

To the Graduate Council:

I am submitting herewith a dissertation written by Benjamin Smith entitled “Insights into the Application of Deep Reinforcement Learning in Healthcare and Materials Science.” I have examined the final paper copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Data Science and Engineering.

Anahita Khojandi, Major Professor

We have read this dissertation
and recommend its acceptance:

Rama Vasudevan

Russell Zaretski

Amir Sadovnik

Accepted for the Council:

Dixie Thompson

Vice Provost and Dean of the Graduate School

To the Graduate Council:

I am submitting herewith a dissertation written by Benjamin Smith entitled “Insights into the Application of Deep Reinforcement Learning in Healthcare and Materials Science.” I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Data Science and Engineering.

Anahita Khojandi, Major Professor

We have read this dissertation
and recommend its acceptance:

Rama Vasudevan

Russell Zaretzki

Amir Sadovnik

Accepted for the Council:

Dixie Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Insights into the Application of Deep Reinforcement Learning in Healthcare and Materials Science

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Benjamin Smith

August 2023

© by Benjamin Smith, 2023
All Rights Reserved.

Acknowledgements

I would first like to thank my advisors, Dr. Anahita Khojandi and Dr. Rama Vasudevan. Their knowledge, support, and guidance were crucial to this research and the completion of my dissertation. I would also like to thank my family, friends, and especially my dog, Mac, for her tireless support and encouragement during this time.

Research in Chapters 1 & 2 is supported by Science Alliance, The University of Tennessee, and the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy.

Research in Chapter 3 is supported by Science Alliance and The University of Tennessee. The material in Chapter 3 is reprinted from [Smith et al. \(2023\)](#) with permission from Theime publishers: Smith, B., Van Steelandt, S., and Khojandi, A. (2023). Evaluating the Impact of Health Care Data Completeness for Deep Generative Models. *Methods of Information in Medicine*, 62(1-02), 31–39. <https://doi.org/10.1055/a-2023-9181>. This research was performed alongside Senne Van Steelandt.

Research in Chapter 4 is supported by the Center for Nanophase Materials Science (CNMS), which is a US Department of Energy, Office of Science User Facility at Oak Ridge National Laboratory.

Abstract

Reinforcement learning (RL) is a type of machine learning designed to optimize sequential decision-making. While controlled environments have served as a foundation for RL research, due to the growth in data volumes and deep learning methods, it is now increasingly being applied to real-world problems. In our work, we explore and attempt to overcome challenges that occur when applying RL to solve problems in healthcare and materials science.

First, we explore how issues in bias and data completeness affect healthcare applications of RL. To understand how bias has already been considered in this area, we survey the literature for existing methods of bias mitigation related to healthcare data and RL. Based on this review, we conduct our own study where RL is used to learn heparin treatment plans from specific patient populations to demonstrate how healthcare data can affect RL policies and cause bias. Additionally, because synthetic samples are often used to improve the utility of incomplete datasets, we evaluate the impact of data completeness in both panel and longitudinal medical data on the ability of deep generative models to generate similar synthetic samples to augment existing medical datasets.

Next, we focus on materials science applications where RL can potentially be used to automate certain processes. In scanning probe microscopy (SPM), electrical charges applied near ferroelectric domain walls cause the domain wall structure to change. In order to apply RL to control domain walls, we develop a dynamics model trained with SPM images and known physical properties to predict domain wall

displacements following an electrical pulse. This dynamics model is then integrated into an environment used to train RL agents that learn how to selectively apply charges that alter domain walls towards a particular structure.

Table of Contents

1	Bias in Reinforcement Learning: A Review in Healthcare Applications	1
1.1	Introduction	3
1.2	Background	5
1.2.1	RL Fundamentals	5
1.2.2	Electronic Health Records (EHRs)	6
1.2.3	RL in healthcare	8
1.2.4	Challenges	10
1.3	Literature Review	11
1.3.1	Data Management in Statistical Learning to Reduce Bias	13
1.3.2	Algorithm Design to Reduce Bias	17
1.3.3	Clinical Implementation of Treatment Planning Models to Reduce Bias	21
1.4	Conclusions	25
2	Uncovering Bias in Reinforcement Learning for Heparin Treatment Planning	26
2.1	Introduction	28
2.1.1	Background	29
2.1.2	Related Works	31
2.2	Motivation	32

2.3	Objectives and Contributions	34
2.4	Methodology	35
2.4.1	Dataset and Patient Cohorts	35
2.4.2	RL Problem Formulation	36
2.4.3	Evaluation	42
2.5	Results	43
2.5.1	Gender-based cohorts	43
2.5.2	Weight-based cohorts	45
2.6	Discussion	48
2.7	Conclusions	49
3	Evaluating the Impact of Healthcare Data Completeness for Deep Generative Models	51
3.1	Introduction	53
3.1.1	Motivation	55
3.1.2	Objectives	56
3.2	Methods	57
3.2.1	Dataset Description	57
3.2.2	Deep Generative Models	58
3.2.3	Model Evaluation	60
3.3	Results	61
3.4	Discussion	65
3.5	Conclusions	66
4	Towards Automating Ferroelectric Domain Wall Control using Reinforcement Learning	67
4.1	Introduction	69
4.2	Overall workflow	71
4.3	Dynamics Model	73
4.3.1	Physics based loss function	75

4.3.2	Dynamics model results	76
4.4	Reinforcement Learning Environment	79
4.4.1	Reinforcement Learning Results	81
4.5	Conclusions	83
Vita		115

List of Tables

1.1	Summary of related works reviewed and their particular areas of contribution and relevance to bias mitigation.	23
2.1	Gender-based cohorts used in this study.	37
2.2	Weight-based cohorts used in this study.	37
3.1	Summary of the features in the sample real-world datasets.	59

List of Figures

1.1	An overview of the RL framework.	9
1.2	RL in a healthcare setting.	9
2.1	Distribution of the weight of patients in the dataset.	37
2.2	Rewards for men’s and women’s cohorts.	46
2.3	Rewards for weight-based cohorts.	46
3.1	The overview of the study framework.	59
3.2	Study design. [GAN: Generative adversarial network, VAE: Variational autoencoder, Transformer]	63
3.3	MMAE scores with the heart rate dataset using GAN, VAE, and Transformer models from both perspectives of missingness and subset rates.	63
3.4	MMAE scores with the heart rate dataset using GAN, VAE, and Transformer models from both perspectives of missingness and subset rates.	64
4.1	Overall workflow combining experiments with computation.	74
4.2	Overall workflow of the surrogate model	77
4.3	Predicted displacements for local domain wall regions using the phase field simulations	77
4.4	Predicted displacements for local domain wall regions using the dynamics model	80

4.5	Overall design of the A2C model	82
4.6	The final domain wall structures are shown after each RL policy is applied	84

Chapter 1

Bias in Reinforcement Learning: A Review in Healthcare Applications

A version of this chapter was originally submitted to ACM Computing Surveys by Benjamin Smith, Dr. Khojandi, and Dr. Vasudevan:

Smith B, Khojandi A, Vasudevan R. (2022) *Bias in Reinforcement Learning: A Review in Healthcare Applications*. [Manuscript Submitted for Publication] Bredeesen Center, University of Tennessee.

Abstract

Reinforcement learning (RL) can assist in medical decision-making using patient data collected in electronic health record (EHR) systems. RL, a type of machine learning, can use these data to develop treatment policies. However, RL models are typically trained using imperfect retrospective EHR data. Therefore, if care is not taken in training, RL policies can propagate existing bias in healthcare. Literature that considers and addresses the issues of bias and fairness in sequential decision making are reviewed. The major themes to mitigate bias that emerge relate to (1) data management; (2) algorithmic design; and (3) clinical understanding of the resulting policies.

1.1 Introduction

Reinforcement learning (RL) is a branch of machine learning (ML) focused on sequential decision-making over time. In RL, an ‘agent’ interacts with an ‘environment,’ and takes ‘actions’ that can affect the ‘state’ of the environment, and in turn, receives ‘rewards.’ The goal of RL is to learn an action strategy or ‘policy,’ i.e., a mapping of states to actions, that maximizes the total reward [Li \(2017\)](#). RL has traditionally been used in robotics and in gameplay, where the agent tries to navigate through or control actions within a well-defined game environment such as Pong or Space Invaders [Levine et al. \(2020\)](#); [Mnih et al. \(2013\)](#). Here, the agent attempts to learn the best action at each time point that leads to winning the game or conversely, maximizing the reward (total points). While games have served as a foundation for RL research, it is now increasingly being applied to real-world problems that are focused on sequential decision-making, given the growth of data volumes that make learning feasible.

Recently, RL has shown great promise in healthcare applications, and in particular, in medical decision-making. In healthcare, the goal for clinicians is generally to identify a treatment policy that maximizes the patients’ benefit, e.g., total expected quality-adjusted lifetime (QALY). In the past, developing such treatment policies mostly relied on physician intuition and simple rule-based methods developed through clinical trials [Bekker \(2015\)](#). However, through the big data revolution, in many cases, researchers now have access to temporal measurements of different patient metrics over time that can be used to learn treatment policies using data-driven approaches. This is mostly enabled through the digital records of patient health information that are collected at the bedside in hospitals, especially in the data-rich intensive care units (ICUs), and stored in electronic health record (EHR) systems. The data compiled in these systems are either entered by practitioners or other personnel, or are directly pulled into the EHRs using other means such as bedside monitors. The corresponding data track patient health status, the interventions made by practitioners, and the consequent patient responses to these interventions.

Together, this information can create useful datasets of retrospective temporal data to train RL models [Landi et al. \(2020\)](#).

Many RL studies exist that use EHR and other health data on an array of diseases, including AIDS [Parbhoo et al. \(2017\)](#), sepsis management [Riachi et al. \(2021\)](#); [Liu et al. \(2020b\)](#), Parkinson’s disease [Baucum et al. \(2021\)](#), implanted cardiac device follow-up care [Khojandi et al. \(2018, 2014\)](#), and Heparin dosing [Baucum et al. \(2020\)](#). In these studies, generally, RL agents leverage historical data to devise treatment policies that can be generalized to new patients. While the RL-prescribed policies are often not expected to be immediately and directly used in clinical settings, in the future they can provide new insights for practitioners and augment factors considered in medical decision-making.

Although using RL in healthcare has the potential to provide new insights and improve patient outcomes, care should be taken to avoid *algorithmic bias*. Algorithms that make decisions differently among patient groups may exhibit algorithmic bias [Akter et al. \(2021\)](#). Recent research highlights the potential for algorithmic bias in supervised classification [Lin et al. \(2019\)](#). Some prominent examples include misidentifying faces of individuals from a certain race [Kärkkäinen and Joo \(2019\)](#) or Warfarin dosing classification based on patient ethnicity [Syn et al. \(2018\)](#); [Sharabiani et al. \(2015\)](#). For ML-based classification problems, instances of algorithmic bias can often be readily observed due to the non-temporal nature of these problems [Pessach and Shmueli \(2022\)](#). As a branch of ML, RL problems are not immune to algorithmic bias, and yet, the issue of bias in RL has received considerably less attention in the literature. Here, bias is less obvious, but equally concerning as it can impact the course of treatment for patients. Hence, there is a timely need to investigate bias in RL and discuss the potential methods to identify and mitigate bias in practice.

1.2 Background

Here, we discuss the fundamentals of RL and training RL with EHR data. We also describe applications of RL in healthcare and discuss active challenges.

1.2.1 RL Fundamentals

ML, an artificial intelligence (AI) method, can be implemented to detect patterns in data. This is largely possible through leveraging different function approximation methods [François-Lavet et al. \(2018\)](#). Within ML exist different techniques and of these methods, RL applies to sequential decision making [Li \(2017\)](#).

The main components of the RL system are an agent (or agents) and an environment. The agent(s) acts in the environment and learns the relationship between states and actions to maximize some rewards. In healthcare, the state of the environment can capture patient health state, e.g., blood pressure or the state of all vitals, and the action is the intervention performed, e.g., the amount and timing of intravenous (IV) medicine administration. Consequently, the rewards relate to the health outcomes of the patient, e.g., a successful discharge or normal vital sign measurements.

To understand how an agent learns an environment and maximizes rewards, it is useful to understand the formal mathematical problem. Markov decision processes (MDPs) formalize the sequential agent and environmental interactions [Otterlo and Wiering \(2012\)](#); [Sutton and Barto \(2018\)](#). A fully observable MDP is defined by a tuple of states, actions, transition function, reward function, and discount factor, $\langle S, A, P, R, \gamma \rangle$. States, $s_t \in S$, and actions, $a_t \in A$, are used to create a transition probability function from the current state s_t to next state s_{t+1} and receive reward r after taking action a_t . This is given by $P(s_{t+1}, r | s_t, a_t)$ a conditional probability for each state and action pair [Littman \(1994\)](#). The agent receives a reward (or incurs a cost) as a result of taking an action in a given state. The reward function $R(s_t, a_t) : S \times A \mapsto \mathbb{R}^1$ returns the immediate reward for each state and action pair [Yu et al. \(2021\)](#). By exploring states and actions, the agent attempts to find a policy

to take an action in each state that maximizes the reward function. Specifically, the agent seeks to act to maximize the expected sum of discounted rewards where the discount factor, $\gamma \in [0, 1]$, is used to determine the trade-off between valuing the immediate reward of taking an action over future expected rewards under the current policy. Figure 1 shows the overview of the RL framework.

RL models can be applied and solved in different ways. An RL algorithm that can use its own policy to generate samples is considered ‘on-policy.’ In on-policy learning, the agent interacts with the environment according to policy, π , that it attempts to improve. Otherwise, algorithms that learn from actions of previous policies are considered ‘off-policy’ as they use other policies to develop a new one. Similarly, ‘off-line’ RL occurs with no additional data collection, and when implemented, it solely uses historical/retrospective data. ‘Off-line’ RL works with a fixed data set, while ‘on-line’ learning is able to learn from data as they are collected. The distributional shift between the historical data and the real environment presents a major challenge in ‘off-line’ RL and oftentimes results in poor estimations of rewards for RL policies. However, some recent progress has been made [Levine et al. \(2020\)](#).

1.2.2 Electronic Health Records (EHRs)

EHRs digitally store clinical data regarding patient health and can be leveraged by RL to evaluate past treatments and prescribe future treatment interventions for new patients [Li et al. \(2021\)](#); [Schaefer and Bailey \(2005\)](#). EHRs generally provide rich data, collected across different patients with a wide range of characteristics and comorbidities. In addition, they typically include longitudinal data, which can enable the prediction of patient health state evolution under various interventions [Wu et al. \(2010\)](#); [Gold et al. \(2017\)](#).

EHRs usually contain high-dimensional time series data and static patient information with heterogeneous data types [Pollard et al. \(2018\)](#). These datasets can describe patient demographic information and known pre-existing conditions across

patients. Dynamic state info such as vital sign measurements, diagnostic codes, physician observations, and the potential treatment information is also often recorded [Asan and Montague \(2012\)](#); [Liu et al. \(2020a\)](#); [Rodriguez et al. \(2021\)](#). While some collection is automatic and includes objective data (e.g., data collected directly from heart rate monitors), other information is dependent on medical personnel. Manual data collection is prone to human factor errors [Carayon et al. \(2014\)](#); [Rosenbloom et al. \(2010\)](#).

Various EHR datasets are currently publicly available and have been successfully used in the past to develop RL models. Some of these datasets focus on the data-rich environments of ICU, e.g., MIMIC [Johnson et al. \(2016\)](#) and the eICU dataset [Pollard et al. \(2018\)](#). The ICU is a particularly suitable setting for collecting high-frequency data and training and deploying RL models, due to continuous monitoring of patients (unless patients are moved for testing or imaging, etc.). In particular, these datasets have been heavily used in off-line RL research for improving treatment planning; they allow for examining past treatment actions and patient state information to evaluate current practices and explore potential improvements.

While EHRs can enable RL modeling, there are often caveats. Often the time increments for recording patient information is not consistent. Thus, data often contains many missing values, which makes working with the data more challenging and may require imputation of missing values [Beaulieu-Jones et al. \(2017\)](#). Also, certain identifiers of patients are not always included because of ethical or privacy concerns [Pollard et al. \(2018\)](#). EHRs also are affected by external factors not demonstrable in the data including logistical constraints for the hospital and physician availability for any given patient. In this way, external bias is introduced to EHRs data. There are efforts towards standardization of EHRs that include social determinants of health, but this is not yet widely available [Rea et al. \(2012\)](#). Due to these issues, learning treatment policies from EHRs should consider how the data can differ from those of the actual patient populations, and care must be taken not to introduce bias into the RL-prescribed treatment policies.

1.2.3 RL in healthcare

When using RL in healthcare applications, the state of the environment is generally set up to capture the patient’s health state, provided through EHRs. The most comprehensive states provide the best description of the patient and can prevent mistaken correlations of variables. Actions are then whether or not some treatment is provided at a certain level at a specific time. Transition probabilities are learned based on the observed transitions in the data from certain states to another, given the actions. Given data availability, states and actions can be defined in a very granular manner. However, if needed, patient states can also be clustered to approximate policies for patients in similar medical states [Vyas et al. \(2020\)](#). In addition, while discrete actions are more widely used in practice, continuous actions are more representative of certain medical treatments [Riachi et al. \(2021\)](#). Lastly, the agent attempts to learn interventions in each state to maximize rewards to improve the long-term health of a patient. A reward could be for maintaining some vital measurement within a certain range or for the patient getting discharged from the ICU to the floor, among others.

In healthcare, generally, the goal is to develop a *clinician-in-the-loop* framework where RL acts as a decision support system [Liu et al. \(2020a\)](#). That is, RL is trained to prescribe actions, given the observed state of the patient in real time. However, treatment decisions are ultimately left to clinicians at the bedside rather than being automatically acted upon by the agent [Tang et al. \(2020\)](#). Figure 1.2 provides a schematic of this approach.

There have been numerous successful studies, adapting different RL techniques for healthcare applications [Liu et al. \(2018\)](#). This includes developing RL models that could be implemented in the ICU to improve glucose control of septic patients [Weng et al. \(2017\)](#); [Raghu et al. \(2017\)](#), and to recommend the dosing of heparin, an anticoagulant medication [Ghassemi et al. \(2014\)](#); [Nemati et al. \(2016\)](#); [Baucum et al. \(2020\)](#). Similarly, in the outpatient setting, RL has been paired with wearable sensor data

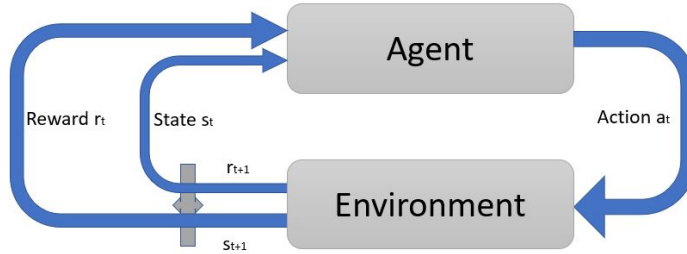


Figure 1.1: An overview of the RL framework. An agent in state s_t takes action a_t . The environment then transitions to state s_{t+1} and receives the reward r_{t+1} . The process then continues starting from the new state.

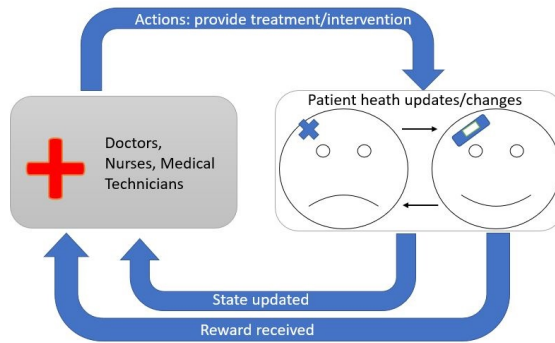


Figure 1.2: RL in a healthcare setting. Similar to the traditional RL diagram, a clinician-in-the-loop RL framework is shown. Clinicians take an action in the form of treatment or testing that causes the patient's state to transition. The patient's health could then change. Accordingly, the updated state and reward are returned to clinicians who repeat the process and take the appropriate actions at each time step.

that track motor functions in response to taking medications to optimize medication regimen for Parkinson’s disease patients [Watts et al. \(2020\)](#); [Baucum et al. \(2021\)](#).

1.2.4 Challenges

RL can play an important role in creating patient-specific treatment models. However, learning such data-driven models come with certain unique challenges. In creating a treatment policy, an RL agent needs sufficient data to describe patient states to learn patient-specific transitions. But healthcare data are often retrospective and clearly, running live experiments where the RL agent directly interacts with patients to collect more data is not ethically possible. Therefore, RL treatment models are typically trained in an off-line setting using existing EHRs.

As discussed in Section 1.2.2, the data used to train the model might not truly represent the patient population being treated for said disease. This directly relates to the ‘domain shift’ problem in off-line RL; a potential patient population shift would decrease effectiveness of RL treatment policies due to overestimation of performance [Swazinna et al. \(2021\)](#).

RL also deserves attention because it can make decisions based on more factors than considerable by a clinician [Li et al. \(2021\)](#); however, this makes some decisions less understandable with human intuition. Explaining why RL recommends certain policies is difficult. Therefore, understanding whether the resulting patient-specific policies are accurate and equitable when put to use is extremely challenging [Puiutta and Veith \(2020\)](#); [Kelly et al. \(2019\)](#). As of 2018, the EU protects individuals’ rights to explanation in automated decision-making [Goodman and Flaxman \(2017\)](#). RL policies should be explainable to people with diverse backgrounds [Heuillet et al. \(2020\)](#) before they are effectively ready for adoption in practice.

Furthermore, it is often challenging to quantify how a certain patient is doing at any given point in time. RL transitions sometimes fail to capture a patient’s entire health trajectory, which is further complicated due to patient covariates that may

not be tracked in the EHR [Tang et al. \(2020\)](#). In addition, when RL rewards are based on medical measurements that are taken infrequently, this directly increases the difficulty in designing appropriate reward functions to guide the agent’s learning. Because of these challenges, RL treatment policies must be carefully trained with the involvement of domain experts as to not overstate algorithmic assumptions or impede the role of physicians.

1.3 Literature Review

In this section, we review the literature to draw insights as to the approaches that can help manage bias in healthcare.

Current Approaches to Bias Management in Healthcare

The medical field has considered differences in patient responses to treatments for years, directly or indirectly [Alosh et al. \(2015\)](#). Early statistical studies developed policies that were optimized for different patient sub-populations to demonstrate these differences [Tanniou et al. \(2016\)](#). Despite this, current medical practices may exhibit bias in treatments due to the heterogeneity of patients and their responses [Lesko et al. \(2018\)](#). The potential for discrimination certainly poses challenges for the adoption of any new treatment plan in healthcare [Panch et al. \(2019\)](#). Healthcare bias is seen in unequal asset allocation or directly in treatment plan efficacy between patients [Furukawa et al. \(2010\)](#).

Currently, there exist disparities in quality of care across patients, particularly as they relate to characteristics such as gender, race, ethnicity, etc. [Riley \(2012\)](#). For instance, in Veteran Affairs hospitals, black veterans are less likely to receive diagnostic testing for cardiovascular issues or undergo bypass procedures than their white counterparts [Whittle et al. \(1993\)](#). This is just one example of how similar patients receive different treatments, even when financial costs are irrelevant.

In addition to differences in quality of care due to overt patient characteristics, non-clinical influences also affect medical decision-making [Hajjaj et al. \(2010\)](#). For example, socio-economic status is directly correlated with the likelihood of receiving the best, and often most expensive, care [Bernheim et al. \(2008\)](#). Hospital equipment availability, physician time constraints, or the cost of treatment have been shown to contribute to treatment decisions, adversely impacting poorer patients [Li et al. \(2020\)](#). This does not mean that physicians are actively discriminating against patients but that external factors affect how clinical decisions are made [Hajjaj et al. \(2010\)](#). The impact of these non-clinical influences on treatment actions and patient responses are further confounded by patient characteristics such as race, gender, or age [Bernheim et al. \(2008\)](#); [Bond et al. \(2003\)](#).

Classically, medical policies are developed through clinical trials [Syn et al. \(2018\)](#). In this setting, one way to reduce biased medical treatments is through sub-group analysis [Liu et al. \(2020a\)](#); [Kreif et al. \(2012\)](#). Sub-groups may be used to develop different treatment models and also in analysis of treatment policies. Using sub-groups to develop different treatment plans can improve statistical effectiveness among patient populations using factors such as patient demographics [Cook et al. \(2004\)](#). Clinical trials use adaptive treatment strategies for individuals as an advanced way to differentiate patient groups [Murphy \(2005\)](#). Sequential multiple-assignment randomized trials incorporate initial patient responses to treatments to further define sub-groups to create different treatment policies.

Although sub-group analysis has shown the potential to reduce bias in treatment planning [Alosh et al. \(2015\)](#), such analysis is sometimes considered after the development process or is complicated due to the interaction between many factors. Despite statistical observations, sometimes creating separate models for sub-groups is challenging due to the requisite sensitive demographic information [Vyas et al. \(2020\)](#). In addition, recent studies identified a lack of biological evidence to create treatment policies by certain characteristics, such as race, for certain diseases due to the potential for discrimination [Jorde and Wooding \(2004\)](#). Regardless of if demographic data

are used to group patients in the development of separate treatment plans, sub-groups are usually used to compare effectiveness of recommended treatment policies between patients [Tanniou et al. \(2016\)](#). The trade-off between statistical advantage of using large patient sub-groups for treatment plans against policies for smaller, more individualized sub-groups has been considered in clinical trial settings [Cook et al. \(2004\)](#). However, the definition of sub-groups in medical studies can be expanded when more data are available.

1.3.1 Data Management in Statistical Learning to Reduce Bias

As discussed in Section [1.2.3](#), retrospective data can be leveraged to build advanced statistical models that can guide treatment planning. However, as expanded in Sections [1.2.4](#) and [1.3](#), bias in current medical practices can affect the recorded EHR data [Vincent \(2014\)](#). Therefore, it is imperative to account for data issues before developing models to ensure that bias does not propagate into the resulting treatment plans [Akter et al. \(2021\)](#); [Pessach and Shmueli \(2022\)](#). In the literature, various data management techniques have been introduced that can be exploited to address bias in advanced statistical models. In the remainder, we first highlight how bias in data can affect treatment plans and consequently, we expand on data management techniques that can be used to address them.

One potential source for bias in EHRs is from imputation of missing data [Jakobsen et al. \(2017\)](#). In medical data, some observations are recorded almost constantly such as from a heart rate monitor. Other features, such as medical testing, are not available at as frequent time intervals. Consequently, EHRs do not have data points for every observation at every time [Beaulieu-Jones et al. \(2017\)](#). To use these features in statistical learning, the missing observations are often filled in, or imputed in the dataset based on different estimates of missing values [Chang et al. \(2020\)](#). Practical considerations for how data are missing influence the choice of imputation methods

and affects the potential introduction of bias [Sterne et al. \(2009\)](#). Note that as an alternative to imputing data, records in EHR data with missing observations may be removed. However, this may negatively impact sample size, and in some cases, has been shown to reduce model performance compared with using imputed data [Jakobsen et al. \(2017\)](#). Also, removal of incomplete records could favor patients with comprehensive observations (partly because of their better access to more resources) [Bernheim et al. \(2008\)](#). EHRs do often contain a high rate of missing data; as such, the specific methods used to impute individual feature observations should be done in a way to best represent the missing data to prevent bias.

The more random the distribution and occurrences of missing (feature) observations, the less likely it is that data imputation can result in a systematic bias [Sterne et al. \(2009\)](#). Generally, missing data can be categorized as missing at random (MAR), missing completely at random (MCAR), or missing not at random (MNAR), each of which should be considered differently for imputation [Hu et al. \(2017\)](#). MAR and MCAR values can usually be estimated without introducing bias by comparing different imputation methods and choosing the method that best represents the missing data [Nguyen et al. \(2021\)](#).

Most imputation methods estimate values based on the assumption that the values are missing randomly; however, sometimes features are MNAR and are instead caused by external non-computable factor(s) [Jakobsen et al. \(2017\)](#). Therefore, if care is not taken, imputing MNAR values can potentially introduce bias. Unfortunately, these features are hard to identify and require a comprehensive understanding of the clinical setting where data were recorded. Indeed, there is no current solution to the bias caused by MNAR values, and hence, increased instances of missing values in EHRs have the potential to exacerbate the risk of bias in statistical models developed with imputed data. As such, it is imperative to acknowledge potential limitations of statistical studies when dealing with missing data, especially of type MNAR, and discuss best and worse case models when these values are imputed to account for potential bias [Jakobsen et al. \(2017\)](#).

Another key data-related factor that can lead to bias in RL models learned from EHR data is sample representation. Classification models have shown to present bias when trained on imbalanced datasets (i.e., where samples are unequally represented in the dataset) [Kärkkäinen and Joo \(2019\)](#). Specifically, if not adjusted for, models can favor the more represented group [Zhu et al. \(2020a\)](#). As an example, ML models developed to predict mortality in the ICU have been shown to present differences in accuracy among ethnic groups when trained on imbalanced data [Chen et al. \(2018\)](#). Such bias can often be identified by comparing error rates across groups similar to sub-group analysis.

To alleviate introducing bias to the model due to sample representation issues, data sampling techniques can be successfully used [Kelly et al. \(2019\)](#). That is, instead of training the model on imbalanced datasets, balanced sub-datasets are curated through oversampling and/or undersampling techniques. For instance, recent studies consider the number of samples from each group and the impact of a skewed population for face classification [Kärkkäinen and Joo \(2019\)](#). This work shows that when resampling data to contain equal samples from each race, the trained model was more accurate in classifying race, gender and age from images. Resampling has been used outside of facial recognition, notably in healthcare, and its use has improved detection accuracy of medical events using EHR data [Zhao et al. \(2018\)](#). For instance, oversampling instances of rare medication administration errors in medical datasets for the purpose of detection and classification of such events has shown model performance improvements.

Resampling healthcare data can be done in two general ways. Minority oversampling is performed by defining some sub-group, then using the existing distributions of variables in that sub-group to either choose more of these samples in training or to generate synthetic data for this group [Chawla et al. \(2002\)](#). Oversampling can increase the representation of less commonly seen patients or health states, and force treatment models to put higher ‘value’ on their corresponding rewards during policy training. Majority undersampling is the opposite of this process; where some observations from

a majority group are ignored to balance the data between groups [Sánchez-Hernández et al. \(2019\)](#). A combination of these oversampling and undersampling techniques can be used as needed to create balanced datasets for RL.

Performing over- and undersampling is not difficult; what is challenging is to identify sub-groups to resample, especially with high-dimensional feature subsets [Hooker \(2021\)](#). This is complicated because the stratifying features are not always even included in EHRs [Pollard et al. \(2018\)](#). Hence, properly defining the sub-groups to balance the data can be quite challenging.

Defining the sub-groups can be further complicated because the data required for treatment planning is generally temporal. Mechanistically, there exists sampling methods to combat high-dimensional imbalanced time-series data that could be used for EHR data [Zhu et al. \(2020a\)](#). This method generates samples that preserve the correlation between features within minority classes which are established via high-dimensional density ratio-based shared nearest neighbors clustering. While this method has been established using time series data, it was proposed as a classification problem and its adoption in RL practices and healthcare still needs to be explored.

Future data management techniques may go beyond simple classification of sub-groups to represent patient populations by expanding sub-group definitions to more than just ethnic, age, or gender [Akter et al. \(2021\)](#). Systematic racism and sexism are major societal issues and are key components to understanding discrimination in medicine. However, these descriptors are not the complete definitions of sub-groups or minorities in this environment. Rather, sub-groups can be defined by multiple variables to represent a patient’s state in high-dimensional latent space from EHR data [Wong et al. \(2018\)](#). Meaning that clinical feature observations can also differentiate populations beyond basic characteristics [Murphy \(2005\)](#). However, there is always a balance between defining more granular sub-groups and data availability. This is because training a model on too small of a sub-group increases the potential risk of model overfitting, not allowing the learned model to properly generalize to future patients that be assigned to the sub-group [Masino et al. \(2019\)](#). As the volume

of data available in EHRs increases, statistical uncertainty and the need for resampling or imputation decreases. Hence, EHRs can potentially allow for the establishment of more comprehensive sub-group definitions with sufficient observations to reduce bias in treatment planning [Akter et al. \(2021\)](#).

1.3.2 Algorithm Design to Reduce Bias

While data management methods aim to mitigate bias prior to learning a statistical model, there exist alternative and/or complementary techniques [Hooker \(2021\)](#) to reduce bias in treatment planning [Parikh et al. \(2019\)](#). Statistical learning model designs can be instrumental where other techniques may fail to address bias towards certain patients [Hooker \(2021\)](#). Unlike with ML classification tasks where differences in sub-group accuracies can be easily shown, it is oftentimes difficult to represent bias in the decision-making of temporal RL policies [Kelly et al. \(2019\)](#). In the following, we describe ML learning methods that can be applicable in RL as well as specific RL techniques that can be potentially leveraged to mitigate bias when training a treatment planning model.

In ML, algorithms generally attempt to minimize the average error or improve prediction accuracy, which converges to the minimum error for the majority population [Roth \(2020\)](#). *Cost-sensitive* methods apply a higher penalty to errors incurred in minority groups and are common in ML studies for imbalanced classification [Kukar et al. \(1998\)](#); [Brownlee \(2020\)](#). Most related to temporal RL problems, these methods have been implemented in time-series classification [Geng and Luo \(2018\)](#).

Cost-sensitivity has also been used when training RL policies. For example, [Ge et al. \(2021\)](#) use fairness-constrained policy optimization to ensure equal performance for dynamic recommendations. Here, the increased cost of mistakes on minority samples results in RL policies that exhibit parity across groups. Recently, cost-sensitive learning has been successfully applied to reward design in RL for different group samples in an imbalanced data set [Lin et al. \(2019\)](#). Specifically, this

imbalanced classification MDP framework changes the reward function so that samples from minority groups have larger positive and negative values for rewards and penalties, respectively, compared with those that are from majority groups. Therefore, the reward function is optimized in a way that the resulting policy is not overly influenced by samples from majority groups despite their higher representation.

Note that in general, applying cost-sensitive RL requires some method of identification of sub-groups, which can be difficult with high-dimensional patient data. For this reason, cost-sensitivity is usually applied to ensure fairness between easily identifiable demographic groups. There are, however, successful instances that use cost-sensitive RL on multiple sub-groups defined using different or multiple features to ensure that rewards are proportionate to that group’s presence in the training data [Yang et al. \(2022a\)](#). High dimensional clustering of patient data has been shown to improve personalized policy performance when sufficient samples exist for each cluster [el Hassouni et al. \(2018\)](#). The advancement beyond binary classification in patients greatly increases the potential effectiveness of bias management in cost-sensitive learning.

Additionally, models can be trained to ensure equal treatment performance, asset allocation, and outcomes between patient groups to increase fairness in treatment policies. For instance, [Rajkomar et al. \(2018\)](#) recommends to define protected sub-groups by basic patient characteristics subject to potential discrimination. Once patients are grouped, models are trained with respect to different fairness goals while ensuring that outcomes are consistent among the sub-groups. Similar fairness constraints in RL models have shown promise in non-healthcare applications. For instance, [Wen et al. \(2021\)](#) considers approving loan applications and shows that fairness constraints reduce biased decision-making. Another study [Ge et al. \(2022\)](#) develops a multi-objective RL framework to balance the trade-off between fairness and accuracy in RL policies for recommender systems. Such frameworks could be adapted to the healthcare domain to incorporate equitable rewards for sub-groups as an objective. Introducing model constraints for fairness requires researchers to group

patients based on potentially sensitive demographic or medical information. Hence, the aforementioned methods for identifying sub-groups who may be underrepresented in the data [Yang et al. \(2022a\)](#); [el Hassouni et al. \(2018\)](#) can pave the way to creating clinically relevant groups to be protected by fairness requirements.

Given adequate data and the information within them, RL treatment policies can be learned rather well for patients in the majority; however, these exact policies do not necessarily generalize well to underrepresented patients as their characteristics may be different [Baucum et al. \(2022\)](#). Note that learning policies directly from such underrepresented patients runs the risk of overfitting due to the small sample size and may not necessarily provide a good alternative [Zha et al. \(2022\)](#). Hence, different techniques can be used to adapt the learned policies from the majority to such underrepresented patients. For instance, [Baucum et al. \(2022\)](#) learns a policy for the majority of patients, and carefully adjusts this policy using variational inference for underrepresented patients who may respond differently to treatments. The results show that (noisy) Bayesian policy updates can identify effective policies for underrepresented patients.

Transfer learning in RL can also be used to adapt policies to underrepresented patients. Transfer learning is the general practice of incorporating outside information to accelerate learning, particularly when live testing and data collection are not possible [Zhu et al. \(2020b\)](#). Transfer learning has been successfully used for bias management in healthcare applications. For instance, [Parbhoo et al. \(2020\)](#) learns policies from data-rich European HIV patient information. Then, using clustered latent representations of patients, policy knowledge is transferred to effectively treat African HIV patients, from whom fewer data are available.

Some learning techniques alter the RL framework itself to adapt policies to new environments. Hidden Parameter Markov models (HiP-MDPs) can be used to transfer policy information to similar domains [Yang et al. \(2019\)](#). HiP-MDPs learn underlying global parameters using a latent representation of the data to efficiently transfer policy information to new data during testing [Yao et al. \(2018\)](#). Meta-RL is another

technique designed to leverage prior experiences to quickly adapt to similar tasks. It is currently used to improve autonomous vehicle driving for different drivers and traffic environments [Zhang et al. \(2021\)](#). Hence, such RL frameworks that effectively use limited data could be further explored in treatment planning for underrepresented groups in large EHRs.

Reward shaping is another technique that incorporates domain knowledge into rewards to assist learning in RL [Hu et al. \(2020\)](#). Specifically, reward shaping allows for guiding how an optimal policy is learned by introducing short-term rewards to the reward function rather than solely relying on long-term rewards [Grzes \(2017\)](#). In this way, insights from clinicians can influence RL by providing feedback when observed rewards in EHRs may be sparse, such as when dealing with infrequent test results. Although the existing literature has not specifically exploited reward shaping to manage and mitigate bias, such extension seems an immediate next step. In this way, primary rewards are still related to long-term patient health, but simple, short-term rewards can incorporate fairness goals based on domain knowledge from clinicians [Okudo and Yamada \(2021\)](#); [Hébert-Johnson et al. \(2017\)](#).

Conservative methods for reward estimation can also be used to reduce bias in RL policies. Limiting the value of estimated rewards in off-line settings has shown promise in risk-averse RL, particularly in robotics [Kumar et al. \(2020, 2019\)](#). Estimating rewards with a pessimistic approach, particularly for new state-action pairs, prevents overestimation of policy effectiveness when applying RL to the real world [Li et al. \(2022a\)](#). Conservative reward estimates have shown to create safer policies for robotics simulation environments [He and Hou \(2020\)](#); [Cicek et al. \(2021\)](#). Such an approach could be extended to healthcare applications to prevent RL models from overestimating the effect of treatment policies in new patients. Other conservative off-line learning approaches limit policy actions to those similar in the data to prevent actions where there are high levels of uncertainty [Swazinna et al. \(2021\)](#). RL methods that prevent off-line estimation bias can be leveraged in the future to prevent potential bias due to the domain shift in EHRs and real patients.

1.3.3 Clinical Implementation of Treatment Planning Models to Reduce Bias

Explainable and understandable treatment plans can allow clinicians to manage potential bias especially if it is not previously mitigated using data management and algorithmic design techniques [Yu et al. \(2021\)](#). Such treatment plans can reduce perceived and actual uncertainty of bias [Ho et al. \(2020\)](#). Here, we discuss how various concepts can be leveraged to further explain RL treatment plans and improve the understanding of potential bias.

As discussed in Section [1.2.4](#), explaining decision-making in RL policies is a challenge. Interdisciplinary collaboration between those who develop RL models and domain experts has been suggested to improve explainability [Puiutta and Veith \(2020\)](#); [Heuillet et al. \(2020\)](#). Therefore, when creating RL treatment policies, clinicians should be actively involved to provide insights and to understand how policies are learned.

Clinical considerations can be incorporated into learning RL policies to improve model development [Hooker \(2021\)](#). For instance, imposing certain constraints on the model can ensure that recommended actions meet pre-defined criteria set by clinicians [Lu et al. \(2020\)](#). This allows for recommendations to be made based on high-dimensional data, yet clinical criteria still be met in an understandable way. Hence, clinical criteria that promote equality can be incorporated into the model as constraints to achieve bias mitigation.

A clinician-in-the-loop framework, which uses patient health states to recommend treatment options to clinicians, rather than implementing the actions directly, can be further used to discourage bias [Tang et al. \(2020\)](#). This framework is useful for multiple reasons. Primarily, there are ethical concerns with automatic administration of any medicine [Escobar-Rodríguez et al. \(2012\)](#). In addition, patient health reward structures are hard to fully define [Tang et al. \(2020\)](#). Because patient health rewards are complex, algorithms may suggest multiple near-optimal options for the clinician

to choose from when administering treatment. This framework allows clinicians to use information such as patient preference or pain tolerance that is not available in EHRs when making decisions. In addition, this framework facilitates on-line learning as actions suggested by RL policies are used to collect further data while alleviating some ethical concerns of applying learned treatment plans. Table 1 summarizes the related works reviewed and their particular areas of contribution and relevance to bias mitigation.

Discussion

This study reviews the existing literature relating to bias and bias management when applying RL for treatment planning. RL can effectively learn from retrospective EHR data to create treatment policies [Li \(2017\)](#); [Li et al. \(2021\)](#); [Schaefer and Bailey \(2005\)](#). Although temporal EHR data lend themselves well to RL, the potential for bias is a major challenge when implementing RL systems. Notably, clinical and non-clinical influences affect the data collected in EHRs and result in sample representation issues and increased missing data [Kelly et al. \(2019\)](#); [Beaulieu-Jones et al. \(2017\)](#).

Through a comprehensive survey of the literature, this review highlights past and ongoing efforts to mitigate bias in RL when applied to treatment planning. The major themes that emerge relate to (1) various strategies to better manage the data from which treatment policies are learned; (2) altering the design of learning algorithms to ensure that they are learned in an equitable manner; and (3) making sure that the resulting policies are explainable and easily understood in a clinical setting.

Data can be managed to reduce bias in treatment recommendations [Zhu et al. \(2020a\)](#); [Chawla et al. \(2002\)](#). Two key challenges in data that can lead to bias are data missingness and lack of sample representation. Missing values are often imputed using different techniques and based on their missingness type. However, care must be taken during imputation not to introduce potential bias. For problems with a lack of sample representation in datasets, minority oversampling and/or majority

Table 1.1: Summary of related works reviewed and their particular areas of contribution and relevance to bias mitigation.

Bias Management Method	References
Missing data	Chang et al. (2020), Furukawa et al. (2010), Hajjaj et al. (2010), Hu et al. (2017), Jakobsen et al. (2017), Nguyen et al. (2021), Sterne et al. (2009)
Resampling	Akter et al. (2021), Chen et al. (2018), Hajjaj et al. (2010), Kärkkäinen and Joo (2019), Panch et al. (2019), Parikh et al. (2019), Rajkomar et al. (2018), Sánchez-Hernández et al. (2019), Roth (2020), Zhao et al. (2018), Zhu et al. (2020a)
Sub-group analysis	Alosh et al. (2015), Baucum et al. (2022), Bernheim et al. (2008), Bond et al. (2003), Cook et al. (2004), el Hassouni et al. (2018), Ge et al. (2021), Hébert-Johnson et al. (2017), Hooker (2021), Kreif et al. (2012), Lesko et al. (2018), Murphy (2005), Panch et al. (2019), Rajkomar et al. (2018), Riley (2012), Sharabiani et al. (2015), Tanniou et al. (2016), Wen et al. (2021), Whittle et al. (1993), Wong et al. (2018)
Cost-sensitive learning	Brownlee (2020), Geng and Luo (2018), Kukar et al. (1998), Zhao et al. (2018)
Transfer learning	Parbhoo et al. (2020), Rajapakshe et al. (2021), Yang et al. (2019), Yao et al. (2018), Zhang et al. (2021)
Reward shaping	Ge et al. (2021), Ge et al. (2022), Grzes (2017), Hu et al. (2020), Li et al. (2022a), Lin et al. (2019), Lu et al. (2020), Okudo and Yamada (2021), Wen et al. (2021), Yang et al. (2022a)
Domain shift	Baucum et al. (2022), Cicek et al. (2021), Kumar et al. (2020), Li et al. (2022a), Swazinna et al. (2021),
Explainability	Akter et al. (2021), Furukawa et al. (2010), Heuillet et al. (2020), Kelly et al. (2019), Li et al. (2020), Liu et al. (2020a), Lu et al. (2020), Okudo and Yamada (2021), Parikh et al. (2019), Puiutta and Veith (2020), Riley (2012), Tang et al. (2020), Vincent (2014)

undersampling data can be used [Koivu et al. \(2020\)](#); [Sánchez-Hernández et al. \(2019\)](#). However, these techniques are not panaceas as they still require identification of appropriate sub-groups from high-dimensional EHR data [Rajkomar et al. \(2018\)](#).

Further, algorithmic design can be leveraged to mitigate bias when learning treatment policies [Lin et al. \(2019\)](#). Cost-sensitive methods can be used to ensure that learning remains sensitive with respective minority sub-groups [Geng and Luo \(2018\)](#). Despite the exploration of this concept within RL, much of past work is grounded in classification where the ground truth is available and minority groups are easily identifiable. Hence, there exists a gap in research on using cost-sensitivity in reward design for developing temporal policies. Similarly, reward shaping could potentially be used to incorporate fairness goals based on domain knowledge from clinicians. Additionally, different techniques currently exist for adapting policies for new circumstances. This includes variational and Bayesian inference, transfer learning, HiP-MDPs, and meta-RL [Zhu et al. \(2020b\)](#); [Baucum et al. \(2022\)](#); [Yao et al. \(2018\)](#); [Yang et al. \(2019\)](#). However, more work is needed to further these techniques and ensure that they produce quality policies for any underrepresented patient with their own unique characteristics. Conservative reward estimation methods also have been shown to reduce bias in off-line RL [Li et al. \(2022a\)](#). However, further research is needed to investigate the extent to which implementing and potentially combining this approach with other techniques mitigates bias when learning treatment plans.

Lastly, RL policies need to be explainable and understandable to gain patient and physician buy-in and to ensure that no obvious bias is introduced during learning. Unfortunately, due to the temporal nature of RL policies and the complexity of patient health data, explaining the policies and identifying potential bias can be difficult [Puiutta and Veith \(2020\)](#). However, imposing certain criteria as constraints and involving physicians as part of model development efforts have the potential to lead to more explainable, understandable, and ultimately equitable treatment plans. A clinician-in-the-loop framework that allows for investigating near-optimal policies and

selecting one that most closely matches clinical intuition is another way to improve explainability of the policies and their potential fairness.

1.4 Conclusions

This study reviews the literature on bias and bias management in RL when applied to treatment planning and identifies major themes that can be used to mitigate bias. Despite the growing literature on bias and bias management in RL when applied to health care, there still is considerable work to be done to improve bias management in RL. Future work should focus on specific methods to identify bias in RL policies and to mitigate it, especially when RL is applied to retrospectively collect EHR data. Improving bias management is the key to successful implementation of RL treatment planning policies in practice.

Chapter 2

Uncovering Bias in Reinforcement Learning for Heparin Treatment Planning

A version of this chapter was originally submitted to IEEE Journal of Biomedical and Health Informatics by Benjamin Smith, Dr. Khojandi, Dr. Vasudevan, Dr. Shafi, and Dr. Davis:

Smith B, Khojandi A, Vasudevan R, Shafi N.I, Davis, R. (2023) *Uncovering Bias in Reinforcement Learning for Heparin Treatment Planning*. [Manuscript submitted for publication] Bredeesen Center, University of Tennessee.

Abstract

Dynamic treatment policies have the potential to greatly improve medical outcomes for patients. Algorithmic bias has not been fully explored for its impact on data-driven temporal treatment planning. Therefore, we investigate how bias can affect heparin treatment policies when learning from observational healthcare data. To do so, we apply both model-based and model-free reinforcement learning (RL) algorithms that attempt to develop policies to recommend heparin doses for patients. When learning these heparin dose policies, we filter the available training data to only include patients of a particular cohort. This process allows us to show how reinforcement learning-based treatment policies are affected by domain shifts in patient populations. Additionally, we explore the possibility of creating more personalized treatment plans using specific cohorts of patients. We observe the difficulty of learning effective treatment policies when too few samples exist from a patient cohort, but we also see how treatment policies are less effective when more dissimilar patients are included in the training data. This study highlights that care must be taken in the broad application of RL-based policies derived from observational healthcare datasets and calls for new strategies to ensure that RL-based treatments are not biased.

2.1 Introduction

Heparin is a rapid-onset anticoagulant often used in intensive care units (ICU) to treat acute thrombosis [Alquwaizani et al. \(2013\)](#). Administered heparin binds to the enzyme antithrombin to inhibit its function and prevent the formation or propagation of blood clots [Jin et al. \(1997\)](#); [Hirsh and Raschke \(2004\)](#). Despite its long history and widespread clinical use, determining the proper dose of heparin remains notoriously challenging, and observed patient responses are highly varied [Basu et al. \(1972\)](#). In addition, misdosing heparin can have serious health consequences. Supratherapeutic heparin can result in excessive bleeding while subtherapeutic dosing does not sufficiently reduce thrombosis risk or treat existing clots [Nemati et al. \(2016\)](#). Existing heparin treatment protocols do not benefit all patients equally [Keeling et al. \(2006\)](#), and these discrepancies result in bias against subgroups of patients [Yu et al. \(2021\)](#).

Currently, heparin is administered using treatment protocols developed from clinical trials that recommend weight-based dosing [Sholzberg et al. \(2021\)](#). Following an initial loading dose of heparin, clinicians monitor patients' activated partial thromboplastin time (aPTT) or factor Xa activity to adjust heparin infusions as needed to keep patients within a therapeutic range. Although the goal ranges for aPTT or Xa activity can vary by patient, the broad goal of reaching relatively stable anticoagulation is consistent [Basu et al. \(1972\)](#). Despite efforts to improve the efficiency of arriving at the required heparin dose, patients often respond differently to similar treatment regimes [Toulon et al. \(2021\)](#). One cause is that differences in weight may not be a sufficient basis for current protocols, which consequently fail to provide personalized effective treatment recommendations [Gottesman et al. \(2018\)](#). Thus, some patients do not receive an equal quality of care [Toulon et al. \(2021\)](#).

Observational studies have shown that population characteristics beyond weight matter [Sholzberg et al. \(2021\)](#); [Kuhle et al. \(2007\)](#). For instance, heparin protocols learned from a population of patients in the US are less applicable for clinical settings

in Japan [Kawase et al. \(2009\)](#). Additionally, adverse reactions to heparin treatments such as bleeding are significantly more common in women, particularly those over the age of 60 [Walker and Jick \(1980\)](#). A dynamic treatment protocol that considers a range of patient descriptors and the patient’s entire medical state to recommend a heparin dose could greatly reduce existing differences in the quality of care for patients.

2.1.1 Background

In recent years, electronic health records (EHRs) have greatly expanded the volume of medical data available to learn new treatment policies [Liu et al. \(2019\)](#). EHRs can provide much more comprehensive datasets than do clinical trials, although just as in clinical trials, existing issues of bias within the data can persist if not carefully considered. For our purposes, EHRs contain static and temporal clinical data from many patients that can be used to improve future treatment recommendations [Liu et al. \(2019\)](#). Static features include demographic data or known pre-existing conditions that do not change during a patient’s hospital stay. Other features are dynamic. For instance, in an ICU, EHRs store temporal medical data from sensors, monitors, and other measurements. Temporal data from heart rate monitors, instances of treatments being applied, or even clinicians’ comments can help describe a patient’s current health state. Although certain factors can limit the quality of EHRs, the increased volume of observations (compared with clinical trials) greatly improves the statistical power when developing data-driven treatment policies.

One way to develop data-driven treatment models is through Reinforcement learning (RL), a type of machine learning (ML) that is used to solve problems involving sequential decision-making [Sutton and Barto \(2018\)](#). RL algorithms learn through the interactions of an ‘environment’ with an ‘agent.’ In healthcare, RL can use temporal EHR data to represent an ‘environment’ in which patients transition from one medical ‘state’ to another based on the treatment or ‘action’ taken by the ‘agent’ [Yu et al. \(2021\)](#). Using sequential patient observations to extrapolate patient

trajectories, RL algorithms learn the relationship between ‘states’ and ‘actions’ in order to derive a ‘policy’ of actions that maximizes the collective ‘reward’ for every state. For medical purposes, ‘rewards’ are usually related to improvements in patient health metrics [Liu et al. \(2018\)](#). In the context of our study, that means being closer to the desired therapeutic range of anticoagulation.

RL is based on the mathematical framework of Markov Decision Processes (MDPs). MDPs are defined by tuples $\langle S, A, P, R, \gamma \rangle$ containing states, actions, transition function, rewards, and discount factor, respectively [Baucum et al. \(2020\)](#). The agent learns the transition function by mapping the current state s_t to the next state s_{t+1} under action a_t . The reward function $r(s_t, a_t)$ returns the immediate rewards, discounted by γ based on the horizon of the decision process. In RL, the goal is often to learn an optimal policy π that prescribes the actions to take in each state to ultimately maximize the total expected discounted reward over the planning horizon. Importantly, when using MDPs, one must assume the Markov property is applicable and that the evolution of the process starting from each state is independent of the past states [Baucum et al. \(2020\)](#). In this way, actions can be learned for each state independent of the past.

RL has been used to learn treatment policies for Parkinson’s disease [Baucum et al. \(2021\)](#); [Watts et al. \(2020\)](#), sepsis [Raghu et al. \(2018\)](#), diabetes [Zhu et al. \(2021\)](#); [Lee et al. \(2021\)](#); [Noaro et al. \(2023\)](#), and importantly and most relevant to this work, for heparin administration [Ghassemi et al. \(2014\)](#); [Nemati et al. \(2016\)](#); [Ghassemi et al. \(2018\)](#); [Baucum et al. \(2020\)](#); [Lin et al. \(2018\)](#). For healthcare applications, RL agents generally leverage historical data to learn treatment policies that can be generalized to new patients. RL studies of this type are performed in an observational setting due to ethical considerations [Gottesman et al. \(2018\)](#). Applying RL with EHR data can be a useful method for developing improved treatment policies. When doing so, one must carefully consider the population characteristics of the patients in the EHRs as well as how an RL algorithm optimizes its policy.

2.1.2 Related Works

EHRs can be used for RL in two distinct ways, model-based and model-free. In model-based RL, EHR data are used to create a mathematical model of the environment that predicts a patient’s next state and its accompanying rewards given a current state and action. The agent then interacts with this environment model to learn which actions to take in each state that maximizes returns (an optimal policy) [Kaiser et al. \(2019\)](#); [Atkeson and Santamaria \(1997\)](#). One advantage of model-based RL is that new policies can be tested on the environment when live data collection is not possible (on-policy RL). However, because environment models are developed using static EHR data, their validity is limited by the volume and quality of available training data. Previous studies have used different environment models to train RL agents to learn discrete heparin dose recommendations that attempt to keep a patient’s aPTT values within a therapeutic range [Baucum et al. \(2020\)](#); [Nemati et al. \(2016\)](#).

In model-free RL, the EHR data are organized to train an RL policy directly from existing observations. EHR data are collected as a sequence of ‘state, action, next state’ tuples in a replay buffer. Observations of patients who receive heparin have been used to create an experience replay buffer for RL [Lin et al. \(2018\)](#). Here, RL learns the optimal policy directly from transitions in the experience replay although the data were collected using a pre-existing, potentially unknown, policy [Qiu et al. \(2022\)](#). In such a case, model-free RL is also an instance of offline and off-policy RL that learns entirely from EHR data without simulating interactions. Model-free RL applied to heparin employs a similar reward function as model-based studies. There are limitations to offline RL policies that use observational data, however. Issues often occur when the patient population to which an RL policy is applied is different from the population with which it was trained [Prudencio et al. \(2022\)](#) or when confounding variables are not included in the data.

While others have successfully used RL with EHR data to develop heparin treatment policies, to the best of our knowledge, none of the past works fully

considered population characteristics of their training data, the differences in patient responses to heparin, and subsequent potential for bias in treatment policies [Nemati et al. \(2016\)](#); [Lin et al. \(2018\)](#); [Baucum et al. \(2020\)](#). Instead, the population of patients in the EHR was considered as a whole, and the policy performance was estimated across *all* patients. In other words, these studies do not evaluate the learned RL policies for potential biases when applied to different groups of patients. Although some of these studies identify underrepresented patients and attempt to adapt an RL policy for them [Baucum et al. \(2022\)](#), to the best of our knowledge, no RL-based heparin treatment planning study has shed light on the differences across patient subgroups to potentially reduce bias or investigated training data homogeneity and sample size when learning the corresponding RL policies.

2.2 Motivation

Algorithmic bias in ML is a well-studied issue [Mehrabi et al. \(2021\)](#), but much less so in RL due to the difficulty of observing and quantifying bias in temporal decision-making problems. This is particularly true for heparin dosing algorithms. RL algorithms may estimate the potential returns for certain states differently based on the observed rewards of individual patients who visited this state. This could cause an algorithm to be more pessimistic for certain states if most observations in those states come from patients who are not treated well by weight-based protocols. In this study, we investigate and demonstrate how the differences in patients and their responses captured in EHR data potentially give rise to bias in heparin treatment policies.

Despite the potential advantages of using EHR data, certain factors can limit the effectiveness of learned treatment plans derived from them. EHR data are generally collected under the standard of care, which may not be fully personalized. Thus, the corresponding quality of care and data across individuals can vary, making learning how to correctly recommend doses for all patients difficult. An RL algorithm learns

a policy that maximizes the total reward for patients included in the training data [Sutton and Barto \(2018\)](#). However, an effective policy for the majority of patients may not be effective for patients’ subsets who respond differently. A truly optimal RL policy would benefit all potential patients and would not sacrifice performance on certain patients [Baucum et al. \(2021\)](#). When considering heparin dosing, observed treatment responses differ between gender, age, and race [Campbell et al. \(1998\)](#); [Ignjatovic et al. \(2006\)](#); [Gibson and Yuet \(2021\)](#). However, partially due to a lack of genealogical explanation for these differences and partially out of a concern for discriminatory algorithms demographic features are not always used in patient state definitions [Lin et al. \(2018\)](#).

Additionally, in off-policy RL, distributional (i.e. domain) shifts between training and testing data can result in policies that are ineffective in practice [Kumar et al. \(2020\)](#). When these domain shift errors occur, RL algorithms typically overestimate performance on previously unseen states [An et al. \(2021\)](#). While protocols may have been learned from patients present in one EHR, the actual population of patients being treated could be different. Although RL policies may estimate higher rewards, domain shifts can result in RL policies that have unknown effectiveness for patients who are under-represented in the training dataset. Then, these policies would exhibit bias against these patients [Gottesman et al. \(2018\)](#). Similar data issues also limit the effectiveness of environmental models used for on-policy learning [Nguyen et al. \(2012\)](#).

Some of these estimation errors can be mitigated by increasing the samples in the training population, although often obtaining more observations may not necessarily be possible. Recent efforts to constrain predictions in offline RL to reduce domain shift errors have improved performance [Qiu et al. \(2022\)](#); [Kumar et al. \(2020\)](#). However, direct consideration of the training population is necessary to identify and mitigate potential biases in heparin treatment plans.

2.3 Objectives and Contributions

The potential for bias within RL-derived treatment plans has not been fully investigated. Here, we explore the impact of training sub-populations when learning RL-based heparin policies from observational data. We control which patients’ data are used to train RL policies to create more personalized policies and show how treatment policies can exhibit bias. Our work aims to highlight gaps in the current heparin RL-policy development paradigm and we suggest how future researchers could consider patient differences when learning RL treatment policies.

We filter the data to create training datasets that contain only patients of a particular cohort and then derive RL policies from it. Theoretically, if sufficient samples exist for a particular cohort, then an effective RL policy can be learned and then applied to similar patients in the future. However, these policies will likely be less effective when applied to patients dissimilar to those included in the training data.

Although limiting the training data to rather homogeneous groups of patients allows us to explore the possibility of creating more tailored treatment policies, it precipitates another issue related to the effect of sample size when learning RL policies: the fewer observations available from a cohort of patients, the higher the likelihood of overfitting the treatment policy. Fewer observations in the training data also require more estimation of unseen state-action values which is often associated with domain shift errors. Sampling more observations of similar patients should improve the certainty of policy effectiveness. However, since no new data are collected, eventually only observations from dissimilar patients are available. Policies learned from larger heterogeneous populations would be less personalized and consequently, not optimized for all different patients. That is, although these generalized policies are often claimed to be more applicable to new patients *on average*, they are less likely to fully capture the unique differences in patient responses to heparin. In this way,

our study also explores the relationship between sample size and patient population heterogeneity on the quality and robustness of learned heparin treatment policies.

2.4 Methodology

2.4.1 Dataset and Patient Cohorts

We use the Medical Information Mart for Intensive Care (MIMIC-III) database [Johnson et al. \(2016\)](#), which has been widely used in RL heparin treatment planning studies [Nemati et al. \(2016\)](#); [Baucum et al. \(2020\)](#); [Lin et al. \(2018\)](#); [Ghassemi et al. \(2014, 2018\)](#); [Qiu et al. \(2022\)](#). We particularly chose to use the same dataset as previous RL studies to investigate the gap in understanding bias in heparin treatment protocols. The MIMIC-III heparin dataset contains hourly observations for different medical measurements of 2067 patients, with a mean patient length-of-stay in the ICU of 26.6 hours.

The dataset contains different patient observations that are used to define states, actions, and rewards. Consistent with past studies [Nemati et al. \(2016\)](#); [Baucum et al. \(2020\)](#), fourteen features including heart rate, aPTT, temperature, creatinine, Glasgow Coma Score, hematocrit, hemoglobin, normalized ratio of prothrombin, platelet count, urea, arterial oxygen saturation, white blood cell count, and arterial carbon dioxide levels define the state of a patient in our environment. The actions are defined by discrete doses of heparin (in mL/kg). Heparin doses are discretized $a_t \in \{0, 1, 2, 3, 4, 5\}$ based on the percentiles of heparin doses administered relative to the patient’s weight. Missing heparin doses were imputed using sample-and-hold interpolation, and missing aPTT values were estimated using a neural network that used the other thirteen features (plus weight, age, and gender) to estimate missing aPTT observations. Together these patient trajectories can be organized into patient health transitions to either train an environment model (for on-policy) or collected in a replay buffer (for off-policy) from which heparin treatment policies can be learned.

In this study, we create two sets of cohorts based on patient gender and weight.

First, we used gender (men/women as encoded in the dataset) as a feature to bin cohorts to investigate domain shift errors and how policy performance uncertainty can increase in RL heparin policies. Both gender groups are well represented in the dataset and hence, it is likely that reasonably generalizable policies can be learned using each cohort’s data. Although distributions of features such as weight differ slightly between men and women, the features are distributed widely in both groups (see Table 2.1), making it conceivable that each of these cohorts would produce policies that could be applicable to the other. This is consistent with the implicit assumptions made by past RL studies about the generalizability of the learned policies from a given population but is at the core of our investigation.

Next, we use weight to create patient cohorts, because heparin doses are currently recommended based on patient weight and weight has been reported to have a significant influence on heparin response [Wilson et al. \(2001\)](#). Therefore, we use weight to bin cohorts to investigate the trade-off between sample size and cohort heterogeneity when training and testing within and across cohorts. We do so by defining a range around the mean patient weight (182 lbs) from which to include patients and then gradually increasing the range to include more patients. The distribution of patient weights (in lbs) for all individuals in the dataset is shown in Figure 2.1. The details of our weight-based cohorts are provided in Table 2.2.

2.4.2 RL Problem Formulation

We implement various RL algorithms, all of which attempt to learn a discrete heparin treatment policy $\pi(a_t|s_t)$ to recommend heparin doses for each state in the environmental model or dataset that maximizes total expected returns (which is the immediate reward $r_t(s_t, a_t)$ at the current time t plus the discounted sum of all future rewards). We use the same reward function used in previous studies [Baucum et al. \(2020\)](#); [Nemati et al. \(2016\)](#); [Lin et al. \(2018\)](#); [Qiu et al. \(2022\)](#) which is calculated

Table 2.1: Gender-based cohorts used in this study.

Cohort	Men	Women
Patients	1,193	873
Observations	31,612	23,265

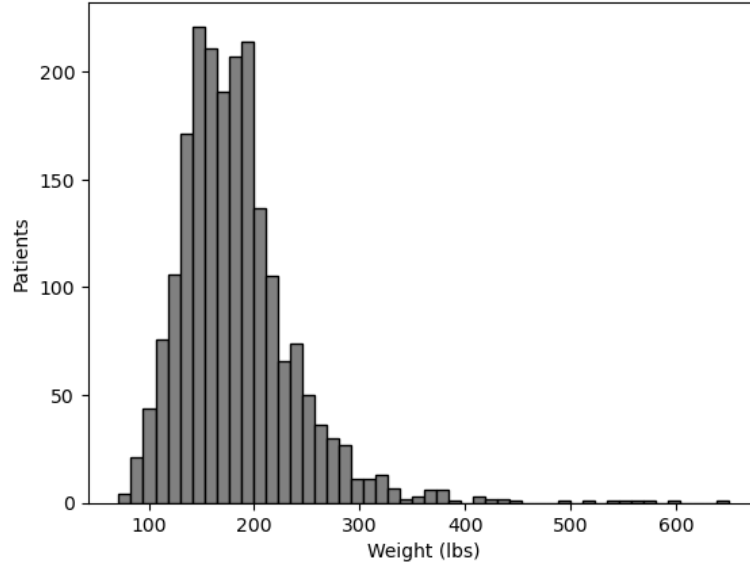


Figure 2.1: Distribution of the weight of patients in the dataset.

Table 2.2: Weight-based cohorts used in this study.

Weight Range (lbs)	173-191	164-200	146-218	128-236
Variance (lbs)	± 9	± 18	± 36	± 54
Patients	331	633	1,136	1,550
Observations	8,303	16,327	29,900	41,157

Weight Range (lbs)	110-254	92-272	74-290	56-308
Variance (lbs)	± 72	± 90	± 108	± 126
Patients	1,797	1,924	1,986	2,006
Observations	47,678	51,057	52,763	53,216

using the patient’s aPTT values in the current state. The reward function essentially returns a value of +1 for aPTT values between 60-100 seconds (therapeutic levels). If a patient’s aPTT value at a particular time is outside of this range, then the reward function returns a -1. The details of this function are provided in Equation (2.1), as follows:

$$r(s_t, a_t) = \frac{2}{1 + e^{-(aPTT-60)}} - \frac{2}{1 + e^{(aPTT-100)}} - 1. \quad (2.1)$$

We train our RL policies in two distinct ways to observe the effects in both model-based and model-free paradigms. This is done to examine whether populations in EHR data can affect RL policies regardless of how treatment policies are learned. In the model-based setting, we first train an environmental model and then apply on-policy RL algorithms to learn treatment policies. Alternatively, we employ off-policy RL algorithms directly to the transitions in the dataset to learn heparin dose policies.

Model-based

For model-based RL, we model the environment using a transitional Variational Autoencoder (tVAE). The tVAE adapts the traditional VAE model to model patient transitions and has been shown to outperform other environmental models for this particular heparin dataset [Baucum et al. \(2020\)](#). The tVAE models transitions between sequential patient observations. A 14-length state vector s_t and the one-hot-encoded action a_t at time t are inputs to the tVAE, and it outputs the predicted state at time $t + 1$. Within the tVAE architecture, one network encodes the current state into a latent space vector, and then a decoder network reconstructs and predicts this vector in the state space for the next state. tVAE models are trained using gradient descent that minimizes the sum of prediction loss and Kullback-Leibler divergence on the latent and prior distributions. Using this environmental model, we are able to apply on-policy RL algorithms that interact with the environment to learn a heparin dose policy. The mathematical details of the tVAE environment are provided in [Baucum et al. \(2020\)](#).

We then apply two state-of-the-art on-policy algorithms, namely Asynchronous Advantage Actor Critic (A3C) and Proximal Policy Optimization (PPO) algorithms. The A3C algorithm learns the treatment policies $\pi_\theta(a_t|s_t)$ that maximize the total expected returns in the tVAE environments Babaeizadeh et al. (2016). Deep RL methods implement neural networks to approximate state-action values, or Q-values, to efficiently learn problems in complex tasks when tabular methods are not feasible. A3C uses two neural networks to represent the value function $V(s_t; \theta_v)$ (critic) and the policy $\pi_\theta(a_t|s_t)$ (actor). A3C attempts to find θ that maximizes $\arg \max_\theta \mathbb{E}[\sum_{t=0}^T \gamma^t r(s_t, \pi_\theta(s_t))]$. For each state, s_t , and action a_t , the agent calculates the advantage $A(s_t, a_t)$ by comparing the difference in the state-action value and the average state value for all possible actions in this state. The advantage is calculated as shown in Equation (2.2):

$$A(s_t, a_t) = \sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k}) - V(s_t). \quad (2.2)$$

where $V(s_t)$ is the total expected value-to-go starting in state s_t , γ is the discount factor (which we set to 0.99 Baucum et al. (2020)), and k is bounded above by the length of the decision process T . The advantage is then used to update the policy network parameters θ using gradient descent by $\nabla_\theta \log(\pi_\theta(a_t|s_t)A(s_t, a_t)) + \eta \nabla_\theta H(\pi_\theta(s_t))$ where $H(\pi_\theta(s_t))$ is the entropy of the policy in state s_t which encourages exploration. A3C applies parallel agents interacting with the environment which act to stabilize policy learning which makes the algorithm asynchronous Babaeizadeh et al. (2016).

Similarly, we apply the PPO algorithm to learn treatment policies using the tVAE environment Schulman et al. (2017). PPO adapts concepts from trust region policy optimization that prevent large policy updates and encourage smaller policy updates to find an optimal policy. PPO constrains the policy update by using a clipped surrogate objective function. The main objective function of PPO is shown

in Equation (2.3):

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t[\min(\rho_t(\theta)\hat{A}_t, \text{clip}(\rho_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)]. \quad (2.3)$$

Here, ρ_t represents the probability of taking action a_t in state s_t with the updated policy π_θ divided by the same probability but under the previous policy $\pi_{\theta_{old}}$. With $\epsilon = 0.2$, the clipped term $\text{clip}(\rho_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t$ removes incentives for any policy update ratios beyond [0.8-1.2] to prevent larger policy updates Schulman et al. (2017). The minimum of the clipped and non-clipped objective functions is selected so that the final objective value is the pessimistic bound of the non-clipped objective function.

Model-free

In the model-free setting, no environmental model exists and only off-policy RL algorithms can be applied to learn from observational data. First, the experience replay tuples (s_t, a_t, r_t, s_{t+1}) for all transitions in the dataset D are collected in replay memory Mnih et al. (2013). Then by sampling transitions from the replay memory, off-policy RL algorithms leverage these data to learn heparin treatment policies.

We then apply three state-of-the-art off-policy algorithms, namely Deep Q-network (DQN), Double Deep Q-network (DDQN), and Conservative Q-learning (CQL). DQN uses two neural networks, a Q-network and a target network, to estimate state-action values Yang et al. (2022b). The Q-network predicts the value for each state-action pair and the target network predicts the best Q-value of all actions for each state. The Q-network $Q(s, a; \theta_i)$ is updated for network parameters θ every i iterations Mnih et al. (2015). The target network is only updated after τ time steps and acts to help stabilize the targets of the Q-network. These networks are iteratively updated until the optimal policy, $\pi_{\theta^*}(s_t, a_t)$, is found. DQN samples uniformly from the replay memory to incrementally update the network parameters to minimize the

loss function shown in Equation (2.4):

$$L(\theta) = \mathbb{E}_{s_t, a_t, r_t, s_{t+1} \sim D} [(r_{t+1} + \gamma \max_a Q_{\theta'}(s_{t+1}, a) - Q_{\theta}(s_t, a_t))^2], \quad (2.4)$$

where we set γ to 0.99 to discount the agent’s horizon on future rewards and θ' refers to the parameters of the target network.

Analogously, DDQN adapts a similar architecture to DQN such that actions are taken from the current Q-network rather than the target network to reduce the overestimation bias [Van Hasselt et al. \(2016\)](#). The loss function in Equation (2.5) shows this difference compared with DQN:

$$L(\theta) = \mathbb{E}_{s_t, a_t, r_t, s_{t+1} \sim D} [(r_{t+1} + \gamma Q_{\theta'}(s_{t+1}, \arg \max_a Q_{\theta}(s_{t+1}, a) - Q_{\theta}(s_t, a_t))^2]. \quad (2.5)$$

Here, different from DQN, the value is calculated by the target network, while the actions are selected by the original Q-network [Noaro et al. \(2023\)](#).

Finally, we implement CQL, an off-policy RL algorithm adapted from the DDQN architecture. CQL is designed to reduce bias in off-policy (and offline) RL by providing conservative estimates of the value of states [Kumar et al. \(2020\)](#). In off-policy RL, because the Q-value of unseen (s_t, a_t, s_{t+1}) tuples must be calculated via bootstrapping, they are prone to overestimation. CQL bounds the Q-function for unseen state-action pairs to prevent this issue. Mathematically, CQL adds a regularization term to minimize Q-values on unseen actions while simultaneously maximizing the Q-function over the entire dataset as shown in Equation (2.6) [Kaushik et al. \(2022\)](#); [Shiranthika et al. \(2022\)](#):

$$L(\theta) = \alpha \mathbb{E}_{s_t \sim D} [\log \sum_a \exp Q_{\theta}(s_t, a_t) - \mathbb{E}_{a \sim D} [Q_{\theta}(s_t, a_t)]] + L_{DoubleDQN}(\theta), \quad (2.6)$$

where α is the trade-off factor between the regularization term and DDQN loss which we set to 1 [Kumar et al. \(2020\)](#). CQL is designed to reduce offline estimation

issues and hence, the policies learned by CQL are expected to represent the best-case scenario for off-policy value estimation when distributional shifts occur in offline RL.

2.4.3 Evaluation

We compare each cohort-specific RL policy’s estimated performance with the performance of the weight-based treatment policy derived from the dataset (observed clinical performance). For consistency and to enable comparisons, we use the same reward function as in our RL formulations to calculate clinical performance. We calculate the cumulative rewards for each patient’s hospital stay and average them across all patients in that cohort. We then compare the average clinical rewards with our cohort-specific RL policy rewards to examine the extent to which RL policies provide benefits beyond each cohort’s observed outcomes.

For model-based RL, the tVAE environment is trained using data from a cohort and then a policy is learned from the cohort’s environment. Using the learned heparin treatment policies, we simulate patients’ trajectories for 27 hours (closest to the average length of patient observations) with the environment and average the cumulative rewards across patients in that cohort. We simulate 5000 trajectories for each cohort five times and report the mean of their average returns.

For model-free RL, the algorithms learn from each cohort’s data directly. After training, we evaluate RL policies on patients who belong to the respective cohort but were withheld from the training data for testing by an 80/20 split, respectively. Each policy can be evaluated using the average value estimation function of states given by $\mathbb{E}_{s_t \sim D}[\max_a Q_\theta(s, a)]$. Similar to on-policy evaluation, we repeat policy training five times for each cohort and report the mean of their average returns.

Additionally, heparin dose policies are tested on patients who belong to different cohorts than the training data. When testing policies this way, we essentially examine

the extent to which RL policies generalize to different patient populations. Such cross-testing across cohorts demonstrates how distributional shifts in patient populations affect the predicted performance of RL policies. Because we know which cohort each policy is learned from and should be optimized for, this cross-testing allows us to elicit the potential mismatches between policy performance estimations and enables us to predict any potential inaccuracies.

2.5 Results

2.5.1 Gender-based cohorts

Figure 2.2 presents the average rewards for policies learned using gender-based cohorts. The clinical data bars represent average clinical rewards using the weight-based treatment policy which are calculated from the dataset directly. In our dataset, men had a mean reward of -1.81, while women had a mean reward of 0.31. We compare the clinical data rewards to the estimated rewards of RL policies learned from men’s and women’s data under on-policy model-based RL algorithms (Figure 2.2(a)) and off-policy model-free algorithms (Figure 2.2(b)). All RL algorithms improve upon clinical policies, which is consistent with the published literature [Baucum et al. \(2020\)](#); [Nemati et al. \(2016\)](#); [Lin et al. \(2018\)](#); [Ghassemi et al. \(2014, 2018\)](#).

As seen in the figure, interestingly, the rewards under a policy that is learned and tested on the same cohort (cohort policy) are different from those under a policy that is learned and tested on the opposite cohort (opposing policy). The generic expectation is that policies trained and evaluated on the same cohort will outperform those trained on different cohorts. Overall, the rewards under cohort policies for men are lower than those for women. This potentially implies that women benefit more from the learned RL cohort policies.

Unexpectedly, for men, rewards under policies trained and tested on men are consistently lower than under policies trained on women (opposing policies). This is

consistent across all on-policy and off-policy RL algorithms, though the discrepancy is greater for DQN and A3C. Note that despite this *seemingly* improved performance, we speculate that the opposing policies are not truly optimized for men. Rather, this observation might be due to poor reward estimations caused by the induced domain shift in training and testing populations. That is, the higher predicted rewards could be due to the misestimation of rewards when using women’s policy applied to unknown spaces in the state-action space of men. Additionally, the differences in the average clinical data show how women have higher rewards on average than men. This difference indicates that men and women respond differently to, and benefit unequally from, the same weight-based treatment protocol. Also, this difference could help cause the overestimation of rewards when testing the policy learned from women on men. The policies trained only with data from women likely fail to identify the differences in heparin response dynamics between the two populations and assume that men will respond similarly to women. Such results point toward the trouble with biased estimators caused by underlying cohort shifts. DQN and A3C seem to be more vulnerable to this domain shift issue, whereas DDQN and PPO appear slightly more robust, which can be partly explained because of their design that prevents overestimation bias. Unexpectedly, CQL was more vulnerable than DDQN, but less vulnerable than DQN to domain shifts.

Misestimating policy performance when the population is shifted shows how RL can be vulnerable to, and further propagate, biases in data. Despite performing seemingly well on new patient populations, because of our study design, we know that there is more uncertainty in the policy’s performance and that the policy assumes that patients in the test population respond similarly to those in the training population. In practice, changes in the patient population are less obvious. Therefore, when applying RL one must consider that differences in patient responses to heparin could decrease the effectiveness of learned treatment policies.

Lastly, we observe certain differences between the on-policy and off-policy implementations. Estimated rewards for on-policy algorithms are in a similar range as

they both use the tVAE environment to model transitions. The off-policy algorithms, however, on average result in higher estimated rewards compared with the on-policy ones. Despite that, the rewards under off-policy algorithms are themselves on a similar scale as these policies are learned by sampling observations from the same replay buffer. This allows us to reasonably compare policies from different algorithms that are developed through the same process (either off- or on-policy).

2.5.2 Weight-based cohorts

Figure 2.3 presents the average rewards for policies learned using the weight-based cohorts. The bars represent the population size of each cohort, which increases in *weight range* from left to right, and the line and dots represent the average rewards of each RL policy for the corresponding cohort. We compare the estimated rewards of RL policies learned from each weight-based cohort’s data under on-policy model-based RL algorithms (Figure 2.3(a)) and off-policy model-free algorithms (Figure 2.3(b)).

As seen in the figure, when using the tightest cohort weight range data in training (i.e., a rather homogeneous cohort with respect to weight but the fewest number of patients), we generally observe the lowest estimated rewards across the algorithms. As the weight ranges increase (i.e., cohorts become less homogeneous with respect to weight but now include more patients), in general, estimated rewards first increase and then start to decline. PPO and DQN achieve their highest estimated rewards when trained with and applied to patients between 128 and 236 lbs, and CQL achieves this with patients between 110 and 254 lbs. Finally, A3C achieves its highest estimated rewards for patients between 92 and 272 lbs, the widest weight range for which we observe a maximum reward across all algorithms.

The observations in Figure 2.3 indicate a trade-off between sample size and heterogeneity in training data. That is, the drop in performance for wider weight ranges is likely due to differences in the newly included patients’ responses to heparin which increases the heterogeneity of the training population. These patients who

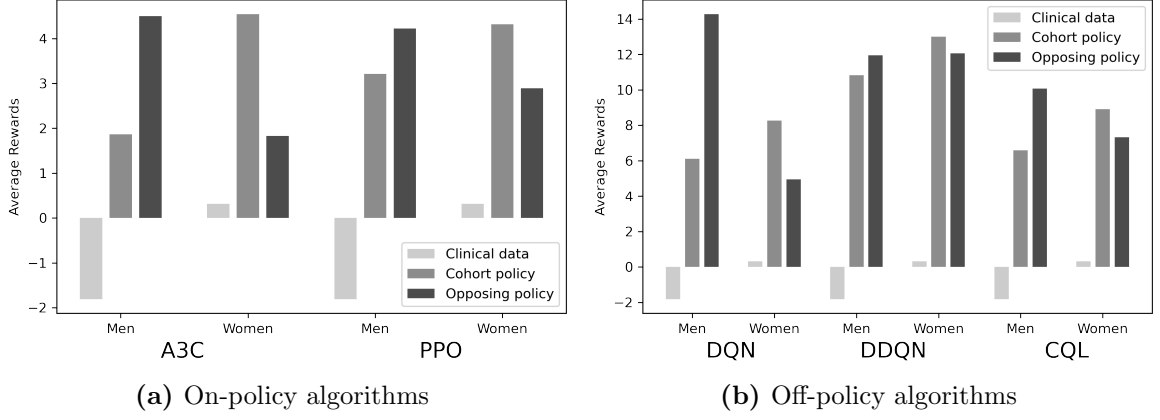


Figure 2.2: Rewards for men's and women's cohorts under on- and off-policy algorithms. Average observed rewards for each cohort from the dataset are shown by Clinical data. Cohort policy refers to the estimated rewards for the RL policy trained using that cohort's respective data. Opposing policy refers to the estimated rewards of the policy learned from the other gender's data.

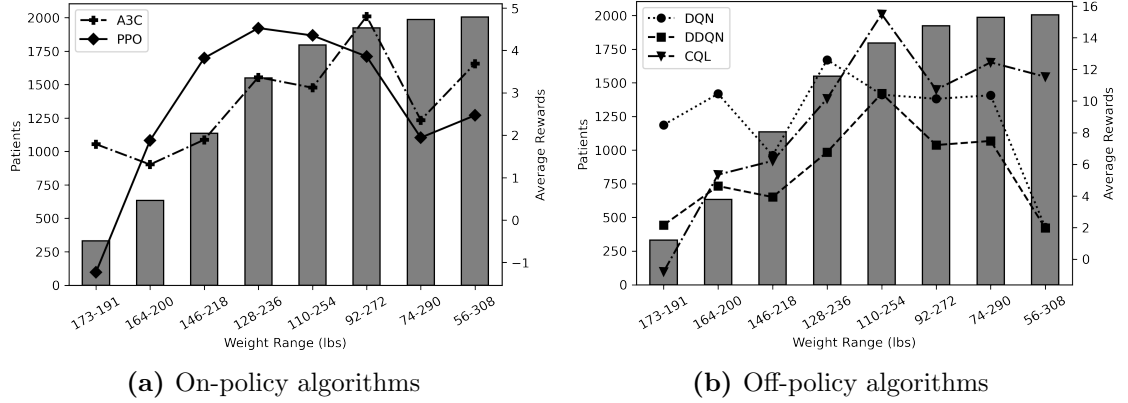


Figure 2.3: Rewards for weight-based cohorts. The bars correspond to the number of patients in each cohort given on the left-side axis. The average rewards for each cohort are shown by the line plot and values on the right-side axis.

serve as outliers with respect to weight could potentially transition differently than the ‘majority,’ affecting how RL recommends doses. Therefore, the policy’s overall performance could be affected partly because the policy is now less effective on the ‘majority’ and partly because the policy performs potentially poorly on the newly, but less represented ‘minority.’

The general trend of increasing and then decreasing rewards is present for all the algorithms implemented, but there are some notable differences. When training policies using tVAE environmental models (i.e., on-policy algorithms), we observe similar values for most cohorts, but a significant discrepancy in rewards when learning from patients within 173-191 lbs. PPO estimates very low performance for this cohort, but A3C predicts a higher value. Because this cohort contains rather few patients (i.e., 331 patients as provided in Table 2.2), it is possible that A3C overfits its policy to the training data. Also, for these two algorithms, we observe an increase in rewards for the largest weight range.

As was the case for on-policy algorithms, there are differences in estimated rewards among off-policy algorithms. Using patient data from the two most tight weight range cohorts, DQN predicts a higher value compared to DDQN and CQL. Again, we attribute this to the policy likely overfitting to the smaller population. Estimated rewards follow a similar trend until the cohort with the largest weight range. When learning from all patients between 56 and 308 lbs, the estimated policy performance for DQN and DDQN drastically decreases which indicates that these methods may be more sensitive to outlier observations. In contrast, CQL’s performance decreases, but less drastically than the other two, which could be because its conservative value estimation makes it robust against changing the policy too significantly for outlier patients.

2.6 Discussion

This study investigates how patient populations can affect policy learning. It particularly demonstrates the subtle ways bias can occur in RL, which is different from the overt way it presents itself in non-temporal ML problems. We explore the potential for bias in RL policies based on the patient population present in EHRs and attempt to use cohorts to create more personalized treatment policies. We explain how domain shifts can occur when learning from EHRs and then observe the effects when changing the patient population using cohorts based on gender. We also observe how the representation of particular patient types limits the ability of RL to learn effective heparin treatment policies using weight-based patient cohorts.

Domain shifts in EHR populations can affect RL heparin treatment policies. Our results show that simple features, such as gender in this case, can potentially cause domain shifts. While our choice of cohorts is made for illustrative purposes, this shows how factors in the data limit the effectiveness of RL treatment plans. Cross-testing policies for the opposite gender allows us to highlight what uncertainties exist when applying RL policies on a similar, but not identical patient population. Because we intentionally change the training population, we can predict when the estimated performance of RL policies is likely inaccurate. Importantly, our results show that domain shifts can affect model-based on-policy RL as well as model-free off-policy RL algorithms. However, they also show that not all algorithms are created equally, and some may be more or less prone to certain biases in data.

Furthermore, our results show that there is a trade-off between sample size and heterogeneity that can affect RL heparin treatment policies. Using weight-based cohorts, we show that heparin dose policies can be learned for particular groups of patients when sufficient samples exist in the data. We observe the differences in estimated RL performance when changing the patients included in the training population. Using weight-based cohorts, we successfully trained policies to learn improved

treatment policies for patients near the ‘average’ weight. However, including less similar patients (and outliers) causes the policy performance to suffer. This is likely due to the differences in transition dynamics of patients less similar in weight. Unfortunately, for groups of patients outside the normal weight range, we are unable to confidently and consistently learn an improved treatment policy using their observations alone. Hence, transfer learning and policy adaptation methods [Baucum et al. \(2021\)](#) for under-represented patients may be the best approach to optimize RL for minority populations that are significantly dissimilar from most patients in the training data.

Our study has several limitations. Importantly, our goal was not to develop an improved method for learning heparin dosing through RL. Rather, our work simply attempts to highlight how EHR data can introduce bias in RL policies. We follow methods previously applied to demonstrate the effect of patient populations in the current paradigm for RL heparin policy development. Our evaluation focuses on how RL algorithms estimate performance as the training and testing patient population changes. Additionally, we chose straightforward cohort definitions for our training populations. High dimensional identification of patient cohorts would likely be more appropriate for developing personalized treatment plans, but our cohorts are chosen to more clearly demonstrate how population shifts occur in practice.

2.7 Conclusions

This study investigates bias in RL for heparin treatment planning. It leverages state-of-the-art model-based and model-free RL algorithms to successfully train RL-based heparin treatments. The study uncovers the subtle ways that bias can present itself in heparin treatment planning by leveraging specifically selected cohorts from a widely-used EHR dataset. As such, this study demonstrates the potential for bias in RL policies due to various data-related issues including domain shifts, and sample size/heterogeneity trade-off. No single RL policy used in this study seems to be fully immune to such biases across all data issues. As such, care must be taken when using

EHR data and RL to devise treatments. Furthermore, new RL algorithms need to be developed to further address the gap that currently exists to ensure that potential RL-based treatments are not biased.

In the future, it will be important to investigate the relationship between heterogeneity and sample size for RL policies. This could potentially include calculating new metrics to capture patient heterogeneity in training data to determine if a patient’s observations should be included during policy learning. Future work should also include developing an effective way to identify if a new patient is not well represented in the training data used to develop the models in order to prevent applying inappropriate treatment policies.

Chapter 3

Evaluating the Impact of Healthcare Data Completeness for Deep Generative Models

A version of this chapter was originally published by Benjamin Smith, Senne Van Steelandt, and Dr. Khojandi:

Smith, B., Van Steelandt, S., and Khojandi, A. (2023). Evaluating the Impact of Health Care Data Completeness for Deep Generative Models. *Methods of Information in Medicine*, 62(1-02), 31–39. <https://doi.org/10.1055/a-2023-9181>.

Abstract

Background: Deep generative models (DGMs) present a promising avenue for generating realistic, synthetic data to augment existing healthcare datasets. However, exactly how the completeness of the original dataset affects the quality of the generated synthetic data is unclear.

Objectives: In this paper, we investigate the effect of data completeness on samples generated by the most common DGM paradigms.

Methods: We create both cross-sectional and panel datasets with varying missingness and subset rates and train generative adversarial networks (GANs), variational autoencoders (VAEs) and autoregressive models (Transformers) on these datasets. We then compare the distributions of generated data with original training data to measure similarity.

Results: We find that increased incompleteness is directly correlated with increased dissimilarity between original and generated samples produced through DGMs.

Conclusions: Care must be taken when using DGMs to generate synthetic data as data completeness issues can affect the quality of generated data in both panel and crossectional datasets.

Keywords: data quality, data completeness, case completeness, missingness, deep generative models

3.1 Introduction

The impact of data-driven healthcare research can be limited due to the volume, quality, and privacy concerns of available datasets [Chen et al. \(2021\)](#). Therefore, using existing healthcare datasets and statistical models to generate additional synthetic samples is considered a potential avenue for furthering healthcare research [Wang et al. \(2021\)](#). While synthetic data can improve model performance, limitations due to the original dataset still arise. In healthcare, this relates to patient representation, missing data, and the potential for biases in treatments [Bhanot et al. \(2021\)](#). Our work directly compares synthetic data produced through popular and highly regarded deep generative models (DGMs) [Kusner et al. \(2017\)](#); [Isola et al. \(2017\)](#), when using real healthcare data of different qualities. Our work demonstrates the limitations in accuracy when generating data using incomplete datasets and explores the mechanisms with in which data completeness affects generative model performance.

DGMs are a family of deep learning architectures that use deep neural networks to infer high dimensional distributions from large datasets [Eigenschink et al. \(2021\)](#). They have displayed state-of-the-art performance across a variety of tasks where mapping these complex distributions is essential, such as speech generation and compression [Huzaifah and Wyse \(2021\)](#). Similarly, DGMs have increasingly been applied in healthcare to generate synthetic data sampled from the inferred probability distribution as a means of augmenting existing datasets or creating synthetic, fully anonymous datasets [Piacentino et al. \(2021\)](#).

The validity of inferred probability distributions highly depends on the quality of the dataset that the DGM is trained on [Raab et al. \(2017\)](#). The limited size of medical datasets, the underrepresentation of minorities, and missing values resulting from patient dropout are only a few of the issues that have been thoroughly described to affect the completeness of datasets [Chen et al. \(2021\)](#). This concept of completeness is typically quantified through two components, namely, case completeness and missingness [Weiskopf et al. \(2013\)](#).

Case completeness refers to a dataset’s coverage of the true population distribution. In other words, it captures the extent to which the sample dataset contains sufficient randomly collected observations such that the model of interest can reconstruct a dataset with the true population statistics [Burkhart and Androwich \(2004\)](#). Case completeness, from a practical perspective, somewhat depends on the presence of a diverse sample population in a dataset. Failure to adequately sample patient observations would bias the resulting generated data from representing characteristics of all different patient types [Wright et al. \(2015\)](#). That is, a small sample size that does not include all patient types would limit the effectiveness of generative models.

Missingness refers to the completeness of the recorded data. Observed features and/or characteristics can have different rates of missingness depending on how often observations were not appropriately recorded in a dataset [Burkhart and Androwich \(2004\)](#). Unless all observations are available, depending on the missingness rate and pattern, incomplete observations can either be removed from the dataset or missing values can be estimated and replaced through imputation [Beaulieu-Jones et al. \(2017\)](#). Increased missingness directly correlates with uncertainty and decreased quality of models built using incomplete observations.

There’s a vast body of healthcare research that has highlighted how the performance of deep neural networks is affected by case completeness and missingness. However, these studies usually focus on these aspects of data quality separately. In addition, existing studies focus almost exclusively on the discriminative paradigm. The generative paradigm remains rather unexplored, and while attempts have been made to either understand the effect of small datasets or missing value on the performance of DGMs, few have looked at the joint relationship, particularly in the context of healthcare.

To the best of our knowledge, this study is the first to examine the combined effect of case completeness and missingness on the ability of DGMs to correctly approximate the data’s probability distributions in a healthcare context. We demonstrate the

challenges associated with generating synthetic samples from incomplete data using three of the most popular and powerful DGM approaches, namely, generative adversarial networks (GANs), variational autoencoders (VAEs) and autoregressive models (Transformers) [Kusner et al. \(2017\)](#); [Isola et al. \(2017\)](#); [Katharopoulos et al. \(2020\)](#); [Vaswani et al. \(2017\)](#).

3.1.1 Motivation

Completeness is one of the most significant issues in healthcare data and directly limits the effectiveness of treatment models built using the data (e.g., using machine learning models) [Hilsenbeck et al. \(1992\)](#); [Kodra et al. \(2017\)](#). While synthetic data can augment healthcare data sources to improve the completeness, their efficacy may be limited by the quality of original data [Reiter \(2004\)](#). Our work focuses on the quality of generated samples from incomplete datasets, rather than the accuracy of treatment models built using the data. Because our work concerns the quality of generated samples, we compare the data generated through DGMs as a function of case completeness and missingness.

Case completeness in training data inhibits DGMs from generating diverse synthetic samples [Wright et al. \(2015\)](#). Completeness, in this way, requires datasets to adequately contain instances of different patients’ characteristics or responses to treatments [Raab et al. \(2017\)](#). Failure to include all patient types in healthcare data prevents DGMs from generating synthetic samples to represent minority populations [Weiskopf et al. \(2013\)](#). Poor case coverage in training data may still result in synthetic data that *on average* are similar to the training population. However, with fewer samples for DGMs to learn from, synthetic samples are likely more similar to ‘average’ patients in the training data rather than being representative of the true patient population [Bhanot et al. \(2021\)](#). This trend results in bias [Dietterich and Kong \(1995\)](#) by favoring certain cases to generate synthetic samples [Wang et al. \(2020\)](#).

Missingness refers to missing observations across features in a dataset [Little et al. \(2012\)](#). Missing observations in data generally necessitate increased approximations when training generative models [Faris et al. \(2002\)](#). Because missing values are often estimated or ignored, DGMs may learn inaccurate distributions of the data and result in synthetic data that are statistically biased [Kelly et al. \(2017\)](#). Missingness has been shown to decrease model performance in discriminative and generative models [Markey et al. \(2006\)](#). Importantly, data missingness limits DGMs’ ability to consistently generate samples similar to true patient populations [Li et al. \(2019\)](#); [Hu et al. \(2021\)](#). As such, the observed variance generally limits the effectiveness of synthetic data to augment incomplete healthcare datasets.

We investigate the impact of completeness in healthcare data for DGMs in the dimensions of case completeness and missingness. Case completeness in healthcare data most relates to the patient sample size, and incompleteness in this way results in DGMs that fail to represent minority patient populations [Xu et al. \(2021\)](#). Missingness relates to the statistical definition of bias due to increased approximations required when generating synthetic data [Feldman et al. \(2017\)](#). These dimensions of completeness have not been explored in unison for generative healthcare data [Mattei and Frellsen \(2019\)](#). Previous work has considered these issues in classification algorithms or for generative methods in one dimension of completeness. In this study, we demonstrate how case completeness and missingness affects generative models’ effectiveness in generating appropriate synthetic samples. Figure 1 provides an overview of our study framework.

3.1.2 Objectives

This study examines the impact of completeness in healthcare data on the quality of synthetic samples produced from popular DGMs. The study demonstrates this impact using real-world cross-sectional and panel healthcare datasets. Specifically, we evaluate the similarity of produced synthetic data to the true, complete data after

introducing different rates of case incompleteness and missingness to the original data. As such, our study demonstrates the potential challenges of using incomplete healthcare data for generating synthetic samples.

In the remainder of this paper, we first describe the datasets and DGMs used in our analysis. We then provide an overview of our evaluation process. We next present the results for DGMs applied to datasets at different rates of completeness. Lastly, we discuss the implications and limitations of our analysis and comment on future work.

3.2 Methods

3.2.1 Dataset Description

The datasets used for the experiments in this paper are derived from the MIMIC III database, a peer-reviewed and widely used database containing a wide array of deidentified data on ICU patients staying at the Beth Israel Deaconess Medical Center between 2001 and 2012 [Johnson et al. \(2016\)](#).

For our experiments, we focus on those patients that received heparin during their stay in the intensive care unit (ICU). Heparin is an anticoagulant that is frequently administered to patients to treat various medical conditions, such as, among others, the prevention of blood clots [Baucum et al. \(2020\)](#). We separate the available data of heparin patients to create two datasets. The first dataset contains the demographic cross-section of patients that receive heparin treatment in the ICU, with features consisting of age and height. The other dataset contains, for each patient, the average heart rate per 4-hour interval during the first 40 hours of heparin administration. The details of the variables in the datasets and their descriptions are provided in [Table 3.1](#). Of all the variables of interest, these variables are particularly included in our demonstrations because of their high rate of completeness in the MIMIC III dataset. In this way, we are able to construct real-world sample datasets to which we can

introduce various rates of completeness for evaluating the corresponding impact on DGM performance.

We perform minimal cleaning and pre-processing, solely consisting of normalizing the data and removing the incomplete cases from both datasets. Following this process, a total of 3,000 patients are present in each dataset.

For both datasets, we leave out 10% of observations for validation, totaling 300 samples. From the remaining 90%, we create 42 cases, each differing in their number of observations (7 levels) and missingness rate (6 levels). Specifically, we consider subset rates of 1.6%, 3.2%, 6.0%, 12.5%, 25%, 50%, and 100% to create cases ranging from 48 to 3,000 in sample size and consider missingness rates of 0%, 2%, 5%, 10%, 25%, and 50%. The missingness rate is introduced by censoring the original dataset based on a missing completely at random (MCAR) pattern.

3.2.2 Deep Generative Models

In this study, we use three deep generative frameworks, namely GANs, VAEs, and Transformers. For the purpose of generalization, we focus on the generic implementation of each DGM and not on task-specific implementations that can potentially achieve better performance at the cost of generalization.

GANs use two competing neural networks to generate realistic synthetic data. GANs are made of a generator network that is trained to take in random noise and transform it into realistic samples. Alternatively, the discriminator network takes in both real and generated samples and attempts to classify each. These two networks train against each other until the classification of the discriminator is around 50%, meaning that it can no longer distinguish between real and generated data samples. At this point, the generator network can be used to create synthetic observations of a particular data type. GANs have been used extensively in healthcare, achieving state-of-the-art performance on tasks such as synthetic health record generation [Torfi](#)

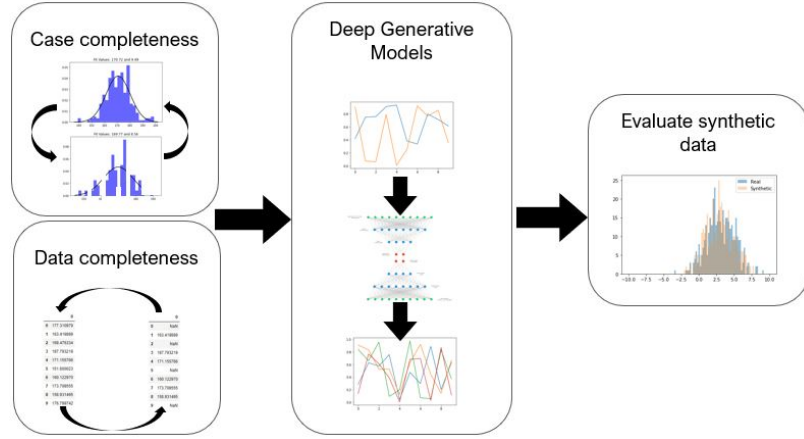


Figure 3.1: The overview of the study framework. First, the two components of completeness introduced to datasets used to train deep generative models to create synthetic samples. Next, these synthetic data are evaluated for similarity of generated samples to true data.

Table 3.1: Summary of the features in the sample real-world datasets, drawn from the MIMIC database, used in this analysis.

Dataset	Variable	Description
Demographic	Age	The age of the patient when admitted to the hospital.
	Height	The height of the patient when admitted to the hospital.
Physiological	Heart rate	The average heart rate of the patient over a 4-hour interval.

and Fox (2020), data imputation Suo et al. (2019) and image generation Lee et al. (2019).

VAEs are based on two neural networks: An encoder and a decoder. The encoder takes in real data and attempts to represent these data as a distribution over the latent space. The decoder then attempts to re-represent the encoded data in the original, real data space. The joint optimization of these two networks results in a model that can generate realistic synthetic samples from real datasets.

Transformer models are neural networks that learn relationships to track sequential data. Transformers are most commonly used in speech recognition and generation and use a term called self-attention to learn the relationship between values such as words in a sentence. Transformers provide a computationally efficient method for learning temporal relationships between data points. Transformers have been applied to solve a variety of predictive healthcare tasks, including opioid use Fouladvand et al. (2021), Covid-19 Shome et al. (2021), suicide risk Salmi et al. (2022) and asthma exacerbation prediction Zeng et al. (2022), as well as an increasing number of generative tasks, such as clinical text generation Amin-Nejad et al. (2020).

3.2.3 Model Evaluation

In our context, generative models’ performance is determined by their ability to recreate the original dataset that they are believed to represent. In other words, if we sample from these generative models, we expect that for a well-performing model, given sufficient sample size, the distribution of these samples to be similar to that of the original population.

Thus, unlike the popular point prediction where a score can be calculated for each observation, we examine the entire generated dataset against a validation dataset (without any missing values) to understand a DGM performance. To do this, we sample a new dataset from a learned DGM, equal in size to that of the validation dataset against which we want to evaluate our DGM performance. We then calculate

the Mean Absolute Error between sample vectors from the generated dataset and the validation dataset, where the samples are paired such that the summed minimum Mean Absolute Error across all pairs is minimized [Jonker and Volgenant \(1987\)](#); [Kuhn \(1955\)](#). We use this summed distance between each pair of points, which we refer to as the Minimum Mean Absolute Error (MMAE), as a measure of a model’s ability to recreate the original dataset. The MMAE is mathematically represented by,

$$\min_{i,j} \sum_{i=1}^n |y_i - \hat{y}_j|$$

where y_i and $\hat{y}_i$ *i*th *paired* samples from the validation and generated datasets respectively, and n denotes the number of samples in each of the two datasets.

Because the MMAE represents the shortest possible distance between the validation and generated datasets, we use it throughout this paper as a metric to describe the similarity between the datasets, where the higher the MMAE, the lower the similarity.

Throughout this manuscript, we present the 95% confidence interval for the results, using the t distribution instead of the normal distribution due to the small cross-validation sample size. Figure 2 summarizes our study design and model evaluation scheme.

3.3 Results

We compare generated, synthetic samples with original data at varying both missingness and subset rates. We implement the VAE and GAN for both the cross-sectional and temporal dataset, while the Transformer model is only implemented for the temporal dataset because this model is only generative when modeling sequential data [Gao et al. \(2020\)](#).

Figure 3 presents the MMAE scores for VAE and GAN models when applied to patient demographic data across varying missingness and subset rates. In general,

for any given missingness rate, the MMAE score decreases (i.e., the performance increases) as the subset rate (hence, the sample size) increases for both VAE and GAN. Similarly, in general, for any given sample size, the MMAE score increases (i.e., the performance decreases) as the missingness rate increases for VAE. Interestingly, this relationship is less pronounced in GANs where for any given sample size, the MMAE score is less sensitive to the missingness rate. This suggests that GANs may be generally more robust against missingness than VAEs in the dataset examined. Despite this decreased sensitivity to missingness, GANs were consistently outperformed by VAE models as shown with higher MMAE scores, particularly at low subset rates.

Figure 4 presents the MMAE scores for GAN, VAE, and Transformer models when applied to temporal heart rate data across varying missingness and subset rates. Similar to the observations made in Figure 3, in general, for any given missingness rate, the MMAE score decreases (i.e., the performance increases) as the subset rate increases for VAEs. For GANs, MMAE scores do not decrease until the subset rate is greater than 12.5%. Interestingly, this pattern is less pronounced for Transformers. That is, at very low subset rates (i.e., at 1.6% and 3.2% where the resulting samples are too small), the MMAE score is rather large. However, for larger subset rates (i.e., larger sample sizes), the MMAE score is consistently lower, meaning that Transformers are generally less sensitive to sample size as long as it is not too small. Furthermore, interestingly, in general, for any given subset rate, the MMAE scores are relatively consistent across various missingness rates for all three models. This relationship (or lack thereof) is more pronounced in VAEs where for any given subset rate, the MMAE score is rather insensitive to any changes in the missingness rate. Overall, this suggests that both VAEs and Transformers are generally robust against missingness rate in the dataset examined.

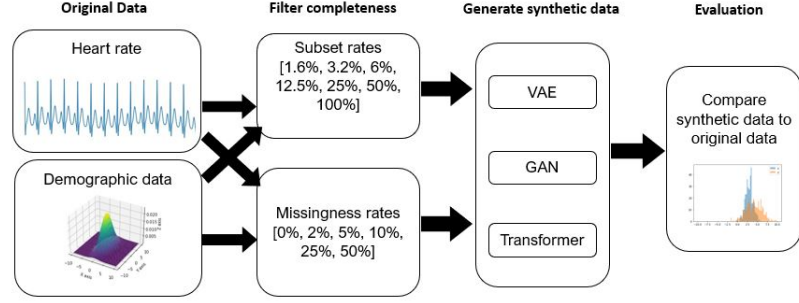


Figure 3.2: Study design. [GAN: Generative adversarial network, VAE: Variational autoencoder, Transformer]

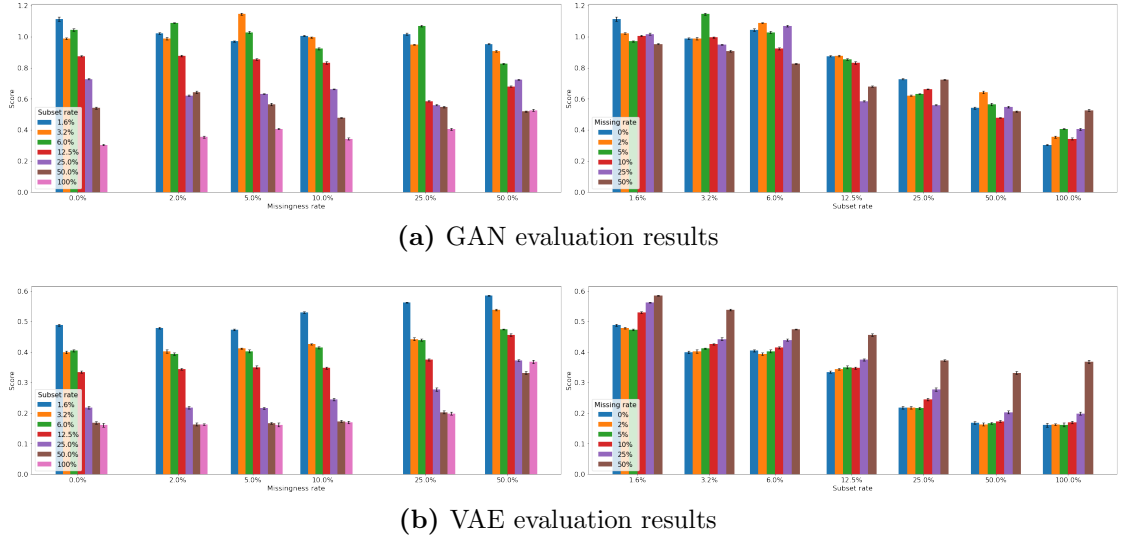
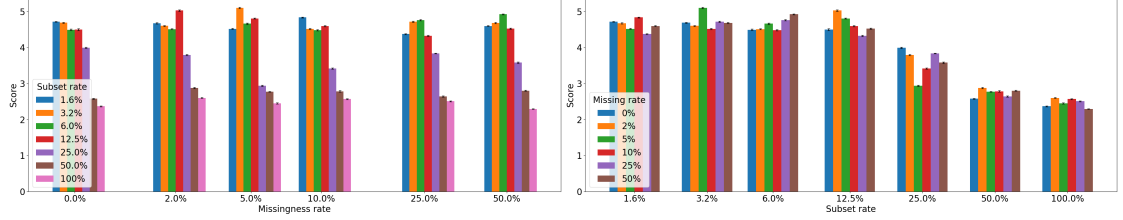
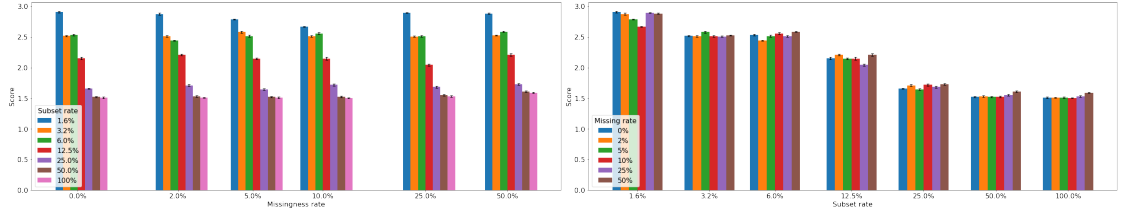


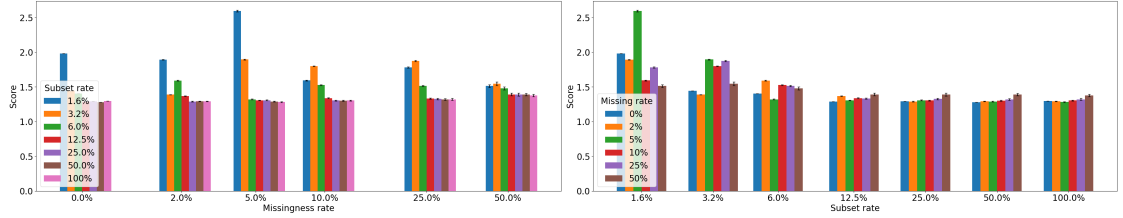
Figure 3.3: MMAE scores with the heart rate dataset using GAN, VAE, and Transformer models from both perspectives of missingness and subset rates.



(a) GAN evaluation results



(b) VAE evaluation results



(c) Transformer evaluation results

Figure 3.4: MMAE scores with the heart rate dataset using GAN, VAE, and Transformer models from both perspectives of missingness and subset rates.

3.4 Discussion

In general, the MMAE score increases (i.e., the performance decreases) as the sample size decreases, although differing significantly between model and data types. Compared with the demographic data, the heart rate data are generally less affected by sample size. That is, the VAE and GAN models trained on demographic data present a stronger effect of sample size on model performance, as compared with the VAE, GAN and Transformer models trained on heart rate data.

Second, relative model performance is in line with the current literature. For the demographic data, VAEs outperform GANs, which are notoriously data inefficient [Ghadirzadeh et al. \(2020\)](#). Similarly, for the heart rate data, the Transformer model outperforms the VAE, while the GAN is the worst performing model. Interestingly, the the Transformer model shows minimal degradation in performance as the sample size decreases.

The relationship between the missingness rate and the MMAE is less pronounced. For the demographic dataset, the VAE models show a positive relationship between the missingness rate and the MMAE score, especially for the larger sample sizes, while the GAN models do not show any strong relationship between MMAE scores and missingness. For the heart rate dataset, no strong relationship between the missingness rate and the MMAE score is observed.

The mechanisms of completeness generally affect DGMs similarly based on the type of data considered. Demographic data are consistently more affected by smaller sample sizes than heart rate data. This suggests that temporal heart rate data can be more easily represented with fewer samples (e.g., where sensor readings may be missing every now and then) than demographic data, and that they can more easily be augmented with generative models. Demographic data are generally more representative of patient population coverage, which is why small sample size appears to cause more of an issue. In this way, case completeness may be a more important concern when learning from non-temporal data.

While preliminary, this work directly relates to the impact of data-driven healthcare research. Synthetic sample generation may improve model performance when learning from incomplete datasets, but the quality of synthetic data generally decreases as the completeness of training data decreases. Importantly, our work highlights the differences in how the mechanisms of completeness affect generative models’ performance using healthcare data of different types. As such, data-driven population-level research should be more considerate of case completeness when using generative models.

3.5 Conclusions

We successfully demonstrate the impact of the completeness of healthcare data on the DGMs used to create synthetic samples. We observe the effects of completeness in the dimensions of case completeness and missingness by introducing these issues to a complete training data and comparing the similarity of the generated data with the ground truth. We implement multiple DGMs on two different real-world datasets designed to represent two common types of healthcare data.

In addition, we use relatively naive implementations of DGMs to show how different types of generative models are affected. Despite these limitations, we effectively study both components of completeness for their effect when generating synthetic samples to augment existing healthcare data.

Future work may consider extending this investigation of data quality beyond completeness. Specifically, it may consider measurement error or the effect of granularity of time series data. This may shed further light on the nuances of data quality’s effect on the generation of synthetic samples through DGMs.

Chapter 4

Towards Automating Ferroelectric Domain Wall Control using Reinforcement Learning

A version of this chapter is in preparation for journal submission by Benjamin Smith, Bharat Pant, Dr. Yongtao Liu, Dr. Yu-Chen Liu, Dr. Jan-Chi Yang, Dr. Stephen Jesse, Dr. Khojandi, Dr. Sergei V. Kalinin, Dr. Ye Cao, and Dr. Vasudevan:

Abstract

Elucidating the interactions and dynamics of domain walls in ferroelectrics is critical on both a fundamental level, for understanding interfacial dynamics in disordered media, but also on a more practical level, as it can enable engineering of new metastable states with novel properties. Piezoresponse force microscopy (PFM) enables both imaging and writing of ferroelectric domain walls via a biased scanning probe. However, control over positioning of individual domain wall segments to engineer domain wall structures over large areas reproducibly remains challenging. Here, we present an experimental workflow via autonomous PFM that enables automated data collection of domain walls interacting with defects in prototypical (110) $PbTiO_3$ thin films. These data are used to construct physics-informed surrogate models of local domain wall response in response to applied electric fields by the PFM tip and are verified using phase-field simulations. The surrogate enables generation of ‘phase diagrams’ of the domain wall, conditional on initial structure, and highlights the importance of domain wall pinning and elastic effects on repeated wall modification attempts. Subsequently, reinforcement learning is used to optimize tip-modification trajectories for obtaining desired domain wall structures in simulated environments utilizing the surrogate model for the environment dynamics. This workflow shows how automated data collection and autonomous agents can be orchestrated towards realizing domain wall manipulations with precision in scanning probe studies, and how such surrogates can aid in understanding domain wall-defect interactions in ferroelectrics.

4.1 Introduction

Domain walls in ferroelectric and ferroelastic materials have garnered attention recently as standalone elements within the bulk material with unique static and dynamic properties [Meier and Selbach \(2022\)](#). By virtue of being topological defects within the order parameter field [Mermin \(1979\)](#), they possess different symmetry to the bulk they inhabit, as they must span degenerate polarization variants. Additionally, due to ferroelastic coupling, they are typically extremely thin, possessing high elastic energy [Yang et al. \(1999\)](#), and can be charged [Vul et al. \(1973\)](#), with possible changes in electronic band structure at the wall [Seidel et al. \(2009\)](#), and even their own phase transitions and 1-D defects [Nataf et al. \(2020\)](#). Such phenomena can lead to interesting properties, including changes in AC and DC conductance compared with the bulk [Meier and Selbach \(2022\)](#); [Vasudevan et al. \(2013\)](#); [Schultheiß et al. \(2022\)](#); [Tselev et al. \(2016\)](#), significant contributions to static permittivity [Xu et al. \(2014\)](#), large extrinsic contributions to the piezo- and dielectric coefficients of ferroelectric ceramics [Xu et al. \(2001\)](#); [Li et al. \(1993\)](#) and changes to opto-mechanical (e.g., photostrictive [Rubio-Marcos et al. \(2023\)](#)) behaviors. Given the breadth of possible novel properties that can be realized at ferroic domain walls, and the fact that it is possible to write domain patterns, inject domain walls [Whyte et al. \(2014\)](#), and control polarization orientation with the scanning probe of an atomic force microscope [Balke et al. \(2009\)](#), this has led towards whole new areas of domain wall engineering and domain wall nanoelectronics [Catalan et al. \(2012\)](#), seeking to tailor these properties for desired materials applications.

The engineering of domain walls and topological defects with the SPM probe has been ongoing for more than two decades, with early work exploring simple domain poling and observations of events such as the pinning of written ferroelectric domain walls by interfaces [Ganpule et al. \(2002\)](#); [Harnagea \(2001\)](#) as well as the occurrence of faceting [Ganpule et al. \(2001\)](#) and aging effects [Shvartsman et al. \(2005\)](#). The ability to pattern more in-plane polarization orientations was opened up by several

studies that showed how the tip-induced electric field’s radial symmetry could be broken, either by motion, as demonstrated for instance by [Balke et al. \(2009\)](#) to write more complex topological patterns with the SPM tip, or by pre-existing defects, e.g. by [Vasudevan et al. \(2011\)](#), who used pulsing at domain walls to produce novel topological defect states. Device-level methods were also discovered, including strategies to “inject” domain walls into ferroelectric devices for memory applications which were explored via [Whyte et al. \(2014\)](#), while [McGilly et al. \(2015\)](#) utilized resistive Pt electrodes to enable writing of domain walls into capacitor structures, with the goal of forming circuitry. Many more studies through the years have explored other methods of tip-based writing of ferroelectric domain walls and more complex domain configurations [Kim et al. \(2019\)](#); [Li et al. \(2017, 2022b\)](#); [Damodaran et al. \(2017\)](#); [Chen et al. \(2019\)](#).

In all cases, the tip follows a user-defined (implicit, e.g., via masks, or explicit) trajectory, and voltage is applied to the tip for specific moments to induce changes in the underlying domain structure. However, this procedure requires human interventions to determine the correct locations to position the tip, and the right values of the excitation pulse to apply to facilitate the desired domain structure modification. This comprises a sequential decision-making task that is conditional on the pre-existing domain structure, the existence of defects, and the current state of the SPM tip, along with the type of domain wall itself, all of which can change spatially, temporally, or both.

It should be noted that the response of domain walls to applied electric potentials has been well studied not only experimentally, but also through a variety of different theoretical means. Numerous works have explored the problem in both pristine [Aravind et al. \(2010\)](#) and defect settings [Yang et al. \(1999\)](#); [Kalinin et al. \(2007\)](#) for example analytically with Landau theory, or through numerical solutions via phase-field methods [Chen \(2008\)](#) and even reactive force fields [Akbarian et al. \(2019\)](#), exploring everything from the impact of pre-existing domain walls on switching [Jesse et al. \(2008\)](#), switching in the vicinity of grain boundaries [Rodriguez et al. \(2009\)](#) and

individual defects [Kalinin et al. \(2008\)](#), to ferroelastic wall deformations [Anbusathaiah et al. \(2010\)](#) and switching of hierarchical domain patterns [Ivry et al. \(2012\)](#). However, to date, the models used are generally compared semi-quantitatively (at best) with experimental observations in piezoresponse force microscopy, largely due to unknowns in tip condition, presence of defects, etc. In order to facilitate appropriate predictions to guide decision-making for autonomous domain wall engineering in ferroelectrics via SPM, what is needed is a reliable model that can predict changes to domain wall structure for arbitrary applications of applied voltages. Such a model could then be used within existing machine learning algorithms for sequential planning and decision-making tasks, such as reinforcement learning (RL), to create agents that can autonomously create and measure novel domain wall patterns in ferroelectrics. In the process, the surrogate model generated can be interrogated for yielding insights into the dynamics of domain walls within the sample.

Here, we present such a workflow towards achieving autonomous domain wall manipulations, showcasing the ability to automatically manipulate domain walls in a prototypical (110) $PbTiO_3$ thin film, develop a physics-informed dynamics model from the ensuing data, and then use this model to both better understand wall displacements in the sample, as well as train RL agents to develop strategies for wall manipulation to achieve desired outcomes. We verify our physics-informed model with phase-field simulations, and find qualitative agreement on wall displacement profiles, and further note the ability to construct functional ‘phase diagrams’ of wall structure as a function of input actions.

4.2 Overall workflow

The overall workflow is illustrated in Figure 4.1(a). The experiments are performed with a commercial AFM system (Oxford instruments, Cypher) with a Pt/Ir coated AFM tip (BudgetSensors) on a 200nm-thick $PbTiO_3$ thin film grown on (110)-oriented $SrTiO_3$ with a small buffer layer (electrode) of $La_{0.7}Sr_{0.3}MnO_3$. We begin by imaging

a virgin region and then poling a domain wall in the film, by applying, alternately, -8V and +8V to the AFM tip while scanning, on the left and right side, with the sample grounded through the bottom LSMO electrode. The sample is then imaged again, either with single frequency or band-excitation piezoresponse force microscopy (PFM), and then the position of the written domain wall in the image is extracted from the PFM phase image. This generates the ‘initial’ state of the system which consists of a single written domain wall through the middle of image. Although the written wall is nominally a 180° domain wall, given the (110) orientation of the film, this means there are both vertical and lateral changes to polarization across this wall.

The tip is then moved to a random location along this domain wall, and then a bias pulse is applied. The bias parameters are random, uniformly sampling in the interval (-10,10) V for the voltage amplitude, and (50,500)ms for the pulse width. The sample is then imaged again with PFM. This sequence (image, action, image) is repeated a set number of times until the wall is reset, i.e., until the wall is rewritten in the original configuration. Note that in our dataset, we collected one dataset where the reset frequency was 10, i.e., every ten actions, the wall would be rewritten, as well as a second dataset where there was no resetting applied. These transitions are stored and then used subsequently to train a dynamics model, which is used as the basis for a reinforcement learning environment that enables agents to learn how to modify domain walls in the system towards desired morphologies.

This experimental procedure is coded with python into a Jupyter notebook, which communicates to an FPGA device (Figure 4.1(b)) that then executes the workflow steps on the microscope. Details of the FPGA device operation will be provided elsewhere, but essentially, this acts as an alternate controller for the microscope enabling customized scan paths and excitations to be applied and controlled via python scripts. An example of the initial image, the bias pulse location and parameters, and the result of the bias pulse, are shown in Figure 4.1(c). It is observed that applying voltages above a certain threshold produces wall displacements, as expected. Although the wall writing procedure produces nominally a 180° domain

wall in these films, with both vertical and lateral components to the polarization; however, we are only probing the vertical PFM signal for simplicity. Note also that numerous ‘needle’ domains can be seen in the vicinity of the written wall.

4.3 Dynamics Model

Our experiment provided us with 801 usable transitions, i.e., circumstances where the imaging quality was sufficient for use. We developed a dynamics model to predict the displacement of domain walls structures when bias pulses are applied via PFM. The workflow of the dynamics (aka surrogate) model is provided in Figure 4.2. PFM is used to generate training data for the model that contains sequences of images, and their corresponding actions. Each image is represented by a 128x128 pixel array from which a 128-pixel length 1-D domain wall vector S_t is extracted. Actions are given by three continuous normalized values that correspond to bias location, voltage, and pulse width in that order. Since the wall displacement mostly occurs in the section of the wall around where the bias was applied, the dynamics model only predicts displacements for the wall in these local regions. Therefore, the first action value (the one that corresponds to the location of the bias pulse) is thus used to index the local region of the domain wall. Around the location bias is applied, a local region of the wall is extracted as a 14-pixel length vector, represented by S'_t in Figure 4.2. The inputs to the surrogate model are the local region S'_t as well as the voltage and pulse width applied. Using these inputs, a neural network with two branches works to predict the displacement for the domain wall in the local region. The first branch takes the domain wall configuration, i.e., S'_t , while the second branch takes the bias amplitude and pulse width as inputs, and both contain only dense layers. These branches are subsequently concatenated before being passed through two more dense layers to obtain the output. The predicted displacements in local region S'_{t+1} can then be added to the original entire wall structure to generate S_{t+1} . Because images are also captured immediately following when the bias is applied, our model’s

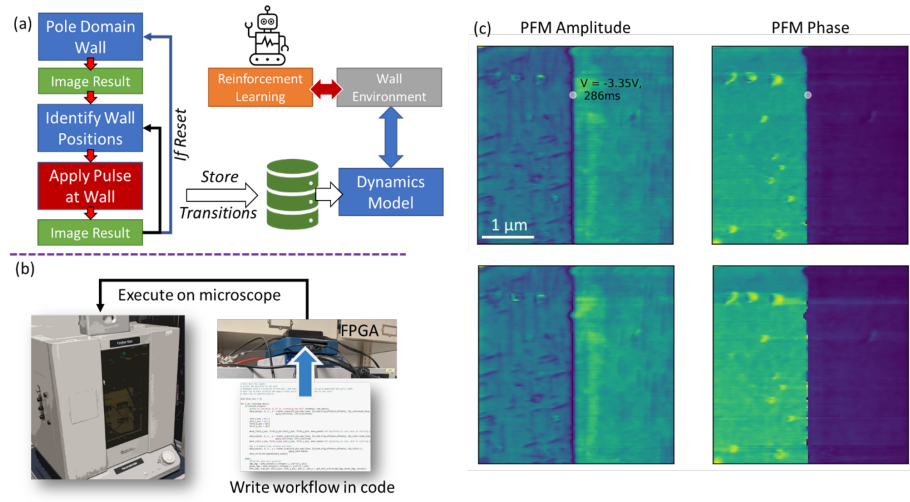


Figure 4.1: (a) Overall workflow combining experiments with computation. The images (from which *states* are derived) are in green; *actions* are in red. (b) A Jupyter notebook with code cells is used to communicate to a home-built program running on an FPGA, that an enables instrument control with fast inputs and outputs for automated experiments. (c) Example image collected with this workflow, with the original image above and the bias details and location indicated by the white circle. After applying the bias pulse, the sample is reimaged, and clearly shows a displaced domain wall at the location of pulsing.

predicted domain wall structures can be compared to the actual, observed domain wall structures. This allows the dynamics model to be trained through a supervised learning method, with a traditional loss function such as mean squared error (MSE).

4.3.1 Physics based loss function

Due to the limited data available, we found it was necessary to introduce two additional regularizations into the loss function to prevent over-fitting and retain reasonable predictions in accordance with known or postulated physics of the system. We introduced two additional regularizers to the loss function for the dynamics model. The first additional term we add to the loss function is a term that emphasizes the agreement of the prediction with known local physics. In this experiment, the direction of displacement for the domain wall should align with the sign of the voltage for the bias that was applied. When this is not the case and the surrogate model’s predictions do not agree with this physical prior, such predictions should be penalized. Therefore, for any pixels in S'_{t+1} whose direction of displacement is opposite the sign (+/-) for the magnitude of the voltage amplitude applied, a penalty coefficient is multiplied to the difference between the observed and predicted wall locations. The mean-squared-error of S'_{t+1} is then calculated using this penalty to determine the total loss for S'_{t+1} which our dynamics model then attempts to minimize in its displacement predictions.

Additionally, we add an additional ‘global’ physics loss which is based on the expectation that the magnitude of displacement of the surrogate model should be monotonic for both the bias voltage and pulse width values. It should be noted that the increased displacement with increased voltage is consistently observed in the raw data. However, the same was not true for pulse width, potentially due to changes in tip condition or other latent variables. To account for this, we implement an additional term into the loss function that acts to promote the monotonicity of the dynamics model with respect to pulse width. During training, for each transition, we

also sample across the action space of pulse widths to make additional predictions. We then compare the predicted area displaced for each pulse width and add a penalty when the area does not either remain constant or increase for longer pulse widths i.e., when monotonicity is not observed. This loss function acts as a regularization term to help ensure that known physical trends are implemented into the dynamics (surrogate) model. In total, the loss function can be written as

$$L = \frac{1}{n} \sum (y - \hat{y})^2 + C_{local} \left(\frac{1}{n} \sum (y - \hat{y})^2 \right) + L_{global} \quad (4.1)$$

$$C_{local} = \begin{cases} 1 & \text{sign}(V) \neq \text{sign}(\hat{y}) \\ 0 & \text{otherwise} \end{cases}$$

$$L_{global} = \begin{cases} \frac{1}{s} & \int_0^n f_k(y) < \int_0^n f_{k-1}(y) \\ 0 & \text{otherwise} \end{cases}$$

where y is the actual domain wall position, $\hat{y}$ is the predicted position, and s is the training step number.

4.3.2 Dynamics model results

The dynamics model enables domain wall displacement prediction to be made for arbitrary wall shapes as well as arbitrary pulse amplitudes and pulse widths. As a point of comparison, we turned to physics-based modeling, i.e., phase field simulations for an understanding of how this system would behave for a few different bias parameters.

The results of the simulation are shown in Figure 4.3 (left) with the domain structure after 3.25V are applied shown in Figure 4.3(a) (right). Specifically, we start with a ground state of the (110) PTO thin film and then apply voltage to the AFM tip when it is placed at a pre-existing domain wall. This causes a large bulge in the domain wall which increases in size with increasing pulse voltage (Figure 4.3(a)).

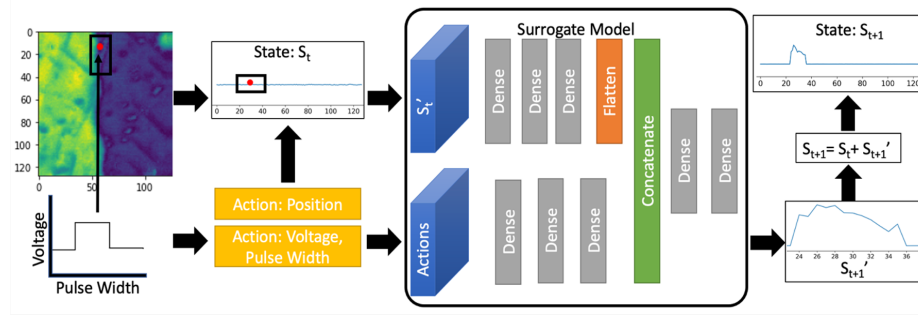


Figure 4.2: Overall workflow of the surrogate model. Domain wall structures are extracted from PFM images. Using these structures and bias pulse information, the model predicts the displacement of the domain wall in the region around where bias was applied. Dense layers for the S'_t branch contain 512, 256, and 256 nodes for each layer in order. For the action branch all layers have 256 nodes. After branches are concatenated, the dense layers contain 512 and then 256 nodes respectively.

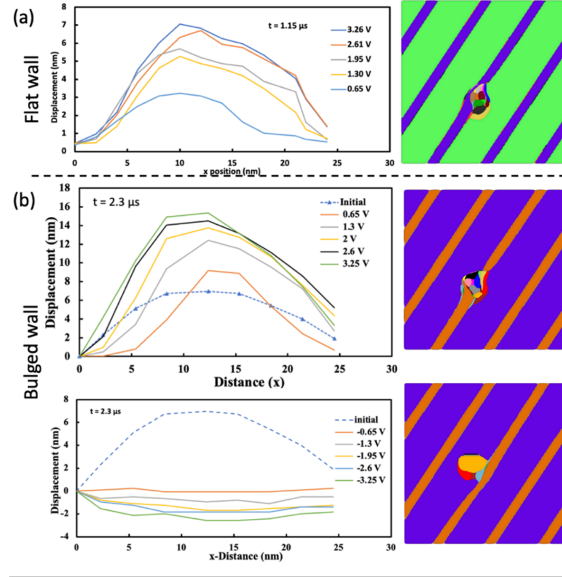


Figure 4.3: Predicted displacements for local domain wall regions using the phase field simulations sampled across a range of voltages and pulse widths. Displacements predicted for a ground-state wall are provided in (a) and the displacements from a wall already pushed in the positive direction are shown in (b). The domains' structure modeled by the phase field model after bias application is shown on the diagrams on the right-hand side. For (b), even a very small (-0.65V) pulse is sufficient to flatten the wall from a bulged state.

The predictions from our data-driven dynamics model are shown in Figure 4.4. The comparison with our dynamics model is shown in the right columns in Figure 4.4(a), which shows how the wall displaces from an initial flat profile after different bias voltages. Overall, we observe the maximum displacement increasing as the magnitude of voltage and pulse width increases as would be expected, though there is some discrepancy at higher voltage ranges. Low voltages also seem to not cause significant displacement, i.e. there is some pinning. The ‘phase diagram’ of the domain wall displacement as a function of the voltages and pulse widths is shown in the far-right column of Figure 4.4 and indicates (in general) increasing domain wall displacements towards higher voltages and pulse widths as would be expected.

At the same time, the dynamics model allows us to explore the reaction of the domain wall in arbitrary configurations, i.e., to investigate the state dependence. As a simple example, consider the circumstance when the domain wall is already bulged: one would expect a very different response to repeated pulsing in this state, given that the wall configuration is likely already very energetically unfavorable given it will prefer to be flat and reduce elastic energy. This circumstance is modeled in the phase-field simulations in Figure 4.3(b) and indicates that when starting with a positive displacement, increasing the positive voltage still does increase the wall displacement in the positive direction, but critically, applying negative bias to this situation very quickly returns the wall to a flat position, even for very low voltages, as seen in Figure 4.3(b). That is, there is an asymmetry where further growth is more difficult, whereas returning to a flat profile is more energetically favorable and occurs for lower voltages. This is also seen in the data-driven dynamics model, in Figure 4.4(b), and particularly in the ‘phase diagram’ on the right with brighter colors on the left side of the diagram indicating more propensity to quickly shift the wall to the negative direction, and darker colors on the right indicating resistance to further bulging the wall in the positive direction. These insights suggest that this method can be effectively used to explore the dependence of domain wall dynamics on the existing

domain wall configuration, providing another method for domain wall investigations via PFM and automated experiments.

4.4 Reinforcement Learning Environment

It is also possible to utilize the learned dynamics model as the basis for a reinforcement learning environment to train agents to manipulate domain wall structures *in-silico*. Notably, such a scenario would be prohibitively expensive with traditional phase-field methods, given that RL typically requires thousands to millions of state transitions for learning policies on meaningful tasks. Note also that while the surrogate model only acts in a local region, we integrate the model into an environment for the entire domain wall. Using this environment, we can then apply RL to learn policies for how to iteratively apply biases to alter the domain wall towards a target structure.

To learn policies to control the domain wall structure, we employ an Advantage Actor-Critic (A2C) RL algorithm as shown in Fig. 5. The A2C algorithm contains two networks: an actor and a critic network to learn through interactions with the environment as described by Mnih et al. (2016). The A2C algorithm takes in the state for the entire domain wall structure, S_t , and the actor network chooses an action. The chosen action is then applied to the environment which uses the dynamics model to output the updated state of the domain wall and the return of the reward function. The environment’s reward function is calculated by the negative mean-squared error between the current wall structure and a pre-defined target wall structure. Based on the reward returned by the environment, the critic network indicates the value of the action in that state, and the actor adjusts its policy based on this feedback from the critic. More specifically, the critic calculates the advantage of each action taken. The advantage refers to difference in estimated value for the chosen action in each state compared to the average value of all possible actions that could be taken in that state this is detailed by Peng et al. (2018). As the wall structure moves closer to the target wall by applying biases, the reward function approaches 0. The algorithm attempts

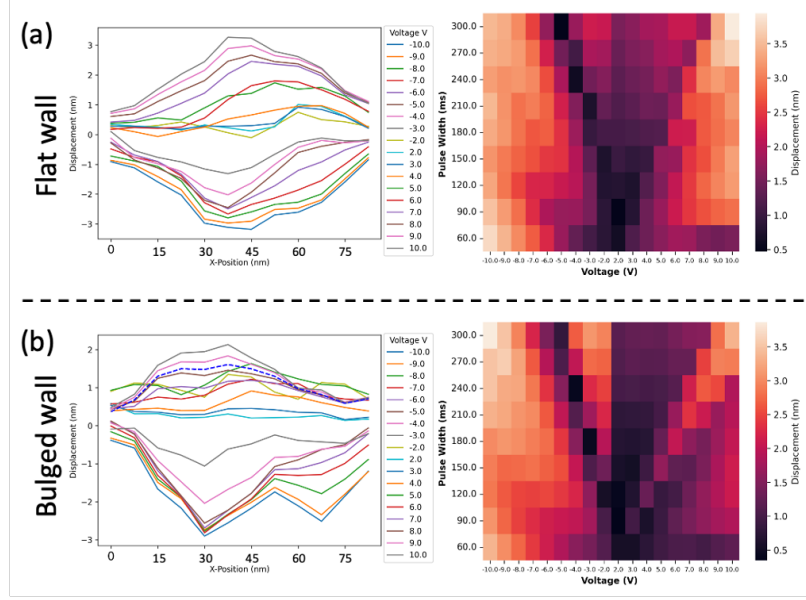


Figure 4.4: Predicted displacements for local domain wall regions using the dynamics model, sampled across the range of voltages and pulse widths. Displacements predicted for a relatively flat wall are provided in (a) and the displacements from a wall already pushed in the positive direction are shown in (b). the considerable difference between the two shows that conditional on the current wall structure, the expected displacements vary significantly, likely due to elastic energy penalties of the already displaced wall.

to maximize rewards which, in this case, is the least negative reward that indicates that the wall positions in S_t are closest to the ideal wall structure.

When training the RL policies there are several different environmental hyperparameters and training methods available to modify. In the environment, it is straightforward to modify the target wall structure and the number of steps available to the agent in any episode before the simulation resets. Additionally, the initial starting wall configuration can be changed to so that the policy must learn to act appropriately for different wall structures. The rewards can either be calculated after each step or just based on the final wall structure, after all steps have been taken in each simulation. Our policy is trained using either the traditional policy gradient (A2C) approach, or in some cases, the Stein Variational Policy Gradient (SVPG) method introduced by [Liu et al. \(2017\)](#). SVPG allows for simultaneous training of multiple policies modeled as a distribution of policy parameters and has shown to have state of the art performance for continuous control problems. Within SVPG optimization, a hyperparameter exists called ‘temperature’ that can be changed to control the amount of policy exploration compared to exploitation in the parameter space. Using SVPG optimization methods, we train policies on an NVidia DGX-2 server utilizing one Tesla V-100 GPU.

4.4.1 Reinforcement Learning Results

We train policies for multiple target wall structures and number of pulses that can be applied. First, the simulations begin with a relatively flat wall and using an RL policy, actions are taken to move the wall towards the target wall structure. An example of the resulting domain wall structure and actions chosen by the policy are shown in Figure 4.6(a). We observe the agent learns to take actions in regions where the ideal wall is further away from the original wall (i.e., where rewards can increase most significantly). As shown in Figure 4.6(a), we see that the agents learn to repeatedly

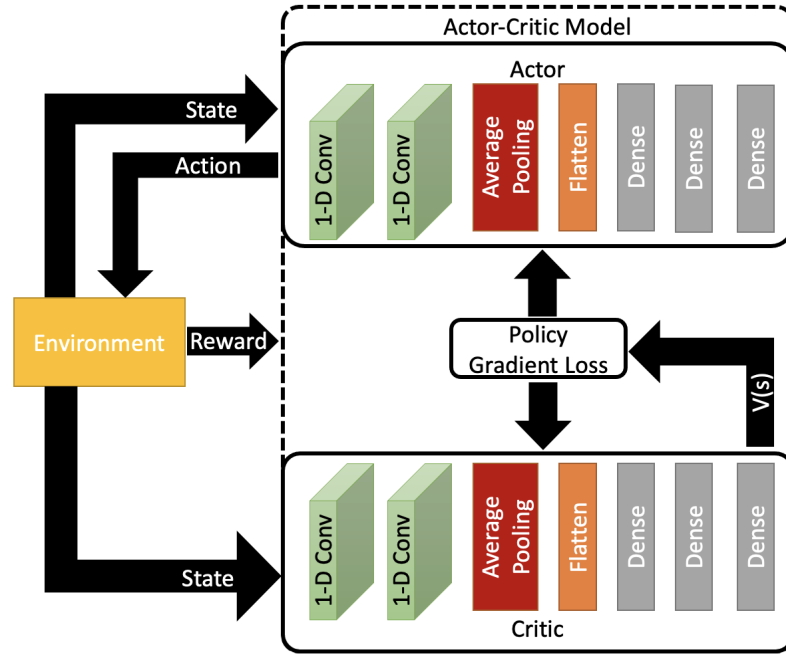


Figure 4.5: Overall design of the A2C model. The actor chooses the location, voltage, and pulse width that bias is applied, and the environment employs the surrogate model to predict the updated domain wall structure. The environment returns rewards to the A2C model, and the critic calculates the advantage to update the policies.

take actions in adjacent locations to iteratively move the entire domain wall structure closer to the target wall.

To illustrate that the agent can learn state dependent actions, instead of each simulation beginning with a flat wall, for the simulation in Figure 4.6(b), 20 random actions are first applied to a flat domain wall structure. By randomizing the initial structure, this demonstrates that our policy is not simply memorizing to take actions in a particular region of the domain wall. In this case, the RL policy is not trained with a consistent starting structure and the policy must be more flexible to properly shift the domain wall. Additionally, in Figure 4.6(a) voltages were limited to just the positive range, but in Figure 4.6(b) negative voltages are also included in the range. Thus, policies learned as in Figure 4.6(b) are more indicative of the utility of our RL algorithm for its ability to take actions based on the current state of the wall. In practice, there are infinitely many initial wall configurations and thus each policy should choose actions differently based on the current structure of the wall. For instance, in Figure 4.6(b) for the section of wall on the far right, the initial wall happened to be displaced downwards during initialization. However, our policy seemingly recognized that significant rewards could be gained in this area, and as such it selected multiple biases to be applied in this local region to move it towards the ideal wall structure.

While the examples provided in Figure 4.6 do not appear to select a ‘perfect’ sequence of actions, both select actions that successfully increase the rewards in the environment. These examples show that effective policies can be learned for a different number of steps, ideal wall structures, and initial wall configurations.

4.5 Conclusions

Our work has shown that RL could potentially be used to control PFM to automate the alteration of ferroelectric domain walls. Using PFM image data, we successfully trained a dynamics model to predict the displacement of domain walls for different

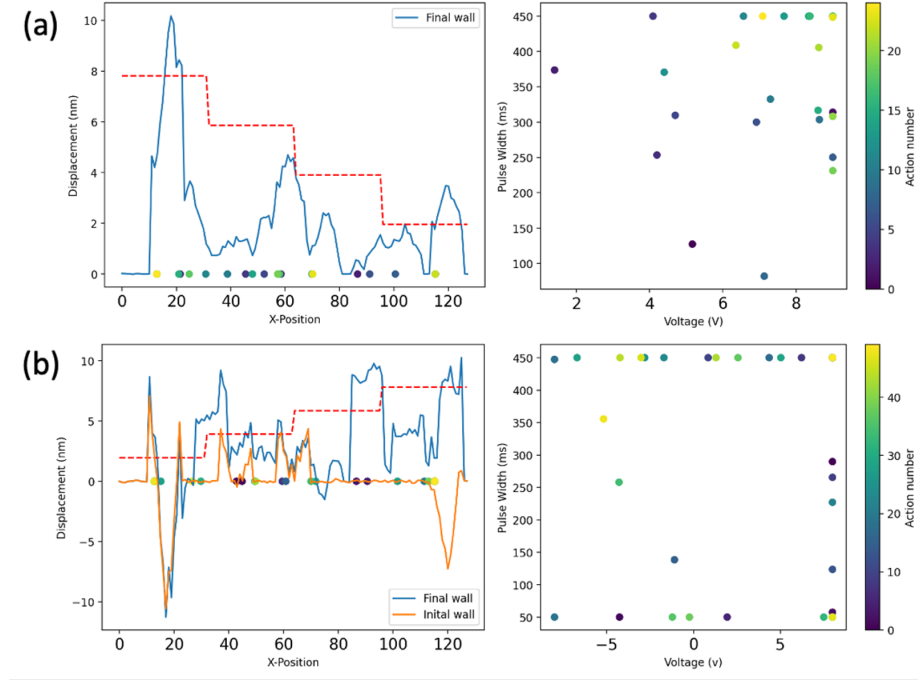


Figure 4.6: The final domain wall structures are shown after each RL policy is applied for the target wall structures represented by the red dashed line are shown on the left. In the plots on the left, the points along the X-axis indicate the locations selected by our policies and on the right, the magnitude of selected actions for both voltage and pulse width are shown. Actions are color coded by the order in which they were selected. In (a) the simulation started with a flat wall and in (b) the wall was displaced randomly prior to the simulation.

pulses. Then, with a dynamics model, we were able to train RL agents to learn policies to select action that move the domain wall towards a particular structure.

There still exist challenges to fully integrate RL automation for domain wall alteration using PFM. Currently, the dynamics model only predicts the 1-D structure of the domain wall. While we can encode information for the areas surrounding the domain wall from PFM images when making predictions, the model itself would have to predict the entire 2-D image for RL policies to be implemented directly. This type of prediction is not feasible due to the limited data currently available. However, it may be possible with many more image transitions, potentially with methods such as fast scanning via compressive sensing approaches [Kelley et al. \(2020\)](#). Interestingly, our data was captured on two different instruments (albeit of the same make and model) with different tips on different days, but this did not seem to introduce a level of distribution shift that severely impacted the learning process of the dynamics model. Should this be a problem, one strategy to counter it may be to learn a simple linear model that maps the applied potentials in previous data to applied potentials in the new data that minimizes the discrepancy between the predictions. This is under the assumption that the major tip change will be to affect the applied potential, but not the shape of the domains themselves.

Moreover, ideally one would use the phase-field simulations directly as the physics regularizer, using methods such as structured Gaussian processes, rather than incorporating additional loss terms that are ad-hoc defined and may not always be suitable depending on the circumstance. At the same time, the dynamics model learned provides significant insight into the dynamics of domain walls in a state-dependent way that is difficult to recover from more traditional PFM spectroscopic methods, and as such can be convenient for investigating ‘domain wall phase diagrams’. Finally, this approach could be used to optimize for specific properties of the material rather than the specific structure of the domain wall itself, i.e., to solve the inverse design problem of maximizing e.g., photoconductance of domain walls by trialing different wall configurations that maximize the photoconductance reward

given to the agent, as opposed to rewarding the generation of a specific structure per se.

Bibliography

- Akbarian, D., Yilmaz, D. E., Cao, Y., Ganesh, P., Dabo, I., Munro, J., Van Ginhoven, R., and Van Duin, A. C. (2019). Understanding the influence of defects and surface chemistry on ferroelectric switching: a reaxff investigation of batio 3. *Physical Chemistry Chemical Physics*, 21(33):18240–18249. [70](#)
- Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D’Ambra, J., and Shen, K. (2021). Algorithmic bias in data-driven innovation in the age of ai. [4](#), [13](#), [16](#), [17](#), [23](#)
- Alosh, M., Fritsch, K., Huque, M., Mahjoob, K., Pennello, G., Rothmann, M., Russek-Cohen, E., Smith, F., Wilson, S., and Yue, L. (2015). Statistical considerations on subgroup analysis in clinical trials. *Statistics in Biopharmaceutical Research*, 7(4):286–303. [11](#), [12](#), [23](#)
- Alquwaizani, M., Buckley, L., Adams, C., and Fanikos, J. (2013). Anticoagulants: a review of the pharmacology, dosing, and complications. *Current emergency and hospital medicine reports*, 1(2):83–97. [28](#)
- Amin-Nejad, A., Ive, J., and Velupillai, S. (2020). Exploring transformer text generation for medical dataset augmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4699–4708. [60](#)
- An, G., Moon, S., Kim, J.-H., and Song, H. O. (2021). Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in neural information processing systems*, 34:7436–7447. [33](#)

- Anbusathaiah, V., Jesse, S., Arredondo, M., Kartawidjaja, F., Ovchinnikov, O., Wang, J., Kalinin, S., and Nagarajan, V. (2010). Ferroelastic domain wall dynamics in ferroelectric bilayers. *Acta materialia*, 58(16):5316–5325. [71](#)
- Aravind, V. R., Morozovska, A. N., Bhattacharyya, S., Lee, D., Jesse, S., Grinberg, I., Li, Y., Choudhury, S., Wu, P., Seal, K., et al. (2010). Correlated polarization switching in the proximity of a 180 domain wall. *Physical Review B*, 82(2):024111. [70](#)
- Asan, O. and Montague, E. (2012). Physician interactions with electronic health records in primary care. *Health systems*, 1(2):96–103. [7](#)
- Atkeson, C. G. and Santamaria, J. C. (1997). A comparison of direct and model-based reinforcement learning. In *Proceedings of international conference on robotics and automation*, volume 4, pages 3557–3564. IEEE. [31](#)
- Babaeizadeh, M., Frosio, I., Tyree, S., Clemons, J., and Kautz, J. (2016). Reinforcement learning through asynchronous advantage actor-critic on a gpu. *arXiv preprint arXiv:1611.06256*. [39](#)
- Balke, N., Choudhury, S., Jesse, S., Huijben, M., Chu, Y. H., Baddorf, A. P., Chen, L.-Q., Ramesh, R., and Kalinin, S. V. (2009). Deterministic control of ferroelastic switching in multiferroic materials. *Nature nanotechnology*, 4(12):868–875. [69](#), [70](#)
- Basu, D., Gallus, A., Hirsh, J., and Cade, J. (1972). A prospective study of the value of monitoring heparin treatment with the activated partial thromboplastin time. *New England Journal of Medicine*, 287(7):324–327. [28](#)
- Baucum, M., Khojandi, A., and Vasudevan, R. (2020). Improving deep reinforcement learning with transitional variational autoencoders: A healthcare application. *IEEE Journal of Biomedical and Health Informatics*, 25(6):2273–2280. [4](#), [8](#), [30](#), [31](#), [32](#), [35](#), [36](#), [38](#), [39](#), [43](#), [57](#)

- Baucum, M., Khojandi, A., Vasudevan, R., and Davis, R. (2022). Adapting reinforcement learning treatment policies using limited data to personalize critical care. *INFORMS Journal on Data Science*, 1(1):27–49. [19](#), [23](#), [24](#), [32](#)
- Baucum, M., Khojandi, A., Vasudevan, R., and Ramdhani, R. (2021). Optimizing patient-specific medication regimen policies using wearable sensors in parkinson’s disease. *Management Science*. [4](#), [10](#), [30](#), [33](#), [49](#)
- Beaulieu-Jones, B. K., Moore, J. H., and CONSORTIUM, P. R. O.-A. A. C. T. (2017). Missing data imputation in the electronic health record using deeply learned autoencoders. In *Pacific symposium on biocomputing 2017*, pages 207–218. World Scientific. [7](#), [13](#), [22](#), [54](#)
- Bekker, H. (2015). Using decision making theory to inform clinical practice. *Shared Decision Making in Healthcare: Achieving Evidence-based Patient Choice*. [3](#)
- Bernheim, S. M., Ross, J. S., Krumholz, H. M., and Bradley, E. H. (2008). Influence of patients’ socioeconomic status on clinical management decisions: a qualitative study. *The Annals of Family Medicine*, 6(1):53–59. [12](#), [14](#), [23](#)
- Bhanot, K., Qi, M., Erickson, J. S., Guyon, I., and Bennett, K. P. (2021). The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165. [53](#), [55](#)
- Bond, M., Bowling, A., McKee, D., Kennelly, M., Banning, A. P., Dudley, N., Elder, A., and Martin, A. (2003). Does ageism affect the management of ischaemic heart disease? *Journal of health services research & policy*, 8(1):40–47. [12](#), [23](#)
- Brownlee, J. (2020). Cost-sensitive learning for imbalanced classification. *Machine Learning Mastery*. [17](#), [23](#)
- Burkhart, L. and Androwich, I. (2004). Measuring the domain completeness of the nursing interventions classification in parish nurse documentation. *CIN: Computers, Informatics, Nursing*, 22(2):72–82. [54](#)

- Campbell, N. R., Hull, R. D., Brant, R., Hogan, D. B., et al. (1998). Different effects of heparin in males and females. *Clinical and investigative medicine*, 21(2):71. [33](#)
- Carayon, P., Wetterneck, T. B., Rivera-Rodriguez, A. J., Hundt, A. S., Hoonakker, P., Holden, R., and Gurses, A. P. (2014). Human factors systems approach to healthcare quality and patient safety. *Applied ergonomics*, 45(1):14–25. [7](#)
- Catalan, G., Seidel, J., Ramesh, R., and Scott, J. F. (2012). Domain wall nanoelectronics. *Reviews of Modern Physics*, 84(1):119. [69](#)
- Chang, C., Deng, Y., Jiang, X., and Long, Q. (2020). Multiple imputation for analysis of incomplete data in distributed health data networks. *Nature communications*, 11(1):1–11. [13](#), [23](#)
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357. [15](#), [22](#)
- Chen, I., Johansson, F. D., and Sontag, D. (2018). Why is my classifier discriminatory? [15](#), [23](#)
- Chen, L.-Q. (2008). Phase-field method of phase transitions/domain structures in ferroelectric thin films: a review. *Journal of the American Ceramic Society*, 91(6):1835–1844. [70](#)
- Chen, M., Ma, J., Peng, R.-C., Zhang, Q., Wang, J., Liang, Y., Wu, J., Chen, L.-Q., Ma, J., and Nan, C.-W. (2019). Robust polarization switching in self-assembled bifeo₃ nanoislands with quad-domain structures. *Acta Materialia*, 175:324–330. [70](#)
- Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., and Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497. [53](#)

- Cicek, D. C., Duran, E., Saglam, B., Kaya, K., Mutlu, F., and Kozat, S. S. (2021). Awd3: Dynamic reduction of the estimation bias. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 775–779. IEEE. [20](#), [23](#)
- Cook, D. I., Gebski, V. J., and Keech, A. C. (2004). Subgroup analysis in clinical trials. *Medical Journal of Australia*, 180(6):289. [12](#), [13](#), [23](#)
- Damodaran, A. R., Clarkson, J., Hong, Z., Liu, H., Yadav, A. K., Nelson, C. T., Hsu, S.-L., McCarter, M., Park, K.-D., Kravtsov, V., et al. (2017). Phase coexistence and electric-field control of toroidal order in oxide superlattices. *Nature materials*, 16(10):1003–1009. [70](#)
- Dietterich, T. G. and Kong, E. B. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Technical report, Department of Computer Science, Oregon State University. [55](#)
- Eigenschink, P., Vamosi, S., Vamosi, R., Sun, C., Reutterer, T., and Kalcher, K. (2021). Deep generative models for synthetic data. *ACM Computing Surveys*. [53](#)
- el Hassouni, A., Hoogendoorn, M., van Otterlo, M., and Barbaro, E. (2018). Personalization of health interventions using cluster-based reinforcement learning. In *PRIMA*. [18](#), [19](#), [23](#)
- Escobar-Rodríguez, T., Monge-Lozano, P., and Romero-Alonso, M. M. (2012). Acceptance of e-prescriptions and automated medication-management systems in hospitals: An extension of the technology acceptance model. *Journal of Information Systems*, 26(1):77–96. [21](#)
- Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D., Knudtson, M. L., Investigators, A., et al. (2002). Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Journal of clinical epidemiology*, 55(2):184–191. [56](#)

- Feldman, K., Faust, L., Wu, X., Huang, C., and Chawla, N. V. (2017). Beyond volume: The impact of complex healthcare data on the machine learning pipeline. In *Towards integrative machine learning and knowledge extraction*, pages 150–169. Springer. [56](#)
- Fouladvand, S., Talbert, J., Dwoskin, L. P., Bush, H., Meadows, A. L., Peterson, L. E., Kavuluru, R., and Chen, J. (2021). Predicting opioid use disorder from longitudinal healthcare data using multi-stream transformer. *arXiv preprint arXiv:2103.08800*. [60](#)
- François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., and Pineau, J. (2018). An introduction to deep reinforcement learning. *CoRR*, abs/1811.12560. [5](#)
- Furukawa, M. F., Raghu, T., and Shao, B. B. (2010). Electronic medical records, nurse staffing, and nurse-sensitive patient outcomes: Evidence from california hospitals, 1998–2007. *Health services research*, 45(4):941–962. [11](#), [23](#)
- Ganpule, C., Nagarajan, V., Hill, B., Roytburd, A., Williams, E., Ramesh, R., Alpay, S., Roelofs, A., Waser, R., and Eng, L. (2002). Imaging three-dimensional polarization in epitaxial polydomain ferroelectric thin films. *Journal of applied physics*, 91(3):1477–1481. [69](#)
- Ganpule, C., Roytburd, A., Nagarajan, V., Hill, B., Ogale, S., Williams, E., Ramesh, R., and Scott, J. (2001). Polarization relaxation kinetics and 180 domain wall dynamics in ferroelectric thin films. *Physical Review B*, 65(1):014101. [69](#)
- Gao, N., Xue, H., Shao, W., Zhao, S., Qin, K. K., Prabowo, A., Rahaman, M. S., and Salim, F. D. (2020). Generative adversarial networks for spatio-temporal data: A survey. *CoRR*, abs/2008.08903. [61](#)
- Ge, Y., Liu, S., Gao, R., Xian, Y., Li, Y., Zhao, X., Pei, C., Sun, F., Ge, J., Ou, W., et al. (2021). Towards long-term fairness in recommendation. In *Proceedings*

- of the 14th ACM International Conference on Web Search and Data Mining, pages 445–453. [17](#), [23](#)
- Ge, Y., Zhao, X., Yu, L., Paul, S., Hu, D., Hsieh, C.-C., and Zhang, Y. (2022). Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 316–324. [18](#), [23](#)
- Geng, Y. and Luo, X. (2018). Cost-sensitive convolution based neural networks for imbalanced time-series classification. [17](#), [23](#), [24](#)
- Ghadirzadeh, A., Poklukar, P., Kyrki, V., Kragic, D., and Björkman, M. (2020). Data-efficient visuomotor policy training using reinforcement learning and generative models. *arXiv preprint arXiv:2007.13134*. [65](#)
- Ghassemi, M. M., Alhanai, T., Westover, M. B., Mark, R. G., and Nemati, S. (2018). Personalized medication dosing using volatile data streams. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*. [30](#), [35](#), [43](#)
- Ghassemi, M. M., Richter, S. E., Eche, I. M., Chen, T. W., Danziger, J., and Celi, L. A. (2014). A data-driven approach to optimized medication dosing: a focus on heparin. *Intensive care medicine*, 40(9):1332–1339. [8](#), [30](#), [35](#), [43](#)
- Gibson, C. M. and Yuet, W. C. (2021). Racial and ethnic differences in response to anticoagulation: A review of the literature. *Journal of Pharmacy Practice*, 34(5):685–693. [33](#)
- Gold, R., Cottrell, E., Bunce, A., Middendorf, M., Hollombe, C., Cowburn, S., Mahr, P., and Melgar, G. (2017). Developing electronic health record (ehr) strategies related to health center patients’ social determinants of health. *The Journal of the American Board of Family Medicine*, 30(4):428–447. [6](#)
- Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine*, 38(3):50–57. [10](#)

- Gottesman, O., Johansson, F. D., Meier, J., Dent, J., Lee, D., Srinivasan, S., Zhang, L., Ding, Y., Wihl, D., Peng, X., Yao, J., Lage, I., Mosch, C., Lehman, L. H., Komorowski, M., Faisal, A., Celi, L. A., Sontag, D. A., and Doshi-Velez, F. (2018). Evaluating reinforcement learning algorithms in observational health settings. *CoRR*, abs/1805.12298. [28](#), [30](#), [33](#)
- Grzes, M. (2017). Reward shaping in episodic reinforcement learning. [20](#), [23](#)
- Hajjaj, F., Salek, M., Basra, M., and Finlay, A. (2010). Non-clinical influences on clinical decision-making: a major challenge to evidence-based practice. *Journal of the Royal Society of Medicine*, 103(5):178–87. [12](#), [23](#)
- Harnagea, C. (2001). *Local piezoelectric response and domain structures in ferroelectric thin films investigated by voltage-modulated force microscopy*. Universitäts-und Landesbibliothek. [69](#)
- He, Q. and Hou, X. (2020). Wd3: Taming the estimation bias in deep reinforcement learning. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 391–398. IEEE. [20](#)
- Hébert-Johnson, Ú., Kim, M. P., Reingold, O., and Rothblum, G. N. (2017). Calibration for the (computationally-identifiable) masses. *CoRR*, abs/1711.08513. [20](#), [23](#)
- Heuillet, A., Couthouis, F., and Rodríguez, N. D. (2020). Explainability in deep reinforcement learning. *CoRR*, abs/2008.06693. [10](#), [21](#), [23](#)
- Hilsenbeck, S. G., Kurucz, C., and Duncan, R. C. (1992). Estimation of completeness and adjustment of age-specific and age-standardized incidence rates. *Biometrics*, pages 1249–1262. [55](#)
- Hirsh, J. and Raschke, R. (2004). Heparin and low-molecular-weight heparin: the seventh accp conference on antithrombotic and thrombolytic therapy. *Chest*, 126(3):188S–203S. [28](#)

- Ho, C. W., Ali, J., and Caals, K. (2020). Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance. *Bulletin of the World Health Organization*, 98(4):263. [21](#)
- Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". [16](#), [17](#), [21](#), [23](#)
- Hu, J., Akande, O., and Wang, Q. (2021). Multiple imputation and synthetic data generation with npbayesimputecat. *R Journal*, 13(2). [56](#)
- Hu, Y., Wang, W., Jia, H., Wang, Y., Chen, Y., Hao, J., Wu, F., and Fan, C. (2020). Learning to utilize shaping rewards: A new approach of reward shaping. *CoRR*, abs/2011.02669. [20](#), [23](#)
- Hu, Z., Melton, G. B., Arsoniadis, E. G., Wang, Y., Kwaan, M. R., and Simon, G. J. (2017). Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of biomedical informatics*, 68:112–120. [14](#), [23](#)
- Huzaifah, M. and Wyse, L. (2021). Deep generative models for musical audio synthesis. In *Handbook of Artificial Intelligence for Music*, pages 639–678. Springer. [53](#)
- Ignjatovic, V., Furmedge, J., Newall, F., Chan, A., Berry, L., Fong, C., Cheng, K., and Monagle, P. (2006). Age-related differences in heparin response. *Thrombosis research*, 118(6):741–745. [33](#)
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134. [53](#), [55](#)
- Ivry, Y., Scott, J. F., Salje, E. K., and Durkan, C. (2012). Nucleation, growth, and control of ferroelectric-ferroelastic domains in thin polycrystalline films. *Physical Review B*, 86(20):205428. [71](#)

- Jakobsen, J. C., Gluud, C., Wetterslev, J., and Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. *BMC medical research methodology*, 17(1):1–10. [13](#), [14](#), [23](#)
- Jesse, S., Rodriguez, B. J., Choudhury, S., Baddorf, A. P., Vrejoiu, I., Hesse, D., Alexe, M., Eliseev, E. A., Morozovska, A. N., Zhang, J., et al. (2008). Direct imaging of the spatial and energy distribution of nucleation centres in ferroelectric materials. *Nature materials*, 7(3):209–215. [70](#)
- Jin, L., Abrahams, J. P., Skinner, R., Petitou, M., Pike, R. N., and Carrell, R. W. (1997). The anticoagulant activation of antithrombin by heparin. *Proceedings of the National Academy of Sciences*, 94(26):14683–14688. [28](#)
- Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9. [7](#), [35](#), [57](#)
- Jonker, R. and Volgenant, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340. [61](#)
- Jorde, L. B. and Wooding, S. P. (2004). Genetic variation, classification and ‘race’. *Nature genetics*, 36(11):S28–S33. [12](#)
- Kaiser, L., Babaeizadeh, M., Milos, P., Osinski, B., Campbell, R. H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., et al. (2019). Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*. [31](#)
- Kalinin, S., Rodriguez, B., Jesse, S., Chu, Y., Zhao, T., Ramesh, R., Choudhury, S., Chen, L., Eliseev, E., and Morozovska, A. (2007). Intrinsic single-domain switching in ferroelectric materials on a nearly ideal surface. *Proceedings of the National Academy of Sciences*, 104(51):20204–20209. [70](#)

- Kalinin, S. V., Jesse, S., Rodriguez, B. J., Chu, Y., Ramesh, R., Eliseev, E., and Morozovska, A. (2008). Probing the role of single defects on the thermodynamics of electric-field induced phase transitions. *Physical review letters*, 100(15):155703. [71](#)
- Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. (2020). Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR. [55](#)
- Kaushik, P., Kummetha, S., Moodley, P., and Bapi, R. S. (2022). A conservative q-learning approach for handling distribution shift in sepsis treatment strategies. *arXiv preprint arXiv:2203.13884*. [41](#)
- Kawase, K., Okazaki, S., Toyoda, K., Toratani, N., Yoshimura, S., Kawano, H., Nagatsuka, K., Matsuo, H., Naritomi, H., and Minematsu, K. (2009). Sex difference in the prevalence of deep-vein thrombosis in japanese patients with acute intracerebral hemorrhage. *Cerebrovascular diseases*, 27(4):313–319. [29](#)
- Keeling, D., Davidson, S., Watson, H., Haemostasis, and of the British Committee for Standards in Haematology, T. T. F. (2006). The management of heparin-induced thrombocytopenia. *British journal of haematology*, 133(3):259–269. [28](#)
- Kelley, K. P., Ziatdinov, M., Collins, L., Susner, M. A., Vasudevan, R. K., Balke, N., Kalinin, S. V., and Jesse, S. (2020). Fast scanning probe microscopy via machine learning: non-rectangular scans with compressed sensing and gaussian process optimization. *Small*, 16(37):2002878. [85](#)
- Kelly, B., Matthews, T. P., and Anastasio, M. A. (2017). Deep learning-guided image reconstruction from incomplete data. *arXiv preprint arXiv:1709.00584*. [56](#)
- Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. [10](#), [15](#), [17](#), [22](#), [23](#)

- Khojandi, A., Maillart, L. M., Prokopyev, O. A., Roberts, M. S., Brown, T., and Barrington, W. W. (2014). Optimal implantable cardioverter defibrillator (icd) generator replacement. *INFORMS Journal on Computing*, 26(3):599–615. [4](#)
- Khojandi, A., Maillart, L. M., Prokopyev, O. A., Roberts, M. S., and Saba, S. F. (2018). Dynamic abandon/extract decisions for failed cardiac leads. *Management Science*, 64(2):633–651. [4](#)
- Kim, J., You, M., Kim, K.-E., Chu, K., and Yang, C.-H. (2019). Artificial creation and separation of a single vortex–antivortex pair in a ferroelectric flatland. *npj Quantum Materials*, 4(1):29. [70](#)
- Kodra, Y., Posada de la Paz, M., Coi, A., Santoro, M., Bianchi, F., Ahmed, F., Rubinstein, Y. R., Weinbach, J., and Taruscio, D. (2017). Data quality in rare diseases registries. *Rare diseases epidemiology: update and overview*, pages 149–164. [55](#)
- Koivu, A., Sairanen, M., Airola, A., and Pahikkala, T. (2020). Synthetic minority oversampling of vital statistics data with generative adversarial networks. *Journal of the American Medical Informatics Association*, 27(11):1667–1674. [24](#)
- Kreif, N., Grieve, R., Radice, R., Sadique, Z., Ramsahai, R., and Sekhon, J. S. (2012). Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Medical decision making*, 32(6):750–763. [12](#), [23](#)
- Kuhle, S., Eulmesekian, P., Kavanagh, B., Massicotte, P., Vegh, P., Lau, A., and Mitchell, L. G. (2007). Lack of correlation between heparin dose and standard clinical monitoring tests in treatment with unfractionated heparin in critically ill children. *haematologica*, 92(4):554–557. [28](#)
- Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97. [61](#)

- Kukar, M., Kononenko, I., et al. (1998). Cost-sensitive learning with neural networks. In *ECAI*, volume 15, pages 88–94. Citeseer. [17](#), [23](#)
- Kumar, A., Fu, J., Soh, M., Tucker, G., and Levine, S. (2019). Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems*, 32. [20](#)
- Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191. [20](#), [23](#), [33](#), [41](#)
- Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. (2017). Grammar variational autoencoder. In *International conference on machine learning*, pages 1945–1954. PMLR. [53](#), [55](#)
- Kärkkäinen, K. and Joo, J. (2019). Fairface: Face attribute dataset for balanced race, gender, and age. [4](#), [15](#), [23](#)
- Landi, I., Glicksberg, B. S., Lee, H.-C., Cherng, S., Landi, G., Danieleto, M., Dudley, J. T., Furlanello, C., and Miotto, R. (2020). Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ digital medicine*, 3(1):1–11. [4](#)
- Lee, D., Kim, J., Moon, W.-J., and Ye, J. C. (2019). Collagan: Collaborative gan for missing image data imputation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2487–2496. [60](#)
- Lee, S., Kim, J., Park, S. W., Jin, S.-M., and Park, S.-M. (2021). Toward a fully automated artificial pancreas system using a bioinspired reinforcement learning design: In silico validation. *IEEE Journal of Biomedical and Health Informatics*, 25(2):536–546. [30](#)

- Lesko, C. R., Henderson, N. C., and Varadhan, R. (2018). Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. *Journal of clinical epidemiology*, 100:22–31. [11](#), [23](#)
- Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643. [3](#), [6](#)
- Li, P., yan Cheng, Z., and lin Liu, G. (2020). Availability bias causes misdiagnoses by physicians: Direct evidence from a randomized controlled trial. *Internal Medicine*, 59(24):3141–3146. [12](#), [23](#)
- Li, Q., Zhou, W., Lu, Z., and Li, H. (2022a). Simultaneous double q-learning with conservative advantage learning for actor-critic methods. *arXiv preprint arXiv:2205.03819*. [20](#), [23](#), [24](#)
- Li, S., Bhalla, A., Newnham, R., and Cross, L. (1993). Quantitative evaluation of extrinsic contribution to piezoelectric coefficient d33 in ferroelectric pzt ceramics. *Materials Letters*, 17(1-2):21–26. [69](#)
- Li, S. C.-X., Jiang, B., and Marlin, B. (2019). Misgan: Learning from incomplete data with generative adversarial networks. *arXiv preprint arXiv:1902.09599*. [56](#)
- Li, T.-H., Wang, Z.-S., Lu, W., Zhang, Q., and Li, D.-F. (2021). Electronic health records based reinforcement learning for treatment optimizing. *Information Systems*, page 101878. [6](#), [10](#), [22](#)
- Li, Y. (2017). Deep reinforcement learning: An overview. *CoRR*, abs/1701.07274. [3](#), [5](#), [22](#)
- Li, Y., Jin, Y., Lu, X., Yang, J.-C., Chu, Y.-H., Huang, F., Zhu, J., and Cheong, S.-W. (2017). Rewritable ferroelectric vortex pairs in bifeo3. *npj Quantum Materials*, 2(1):43. [70](#)

- Li, Z., Shen, H., Dawson, G., Zhang, Z., Wang, Y., Nan, F., Song, G., Li, G., Wu, Y., and Liu, H. (2022b). Uniform arrays of centre-type topological domains in epitaxial ferroelectric thin films. *Journal of Materials Chemistry C*, 10(8):3071–3080. [70](#)
- Lin, E., Chen, Q., and Qi, X. (2019). Deep reinforcement learning for imbalanced classification. *CoRR*, abs/1901.01379. [4](#), [17](#), [23](#), [24](#)
- Lin, R., Stanley, M. D., Ghassemi, M. M., and Nemati, S. (2018). A deep deterministic policy gradient approach to medication dosing and surveillance in the icu. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4927–4931. IEEE. [30](#), [31](#), [32](#), [33](#), [35](#), [36](#), [43](#)
- Little, R. J., D’Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360. [56](#)
- Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *In Proceedings of the Eleventh International Conference on Machine Learning*, pages 157–163. Morgan Kaufmann. [5](#)
- Liu, N., Liu, Y., Logan, B., Xu, Z., Tang, J., and Wang, Y. (2018). Deep reinforcement learning for dynamic treatment regimes on medical registry data. *CoRR*, abs/1801.09271. [8](#), [30](#)
- Liu, S., Ngiam, K. Y., and Feng, M. (2019). Deep reinforcement learning for clinical decision support: A brief survey. *CoRR*, abs/1907.09475. [29](#)
- Liu, S., See, K. C., Ngiam, K. Y., Celi, L. A., Sun, X., Feng, M., for Integrative Science, N. G. S., Engineering, and Feng (2020a). Reinforcement learning for clinical decision support in critical care comprehensive review. [7](#), [8](#), [12](#), [23](#)

- Liu, Y., Ramachandran, P., Liu, Q., and Peng, J. (2017). Stein variational policy gradient. *arXiv preprint arXiv:1704.02399*. [81](#)
- Liu, Z., Khojandi, A., Li, X., Davis, R. L., and Kamaleswaran, R. (2020b). A machine learning-enabled partially observable markov decision process framework for early sepsis prediction. [4](#)
- Lu, M., Shahn, Z., Sow, D., Doshi-Velez, F., and Lehman, L.-w. H. (2020). Is deep reinforcement learning ready for practical applications in healthcare? a sensitivity analysis of duel-ddqn for hemodynamic management in sepsis patients. In *AMIA Annual Symposium Proceedings*, volume 2020, page 773. American Medical Informatics Association. [21](#), [23](#)
- Markey, M. K., Tourassi, G. D., Margolis, M., and DeLong, D. M. (2006). Impact of missing data in evaluating artificial neural networks trained on complete data. *Computers in Biology and Medicine*, 36(5):516–525. [56](#)
- Masino, A. J., Harris, M. C., Forsyth, D., Ostapenko, S., Srinivasan, L., Bonafide, C. P., Balamuth, F., Schmatz, M., and Grundmeier, R. W. (2019). Machine learning models for early sepsis recognition in the neonatal intensive care unit using readily available electronic health record data. *PloS one*, 14(2):e0212665. [16](#)
- Mattei, P.-A. and Frellsen, J. (2019). Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR. [56](#)
- McGilly, L., Yudin, P., Feigl, L., Tagantsev, A., and Setter, N. (2015). Controlling domain wall motion in ferroelectric thin films. *Nature nanotechnology*, 10(2):145–150. [70](#)
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35. [32](#)

- Meier, D. and Selbach, S. M. (2022). Ferroelectric domain walls for nanotechnology. *Nature Reviews Materials*, 7(3):157–173. [69](#)
- Mermin, N. D. (1979). The topological theory of defects in ordered media. *Reviews of Modern Physics*, 51(3):591. [69](#)
- Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR. [79](#)
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. A. (2013). Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602. [3](#), [40](#)
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533. [40](#)
- Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481. [12](#), [16](#), [23](#)
- Nataf, G., Guennou, M., Gregg, J., Meier, D., Hlinka, J., Salje, E., and Kreisel, J. (2020). Domain-wall engineering and topological defects in ferroelectric and ferroelastic materials. *Nature Reviews Physics*, 2(11):634–648. [69](#)
- Nemati, S., Ghassemi, M. M., and Clifford, G. D. (2016). Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 2978–2981. IEEE. [8](#), [28](#), [30](#), [31](#), [32](#), [35](#), [36](#), [43](#)
- Nguyen, C. D., Carlin, J. B., and Lee, K. J. (2021). Practical strategies for handling breakdown of multiple imputation procedures. *Emerging themes in epidemiology*, 18(1):1–8. [14](#), [23](#)

- Nguyen, T., Silander, T., and Leong, T. (2012). Transferring expectations in model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 25. [33](#)
- Noaro, G., Zhu, T., Cappon, G., Facchinetti, A., and Georgiou, P. (2023). A personalized and adaptive insulin bolus calculator based on double deep q-learning to improve type 1 diabetes management. *IEEE Journal of Biomedical and Health Informatics*, pages 1–10. [30](#), [41](#)
- Okudo, T. and Yamada, S. (2021). Subgoal-based reward shaping to improve efficiency in reinforcement learning. *CoRR*, abs/2104.06411. [20](#), [23](#)
- Otterlo, M. and Wiering, M. (2012). Reinforcement learning and markov decision processes. *Reinforcement Learning: State of the Art*, pages 3–42. [5](#)
- Panch, T., Mattie, H., and Atun, R. (2019). Artificial intelligence and algorithmic bias: implications for health systems. [11](#), [23](#)
- Parbhoo, S., Bogojeska, J., Zazzi, M., Roth, V., and Doshi-Velez, F. (2017). Combining kernel and model based learning for hiv therapy selection. [4](#)
- Parbhoo, S., Wieser, M., Roth, V., and Doshi-Velez, F. (2020). Transfer learning from well-curated to less-resourced populations with hiv. In *Machine Learning for Healthcare Conference*, pages 589–609. PMLR. [19](#), [23](#)
- Parikh, R. B., Teeple, S., and Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. *Jama*, 322(24):2377–2378. [17](#), [23](#)
- Peng, B., Li, X., Gao, J., Liu, J., Chen, Y.-N., and Wong, K.-F. (2018). Adversarial advantage actor-critic model for task-completion dialogue policy learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6149–6153. IEEE. [79](#)

- Pessach, D. and Shmueli, E. (2022). A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3). [4](#), [13](#)
- Piacentino, E., Guarner, A., and Angulo, C. (2021). Generating synthetic ecgs using gans for anonymizing healthcare data. *Electronics*, 10(4):389. [53](#)
- Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. (2018). The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13. [6](#), [7](#), [16](#)
- Prudencio, R. F., Maximo, M. R. O. A., and Colombini, E. L. (2022). A survey on offline reinforcement learning: Taxonomy, review, and open problems. [31](#)
- Puiutta, E. and Veith, E. M. S. P. (2020). Explainable reinforcement learning: A survey. *CoRR*, abs/2005.06247. [10](#), [21](#), [23](#), [24](#)
- Qiu, X., Tan, X., Li, Q., Chen, S., Ru, Y., and Jin, Y. (2022). A latent batch-constrained deep reinforcement learning approach for precision dosing clinical decision support. *Knowledge-Based Systems*, 237:107689. [31](#), [33](#), [35](#), [36](#)
- Raab, G. M., Nowok, B., and Dibben, C. (2017). Guidelines for producing useful synthetic data. *arXiv preprint arXiv:1712.04078*. [53](#), [55](#)
- Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., and Ghassemi, M. (2017). Continuous state-space models for optimal sepsis treatment - a deep reinforcement learning approach. *CoRR*, abs/1705.08422. [8](#)
- Raghu, A., Komorowski, M., and Singh, S. S. (2018). Model-based reinforcement learning for sepsis treatment. *CoRR*, abs/1811.09602. [30](#)
- Rajapakshe, T., Rana, R., Khalifa, S., Schuller, B. W., and Liu, J. (2021). A novel policy for pre-trained deep reinforcement learning for speech emotion recognition. *CoRR*, abs/2101.00738. [23](#)

- Rajkomar, A., Howell, M., and Hardt, M. (2018). Ensuring fairness in machine learning to advance health equity. [18](#), [23](#), [24](#)
- Rea, S., Pathak, J., Savova, G., Oniki, T. A., Westberg, L., Beebe, C. E., Tao, C., Parker, C. G., Haug, P. J., Huff, S. M., et al. (2012). Building a robust, scalable and standards-driven infrastructure for secondary use of ehr data: the sharpn project. *Journal of biomedical informatics*, 45(4):763–771. [7](#)
- Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30(2):235–242. [55](#)
- Riachi, E., Mamdani, M., Fralick, M., and Rudzicz, F. (2021). Challenges for reinforcement learning in healthcare. [4](#), [8](#)
- Riley, W. J. (2012). Health disparities: gaps in access, quality and affordability of medical care. *Transactions of the American Clinical and Climatological Association*, 123:167. [11](#), [23](#)
- Rodriguez, B. J., Choudhury, S., Chu, Y., Bhattacharyya, A., Jesse, S., Seal, K., Baddorf, A. P., Ramesh, R., Chen, L.-Q., and Kalinin, S. V. (2009). Unraveling deterministic mesoscopic polarization switching mechanisms: spatially resolved studies of a tilt grain boundary in bismuth ferrite. *Advanced Functional Materials*, 19(13):2053–2063. [70](#)
- Rodriguez, P. J., Ward, Z. J., Long, M. W., Austin, S. B., and Wright, D. R. (2021). Applied methods for estimating transition probabilities from electronic health record data. *Medical Decision Making*, 41(2):143–152. [7](#)
- Rosenbloom, S., Stead, W., Denny, J., Giuse, D., Lorenzi, N., Brown, S., and Johnson, K. (2010). Generating clinical notes for electronic health record systems. *Applied clinical informatics*, 1:232–243. [7](#)
- Roth, A. C. A. (2020). A snapshot of the frontiers of fairness in machine learning. [17](#), [23](#)

- Rubio-Marcos, F., Pamies, P., Del Campo, A., Tiana, J., Ordoñez-Pimentel, J., Venet, M., Rojas-Hernandez, R. E., Ochoa, D. A., Fernández, J. F., and García, J. E. (2023). Light-induced strain and its correlation with the optical absorption at charged domain walls in polycrystalline ferroelectrics. *Applied Materials Today*, 32:101838. [69](#)
- Salmi, S., Mérelle, S., Gilissen, R., van der Mei, R., and Bhulai, S. (2022). Detecting changes in help seeker conversations on a suicide prevention helpline during the covid- 19 pandemic: in-depth analysis using encoder representations from transformers. *BMC public health*, 22(1):1–10. [60](#)
- Sánchez-Hernández, F., Ballesteros-Herráez, J. C., Kraiem, M. S., Sánchez-Barba, M., and Moreno-García, M. N. (2019). Predictive modeling of icu healthcare-associated infections from imbalanced data. using ensembles and a clustering-based undersampling approach. *Applied Sciences*, 9(24):5287. [16](#), [23](#), [24](#)
- Schaefer, A. and Bailey, M. (2005). Modeling medical treatment using markov decision processes. [6](#), [22](#)
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *CoRR*, abs/1707.06347. [39](#), [40](#)
- Schultheiß, J., Rojac, T., and Meier, D. (2022). Unveiling alternating current electronic properties at ferroelectric domain walls. *Advanced Electronic Materials*, 8(6):2100996. [69](#)
- Seidel, J., Martin, L. W., He, Q., Zhan, Q., Chu, Y.-H., Rother, A., Hawkrige, M., Maksymovych, P., Yu, P., Gajek, M., et al. (2009). Conduction at domain walls in oxide multiferroics. *Nature materials*, 8(3):229–234. [69](#)
- Sharabiani, A., Bress, A., Douzali, E., and Darabi, H. (2015). Revisiting warfarin dosing using machine learning techniques. *Computational and mathematical methods in medicine*, 2015. [4](#), [23](#)

- Shiranthika, C., Chen, K.-W., Wang, C.-Y., Yang, C.-Y., Sudantha, B. H., and Li, W.-F. (2022). Supervised optimal chemotherapy regimen based on offline reinforcement learning. *IEEE Journal of Biomedical and Health Informatics*, 26(9):4763–4772. [41](#)
- Sholzberg, M., Tang, G. H., Rahhal, H., AlHamzah, M., Kreuziger, L. B., Áinle, F. N., Alomran, F., Alayed, K., Alsheef, M., AlSumait, F., et al. (2021). Effectiveness of therapeutic heparin versus prophylactic heparin on death, mechanical ventilation, or intensive care unit admission in moderately ill patients with covid-19 admitted to hospital: Rapid randomised clinical trial. *bmj*, 375. [28](#)
- Shome, D., Kar, T., Mohanty, S. N., Tiwari, P., Muhammad, K., AlTameem, A., Zhang, Y., and Saudagar, A. K. J. (2021). Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. *International Journal of Environmental Research and Public Health*, 18(21):11086. [60](#)
- Shvartsman, V., Kholkin, A., Verdier, C., Yong, Z., and Lupascu, D. (2005). Investigation of fatigue mechanism in ferroelectric ceramic via piezoresponse force microscopy. *Journal of the European Ceramic Society*, 25(12):2559–2561. [69](#)
- Smith, B., Van Steelandt, S., and Khojandi, A. (2023). Evaluating the impact of health care data completeness for deep generative models. *Methods of Information in Medicine*. [iii](#)
- Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338. [14](#), [23](#)
- Suo, Q., Zhong, W., Ma, F., Yuan, Y., Gao, J., and Zhang, A. (2019). Metric learning on healthcare data with incomplete modalities. In *IJCAI*, pages 3534–3540. [60](#)
- Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, second edition. [5](#), [29](#), [33](#)

- Swazinna, P., Udluft, S., and Runkler, T. (2021). Overcoming model bias for robust offline deep reinforcement learning. *Engineering Applications of Artificial Intelligence*, 104:104366. [10](#), [20](#), [23](#)
- Syn, N. L., Wong, A. L.-A., Lee, S.-C., Teoh, H.-L., Yip, J. W. L., Seet, R. C., Yeo, W. T., Kristanto, W., Bee, P.-C., Poon, L., et al. (2018). Genotype-guided versus traditional clinical dosing of warfarin in patients of asian ancestry: a randomized controlled trial. *BMC medicine*, 16(1):1–10. [4](#), [12](#)
- Tang, S., Modi, A., Sjoding, M. W., and Wiens, J. (2020). Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies. *CoRR*, abs/2007.12678. [8](#), [11](#), [21](#), [23](#)
- Tanniou, J., Van Der Tweel, I., Teerenstra, S., and Roes, K. C. (2016). Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes. *BMC medical research methodology*, 16(1):1–15. [11](#), [13](#), [23](#)
- Torfi, A. and Fox, E. A. (2020). Cor-gan: Correlation-capturing convolutional neural networks for generating synthetic healthcare records. [58](#)
- Toulon, P., Smahi, M., and De Pooter, N. (2021). Aptt therapeutic range for monitoring unfractionated heparin therapy. significant impact of the anti-xa reagent used for correlation. *Journal of Thrombosis and Haemostasis*, 19(8):2002–2006. [28](#)
- Tselev, A., Yu, P., Cao, Y., Dedon, L. R., Martin, L. W., Kalinin, S. V., and Maksymovych, P. (2016). Microwave ac conductivity of domain walls in ferroelectric thin films. *Nature communications*, 7(1):11630. [69](#)
- Van Hasselt, H., Guez, A., and Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30. [41](#)

- Vasudevan, R. K., Chen, Y.-C., Tai, H.-H., Balke, N., Wu, P., Bhattacharya, S., Chen, L.-Q., Chu, Y.-H., Lin, I.-N., Kalinin, S. V., et al. (2011). Exploring topological defects in epitaxial bifeo₃ thin films. *Acs Nano*, 5(2):879–887. [70](#)
- Vasudevan, R. K., Wu, W., Guest, J. R., Baddorf, A. P., Morozovska, A. N., Eliseev, E. A., Balke, N., Nagarajan, V., Maksymovych, P., and Kalinin, S. V. (2013). Domain wall conduction and polarization-mediated transport in ferroelectrics. *Advanced Functional Materials*, 23(20):2592–2616. [69](#)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. [55](#)
- Vincent, R. (2014). Reinforcement learning in models of adaptive medical treatment strategies. [13](#), [23](#)
- Vul, B., Guro, G., and Ivanchik, I. (1973). Encountering domains in ferroelectrics. *Ferroelectrics*, 6(1):29–31. [69](#)
- Vyas, D. A., Eisenstein, L. G., and Jones, D. S. (2020). Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. [8](#), [12](#)
- Walker, A. M. and Jick, H. (1980). Predictors of bleeding during heparin therapy. *Jama*, 244(11):1209–1212. [29](#)
- Wang, X., Lyu, Y., and Jing, L. (2020). Deep generative model for robust imbalance classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14124–14133. [55](#)
- Wang, Z., Myles, P., and Tucker, A. (2021). Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Computational Intelligence*, 37(2):819–851. [53](#)

- Watts, J., Khojandi, A., Vasudevan, R., and Ramdhani, R. (2020). Optimizing individualized treatment planning for parkinson’s disease using deep reinforcement learning. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5406–5409. [10](#), [30](#)
- Weiskopf, N. G., Hripcsak, G., Swaminathan, S., and Weng, C. (2013). Defining and measuring completeness of electronic health records for secondary use. *Journal of biomedical informatics*, 46(5):830–836. [53](#), [55](#)
- Wen, M., Bastani, O., and Topcu, U. (2021). Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pages 1144–1152. PMLR. [18](#), [23](#)
- Weng, W., Gao, M., He, Z., Yan, S., and Szolovits, P. (2017). Representation and reinforcement learning for personalized glycemic control in septic patients. *CoRR*, abs/1712.00654. [8](#)
- Whittle, J., Conigliaro, J., Good, C., and Lofgren, R. P. (1993). Racial differences in the use of invasive cardiovascular procedures in the department of veterans affairs medical system. *New England Journal of Medicine*, 329(9):621–627. [11](#), [23](#)
- Whyte, J. R., McQuaid, R. G., Sharma, P., Canalias, C., Scott, J. F., Gruverman, A., and Gregg, J. M. (2014). Ferroelectric domain wall injection. *Advanced Materials*, 26(2):293–298. [69](#), [70](#)
- Wilson, S. J.-A., Wilbur, K., Burton, E., and Anderson, D. R. (2001). Effect of patient weight on the anticoagulant response to adjusted therapeutic dosage of low-molecular-weight heparin for the treatment of venous thromboembolism. *Pathophysiology of Haemostasis and Thrombosis*, 31(1):42–48. [36](#)
- Wong, E. S., Yoon, J., Piegari, R. I., Rosland, A.-M. M., Fihn, S. D., and Chang, E. T. (2018). Identifying latent subgroups of high-risk patients using risk score trajectories. *Journal of general internal medicine*, 33(12):2120–2126. [16](#), [23](#)

- Wright, A., McCoy, A. B., Hickman, T.-T. T., Hilaire, D. S., Borbolla, D., Bowes III, W. A., Dixon, W. G., Dorr, D. A., Krall, M., Malholtra, S., et al. (2015). Problem list completeness in electronic health records: a multi-site study and assessment of success factors. *International journal of medical informatics*, 84(10):784–790. [54](#), [55](#)
- Wu, J., Roy, J., and Stewart, W. F. (2010). Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, pages S106–S113. [6](#)
- Xu, F., Trolier-McKinstry, S., Ren, W., Xu, B., Xie, Z.-L., and Hemker, K. (2001). Domain wall motion and its contribution to the dielectric and piezoelectric properties of lead zirconate titanate films. *Journal of Applied Physics*, 89(2):1336–1348. [69](#)
- Xu, L., Zeng, X., Li, W., and Bai, L. (2021). Idhashgan: Deep hashing with generative adversarial nets for incomplete data retrieval. *IEEE Transactions on Multimedia*, 24:534–545. [56](#)
- Xu, R., Karthik, J., Damodaran, A. R., and Martin, L. W. (2014). Stationary domain wall contribution to enhanced ferroelectric susceptibility. *Nature communications*, 5(1):3120. [69](#)
- Yang, J., Petersen, B. K., Zha, H., and Faissol, D. M. (2019). Single episode policy transfer in reinforcement learning. *CoRR*, abs/1910.07719. [19](#), [23](#), [24](#)
- Yang, J., Soltan, A. A., and Clifton, D. A. (2022a). Algorithmic fairness and bias mitigation for clinical machine learning: A new utility for deep reinforcement learning. *medRxiv*. [18](#), [19](#), [23](#)
- Yang, T., Gopalan, V., Swart, P., and Mohideen, U. (1999). Direct observation of pinning and bowing of a single ferroelectric domain wall. *Physical review letters*, 82(20):4106. [69](#), [70](#)

- Yang, Z., Tian, Y., Zhou, T., Zhu, Y., Zhang, P., Chen, J., and Li, J. (2022b). Optimization of dry weight assessment in hemodialysis patients via reinforcement learning. *IEEE Journal of Biomedical and Health Informatics*, 26(10):4880–4891. [40](#)
- Yao, J., Killian, T., Konidaris, G., and Doshi-Velez, F. (2018). Direct policy transfer via hidden parameter markov decision processes. In *LLARLA Workshop, FAIM*, volume 2018. [19](#), [23](#), [24](#)
- Yu, C., Liu, J., Nemati, S., and Yin, G. (2021). Reinforcement learning in healthcare: A survey. *ACM Comput. Surv.*, 55(1). [5](#), [21](#), [28](#), [29](#)
- Zeng, X., Linwood, S. L., and Liu, C. (2022). Pretrained transformer framework on pediatric claims data for population specific tasks. *Scientific Reports*, 12(1):1–13. [60](#)
- Zha, D., Lai, K.-H., Tan, Q., Ding, S., Zou, N., and Hu, X. B. (2022). Towards automated imbalanced learning with deep hierarchical reinforcement learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2476–2485. [19](#)
- Zhang, S., Wen, L., Peng, H., and Tseng, H. E. (2021). Quick learner automated vehicle adapting its roadmanship to varying traffic cultures with meta reinforcement learning. *CoRR*, abs/2104.08876. [20](#), [23](#)
- Zhao, Y., Wong, Z. S.-Y., and Tsui, K. L. (2018). A framework of rebalancing imbalanced healthcare data for rare events’ classification: A case of look-alike sound-alike mix-up incident detection. [15](#), [23](#)
- Zhu, T., Li, K., Herrero, P., and Georgiou, P. (2021). Basal glucose control in type 1 diabetes using deep reinforcement learning: An in silico validation. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1223–1232. [30](#)

- Zhu, T., Lin, Y., and Liu, Y. (2020a). Oversampling for imbalanced time series data. *CoRR*, abs/2004.06373. [15](#), [16](#), [22](#), [23](#)
- Zhu, Z., Lin, K., and Zhou, J. (2020b). Transfer learning in deep reinforcement learning: A survey. *CoRR*, abs/2009.07888. [19](#), [24](#)

Vita

Benjamin Smith graduated from the University of Tennessee: Chattanooga in 2020 with a Bachelor of Science in Physics: Biophysics. He received his degree with departmental honors and graduate Cum Laude for his academic performance and research thesis. He then began his Ph.D. in Data Science and Engineering through the Bredesen Center for Interdisciplinary Research and Graduate Education at the University of Tennessee: Knoxville, graduating (hopefully) in 2023. His doctoral research focused on applications of reinforcement learning in healthcare and materials science. His research explores the use of reinforcement learning for heparin treatments as well as for materials science applications such as in ferroelectric domain wall control.