of propensities. IPW bound clamps the propensities if the values are greater than the predefined threshold. A plausible IPW bound value is typically confined by $\tau \in [propensity_{min}, propensity_{mean}]$. IPW bound percentage can be calculated as $\tau\% = \frac{\tau - propensity_{min}}{propensity_{mean} - propensity_{min}}$. In Multi-DR, imputation model will introduce the unclicked items to the training set. Empirically, most of the unclicked items will not be purchased by customers even if they were clicked. Therefore, including these unclicked items in training set will skew the data distribution and make the class imbalance issue worse. Therefore, Instead of adding all the unclicked samples, we under-sample them with a sampling rate λ. For example, if the number of clicked samples ($N_{clicked}$) is 100 and the batch size is 1000, $\lambda = 1.5$ means that after under-sampling the samples we used to train Multi-DR is $N_{clicked} \times \lambda = 100 \times 1.5 = 150$. Note that without under-sampling, Multi-DR takes all samples in the batch as training samples.

7.4.4 Empirical study on hyper-Parameter sensitivity (Q3)

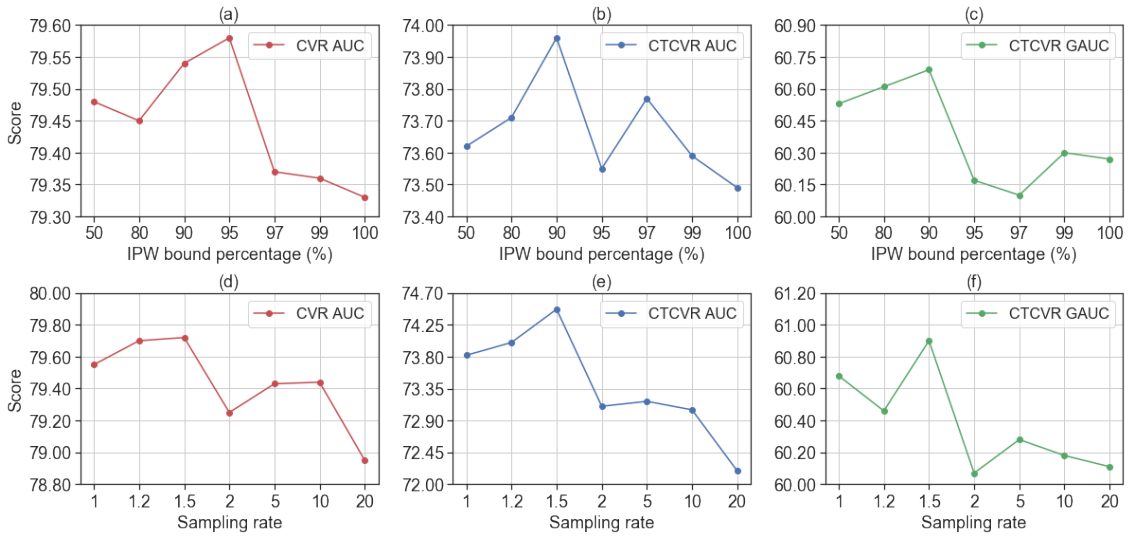


Figure 7.4: Results of parameter sensitivity experiments.

In this section, we investigate how Multi-IPW and Multi-DR are affected by two important

hyper-parameters in our model implementation, IPW bound τ and sampling rate λ. We evaluate the performance of Multi-IPW with varying IPW bound τ. We observe that in Figure 7.4, when IPW bound $50\% < \tau\% < 90\%$, prediction performance eventually improves as IPW bound increases. We can clearly see the performance drop of CVR AUC if the threshold is greater than 95%. We reason that larger IPW bound allows undesired higher variability of propensity scores, which may lead to sub-optimal prediction performance.

We evaluate the performance of Multi-DR with varying sampling rate λ. We observe that $\lambda = 1.5$ produces the best prediction, and the model performance starts decreasing when $\lambda > 5$. We argue that, as the sampling rate increases, more unclicked samples are included to our training set, and it inevitably worsens the class imbalance issue, which typically causes predictive models to generalize poorly. On the contrary, introducing a small number of unclicked samples from the imputation model can boost our CVR prediction (see figure (d) when $\lambda \in [1.0, 1.5]$).

7.5 Conclusion and future works

In this paper, we proposed Multi-IPW and Multi-DR CVR estimators for industrial recommender system. Both CVR estimators aim to counter the inherent issues in practice: 1) selection bias, and 2) data sparsity. Extensive experiments with billions of data samples demonstrate that our methods outperform the state-of-the-art CVR predictive models, and handle CVR estimation task in a principled, highly effective and efficient way. Although our methods are devised for CVR estimation, the idea can be generalized to debiasing CTR estimation by exploiting the sequential pattern "item pool $\rightarrow$ exposure $\rightarrow$ click".

CHAPTER 8

Curse of dimensionality and Causal discovery

8.1 Introduction

Healthcare data analytics has been longing for interpretable and transparent learning methods. Identification of the evidence of causality from observational dataset can keep healthcare practitioners from unreliable decision making due to confounding issues, which are commonly observed [25, 24]. Meanwhile, causal discovery also enables domain adaptation and model fairness research [225, 226]. Existing works of "causal discovery theory" for causal relations identification are far from practical for healthcare analytics due to the several theoretical challenges such as data high dimensionality and heterogeneity.

The objective of this project is to advance the casual discovery research in high dimensional setting. Casual relations are often encoded in directed acyclic graphs (DAGs), known as causal graphs. However, a major barrier deterring the causal discovery application in healthcare domain is the complexity in constructing causal graphs. In high dimensional datasets, say, the protein interaction data, search problem of identifying causal diagrams from observational data is deemed computationally intensive since the search space of directed acyclic graphs scales super-exponentially with the number of nodes [227]. In this chapter, we proposed a novel continuous constraint optimization method for causal discovery. The proposed constraint has better computational complexity compared to the exponential constraint with an exponential term used in [167].

Issues with causal discovery in high dimension space

The mainstream causal structure learning with high dimensional dataset is widely acknowledged challenging [164, 168, 167, 228, 229, 230, 231]. Recall that constraint-based methods start with a fully connected graph and perform conditional independence tests to remove the edges between variables if the vertices are not related. The score-based methods, on the contrary, start with an empty graph and evaluate whether an edge addition, deletion, and direction orientation improves the fitness of graph using likelihood-based score (e.g., Bayesian Information Criteria). Both the constraint-based methods and score-based methods face the challenge that the number of directed acyclic graphs (e.g., DAGs) grows super-exponentially with the number of nodes [164, 229, 230]. The number of directed acyclic graphs with N vertices is lower bounded by the number of undirected acyclic graphs with N vertices as every DAG can be converted into a corresponding undirected graph by removing the edge orientation. Given N vertices, the number of distinct undirected edge is $\binom{n}{2}$, and the number of undirected graphs is hereby $2^{\binom{n}{2}}$. Therefore, the number of DAGs with N vertices is lower bounded by $2^{\binom{n}{2}}$, and searching for the best Bayesian network cannot be decided in polynomial time.

8.2 Methodology

This section introduce the formulation of the causal discovery as a continuous constraint problem with a polynomial constraint.

8.2.1 Preliminary

The data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ contains n observations of a random vector $\vec{x} = (x_1, x_2, ..., x_d)$. The space of directed acyclic graphs (DAGs) is denoted by $\mathbb{D}$, where each DAG is represented by $G =< V, E >$ with d nodes. This paper focus on learning linear structural equation modeling (SEM) under additive noise models, which the same data generation method as described in [167]. Specifically, the linear SEM is defined as follows,

$$x_j = a_j^T \vec{x} + z_j \tag{8.1}$$

where z_j is an additive noise. We assume the noises are under Gaussian distribution with zero mean for simplicity. a_j is a column vector of coefficients and $A =< a_1, a_2, ..., a_d >$ is the weighted adjacency matrix of the linear SEM. To learn this linear causal model, we use least-square loss function following the work in [232, 167],

$$F(A) = \frac{1}{2n}||\mathbf{X} - \mathbf{X}A||_F^2 + \lambda||A||_1 \tag{8.2}$$

The $l1$ regularization term $||A||_1$ is added to the least-square loss function for learning a sparse DAG.

8.2.2 Characterization of acyclicity

In this study, we utilize a non-negative adjacency matrix A to establish a DAG. The presence of a directed edge between node i and node j is indicated by a non-zero value in the element a_{ij}. Conversely, if $a_{ij} = 0$, it signifies the absence of an edge between node i and node j.

In essence, our objective is to compute the adjacency matrix A, with the added condition that the resulting graph must be acyclic. Recall the theorem in [233, 167] states that the entry (i, j) in the k-th power of a non-negative adjacency matrix A^k (i.e., $a_{i,j}^k$) says the there is a walk of length k between nodes i and j. Similarly, the diagonal elements of the adjacency matrix, $a_{i,i}^k$, indicates the number of length-k closed walk starting from and ending at the same vertex i.

Therefore, $a_{i,i}^k = 0(\forall i \in \{1, ..., d\})$ in $A^k(W \in \mathbb{R}^{d\times d})$ guarantees no length-k closed walk starting at node i in the graph. Equivalently, the constraint of acyclicity for any graph represented by adjacency matrix A can be written as follows.

$$trace(A^1) + trace(A^2) + trace(A^3) + ... + trace(A^d) = 0 \tag{8.3}$$

Now we use proof by contradiction to prove that equation 8.3 is sufficient to eliminate any closed walk in a directed graph with d nodes. Recall in graph theory, a cycle is a closed walk without repeated interim nodes but the starting node and ending node are the same. Assume equation 8.3 is not sufficient, and there is a closed walk of length $d+1$ in the graph. Then there must be repeated interim nodes on that walk. This contradicts the definition of the cycles in the graph theory. Hence the longest cycle in any directed graph with d nodes is d.

Note that our assumption is that A is non-negative. In order to make this assumption more general, we can consider a matrix $A = W \circ W$, where each element of A is the square of the corresponding element in W. By setting W as an adjacency matrix of a directed graph with dimensions $\mathbb{R}^{d\times d}$, we can maintain the characteristic of acyclicity that was previously mentioned.

Let $A = W \circ W$, and we can rewrite equation 8.3 as,

$$\sum_{i=1}^{d} tr((W \circ W)^i) = tr(W \circ W) + tr((W \circ W)^2) + ... + tr((W \circ W)^d) = 0 \tag{8.4}$$

Theorem 5. *Let $W \in \mathbb{R}^{d\times d}$ be the weighted adjacency matrix of a directed graph. The graph is acyclic if and only if,*

$$\sum_{i=1}^{d} tr((W \circ W)^i) = 0 \tag{8.5}$$

8.2.3 Problem formulation

The causal discovery algorithm is formulated as a continuous constraint optimization problem as follows,

$$\min_{W\in\mathbb{R}^{d\times d}} F(W) \tag{8.6}$$

$$\text{subject to: } \sum_{i=1}^{d} tr((W \circ W)^i) = 0 \tag{8.7}$$

where $F(W) = \frac{1}{2n}||\mathbf{X} - \mathbf{X}W||_F^2 + \lambda||W||_1$

Note that the constraint in equation 8.7 is a polynomial term. Compare with the exponential constraint with exponential term in [167], the proposed constraint could be optimized and have better computational complexity.

The summation of the geometric series in equation 8.7 can be rewritten as the following,

$$\sum_{i=1}^{d} tr((W \circ W)^i) = tr(W \circ W) + tr((W \circ W)^2) + ... + tr((W \circ W)^d) \tag{8.8}$$

$$= tr([I - (W \circ W)]^{-1}[I - (W \circ W)^{d+1}]) \tag{8.9}$$

Then the computation optimized problem formulation can be written as follows,

$$\min_{W \in \mathbb{R}^{d \times d}} F(W) \tag{8.10}$$

$$\text{subject to: } h(W) \equiv tr([I - (W \circ W)]^{-1}[I - (W \circ W)^{d+1}]) = 0 \tag{8.11}$$

where $F(W) = \frac{1}{2n}||\mathbf{X} - \mathbf{X}W||_F^2 + \lambda||W||_1$.

Complexity analysis of the proposed constraint

The computational complexity of the constraint with matrix exponetial term ($e^{(W \circ W)}$) in [167] is $O(d^3)$, where d is the dimension of feature space. The computational complexity of equation 8.11 can be optimized to $O(d^{2.37})$ if $W \circ W$ can be diagonalized [234]. Specifically, the complexity of $(I - W \circ W)^{-1}$ is $O(d^{2.37})$ with two steps: 1) Hadamard product takes $O(d^2)$, 2) given $(W \circ W)$, computing the inversion takes $O(d^3)$ in general. However, this operation can be optimized to $O(d^{2.37})$ with optimized CW-like algorithm [234]. Similarly, the complexity of $(I - (W \circ W)^{d+1})$ can be optimized to $O(d^{2.37})$ if $(W \circ W)$ can be diagnolized using Single Value Decomposition [235]. Therefore, the overall complexity of computing equation 8.11 can be optimized to $O(d^{2.37})$.

8.2.4 Training

In the previous section, we have formulated the causal discovery as a constrained problem in equation 8.10 - 8.11. Extensive research has been conducted on non-linear equality-constrained problem, and one commonly employed method for their solution is the augmented Lagrangian approach. We give a brief summary of the solution in this section and readers are referred to the textbooks [236, 237].

In this work, we convert the proposed constrained optimization problem into a series of unconstrained minimization problems using augmented Lagrangian method. The unconstrained optimization problem can be formed by adding a penalty term to the objective function (equation 8.10). The converted unconstrained problems can be written as follows,

$$D(\alpha) = \min_{W \in \mathbb{R}^{d \times d}} L^{\rho}(W, \alpha), \tag{8.12}$$

where $L^{\rho}(W, \alpha) = F(W) + \frac{\rho}{2}|h(W)|^2 + \alpha h(W)$, and α is the Lagrange multiplier and ρ is the penalty term. The penalty term is a measure of violation of the constraint. Its value is non-zero if the constraint is violated and is zero if the constraint is satisfied [238].

The unconstrained optimization problem 8.12 can be solved efficiently by standard gradient methods algorithms [239].

8.3 Experimentation

The proposed method is compared against the no-tears method [167] and FGES method [240] in this study, using both synthetic and real-world datasets. The aim is to showcase superior performance in terms of metrics such as true positive rate and false positive rate, as well as greater efficiency in terms of time units such as seconds.

8.3.1 Synthetic dataset

To generate simulation datasets in this experiment, we follow a series of steps. First, we create graphs, which also serve as the ground truth, using a random graph model called Erdos-Renyi (ER) [241]. We then assign random weights to the edges in the graph, generating the adjacency matrix W. Using W, we sample data points according to $X = XW + Z$ (Z being the Gaussian noise parameter). The resulting data samples are independent and identically distributed (*i.i.d.*). The simulation datasets can be generated with varying data sizes, such as $n \in \{20, 1K\}$, and feature dimensions, such as $d \in \{20, 50, 100\}$. A data size of $n = 20$ simulates a high-dimension scenario ($d \gg n$), while $n = 1K$ simulates a low-dimension case [167, 242].

8.3.1.1 Parameter estimation

This section is a qualitative study of the estimated adjacency matrix obtained from the proposed method, compared with the ground truth in the form of heatmap side by side. The results in Figure 8.1 are generated with $n = 1K$ and $d = 20$ for the ease of visualization. This qualitative study attempts to demonstrate that the proposed model is as accurate as the no-tears while outperforms the classic causal discovery algorithm FGES.

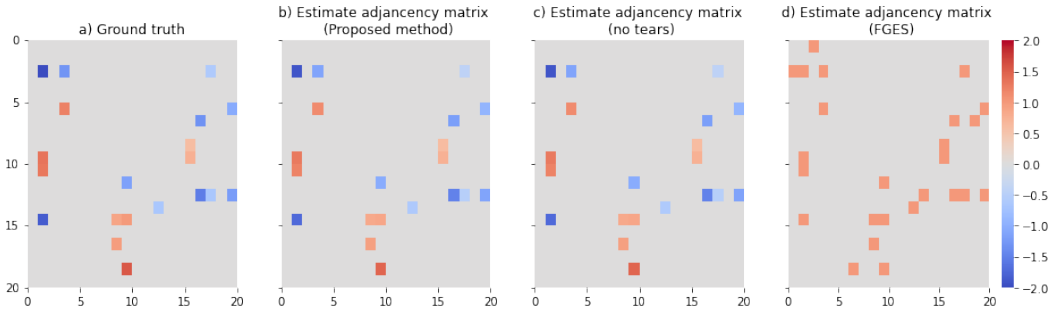


Figure 8.1: According to the results of the qualitative study, the proposed method exhibits accuracy that is on par with no-tears, while also outperforming FGES. These results were obtained using a dataset with a size of $n = 1000$ and feature dimensions of $d = 20$.

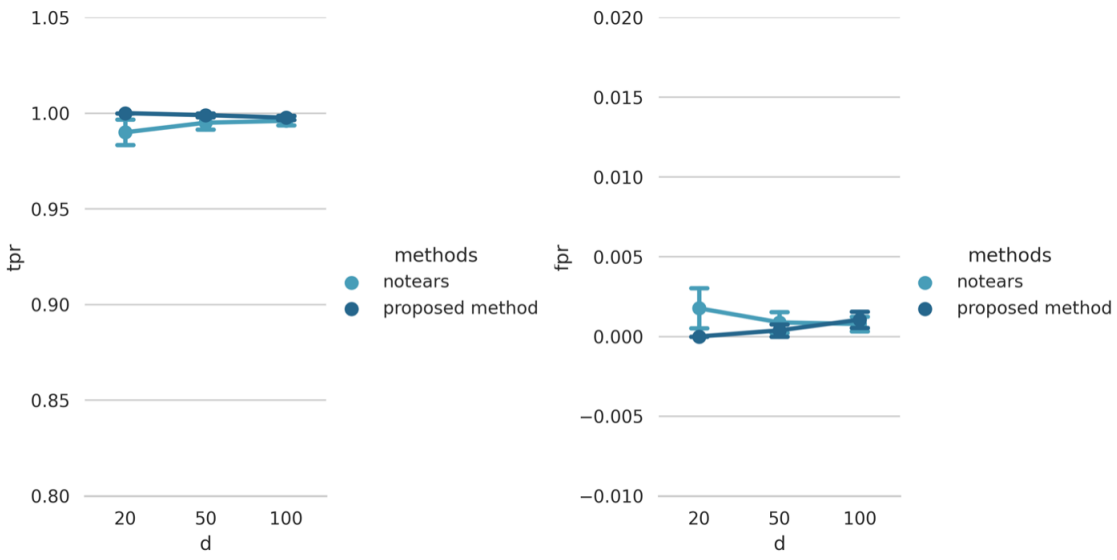


Figure 8.2: Results of qualitative study shows the proposed method has comparable accuracy as no-tears in terms of true positive rate (tpr) and false positive rate (fpr), $n = 20, d = \{20, 50, 100\}$

.

8.3.1.2 Structure learning

In this section, we quantitatively investigate how the performance of each model varies as the data samples and feature dimensions increase. The results in Figure 8.2 are obtained using a synthetic dataset with a data size of $n = 20$ and feature dimensions of $d \in \{20, 50, 100\}$. As the experiments with FGES can be time-consuming (taking over 12 hours) and memory-intensive when $d = 100$ even with small sample sizes, we only present the results for the proposed method and no-tears.

8.3.1.3 Model complexity

In this section, we conduct a quantitative analysis (see Figure 8.3) of the computational efficiency of each model on datasets of different sizes and feature dimensions. Specifically, we compare the time efficiency of computing the constraints of our proposed method with those of the baseline model notears. It is worth noting that this comparison serves as an experimental verification of our proposed constraint ($O(n^{2.37})$), which has a faster computation time than the constraint in notears

with a complexity of $O(n^3)$, as reported in [167].

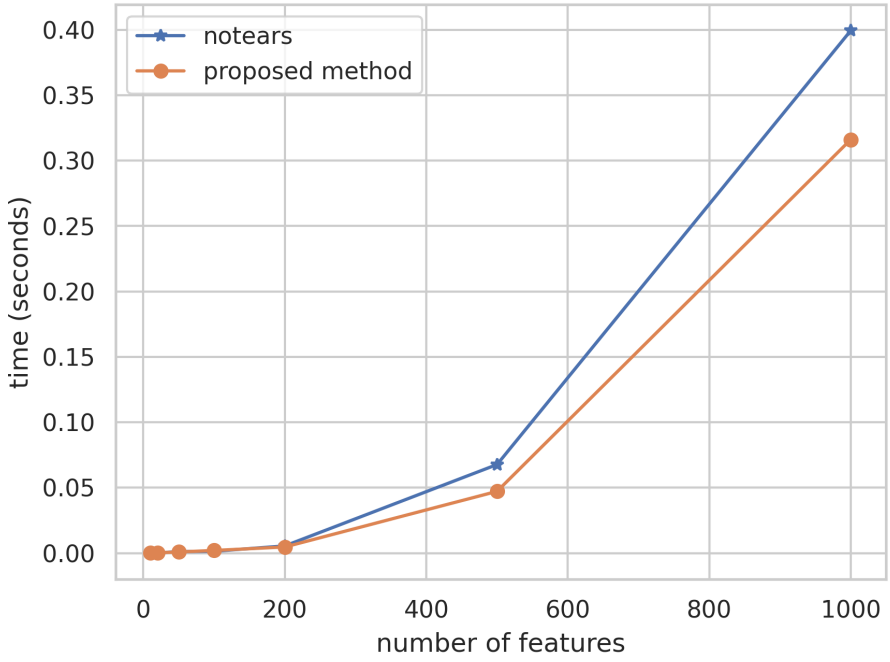


Figure 8.3: Comparison study shows that our proposed constraint ($O(n^{2.37})$), which has a faster computation time than the constraint in notears with a complexity of $O(n^3)$.

8.3.2 Real-world dataset

8.3.2.1 Protein Signaling Network Dataset

The Protein Signaling Network Dataset provides measurements of various phosphorylated protein and phospholipid components in thousands of primary human immune system cells. This data enables the reconstruction of the fundamental framework of a classical signaling network, which interconnects several crucial phosphorylated proteins involved in human T cell signaling [243]. The dataset comprises 7466 data samples, 11 features, and 20 edges.

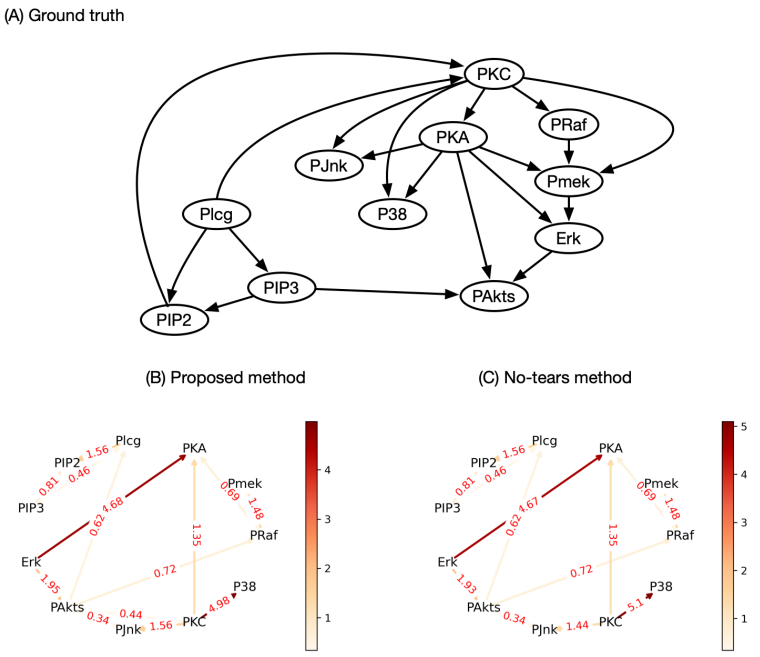


Figure 8.4: Comparison study on Protein Signaling Network Dataset

8.3.2.2 Case study

Figure 8.4 demonstrates that the proposed method detects an additional causal link compared to the baseline model (no-tears). Specifically, the proposed method identifies a causal relationship between *PKC* and *PAkts*, which suggests that these two proteins may be connected. It is worth noting that the ground truth includes two indirect causal paths from *PKC* to *PAkts*, as depicted in Figure 8.4 (A): *PKC*→*PRaf*→*Pmek*→*Erk*→*PAkts* and *PKC*→*PKA*→*PAkts*. Additionally, the proposed method identifies the causal connection between *Erk* and *Akt*, which was not well-established in the literature until confirmed by both machine learning modeling and experimental validation in this study [243]. Notably, the proposed method identifies this significant causal link automatically, without any prior domain knowledge.

8.4 Conclusion

In our research, we present a novel approach to causal discovery through the introduction of a new continuous constraint optimization method. Unlike previous methods that relied on an exponential constraint with high computational complexity (as mentioned in [167]), our proposed constraint offers significantly improved computational efficiency. Through our experimentation, we have demonstrated that our method achieves comparable performance to existing state-of-the-art causal discovery methods, while also requiring less computational time.

Building on the current research, our future work aims to enhance the objective function even further. By refining the underlying optimization process, we seek to unlock additional insights and improve the accuracy of causal discovery. We anticipate that these advancements will contribute to the development of more robust and efficient causal discovery techniques, with potential applications in various domains including healthcare data analytics.

CHAPTER 9

Conclusion and future work

9.1 Summary of the Thesis

In this thesis, we have explored various aspects of healthcare analytics and machine learning, with a particular emphasis on causal inference. In Chapter 1, we provided a brief introduction to the objectives, motivations, and contributions of this work. In Chapter 2, we discussed the background of data analytics in healthcare and presented a literature review on imbalanced learning and arrhythmia classification using deep transfer learning with electrocardiogram datasets.

Chapter 3 provided an overview of the concepts of causality, structural causal models, causal graphs, and intervention using do-calculus. Additionally, we explored the issues of spurious correlation and confounding, including the Simpson paradox, and discussed methods for discovering causal relationships from data.

In Chapter 4, we demonstrated the ability of the Sensing At-Risk Population system to provide deeper insights into the health conditions of patients by using wearable technology with sophisticated physical activity tracking algorithms. This study has potential applications in identifying patients who are at risk of re-admission to the hospital and monitoring the effectiveness of rehabilitation.

Chapter 5 introduced a new approach called WOT-Boost, which combines a Weighted Over-sampling Technique with an ensemble Boosting method to enhance the accuracy of minority data classification without compromising the accuracy of the majority class. In Chapter 6, we proposed a new deep transfer learning framework for arrhythmia classification using limited training data.

In Chapter 7, we proposed two principled, efficient, and highly effective CVR estimators for industrial CVR estimation from a causal perspective, accounting for the causes of missing not at

random. These models were based on the multi-task learning framework and mitigated the data sparsity issue.

Finally, in Chapter 8, we presented a novel approach to identifying causal relations in high-dimensional space.

In conclusion, this thesis has contributed to the development of advanced techniques for healthcare analytics and machine learning, with a particular emphasis on causal inference. The proposed methods have potential applications in identifying patients at risk of re-admission, enhancing the accuracy of minority data classification, and improving CVR estimation. The results of this work have the potential to make a significant impact on the field of healthcare analytics and machine learning.

9.2 Future work

Imbalanced learning is a common challenge in constructing medical predictive models. While the approach presented in Chapter 5 focuses on tabular datasets such as electronic health records, medical datasets frequently incorporate diverse modalities such as medical imaging and videos. As a result, it is necessary for new algorithms to be capable of learning from minority data samples that may not conform to tabular formats. Several techniques have been proposed to improve imbalanced learning when dealing with medical datasets that include medical imaging and videos. One of the most promising approaches is the use of generative adversarial networks (GANs). Given the success of GANs in generating artificial images, it is reasonable to explore their potential in oversampling minority classes by generating synthetic images.

Chapter 6 discusses arrhythmia classification and detection using a transfer learning model that fine-tunes a pre-trained model, ResNet-18, with a small dataset. In the future, we aim to develop a lightweight version of such classifiers that can be deployed on wearables like smartwatches. This requires an end-to-end predictive model that reads a single-lead ECG signal and generates outcomes. Such a model can enhance the capabilities of patient remote health monitoring systems that incorporate wearables.

Chapter 8 discusses a new approach to identifying causal relations in high-dimensional space. The proposed method formulates the causal discovery as a continuous constrained problem. A novel aspect of the approach is the improvement of the constraint formulation from a matrix exponential term proposed in work [167] to a polynomial term that is more efficient. However, the research assumes that the causal models are linear. We aim to explore non-linear causal relations. For instance, [168] has investigated the use of auto-encoders for modeling non-linear causal-effect relations.

REFERENCES

[1] Franz H. Messerli. Chocolate consumption, cognitive function, and nobel laureates. New England Journal of Medicine, 367(16):1562–1564, 2012. PMID: 23050509.

[2] Elias Bareinboim – Causal Data Science. `https://www.youtube.com/watch?v=dUsokjG4DHc`, 2019.

[3] Craig K Enders. Applied missing data analysis. Guilford press, 2010.

[4] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In Advances in neural information processing systems, pages 1277–1285, 2013.

[5] Judea Pearl and Dana Mackenzie. The book of why: the new science of cause and effect, chapter Beyond Adjustment: The Conquest of Mount Intervention, page 234. Basic Books, 2018.

[6] Judea Pearl et al. Causal inference in statistics: An overview. Statistics surveys, 3:96–146, 2009.

[7] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. The deconfounded recommender: A causal inference approach to recommendation. arXiv preprint arXiv:1808.06581, 2018.

[8] Judea Pearl. Causality: models, reasoning and inference, volume 29. Springer, 2000.

[9] Stephen L Morgan and Christopher Winship. Counterfactuals and causal inference. Cambridge University Press, 2015.

[10] Lakmini P Malasinghe, Naeem Ramzan, and Keshav Dahal. Remote patient monitoring: a comprehensive study. Journal of Ambient Intelligence and Humanized Computing, 10:57–76, 2019.

[11] Zachary Tran, Wenhao Zhang, Arjun Verma, Alan Cook, Dennis Kim, Sigrid Burruss, Ramin Ramezani, and Peyman Benharash. The derivation of an icd-10-based trauma-related mortality model utilizing machine learning. The Journal of Trauma and Acute Care Surgery, 2021.

[12] Bryan P Bednarski, Akash Deep Singh, Wenhao Zhang, William M Jones, Arash Naeim, and Ramin Ramezani. Temporal convolutional networks and data rebalancing for clinical length of stay and mortality prediction. Scientific Reports, 12(1):21247, 2022.

[13] Ramin Ramezani, Wenhao Zhang, Pamela Roberts, John Shen, David Elashoff, Zhuoer Xie, Annette Stanton, Michelle Eslami, Neil S Wenger, Jacqueline Trent, et al. Physical activity behavior of patients at a skilled nursing facility: Longitudinal cohort study. JMIR mHealth and uHealth, 10(5):e23887, 2022.

[14] Ramin Ramezani, Wenhao Zhang, Zhuoer Xie, John Shen, David Elashoff, Pamela Roberts, Annette Stanton, Michelle Eslami, Neil Wenger, Majid Sarrafzadeh, et al. A combination of indoor localization and wearable sensor–based physical activity recognition to assess older patients undergoing subacute rehabilitation: Baseline study results. JMIR mHealth and uHealth, 7(7):e14090, 2019.

[15] Haibo He and Edwardo A Garcia. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9):1263–1284, 2009.

[16] Wenhao Zhang, Ramin Ramezani, and Arash Naeim. Wotboost: Weighted oversampling technique in boosting for imbalanced learning. In 2019 IEEE International Conference on Big Data (Big Data), pages 2523–2531. IEEE, 2019.

[17] Paul D Allison. Missing data. Sage publications, 2001.

[18] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6):1–45, 2017.

[19] Valentin Amrhein, Sander Greenland, and Blake McShane. Scientists rise up against statistical significance. Nature, 567(7748):305–307, 2019.

[20] Adeline Lo, Herman Chernoff, Tian Zheng, and Shaw-Hwa Lo. Why significant variables aren't automatically good predictors. Proceedings of the National Academy of Sciences, 112(45):13892–13897, 2015.

[21] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. Causal inference in statistics: A primer, chapter The effects of Interventions, pages 53–55. John Wiley & Sons, 2016.

[22] David Hume. A treatise of human nature. Courier Corporation, 2003.

[23] Fujin Zhu. On Causal Discovery and Inference from Observational Data. PhD thesis, 2019.

[24] Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. Large-scale causal approaches to debiasing post-click conversion rate estimation with multi-task learning. In Proceedings of The Web Conference 2020, pages 2775–2781, 2020.

[25] Wenhao Zhang, Ramin Ramezani, and Arash Naeim. Causal inference in medicine and in health policy: A summary. In HANDBOOK ON COMPUTER LEARNING AND INTELLIGENCE: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation, pages 263–302. World Scientific, 2022.

[26] Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. Journal of Machine Learning Research, 11(1), 2010.

[27] Sandeep Yaramakala and Dimitris Margaritis. Speculative markov blanket discovery for optimal feature selection. In Fifth IEEE International Conference on Data Mining (ICDM'05), pages 4–pp. IEEE, 2005.

[28] Ramin Ramezani, Babak Moatamed, Arash Naeim, Majid Sarrafzadeh, et al. Subject assessment using localization, activity recognition and a smart questionnaire, March 2 2021. US Patent 10,937,547.

[29] Jennifer M Ortman, Victoria A Velkoff, Howard Hogan, et al. An aging nation: the older population in the united states. 2014.

[30] Babak Moatamed, Farhad Shahmohammadi, Ramin Ramezani, Arash Naeim, Majid Sarrafzadeh, et al. Low-cost indoor health monitoring system. In 2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN), pages 159–164. IEEE, 2016.

[31] Stephen J Brown. Remote health monitoring and maintenance system, December 14 2010. US Patent 7,853,455.

[32] Myung-kyung Suh, Chien-An Chen, Jonathan Woodbridge, Michael Kai Tu, Jung In Kim, Ani Nahapetian, Lorraine S Evangelista, and Majid Sarrafzadeh. A remote patient monitoring system for congestive heart failure. Journal of medical systems, 35:1165–1179, 2011.

[33] Michael Schwenk, Jane Mohler, Christopher Wendel, Karen D"Huyvetter, Mindy Fain, Ruth Taylor-Piliae, and Bijan Najafi. Wearable sensor-based in-home assessment of gait, balance, and physical activity for discrimination of frailty status: baseline results of the arizona frailty cohort study. Gerontology, 61(3):258–267, 2015.

[34] Nima Toosizadeh, Bijan Najafi, Eric M Reiman, Reine M Mager, Jaimeson K Veldhuizen, Kathy O'Connor, Edward Zamrini, and Jane Mohler. Upper-extremity dual-task function: an innovative method to assess cognitive impairment in older adults. Frontiers in aging neuroscience, 8:167, 2016.

[35] Bijan Najafi, Kamiar Aminian, Anisoara Paraschiv-Ionescu, François Loew, Christophe J Bula, and Philippe Robert. Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly. IEEE Transactions on biomedical Engineering, 50(6):711–723, 2003.

[36] Javad Razjouyan, Aanand D Naik, Molly J Horstman, Mark E Kunik, Mona Amirmazaheri, He Zhou, Amir Sharafkhaneh, and Bijan Najafi. Wearable sensors and the assessment of frailty among vulnerable older adults: an observational cohort study. Sensors, 18(5):1336, 2018.

[37] Michael K Ong, Patrick S Romano, Sarah Edgington, Harriet U Aronow, Andrew D Auerbach, Jeanne T Black, Teresa De Marco, Jose J Escarce, Lorraine S Evangelista, Barbara Hanna, et al. 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, 176(3):310–318, 2016.

[38] Susan Michie, Maartje M Van Stralen, and Robert West. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implementation science, 6(1):1–12, 2011.

[39] Kevin Bouchard, Ramin Ramezani, and Arash Naeim. Features based proximity localization with bluetooth emitters. In 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pages 1–5. IEEE, 2016.

[40] Kevin Bouchard, Mahir Rafi Eusufzai, Ramin Ramezani, and Arash Naeim. Generalizable spatial feature for human positioning based on bluetooth beacons. In 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pages 1–5. IEEE, 2016.

[41] Kevin Bouchard, Ramin Ramezani, Arash Naeim, et al. Evaluation of bluetooth beacons behavior. In 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pages 1–3. IEEE, 2016.

[42] Andrea Mannini, Stephen S Intille, Mary Rosenberger, Angelo M Sabatini, and William Haskell. Activity recognition using a single accelerometer placed at the wrist or ankle. Medicine and science in sports and exercise, 45(11):2193, 2013.

[43] Ling Bao and Stephen S Intille. Activity recognition from user-annotated acceleration data. In Pervasive Computing: Second International Conference, PERVASIVE 2004, Linz/Vienna, Austria, April 21-23, 2004. Proceedings 2, pages 1–17. Springer, 2004.

[44] A Bhattacharya, EP McCutcheon, E Shvartz, and JE Greenleaf. Body acceleration distribution and o2 uptake in humans during running and jumping. Journal of Applied Physiology, 49(5):881–887, 1980.

[45] Margarita S Treuth, Kathryn Schmitz, Diane J Catellier, Robert G McMurray, David M Murray, M Joao Almeida, Scott Going, James E Norman, and Russell Pate. Defining accelerometer thresholds for activity intensities in adolescent girls. Medicine and science in sports and exercise, 36(7):1259, 2004.

[46] John Staudenmayer, David Pober, Scott Crouter, David Bassett, and Patty Freedson. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. Journal of applied physiology, 2009.

[47] Emmanuel Munguia Tapia, Stephen S Intille, and Kent Larson. Activity recognition in the home using simple and ubiquitous sensors. In Pervasive Computing: Second International Conference, PERVASIVE 2004, Linz/Vienna, Austria, April 21-23, 2004. Proceedings 2, pages 158–175. Springer, 2004.

[48] SE Crouter. Validity of 10 electronic pedometers for measuring steps, distance, and energy cost. Med Sci Sports Exerc., 36:331–335, 2004.

[49] George Mammen, Sarah Gardiner, Arrani Senthinathan, Laura McClemont, Michelle Stone, and Guy Faulkner. Is this bit fit? measuring the quality of the fitbit step-counter. The Health & Fitness Journal of Canada, 5(4):30–39, 2012.

[50] Meredith A Case, Holland A Burwick, Kevin G Volpp, and Mitesh S Patel. Accuracy of smartphone applications and wearable devices for tracking physical activity data. Jama, 313(6):625–626, 2015.

[51] Yuri Feito, David R Bassett, and Dixie L Thompson. Evaluation of activity monitors in controlled and free-living environments. Medicine & Science in Sports & Exercise, 44(4):733–741, 2012.

[52] David R Bassett, Lindsay P Toth, Samuel R LaMunion, and Scott E Crouter. Step counting: a review of measurement considerations and health-related applications. Sports Medicine, 47:1303–1315, 2017.

[53] Foster Provost. Machine learning from imbalanced data sets 101. In Proceedings of the AAAI 2000 workshop on imbalanced data sets, volume 68, pages 1–3. AAAI Press, 2000.

[54] T Elhassan and M Aljurf. Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method.". 2016.

[55] Tom Fawcett and Foster J Provost. Combining data mining and machine learning for effective user profiling. In KDD, pages 8–13, 1996.

[56] Miroslav Kubat, Robert C Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. Machine learning, 30(2-3):195–215, 1998.

[57] Wenke Lee and Salvatore Stolfo. Data mining approaches for intrusion detection. 1998.

[58] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In Machine learning proceedings 1994, pages 148–156. Elsevier, 1994.

[59] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

[60] Erin L Allwein, Robert E Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of machine learning research, 1(Dec):113–141, 2000.

[61] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61:863–905, 2018.

[62] Aida Ali, Siti Mariyam Shamsuddin, and Anca L Ralescu. Classification with class imbalance problem: a review. Int. J. Advance Soft Compu. Appl, 7(3):176–204, 2015.

[63] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing, pages 878–887. Springer, 2005.

[64] Krystyna Napierala and Jerzy Stefanowski. Types of minority class examples and their influence on learning classifiers from imbalanced data. Journal of Intelligent Information Systems, 46(3):563–597, 2016.

[65] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pages 1322–1328. IEEE, 2008.

[66] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In Pacific-Asia conference on knowledge discovery and data mining, pages 475–482. Springer, 2009.

[67] Vishwa Karia, Wenhao Zhang, Arash Naeim, and Ramin Ramezani. Gensample: A genetic algorithm for oversampling in imbalanced datasets, 2019.

[68] Nguyen Thai-Nghe, DT Nghi, and Lars Schmidt-Thieme. Learning optimal threshold on resampling data to deal with class imbalance. In Proc. IEEE RIVF International Conference on Computing and Telecommunication Technologies, pages 71–76, 2010.

[69] Ivan Tomek. An experiment with the edited nearest-neighbor rule. IEEE Transactions on systems, Man, and Cybernetics, (6):448–452, 1976.

[70] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23(04):687–719, 2009.

[71] Tasadduq Imam, Kai Ming Ting, and Joarder Kamruzzaman. z-svm: an svm for improved classification of imbalanced data. In Australasian Joint Conference on Artificial Intelligence, pages 264–273. Springer, 2006.

[72] Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. Svms modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(1):281–288, 2009.

[73] Chu-Hong Hoi, Chi-Hang Chan, Kaizhu Huang, Michael R Lyu, and Irwin King. Biased support vector machine for relevance feedback in image retrieval. In 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541), volume 4, pages 3189–3194. IEEE, 2004.

[74] Larry M Manevitz and Malik Yousef. One-class svms for document classification. Journal of machine Learning research, 2(Dec):139–154, 2001.

[75] Debashree Devi, Saroj K Biswas, and Biswajit Purkayastha. Learning in presence of class imbalance and class overlapping by using one-class svm and undersampling technique. Connection Science, pages 1–38, 2019.

[76] Bianca Zadrozny and Charles Elkan. Learning and making decisions when costs and probabilities are both unknown. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 204–213. ACM, 2001.

[77] Dragos D Margineantu. Class probability estimation and cost-sensitive classification decisions. In European Conference on Machine Learning, pages 270–281. Springer, 2002.

[78] Thomas G Dietterich. Ensemble methods in machine learning. In International workshop on multiple classifier systems, pages 1–15. Springer, 2000.

[79] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In icml, volume 96, pages 148–156. Citeseer, 1996.

[80] Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In European conference on principles of data mining and knowledge discovery, pages 107–119. Springer, 2003.

[81] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 40(1):185–197, 2010.

[82] Hongyu Guo and Herna L Viktor. Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. ACM Sigkdd Explorations Newsletter, 6(1):30–39, 2004.

[83] Ashutosh Kumar, Roshan Bharti, Deepak Gupta, and Anish Kumar Saha. Improvement in boosting method by using rustboost technique for class imbalanced data. In Recent Developments in Machine Learning and Data Analytics, pages 51–66. Springer, 2019.

[84] Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. Ecg heartbeat classification: A deep transferable representation. In 2018 IEEE international conference on healthcare informatics (ICHI), pages 443–444. IEEE, 2018.

[85] Eduardo José da S Luz, William Robson Schwartz, Guillermo Cámara-Chávez, and David Menotti. Ecg-based heartbeat classification for arrhythmia detection: A survey. Computer methods and programs in biomedicine, 127:144–164, 2016.

[86] Taissir Fekih Romdhane and Mohamed Atri Pr. Electrocardiogram heartbeat classification based on a deep convolutional neural network and focal loss. Computers in Biology and Medicine, 123:103866, 2020.

[87] Eduardo Luz and David Menotti. How the choice of samples for building arrhythmia classifiers impact their performances. In 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 4988–4991. IEEE, 2011.

[88] Mi Hye Song, Jeon Lee, Sung Pil Cho, Kyoung Joung Lee, and Sun Kook Yoo. Support vector machine based arrhythmia classification using reduced features. 2005.

[89] MIT-BIH ECG database. https://physionet.org/content/mitdb/1.0.0/.

[90] Selcan Kaplan Berkaya, Alper Kursat Uysal, Efnan Sora Gunal, Semih Ergin, Serkan Gunal, and M Bilginer Gulmezoglu. A survey on ecg analysis. Biomedical Signal Processing and Control, 43:216–235, 2018.

[91] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature medicine, 25(1):65–69, 2019.

[92] Omid Sayadi and Mohammad Bagher Shamsollahi. Ecg denoising and compression using a modified extended kalman filter structure. IEEE transactions on biomedical engineering, 55(9):2240–2248, 2008.

[93] Cuiwei Li, Chongxun Zheng, and Changfeng Tai. Detection of ecg characteristic points using wavelet transforms. IEEE Transactions on biomedical Engineering, 42(1):21–28, 1995.

[94] Juan Pablo Martínez, Rute Almeida, Salvador Olmos, Ana Paula Rocha, and Pablo Laguna. A wavelet-based ecg delineator: evaluation on standard databases. IEEE Transactions on biomedical engineering, 51(4):570–581, 2004.

[95] Mohammed Bahoura, M Hassani, and M Hubin. Dsp implementation of wavelet transform for real time ecg wave forms detection and heart rate analysis. Computer methods and programs in biomedicine, 52(1):35–44, 1997.

[96] Jingshan Huang, Binqiang Chen, Bin Yao, and Wangpeng He. Ecg arrhythmia classification using stft-based spectrogram and convolutional neural network. IEEE access, 7:92871–92880, 2019.

[97] Chun-Cheng Lin and Chun-Min Yang. Heartbeat classification using normalized rr intervals and morphological features. Mathematical Problems in Engineering, 2014:1–11, 2014.

[98] Douglas A Coast, Richard M Stern, Gerald G Cano, and Stanley A Briller. An approach to cardiac arrhythmia analysis using hidden markov models. IEEE Transactions on biomedical Engineering, 37(9):826–836, 1990.

[99] Sajad Mousavi and Fatemeh Afghah. Inter-and intra-patient ecg heartbeat classification for arrhythmia detection: a sequence to sequence deep learning approach. In ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 1308–1312. IEEE, 2019.

[100] Philip De Chazal, Maria O'Dwyer, and Richard B Reilly. Automatic classification of heartbeats using ecg morphology and heartbeat interval features. IEEE transactions on biomedical engineering, 51(7):1196–1206, 2004.

[101] Daniel E Geer Jr. Correlation is not causation. IEEE Security & Privacy, 9(2):93–94, 2011.

[102] KE Havens. Correlation is not causation: a case study of fisheries, trophic state and acidity in florida (usa) lakes. Environmental Pollution, 106(1):1–4, 1999.

[103] David Hume. An enquiry concerning human understanding. In Seven masterpieces of philosophy, pages 191–284. Routledge, 2016.

[104] Alfredo Morabia. Epidemiological causality. History and philosophy of the life sciences, pages 365–379, 2005.

[105] Mark Parascandola. Causes, risks, and probabilities: probabilistic concepts of causation in chronic disease epidemiology. Preventive Medicine, 53(4-5):232–234, 2011.

[106] Tyler J VanderWeele and Ilya Shpitser. On the definition of a confounder. Annals of statistics, 41(1):196, 2013.

[107] Steven A Julious and Mark A Mullee. Confounding and simpson's paradox. Bmj, 309(6967):1480–1481, 1994.

[108] Paul W Holland and Donald B Rubin. On lord's paradox. Principals of modern psychological measurement, pages 3–25, 1983.

[109] Sander Greenland and Hal Morgenstern. Confounding in health research. Annual review of public health, 22(1):189–212, 2001.

[110] Paul Hünermund and Elias Bareinboim. Causal inference and data-fusion in econometrics. arXiv preprint arXiv:1912.09104, 2019.

[111] Zoraida Verde, Catalina Santiago, José Miguel Rodríguez González-Moro, Pilar de Lucas Ramos, Soledad López Martín, Fernando Bandrés, Alejandro Lucia, and Félix Gómez-Gallego. 'smoking genes': a genetic association study. PloS one, 6(10):e26668, 2011.

[112] James MacKillop, Ezemenari M Obasi, Michael T Amlung, John E McGeary, and Valerie S Knopik. The role of genetics in nicotine dependence: mapping the pathways from genome to syndrome. Current cardiovascular risk reports, 4(6):446–453, 2010.

[113] Judea Pearl. The do-calculus revisited. arXiv preprint arXiv:1210.4852, 2012.

[114] Robert R Tucci. Introduction to judea pearl's do-calculus. arXiv preprint arXiv:1305.5506, 2013.

[115] Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. arXiv preprint arXiv:1206.6831, 2012.

[116] Dan Geiger, Thomas Verma, and Judea Pearl. Identifying independence in bayesian networks. Networks, 20(5):507–534, 1990.

[117] Judea Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA, USA, pages 15–17, 1985.

[118] Judea Pearl. Bayesian networks, causal inference and knowledge discovery. UCLA Cognitive Systems Laboratory, Technical Report, 2001.

[119] David Heckerman, Christopher Meek, and Gregory Cooper. A bayesian approach to causal discovery. Computation, causation, and discovery, 19:141–166, 1999.

[120] Tyler J VanderWeele and Ilya Shpitser. On the definition of a confounder. Annals of statistics, 41(1):196, 2013.

[121] Sander Greenland, James M Robins, and Judea Pearl. Confounding and collapsibility in causal inference. Statistical science, pages 29–46, 1999.

[122] Clive R Charig, David R Webb, Stephen Richard Payne, and John E Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. Br Med J (Clin Res Ed), 292(6524):879–882, 1986.

[123] Kevin P Murphy. Machine learning: a probabilistic perspective. MIT press, 2012.

[124] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences, 113(27):7345–7352, 2016.

[125] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In Artificial Intelligence and Statistics, pages 100–108, 2012.

[126] Jeremy Berke. A statewide antibody study estimates that 21% of new york city residents have had the coronavirus, cuomo says, 2020.

[127] Cassie Kozyrkov. Were 21% of new york city residents really infected with the novel coronavirus?, 2020.

[128] Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. A causal perspective to unbiased conversion rate estimation on data missing not at random, 2019.

[129] Anthony B Ryder, Anna V Wilkinson, Michelle K McHugh, Katherine Saunders, Sumesh Kachroo, Anthony D'Amelio, Melissa Bondy, and Carol J Etzel. The advantage of imputation of missing income data to evaluate the association between income and self-reported health status (srh) in a mexican american cohort study. Journal of immigrant and minority health, 13(6):1099–1109, 2011.

[130] Donald B Rubin. Inference and missing data. Biometrika, 63(3):581–592, 1976.

[131] Judea Pearl and Karthika Mohan. Recoverability and testability of missing data: Introduction and summary of results. Available at SSRN 2343873, 2013.

[132] Karthika Mohan and Judea Pearl. Missing data from a causal perspective. In Workshop on Advanced Methodologies for Bayesian Networks, pages 184–195. Springer, 2015.

[133] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Gain: Missing data imputation using generative adversarial nets. arXiv preprint arXiv:1806.02920, 2018.

[134] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. Journal of statistical software, pages 1–68, 2010.

[135] Yi Deng, Changgee Chang, Moges Seyoum Ido, and Qi Long. Multiple imputation for general missing data patterns in the presence of high-dimensional data. Scientific reports, 6:21689, 2016.

[136] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. Psychological methods, 7(2):147, 2002.

[137] Roderick JA Little and Donald B Rubin. Statistical analysis with missing data, volume 793. John Wiley & Sons, 2019.

[138] Judea Pearl. On the testability of causal models with latent and instrumental variables. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pages 435–443. Morgan Kaufmann Publishers Inc., 1995.

[139] Martin Ford. Architects of Intelligence: The truth about AI from the people building it. Packt Publishing Ltd, 2018.

[140] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, page 1670–1679. JMLR.org, 2016.

[141] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, page 1137–1140, New York, NY, USA, 2018. Association for Computing Machinery.

[142] Arnaud De Myttenaere, Bénédicte Le Grand, Boris Golden, and Fabrice Rossi. Reducing offline evaluation bias in recommendation systems. In 23rd annual Belgian-Dutch Conference on Machine Learning (Benelearn 2014), pages 55–62, Bruxelles, Belgium, June 2014.

[143] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In Advances in Neural Information Processing Systems, pages 2643–2651, 2013.

[144] Han Zhu, Junqi Jin, Chang Tan, Fei Pan, Yifan Zeng, Han Li, and Kun Gai. Optimized cost per click in taobao display advertising. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 2191–2200, 2017.

[145] Yuriko Yamaguchi, Mimpei Morishita, Youichi Inagaki, Reyn Nakamoto, Jianwei Zhang, Junichi Aoi, and Shinsuke Nakajima. Web advertising recommender system based on estimating users' latent interests. In Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services, pages 42–49, 2016.

[146] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In Advances in Neural Information Processing Systems, pages 601–608, 2007.

[147] RJA Little and DB Rubin. Statistical analysis with missing data. wiley. New York, 2002.

[148] Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[149] Harald Steck. Training and testing of recommender systems on data missing not at random. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 713–722. ACM, 2010.

[150] Qingyao Ai, Keping Bi, Cheng Luo, Jiafeng Guo, and W Bruce Croft. Unbiased learning to rank with unbiased propensity estimation. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pages 385–394. ACM, 2018.

[151] Kuang-chih Lee, Burkay Orten, Ali Dasdan, and Wentong Li. Estimating conversion rate in display advertising from past erformance data. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 768–776. ACM, 2012.

[152] Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee. Billion-scale commodity embedding for e-commerce recommendation in alibaba. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 839–848. ACM, 2018.

[153] Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. Proceedings of Machine Learning Research, 2019.

[154] Zhifeng Hao, Hao Zhang, Ruichu Cai, Wen Wen, and Zhihao Li. Causal discovery on high dimensional data. Applied Intelligence, 42(3):594–607, 2015.

[155] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8):1226–1238, 2005.

[156] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. Artificial Intelligence, 182:1–31, 2012.

[157] George Kour and Raid Saabne. Fast classification of handwritten on-line arabic characters. In Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of, pages 312–318. IEEE, 2014.

[158] Craig Silverstein, Sergey Brin, Rajeev Motwani, and Jeff Ullman. Scalable techniques for mining causal structures. Data Mining and Knowledge Discovery, 4(2):163–192, 2000.

[159] Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. Algorithms for large scale markov blanket discovery. In FLAIRS conference, volume 2, pages 376–380, 2003.

[160] Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, and Bingyu Sun. Mining causal association rules. In 2013 IEEE 13th International Conference on Data Mining Workshops, pages 114–123. IEEE, 2013.

[161] Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, Bingyu Sun, and Saisai Ma. From observational studies to causal rule mining. ACM Transactions on Intelligent Systems and Technology (TIST), 7(2):1–27, 2015.

[162] Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. J. Mach. Learn. Res., 15(1):3741–3782, 2014.

[163] Daniela Marella et al. Pc complex: Pc algorithm for complex survey data. Technical report, Department of Economics-University Roma Tre, 2018.

[164] Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. IEEE/ACM transactions on computational biology and bioinformatics, 16(5):1483–1495, 2016.

[165] Ramakrishnan Srikant and Rakesh Agrawal. Mining generalized association rules, 1995.

[166] Wenhao Zhang, Ramin Ramezani, and Arash Naeim. An introduction to causal reasoning in health analytics. arXiv preprint arXiv:2105.04655, 2021.

[167] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. arXiv preprint arXiv:1803.01422, 2018.

[168] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In International Conference on Machine Learning, pages 7154–7163. PMLR, 2019.

[169] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In Proceedings of the IEEE international conference on computer vision, pages 2942–2950, 2017.

[170] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning-problems, methods and evaluation. ACM SIGKDD Explorations Newsletter, 22(1):18–33, 2020.

[171] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In international conference on machine learning, pages 1670–1679. PMLR, 2016.

[172] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. arXiv preprint arXiv:1703.06856, 2017.

[173] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. arXiv preprint arXiv:1805.05859, 2018.

[174] Leon Kopitar, Primoz Kocbek, Leona Cilar, Aziz Sheikh, and Gregor Stiglic. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports, 10(1):1–12, 2020.

[175] Minimum data set 3.0 public reports. cms.gov. https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/Minimum-Data-Set-3-0-Public-Reports. Accessed: 2023-05-27.

[176] Joseph J Locascio and Alireza Atri. An overview of longitudinal data analysis methods for neurological research. Dementia and geriatric cognitive disorders extra, 1(1):330–357, 2011.

[177] Zoë Tieges, Gillian Mead, Mike Allerhand, Fiona Duncan, Frederike Van Wijck, Claire Fitzsimons, Carolyn Greig, and Sebastien Chastin. Sedentary behavior in the first year after stroke: a longitudinal cohort study with objective measures. Archives of physical medicine and rehabilitation, 96(1):15–23, 2015.

[178] Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. Journal of the American statistical Association, 88(421):9–25, 1993.

[179] Guido Van Rossum et al. Python programming language. In USENIX annual technical conference, volume 41, pages 1–36. Santa Clara, CA, 2007.

[180] Wes McKinney. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.", 2012.

[181] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. the Journal of machine Learning research, 12:2825–2830, 2011.

[182] Regina Nuzzo. Statistical errors. Nature, 506(7487):150, 2014.

[183] Matthew Barrett, John Charles Snow, Megan C Kirkland, Liam P Kelly, Maria Gehue, Matthew B Downer, Jason McCarthy, and Michelle Ploughman. Excessive sedentary time during in-patient stroke rehabilitation. Topics in stroke rehabilitation, 25(5):366–374, 2018.

[184] Rafael Mesquita, Kenneth Meijer, Fabio Pitta, Helena Azcuna, Yvonne MJ Goërtz, Johannes MN Essers, Emiel FM Wouters, and Martijn A Spruit. Changes in physical activity and sedentary behaviour following pulmonary rehabilitation in patients with copd. Respiratory medicine, 126:122–129, 2017.

[185] Ailar Ramadi and Robert G Haennel. Sedentary behavior and physical activity in cardiac rehabilitation participants. Heart & Lung, 48(1):8–12, 2019.

[186] Henri Vähä-Ypyä, Tommi Vasankari, Pauliina Husu, Jaana Suni, and Harri Sievänen. A universal, accurate intensity-based classification of different physical activities using raw data of accelerometer. Clinical physiology and functional imaging, 35(1):64–70, 2015.

[187] Mary E Rosenberger, William L Haskell, Fahd Albinali, Selene Mota, Jason Nawyn, and Stephen Intille. Estimating activity and sedentary behavior from an accelerometer on the hip or wrist. Medicine and science in sports and exercise, 45(5):964, 2013.

[188] ACRMDOG Godfrey, Richard Conway, David Meagher, and Gearoid ÓLaighin. Direct measurement of human movement by accelerometry. Medical engineering & physics, 30(10):1364–1386, 2008.

[189] Der-Chiang Li, Chiao-Wen Liu, and Susan C Hu. A learning method for the class imbalance problem with medical data sets. Computers in biology and medicine, 40(5):509–518, 2010.

[190] Greta Nasi, Maria Cucciniello, and Claudia Guerrazzi. The role of mobile technologies in health care processes: the case of cancer supportive care. Journal of medical Internet research, 17(2):e26, 2015.

[191] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1):119–139, 1997.

[192] Catherine L Blake and Christopher J Merz. Uci repository of machine learning databases, 1998, 1998.

[193] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[194] M Elter, R Schulz-Wendtland, and T Wittenberg. The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process. Medical physics, 34(11):4164–4172, 2007.

[195] Michel Verleysen, J-L Voz, Philippe Thissen, and J-D Legat. A statistical neural network for high-dimensional vector classification. In Proceedings of ICNN'95-International Conference on Neural Networks, volume 2, pages 990–994. IEEE, 1995.

[196] I-Cheng Yeh, King-Jang Yang, and Tao-Ming Ting. Knowledge discovery on rfm model using bernoulli sequence. Expert Systems with Applications, 36(3):5866–5871, 2009.

[197] J Sayyad Shirabad and Tim J Menzies. The promise repository of software engineering databases. School of Information Technology and Engineering, University of Ottawa, Canada, 24, 2005.

[198] Jock A Blackard and Denis J Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. Computers and electronics in agriculture, 24(3):131–151, 1999.

[199] John A Swets. Measuring the accuracy of diagnostic systems. Science, 240(4857):1285–1293, 1988.

[200] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In Icml, pages 179–186, 1997.

[201] Gabriel Garcia, Gladston Moreira, David Menotti, and Eduardo Luz. Inter-patient ecg heartbeat classification with temporal vcg optimized by pso. Scientific reports, 7(1):10543, 2017.

[202] Varun Gupta and Monika Mittal. Qrs complex detection using stft, chaos analysis, and pca in standard and real-time ecg databases. Journal of The Institution of Engineers (India): Series B, 100:489–497, 2019.

[203] Sakshi Ahuja, Bijaya Ketan Panigrahi, Nilanjan Dey, Venkatesan Rajinikanth, and Tapan Kumar Gandhi. Deep transfer learning-based automated detection of covid-19 from lung ct scan slices. Applied Intelligence, 51:571–585, 2021.

[204] A Rajkumar, M Ganesan, and R Lavanya. Arrhythmia classification on ecg using deep learning. In 2019 5th international conference on advanced computing & communication systems (ICACCS), pages 365–369. IEEE, 2019.

[205] Yichao Lu, Ruihai Dong, and Barry Smyth. Coevolutionary recommendation model: Mutual learning between ratings and reviews. In Proceedings of the 2018 World Wide Web Conference, pages 773–782. International World Wide Web Conferences Steering Committee, 2018.

[206] Prem K Gopalan, Laurent Charlin, and David Blei. Content-based recommendations with poisson factorization. In Advances in Neural Information Processing Systems, pages 3176–3184, 2014.

[207] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: a factorization-machine based neural network for ctr prediction. arXiv preprint arXiv:1703.04247, 2017.

[208] Steffen Rendle. Factorization machines. In 2010 IEEE International Conference on Data Mining, pages 995–1000. IEEE, 2010.

[209] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. Fast matrix factorization for online recommendation with implicit feedback. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pages 549–558. ACM, 2016.

[210] Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. Multi-task feature learning for knowledge graph enhanced recommendation. In The World Wide Web Conference, pages 2000–2010. ACM, 2019.

[211] Guy Hadash, Oren Sar Shalom, and Rita Osadchy. Rank and rate: multi-task learning for recommender systems. In Proceedings of the 12th ACM Conference on Recommender Systems, pages 451–454. ACM, 2018.

[212] Yabo Ni, Dan Ou, Shichen Liu, Xiang Li, Wenwu Ou, Anxiang Zeng, and Luo Si. Perceive your users in depth: Learning universal user representations from multiple e-commerce tasks. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 596–605. ACM, 2018.

[213] Guido W Imbens and Donald B Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.

[214] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4):1161–1189, 2003.

[215] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3):399–424, 2011.

[216] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, page 1097–1104, Madison, WI, USA, 2011. Omnipress.

[217] Karel Vermeulen and Stijn Vansteelandt. Bias-reduced doubly robust estimation. Journal of the American Statistical Association, 110(511):1024–1036, 2015.

[218] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In Jennifer Dy and Andreas Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 1447–1456, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[219] Gary M Weiss. Mining with rarity: a unifying framework. ACM Sigkdd Explorations Newsletter, 6(1):7–19, 2004.

[220] Wenhao Zhang, Ramin Ramezani, and Arash Naeim. WOTBoost: Weighted oversampling technique in boosting for imbalanced learning. arXiv e-prints, page arXiv:1910.07892, Oct 2019.

[221] Tom Fawcett. An introduction to roc analysis. Pattern Recognition Letters, 27(8):861–874, 2006.

[222] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018.

[223] Ronald A Fisher. Statistical methods and scientific inference. Hafner Publishing Co., 1956.

[224] Anne Whitehead and John Whitehead. A general parametric approach to the meta-analysis of randomized clinical trials. Statistics in Medicine, 10(11):1665–1677, 1991.

[225] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In Proceedings of the conference on fairness, accountability, and transparency, pages 349–358, 2019.

[226] Zhongqi Yue, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Transporting causal mechanisms for unsupervised domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8599–8608, 2021.

[227] Max Chickering, David Heckerman, and Chris Meek. Large-sample learning of bayesian networks is np-hard. Journal of Machine Learning Research, 5, 2004.

[228] Dennis Wei, Tian Gao, and Yue Yu. Dags with no fears: A closer look at continuous optimization for learning bayesian networks. arXiv preprint arXiv:2010.09133, 2020.

[229] David Maxwell Chickering. Learning bayesian networks is np-complete. In Learning from data, pages 121–130. Springer, 1996.

[230] David Maxwell Chickering. Optimal structure identification with greedy search. Journal of machine learning research, 3(Nov):507–554, 2002.

[231] Ashlynn N Fuccello, Daniel Y Yuan, Panayiotis V Benos, and Vineet K Raghu. Improving constraint-based causal discovery from moralized graphs.

[232] Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. A graph autoencoder approach to causal structure learning. arXiv preprint arXiv:1911.07420, 2019.

[233] Frank Harary, Robert Zane Norman, and Dorwin Cartwright. Structural models: An introduction to the theory of directed graphs. Wiley, 1965.

[234] Virginia Vassilevska Williams. Multiplying matrices in o (n2. 373) time. preprint, pages 0–105698, 2014.

[235] Kirk Baker. Singular value decomposition tutorial. The Ohio State University, 24, 2005.

[236] Michel Fortin and Roland Glowinski. Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems. Elsevier, 2000.

[237] Dimitri P Bertsekas. Nonlinear programming. Journal of the Operational Research Society, 48(3):334–334, 1997.

[238] Brian Beavis and Ian Dobbs. Optimisation and stability theory for economic analysis. Cambridge university press, 1990.

[239] Stephen P Boyd and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.

[240] Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. International journal of data science and analytics, 3(2):121–129, 2017.

[241] Sourav Chatterjee and SR Srinivasa Varadhan. The large deviation principle for the erdős-rényi random graph. European Journal of Combinatorics, 32(7):1000–1017, 2011.

[242] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. The elements of statistical learning: data mining, inference, and prediction, volume 2. Springer, 2009.

[243] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. Science, 308(5721):523–529, 2005.