Towards Precision Oncology: A Predictive and Causal Lens

by

Zeshan Hussain

B.S., Stanford University (2016)
M.S., Stanford University (2017)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Authored by: Zeshan Hussain
Department of Electrical Engineering and Computer Science
June 30, 2023

Certified by: David Sontag
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by: Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

بسم الله الرحمن الرحيم

Towards Precision Oncology: A Predictive and Causal Lens

by

Zeshan Hussain

Submitted to the Department of Electrical Engineering and Computer Science
on June 30, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Precision oncology promises personalized care for each patient based on a holistic view of their data. However, several methodological and translational advances are required for successful implementation of this vision in the clinic. These include building temporal models to predict a patient's survival outcomes in response to therapy, validating these methods with experimental data from Randomized Controlled Trials (RCTs), quantifying the uncertainty in the predictions, and finally, exploring how these elements can be woven together into a clinical decision support tool. In this thesis, I explore each of these aspects in turn: i) first, I build different models of clinical time-series data, with a focus on prediction of survival outcomes and forecasting of core biomarkers, ii) next, I design methods to give additional "context" for these models, including uncertainty quantification of causal estimates and validation of these estimates using RCT data, and iii) finally, I study how these elements affect treatment decision-making via a controlled user study of a decision support tool prototype.

Thesis Supervisor: David Sontag
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

I began my doctoral studies within the broader structure of the MD-PhD program, with the goal of developing a robust technical toolset through which I could think meaningfully about clinical problems. There have been numerous mentors, friends, and family members who have played a pivotal role in this journey and to them I am eternally grateful.

I would like to begin by thanking my thesis supervisor and PhD advisor, David Sontag, whose boundless energy, intellectual rigor, and unbending support has motivated and inspired me in my work. His questions forced me to elevate my research and made me think deeply about why something did (or did not) work. He cultivated a lab environment that was collaborative and supportive, and it was because of this environment that I met and worked with some of the most brilliant researchers. I was lucky to have him as my advisor.

I would also like to thank the rest of my thesis committee for their helpful mentorship throughout the latter half of my PhD. Rahul Krishnan, now a Professor at the University of Toronto, was my first mentor when he was a senior PhD student in David's lab. It is no exaggeration to say that one of the reasons I pursued a PhD in the first place was to eventually be a researcher like Rahul. I admired his immense technical skill and ability to explain difficult concepts in accessible ways. Next, I would like to thank Marzyeh Ghassemi, whose knowledge of the literature, especially in Human-AI interaction, helped contextualize many aspects of this dissertation. Finally, I am indebted to Eli Van-Allen, whose experience as a physician-scientist in oncology was and will continue to be immensely helpful in my own career aspirations.

I have been blessed with many mentors and collaborators with whom I have worked over the past 4+ years who have both published papers with me and expanded my knowledge of machine learning, causal inference, and statistics. These esteemed individuals include: Ming-Chieh Shih, Michael Oberst, Fredrik Johansson, Rickard Karlsson, Martin Willbo, Ahmed Alaa, Ilker Demirel, Edward De Brouwer, Rebecca Boiarsky, Andrew Yee, Barbara Lam, and Maia Jacobs. Thank you for the many

all-nighters, coffee runs, endless whiteboarding, and innumerable research discussions. I would also like to thank Fernando Acosta Perez, whom I mentored, though he taught me more than I taught him.

I was also honored to have worked with immensely intelligent and kind labmates in the clinical ML group: Chandler Squires, Hunter Lang, Monica Agrawal, Christina Ji, Hussein Mozannar, Shannon Shen, Liz Bondi-Kelly, Alejandro Buendia, Niklas Mannhardt, Sharon Jiang, and Penny Brant. I am grateful for their insights and companionship and have fond memories of our lab hikes, dinners, and all-important tennis match viewings.

I would like to thank my many friends who have supported me throughout my academic career, starting in high school and then in Stanford and now finally in Longwood and Cambridge. Our weekly gatherings of remembrance, basketball games, dinner outings, and musings on the subtleties of life have been instrumental in helping me stay positive throughout this incredibly long academic path.

I am incredibly grateful to my two brothers, Naumaan and Imraan, as well my sister, Hafsa, for their unwavering friendship. I would like to thank Naumaan, especially, for his support during my lowest points. I am grateful for my in-laws for their support and optimism. Most of all, I would like to thank my parents, without whom I could not be where I am. It is their sacrifice, having come to this country with nothing, as well as their constant love and support that has kept me going. Finally, I would like to thank my wife, Juvaria, who has been nothing short of a shining light in my life and my best friend. I could not have completed this dissertation without her devotion and love. May God preserve each and every person who has played a role in my journey.

Wa-l-Ḥamdu li-l-Lāhi Rabb al-’Ālamīn

This doctoral thesis has been examined by a Committee of the Department of Electrical Engineering and Computer Science as follows:

Professor David Sontag . . . .
Thesis Supervisor
Professor of Electrical Engineering and Computer Science, MIT

Professor Marzyeh Ghassemi . . . .
Member, Thesis Committee
Assistant Professor of Electrical Engineering and Computer Science, MIT

Professor Eliezer Van Allen . . . .
Member, Thesis Committee
Associate Professor of Medicine, Harvard Medical School

Professor Rahul G. Krishnan . . . .
Member, Thesis Committee
Assistant Professor of Computer Science, The University of Toronto

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

List of Figures

List of Tables

Glossary of Recurring Acronyms

- AI: artificial intelligence
- AUC: area under the receiver operating characteristic curve (also AUROC)
- CDSS: clinical decision support system
- EHR: electronic health record
- LSTM: Long Short-term Memory
- ML: machine learning
- OBS: observational data
- RCT: randomized controlled trial
- RNN: Recurrent Neural Network
- UI: user interface

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

1.1 A Vision of Precision Oncology

Patient data, including biomarkers, genetics, labs, and demographics, have become increasingly available through Electronic Health Records (EHRs) and other observational datasets [158]. The advent of powerful data-driven models trained on such datasets has opened the doors for precision medicine, where clinicians can tailor their treatment decisions to the patient at hand [292, 2, 133]. The promise of personalized care is relevant for oncology, where a plethora of possible treatments exists for each cancer and the uniqueness of each patient often leads to significant heterogeneity in treatment response [95].

The workflow of an oncologist who has access to tools that enable precision medicine will look radically different than workflows in current clinical practice. Imagine an oncologist seeing a patient who has been newly diagnosed with multiple myeloma. Suppose that the oncologist has access to a clinical decision support system (CDSS) of the kind shown in Figure 1-1. In trying to determine the best treatment for their patient, the oncologist can select two or more treatments to consider. A holistic view of what the patient's disease progression might look like under each potential treatment would be given. For example, the CDSS would show the projected progression free and overall survival as well as the adverse event profile under each treatment. Additionally, the predicted evolution of important biomarkers, such as the serum immunoglobulins

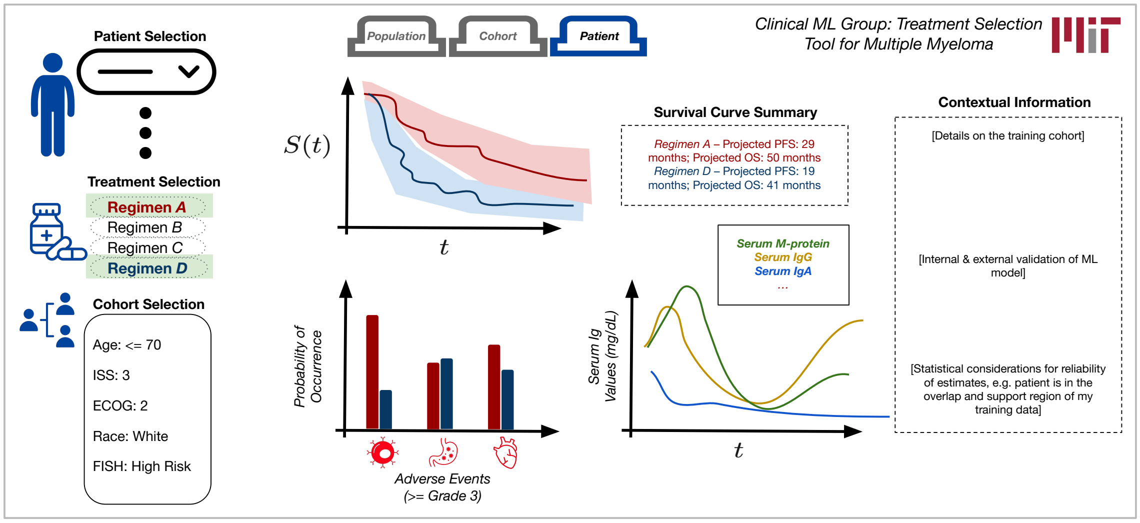

Patient Selection
Treatment Selection
Regimen A
Regimen B
Regimen C
Regimen D
Cohort Selection
Age: <= 70
ISS: 3
ECOG: 2
Race: White
FISH: High Risk
Population
Cohort
Patient
Clinical ML Group: Treatment Selection Tool for Multiple Myeloma
$S(t)$
t
Probability of Occurrence
Adverse Events (>= Grade 3)
Survival Curve Summary
Regimen A – Projected PFS: 29 months; Projected OS: 50 months
Regimen D – Projected PFS: 19 months; Projected OS: 41 months
Serum M-protein
Serum IgG
Serum IgA
...
Serum Ig Values (mg/dL)
t
Contextual Information
[Details on the training cohort]
[Internal & external validation of ML model]
[Statistical considerations for reliability of estimates, e.g. patient is in the overlap and support region of my training data]

Figure 1-1: A vision for a clinical decision support system for oncologists.

in the case of multiple myeloma, over several months might be shown. While these predictions are given by underlying machine learning models tailored to the patient, the oncologist also has access to population-level estimates given by Randomized Controlled Trials (RCTs), which comprise the current gold-standard evidence for establishing efficacy of a therapy [39]. Importantly, the CDSS could also give the oncologist "sanity checks" of the predictions that they see. For example, confidence intervals for the survival curves and the adverse event estimates could be given to quantify the uncertainty in the predictions. "Contextual information", such as details of the model training cohort, what covariates were included as input, internal and external validation results of the model, as well as other statistical considerations such as whether the patient was well-represented in the training data, could also be included.

This vision stands in contrast to current clinical practice, where clinicians and committees that form treatment guidelines rely on RCTs to make treatment decisions. RCTs, though seen as the highest calibre of evidence, often only give unbiased estimates of treatment effect averaged across a narrow population. RCT results may not always apply to every patient, and concerns about the generalizability of some clinical trials have been posited [230]. Thus, allowing a physician to leverage both RCTs and ML-based systems that provide personalized predictions has the potential to better meet the goal of managing cancer, which is to maximize the survival of the patient while maintaining their quality of life.

1.2 Challenges

However, the path to reaching the vision described above is difficult. The first challenge is with respect to the underlying clinical data and how one models it. The data can be very sparse, containing many missing values depending on the nature of the patient's disease state [158]. For example, in multiple myeloma, getting a complete blood count for the patient may not be necessary at each visit, especially if they are in remission. Underlying relationships between treatments and biomarkers can be highly nonlinear and may differ across patient subgroups. Finally, sample sizes may be small, i.e. on the order of a few thousand patients, especially for rare diseases. All of these properties motivate the need for data-efficient, expressive modeling architectures that can both handle the nature of clinical temporal data **and** are clinically useful.

A second challenge is the need for contextualizing and giving meaningful "sanity checks" of any ML-based system to a clinician. There are several technical problems that arise when attempting to do this assessment, which motivates a portion of the methods found in this thesis. For one, a prediction shown in a CDSS should have confidence intervals that give a lens into the uncertainty of that prediction. Although uncertainty quantification is straightforward when reporting a population-level result from an RCT, it is less so for a prediction from an ML model. Ideally, a confidence interval for the latter would be interpretable not just in an average sense with respect to the entire population but would also be valid for the specific patient for whom the prediction has been made. This type of conditional validity may be difficult to achieve without further assumptions on the data distribution or the modeling architecture. A second technical issue involves using a small amount of experimental data to verify causal estimates derived from ML models trained on observational data. Such an approach is inspired by the medical community's acknowledgment of the strength of sound RCT-based evidence. Consequently, employing RCT estimates as a benchmark to validate estimates derived from observational data on a cohort with similar attributes can prove to be effective.

Still, in spite of any methodological advancements we might make to overcome the above challenges, what remains unclear is how physicians may interact with a CDSS that uses these methods, especially in the context of their current clinical workflows. This challenge motivates conducting studies of physician-AI interaction, particularly in the oncology setting, to understand better how the decision-making of a clinician is affected.

1.3 Contributions

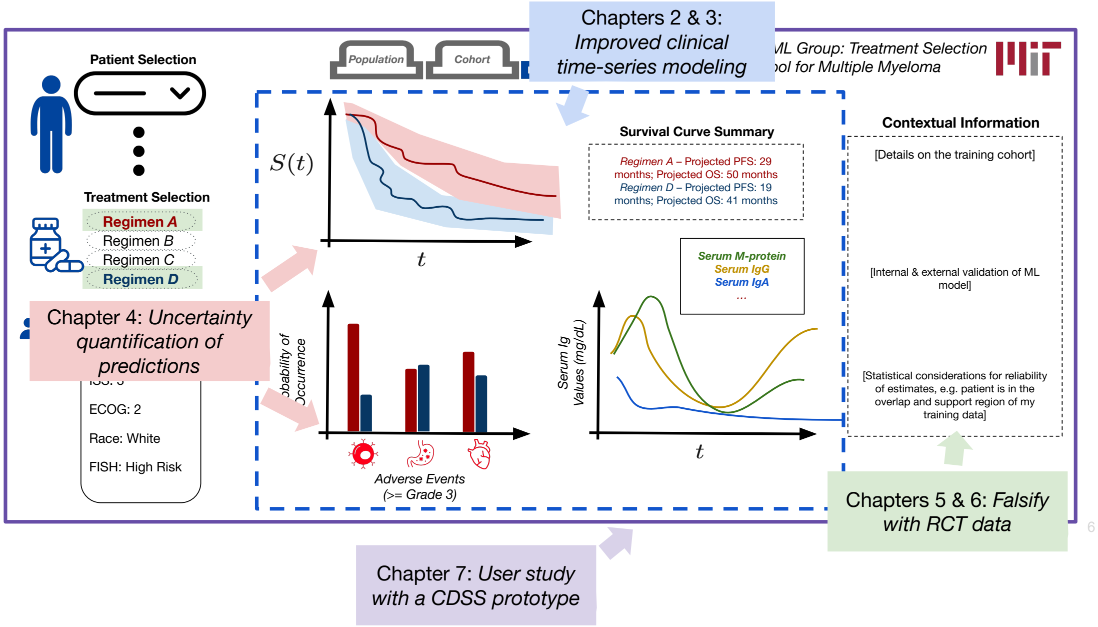

Chapters 2 & 3: *Improved clinical time-series modeling*
ML Group: Treatment Selection
ool for Multiple Myeloma
Patient Selection
Population
Cohort
Treatment Selection
Regimen A
Regimen B
Regimen C
Regimen D
Chapter 4: *Uncertainty quantification of predictions*
ECOG: 2
Race: White
FISH: High Risk
$S(t)$
t
obability of Occurrence
Adverse Events (>= Grade 3)
Survival Curve Summary
Regimen A – Projected PFS: 29 months; Projected OS: 50 months
Regimen D – Projected PFS: 19 months; Projected OS: 41 months
Serum M-protein
Serum IgG
Serum IgA
...
Serum Ig Values (mg/dL)
t
Contextual Information
[Details on the training cohort]
[Internal & external validation of ML model]
[Statistical considerations for reliability of estimates, e.g. patient is in the overlap and support region of my training data]
Chapters 5 & 6: *Falsify with RCT data*
Chapter 7: *User study with a CDSS prototype*
6

Figure 1-2: Annotation of clinical decision support system according to thesis chapters.

In this thesis, I present machine learning and statistical methods that can form the underlying machinery for a CDSS in clinical oncology. I also provide an initial, simulated exploration of how physicians might engage with such a tool. In the chapters that follow, I address the challenges listed above by focusing on the following themes: building better models of clinical temporal data, which will enable prediction of important clinical outcomes over time in response to treatment, developing statistical methods that provide additional "context" to these models, and studying how physicians interact with an ML-based CDSS. The chapters in this thesis are organized as follows and connect back to the CDSS as shown in Figure 1-2.

Building predictive models of clinical temporal data: Oncologists approach treatment decision making for their patients multifactorially, aiming to maximize survival of the patient while limiting any adverse events. Over time, they will look at the progression of clinical biomarkers to keep track of how their patient is doing. I develop models that can do each of these tasks for the clinician, from prediction of survival outcomes to forecasting of biomarkers.

- Chapter 2 (originally published at ICML 2021 as [122]) focuses on the problem of forecasting clinically-relevant biomarkers conditioned on a patient's labs, genetics, demographics, and treatment regimen. We demonstrate how to incorporate known or conjectured responses of disease biomarkers to treatment as inductive biases in state space models. The proposed model shows strong improvements in generalization compared to several baselines and prior state of the art. We show how the model can provide insights into multiple myeloma progression.

- Chapter 3 (submitting as a journal article titled "Joint attention-based event prediction and longitudinal modeling in newly diagnosed and relapsed multiple myeloma") turns to attention-based architectures instead of sequential state space models and looks at how to provide a more holistic view of a patient's disease progression. We propose a deep-learning approach that jointly models several important tasks for the clinical management of multiple myeloma. These include (1) predicting progression-free survival (PFS), overall survival (OS), and adverse events (AE), (2) forecasting key disease biomarkers, and (3) estimating the effect of using different treatment strategies. Importantly, our model dynamically adapts to newly collected clinical data and performs the above tasks with observed clinical history of arbitrary lengths.

Developing statistical methods as "sanity checks" for predictive and causal estimates: I develop methods that provide additional context for predictions from ML models. One approach is through uncertainty quantification, where we show how to give valid confidence intervals for ML model predictions. The second is through using RCT data as a means of assessing the reliability of causal estimates from observational data.

- Chapter 4 (published as a part of [4] at AISTATS 2023) builds on the framework of conformal prediction to develop a method that communicates model uncertainty for specific prediction instances in an adaptive and transparent manner. Specifically, we construct a predictive inference procedure that *adapts* the length of its issued confidence intervals based on the varying level of uncertainty across different prediction instances, and reports a coverage guarantee that is "relevant" to each specific instance.

- Chapter 5 (originally published as [120] at NeurIPS 2022) proposes a meta-algorithm that attempts to reject, or *falsify*, observational estimates that are biased. We do so using *validation effects*, causal effects that can be inferred from both RCT and observational data. After rejecting estimators (e.g. from multiple observational studies) that do not pass this test, the algorithm generates conservative confidence intervals on the *extrapolated* causal effects for subgroups not observed in the RCT. Under the assumption that at least one observational estimator is asymptotically normal and consistent for both the validation and extrapolated effects, guarantees are provided on the coverage probability of the intervals output by our algorithm. We illustrate the properties of this approach on semi-synthetic and real world datasets, and show that it compares favorably to standard meta-analysis techniques.

- Chapter 6 (originally published as [121] at AISTATS 2023) continues to study the problem of *falsification*, or using limited RCT data to "falsify" causal assumptions made in observational studies. Our method can be applied even when the RCT data does not cover the entire observational population, and hence cannot be used on its own to estimate causal effects. Assuming that the RCT can yield unbiased treatment effects, we show that certain causal assumptions of the observational study have a testable implication in the form of a set of conditional moment restrictions (CMRs). We propose a *falsification algorithm* that tests whether or not these CMRs hold, thereby providing an opportunity to reject these assumptions when they fail to hold. This procedure allows us to take advantage of approaches developed in the econometrics literature for testing CMRs. We use a Maximum Moment Restriction (MMR)-based test [185] for this purpose.

Physician-AI interaction for oncology: Finally, I build a CDSS prototype with which I run a user study to assess how physicians interact with synthetic versions of several of the aforementioned ML-based components.

- Chapter 7 (submitting as a journal article titled "Evaluating Physician-AI Interaction for Cancer Management: Paving the Path towards Precision Oncology") seeks to understand how physicians interact with an ML-based system while also having access to accepted evidence on which current clinical practice is based, e.g. findings from an RCT, via a user study of 32 physicians. Our findings suggest that future ML-based CDSS systems do have the potential to change treatment decisions in cancer management, but that meticulous development and validation of these systems before deployment is crucial.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Background

In this chapter, we introduce core concepts in Bayesian networks, probability theory, and causal inference that will be important background material for the rest of this thesis. For a more rigorous treatment of causal inference, we refer the reader to [112]. We use some of the background material on probability and Bayesian networks from [97]. Finally, some of the background material in the causal inference section (Section 2.3) is based on [186].

2.1 Basic Probability Theory

Probability theory forms the basis for statistical analysis and decision-making. It is a framework used to model uncertain events and quantify their likelihood. We begin with the definition of key concepts. A sample space is the set of all possible outcomes of a random experiment, denoted by Ω. An event is a subset of the sample space, representing a particular outcome or combination of outcomes. Events are denoted by capital letters, such as X, Y, or Z. A probability measure assigns a numerical value between 0 and 1 to each event in the sample space. It captures the likelihood of an event occurring. The probability of an event A is denoted as $P(A)$.

A random variable is a variable whose value, corresponding to some event of interest, is unknown but may be one of many values. The domain of a random variable may be discrete (e.g. the side of a die) or continuous (e.g. diastolic blood pressure after an intervention). Thus, a probability distribution describes the likelihood of different values occurring for a random variable. For discrete random variables, it is represented by a probability mass function (PMF). For continuous random variables, it is represented by a probability density function (PDF). Notationally, $P(Y = 2)$ denotes the chances that the random variable Y takes on an assignment of 2.

Conditional probability measures the likelihood of an event occurring given that another event has already occurred. It is denoted by $P(A|B)$, where A and B are events. Two random variables X and Y are independent if and only if their joint probability distribution is equal to the product of their individual probability distributions. Mathematically, for independent random variables X and Y, $P(X, Y) = P(X)P(Y)$. Similarly, two random variables X and Y are conditionally independent given a third random variable if they are independent in their conditional probability distributions given this random variable, i.e. $X \perp\!\!\!\perp Y|Z$ if and only if $P(X, Y|Z) = P(X|Z)P(Y|Z)$.

Now, we can introduce a few basic probability rules based on the definitions above.

- Sum Rule – For two random variables X and Y, the sum rule states that $P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$, where $\cup$ denotes the union of the events spanned by the random variables X, Y and $\cap$ denotes the intersection of the events.

- Chain Rule – The chain rule enables us to calculate the joint probability distribution of multiple random variables. It states that the joint probability distribution of a sequence of random variables $X_1, X_2, \ldots, X_n$ can be obtained by multiplying the conditional probability distributions of each random variable given the occurrence of the previous random variables. For example, we have $P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2) \ldots P(X_n|X_1, X_2, \ldots, X_{n-1})$

- Bayes' Rule – An immediate consequence of the Chain Rule is Bayes' Rule, which states, $P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$.

These basic rules of probability provide a solid foundation for analyzing and solving problems involving uncertain events.

2.2 Bayesian Networks

2.2.1 Preliminaries

Probabilistic graphical models (PGMs) [146] use graphs, which consist of nodes and edges between them, to represent the complex relationships among potentially many random variables. A Bayesian network, or directed graphical model, is a type of PGM that represents a set of variables and their probabilistic dependencies using a directed acyclic graph (DAG). The nodes of the graph represent random variables, and the edges capture the conditional dependencies among the variables. Namely, an edge pointing from node A to node B in a Bayesian network indicates that A is a direct parent of B and that B depends on A. PGMs may also be undirected, i.e. undirected graphical models, but we will be only dealing with directed graphical models in this thesis. We will interchangeably use the terms Bayesian network, directed graphical model, and directed graph in this section.

One distinct advantage of using the language of graphical models is that it allows the practitioner to encode dependencies between random variables that imply a certain factorization over the joint distribution of the random variables. This follows from an assumption that we will now introduce, and an example illustrating factorization of a Bayesian network that follows from this assumption. The development that follows is taken from [186].

Assumption 2.2.1 (Local Markov Assumption)**.** Given its parents in a directed acyclic graph (DAG), a node X is independent of all its non-descendants.

One of the main consequences of Assumption 2.2.1 is Bayesian network factorization, which we define below:

Definition 2.2.1 (Bayesian Network Factorization)**.** Given a probability distribution, P, and a DAG, G, P factorizes according to G if,

$$P(X_1, \dots, X_n) = \prod_i P(X_i | \mathrm{pa}(X_i)), \tag{2.1}$$

where pa(·) refers to the parents of a node.

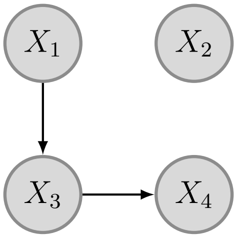


Figure 2-1: Four node directed graph

Given the above definition and assumption, we can factorize the example in Figure 2-1, as follows,

$$P(X_1, \ldots, X_4) = P(X_1)P(X_2)P(X_3|X_1)P(X_4|X_3) \tag{2.2}$$

Note that we are able to more efficiently model the joint distribution of these four binary random variables when taking into account the local dependencies between them. Namely, instead of needing $2^n - 1$ parameters to model the joint distribution of n variables, we need only the number of parameters necessary to model each conditional probability in the factorization. As n grows larger, this idea becomes increasingly important.

Another advantage of using this framework is that it allows a domain expert to specify which local dependencies exist between variables and which ones do not. For example, when modeling the relationship between random variables representing cardiac markers and other physiologic parameters such as blood pressure, stroke volume, and pulse, a cardiologist would be well equipped to construct a directed graphical model mapping out the relationship between these variables.

2.2.2 Independence in Directed Graphical Models

Depending on the structure of the directed graphical model, we are able to glean which variables are independent of each other in the graph. These notions of independence will prove useful when thinking about causal graphs and causal inference more generally, which we introduce later on.

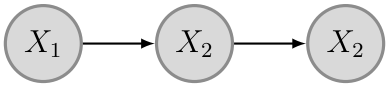


Figure 2-2: 3-node chain graph

Two variables are marginally independent in a directed graph if there is no directed path between the two variables. The simplest case is with two disconnected nodes, X_1 and X_2. Under Assumption 2.2.1, we can factorize the joint distribution of these random variables as follows, $P(X_1, X_2) = P(X_1)P(X_2)$, which immediately gives us independence.

Conditioning on a variable in a directed graph may also result in two other variables becoming conditionally independent given the variable that was conditioned on. An example is given in Figure 2-2, where $X_1 \perp\!\!\!\perp X_3|X_2$. This can be easily proven by factorizing the directed graph under Assumption 2.2.1 and showing that $P(X_1, X_3|X_2) = P(X_1|X_2)P(X_3|X_2)$.

There are instances where conditioning on a variable may not result in conditional independence, but rather induces dependence between two other variables. This phenomenon occurs most prominently in “v-structures” or “colliders”, where a child node has two parents that are not connected. In fact, not conditioning on the chid renders the parents independent. The following example gives intuition as to why two variables become dependent when conditioning on their child: let X_1 denote whether a sprinkler is on or not, X_2 denotes whether it is raining or not, and X_3 denotes whether the grass is wet or not. Knowing X_3, i.e. the grass is wet, implies that either it rained, meaning that it is less likely that the sprinkler was on, or that the sprinkler was on, implying that it is less likely that it rained.

2.2.3 Parameterizations of Bayesian Networks

There are various ways of parameterizing Bayesian networks, which all result in different models that can be learned from data. We discuss two different settings below.

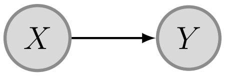


Figure 2-3: 2-node graph.

Supervised Learning In supervised learning, we are given a dataset, $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where X_i represents a set of potentially multi-dimensional covariates and Y_i represents the corresponding label, which can be discrete, binary, or continuous. The goal of supervised learning is to learn a model that, when given a new instance, X_k, can predict the correct outcome, Y_k. One example of a supervised learning problem is to predict whether a patient has pneumonia or not from a chest X-ray.

Figure 2-3 depicts the Bayesian network that corresponds to several common models used for supervised learning. We let θ denote the parameters of the model for the conditional distribution, $P(Y|X;\theta)$. Two common choices for parameterization are *linear regression*, where $P(Y|X) = \mathcal{N}(W^T X + b, \mathbf{I})$, and *logistic regression*, where $P(Y|X) = \frac{1}{1+\exp W^T X + b}$. In both of these cases, the parameters of the model are W and b, i.e. $\theta = \{W, b\}$. We may also parameterize the conditional distribution with nonlinear functions, which we do often throughout this thesis, e.g. $P(Y|X) = \mathcal{N}(f(X;\theta), \mathbf{I})$, where $f(\cdot)$ can be a neural network.

Unsupervised Learning Unsupervised learning encompasses a range of methods designed to model the likelihood of the data, given samples, $(X_1, Y_1), \ldots, (X_n, Y_n)$, from a dataset, $\mathcal{D}$. The objective is to construct a model that approximates the true distribution of the data, $P(X)$. This process is commonly known as density estimation. The fundamental principle of unsupervised learning is that a successful model for density estimation captures the essential characteristics of the dataset, identifying the key patterns and structures within the data.

2.2.4 Causal Graphs

So far, we have interpreted the relationships implied by directed graphs as one of statistical dependencies or independencies. If we assume that each edge in the graph takes a causal meaning, whereby a parent node is the direct cause of its children (see

Causal Edges Assumption in [186]), then the directed graph becomes a *causal* directed graph.

This interpretation becomes important when we want to understand the causal relationships between variables, beyond the associational or statistical relationship between them. The type of independence statements made in Section 2.2.2 will be relevant when discussing how to estimate the causal effect of an intervention, A, on an outcome, Y, while controlling for other paths between A and Y in the directed graph. In the next section, we will discuss important concepts in causal inference, which give us the tools to determine if such causal effects are even identifiable (e.g. there is only one path from A to Y and we can "block" all other paths) from data.

2.3 Causal Inference from Observational Data

In this section, we cover the basics of causal inference via the potential outcomes framework proposed by Rubin in the late 1970s [232]. However, these same concepts can be made via the framework based on causal graphs, which we will briefly discuss at the end of this section.

2.3.1 Neyman-Rubin Potential Outcomes Framework

In this framework, we assume that there are multiple treatments or interventions available. For example, we may have two treatments, represented by the random variable $A \in \{0, 1\}$. Each patient has a *potential outcome* under each treatment $A = a$, which is the outcome that a patient would have if they *potentially* receive treatment a. In reality, we observe only one of these potential outcomes, called the *observed outcome*, since a patient would only have been given either $A = 0$ or $A = 1$. The other potential outcome would be referred to as the *counterfactual outcome.* Thus, letting Y_a be the random variable denoting the potential outcome under $A = a$, we never actually observe the treatment effect for the individual patient, i.e. individual treatment effect: $Y_1 - Y_0$. Similarly, we could also try and find the average treatment effect (ATE) for the population, where we are essentially trying to compute the difference in average outcome of the patient population had they all been given $A = 1$ and the average outcome of the patient population had they all been given $A = 0$. Mathematically, this would be,

$$\text{ATE} := \mathbb{E}[Y_1 - Y_0], \tag{2.3}$$

where the expectation is with respect to population distribution. Importantly, the notion of assigning a single intervention to the entire population is *counterfactual* since it is not done in reality but rather imagined by the researcher. As such, both the ITE and the ATE are impossible to compute without additional assumptions, since we do not observe both potential outcomes of the patients in the real world.

The assumptions that are required to make the estimation of the ATE possible are required *identifiability assumptions*. We formalize these assumptions now. In addition to our binary random variable A denoting the treatment and the random variable Y_a denoting the potential outcome had the patient received treatment a, we also introduce X, which denotes a vector of potential confounders. A confounder is defined as a variable that affects both the treatment assignment and the outcome. The ATE is identifiable if the following assumptions are satisfied [112],

- (*Conditional Exchangeability between Treatments*) $Y_a \perp\!\!\!\perp A|X, \forall a \in \{0, 1\}$
- (*Consistency*) $Y = Y^A$
- (*Positivity of Treatment Assignment*) $P(X) > 0 \implies P(A = a|X)) > 0, \forall a \in \{0, 1\}$

The first assumption implies that there are *no unobserved confounders*, i.e. we include all confounders in X. This assumption is important because it allows us to control for any confounding effects that may mask the true causal effect. The second assumption states that when a patient is given a treatment a, the outcome is the same as the potential outcome under a. In other words, the way of assigning the treatment is consistent and results in the same outcome. Finally, the third assumption assumption states that within any covariate strata in the patient population, each individual has a nonzero chance of receiving either treatment.

Similar assumptions need to be made in order to estimate the ATE for a particular subgroup of patients, who have covariates $X = x$. This quantity is also referred to as the conditional average treatment effect (CATE) and can be written as,

$$\text{CATE} := \mathbb{E}[Y_1 - Y_0 | X = x] \tag{2.4}$$

2.3.2 Estimators of Causal Effects

Under the identifiability assumptions introduced in the prior section, numerous estimators have been proposed to provide asyptotically unbiased estimates of the ATE and CATE. One approach is to model the average outcome, conditioned on the treatment and covariates, $E[Y|A = a, X = x]$, termed the response function. Let $\hat{\mu}_a(x)$ be an asymptotically unbiased estimator of $E[Y|A = a, X = x]$. We could use any model for the response surface, such as a regression model, random forest, or neural network. Then, an asymptotically unbiased estimator for the ATE is,

$$\mathbb{E}_n[\hat{\mu}_1(X) - \hat{\mu}_0(X)], \tag{2.5}$$

where $\mathbb{E}$ is the empirical average of the computed treatment effects for each patient. This estimation procedure goes by many names, including *response surface modeling*, *conditional outcome modeling*, and *g-computation*.

Another estimation approach is weighting by estimates of the propensity score, which is the probability that a patient receives a specific treatment given their covariates: $P(A = a|X = x)$. We refer to an estimate of the propensity score as $\hat{e}_a(X)$. When X is low-dimensional, logistic regression is often used as the model of choice for propensity score estimation. The propensity score estimates are then plugged into an inverse probability of treatment weighting (IPTW) estimator,

$$\mathbb{E}_n\left[\left(\frac{\mathbf{1}(A=1)}{\hat{e}_1(X)} - \frac{\mathbf{1}(A=0)}{\hat{e}_0(X)}\right)Y\right], \tag{2.6}$$

which gives an asymptotically unbiased estimate of the ATE, as long as the estimate of the propensity score function is also asymptotically unbiased. Intuitively, we are weighting all those who receive $A = 1$ inversely proportionally to the probability of their receiving $A = 1$, and similarly for $A = 0$.

Although the response surface modeling estimator and the IPTW estimator are both valid estimators, they only yield asymptotically unbiased estimates if the underlying response functions and propensity score functions are estimated in an asymptotically unbiased fashion. Thus, for the response surface modelling estimator, the functional form of the response surface must be correctly specified, and the same holds for the propensity score function in the IPTW estimator. As a result, a *doubly robust* estimator that combines both the response surface and IPTW estimators has been proposed,

$$\mathbb{E}_n\left[\hat{\mu}_1(X) - \hat{\mu}_0(X) + \left(\frac{\mathbf{1}(A=1)}{\hat{e}_1(X)} - \frac{\mathbf{1}(A=0)}{\hat{e}_0(X)}\right)(Y - \hat{\mu}_A(X)\right], \tag{2.7}$$

The main advantage of a doubly robust estimator is that it yields a valid, asymptotically unbiased estimator if either the the estimated response surface functions or the estimated propensity scores are asymptotically unbiased [112]. However, it is not required for *both* estimated functions to be correctly specified.

Finally, we denote the propensity score functions and response surface functions as *nuisance functions*, in that they need to be estimated to get a final estimate of the causal effect, but they themselves are not the targets of estimation.

2.3.3 Causal Graph Perspective: Backdoor and Frontdoor Adjustment

Identification of the causal effect, which required the assumptions introduced in the beginning of this section, can also be thought about from the perspective of causal graphs. The following development is taken from [186]. As before, let A denote treatment and Y denote the outcome. Suppose we have the causal graph shown in Figure 2-4. First, we have the following definitions,

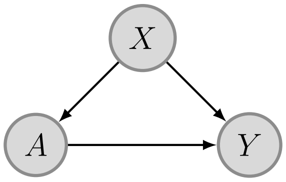


Figure 2-4: Canonical causal graph

Definition 2.3.1 (Blocked Path)**.** A path between nodes A and Y is blocked by a (potentially empty) conditioning set X if either of the following is true: 1. Along the path, there is a chain $\ldots \rightarrow W \rightarrow$ or a fork $\ldots \leftarrow W \rightarrow$, where W is conditioned on $(W \in X)$. 2. There is a collider W on the path that is not conditioned on $(W \notin X)$ and none of its descendants are conditioned on.

Definition 2.3.2 (Backdoor Criterion)**.** A set of variables X satisfies the backdoor criterion relative to A and Y if the following are true:

- X blocks all backdoor paths from A to Y.
- X does not contain any descendants of A.

Now, in the context of the causal graph in Figure 2-4, where X fulfills the backdoor criterion given in Definition 2.3.2, one can prove that the causal effect of A on Y can be identified as long as X is conditioned on. Graphically, conditioning on X amounts to blocking the flow of association from A to Y through X. As long as X contains all variables such that there are no other backdoor paths, which is equivalent to there being no unobserved confounders, then we can identify the causal effect. This procedure is famously known as *backdoor adjustment*.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Incorporating Structure of Longitudinal Clinical Data in Sequential Models

Acknowledgement of Co-authors I would like to acknowledge Rahul Krishnan, who was instrumental in driving the initial stages of this work, and additionally for his guidance and mentorship.

In this chapter, I begin with an initial foray into building a model of structured, longitudinal clinical data, with a focus on biomarker forecasting. I study how to incorporate known or conjectured responses of disease biomarkers to treatment as inductive biases in state space models. This chapter provides one methodological approach to modeling clinical data temporally.

3.1 Introduction

Clinical biomarkers capture snapshots of a patient's evolving disease state as well as their response to treatment. However, these data can be high-dimensional, exhibit missingness, and display complex nonlinear behaviour over time as a function of time-varying interventions. Good unsupervised models of such data are key to discovering new clinical insights. This task is commonly referred to as disease progression modeling

[283, 273, 239, 75, 172, 3, 246].

Reliable unsupervised models of time-varying clinical data find several uses in healthcare. One use case is enabling practitioners to ask and answer counterfactuals using observational data [231, 205, 30]. Other use cases include guiding early treatment decisions based on a patient's biomarker trajectory, detecting drug effects in clinical trials [183], and clustering patterns in biomarkers that correlate with disease sub-type [300]. To do these tasks well, understanding how a patient's biomarkers evolve over time given a prescribed treatment regimen is vital, since a person's biomarker profile is often the only observed proxy to their true disease state. Like prior work [3, 246, 148], we frame this problem as a conditional density estimation task, where our goal is to model the density of complex multivariate time-series conditional on time-varying treatments.

Representation learning exposes a variety of techniques for good conditional density estimation [44, 181, 55, 260]. For sequential data, a popular approach has been to leverage black-box, sequential models (e.g. Recurrent Neural Networks (RNNs)), where a time-varying representation is used to predict clinical biomarkers. Such models are prone to overfitting, particularly on smaller clinical datasets. More importantly, such models often make simplistic assumptions on how time-varying treatments affect downstream clinical biomarkers; for example, one choice is to concatenate treatments to the model's hidden representations [3, 148]. The assumption here is that the neural network learns how treatments influence the representation. We argue that this choice is a missed opportunity and better choices exist. Concretely, we aim to encourage neural models to learn representations that encode a patient's underlying disease burden by specifying how these representations evolve due to treatment. We develop a new disease progression model that captures such insights by using inductive biases rooted in the biological mechanisms of treatment effect.

Inductive biases have been integral to the success of deep learning in other domains such as vision, text and audio. For example, convolutional neural networks explicitly learn representations invariant to translation or rotation of image data [157, 131, 272], transformers leverage attention modules [16, 271] that mimic how human vision pays attention to various aspects of an image, and modified graph neural networks can explicitly incorporate laws of physics to generalize better [243]. For some learning problems, the physics underlying the domain are often known, e.g. the laws of motion, and may be leveraged in the design of inductive biases [169, 8, 282]. The same does not hold true in healthcare, since exact disease and treatment response mechanisms are not known. However, physicians often have multiple hypotheses of how the disease behaves during treatment. To capture this intuition, we develop inductive biases that allow for a data-driven selection over multiple neural mechanistic models that dictate how treatments affect representations over time.

3.1.1 Contributions

In this chapter, we present a new attention-based neural architecture, $\mathbb{PKPD}_{\text{Neural}}$, that captures the effect of drug combinations in representation space (Figure 3-1 [top]). It learns to attend over multiple competing mechanistic explanations of how a patient's genetics, past treatment history, and prior disease state influence the representation to predict the next outcome. The architecture is instantiated in a state space model, $\mathbf{SSM}_{\text{PK-PD}}$, and shows strong improvements in generalization compared to several baselines and prior state of the art. We demonstrate the model can provide insights into multiple myeloma progression. Finally, we release a disease progression benchmark dataset called ML-MMRF, comprising a curated, pre-processed subset of data from the Multiple Myeloma Research Foundation CoMMpass study [188]. Our model code can be found at `https://github.com/clinicalml/ief`, and the data processing code can be found at `https://github.com/clinicalml/ml_mmrf`.

3.2 Related Work

Much work has been done across machine learning, pharmacology, statistics and biomedical informatics on building models to characterize the progression of chronic diseases. Gaussian Processes (GPs) have been used to model patient biomarkers over time and estimate counterfactuals over a single intervention [90, 240, 248, 251].

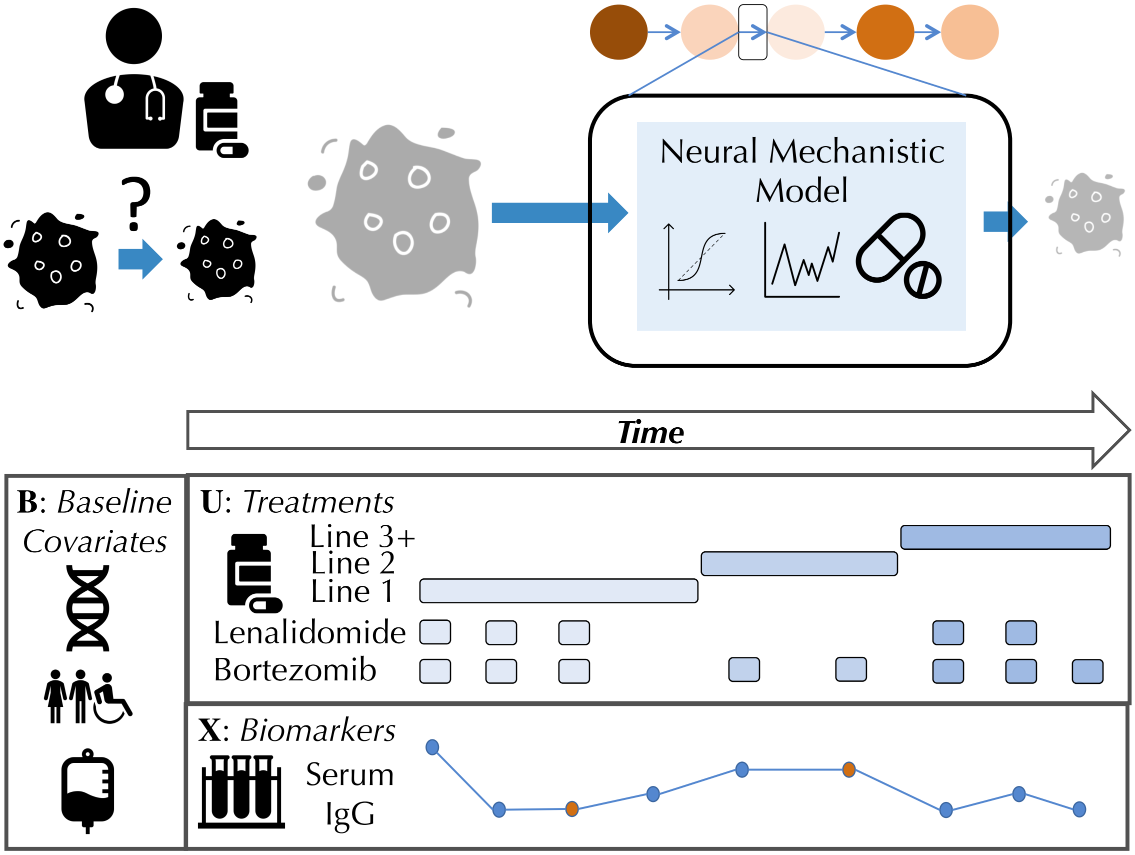


Figure 3-1: Inductive Bias Concept (Top): A clinician often has multiple mechanistic hypotheses as to how the latent tumor burden evolves. Our approach formalizes these hypotheses as neural architectures that specify how representations respond to treatments. **Patient Data (Bottom):** Illustration of data from a chronic disease patient. Baseline (static) data typically consists of genomics, demographics, and initial labs. Longitudinal data typically includes laboratory values (e.g. serum IgG) and treatments (e.g. lenalidomide). Baseline data is usually complete, but longitudinal measurements are frequently missing at various time points.

In each of these cases, the focus is either on a single intervention per time point or on continuous-valued interventions given continuously, both strong assumptions for chronic diseases. To adjust for biases that exist in longitudinal data, Lim et al. [168], Bica et al. [30] use propensity weighting to adjust for time-dependent confounders. However, they concatenate multi-variate treatments to patient biomarkers as input to RNNs; when data is scarce, such approaches have difficulty capturing how the hidden representations respond to treatment.

State space models and other Markov models have been used to model the progression of a variety of chronic diseases. [3] use an attention mechanism over an auto-regressive state space model to build a model of progression for Cystic Fibrosis, and [239] model the predicted forced vital capacity of individuals with scleroderma. In addition, [261] use multi-state Markov models to characterize the progression among breast cancer patients. [283] model the progression of COPD using continuous time Markov jump processes. [259] model clinical observations from patients suffering from Huntington's disease. There has also been much research in characterizing disease trajectories, subtypes, and correlations between risk factors and progression for patients suffering from Alzheimer's Disease [141, 98, 297, 177]. Like us, the above studies pose disease progression as density estimation but in contrast, many of the above models do not condition on time-varying interventions.

Further, [148] use neural networks to parameterize the transition function in nonlinear state space models of diabetic patients. However, this approach, as we study, is susceptible to overfitting. [207] use linear dynamical systems to model the progression of patients suffering from Chronic Kidney Disease. This approach suffers when the biomarkers vary non-linearly.

3.3 Preliminaries & Background

SSMs are a popular model for sequential data and have a rich history in modeling disease progression. They admit flexible parametric forms in their transition and emission distributions, which is a distinct advantage over other common model families such as RNNs, and allow for the latent state to be continuous. Additionally, many prior works have studied Hidden Markov Models (HMMs), and SSMs provide a natural continuous-variable analogue of the same.

Notation: $B \in \mathbb{R}^J$ denotes baseline data that are static, i.e. individual-specific covariates. For chronic diseases, these data comprise a J-dimensional vector, including patients' age, gender, genetics, race, and ethnicity. Let $\mathbf{U} = \{U_0, \ldots, U_{T-1}\}$; $U_t \in \mathbb{R}^L$ be a sequence of L-dimensional interventions for an individual. An element of U_t may be binary, to denote prescription of a drug, or real-valued, to denote dosage.

$\mathbf{X} = \{X_1, \ldots, X_T\}$; $X_t \in \mathbb{R}^M$ denotes the sequence of real-valued, M-dimensional clinical biomarkers. An element of X_t may denote a serum lab value or blood count, which is used by clinicians to measure organ function as a proxy for disease severity. X_t frequently contains missing data. We assume access to a dataset $\mathcal{D} = \{(\mathbf{X}^1, \mathbf{U}^1, B^1), \ldots, (\mathbf{X}^N, \mathbf{U}^N, B^N)\}$. For a visual depiction of the data, we refer the reader to Figure 3-1. Unless required, we ignore the superscript denoting the index of the datapoint and denote concatenation with [].

Model: SSMs capture dependencies in sequential data via a time-varying latent state. When this latent state is discrete, SSMs are also known as Hidden Markov Models (HMM). In our setting, we deal with a continuous latent state. The generative process is:

$$\begin{aligned} p(\mathbf{X}|\mathbf{U}, B) = &\int_Z \prod_{t=1}^{T} p_\theta(Z_t|Z_{t-1}, U_{t-1}, B) p_\theta(X_t|Z_t) dZ \\ Z_t|\cdot \sim &\mathcal{N}(\mu_\theta(Z_{t-1}, U_{t-1}, B), \Sigma_\theta^t(Z_{t-1}, U_{t-1}, B)), \\ X_t|\cdot \sim &\mathcal{N}(\kappa_\theta(Z_t), \Sigma_\theta^e(Z_t)) \end{aligned} \tag{3.1}$$

We denote the parameters of a model by θ, which may comprise weight matrices or the parameters of functions that index θ. SSMs make the Markov assumption *on the latent variables*, Z_t, and we assume that relevant information about past medications are captured by the state or contained in U_{t-1}. We set $\Sigma_\theta^t, \Sigma_\theta^e, \kappa_\theta(Z_t)$ to be functions of a concatenation of their inputs, e.g. $\Sigma_\theta^t(\cdot) = \textbf{softplus}(\mathbf{W}[Z_{t-1}, U_{t-1}, B] + \mathbf{b})$. $\Sigma_\theta^t, \Sigma_\theta^e$ are diagonal matrices where the softplus function is used to ensure positivity. When $\mu_\theta(Z_{t-1}, U_{t-1}, B)$ is a nonlinear function, as is the case in our work, the model is an instance of a Deep Markov Model [148].

Learning: We maximize $\sum_{i=1}^{N} \log p(\mathbf{X}^i|\mathbf{U}^i, B^i)$ with respect to θ. For a nonlinear SSM, this function is intractable, so we learn via maximizing a variational lower bound on it. To evaluate the bound, we perform probabilistic inference using a structured inference network [148]. The learning algorithm alternates between predicting variational parameters using a bi-directional recurrent neural network, evaluating a variational upper bound, and making gradient updates jointly with respect to the parameters of the generative model and the inference network. When evaluating the likelihood of data under the model, if X_t is missing, it is marginalized out. Since the inference network also conditions on sequences of observed data to predict the variational parameters, we use forward fill imputation where data are missing.

3.4 Attentive Pharmacodynamic State Space Model

To make the shift from black-box models to those that capture useful structure for modeling clinical data, we begin with a discussion of PK-PD models and some of the key limitations that practitioners may face when directly applying them to modern clinical datasets.

3.4.1 Limitations of Pharmacokinetic-Pharmacodynamic Modeling

Pharmacology is a natural store of domain expertise for reasoning about how treatments affect disease. We look specifically at pharmacokinetics (PK), which deals with how drugs move in the body, and pharmacodynamics (PD), which studies the body's response to drugs. Consider a classical pharmacokinetic-pharmacodynamic (PK-PD) model used to characterize variation in tumor volume due to chemotherapy [189, 286]. Known as the log-cell kill model, it is based on the hypothesis that a given dose of chemotherapy results in killing a constant fraction of tumor cells rather than a constant number of cells. The original model is an ordinary differential equation but an equivalent expression is:

$$S(t) = S(t-1) \cdot (1 + \rho \log(K/S(t-1)) - \beta_c C(t)), \tag{3.2}$$

$S(t)$ is the (scalar) tumor volume, $C(t)$ is the (scalar) concentration of a chemotherapeutic drug over time, K is the maximum tumor volume possible, ρ is the growth rate, and β_c represents the drug effect on tumor size. Besides its bespoke nature, there are some key limitations of this model that hinder its broad applicability for unsupervised learning:

Single intervention, single biomarker: The model parameterizes the effect of a *single, scalar* intervention on a single, scalar, time-varying biomarker making it impossible to apply directly to high-dimensional clinical data. Furthermore, the quantity it models, tumor volume, is unobserved for non-solid cancers.

Misspecified in functional form: The log-cell-kill hypothesis, by itself, is not an accurate description of the drug mechanism in most non-cancerous chronic diseases.

Misspecified in time: Patients go through cycles of recovery and relapse during a disease. Even if the hypothesis holds when the patient is sick, it may not hold when the patient is in recovery.

In what follows, we aim to mitigate these limitations to build a practical, scalable model of disease progression.

3.4.2 Latent Representations of Disease State

Tackling the first limitation, we use nonlinear SSMs in order to model longitudinal, high-dimensional data. Even though tumor volume may not be observed in observational clinical datasets, various proxies (e.g. lab values, blood counts) of the unobserved disease burden often are. We conjecture that the time-varying latent representation, Z_t, implicitly captures such clinical phenotypes from the observations.

To ensure that the phenotypes captured by Z_t vary over time in a manner akin to clinical intuition, we focus the efforts of our design on the transition function, $\mu_\theta(Z_{t-1}, U_{t-1}, B)$, of the state space model. This function controls the way in which the latent state Z_t in an SSM evolves over time (and through it, the data) when exposed to interventions, U_t; this makes the transition function a good starting point for incorporating clinical domain knowledge.

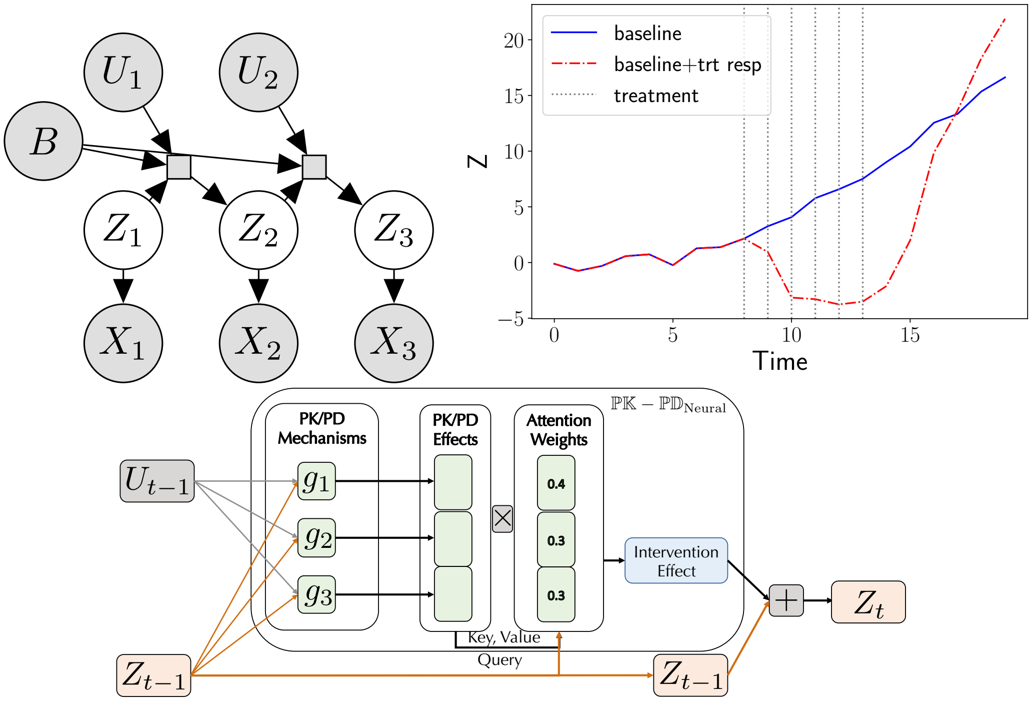


Figure 3-2: Unsupervised Models of Sequential Data (Left): We show a State Space Model (SSM) of $\mathbf{X}$ (the longitudinal biomarkers) conditioned on B (genetics, demographics) and $\mathbf{U}$ (binary indicators of treatment and line of therapy). The rectangle depicts $\mu_\theta(Z_{t-1}, U_{t-1}, B)$. **Neural Architecture for $\mathbb{PKPD}_{\textbf{Neural}}$ (Middle):** Illustration of the neural architecture we design; we use a soft-attention mechanism over the neural PK/PD effects using the current patient representation as a query to decide how the masks should be distributed. **Modeling relapse with the neural treatment exponential response (Right):** The curve depicts a single dimension of the representation and vertical lines denote a single treatment. After maintaining the response with treatments, a regression towards baseline (in blue; depicting what would have happened had no treatment been prescribed) occurs when treatment is stopped.

3.4.3 Neural Attention over Treatment Effect Mechanisms

In order to design a good transition function, we first need to address the second limitation that we may not know the *exact* mechanism by which drugs affect the disease state. However, we often have a set of reasonable hypotheses about the mechanisms that underlie how we expect the dynamics of the latent disease state to behave.

Putting aside the specifics of what mechanisms we should use for the moment, suppose we are given d mechanism functions, $g_1, \ldots, g_d$, each of which is a neural architecture that we believe captures aspects of how a representation should vary as a response to treatment. How a patient's representation should vary will depend on what state the patient is in. e.g. sicker patients may respond less well to treatment than healthier ones. To operationalize this insight, we make use of an attention mechanism [16] to attend to which choice of function is most appropriate.

Attending over mechanisms of effect Attention mechanisms operate by using a "query" to index into a set of "keys" to compute a set of attention weights, which are a distribution over the "values". We propose a soft-attention mechanism to select between $g_1, \ldots, g_d$. At each t, for the query, we have $q = Z_{t-1}W_q$. For the key and value, we have,

$$\tilde{K} = [g_1(Z_{t-1}, U_{t-1}, B); \ldots; g_d(Z_{t-1}, U_{t-1}, B)]^\top W_k$$
$$\tilde{V} = [g_1(Z_{t-1}, U_{t-1}, B); \ldots; g_d(Z_{t-1}, U_{t-1}, B)]^\top W_v.$$

Note that $W_q, W_k, W_v \in \mathbb{R}^{Q\times Q}$ and that $q \in \mathbb{R}^Q$, $\tilde{K} \in \mathbb{R}^{Q\times d}$, and $\tilde{V} \in \mathbb{R}^{Q\times d}$. Then, we have the following,

$$\mu_\theta(Z_{t-1}, U_{t-1}, B) = \left(\sum_{i=1}^{d} \mathbf{softmax}\left(\frac{q \odot \tilde{K}}{\sqrt{Q}}\right)_i \odot \tilde{V}_i\right) W_o \tag{3.3}$$

We compute the attention weights using the latent representation at a particular time point as a "query" and the output of each of $g_1, \ldots, g_d$ as "keys"; see Figure 3-2 (middle). This choice of neural architecture for μ_θ allows us to parameterize heterogeneous SSMs, where the function characterizing latent dynamics changes over time.

3.4.4 Lines of Therapy with Local and Global Clocks

Here, we address a third limitation of classical PK-PD models: a proposed drug mechanism's validity may depend on how long the patient has been treated and what stage of therapy they are in. Such stages, or *lines of therapy*, refer to contiguous plans of multiple treatments prescribed to a patient. They are often a unique structure of clinical data from individuals suffering from chronic diseases. For example, first line therapies often represent combinations prioritized due to their efficacy in clinical trials; subsequent lines may be decided by clinician preference. Lines of therapy index treatment plans that span multiple time-steps and are often laid out by clinicians at first diagnosis. We show how to make use of this information within a mechanism function.

To capture the clinician's intention when prescribing treatment, we incorporate line of therapy as a one-hot vector in $U_t[: K]$ $\forall t$ (K is the maximal line of therapy). Lines of therapy typically change when a drug combination fails or causes adverse side effects. By conditioning on line of therapy, a transition function (of the SSM) parameterized by a neural network can, in theory, infer the length of time a patient has been on that line. However, although architectures such as Neural Turing Machines [99] can learn to count occurrences, they would need a substantial amount of data to do so.

To enforce the specified drug mechanism functions to capture time since change in line of therapy, we use *clocks* to track the time elapsed since an event. This strategy has precedent in RNNs, where Che et al. [44] use time since the last observation to help RNNs learn well when data is missing. Koutnik et al. [147] partition the hidden states in RNNs so they are updated at different time-scales. Here, we augment our interventional vector, U_t, with two more dimensions. A global clock, gc, captures time elapsed since $T = 0$, i.e. $U_t[K] = \mathrm{gc}_t = t$. A local clock, lc, captures time elapsed since a line of therapy began; i.e. $U_t[K+1] = \mathrm{lc}_t = t - p_t$ where p_t denotes the index of time when the line last changed. By using the local clock, $\mu_\theta(Z_{t-1}, U_{t-1}, B)$ can modulate Z_t to capture patterns such as: the longer a line of therapy is deployed, the less or (more) effective it may be.

For the patient in Figure 3-1, we can see that the first dimension of $\mathbf{U}$ denoting line of therapy would be $[0, 0, 0, 0, 1, 1, 2, 2, 2]$. Line 0 was used four times, line 1 used twice, line 2 used thrice. Then, $p = [0, 0, 0, 0, 4, 4, 6, 6, 6, 6]$, $gc = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]$ and $lc = [0, 1, 2, 3, 0, 1, 0, 1, 2, 3]$. To the best of our knowledge, we are the first to make use of lines-of-therapy information and clocks concurrently to capture temporal information when modeling clinical data.

3.4.5 Neural PK-PD Functions for Chronic Diseases

Having developed solutions to tackle some of the limitations of PK-PD models, we turn to the design of three new mechanism functions, each of which captures different hypotheses a clinician may have about how the underlying disease burden of a patient changes (as manifested in their latent states).

Modeling baseline conditional variation: Biomarkers of chronic diseases can increase, decrease, or stay the same. Such patterns may be found in the dose-response to chemotherapy used in solid cancerous tumors [144]. In reality, clinicians find that these changes are often modulated by patient specific features such as age, genetic mutations, and history of illness. Patients who have been in therapy for a long time may find decreased sensitivity to treatments. To capture this variation:

$$g_1(Z_{t-1}, U_{t-1}, B) = Z_{t-1} \cdot \tanh(b_{\text{lin}} + W_{\text{lin}}[U_{t-1}, B]) \tag{3.4}$$

where $b_{\text{lin}} \in \mathbb{R}Q, W_{\text{lin}} \in \mathbb{R}^{Q\times(L+J)}$. Here, the effects on the representation are bounded (via the tanh function) but depend on the combination of drugs prescribed and the patient's baseline data, including genetics.

Modeling slow, gradual relapse after treatment: One of the defining features of many chronic diseases is the possibility of a relapse during active therapy. In cancer, a relapse can happen due to cancerous cells escaping the treatment or a variety of other bio-chemical processes, such as increased resistance to treatment due to mutations. The relapse can result in bio-markers reverting to values that they held prior to the start of treatment; for an example of this, see Figure 3-2 (right). We design the following neural architectures to capture such patterns in a latent representation.

Neural Log-Cell Kill: This architecture is inspired by the classical log cell kill model of tumor volume in solid cell tumors [286] but unlike the original model, scales to high-dimensional representations and takes into account *lines of therapy* via the local clock. This allows the model to effectively reset every time a new line of therapy begins. The functional form of the model is,

$$
\begin{aligned}
g_2(Z_{t-1}, U_{t-1}, B) = Z_{t-1} \cdot (1 - \rho \log(Z_{t-1}^2) \\
- \beta \exp(-\delta \cdot \mathrm{lc}_{t-1})),
\end{aligned} \tag{3.5}
$$

where $\beta = \tanh(W_{lc} U_{t-1} + b_{lc})$. $W_{lc} \in \mathbb{R}^{Q \times L}, b_{lc} \in \mathbb{R}^Q, \delta \in \mathbb{R}^Q$ and $\rho \in \mathbb{R}^Q$ are learned. While diseases may not have a single observation that characterizes the state of the organ system (akin to tumor volume), we hypothesize that representations, Z_t, of the observed clinical biomarkers may benefit from mimicking the dynamics exhibited by tumor volume when exposed to chemotherapeutic agents. We emphasize that unlike Equation 3.2, the function in Equation 3.5 operates over a *vector valued* set of representations that can be modulated by the patient's genetic markers.

Neural Treatment Exponential: Xu et al. [294] develop a Bayesian nonparameteric model to explain variation in creatinine, a single biomarker, due to treatment. We design an architecture inspired by their model that scales to high dimensional representations, allows for the representation to vary as a function of the patient's genetics, and makes use of information in the lines of therapy via the clocks.

$$
\begin{aligned}
& g_3(Z_{t-1}, U_{t-1}, B) \\
&= \begin{cases}
b_0 + \alpha_{1,t-1}/[1 + \exp(-\alpha_{2,t-1}(\mathrm{lc}_{t-1} - \frac{\gamma_l}{2}))], \\
\text{if } 0 \leq \mathrm{lc}_{t-1} < \gamma_l \\
b_l + \alpha_{0,t-1}/[1 + \exp(\alpha_{3,t-1}(\mathrm{lc}_{t-1} - \frac{3\gamma_l}{2}))], \\
\text{if } \mathrm{lc}_{t-1} \geq \gamma_l
\end{cases}
\end{aligned} \tag{3.6}
$$

Despite its complexity, the intermediate representations learned within this architecture have simple intuitive meanings. $\alpha_{1,t-1} = W_d[Z_{t-1}, U_{t-1}, B] + b_d$, where $W_d \in \mathbb{R}^{Q \times (Q+L+J)}, b_d \in \mathbb{R}^Q$ is used to control whether each dimension in Z_{t-1} increases or decreases as a function of the treatment and baseline data. $\alpha_{2,t-1}, \alpha_{3,t-1}$, and γ_l control the steepness and duration of the intervention effect. We restrict these characteristics to be similar for drugs administered under the same line of therapy.

Thus, we parameterize: $[\alpha_2, \alpha_3, \gamma_l]_{t-1} = \sigma(W_e \cdot U_{t-1}[0] + b_e)$. If there are three lines of therapy, $W_e \in \mathbb{R}^{3\times3}, b_e \in \mathbb{R}^3$ and the biases, $b_0 \in \mathbb{R}^Q$ and $b_l \in \mathbb{R}^Q$, are learned. Finally, $\alpha_{0,t-1} = (\alpha_{1,t-1} + 2b_0 - b_l)/(1 + \exp(-\alpha_{3,t-1}\gamma_l/2))$ ensures that the effect peaks at $t = \mathrm{lc}_t + \gamma_l$. Figure 3-2 (right) depicts how a single latent dimension may vary over time for a single line of therapy using this neural architecture.

From $\mathbb{PKPD}_{\textbf{Neural}}$ to the SSM$_{\textbf{PK-PD}}$: When g_1, g_2, g_3, as described in Equations 3.4, 3.5, 3.6, are used in the transition function μ_θ (as defined in Equation 3.3), we refer to the resulting function as $\mathbb{PKPD}_{\text{Neural}}$. Moreover, when $\mathbb{PKPD}_{\text{Neural}}$ is used as the transition function in an SSM, we refer to the resulting model as **SSM**$_{\text{PK-PD}}$, a heterogeneous state space model designed to model the progression of diseases.

3.5 Evaluation of SSM$_{\textbf{PK-PD}}$

3.5.1 Datasets

We study **SSM**$_{\text{PK-PD}}$ on three different datasets – two here, and on a third semi-synthetic dataset in the appendix.

Synthetic Data: We begin with a synthetic disease progression dataset where each patient is assigned baseline covariates $B \in \mathbb{R}^6$. B determines how the biomarkers, $X_t \in \mathbb{R}^2$, behave in the absence of treatment. $U_t \in \mathbb{R}^4$ comprises the line of therapy ($K = 2$), the local clock, and a single binary variable indicating when treatment is prescribed. To mimic the data dynamics described in Figure 3-1, the biomarkers follow second-order polynomial trajectories over time with the underlying treatment effect being determined by the Neural Treatment Exponential (see Equation 3.6). Biomarker 1 can be thought of as a marker of disease burden, while biomarker 2 can be thought of as a marker of biological function. The full generative process for the data is in the supplementary material. To understand generalization of the model as a function of sample complexity, we train on 100 samples and 1000 samples, respectively, and evaluate on five held-out sets of size 50000.

ML-MMRF: The Multiple Myeloma Research Foundation (MMRF) CoMMpass study releases de-identified clinical data for 1143 patients suffering from multiple myeloma, an incurable plasma cell cancer. All patients are aligned to the start of treatment, which is made according to current standard of care (not random assignment). With an oncologist, we curate demographic and genomic markers, $B \in \mathbb{R}^{16}$, clinical biomarkers, $X_t \in \mathbb{R}^{16}$, and interventions, $U_t \in \mathbb{R}^{9}$, with one local clock, a three dimensional one-hot encoding for line of therapy, and binary markers of 5 drugs. Our results are obtained using a 75/25 train/test split. To select hyperparameters, we perform 5-fold cross validation on the training set. Finally, there is missingness in the biomarkers, with 66% of the observations missing. We refer the reader to the appendix for more details on the dataset.

3.5.2 Setup

We learn via: $(\arg\min_{\theta} - \log p(\mathbf{X}|\mathbf{U}, B; \theta))$ using ADAM [142] with a learning rate of 0.001 for 15000 epochs. L1 or L2 regularization is applied in one of two ways: either we regularize all model parameters (including parameters of inference network), or we regularize all weight matrices except those associated with the attention mechanism. We search over regularization strengths of $0.01, 0.1, 1, 10$ and latent dimensions of $16, 48, 64$ and 128. We do model selection using the negative evidence lower bound (NELBO); Appendix B contains details on the derivation of this bound. We use a single copy of g_1, g_2, and g_3 in the transition function. Multiple copies of each function as well as other "mechanistic" functions can be used, highlighting the flexibility of our approach. However, this must be balanced with potentially overfitting on small datasets.

3.5.3 Baselines

SSM$_{\text{Linear}}$ parametrizes $\mu_\theta(Z_{t-1}, U_{t-1}, B)$ with a linear function. This model is a strong, linear baseline whose variants have been used for modeling data of patients suffering from Chronic Kidney Disease [207].

SSM$_{\text{NL}}$: Krishnan et al. [148] use a nonlinear SSM to capture variation in the clinical biomarkers of diabetic patients. We compare to their model, parameterizing the transition function with a 2-layer MLP.

SSM$_{\text{MOE}}$: We use an SSM whose transition function is parameterized via a Mixture-of-Experts (MoE) architecture [130, 135]; i.e. g_1, g_2, g_3 are each replaced with a multi-layer perceptron. This baseline does not incorporate any domain knowledge and tests the relative benefits of prescribing the functional forms via mechanisms versus learning them from data.

SSM$_{\text{Attn.Hist.}}$: We implement a variant of the SSM in Alaa and van der Schaar [3], a state-of-the-art model for disease progression trained via conditional density estimation. The authors use a *discrete* state space for disease progression modeling making a direct comparison difficult. However, **SSM**$_{\text{Attn.Hist.}}$ preserves the structural modeling assumptions they make. Namely, the transition function of the model attends to a concatenation of previous states and interventions at each point in time. We defer specifics to Appendix A.

In addition, we run two simpler baselines, a First Order Markov Model (FOMM) and Gated Recurrent Unit (GRU) [54], on the synthetic data and ML-MMRF but defer those results to Appendix A.

3.5.4 Evaluation Metrics

NELBO On both the synthetic data and ML-MMRF data, we quantify generalization via the negative evidence lower bound (NELBO), which is a variational upper bound on the negative log-likelihood of the data. A lower NELBO indicates better generalization.

Pairwise Comparisons For a fine-grain evaluation of our models on ML-MMRF, we compare held-out NELBO under **SSM**$_{\text{PK-PD}}$ versus the corresponding baseline for each patient. For each held-out point, $\Delta_i = 1$ when the NELBO of that datapoint is lower under **SSM**$_{\text{PK-PD}}$ and $\Delta_i = 0$ when it is not. In Table 3.1 (bottom), we report $\frac{1}{N}\sum_{i=1}^{N}\Delta_i$, the proportion of data for which **SSM**$_{\text{PK-PD}}$ yields better results.

Counts To get a sense for the number of patients on whom **SSM**$_{\text{PK-PD}}$ does much better, we count the number of held-out patients for whom the held-out negative log likelihood (computed via importance sampling) is more than 10 nats lower under

Training Set Size	SSM Linear	SSM NL	SSM MOE	SSM Attn. Hist.	SSM PK-PD	SSM PK-PD (w/o TExp)
100	58.57 +/- .06	69.04 +/- .11	60.98 +/- .04	76.94 +/- .02	**55.34 +/- .03**	**58.39 +/- .05**
1000	53.84 +/- .02	44.75 +/- .02	51.57 +/- .03	73.80 +/- .03	**39.84 +/- .02**	**38.93 +/- .01**

Evaluation Metric	SSM PK-PD vs. SSM Linear	SSM PK-PD vs. SSM NL	SSM PK-PD vs. SSM MOE	SSM PK-PD vs. SSM Attn. Hist.
Pairwise Comparison (↑)	0.796 (0.400)	0.760 (0.426)	0.714 (0.450)	0.934 (0.247)
Counts (↑)	PK-PD: 158, LIN: 6	130, 12	94, 8	272, 0
NELBO (↓)	PK-PD: 61.54, LIN: 74.22	61.54, 79.10	61.54, 73.44	61.54, 105.04
# of Model Parameters	PK-PD: 23K, LIN: 7K	23K, 51K	23K, 77K	23K, 17K

Table 3.1: *Top:* **Synthetic data:** Lower is better. We report the test NELBO with std. dev. for each SSM model to study generalization in the synthetic setting. *Bottom:* **ML-MMRF:** *Pairwise Comparison*: each number is the fraction (with std. dev.) of test patients for which $\mathbf{SSM}_{\text{PK-PD}}$ has a lower NELBO than an SSM that uses a different transition function. *Counts*: We report the number of test patients (out of 282) for which an SSM model (PK-PD or otherwise) has a greater than 10 nats difference in negative log likelihood compared to the alternative model. *NELBO*: We report the test NELBO of each model. Note that we label the metrics associated with $\mathbf{SSM}_{\text{PK-PD}}$ and $\mathbf{SSM}_{\text{Linear}}$ in the first column, but drop these labels in subsequent columns.

$\mathbf{SSM}_{\text{PK-PD}}$ than the corresponding baseline (and vice versa for the baselines).

3.5.5 Results

We investigate three broad categories of questions.

Generalization under different conditions

SSM$_{\textit{PK-PD}}$ *generalizes better in setting with few (∼ 100) samples.* Table 3.1 (top) depicts NELBOs on held-out synthetic data across different models, where a lower number implies better generalization. We see a statistically significant gain in generalization for $\mathbf{SSM}_{\text{PK-PD}}$ compared to all other baselines. $\mathbf{SSM}_{\text{NL}}$, $\mathbf{SSM}_{\text{MOE}}$ overfit quickly on 100 samples but recover their performance when learning with 1000 samples.

SSM$_{\textit{PK-PD}}$ *generalizes well when it is misspecified.* Because we often lack prior knowledge about the true underlying dynamics in the data, we study how $\mathbf{SSM}_{\text{PK-PD}}$ performs when it is misspecified. We replace the Neural Treatment Exponential function, g_3, from $\mathbb{PKPD}_{\text{Neural}}$ with another instance of g_1. The resulting model is now misspecified since g_3 is used to generate the data but no longer lies within the model family. We denote this model as (SSM PK-PD w/o TExp). In Table 3.1 (top), when

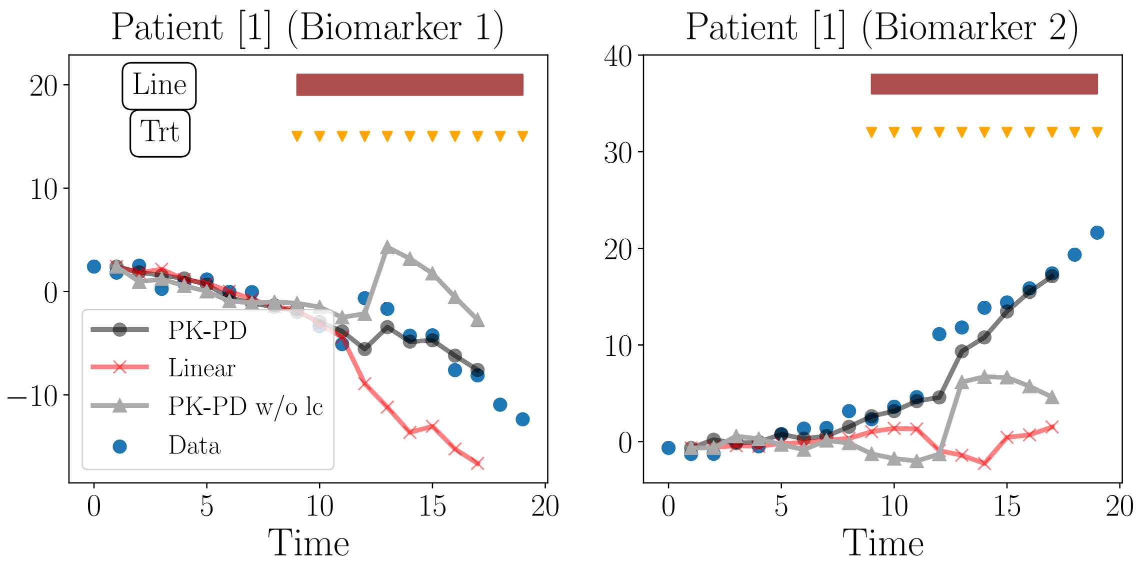


Figure 3-3: *Synthetic*: Forward samples (conditioned only on B) from **SSM**$_{\text{PK-PD}}$ (o), **SSM**$_{\text{Linear}}$ (x), **SSM**$_{\text{PK-PD}}$ without local clocks (△), for a single patient. Y-axis shows biomarker values.

comparing the sixth column to the others, we find that we outperform all baselines and get comparable generalization to **SSM**$_{\text{PK-PD}}$ with the Neural Treatment Exponential function. This result emphasizes our architecture's flexibility and its ability to learn the underlying (unknown) intervention effect through a combination of other, related mechanism functions.

SSM$_{\textit{PK-PD}}$ *generalizes well on real-world patient data.* A substantially harder test of model misspecification is on the ML-MMRF data where we have unknown dynamics that drive the high-dimensional (often missing) biomarkers in addition to combinations of drugs prescribed over time. To rigorously validate whether we improve generalization on ML-MMRF data with **SSM**$_{\text{PK-PD}}$, we study model performance with respect to the three metrics introduced in Section 3.5.4. We report our results in Table 3.1 (bottom). First, we consistently observe that a high fraction of patient data in the test set are explained better by **SSM**$_{\text{PK-PD}}$ than the corresponding baseline (pairwise comparisons). We also note that out of 282 patients in the test set, across all the baselines, we find that the **SSM**$_{\text{PK-PD}}$ generalizes better for many more patients (counts). Finally, **SSM**$_{\text{PK-PD}}$ has lower NELBO averaged across the entire test set compared to all baselines.

Model complexity & generalization

The improvements of $\boldsymbol{SSM}_{PK\text{-}PD}$ are consistent taking model sizes into account. We show in Table 3.1 (bottom) the number of parameters used in each model. We find that more parameters do not imply better performance. Models with the most parameters (e.g. $\mathbf{SSM}_{\mathrm{NL}}$) overfit while those with the lowest number of parameters underfit (e.g. $\mathbf{SSM}_{\mathrm{Linear}}$) suggesting that the gains in generalization that we observe are coming from our parameterization. We experimented with increasing the size of the $\mathbf{SSM}_{\mathrm{Linear}}$ model (via the latent variable dimension) to match the size of the best PK-PD model. We found that doing so *did not outperform* the held-out likelihood of $\mathbf{SSM}_{\mathrm{PK\text{-}PD}}$.

When data are scarce, a Mixture of Experts architecture is difficult to learn: How effective are the functional forms of the neural architectures we develop? To answer this question, we compare the held-out log-likelihood of $\mathbf{SSM}_{\mathrm{PK\text{-}PD}}$ vs $\mathbf{SSM}_{\mathrm{MOE}}$ in the third column of Table 3.1 (bottom). In the ML-MMRF data, we find that the $\mathbf{SSM}_{\mathrm{PK\text{-}PD}}$ outperforms the $\mathbf{SSM}_{\mathrm{MOE}}$. We suspect this is due to the fact that learning diverse "experts" is hard when data is scarce and supports the hypothesis that the judicious choice of neural architectures plays a vital role in capturing biomarker dynamics.

Can $\mathbb{PKPD}_{Neural}$ be used in other model families? In the supplement, we implement $\mathbb{PKPD}_{\mathrm{Neural}}$ in a first-order Markov model and find similar improvements in generalization on the ML-MMRF dataset. This result suggests that the principle we propose of leveraging domain knowledge from pharmacology to design mechanism functions can allow other kinds of deep generative models (beyond SSMs) to also generalize better when data are scarce.

Visualizing Patient Dynamics

In Figure 3-4 (right), to further validate our initial hypothesis that the model is using the various neural PK-PD effect functions, we visualize the attention weights from $\mathbf{SSM}_{\mathrm{PK\text{-}PD}}$ trained on ML-MMRF averaged across time and all patients. The highest weighted component is the treatment exponential model g_3, followed by the bounded

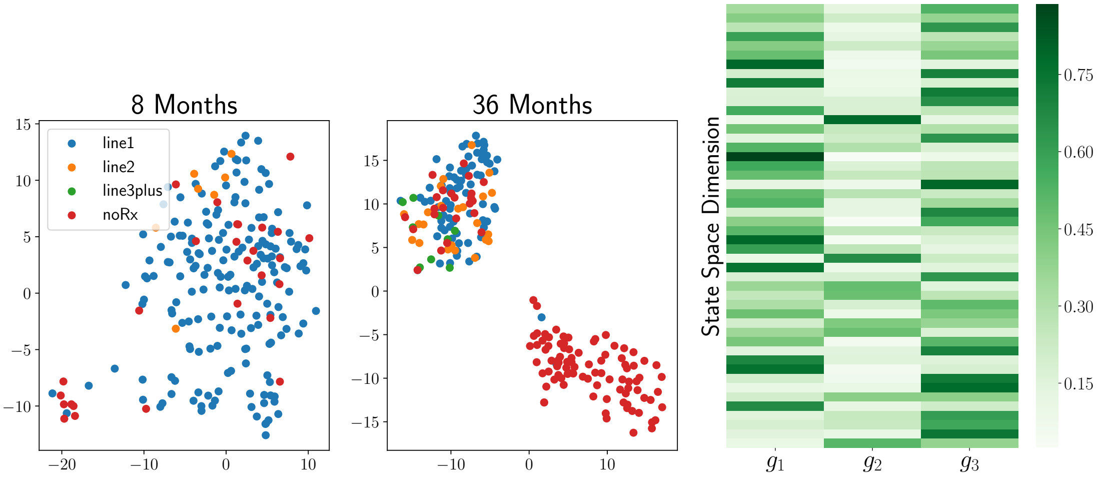


Figure 3-4: *ML-MMRF*: (Left two) We visualize the TSNE representations of each test patient's latent state, Z_t, at the start of treatment and three years in. (Right) For $\textbf{SSM}_{\text{PK-PD}}$, we visualize the attention weights, averaged over all time steps and all patients, on each of the neural effect functions across state space dimensions.

linear model g_1 for many of the latent state dimensions. We also see that several of the latent state dimensions make exclusive use of the neural log-cell kill model g_2.

How do the clocks help model patient dynamics? Figure 3-3 shows samples from three SSMs trained on synthetic data. $\textbf{SSM}_{\text{PK-PD}}$ captures treatment response accurately while $\textbf{SSM}_{\text{Linear}}$ does not register that the effect of treatment can persist over time. To study the impact of clocks on the learned model, we perform an ablation study on SSMs where the local clock in U_t, used by $\mathbb{PKPD}_{\text{Neural}}$, is set to a constant. Without clocks (PK-PD w/o lc), the model does not capture the onset or persistence of treatment response.

***SSM**$_{\textit{PK-PD}}$ learns latent representations that reflect the patient's disease state:* In ML-MMRF, we restrict the patient population to those with at least $T = 36$ months of data. At two different points during their treatment of the disease, we visualize the result of TSNE [176] applied to their latent representations in Figure 3-4 (left).

Early in their treatment, the latent representations of these patients appear to have no apparent structure. As time progresses, we find that the dimensions split into two groups. One group, for the most part, is still being treated, while the other is not being treated. A deeper dive into the untreated patients reveals that this cohort has a less severe subtype of myeloma (via a common risk assessment method known as

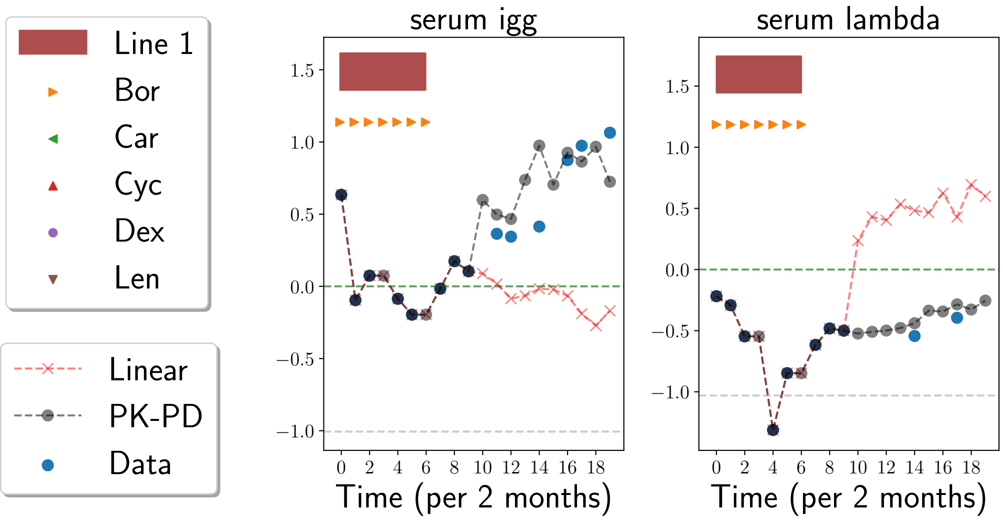


Figure 3-5: *ML-MMRF*: Each column is a different biomarker containing forward samples (conditioned on approximately the first 2 years of a patient's data) from $\mathbf{SSM}_{\text{PK-PD}}$ (o) and $\mathbf{SSM}_{\text{linear}}$ (x) of a single test patient. Blue circles denote ground truth, and the markers above the trajectories represent treatments prescribed across time. Y-axis shows biomarker levels, with the dotted green[gray] line representing the maximum[minimum] healthy value. Car, Cyc, Dex, and Len shown in legend to maintain consistency with plots in Appendix, but are not given in the treatment regimen.

ISS staging). This result suggests that the latent state of $\mathbf{SSM}_{\text{PK-PD}}$ has successfully captured the coarse disease severity of patients at particular time points.

Visualizing patient samples from $\mathbf{SSM}_{\text{PK-PD}}$: Figure 3-5 shows the average of three samples from $\mathbf{SSM}_{\text{Linear}}$ and $\mathbf{SSM}_{\text{PK-PD}}$ trained on ML-MMRF. We track two biomarkers used by clinicians to map myeloma progression. $\mathbf{SSM}_{\text{PK-PD}}$ better captures the evolution of these biomarkers conditioned on treatment. For serum IgG, $\mathbf{SSM}_{\text{PK-PD}}$ correctly predicts the relapse of disease after stopping first line therapy, while $\mathbf{SSM}_{\text{Linear}}$ does not. On the other hand, for serum lambda, $\mathbf{SSM}_{\text{PK-PD}}$ correctly predicts it will remain steady.

3.6 Discussion

$\mathbb{PKPD}_{\text{Neural}}$ leverages domain knowledge from pharmacology in the form of treatment effect mechanisms to quantitatively and qualitatively improve performance of a representation-learning based disease progression model. Bica et al. [31] note the potential for blending ideas from pharmacology with machine learning: our work is among the first to do so.

We highlight several avenues for future work. For example, in machine learning for healthcare, the model we develop can find use in practical problems such as forecasting high-dimensional clinical biomarkers and learning useful representations of patients that are predictive of outcomes. Doing so can aid in the development of tools for risk stratification or in software to aid in clinical trial design. Applying the model to data from other chronic diseases such as diabetes, congestive heart failure, small cell lung cancer and rheumatoid arthritis, brings opportunities to augment $\mathbb{PKPD}_{\text{Neural}}$ with new neural PK-PD functions tailored to capture the unique biomaker dynamics exhibited by patients suffering from these diseases.

Finally, we believe $\mathbb{PKPD}_{\text{Neural}}$ can find use in the design of parameteric environment simulators for different domains. In pharmacology, such simulation based pipelines can help determine effective drug doses [123]. Our idea of attending over multiple dynamics functions can find use in the design of simulators in domains such as economics, where multiple hypothesized mechanisms are used to explain observed market phenomena [96].

In this chapter, I focused on using the structure of clinical temporal data to improve forecasting of biomarkers via state space models, a type of sequential model. In Chapter 4, I will shift gears to developing attention-based architectures, as opposed to sequential architectures, that can jointly predict clinical outcomes and forecast relevant biomarkers.

Chapter 4

Case Study – Investigating Performance of a Transformer-based Architecture trained on the TOURMALINE trial

Acknowledgement of Co-authors I would like to acknowledge Edward De Brouwer, who was integral in the software development for this chapter. Our daily brainstorming sessions were also instrumental in bringing the ideas in this work to maturity.

In this chapter, I continue building out the methods that can serve as the foundation for a CDSS in oncology, as shown in the Introduction of this thesis in Figure 1-2. In particular, I move towards a more general modeling architecture that can both predict important clinical outcomes, e.g. progression free survival and adverse events, and forecast patient labs. I explore a different approach to modeling compared to the last chapter, where I now develop an attention-based architecture and apply it on a clinical trial dataset of multiple myeloma patients.

4.1 Introduction

Multiple myeloma, the second most common blood cancer, has a global incidence of approximately 120,000 cases per year [174]. Clinical management of multiple myeloma patients is complex, both in terms of the factors that impact physician decision-making as well as the overall goals of care. Several factors, including the stage of the disease, transplant status, functional status, and genetic profile, are considered in the initial treatment decision [252, 70]. Furthermore, effective care aims to maximize patient survival and time to disease progression while minimizing adverse events from therapy. When deciding on a treatment, striking an appropriate balance between optimizing patient survival while maintaining quality of life by limiting adverse events is intricate and requires frequent follow-up with patients. Indeed, the International Myeloma Working Group (IMWG) recommends monthly (or bimonthly, depending on the therapy given as part of follow-up or maintenance regimen) monitoring of patients receiving initial chemotherapy for response to treatment, disease complications, and toxic sequelae of therapy [174].

In recent years, the availability of structured clinical data has led to the development of data-driven predictive models aimed at facilitating clinical diagnosis, prognostication, and decision-making. Examples include risk-score development for prostate cancer [21], improved detection of chronic kidney disease from retinal imaging [233], and survival outcome prediction in colorectal cancer [249]. However, previous works in multiple myeloma and other cancers have focused only on individual predictive or prognostic tasks, such as overall survival (OS) [105, 149], progression free survival (PFS) [14], and adverse event (AE) [87] prediction, using genomic, demographic, and lab covariates from structured clinical data as inputs. Despite existing literature on forecasting biomarkers in cancer and other chronic diseases [3, 294], no data-driven models for biomarker forecasting have been developed for multiple myeloma [6]. More generally, to our knowledge, there is a lack of existing literature on data-driven models that provide a comprehensive view of a patient's current and future disease course by jointly inferring a patient's risk of progression, death, and adverse events, as well as clinical trajectory reflected by labs and biomarkers. We argue that such a model is crucial for building decision support tools that effectively assist physicians in balancing the multiple facets of oncologic care when treating a patient.

To address this gap, we propose a deep-learning approach that jointly models the patient's current and future disease course by (1) predicting survival outcomes such as progression-free survival PFS, OS, and AEs, (2) forecasting key disease biomarkers, and (3) estimating the effect of using different treatment strategies. Importantly, our model dynamically adapts to newly collected clinical data and performs the above tasks with observed clinical history of arbitrary lengths. We call our model Transformer-CPH, which uses a transformer as the backbone, coupled with Cox proportional hazards (CPH) prediction heads. Transformer architectures [271] are attention-based models that have emerged as a powerful approach for capturing long-range dependencies and modeling temporal relationships in sequential data. With their ability to dynamically weigh the importance of different elements in a sequence, attention-based models excel at capturing complex patterns and dependencies in longitudinal data. The CPH prediction head additionally affords us with a straightforward parametrization for predicting survival outcomes in the presence of censored data. Transformer-CPH leverages all of these strengths.

To train and evaluate our model, we use data from two randomized controlled TOURMALINE trials, MM1 and MM2. The randomized controlled trials (RCTs) investigated the effect of an ixazomib, lenalidomide, and dexamethasone (IRd) combination regimen on survival outcomes, compared to lenalidomide and dexamethasone (Rd). MM2, comprising newly diagnosed multiple myeloma patients (NDMM), was used for training and internal validation, while MM1, comprising relapsed and refractory multiple myeloma patients (RRMM), was used for external validation in a setting where the patient population was at a different stage in their disease process and was enrolled according to different inclusion criteria. Notably, the randomization of treatment assignment in both datasets provides a unique opportunity to reliably infer treatment effects from our model, which can be used for downstream personalized treatment recommendations.

In addition to assessing predictive performance, we perform introspection of our model and demonstrate how our joint modeling approach automatically segments patients by myeloma subtype and reveals relationships between predicted biomarker trajectories and the risk of progression. Leveraging the individualized treatment effect prediction capabilities of Transformer-CPH, we show that our method can be used as a tool to retrospectively identify patient subgroups that would benefit most from a particular treatment regimen. As a proof of concept, we uncover a subgroup from the MM2 dataset, characterized by an IgA-dominant myeloma subtype and high involved free light chain level, that appears to result in a statistically significant lower hazard ratio for IRd compared to Rd with respect to PFS.

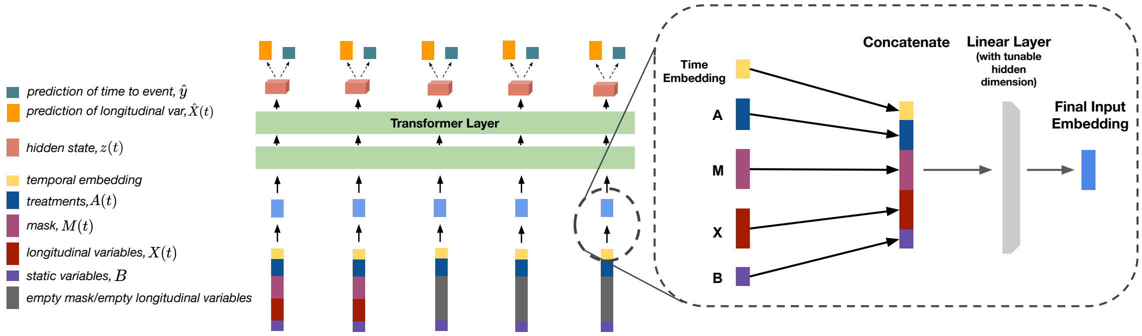

prediction of time to event, $\hat{y}$
prediction of longitudinal var, $\hat{X}(t)$
hidden state, $z(t)$
temporal embedding
treatments, $A(t)$
mask, $M(t)$
longitudinal variables, $X(t)$
static variables, B
empty mask/empty longitudinal variables
Transformer Layer
Time Embedding
A
M
X
B
Concatenate
Linear Layer
(with tunable hidden dimension)
Final Input Embedding

Figure 4-1: Diagram of Transformer-CPH architecture, indicating input variables, transformer layers, the number of which can be tuned, as well as the prediction heads. Additional architectural details can be found in the Methods section.

4.2 Methods

We begin by outlining the baseline methods used in this chapter for event prediction and forecasting followed by more details on the Transformer-CPH architecture. We also include details on the analyses that we discuss later on. For information on data preprocessing, including data normalization, imputation, sample splitting, and covariates used as input, please see Appendix B.

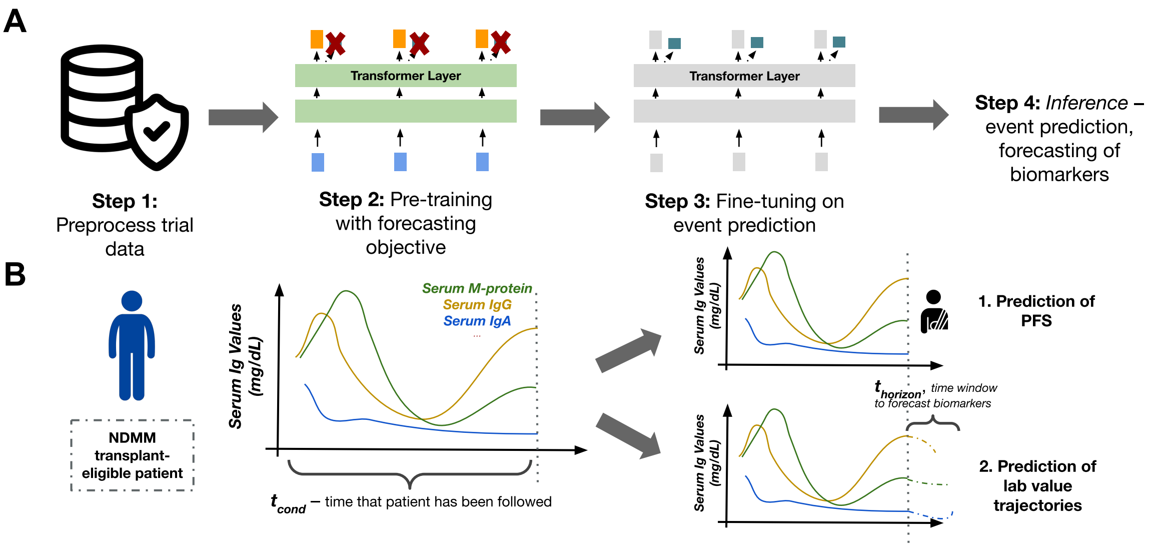

A
Transformer Layer
Transformer Layer
Step 1: Preprocess trial data
Step 2: Pre-training with forecasting objective
Step 3: Fine-tuning on event prediction
Step 4: Inference – event prediction, forecasting of biomarkers
B
NDMM transplant-eligible patient
Serum Ig Values (mg/dL)
Serum M-protein
Serum IgG
Serum IgA
...
t_cond – time that patient has been followed
Serum Ig Values (mg/dL)
1. Prediction of PFS
t_horizon, time window to forecast biomarkers
Serum Ig Values (mg/dL)
2. Prediction of lab value trajectories

Figure 4-2: (A): The training workflow consists of two steps: pre-training and fine-tuning. We pre-train the transformer model on the forecasting objective, and then fine-tune on the event prediction task, keeping the rest of the model parameters frozen. **(B)**: A schematic of a canonical scenario that our model might see during inference. Namely, a patient may have data that is observed for some length of time, the observation window t_{cond}, at which point we are interested in predicting both the survival outcome of the patient as well as how their biomarkers will evolve. We denote the length of the time window on over which we forecast the biomarkers, the forecasting window, as t_{horizon}. We evaluate our model architecture over various combinations of t_{cond} and t_{horizon}.

4.2.1 Baseline Approaches for Event Prediction and Forecasting

Event Prediction Baselines

We use two standard survival models, a Cox Proportional Hazards (CPH) model and a Random Survival Forest (RSF) as baseline approaches for survival prediction (PFS, OS) as well as adverse event prediction. These are standard survival models that have been widely used in a wide variety of risk prediction tasks across several chronic diseases [132, 182, 21, 105]. Additionally, we use a CPH model where the only covariate we add to the model is the ISS stage of a patient, a common risk score used in multiple myeloma that is computed from a patient's serum beta-2 microglobulin and serum albumin values and ranges from one to three [194]. This model is denoted as CPH-ISS. CPH-ISS can be seen as the "standard of care" approach used by oncologists to stratify myeloma patients according to their risk.

A different baseline model is trained for each observation window length. For an observation window of t time steps, we concatenate the observations at the last time step ($X(t)$), the observations at the first time step ($X(0)$), and the baseline variables (B), to form the input vector of these methods. Feeding both $X(t)$ and $X(0)$ to the model provides information about the trajectory of the patient.

Forecasting Baselines

We compare the performance of our method against two state-of-the-art time series forecasting methods (recurrent neural networks and deep markov models) and a standard imputation method (last observation carried forward). Except for the last observation carried forward method, the models use the same inputs and are trained on the same reconstruction task as Transformer-CPH. However, these models do not predict the risk of future events.

Recurrent Neural Networks Recurrent neural networks (RNNs) are a widely used class of neural networks for time series processing [54]. They operate by maintaining an internal memory state that is updated as the network processes each input in the sequence. The memory state is passed from one step of the sequence to the next, allowing the network to "remember" previous inputs and use them to inform the processing of future inputs. At each step of the sequence, the RNN takes in an input and combines it with the current memory state to generate a new memory state and an output. Unlike our approach, which relies on attention based models, RNN process the time series *sequentially.* This restricts their ability to efficiently capture long-term dependencies.

Deep Markov Models Deep Markov Models (DMMs) are a class of probabilistic models dedicated to time series modeling [148]. DMMs take their name from the Markov assumption, which states that the current state of a system depends only on its previous state, and not on the entire history. DMMs extend this idea to deep architectures by adding multiple hidden layers, allowing them to capture complex dependencies between inputs. Unlike RNNs, which process sequences in a deterministic manner, DMMs are stochastic models that can generate multiple possible outputs for a given input. This makes them well-suited for modeling noisy time series data. Akin to RNNs, DMMs process the temporal observation sequentially, which limits their ability to capture long-term dependencies.

Last Observation Carried Forward Last Observation Carried Forward (LOCF) is a simple imputation method used in time series forecasting. LOCF replaces missing values in a time series with the most recent available observation. LOCF is a straightforward and computationally inexpensive approach, making it a popular choice in many applications. However, LOCF does not consider the temporal dependencies between observations and may not capture the underlying patterns and dynamics in the time series data. In contrast, RNNs, DMMs, or attention-based models like Transformer-CPH are designed to work with sequential data and can capture complex temporal dependencies.

4.2.2 Transformer-CPH Architecture and Optimization Details

The transformer architecture in our experiments is built around a transformer encoder with continuous temporal embeddings. For a graphical depiction of the architecture, see Figure 4-1. Notationally, at some time point t, $X(t)$ denotes the longitudinal covariates (e.g., lab values, serum immunoglobulins, etc.), $M(t)$ denotes the mask of observed longitudinal covariates, where $M(t) = 1$ indicates the value is observed and $M(t) = 0$ indicates the value is missing, and $A(t)$ denotes the treatment administered as well as dosages of each drug in the combination regimen. B denotes baseline covariates (e.g., age, sex, ISS stage, etc.). We assume B does not change with time. The dimensions of each vector are as follows: $X(t) \in \mathbb{R}^{N \times D_{\text{lab}}}$, $M(t) \in \mathbb{R}^{N \times D_{\text{lab}}}$, $A(t) \in \mathbb{R}^{N \times D_{\text{treat}}}$, and $B \in \mathbb{R}^{N \times D_{\text{base}}}$, where N is the number of patients in the dataset, D_{lab} is the number of lab covariates, D_{treat} is the number of treatment covariates (e.g., dosage of each drug), and D_{base} is the number of baseline covariates. We let $T_{\max}$ be the maximum follow-up time for any patient, such that $t \leq T_{\max}$.

Input Sequence

Let $E(t)$ be a τ-dimensional time embedding of the following form: $E_{2k}(t) := \sin\left(\frac{t}{T_{\max}^{2k/\tau}}\right)$, $E_{2k+1}(t) := \cos\left(\frac{t}{T_{\max}^{(2k+1)/\tau}}\right)$, where $0 \leq k \leq \tau/2 - 1$ indexes the $E(t)$ vector. Then, the input embedding to the Transformer-CPH is simply a concatenation of $[B, X(t), M(t), A(t), E(t)]$ taken through a linear layer. Both the time embedding dimension, τ, and the input embedding dimension are tuned on the validation set. See Table 4.1 for a full set of hyperparameters and what values we search over for each.

Name	Description	Values
Hidden dimension (D_z)	Dimension of the hidden vectors in the transformer layers.	$[16, 32]$
Dropout	Probability of dropout in the transformer layers.	$[0.1, 0.2]$
Rollout-window (K)	Number of future hidden used as inputs to the event prediction modules.	$[0, 1]$
Number of layers	Number of transformer layers in the architecture.	2
Non-linearity	Whether to use a linear or non-linear prediction function for the event prediction.	[True,False]

Table 4.1: Hyperparameters used for training Transformer-CPH

Temporal Self-attention

We refer to $z_t^{(0)} = emb([B, X(t), M(t), A(t), E(t)])$ as the input embedding of our model for the observations at time t. The transformer encoder layer performs the following computation. For a temporal embedding at layer l, $z_t^{(l)}$, we compute query (q), key (k), and value (v) vectors as follows,

$$q_t^{(l)} = W_q z_t^{(l)} \tag{4.1}$$

$$k_t^{(l)} = W_k z_t^{(l)} \tag{4.2}$$

$$v_t^{(l)} = W_v z_t^{(l)} \tag{4.3}$$

These vectors are then concatenated into their respective matrices $Q^{(l)}$, $K^{(l)}$, and $V^{(l)}$. The output embedding for time t, $z_t^{(l+1)}$, is given by

$$z_t^{(l+1)} = f_l\Bigg(\text{softmax}\Bigg(\frac{q_t^{(l)}(K^{(l)})^T}{\sqrt{d_k}}\Bigg)V^{(l)}\Bigg), \tag{4.4}$$

where $f_l(\cdot)$ is a multi-layer perceptron (MLP), and d_K is the dimension of the input embedding vector. In the next few sections, we formalize how to do forecasting and event prediction via separate "prediction heads".

Forecasting Heads

We perform forecasting based on the sequence of embeddings after the last transformer layer, $z_t^{(L)}$. In our experiments, we set $L = 2$. We use D_z to refer to the dimension of the hidden vectors. We predict the future values of biomarkers by using a dedicated MLP ($f_{pred}(\cdot)$) with the last embeddings as input:

$$\widehat{X}_{t+1} = f_{\text{pred}}(z_t^{(L)}) \tag{4.5}$$

Long-Range Forecasting

The forecasting prediction heads predict the next values of biomarkers conditioned on the clinical trajectory until time t. To predict over horizons longer than a single step ahead, we re-use the prediction at the previous time step in the next input vector. Writing $z_t^{(L)}(B, X(t), A(t), E(t))$ as the output embedding for time t, where we make the dependence on the inputs explicit, we compute the next embedding as follows,

$$z_{t+1}^{(L)} = z_{t+1}^{(L)}(B, \widehat{X}_{t+1}, A(t), E(t)) \tag{4.6}$$

$$\widehat{X}_{t+2} = f_{\text{pred}}(z_{t+1}^{(L)}) \tag{4.7}$$

where $\hat{X}_{t+1}$ is obtained with Eq. 4.5. This procedure can be repeated to produce forecasts over arbitrary prediction horizons. The forecasts therefore only depend on X until time t. We still use the future treatment assignments, which allow one to generate the trajectories conditioned on a prospective treatment strategy.

Event-prediction Heads

In contrast to state space models such as recurrent neural networks, attention-based models like Transformer-CPH do not readily provide a hidden state vector that acts as a sufficient representation for the entire trajectory. Rather, the hidden vector produced at each time step is trained to predict the values of the clinical variables at the next time step. To circumvent this limitation, we condition our event predictions on several predicted hidden states.

We predict the score for a particular event, conditioned on history up to time t as

$$\hat{y}_t = f_{\text{event}}([z_t^{(L)}, z_{t+1}^{(L)}, ..., z_{t+K}^{(L)}]), \tag{4.8}$$

where K refers to the number of future hidden variables we are conditioning on in the event predictions. We found that using $K = 1$ leads to the best results in our experiments. Importantly, the prediction at time t does not use future values of the time series, since $z_{t+k}^{(L)}$ only uses information up until t, as defined in Eq. 4.6. We use a different head for each type of event: disease progression, overall survival, and adverse events, corresponding to different functions $f_{\text{event}} \in \{f_{\text{PFS}} : \mathbb{R}^{D_z} \to \mathbb{R}, f_{\text{OS}} : \mathbb{R}^{D_z} \to \mathbb{R}, f_{\text{AE}} : \mathbb{R}^{D_z} \to \mathbb{R}^{12}\}$. These functions are all parametrized by neural networks.

Model Training

The training of Transformer-CPH consists of two steps – the first is pre-training the entire architecture on the forecasting task, whose objective function we specify below. The second step consists of fine-tuning the model on only the event prediction tasks, which practically entails training the final non-linear head that does the prediction of the event and freezing the remaining weights. We list the loss function for each step.

For the forecasting task, we train the model autoregressively, minimizing the discrepancy between the predicted values $\widehat{X}_{t+1}$ and true values X_{t+1}. We use the observation mask M to only include the observed values in the loss function. Formally, the loss function for the forecasting objective is,

$$\mathcal{L}_{\text{forecast}} = \frac{1}{N}\sum_i \frac{\sum_t M_i(t) \odot (X_i(t) - \widehat{X}_i(t))^2}{\sum_t \sum M_i(t)}.$$

For the fine-tuning step, we train a linear head (or alternatively an MLP) for event prediction while freezing the remaining weights of the model. We use a Cox proportional hazards loss of the following form for training,

$$\mathcal{L}_{\text{event}} = \sum_{t=1}^{T_{\max}} \frac{1}{\sum_i \mathbf{1}(y_i \leq t)\mathbf{1}(\Delta_i = 1)} \sum_{i:\Delta_i=1} \mathbf{1}(y_i \leq t)\left(\hat{y}_i - \log \sum_{j \in \mathcal{R}(y_i)} \exp(\hat{y}_j)\right), \quad (4.9)$$

where Δ_i is the event indicator variable for patient i. $\Delta_i = 1$ indicates that the event was observed for patient i and $\Delta_i = 0$ indicates that the patient was censored. $\hat{y}_{i,t}$ is the predicted hazard rate outputted by the model for patient i at time t, and $\mathcal{R}$ is the set of patients still at risk for failure at time t. The event loss function can be defined for any event type: $\mathcal{L}_{\text{event}} \in \{\mathcal{L}_{\text{PFS}}, \mathcal{L}_{\text{OS}}, \mathcal{L}_{\text{AE}}\}$. We fine-tune the survival outcome (i.e. PFS or OS) and adverse event prediction heads independently with their corresponding loss functions. Both the forecasting and event prediction losses are minimized using the Adam optimizer [142].

Bootstrap Aggregation

To reduce the variance of our estimator and to provide reliable uncertainty predictions, we use bootstrap aggregation (also known as bagging), a gold standard machine learning technique for uncertainty estimation [36]. In practice, for each fold, we train five different Transformer-CPH models where the training set is composed of a random selection of the original training set (with replacement) sampled to 1.5 times the original size. The validation and test sets are left unchanged.

4.2.3 Transformer introspection

We perform introspection of the transformer hidden states. The hidden states are “joint” representations that are used to predict adverse events, survival, and biomarkers longitudinally. The hidden representations at each time t are denoted as $z_t^{(L)}$. The predictions for disease progression and biomarker forecasting are referred to as $\hat{y}_{t,\text{pfs}} \in \mathbb{R}$ and $\widehat{X}_t$, respectively. We obtain the UMAP plot in Figure 4-7 by using the the $z_1^{(L)}$ vectors (at the first time step) of each patient.

The experiment reported in Figure 4-8 consists of computing the Pearson correlation (ρ) between the different dimensions of the hidden states $z_t^{(L)}$ and the various predictions:

- For PFS, we have $\rho(z_{t,d}^{(L)}, \hat{y}_{t,\text{pfs}}), \forall d \in \{1, \ldots, D_z\}$.
- For each biomarker variable j, we have $\rho(z_{t,d}, \widehat{X}_{t,j}), \forall d \in \{1, \ldots, D_z\}$.

These correlations are computed for t equal to one, three, six, nine, and twelve months.

We also compute a more interpretable risk score for each patient that represents the risk of their experiencing an event. Although the output of the event-prediction head represents such a score in theory, these scores are not limited in range and are only informative relative to the scores of other patients. Indeed, the semi-parametric nature of proportional hazards models only enforces a consistent *ordering* of the scores across the dataset. The absolute value of the score is not meaningful. Thus, to improve the interpretability of the event prediction scores, we compute a normalized score $\tilde{y}$:

$$\tilde{y}_t = \frac{\hat{y}_t - \hat{y}_{min}}{\hat{y}_{min} - \hat{y}_{max}}, \tag{4.10}$$

where $\hat{y}_{min}$ and $\hat{y}_{max}$ are the minimum and maximum values of the score across patients and time in the training set. This normalized score can directly be interpreted as a quantile of the risk of a particular patient with respect to the whole cohort. For instance, $\tilde{y}_t = 0.7$ suggests that 70% of patients in the cohort have a lower risk score. We used these normalized scores in Figure 4-9.

4.2.4 Counterfactuals and subgroup analysis

We use a trained Transformer-CPH model to impute the counterfactuals on the entire MM2 dataset. Counterfactuals are obtained by using our model to perform predictions,

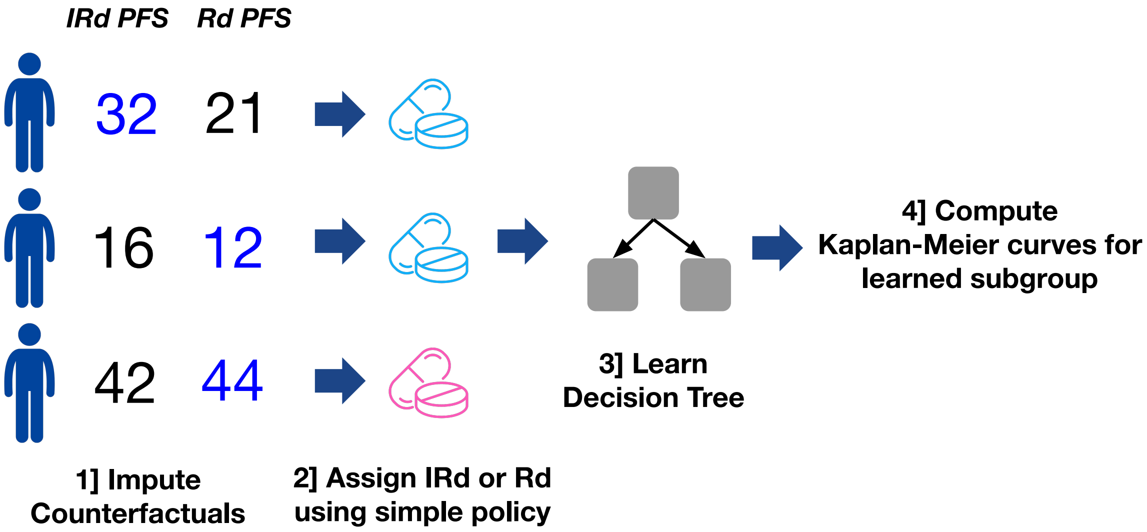

IRd PFS
Rd PFS
32
21
16
12
42
44
1] Impute Counterfactuals
2] Assign IRd or Rd using simple policy
3] Learn Decision Tree
4] Compute Kaplan-Meier curves for learned subgroup

Figure 4-3: Diagram of a proof-of-concept subgroup discovery analysis.

with modified treatment input vectors. Information about the treatment appears both in vector B and in the longitudinal treatment vector A. We thus change both the treatment arm covariate in B as well as the dosages in A to the desired treatment regimen, i.e. either Rd or IRd, before doing inference. We perform predictions of potential outcomes at baseline only (as this is the only time step where randomization occurs). We write the potential outcomes, i.e. the predicted PFS risk score, under IRd and Rd, respectively:

$$\hat{y}^1_{\text{PFS}} = f_{PFS}(z_0^{(L)}(X_0, B^1, A^1))$$
$$\hat{y}^0_{\text{PFS}} = f_{PFS}(z_0^{(L)}(X_0, B^0, A^0)).$$

B^1 and B^0 refer to the baseline vectors where the treatment indicator is set to 1 or 0. Similarly, A^1 and A^0 refer to the IRd and Rd treatment strategies, respectively. Finally, we define potential biomarker trajectories under the two treatment regimens:

$$\widehat{X}^1_{t,\text{PFS}} = f_{\text{pred}}(z_{0+t}^{(L)}(X_0, B^1, A^1))$$
$$\widehat{X}^0_{t,\text{PFS}} = f_{\text{pred}}(z_{0+t}^{(L)}(X_0, B^0, A^0)),$$

where we make explicit that the trajectories are only conditioned on information available at baseline. Examples of biomarkers trajectories conditioned on different treatments are presented in Figure 4-10.

Given $\hat{y}^1_{\text{PFS}}$ and $\hat{y}^0_{\text{PFS}}$, we can define a conditional average treatment effect (CATE) on risk of disease progression for each patient as follows,

$$\text{CATE} = \hat{y}^1_{\text{PFS}} - \hat{y}^1_{\text{PFS}}.$$

p-values in subgroup analysis is highly impacted by the sample size. To provide more reliable estimates, we create an even split of the MM2 cohort into a train and test set. This new train set consists fully of the original train set minus a random selection patients to reach 50% of the cohort. The rest of the patients then comprise the test set. All model training and policy learning are done on the train set, and all Kaplan-Meier curves shown are computed on the test set.

Using this training cohort from MM2, we then learn a policy to assign treatment; specifically, we use the median conditional average treatment effect (CATE) over the training patients as a threshold to determine whether a patient would be assigned to a subgroup for whom IRd would be most efficacious. Because $\hat{y}$ represents the *risk* of disease progression, a positive CATE indicates a negative effect of IRd (the risk of disease progression is higher) while a negative CATE indicates a positive effect of IRd. We then define the subgroup threshold as the median of all CATE values in the training set ($\delta = \text{CATE}_{\frac{1}{2}}$). We obtain the initial subgroup assignment strategy, π^*, to determine if someone is included in the subgroup or not, as follows, for each patient i:

$$\pi^*_i = \begin{cases} 0 & \text{if} \quad \text{CATE}_i > \delta \\ 1 & \text{if} \quad \text{CATE}_i \leq \delta \end{cases}$$

Intuitively, a patient is in the subgroup, i.e., $\pi^* = 1$, if IRd is effective at decreasing their risk of progression.

To improve the interpretability of this subgroup assignment strategy, we learn a shallow decision tree where we used $\pi^*_i, \forall i$ as the labels. We train the decision tree on the training cohort using the baseline clinical variables, B, as covariates. The output of this decision tree is then used as another, more interpretable, subgroup strategy, π. Kaplan-Meier curves are estimated for the discovered subgroups (found via both subgroup assignment strategies, π^* and π) by looking at the actual treatment and control groups in the subgroup (Figure 4-11. This procedure is depicted graphically at a high level in Figure 4-3.

4.3 Results

Our results section is structured as follows. Firstly, we present the statistics of the two patient cohorts, namely MM1 and MM2. Secondly, we provide a comprehensive evaluation of Transformer-CPH and its ability to jointly predict disease progression, overall survival, adverse events, and forecast biomarkers. Thirdly, we conduct an introspective analysis of the hidden states of Transformer-CPH over time, which reveals a strong correlation between the trajectories of clinical biomarkers and the risk of disease progression. This enables us to gain insights into the underlying signals detected by the transformer model. Fourthly, we showcase the usefulness of Transformer-CPH as a response-surface model that can generate factual and counterfactual predictions, facilitating the computation of conditional average treatment effects (CATEs) and the identification of subgroups with heterogeneous treatment effects. Finally, we externally validate the model on a distinct dataset of multiple myeloma patients who are in a different stage of their disease progression, namely relapsed and refractory patients, as opposed to newly diagnosed ones.

4.3.1 Cohort Statistics

Table 4.2 presents a breakdown of the cohort statistics for both MM1 and MM2. The MM2 cohort consists of 703 patients, while the MM1 cohort consists of 722 patients. The inclusion criteria were autologous stem cell transplant (ASCT) ineligibility as well as a new symptomatic myeloma diagnosis for MM2, whereas the inclusion criteria for MM1 included relapse from previously controlled disease. We perform a random 80/20 split of the data into a training set and test set. We further do five random 75/25 splits of the training set: each split consisted of a smaller training set that is used to train the model and a validation set that is used to tune the hyperparameters of the model. All metrics in the subsequent sections are reported on the test set, averaged over the five trained models. The error bars correspond to standard deviation, computed over the five model predictions.

4.3.2 Joint-modeling for multiple myeloma

Transformer-CPH uses a patient's past clinical history and future treatment plan to jointly predict the risk of disease progression and adverse events as well as forecast clinical markers. Notably, it supports an arbitrary duration for the available clinical history (t_{cond}). Our method can thus use the entirety of the available clinical history of a patient, regardless of its specific duration. For each individual prediction task (forecasting and the various event prediction tasks), we compare the performance of Transformer-CPH against methods trained specifically to that particular task and to a particular duration of the clinical history. Despite this experimental setup apparently advantaging the task-specialized methods, we show that Transformer-CPH compares favorably and sometimes outperforms the other approaches.

We first assess our model's event prediction performance, focusing on progression-free survival (PFS), overall survival (OS) and adverse events (AE) as the clinical outcomes of interest. We consider the performance of our model for different durations of clinical history (t_{cond}) and for different forecasting horizons (t_{horizon}). We consider three observation windows and forecasting horizons, one month, six months, and twelve months, since most patients had already experienced the event or were censored after two years. Each unit of time in our results is measured as a "treatment period", corresponding to 28 days, or approximately one month. We evaluate the ability of our model to predict the onset of disease progression using the concordance index for right-censored data based on inverse probability of censoring weights (C-index IPCW)[1] [268]. We compare our approach against the international scoring system for

[1] C-index is 0.5 for random performance

Variable	**MM2 Cohort**	**MM1 Cohort**
Patients	703	720
Sex n(%)		
Female	351 (49.9%)	312 (43.3%)
Male	352 (50.1%)	408 (56.7%)
Race n(%)		
White	575 (81.8%)	614 (85.3%)
Asian	96 (13.7%)	64 (8.9%)
Native Hawaiian or other Pacific islander	1 (0.1%)	4 (0.6%)
Black or African American	23 (3.3%)	13 (1.8%)
American Indian or Alaska native	3 (0.4%)	1 (0.1%)
Other	5 (0.7%)	7 (1.0%)
Not reported	0 (0.0%)	17 (2.4%)
IG Type n(%)		
IGG	403 (57.3 %)	389 (54.0%)
IGA	142 (20.2 %)	123 (17.1 %)
IGD	10 (1.4 %)	7 (1.0 %)
IGE	3 (0.4 %)	15 (2.1 %)
IGM	3 (0.4 %)	1 (0.1 %)
Biclonal	23 (3.3 %)	30 (4.2 %)
No Heavy Chain	119 (16.9 %)	155 (21.5 %)
Age [years] median(range)	73 (48,90)	66 (30,91)
Time from diagnosis [months] median(range)	1.11 (0.3,52.9)	42.8 (3.0,306)
ISS Stage at study entry n(%)		
I	324 (46.1%)	458 (63.6%)
II	263 (37.4%)	176 (24.0 %)
III	115 (16.4%)	86 (12.0 %)
Actual Treatment n(%)		
Rd	349 (49.6%)	359 (49.9%)
IRd	354 (50.4%)	361 (50.1%)
Planned Treatment n(%)		
Rd	353 (50.2%)	362 (50.3%)
IRd	350 (49.8%)	358 (49.7%)

Table 4.2: Summary statistics of the MM1 and MM2 patient cohorts

multiple myeloma [101] (CPH-ISS) and two machine learning models: Cox proportional hazards models (CPH) and a random survival forest (RSF). As these methods cannot incorporate varying lengths of clinical history, a different Cox and RSF model is trained for each clinical history duration. This contrasts with Transformer-CPH, which relies on a single model to predict progression for any clinical history duration.

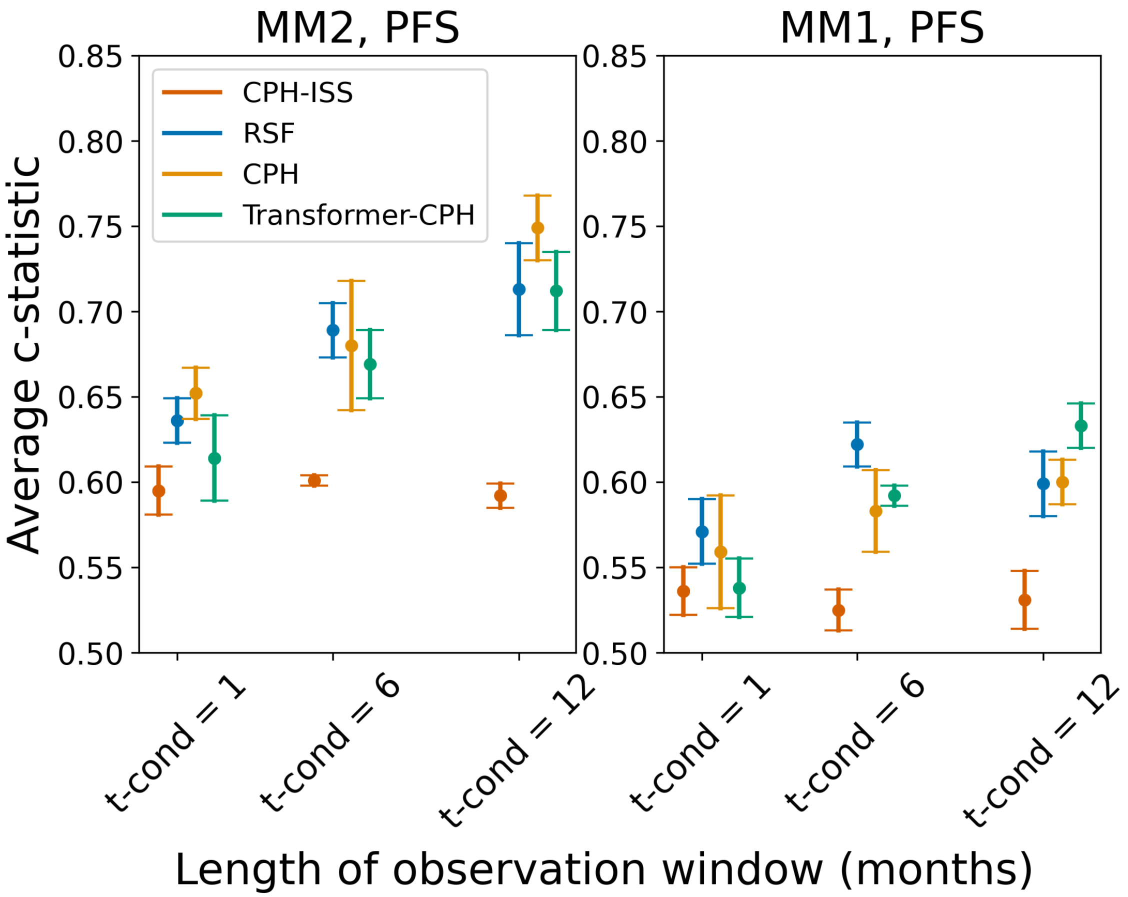


Figure 4-4: We train our model on MM2, which consists of newly diagnosed multiple myeloma patients (NDMM), and evaluate on a held-out portion of MM2. We further evaluate on MM1, which consists of relapsed and refractory multiple myeloma patients (RRMM). We report concordance index based on inverse probability of censoring weights (C-index IPCW) averaged across three time quantiles (25th, 50th, and 75th quantiles) at different observation windows (1 month, 6 months, and 12 months). Looking at both MM2 and MM1 together, we find that the Transformer-CPH has largely comparable performance to the RSF and CPH models, and significantly better performance than CPH-ISS. We note the added benefit of having to train Transformer-CPH only once, compared to the two other model architectures, which require a separate model for each observation window and each event outcome.

Figure 4-4 shows a graphical representation of these results. We find that our model is competitive with the machine learning baselines. Importantly, it outperforms the CPH-ISS clinical baseline, which is based on a version of the risk score used in current clinical practice. For example, for $t_{\text{cond}} = 6$, we have a C-index IPCW of 0.67 (+/– 0.02) for Transformer-CPH, 0.60 (+/– 0.01) for CPH-ISS, 0.68 (+/– 0.04) for the Cox model, and 0.69 (+/– 0.02) for RSF. We note that this trend persists across patient

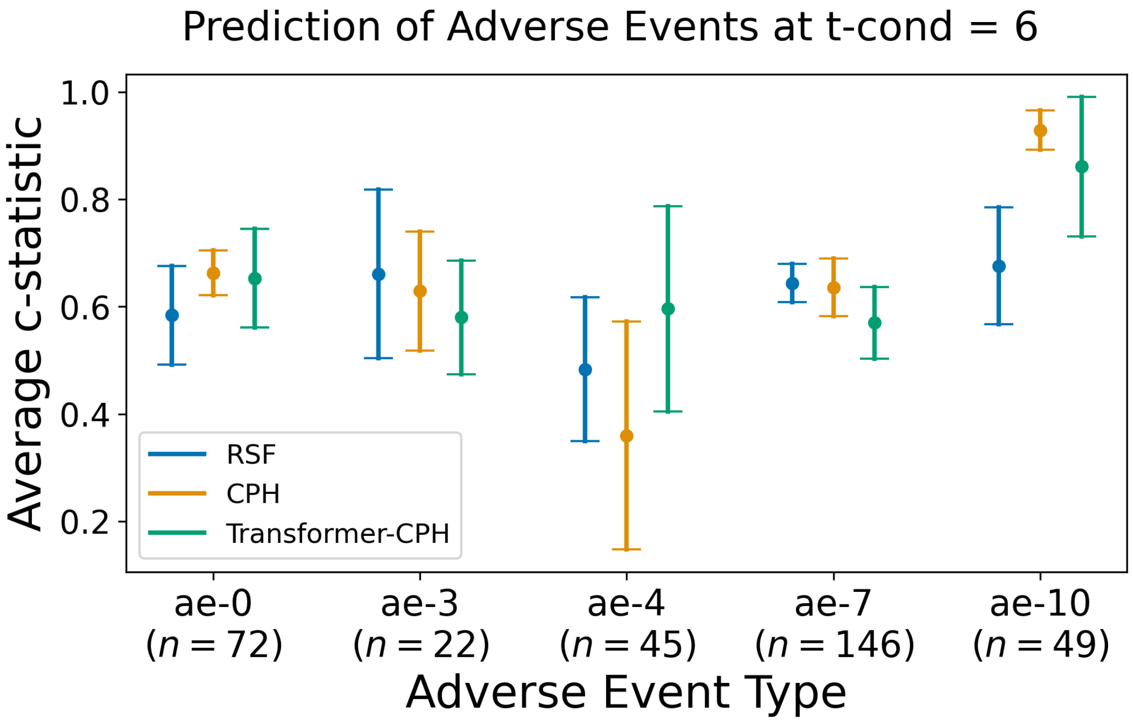


Figure 4-5: We report average concordance index for multiple adverse events (filtering down to only ≥ Grade 2 non-hematalogic events and ≥ Grade 3 hematologic events) at the 6 month observation window (we refer to Appendix B for results at other time points). The adverse events are mapped to shortened names as follows – ae-0: Acute Renal Failure, ae-1: Cardiac Arrhythmias, ae-2: Diarrhea, ae-3: Heart Failure, ae-4: Hypotension, ae-5: Liver Impairment, ae-6: Nausea, ae-7: Neutropenia, ae-8: Peripheral Neuropathies, ae-9: Rash, ae-10: Thrombocytopenia, and ae-11: Vomiting. We find that for those adverse events that were predictable from the data (i.e., heart failure, hypotension, acute renal failure, neutropenia, and thrombocytopenia), Transformer-CPH is competitive with highly-tuned, task-specific CPH and RSF models trained separately on each adverse event.

subgroups with a different dominant immunoglobulin heavy chain (see Appendix B). For overall survival (OS) prediction, we observed a similar result, where Transformer-CPH outperforms CPH-ISS, except for the first time point. The OS results are given in Appendix B. Results for all observation windows and detailed per patient subgroups are also available in Appendix B.

Concurrently, we measure the ability of our model to forecast future clinical trajectories by computing the mean squared error (MSE) over the forecasting horizon. We compare against a recurrent neural network (RNN) and a deep Markov model (DMM), two state-of-the-art time series forecasting models, and a last observation

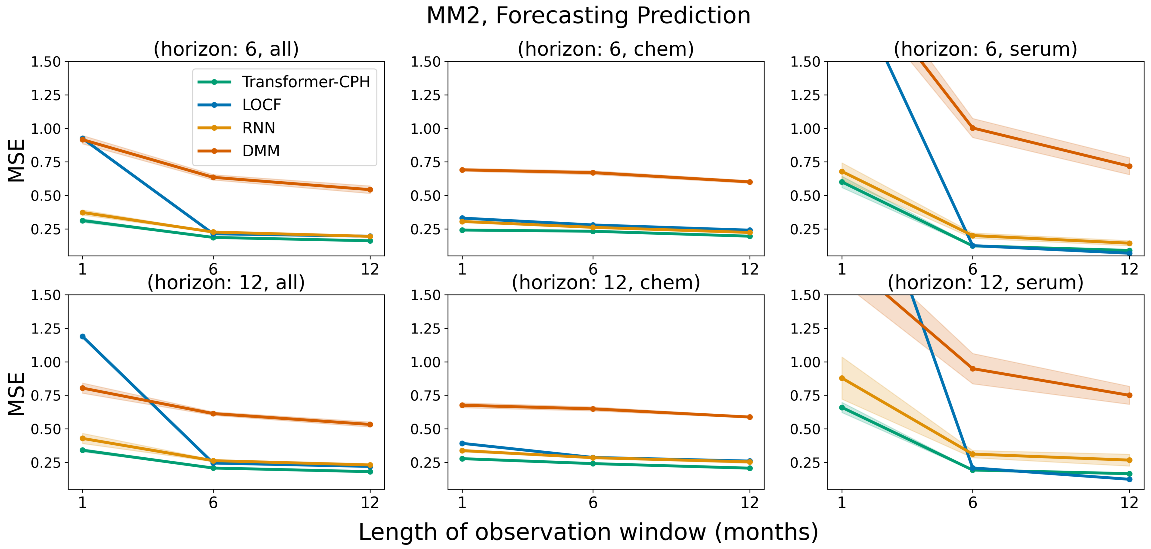


Figure 4-6: Finally, we plot the mean squared error (MSE) of each model on forecasting different sets of variables, chemistry labs, serum immunoglobulins, and all lab values, over two forecasting horizons, 6 months and 12 months. Evaluation is done after having observed all of the patient's data at three different time points (t_{cond}): 1 month, 6 months, and 12 months. We found that Transformer-CPH outperformed the other methods in all cases.

carried forward mechanism (LOCF). The results are shown in Figure 4-6. For an observation window of 1 month (i.e. conditioning on all of the patient's data at 1 month) ($t_{\text{cond}} = 1$) and a horizon of 6 months ($t_{\text{horizon}} = 6$), our method outperforms all other methods, with a MSE of 0.31 (+/– 0.01) for Transformer-CPH, 0.37 (+/– 0.02) for the RNN, 0.92 (+/– 0.03) for the DMM, and 0.93 (+/– 0.0) for LOCF. We note that the performance gap decreases with an increasing length of the observation window. This result may be due to the decreasing variance of the clinical trajectories over time.

In addition to forecasting clinical trajectories and predicting disease progression and overall survival, Transformer-CPH is able to predict the onset of particular adverse events, as shown in Figure 4-5. We find that our model can predict the occurrence of acute renal failure (C-index IPCW 0.62 (+/– 0.07)), hypotension 0.66 (+/– 0.14), neutropenia 0.59 (+/– 0.07), and thrombocytopenia 0.85 (+/– 0.10).

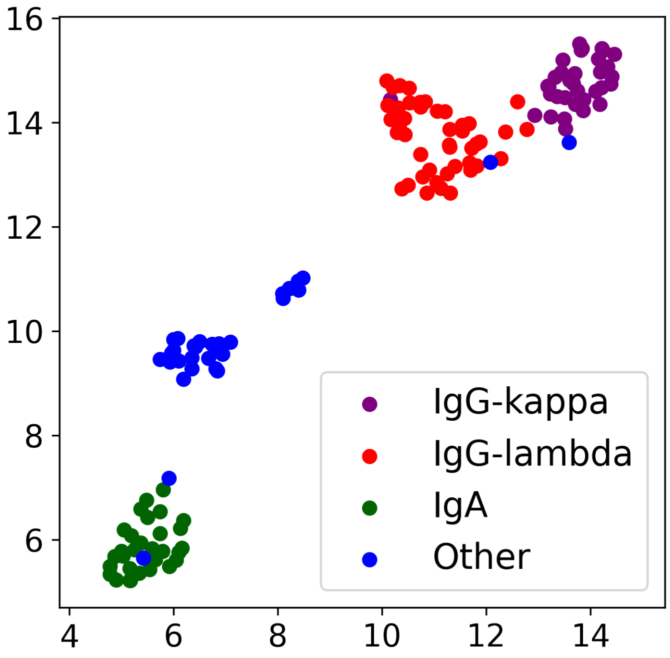


Figure 4-7: We compute a UMAP embedding of the transformer hidden state at the first time point and then color each patient by their myeloma subtype. We find that the hidden state captures the underlying myeloma subtype structure, including subtypes delineated by which heavy and light chains dominate the disease process.

4.3.3 Introspection into Transformer-CPH

We conduct analyses to interpret the learnt hidden states of Transformer-CPH. All analyses are performed with a single model trained on the MM2 training cohort. In Figure 4-7, we show a UMAP visualization of the hidden state vector at $t = 1$ month for each patient. Each point represents a single patient in the test set. We find that the hidden state captures the canonical myeloma subtypes, which are determined by the involved heavy chain and light chain.

We then proceed with a finer analysis of each individual dimension of the hidden state over time. In particular, we investigate what dimensions of the hidden state are associated with a higher event/progression risk and how this relates to trends in the biomarker predictions. For each dimension of the hidden state, and for each time step, we compute the Pearson correlation (1) between the value of the hidden state at said dimension and the predicted risk of progression, and (2) between the value of the hidden state at said dimension and the predicted value at the next time step for all

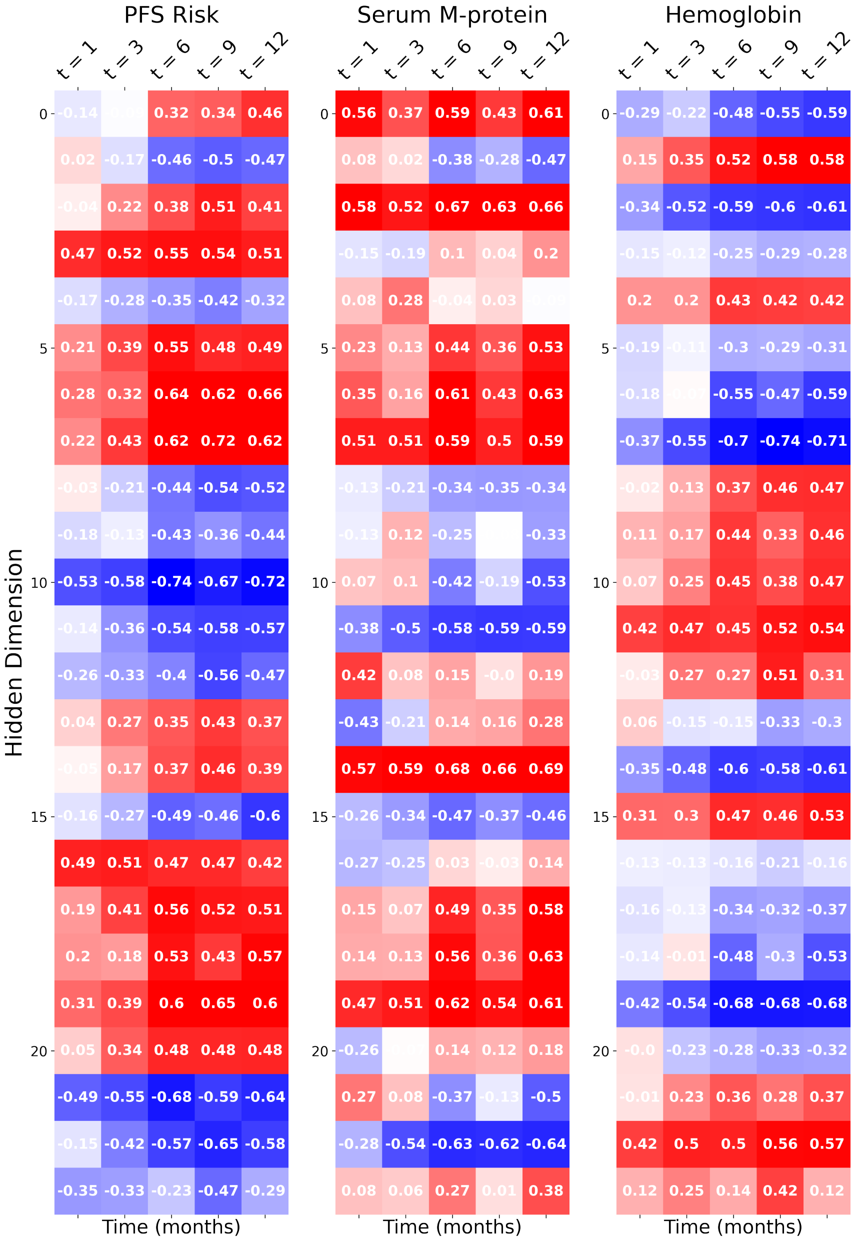
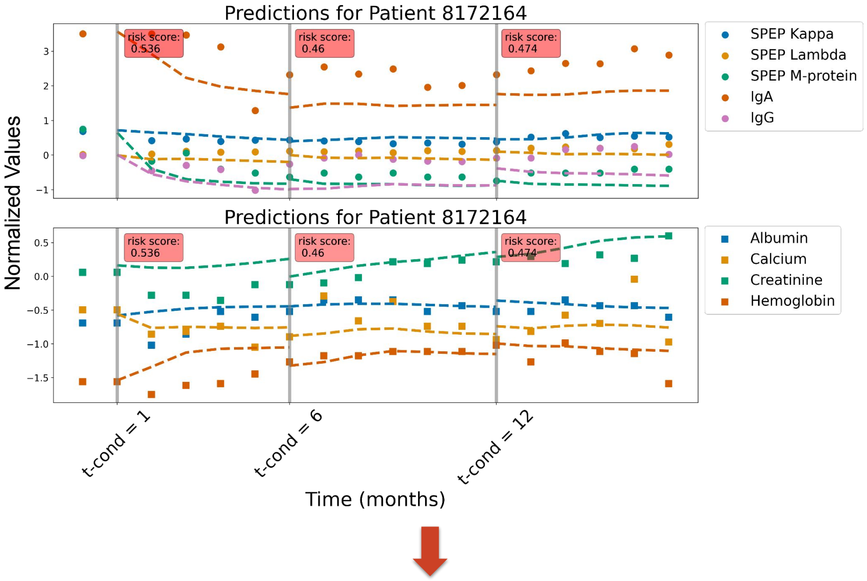


Figure 4-8: At five different time points, we compute the correlation between the hidden state for all patients and the risk of progression (*predicted value*), serum M-protein level (*feature*), and hemoglobin level (*feature*), respectively. The number of hidden dimensions is 64, but only the dimensions that have at least one time point above 0.4 are shown. Red indicates a positive association between the hidden state value and the feature or prediction, whereas blue indicates a negative association. We see that in dimensions where there is a growing risk of progression, a sensible change in the forecasts is noted, i.e., serum M-protein level tends to go up, and hemoglobin tends to go down, indicating anemia.

biomarkers. The resulting correlations for risk of progression, serum M-protein, and hemoglobin are presented in Figure 4-8, for five observation windows (one month, three



71 yo M, ISS stage 2, ECOG 0 presenting with symptomatic, newly diagnosed multiple myeloma, IgA-dominant (baseline IgA level = 31.3 g/L, FLC ratio = 2.55), who was put on an IRd regimen after diagnosis. Baseline labs are 2.22 mmol/L Calcium, 1.0 mg/dL Creatinine, 3.6 g/dL Albumin, 9.3 g/dL Hemoglobin.

ML model forecasting to 6 months (at baseline) – predictions in **blue** and true values in **red**
IgA: trending downwards to ***4.4 g/L (5.4 g/L)*** at 6 months from treatment initiation
CRAB features: stable calcium at ***2.16 mmol/L (2.12 mmol/L)***, hemoglobin trending upwards to ***11.0 g/dL (10.3 g/dL)***, creatinine within normal limits at ***1.09 mg/dL (0.92 mg/dL)***
Albumin stable at ***3.9 g/dL (3.8 g/dL)***.
Progression risk score = *0.54*
Renal failure risk score = *0.27*; Thrombocytopenia risk score = *0.02*

ML model forecasting to 12 months (at 6 months)
IgA: trending downwards to ***3.83 g/L (4.78 g/L)*** at 12 months
CRAB features: stable calcium at ***2.13 mmol/L (2.16 mmol/L)***, hemoglobin remaining stable at ***10.7 g/dL (10.8 g/dL)***, creatinine with slight uptick but still wnl at ***1.1 mg/dL (1.07 mg/dL)***
Albumin stable at ***3.9 g/dL (3.9 g/dL)***.
Progression Risk score = *0.46*
Renal failure risk score = *0.22*; Thrombocytopenia risk score = *0.04*

Figure 4-9: We generate samples of several biomarkers, including immunoglobulins and chemistry labs, from the model at three different conditional time points (one month, six months, and twelve months) for a test patient. The solid dots denote the ground truth values, and the dotted lines are the predictions. At each time point, we also report a risk score for disease progression, and two adverse events: acute renal failure and thrombocytopenia. How we compute these risk scores from the predictions is defined in Section 4.2.3. These predictions enable a clinical assessment of individual patients that can be summarized into a clinical vignette for the physician, which we do manually.

months, six months, nine months, and twelve months). Each cell is colored according to the correlation strength. For clarity, we remove the hidden dimensions that had low correlation (below a threshold of 0.4). We find that for hidden dimensions indicating a higher risk of progression, sensible correlations with M-protein and hemoglobin predictions are appreciated. Namely, serum M-protein level is predicted to go up, given the hidden state as input, and hemoglobin is predicted to go down, indicating anemia, which agrees with basic clinical reasoning. This demonstrates the ability of our model to recover associations between longitudinal biomarkers and clinical events.

Finally, in Figure 4-9, we demonstrate the granularity of the model's predictive capabilities by reporting the predictions for disease progression and forecasts for several biomarkers over multiple horizons and observation windows, for an individual patient. We further show that these predictions can be used to create individual clinical summaries that may assist physicians in care management. We find that our model generally predicts the correct trends for both immunoglobulins and biomarkers related to "CRAB"-like symptoms, e.g. calcium, hemoglobin, and creatinine. Figure 4-9 depicts an example for a patient with IgA-dominant multiple myeloma.

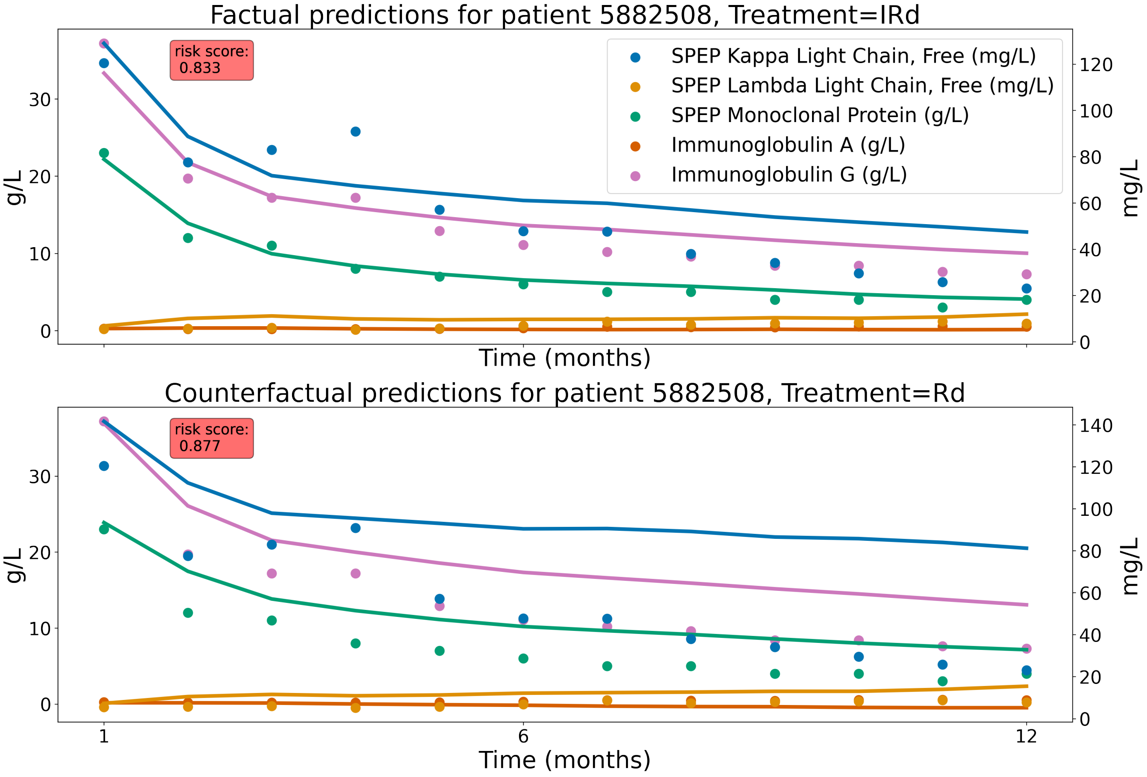


Figure 4-10: We plot the predictions from baseline of a single patient's serum immunoglobulins over both the factual treatment (here IRd) and the counterfactual treatment (Rd), demonstrating Transformer-CPH's ability to model counterfactuals.

4.3.4 Assessing the effect of treatment and uncovering heterogeneous subgroups

Our architecture uses the *planned treatment assignments* to provide forecasts of clinical trajectories and survival predictions for individual patients. Because the MM2 dataset was collected in the context of a randomized clinical trial, our model can be used to investigate the impact of *counterfactual* treatments on individual patients as well. Furthermore, due to its ability to forecast clinical trajectories and to predict survival, Transformer-CPH can provide a comprehensive clinical assessment of the impact of a treatment on individual patients. An example of forecasting with different treatment assignments for an individual patient is presented in Figure 4-10. For this patient, Transformer-CPH assigns a lower risk score when treated with IRd than with Rd. Projections with Rd predict a marked increase in serum kappa light chain, serum M-protein, and IgG over time compared to a scenario where the patient would be treated with IRd.

By producing individual predictions with different treatment assignments, our model can also serve as a useful tool to discover sub-populations of the initial patient cohort that have heterogeneous treatment effects. To demonstrate this ability, we computed the individual treatment effects on risk of disease progression for each patient and identified the subgroup of patients that would benefit the most from the treatment using a simple treatment policy learned from the MM2 training data. We additionally produced a more interpretable view of the identified subgroup using a shallow decision tree. See the "Counterfactuals and subgroup analysis" section under Methods for more details on the procedure. This treatment assignment policy resulted in a subgroup for which the statistical significance in favor of the treatment was vastly improved.

Following the original clinical trial analysis [79], we evaluated the significance of treatment by performing a stratified Cox regression on a held out set of MM2 patients. This analysis resulted in a barely non-significant treatment effect ($N = 351, p = 0.093$), which was slightly higher than the significance level obtained on the whole MM2 cohort

($N = 703, p = 0.073$). In contrast, in the subgroup identified with our approach, the effect of treatment was strongly significant ($N = 182, p = 0.014$), using the same strata in the stratified Cox regression. A shallow decision tree trained to replicate the subgroup assignment strategy of our model resulted in a subgroup with similar significance with respect to treatment efficacy ($N = 153, p = 0.016$). The inclusion criteria for the discovered subgroup, as reflected by the learned decision tree, are presented in Figure 4-12. The relevant clinical variables are the subtype of multiple myeloma, characterized by the dominant heavy chain and light chain, the free light chain concentration, serum creatinine, CG and GG polymorphisms, and the Durie-Salmon stage. Notably, the learned subgroup suggests that IgA patients are more likely to benefit from IRd than IgG and IgM patients.

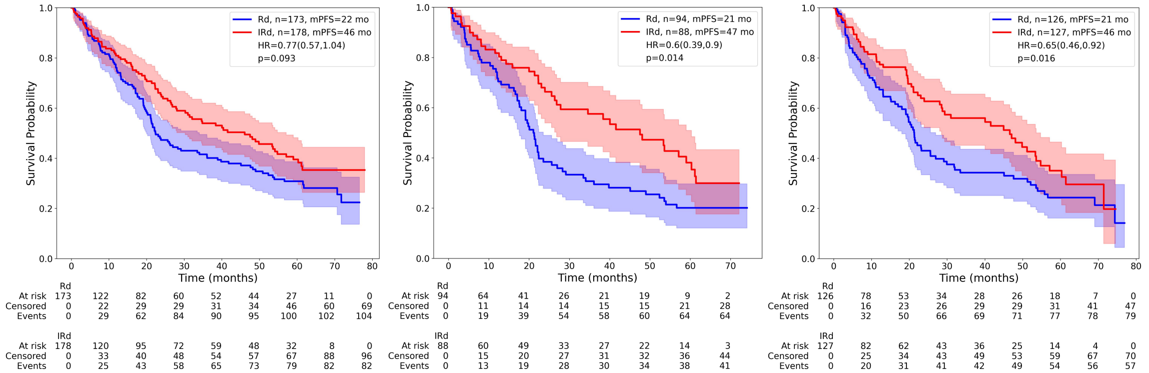


Figure 4-11: On the left, we show the Kaplan-Meier curves for the original treatment and control groups in a left out set of MM2 patients. In the middle, we show the Kaplan-Meier curves for the treatment and control groups in a learned subgroup of MM2 that suggest greater differential survival between IRd and Rd for this patient subgroup. On the right, we show the Kaplan-Meier curves for the interpretable subgroup learned to replicate the subgroup found with our model. The p-values were computed with a log-rank test.

4.3.5 Generalizability to relapsed and refractory multiple myeloma patients

We evaluate our model on the MM1 trial data, which consists of patients with relapsed myeloma who had received prior lines of therapy. This cohort is distinct from the MM2 trial, where patients had newly diagnosed myeloma. The evaluation of our model

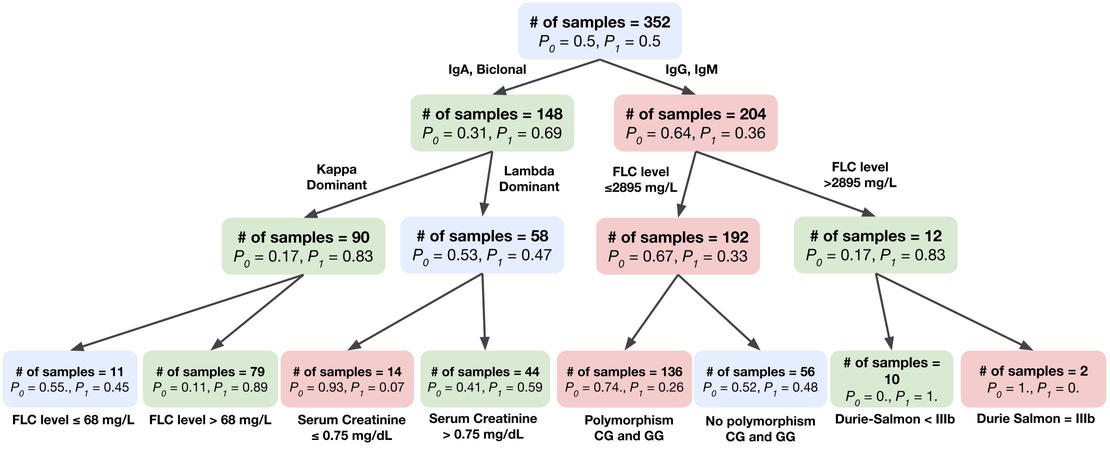


Figure 4-12: Visualization of the learned decision tree, which is trained using the CATEs as labels and patient baseline data as features. Each node contains the number of patients, and the proportion of the samples who are labeled as getting Rd (P_0) or IRd (P_1), respectively. A node is colored red if $P_0 > P_1$, green if $P_0 < P_1$, and blue if $P_0 \approx P_1$.

on an external cohort with different disease characteristics compared to the training population serves as external validation and measures the model's ability to generalize. In the PFS prediction task (Figure 4-4), we observe a decrease in performance across all models, as expected due to the change in cohort characteristics. However, Transformer-CPH has the lowest average decrease in concordance index (0.077) compared to RSF (0.083) and CPH (0.11). CPH-ISS has a comparable decrease (0.076 vs 0.077) to Transformer-CPH, but with poor absolute performance. Additionally, we find that our model outperforms all other models in forecasting performance on the MM1 cohort (see Appendix B), despite a slight increase in mean squared error across each t_{cond}. Our results suggest that Transformer-CPH has better generalizability than the baselines, but caution should be exercised when applying the model to a cohort with different inclusion criteria than the training cohort.

4.4 Discussion & Related Work

In this chapter, we proposed a transformer-based architecture, Transformer-CPH, that jointly forecasts core myeloma biomarkers and predicts risk of events, including disease progression, overall survival, and adverse events. In recent years, there has been a rich line of work leveraging purely attention-based architectures, i.e., transformers, for tasks involving temporal data. For example, several prior papers detail architectural-level and module-level variants of the transformer, including incorporating causal convolutions as well as introducing sparse bias into the attention module, to perform better forecasting [171, 165, 303, 304, 46, 290]. With respect to event prediction, prior work has included adapting transformers for electronic health record (EHR) data to do risk prediction. Examples include BEHRT, Med-BERT, CEHR-BERT, and Hi-BERT [166, 167, 195, 217].

In contrast to these prior approaches, our method is deeply motivated by the clinical management of multiple myeloma, as outlined by the International Myeloma Working Group. Managing multiple myeloma involves a crucial balance between maximizing survival, minimizing adverse events, and keeping track of core biomarkers, which largely serve as a sufficient proxy for disease burden, over time [70]. These principles hold true in a wide range of malignancies and inform oncologic management more generally [134, 235, 179]. Our model reflects these clinical management objectives by jointly modeling time to event and longitudinal biomarkers. Joint modeling of time to event and longitudinal data has a long history in the biostatistics literature, where it has been used to improve the efficiency of estimators [125]. Most approaches for joint modeling have relied on linear mixed effect models and classical survival models, such as Cox regression and Accelerated Failure Time models [220, 267]. For example, Proust-Lima et al. developed and validated a dynamic prognostic tool for prostate cancer recurrence using repeated measures of post-treatment prostate specific antigen (PSA) [214]. However, these prior approaches largely focus on a single event and univariate biomarker predictions, in contrast to our work, where we consider multivariate forecasting and prediction of multiple events. In general, previous literature has predominantly employed less expressive models, such as linear models, in contrast to the transformer architecture that underlies our model.

On the joint modeling tasks, our experiments showed that Transformer-CPH outperforms other state-of-the-art longitudinal models for biomarker forecasting, over all observation windows and prediction horizons. Additionally, our model was comparable to tailored event prediction models, and superior to CPH-ISS, both in terms of prediction of disease progression and overall survival. Unlike these approaches, our model can do inference given data over any observation window. We further showed that our model can accurately predict the occurrence of serious adverse events such as heart failure, hypotension, acute renal failure, neutropenia, and thrombocytopenia. The ability of our model to jointly predict these aspects of the disease presents significant potential for achieving more personalized clinical care. Indeed, with our approach, clinicians are afforded a more holistic picture of a patient's disease progression. Current approaches either have to be trained separately at each time point or are not as expressive, e.g. the linear mixed effect models used in current joint modeling approaches.

In our introspection experiments, we demonstrated the potential of our model towards personalized care by showing how generated predictions for an individual patient can be summarized into a concise, tailored clinical assessment over time. Though we did this summarization manually by looking at the model predictions, recent advances in generative artificial intelligence may enable training of an end-to-end model that can generate predictions and summarize these samples into a clinically meaningful note [302].

A typical limitation to the deployment of machine learning models in clinical practice is their black box nature. We performed an in-depth investigation of the model by studying the hidden representations used in our architecture. First, we showed that the hidden representations tended to cluster by myeloma subtype. We further studied the learnt associations between biomarkers and event predictions, conditioned on the hidden representations. While the model recovered clinically sensible correlations (e.g., higher serum M-protein being positively correlated with higher risk of disease progression), this introspection can also potentially serve as the basis for discovery of novel associations. This can be especially pertinent in the case of genomic marker discovery.

We established the potential of Transformer-CPH as a response-surface model with which we can estimate counterfactual outcomes and thereby estimate personalized treatment effects, i.e. in this case, the effect of IRd compared to Rd for individual patients. Providing counterfactual predictions can both help clinicians make more informed treatment decisions and serve as a tool for uncovering heterogeneous patient subgroups. As a demonstration of the latter, we propose an algorithm that uncovers a patient subgroup, i.e., IgA and biclonal myeloma patients with high baseline kappa light chain levels, for whom IRd leads to statistically significantly better PFS than Rd. In the original pre-specified subgroup analysis done in [79] for the TOURMALINE-MM2 trial, median PFS was statistically significantly higher for patients who had certain high-risk cytogenetics, which included del(17p), t(4;14), t(14;16), and amp(1q21) abnormalities, as well as patients who had low creatinine clearance ($\leq$ 60 mL/min). Subgroups based on the nature of the monoclonal protein (e.g. IgA-dominant, IgG-dominant, etc.) were not assessed. However, we also find, in our learned decision tree, that patients with high serum creatinine, due to potentially low creatinine clearance, were more likely to be assigned to the "IRd subgroup", where IRd would be more efficacious. We emphasize that this proof-of-concept analysis is hypothesis-generating and that further work is necessary to validate this result. Past work in the treatment effect estimation literature had extensively explored using various models, including Bayesian regression trees and neural networks, as response-surface models, to estimate the expected outcome of interest conditioned on a treatment and patient covariates [247, 152, 118, 110]. We see our approach as an important step towards using transformer-based models for heterogeneous treatment effect estimation in *real-world data.*

Data driven models are by definition highly dependent on the input data they were trained on. This can lead to poor performance when these models are evaluated on a patient cohort with different characteristics. In an attempt to quantify the generalization ability of our approach, we evaluated our model on an external patient cohort (MM1) consisting of relapsed and refractory multiple myeloma patients. While Transformer-CPH did incur an expected loss of performance, the results suggested that our model is more robust than other state-of-the-art baselines. Still, caution should be exercised when using our approach (or any other approach) on other patient cohorts not included in the training data. Developing better techniques to improve the out-of-distibrution generalization of our model is an important direction for future work.

The design of our model was motivated by providing clinicians with the most holistic and personalized view of the disease state of multiple myeloma patients. Our model represents a step in that direction, though many hurdles towards clinical deployment remain. While we attempted to mitigate and quantify these limitations, prospective evaluation of the model in a clinical trial will be crucial for an eventual translation to clinical practice.

Independent censoring in the TOURMALINE trial A central assumption of the subgroup analyses done in this chapter is independent censoring, which is defined as the time-to-event being independent of the censoring time conditioned on some set of covariates [145]. In the table below, we show the distribution of reasons for why patients were censored.

Description	*n*
Progressive Disease (PD)	314
Death	64
No documented death or disease progression	165
Alternate therapy	107
Withdrawal of consent	24
Death or PD after more than 1 missed visit	22
Lost to follow-up	5
No Baseline/No Post-baseline	4

Table 4.3: Reasons for censorship in the TOURMALINE-MM2 Trial

The top two rows of Table 4.3 are descriptions for patients who were observed to have the event. The following: “withdrawal of consent”, “death or PD after more than 1 missed visit”, “lost to follow-up” are due to study termination or end of treatment. Under these conditions, independent censoring is reasonable [159]. Of the remaining descriptions, only two are substantial: “No documented death or disease progression” and “Alternate therapy”. We condition on age, ISS, and a pain score in our analysis, and in subgroups stratified by these covariates, it may be reasonable to assume independent censoring for the “Alternate therapy” group. Specifically, patients may be put on another therapy due to intolerability of the original treatment. This may be controlled for by conditioning on the pain score, which indicates how the patient may be tolerating the treatment, or the staging risk score, i.e. ISS. The former is much harder to assess without further information or further assumptions. To go beyond the original analysis done for the trial [79], we may condition on the entire set of covariates (beyond the the three strata described above) via a weighted log-rank test, where the weights are inverse probability of censoring weights (IPCW). This strategy would effectively control for any dependencies between the event times and censoring times through observed covariates.

Latent variable models vs attention-based architectures Finally, it is worth noting an interesting tradeoff uncovered by the results of this chapter and the prior one. In the prior chapter, we found that not encoding inductive biases into the latent variable model in the form of the treatment mechanism resulted in worse generalization and performance. Generally speaking, non-latent variable models tended to perform worse than the latent variable models. However, in this chapter, we see the opposite trend, where non-latent variable models, in this case an attention-based architecture, performed best. This paradoxical result is likely due to the differences in the data used in both chapters to train the respective models. In the previous chapter, we used an observational dataset, where greater diversity in the possible treatment regimens assigned resulted in fewer samples available per therapy ($<$100). In this chapter, we used trial data, where there were only two possible treatment regimens, allowing for $>$3x more samples (≈ 350). Thus, in settings where there are fewer samples per treatment regimen and more heterogeneity exists in the data, as is the case for the observational dataset, encoding inductive biases into the model and using a latent variable model, which has intrinsic stochasticity to capture the noise in the data, would be appropriate. On the other hand, in cases where the data is more homogeneous and the practitioner has more samples per treatment regimen, larger models for representation learning, such as RNNs or transformer architectures, will likely perform better.

Understanding how one can bolster the performance of large attention-based architectures even with highly heterogeneous observational datasets is an interesting direction for future work. One idea is to explore how the PK-PD mechanisms can be encoded into transformers. Another is to consider rigorous and statistically valid methods of pooling observational datasets together to improve sample efficiency and training. Such an approach would require standardizing the metadata across the datasets as well as improved imputation techniques, both of which are explored in [33].

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Uncertainty Quantification of Model Predictions using Conformal Inference

Acknowledgement of Co-authors I am the second author on the paper making up the bulk of this chapter, following first author Ahmed Alaa. While Ahmed and I had weekly meetings about the work described in this chapter, Ahmed took the lead in developing the initial method (for uncensored data) and the writing. I contributed the bulk of the experimental evaluation. As such, I do not include one of the theoretical results of the method in this chapter.

In the previous chapters, I focused on the theme of building predictive models of longitudinal clinical data with the vision of using these methods as the underlying workhorses for a CDSS assisting management of cancer patients. These models are meant to be clinically applicable, enabling prediction of survival outcomes, adverse events, and clinical biomarkers, three facets that are the cornerstones of cancer management. In the next three chapters, I shift to the theme of assessing the reliability of model predictions as well as building "sanity checks" for causal estimates derived from real world data. These methods might manifest in the CDSS, shown in Figure 5-1, as confidence intervals around the survival predictions and additional contextual information, respectively.

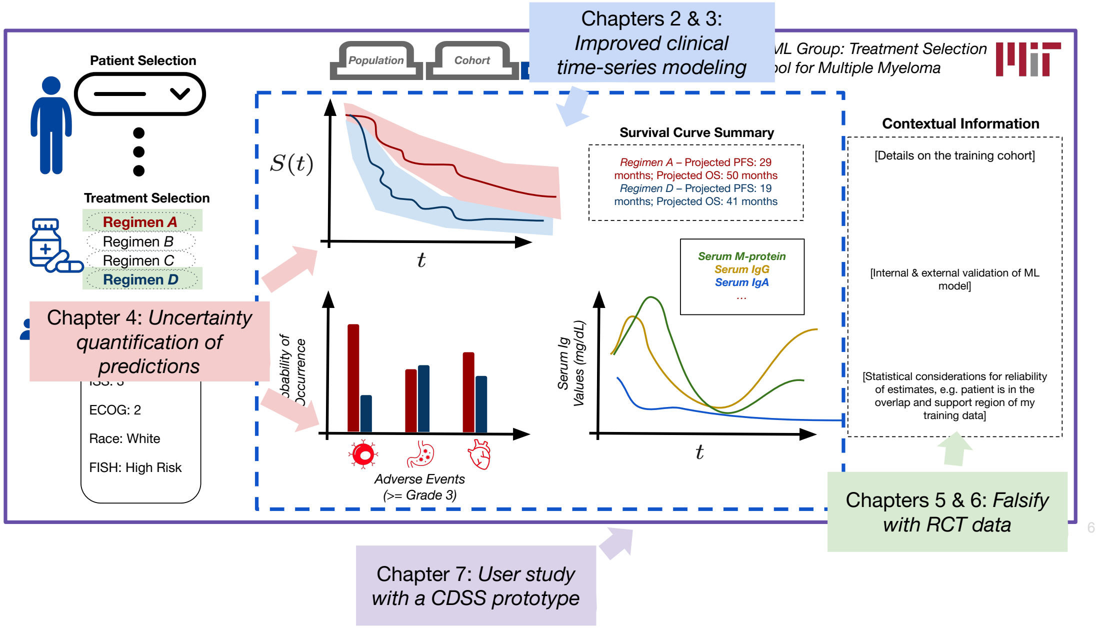

Chapters 2 & 3: *Improved clinical time-series modeling*
ML Group: Treatment Selection
ool for Multiple Myeloma
Population
Cohort
Patient Selection
Treatment Selection
Regimen A
Regimen B
Regimen C
Regimen D
Chapter 4: *Uncertainty quantification of predictions*
ISS: 3
ECOG: 2
Race: White
FISH: High Risk
$S(t)$
t
obability of
Occurrence
Adverse Events
(>= Grade 3)
Survival Curve Summary
Regimen A – Projected PFS: 29 months; Projected OS: 50 months
Regimen D – Projected PFS: 19 months; Projected OS: 41 months
Serum M-protein
Serum IgG
Serum IgA
...
Serum Ig Values (mg/dL)
t
Contextual Information
[Details on the training cohort]
[Internal & external validation of ML model]
[Statistical considerations for reliability of estimates, e.g. patient is in the overlap and support region of my training data]
Chapters 5 & 6: *Falsify with RCT data*
6
Chapter 7: *User study with a CDSS prototype*

Figure 5-1: Recap of clinical decision support system

5.1 Introduction

Consider a training data set $(X_1, Y_1), \ldots, (X_n, Y_n)$, and a test point (X_{n+1}, Y_{n+1}), with the training and test data all drawn independently from the same distribution, i.e.,

$$(X_i, Y_i) \overset{i.i.d}{\sim} P = P_X \times P_{Y|X},\ i = 1, \ldots, n+1. \tag{5.1}$$

Here, each $X_i \in \mathbb{R}^d$ is a covariate vector, while $Y_i \in \mathbb{R}$ is a response variable. The joint distribution over covariates and responses, P, is unknown. In this paper, we tackle the problem of *predictive inference*, where given the covariate vector X_{n+1} for a new test point, the goal is to construct a predictive interval that is likely to contain the true response Y_{n+1} with probability at least $1-\alpha$, for some $\alpha \in (0,1)$. More precisely, our goal is to use the n training sample points to construct a set-valued function:

$$\widehat{C}_n(x) := \widehat{C}_n((X_1, Y_1), \ldots, (X_n, Y_n), x) \subseteq \mathbb{R}, \tag{5.2}$$

such that for a new test point $(X_{n+1}, Y_{n+1}) \sim P$, the response Y_{n+1} falls in $\widehat{C}_n(X_{n+1})$ with probability $1-\alpha$. Intervals satisfying this *coverage* condition are said to be *valid*.

The sense in which predictive inferences are valid determines the relevance of the corresponding coverage guarantees to specific prediction instances. The weakest form of validity is when predictive intervals cover the true response on average—such intervals are said to be *marginally valid.* Formally, marginal validity is satisfied when

$$\mathbb{P}\left[Y_{n+1} \in \widehat{C}_n(X_{n+1})\right] \geq 1 - \alpha, \tag{5.3}$$

where the probability is defined with respect to randomness of both training and testing data. The coverage condition in (5.3) is said to be *distribution-free* if it holds for all P. Conformal prediction (CP), described formally in Section 5.2, is a popular framework for predictive inference that guarantees distribution-free marginal validity in finite samples [278, 277, 196, 161, 160, 162, 223]. In its most basic form, CP achieves marginal validity by issuing a fixed-length interval for all prediction instances.

In many applications, it is important to ensure *transparency* in communicating uncertainty in predictions issued for individual users. For instance, suppose that X_i is a set of risk factors for patient i (e.g., age, blood pressure, etc.), and Y_i is a measure of kidney function (e.g., eGFR). For a new patient, our goal would be to predict a range of values for their future eGFR with a predetermined degree of confidence, i.e., we would like to be able to make a statement along the lines of: "Based on your risk factors, there is a 95% chance that your eGFR will decline by 1.8–5.2 mL/min over the next 3 years". Marginal coverage guarantees that predicted ranges are accurate for 95% of the patients on average, but can be arbitrarily inaccurate for specific prediction instances. Since marginal coverage is defined with respect to P_X, we expect coverage to be violated in regions with few training examples in covariate-space, i.e., instances for which the predictive model has high *epistemic* uncertainty. Hence, the marginally-valid fixed-length intervals issued by vanilla CP may not be informative for prediction instances to which uncertainty quantification matters the most.

Adaptive and transparent CP. In this paper, we address the following question: how can we communicate model uncertainty in specific prediction instances in an *adaptive* and *transparent* manner? That is, we would like to construct a predictive inference procedure that *adapts* the length of its issued intervals based on the varying level of uncertainty across different prediction instances, and reports a coverage guarantee that is "relevant" to each specific instance. Ideally, we would like to develop a predictive inference procedure that achieves the following conditional coverage guarantee:

$$\mathbb{P}\left[Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,\middle|\, X_{n+1} = x\right] \geq 1 - \alpha, \tag{5.4}$$

for almost all[1] $x \in \mathbb{R}^d$. Predictive intervals that satisfy (5.4) are said to be *conditionally valid*. A procedure that satisfies (5.4) is transparent as its guarantee holds for each prediction instance, and is adaptive if the length of $\widehat{C}_n$ for a given x reflects the relative level of uncertainty in this specific instance. It is known that distribution-free validity in the sense of (5.4) is impossible to achieve for non-trivial predictions [276, 160, 88]. Hence, we build on the CP framework to develop an adaptive predictive inference procedure that satisfies an approximate version of (5.4), and transparently reports the granularity of coverage for each prediction instance.

For each new test point, our procedure reports an instance-specific predictive interval and identifies a local region containing the instance $X_{n+1} = x$, over which the procedure is marginally valid, i.e., the inference is reported as follows:

For $X_{n+1} = x$, report $\widehat{C}_n(x), \widehat{\mathcal{S}}_n(x)$ such that:

$$\mathbb{P}\left[Y \in \widehat{C}_n(X) \,\middle|\, X \in \widehat{\mathcal{S}}_n(x)\right] \geq 1 - \alpha, \; \forall x \in \mathbb{R}^d,$$

where $\widehat{\mathcal{S}}_n(x) \subseteq \mathbb{R}^d$, which we call a *relevance subgroup*, is a local region containing x.

In the clinical example discussed earlier, our procedure would communicate uncertainty with an individual patient as follows: "The model predicts that your eGFR will decline by 1.8–5.2 mL/min over the next 3 years. The predictions of the model tend to be accurate 95% of the time for patients *similar* to you defined by the patient

[1] We write "almost all x" to mean that the set of points where the bound fails has measure zero under P_X.

subgroup $\widehat{\mathcal{S}}_n(x)$, but the accuracy will vary from one patient to another within this group". By communicating uncertainty in this more transparent form, the clinician can reason about the relevance of this prediction to the patient at hand by inspecting the reported subgroup, e.g., checking if the relevance subgroup includes patients with different disease phenotypes.

The key idea behind our method is to view localized predictive inference as a marginal CP problem under a "hypothetical" covariate distribution G_x localized around the test point x instead of the true distribution P_X. This is implemented through a commonly used approach in econometrics, known as unconditional quantile regression (UQR), which estimates the marginal quantiles of outcomes within arbitrary local subsets (i.e., relevance subgroups) of the covariate space [84]. It does so by regressing the recentered influence function (RIF) of the quantile functional over covariates, and marginalizing the predicted RIFs within relevance subgroups. This is different from conditional quantile regression [223], for which the regression targets do not recover marginal quantiles when averaged over subgroups.

Our procedure involves two steps. First, we use the UQR model to generate a nested sequence of predictive bands for each relevance subgroup. Next, we select the tightest band that achieves coverage within each subgroup using a held-out calibration set. In the rest of the paper, we explain the two steps of our procedure; we first start by providing a brief background on the standard CP method in the next Section.

5.2 Conformal Prediction

The standard *split* CP procedure relies on sample splitting for constructing predictive intervals that satisfy finite-sample coverage guarantees [278, 277, 196]. Assuming that all data points are exchangeable, the procedure splits the data set into two disjoint subsets: a proper training set $\{(X_i, Y_i) : i \in \mathcal{D}_t\}$, and a *calibration* set $\{(X_i, Y_i) : i \in \mathcal{D}_c\}$. Then, a machine learning model $\widehat{\mu}(x)$ is fit to the training data set $\mathcal{D}_t$, and a *conformity score* $V(.)$ is computed for all samples in $\mathcal{D}_c$—this score measures how unusual the prediction looks relative to previous examples. A typical choice of $V(.)$

is the absolute residual, i.e., $V(x,y) \coloneqq |\,\widehat{\mu}(x) - y\,|$. The conformity scores are evaluated as follows:

$$V_i \coloneqq V(X_i, Y_i) = |\,\widehat{\mu}(X_i) - Y_i\,|, \, \forall i \in \mathcal{D}_c. \tag{5.5}$$

For a given miss-coverage level α, we then compute a quantile of the empirical distribution of the absolute residuals,

$$Q_{\mathcal{V}}(1-\alpha) \coloneqq (1-\alpha)(1 + 1/|\mathcal{D}_c|)\text{-th quantile of } \mathcal{V}, \tag{5.6}$$

where $\mathcal{V} = \{V_i : i \in \mathcal{D}_c\}$. Finally, the prediction interval at a new point $X_{n+1} = x$ is given by

$$\widehat{C}_n(x) = [\,\widehat{\mu}(x) - Q_{\mathcal{V}}(1-\alpha), \, \widehat{\mu}(x) + Q_{\mathcal{V}}(1-\alpha)\,]. \tag{5.7}$$

The CP intervals have a fixed length of $2Q_{\mathcal{V}}(1-\alpha)$, independent of X_{n+1}, which is sufficient for satisfying marginal validity but does not adapt to the varying degrees of uncertainty across different prediction instances.

5.3 Conformalized Unconditional Quantile Regression (CUQR)

In this Section, we describe the two steps involved in our procedure, which we call conformalized unconditional quantile regression (CUQR). Indeed, ours is not the first adaptive variant of CP—we compare our method with existing approaches to adaptive uncertainty quantification in Section 5.4.

5.3.1 Step 1: Unconditional Quantile Regression (UQR)

Consider the true populational counterpart of $\widehat{C}_n$ in (5.4), i.e.,

$$C^*(x) \coloneqq \left[Q\left(\alpha/2, x\right), Q\left(1-\alpha/2, x\right)\right], \tag{5.8}$$

where C^* is the predictive interval given *oracle* knowledge of P, $Q(\alpha, x)$ is the level-α quantile of $Y|X = x$, i.e., $Q(\alpha, x) := \inf\{y \in \mathbb{R} : F(y \,|\, X = x) \geq \alpha\}$, and $F(.)$ is the conditional cumulative density function (CDF), $F(y \,|\, X = x) := \mathbb{P}(Y \leq y \,|\, X = x)$. By definition, the oracle band in (5.8) is conditionally valid in the sense of (5.4). Hence, a sensible guess of an uncertainty band that is both adaptive and transparent can be obtained by directly estimating the conditional quantile $Q(., x)$.

Nested sequence of plug-in estimates.

We use a *plug-in* approach for estimating C^* by replacing the conditional quantile in (5.8) with a consistent estimate $\widehat{Q}$, i.e., $\widehat{C}_n(x) = [\, \widehat{Q}(\alpha/2, x), \widehat{Q}\,(1 - \alpha/2, x)\,]$. While plug-in models can learn accurate estimates of C^*, they do not provide finite-sample coverage guarantees. To take advantage of both the adaptivity of plug-in estimates and the finite-sample coverage of the CP framework, a typical approach is to "conformalize" these plug-in estimates [223]. In what follows, we explain how our procedure creates conformity scores based on plug-in estimates of C^*.

Instead of constructing predictive intervals using a point estimate of $Q(\alpha, .)$ obtained from a single plug-in model $\widehat{Q}(\alpha, .)$, we generate a set of "candidate" estimates of the conditional quantile function and use the calibration set $\mathcal{D}_c$ to pick the narrowest candidate band that achieves the desired coverage. More precisely, we define a set of predictive intervals for covariate x, $\widehat{\mathcal{C}}(x)$, as follows:

$$\widehat{\mathcal{C}}(x) := \{\widehat{\mathcal{C}}_{\tilde{\alpha}}(x) = \widehat{\mu}(x) \pm \widehat{Q}(\tilde{\alpha}, x)\}_{\tilde{\alpha} \in (0,1)}, \tag{5.9}$$

where $\widehat{Q}(\tilde{\alpha}, x)$ is a plug-in estimate of the level-$\tilde{\alpha}$ conditional quantile of the **model residual** at x. We require that the plug-in estimates are monotonic: $\widehat{Q}(\tilde{\alpha}, x) \leq \widehat{Q}(\tilde{\alpha}', x)$ for $\tilde{\alpha} \leq \tilde{\alpha}'$, i.e., no quantile crossing. Thus, $\widehat{\mathcal{C}}(x)$ comprises a nested sequence of candidate intervals, through which we define the following conformity score:

$$V(X_i, Y_i) = \inf\{\tilde{\alpha} \in (0, 1) : Y_i \in \widehat{\mathcal{C}}_{\tilde{\alpha}}(X_i)\}, \tag{5.10}$$

for all $i \in \mathcal{D}_c$. The conformity score in (5.10) checks for the smallest value of $\tilde{\alpha}$ for which the corresponding interval $\widehat{\mathcal{C}}_{\tilde{\alpha}}(X_i)$ in $\widehat{\mathcal{C}}$ covers the response Y_i. We compute the empirical quantile of conformity scores $Q_{\mathcal{V}}(1-\alpha)$ as in (5.6), and construct a predictive interval for $X_{n+1} = x$ as:

$$\widehat{C}_n(x) \coloneqq \{\widehat{\mu}(x) \pm \widehat{Q}(\tilde{\alpha}^*, x)\},\ \tilde{\alpha}^* = Q_{\mathcal{V}}(1-\alpha). \tag{5.11}$$

We drop the dependence of $\widehat{C}_n$ on α to reduce notational clutter. Note that the procedure in (5.11) produces predictive intervals that vary across prediction instances since it picks an entire conditional quantile function from the nested set. The intervals in (5.11) still follow the CP construction: hence, they satisfy the following marginal coverage guarantee.

Proposition 1. *Consider a sequence of plug-in estimates $\{\widehat{Q}(\tilde{\alpha}, .)\}_{\tilde{\alpha}}$ obtained from a sample $\mathcal{D}_{c,1}$, and the corresponding conformity scores $\mathcal{V} = \{V(X_i, Y_i) : i \in \mathcal{D}_{c,2}\}$ obtained from another sample $\mathcal{D}_{c,2}$, where $\mathcal{D}_{c,1}$ and $\mathcal{D}_{c,2}$ are two disjoint subsets of $\mathcal{D}_c$. If $\{(X_i, Y_i) : 1 \leq i \leq n+1\}$ are exchangeable, then the interval in (5.11) satisfies*

$$\mathbb{P}(Y_{n+1} \in \widehat{C}_n(X_{n+1})) \geq 1 - \alpha.$$

Proof is given in Appendix A. Variants of this result appear in the literature [162, 107]. Proposition 1 indicates that, by defining conformity scores over a sequence of bands rather than intervals, we can construct adaptive predictive intervals while retaining the marginal coverage guarantees of CP. **Plug-in estimation via UQR.** We use UQR [84] to fit the nested sequence of plug-in estimates in (5.9). In what follows, we explain how UQR works from a Taylor approximation perspective. UQR approximates the conditional quantiles of $Y \,|\, X = x$, $Q(\alpha, x)$, under the true distribution P as the marginal quantile of Y, $Q(\alpha)$, under an alternative distribution G_x that is "localized" around x. Since the quantile is a statistical functional of the underlying distribution, we can estimate the marginal quantile under G_x given the marginal quantile under P using a von Mises linear approximation (VOM), i.e., a distributional analog of Taylor series of the following form [83]:

$$Q_{G_x}(\alpha) \approx Q_P(\alpha) + \int \mathrm{IF}(y; Q(\alpha), P) \cdot dG_x(y), \tag{5.12}$$

where IF is the influence function of the functional $Q(\alpha)$ at P for a given point (x, y) in the direction of the localized distribution G_x, which is defined as follows:

$$\mathrm{IF}(y; Q(\alpha), P) = \lim_{\epsilon \to 0} \frac{Q_{P^y}(\alpha) - Q_P(\alpha)}{\epsilon}, \tag{5.13}$$

where $P^y = (1 - \epsilon)P + \epsilon\, \delta_y$. The influence function of the quantile measures the contribution of the outcome value y on the marginal quantile statistic $Q_P(\alpha)$. By weighting the contributions of observations sampled from P using the localized density G_x as in (5.12), we obtain a first-order approximation of the marginal quantile functional under G_x. The influence function of the quantile functional is:

$$\mathrm{IF}(y; Q(\alpha), P) = \frac{\alpha - \mathbf{1}\{y \leq Q_P(\alpha)\}}{f_Y(Q_P(\alpha))}. \tag{5.14}$$

Here, $f_Y(.)$ is the (one-dimensional) marginal density of Y. The derivation of the formula in (5.14) is standard, and is provided in Appendix B for completeness. The *re-centered* influence function (RIF) is defined as:

$$\mathrm{RIF}(y; Q(\alpha), P) = Q_P(\alpha) + \mathrm{IF}(y; Q(\alpha), P), \tag{5.15}$$

UQR involves regressing RIF over X to obtain a model for $\mathbb{E}[\,\mathrm{RIF}(Y; Q(\alpha), P) \,|\, X = x\,]$. Note that the influence function in (5.14) is a dichotomous variable as $\mathbf{1}\{y \leq Q_P(\alpha)\}$ is the only term that changes across covariates. Thus, UQR involves fitting a one-dimensional density estimate for f_Y and a binary classifier for the dichotomous variable. UQR is typically used to study the effect of changing the covariate distribution on the marginal quantiles of outcomes, e.g., the effect of unionization on wages [210]. In our setup, we use (5.12) to obtain a plug-in estimate for the predictive band at the test point x as follows:

$$Q_{G_x}(\alpha) \approx \int \mathbb{E}[\,\mathrm{RIF}(Y; Q(\alpha), P) \,|\, X = x\,] \cdot dG_x. \tag{5.16}$$

Constructing the nested sequence using UQR. The approximation in (5.16) inspires a simple regression procedure for constructing the nested sequence in (5.9) while avoiding quantile crossing. Let $\mathrm{RIF}_\alpha(X_i)$ be the RIF of the level-α quantile associated with the i-th data point. For a test point $X_{n+1} = x$, we can predict the value of $Q_{G_x}(\alpha)$ by fitting an ML model on the data set $\{(X_i, \mathrm{RIF}_\alpha(X_i))\}_{i \in \mathcal{D}_c}$. Let the RIF values predicted by the ML model be $\widehat{\mathrm{RIF}}_\alpha(x)$. Then, by repeating this process $K > 0$ times for all values of $\tilde{\alpha}$ in $\tilde{\alpha} = [1/K, \ldots, (K-1)/K]$, we can construct the nested sequence as follows:[2]

$$\widehat{\mathcal{C}}(x) = \{\widehat{\mathcal{C}}_k(x) = \widehat{\mu}(x) \pm |\widehat{\mathrm{RIF}}_{\tilde{\alpha}_k}(x)|\}_{k=1}^{K-1}. \tag{5.17}$$

where $\tilde{\alpha}_k = 1/k$. Here, the RIF is defined with respect to quantiles of the model residual $E = |\,\widehat{\mu}(X) - Y|$ rather than the outcome Y. Note that, combining (5.14) and (5.15), the RIF for the level $\tilde{\alpha}_k$ quantile can be written as:

$$\mathrm{RIF}_{\tilde{\alpha}_k}(x) = Q_P(\tilde{\alpha}_k) + \frac{\tilde{\alpha}_k - \mathbf{1}\{e \leq Q_P(\tilde{\alpha}_k)\}}{f_E(Q_P(\tilde{\alpha}_k))}. \tag{5.18}$$

The 1-D density $f_E(.)$ can be estimated using kernel density estimation (KDE), and $Q_P(\alpha)$ can be estimated as the empirical level-$\tilde{\alpha}_k$ quantile of the residuals. The only term that we need to predict for each test point is $\mathbf{1}\{e \leq Q_P(\tilde{\alpha}_k)\}, \forall k \in \{1, \ldots, K-1\}$. This can be achieved with a single ML model as follows: for each data point (X_i, E_i), define a target $k_i^* := \min k$, s.t. $E_i \leq \widehat{Q}_P(\tilde{\alpha}_k)$, then fit a model $g_\theta(.)$ (e.g., a regression model or a multi-class classifier) on the data set $\{(X_i, k_i^*)\}_i$. For a new test point $X_{n+1} = x$, we predict the RIF at $X_{n+1} = x$ by plugging in the predictions of g_θ into (5.18) as follows:

$$\widehat{\mathrm{RIF}}_{\tilde{\alpha}_k}(x) = \widehat{Q}_P(\tilde{\alpha}_k) + \frac{\tilde{\alpha}_k - \mathbf{1}\{g_\theta(x) \leq k\}}{\widehat{f}_E(\widehat{Q}_P(\tilde{\alpha}_k))}, \tag{5.19}$$

$\forall k \in \{1, \ldots, K-1\}$. The nested set in (5.17) can thus be constructed using the $K-1$ predictions in (5.19). Note that for any $k < k'$, it is sufficient that $\partial \widehat{f}_E(\widehat{F}_E^{-1}(\alpha))/\partial \alpha < 0$ for the monotonicity of the nested intervals to be preserved, i.e., $\widehat{\mathrm{RIF}}_{\tilde{\alpha}_k}(x) < \widehat{\mathrm{RIF}}_{\tilde{\alpha}_{k'}}(x)$. This condition is met by various typical probability distributions (i.e., the exponential

[2] We take the absolute value of $\widehat{\mathrm{RIF}}$ to account for erroneously negative predictions.

distribution). When this condition is violated for any two consecutive intervals in our empirical estimate, we replace $\widehat{f}_E(\widehat{Q}_P(\tilde{\alpha}_k))$ with $\widehat{f}_E(\widehat{Q}_P(\tilde{\alpha}_{k-1}))$ to enforce monotonicity.

Alternative approaches to constructing the set $\{\widehat{Q}(\alpha,.)\}_\alpha$ include fitting K independent quantile regression models or a single distributional model (e.g., a Bayesian nonparameteric regression [53]). The RIF-based construction of the nested set $\{\widehat{Q}(\alpha,.)\}_\alpha$ is more computationally and statistically efficient than either approach as it requires training a single ML model with labels that condense information about the conditional quantiles at levels $[1/K,\ldots,(K-1)/K]$.

Algorithm 1 Conformalized UQR

```
Input: Dataset $\{(X_i,Y_i)\}_{i=1}^n$, test covariate $X_{n+1}$, model $\mu$, parameters $\alpha$, $G$ and $K$.
Output: $\widehat{C}_n(X_{n+1})$ and $\widehat{\mathcal{S}}_n(X_{n+1})$
Split the data set into a proper training set $\mathcal{D}_t$ and 2 disjoint calibration sets $\mathcal{D}_c = \mathcal{D}_{c,1} \cup \mathcal{D}_{c,2}$
Using the training set $\mathcal{D}_t$, do the following:
    Fit the predictive model $\widehat{\mu}: \mathcal{X} \to \mathbb{R}$ using $\mathcal{D}_t$
    Partition $\mathcal{X}$ into relevance subgroups $\{\widehat{\mathcal{S}}_g\}_{g=1}^G$
Using the calibration set $\mathcal{D}_{c,1}$, do the following:
    Fit $\widehat{f}_E(.)$, $\widehat{Q}_P(\tilde{\alpha}_k)$ & $g_\theta$. Plug the estimates in (5.19)
    Construct a nested sequence of intervals $\widehat{\mathcal{C}}(x) = \{\widehat{\mathcal{C}}_k(x) = \widehat{\mu}(x) \pm |\widehat{\mathrm{RIF}}_{\tilde{\alpha}_k}(x)|\}_{k=1}^K$
Using the calibration set $\mathcal{D}_{c,2}$, do the following:
    For all calibration data within each subgroup in $\{\widehat{\mathcal{S}}_g\}_{g=1}^G$, compute the conformity scores in (5.10)
    For each subgroup, select the tightest band in $\widehat{\mathcal{C}}$ that achieves the target coverage as in (5.11). Let $\widehat{\mathcal{C}}_{k(g)}(x)$ be the band selected for the $g$-th subgroup
For a new test covariate $X_{n+1}$, do the following:
    Identify subgroup $g_{n+1}$ for which $X_{n+1} \in \widehat{\mathcal{S}}_{g_{n+1}}$
    $\widehat{C}_n(X_{n+1}) \leftarrow \widehat{\mathcal{C}}_{k(g_{n+1})}(X_{n+1})$, $\widehat{\mathcal{S}}_n(X_{n+1}) \leftarrow \widehat{\mathcal{S}}_{g_{n+1}}$
Return $\widehat{C}_n(X_{n+1})$ and $\widehat{\mathcal{S}}_n(X_{n+1})$.
```

5.3.2 Step 2: Conformalizing UQR within subgroups

In Section 5.3.1, we developed a procedure that fulfills the adaptivity requirement while retaining the marginal validity of the non-adaptive CP method (Proposition 1).

These marginal guarantees, however, do not meet the criteria for being transparent as they do not reflect the accuracy of the issued intervals for any given prediction. To provide more transparent guarantees, the second step selects an interval from the nested set by applying the conformal procedure within local regions in covariate-space as follows:

- Partition the covariate space into G subsets $\{\widehat{\mathcal{S}}_g\}_{g=1}^{G}$.
- Apply the conformal procedure in Section 5.3.1 locally within each subgroup g by picking a subgroup-specific band $\widehat{\mathcal{C}}_{k(g)}(x)$ from the set $\{\widehat{\mathcal{C}}_k(x)\}_{k=1}^{K-1}$.
- For a test point $X_{n+1} = x$, identify the relevance subgroup g_{n+1} in which it belongs and report the subgroup $\widehat{\mathcal{S}}_{g_{n+1}}$ and the corresponding interval $\widehat{\mathcal{C}}_{k(g_{n+1})}(X_{n+1})$.

The steps involved in our conformalized UQR (CUQR) procedure are given in Algorithm 1. The procedure reports both a predictive interval that is specific to each instance, and a subgroup for which a desired level of accuracy is achieved. The relevance subgroups can either be learned from training data (e.g., using a clustering algorithm) or predetermined using application-specific knowledge (e.g., disease subtypes or protected attributes). The subgroup can be reported in the form of a cluster with a "representative" covariate value that is typical for this subgroup. In the clinical example discussed earlier, our algorithm's output can represent the relevance subgroup in terms of a set of "representative" patients.

Increasing the number of subgroups G increases the granularity of the achieved average coverage at the cost of higher variance in realized coverage. G can be set to larger values for larger sample sizes, e.g., $G = O(n)$, so that conditional coverage is achieved asymptotically. In the other extreme case when $G = 1$, we recover the standard marginal coverage guarantee.

5.4 Related Work

Distributional regression models that directly estimate the conditional density $P_{Y|X=x}$ provide adaptive estimates of uncertainty. Broad classes of methods fall under this category; a non-exhaustive list includes: Bayesian non-parametric regression (e.g., using Gaussan processes [250] or regression trees [53]), Bayesian neural nets [143, 116, 219], and deep ensembles [153, 86]. Many of these models can provide accurate pragmatic estimates of predictive variance, but without the finite-sample coverage guarantees enabled by CP. The achieved coverage of distributional regression can be very sensitive to modeling choices (e.g., hyper-parameters or prior distributions). For instance, in Bayesian regression, the frequentist coverage achieved by posterior credible intervals with exact inference depends on the choice of the prior [23]. With the more commonly used approximate inference methods (e.g., dropout or variational inference [91]), the induced posterior distributions may not concentrate asymptotically, resulting in poor coverage behavior [192, 119]. Distributional models are often used in conjunction with CP approaches to satisfy finite-sample coverage while maintaining adaptivity [51].

Conformal prediction in its most basic form achieves finite-sample marginal coverage at the expense of adaptivity [278, 277, 196]. Various approaches to CP-based adaptive predictive inference have been recently proposed [162, 223, 103, 82, 245, 107, 104, 88, 17]. The idea of "conformalizing" a plug-in estimate of the conditional quantile function originated in [223]. In this work, a single quantile regression model $\widehat{Q}(\alpha, x)$ is fit to the training data, and a conformity score that measures the accuracy of $\widehat{Q}(\alpha, x)$ is used to derive an adjustment for these intervals. Conformalized quantile regression provides marginal coverage guarantees, but its empirical conditional coverage depends on the quality of the underlying quantile regression model. Refinements of this approach were later proposed through two different lines of work: the first uses a re-weighting technique to "localize" CP at new test points [103, 104], and the second conformlizes a distributional regression model from which conditional quantiles can be derived [245, 51]. In both lines of work, conditional validity is achieved in an asymptotic sense.

Our work holds subtle connections to these two approaches. In terms of the construction of conformity scores, [245, 51] define the scores based on conditional ranks rather than error residuals—our procedure constructs predictive intervals by selecting among "candidate" estimates of conditional quantiles generated by varying the quantile level α. This is equivalent to constructing the predictive intervals by selecting among estimates of conditional ranks of the full conditional distribution $P_{Y|X}$, but unlike the procedures in [245, 51], ours does not require access to consistent estimates of $P_{Y|X}$. Similar to the localized CP methods in [103, 104], our procedure is effectively a localized version of the marginal CP method. But unlike localized CP (LCP), our procedure can utilize any ML model for obtaining the localized quantile estimates (i.e., Equation (5.19)), whereas LCP is limited to re-weighting estimators (e.g., based on Nadarya-Watson kernel). Additionally, because our procedure selects among multiple quantile functions within each subgroup, it can achieve finite-sample (rather than asymptotic) coverage within local regions. The nested construction of our plug-in estimates falls within the general nested CP formulation developed in [107]. Unlike the formulation in this work, we construct our nested sets by parametrizing a functional form for the predictive band rather than directly parametrizing intervals, which enables selecting a different interval for each instance within a subgroup.

Re-centered influence functions (RIF) are typically used as targets for unconditional quantile regression models, a common modeling tool in econometric studies [84]. To the best of our knowledge, RIF- based regression has not been operationalized as a conformity score prior to this work.

5.5 Experiments

We compare CUQR with various conformal and quantile regression baselines across multiple benchmark data sets. We start by describing our experimental setup below.

Baselines. We consider standard split *conformal prediction* (CP), the *locally adaptive CP* (LACP) method in [197], *conformalized quantile regression (CQR)* [223], and *conformal conditional histograms* (CCH) [245]. We also consider two variants of the standard *quantile regression* (QR) model for estimating conditional quantiles: QR with an underlying random forest model (QR-RF), and QR implemented using a neural network (QR-NN). We consider two ablated versions of CUQR that apply a conformalization procedure within the same relevance subgroups of CUQR. The first baseline, dubbed CQ, constructs the nested set $\mathcal{C}(.)$ using the empirical quantiles of residuals within the relevance subgroups, assigning the same intervals to all units within a subgroup without using the UQR-based plug-in estimates. The second baseline, which we call CQR-S, applies the adaptive CQR method [223] within the relevance subgroups.

Finally, we consider two variants of our method: CUQR which applies conformalization based on the empirical $(1-\alpha)$-th quantiles, and CUQR-PAC, which corrects for the slack term in Theorem 1 to provide a high probability (PAC-style) coverage per subgroup. We select the value of λ in Theorem 1 so that the probability that coverage holds (conditional on training data) is 90%, i.e., $1 - 2\exp(-2\lambda^2) = 0.9$. We implement our method using an XGBoost regression model for g_θ. In all experiments, we create the relevance subgroups using the K-means clustering algorithms. Further experimental details are provided in Appendix D.

Evaluation metrics. We evaluate all baselines with respect to their achieved marginal coverage C_{av}, efficiency quantified via average interval length L_{av} and subgroup-level coverage denoted as $C_{av}(\widehat{\mathcal{S}}_g)$ for all subgroups $\{\widehat{\mathcal{S}}_g\}_{g=1}^{G}$. We also evaluate the worst-case subgroup-level coverage, defined as $C_G^{w.c.} = \min_g C_{av}(\widehat{S}_g)$. All metrics are evaluated on testing data and averaged over 10 runs. Unless otherwise stated, we set the target coverage level to $1 - \alpha = 0.9$.

Data sets. We evaluate all baselines on **9 benchmark data sets** that are commonly used to evaluate CP methods: MEPS-19, MEPS-20, MEPS-21, Facebook-1, Facebook-2, Bio, Kin8nm, Naval, and Blog [82, 223, 82, 57]. Due to space limitations, we highlight results for four data sets (MEPS-19, Facebook-1, Blog and Kin8nm) in this Section and defer further results to the Appendix. Details of all data sets are provided in the Appendix. For each run, we randomly split each data set into disjoint training $\mathcal{D}_t$ (42.5%), calibration $\mathcal{D}_c$ (42.5%) and testing $\mathcal{D}_{test}$ (15%) samples. For the LACP, CCH, CQR, CQR-S, CQ and CUQR baselines, we further split the calibration set $\mathcal{D}_c$ in half to obtain plug-in estimates and conformity scores from different splits. In all experiments, we fit a Gradient Boosting regression model $\widehat{\mu}$

	MEPS-19			**Facebook-1**			**Blog**			**Kin8nm**		
	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$
QR methods												
QR-RF	0.90	1.00	0.54	0.93	0.85	0.78	0.79	0.73	0.76	0.93	1.36	0.89
QR-NN	0.79	0.54	0.67	0.81	0.55	0.68	0.79	0.73	0.76	0.79	0.94	0.74
CP methods												
CP	0.89	1.28	0.19	0.90	1.39	0.72	0.89	1.89	0.57	0.90	2.17	0.83
LACP	0.89	0.61	0.20	0.90	0.69	0.76	0.89	1.06	0.63	0.90	1.09	0.84
CQR	0.89	1.12	0.46	0.90	0.83	0.77	0.90	1.34	0.82	0.90	1.33	0.85
CCH	0.96	5.37	0.79	0.89	0.72	0.65	0.98	5.58	0.96	0.89	1.14	0.86
CQ	0.87	2.02	0.76	0.89	1.34	0.79	0.87	1.81	0.76	0.89	2.16	0.85
CQR-S	0.89	1.54	0.67	0.90	0.77	0.87	0.90	1.37	0.80	0.90	1.33	0.85
CUQR	0.89	1.25	0.73	0.90	1.36	0.87	0.87	1.82	0.67	0.89	2.19	0.85
CUQR-PAC	0.89	2.90	0.88	0.92	1.61	0.90	0.90	2.23	0.82	0.96	3.27	0.93

Table 5.1: Marginal coverage, efficiency and conditional coverage of all baselines on benchmark data sets.

using the training set $\mathcal{D}_t$ and apply the predictive inference baselines on top of the predictions issued by the model $\widehat{\mu}$.

5.5.1 Results

Evaluating transparency. All baselines are calibrated to have a target marginal coverage of 90%, but what does this notion of coverage mean to individual users of the model? In this experiment, we assess the transparency of different baselines by evaluating the worst-case conditional coverage within typical "subgroups" of individuals or prediction instances. We use K-means clustering to identify $G = 10$ relevance subgroups using training samples. Note that the subgroups are not arbitrary—they represent a clustering of the population into "typical" subgroups of similar individuals.

In Table 1, we show the marginal coverage and average lengths of predictive intervals for all baselines across the three data sets. First, we observe that while all conformal methods achieve the target (marginal) coverage levels, the coverage of the QR baselines vary depending on the underlying model specification, which is expected as these baselines do not provide any coverage guarantees. On the contrary,

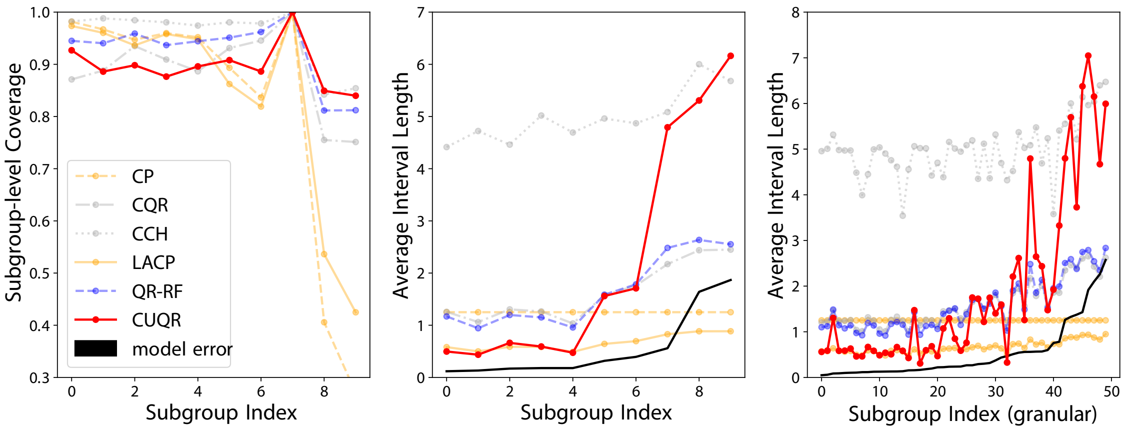


Figure 5-2: Adaptivity of predictive inference baselines on the **MEPS-19** dataset.

all CP-based methods achieve their promised model-agnostic coverage guarantees, but how well do the different CP variants perform when examined on a subgroup level?

As we can see in Table 1, CP baselines that only guarantee marginal coverage (CP and CQR) have a very poor worst-case coverage conditional on a subgroup, i.e., their declared guarantees do not reflect their performance among a large subgroup of "similar" individuals in a given population. Similary, CP methods with asymptotic conditional coverage guarantees can exhibit severe under-coverage in finite samples (LACP), or maintain reasonable conditional coverage but with poor efficiency (CCH), which highlights the importance of controlling for finite-sample conditional coverage. While CUQR achieves subgroup-specific coverage marginally, the worst-case subgroup-level coverage can be significantly lower than the desired target coverage for some data sets (e.g., MEPS-19 and Blog).

The CQ variant of our method, which picks a fixed interval per subgroup rather than a full RIF-based predictive band, achieves better worst-case coverage at the expense of within-subgroup adaptivity and average efficiency. Because the subgroup-specific coverage guarantees for CUQR hold on average with respect to the randomness of calibration data, the variance of the empirically achieved coverage on test data increases as the number of subgroups G increases (i.e., smaller calibration sample per subgroup). Consequently, the worst case subgroup-level coverage achieved by CUQR decreases (in expectation) as the number of relevance subgroups increases.

Evaluating adaptivity. Next, we assess the extent to which baselines are adaptive, i.e., the lengths of their intervals vary according to the true uncertainty of the base model $\widehat{\mu}$. Among the data sets under study, the MEPS-19 data exhibited significant heteroscedasticity, i.e., the average error of the predictive model varies significantly across the subgroups. Hence, we expect the average lengths of intervals issued by adaptive procedures to be greater for subgroups where the model errors are high. In Figure 5-2, we plot the achieved subgroup-level coverage (left) and the corresponding average interval length per subgroup (middle) in the **MEPS-19** dataset. In both Figures, the subgroup indexes on the x-axis are ordered ascendingly according to the model's subgroup-level average error on testing data (i.e., larger indexes correspond to higher model uncertainty). (In Figure 5-2, we exclude baselines that were under-performing to avoid clutter.) As we can see, CUQR maintains the target coverage approximately for all subgroups, and adjusts the lengths of its issued intervals withing each subgroup according to the model uncertainty. On the contrary, competing baselines either fail to recognize the varying uncertainty across subgroups (LACP and CP), or do not adequately adapt the interval length to maintain target coverage (QR-RF and CQR). Finally, to evaluate the adaptivity of CUQR beyond the subgroups on which it was calibrated, we run a K-means clustering algorithm with a different random seed and different number of clusters $G = 50$, and order the subgroup indexes ascendingly according to the base model's uncertainty as before. We evaluate the average interval lengths of CUQR (previously fitted on the $G = 10$ subgroups), along with the other baselines, on the new and more granular 50 subgroups. As we can see in Figure 5-2 (right), CUQR outperforms other baselines in adapting its intervals to subgroup-level uncertainty, indicating better conditional adaptivity properties beyond what is implied by the theoretical guarantees.

5.6 Conclusion

We developed a conformal prediction method that adapts its issued intervals to the level of uncertainty in each prediction instance, while reporting a local region in covariate-space that contains the queried covariate instance and over which the procedure is guaranteed to be accurate on average. Our procedure partitions the covariate space into subgroups, and leverages the re-centered influence function of the quantile functional to construct a nested sequence of predictive bands, from which it selects one band per subgroup. By reporting instance-specific predictive intervals and subgroup-specific coverage guarantees to end-users, our method enables a more transparent approach to communicating uncertainty in the predictions of ML models.

One important direction for future work is to extend our method to give valid intervals that are not just valid in a pointwise sense, but can be uniformly valid over e.g. the entire survival curve, which is more directly applicable in our vision of a CDSS for oncology. Furthermore, more clever ways of dealing with censored data, beyond simply utilizing our approach on the uncensored patients, would also be an interesting direction for future work. Possible avenues include only excluding censored patients whose censored times are below some threshold, with the intuition being that patients with large censored times are close to experiencing the event of interest and can be included in the analysis. However, the covariate shift induced by excluding such patients would have to be accounted for, as is done in [42, 106]. Finally, a straightforward extension of our method that could handle censored data is to use a Cox regression model as the underlying regression model trained in the first step of the procedure.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

Falsification: Assessing Reliability of Causal Estimates with Experimental Data

Acknowledgement of Co-authors For this chapter and the next chapter, I would like to acknowledge Ming-Chieh Shih and Michael Oberst, who were integral in developing the underlying theory for this framework, and additionally for their immense guidance and mentorship. I would also like to acknowledge Ilker Demirel, who assisted with the writing and theoretical framing of the next chapter.

In the previous chapter, I explored how to provide uncertainty quantification of these predictions using conformal inference, giving one assessment of the "reliability" of these predictions. In this chapter, I further explore the theme of "reliability" by introducing the principle of falsification, i.e. using RCT (experimental) data to validate estimates inferred from observational studies whereby an observational study is "falsified" if it fails to replicate the results of a corresponding RCT. Whereas the prior chapters took a predictive lens, I use tools in causal inference to build out the framework of falsification in this chapter. I introduce falsification as a paradigm of building trust in causal estimates from an observational study, particularly in the setting where the study can provide causal effects of subgroups for whom we do not have RCTs. In Chapter 7, I delve more deeply into developing more powerful falsification methods via the framework of conditional moment restrictions. This framework can serve as an additional "sanity check" of causal estimates that will be presented in the running example of the CDSS in this thesis.

6.1 Introduction

Policy guidelines often rely on conclusions from Randomized Controlled Trials (RCTs), whether considering treatment decisions in healthcare, classroom interventions in education, or social programs in economics [140, 58, 212]. In healthcare, when a target population has reasonable overlap with the inclusion criteria of RCTs, current clinical treatment guidelines rely primarily on RCTs [109, 108]. For target populations not well-represented in RCTs, observational studies are often used to infer treatment effects. However, different observational estimates can give conflicting conclusions. We give an example of this tension when looking at a new chemotherapy for multiple myeloma.

Example 6.1.1 (*Carfilzomib-based Therapy for Multiple Myeloma*)**.** Until 2020, the effect of Carfilzomib-based combination therapy in the NDMM subpopulation had not been studied via an RCT. However, a trial (ASPIRE) in 2015 measured the effect of Carfilzomib-based therapy on survival in Relapsed & Refractory Multiple Myeloma (RRMM) patients [257]. The CoMMpass trial, an observational dataset, was also available in which the Carfilzomib regimen was given to both NDMM and RRMM patients [188]. Several analyses on the CoMMpass dataset to estimate the effect of Carfilzomib-based therapy on NDMM patients led to different, sometimes opposing, conclusions on the benefit of the therapy in this subpopulation [164, 155].

A traditional meta-analysis approach would combine observational estimates under the assumption that differences arise only due to random variation, and not e.g., differences in confounding bias [117, Section 10.10.4.1]. This is unlikely to be true in practice. For instance, in Example 6.1.1, the two studies in question made different choices in e.g., how to adjust for confounders. *In this paper, we relax the* *assumption that all observational estimates are valid.* Instead, we assume that at least one observational estimate is valid across all subpopulations. In the context of Example 6.1.1, we might assume that at least one of the candidate observational studies yields consistent and asymptotically normal estimates of the effects in both the NDMM and RRMM populations. While we cannot *verify* that any given estimator is valid for all subpopulations, we can *falsify* this claim of validity if an estimator is inconsistent for the causal effects identified by the RCT (e.g., RRMM). Hence, we use the term *validation effects* to refer to causal effects in subpopulations that overlap between the observational and randomized datasets (e.g., RRMM), and use the term *extrapolated effects* to refer to those only covered by observational datasets (e.g., NDMM).

We propose a meta-algorithm that combines two key ideas: falsification of estimators, and pessimistic combination of confidence intervals. We first aim to falsify candidate estimators using hypothesis testing, rejecting those that fail to replicate the RCT estimates of validation effects.[1] In Section 6.2.2, we motivate this approach with examples of observational estimates based on different causal assumptions, showing that hypothesis tests based on asymptotic normality can be applied even when causal assumptions fail to hold. Then, we combine accepted estimators to get confidence intervals on the extrapolated effects. Since failure to reject does not imply validity,[2] we return an interval that contains every confidence interval of the accepted estimators. We demonstrate theoretically that if *at least one* candidate estimator is consistent for both the validation and extrapolated effects, then the intervals returned by our algorithm provide valid asymptotic coverage of the true effects.

In scenarios where the covariate distribution differs across datasets, estimators that "transport" the causal effect should be used [204, 61, 64]. Furthermore, in the case of high-dimensional covariates, flexible machine learning methods are required to estimate nuisance functions, which can affect the hypothesis tests due to their slower

[1] These can be the average treatment effect for the entire population, or multiple group average treatment effects for stratified subgroups, as frequently reported in clinical trials, e.g. stratification by sex and age in [257].

[2] For instance, we could fail to reject due to low power, or because falsification is impossible, due to differences in causal structure across subpopulations, as discussed in Appendix D.1.

convergence rates. In light of this, we adapt estimators of the average treatment effect in this setting to provide estimates of group-wise treatment effects, and show (via the framework of double machine learning [50, 241]) that this estimator enjoys asymptotic normality under mild conditions on convergence rates of the nuisance function estimators. Our conclusions are supported by semi-synthetic experiments, based on the IHDP dataset, as well as real-world experiments, based on clinical trial and observational data from the Women's Health Initiative (WHI), that demonstrate various characteristics of our meta-algorithm.

6.2 Setup and Motivating Examples

6.2.1 Notation and Assumptions

Let $Y \in \mathcal{Y}$ denote an outcome of interest, and $A \in \{0, 1\}$ denote a binary treatment. We use Y_a to denote the potential outcome of an individual under treatment $A = a$. We use $X \in \mathcal{X}$ to denote all other covariates. To distinguish between different sampling distributions (i.e., datasets), we use the random variable $D \in \{0, \dots J\}$, where $J \geq 1$ is the number of observational datasets, and $D = 0$ is reserved for the sampling distribution of the randomized trial. We let $\mathbb{P}(Y_1, Y_0, Y, A, X, D)$ denote the joint distribution over all variables, including unobserved potential outcomes. For instance, $\mathbb{P}(Y_1, Y_0, X \mid D = 0)$ denotes the distribution of potential outcomes and covariates in the RCT.

We seek to estimate conditional average treatment effects for a finite set of I subgroups $\{\mathcal{G}_i\}_{i=1}^{I}$. We assume subgroups are defined a-priori by a function $G : \mathcal{X} \mapsto \{1, \dots, I\}$, such that $G = i$ indicates that $X \in \mathcal{G}_i$. We use observational data precisely because not all groups are supported on the RCT dataset. To this end, we use $\mathcal{I}_R = \{i : \mathbb{P}(G = i \mid D = 0) > 0\}$ to denote the set of subgroups supported on the RCT dataset, and we let $\mathcal{I}_O$ denote the complement $\{1, \dots, I\} \setminus \mathcal{I}_R$. We use $|\mathcal{I}_R|$ to denote the cardinality of a set, and assume that every observational dataset has support for all groups.

Assumption 6.2.1 (Support)**.** We assume that $\mathbb{P}(G = i, D = j) > 0$ for all $i \in \{1, \dots, I\}$ and $j \in \{1, \dots, J\}$, i.e., all observational datasets ($D \geq 1$) have support for all groups.

Definition 6.2.1 (Validation and Extrapolated Effects)**.** We define the group average treatment effect (GATE)[3] as

$$\tau_i := \begin{cases} \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 0], & \text{if } i \in \mathcal{I}_R \\ \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 1], & \text{if } i \in \mathcal{I}_O \end{cases} \tag{6.1}$$

and refer to τ_i for $i \in \mathcal{I}_R$ as a validation effect, and τ_i for $i \in \mathcal{I}_O$ as an extrapolated effect.

Here, we focus on discrete subgroups, in part to reflect the practical reality of comparing RCTs to observational studies, where we may have large observational datasets with rich covariates but only have access to the published results of the RCT, which often provides estimates (with confidence intervals) for subgroup effects but not the raw data itself [254, Figure 4, for example]. In Def. 6.2.1, we allow for the fact that different datasets may have different distributions of effect modifiers. To have a well-defined effect of interest, we have chosen the reference dataset $D = 1$ arbitrarily, but in principle we could choose any of the observational datasets. We discuss further nuances of this definition under Assumption 6.2.3. By Def. 6.2.1, we often write these effects as a vector $\tau \in \mathbb{R}^I$. We use $\hat{\tau}(k) \in \mathbb{R}^I$ to denote an estimator, where $k \in \{0, \dots, K\}$, with $\hat{\tau}(0)$ reserved to denote the estimator derived from the RCT data. The remainder are observational estimators.[4] In general, we use "hat" notation to refer to estimators, and refer to their population quantities without a hat. We use N_k to denote the number of samples used by each estimator. Throughout, we will assume that the RCT estimator is consistent.

[3] We use this term in line with the literature [49, 128, 199, 241] and to distinguish it from the CATE function.

[4] We define $\hat{\tau}(0)$ as a vector in $\mathbb{R}^I$ for simplicity of notation, allowing the entries $\hat{\tau}_i(0), i \in \mathcal{I}_O$ to be arbitrary.

Assumption 6.2.2. The RCT estimator $\hat{\tau}(0)$ is a consistent estimator of the (supported) dimensions of τ, such that for each $i \in \mathcal{I}_R$, $\hat{\tau}_i(0)$ is consistent for τ_i.

Below, our central assumption states that at least one observational estimator also enjoys consistency. We discuss examples of specific observational estimators in Section 6.2.2.

Assumption 6.2.3. There exists at least one observational estimator $\hat{\tau}(k) \in \mathbb{R}^I$, $k \geq 1$ that is a consistent estimator of $\tau \in \mathbb{R}^I$, such that for each $i \in \{1, \ldots, I\}$, $\hat{\tau}_i(k)$ is consistent for τ_i.

Remark 6.2.1. Assumption 6.2.3 is our primary non-trivial assumption, and in Appendix D.2, we give one example of causal assumptions (for a given observational study) under which the entire GATE vector τ is **identifiable** from observational data, and give an **estimator** of the resulting observational quantity which is asymptotically normal [203, 204, 202, 64, 66]. In order to compare observational estimates with experimental ones, Assumption 6.2.3 requires not only that the observational data is free of confounding, but also that the causal effect can be transported to the RCT population. This can be done so long as relevant effect modifiers are observed in both the RCT and observational study, but the latter requirement is satisfied automatically (without requiring RCT data) if e.g., treatment effects are constant within each subgroup G, or if the distribution of effect modifiers is the same between the RCT and observational study, in which case $\mathbb{E}[Y_1 - Y_0 \mid D, G] = \mathbb{E}[Y_1 - Y_0 \mid G]$. This represents one (conservative) failure mode of our approach, in which we may reject an observational estimator due to failures in transportability, even if it yields unbiased estimates of the extrapolated effects. Additionally, as we note in Appendix D.1, adversarial cases exist where an observational estimator produces correct estimates of the validation effects, but yields incorrect estimates of the extrapolated effects, an inevitable limitation of any approach that seeks to use RCT data to validate observational effects.

Assumptions 6.2.2 and 6.2.3 imply that there exists an observational estimator $\hat{\tau}(k)$ such that both $\hat{\tau}_i(k)$ and the RCT estimate $\hat{\tau}_i(0)$ are both consistent for the validation effects τ_i, $\forall i \in \mathcal{I}_R$. To validate this implication in finite samples, we will construct a statistical test to compare $\hat{\tau}_i(k)$ and $\hat{\tau}_i(0)$. Our general approach could be modified to use any valid test, but to facilitate further analysis, as well as explicit construction of confidence intervals, we additionally assume the following:

Assumption 6.2.4. All GATE estimators are pointwise[5] asymptotically normally distributed.

$$\sqrt{N_k}(\hat{\tau}_i(k) - \tau_i(k))/\hat{\sigma}_i(k) \xrightarrow{d} \mathcal{N}(0,1) \tag{6.2}$$

for all $k \in \{0,, ..., K\}$, and for all i in $\mathcal{I}_R$ if $k = 0$ (the RCT estimator), and otherwise for all $i \in \mathcal{I}_R \cup \mathcal{I}_O$. Here, $\xrightarrow{d}$ denotes convergence in distribution, and $\hat{\sigma}_i^2(k)$ is an estimate of the variance that converges in probability to $\sigma_i^2(k)$, the asymptotic variance of $\sqrt{N_k}(\hat{\tau}_i(k) - \tau_i(k))$.

Assumption E.7.1 requires each estimator $\hat{\tau}(k)$ to be consistent and asymptotically normal for some $\tau(k)$, which may **not** be equal to τ. This is not a particularly strong assumption, as we discuss below.

6.2.2 Asymptotic Normality of Biased Estimators

In this section, we give two simple examples to illustrate the principle that multiple estimators $\hat{\tau}(k)$ may be asymptotically normal, even if they are asymptotically biased (i.e., $\tau(k) \neq \tau$). In both cases, there is a distinction between the *statistical* assumptions required to obtain asymptotic normality, and the *causal* assumptions required for $\tau(k)$ to identify the causal effect τ. For simplicity in both examples, we restrict to the setting of comparing one-dimensional estimates $\tau(k) \in \mathbb{R}$, which estimate the GATE, τ, in a single group $G = 1$ covered by all datasets. The statistical claims here also extend to GATE estimation with multiple groups [241].

Example 6.2.1 (Variation in confounding across datasets)**.** Suppose that there is one estimator of the GATE per observational dataset, and each estimator seeks to estimate the population quantity, $\tau(k) = \mathbb{E}[g_k(1, X_k) - g_k(0, X_k) \mid G = 1, D = k]$, where X_k denotes the controls used in each study, and $g_k(A, X_k) \coloneqq \mathbb{E}[Y \mid A, X_k, D = k]$ and

[5] Here, "pointwise" refers to the fact that each subgroup effect estimate is asymptotically normal.

$m_k(X_k) \coloneqq \mathbb{P}(A = 1 \mid X_k, D = k)$. We assume that $\eta < m_k(x) < 1 - \eta$ for some $\eta > 0$ for all x, k. Note that $\tau(k)$ is only a *statistical* quantity: identifying this with the *causal* quantity (the GATE) requires additional assumptions like unconfoundedness, that $Y_a \perp\!\!\!\perp A \mid X_k$ for the given dataset D. This assumption may hold for some datasets, but not others, particularly if the set of observed confounders X_k differs across datasets.

Regardless of the interpretation of $\tau(k)$, one can construct estimators of it that are consistent and asymptotically normal using flexible machine learning estimators.[6] One approach, given in Chernozhukov et al. [50], is to use double machine learning (DML), which employs cross-fitting to produce estimates $\hat{\tau}(k)$ based on the doubly-robust score [221], while using plug-in estimates $\hat{g}_k, \hat{m}_k$ based on machine learning models. This approach achieves asymptotic normality, $\sqrt{N_k}(\hat{\tau}(k) - \tau(k))/\hat{\sigma}^2(k) \overset{d}{\to} \mathcal{N}(0, 1)$, under regularity conditions that allow for flexible machine learning estimators that converge at slower than parametric rates, and where $\hat{\sigma}^2(k)$ converges in probability to the variance of the doubly robust score [See Theorem 5.1 of 50, for additional details]. These results hold whether or not $\tau(k) = \tau$, as discussed in Footnote 9 of Chernozhukov et al. [50]. For simplicity, we have focused on the case where $\mathbb{E}[Y_1 - Y_0 \mid G = 1]$ is constant across datasets. When this does not hold, certain conditions enable valid transportation of treatment effects across datasets [66] with the use of transported estimators [64] (see Appendix D.2 for details).

Example 6.2.2 (Selection of Adjustment Strategy)**.** Consider the two causal graphs given in Figure 6-3, and assume that all variables are binary. Each graph suggests a different identification strategy for the causal effect, $\mathbb{E}[Y \mid do(A = a), G = 1]$. In Figure 6-1, this is identified by the (observational) quantity $\mathbb{E}[Y \mid A = a, G = 1]$, and in Figure 6-2, by front-door adjustment [201] as $\sum_M P(M \mid a, G = 1) \sum_{A'} \mathbb{P}(Y \mid M, A', G = 1)\mathbb{P}(A' \mid G = 1)$.

These observational quantities will typically differ: the one that represents the true interventional effect depends on which graph reflects the true causal structure.

[6] A rich literature focuses on establishing such results, beyond the approach in this example [10, 80, 279, 191, 11].

However, in the case where all variables are discrete and low-dimensional, we can still construct asymptotically normal estimators for both observational quantities.[7] For more complex settings (e.g., requiring regularized ML models for estimating conditional distributions) asymptotic normality has been established under certain conditions for general graphs [29, 136]

Remark 6.2.2. In each example, there are multiple estimators available, each asymptotically normal under basic statistical assumptions, but potentially biased in the sense that $\tau(k) \neq \tau$. In the first example, this bias occurs if X is not sufficient to control for confounding in all observational datasets. In the second, this bias arises in a given estimator if the causal graph is incorrectly specified. Assumption 6.2.3 corresponds to assuming that both the statistical assumptions and causal assumptions hold for one of the candidate estimators, e.g., X is sufficient to control for confounding in at least one study (Example 6.2.1), or that one of the causal graphs is correct (Example 6.2.2).

6.2.3 Asymptotic Normality of GATE Estimators with Transportation

Example 6.2.1 assumes that $\mathbb{E}[Y_1 - Y_0 \mid G = i]$ is constant across datasets. In practice, it may be necessary to correct for differences (not captured by group indicators) between the observational and RCT populations. There exist estimators for the ATE in this setting under mild additional assumptions [64, 61]. These extend in a straightforward way to estimators of the GATE, but proving asymptotic normality is nuanced in high-dimensional settings when using flexible machine learning methods to estimate nuisance functions. For completeness, inspired by Semenova and Chernozhukov [241], we demonstrate that a doubly-robust GATE estimator for this setting is asymptotically normal under reasonable conditions (Assumption D.3.1 to D.3.5 in Appendix D.3). Details on the estimator, and the corresponding proof of normality, are given in Appendix D.3, and may be of independent interest.

[7] This follows from the use of maximum likelihood (i.e., empirical counts) for estimating each conditional distribution, and applying the delta method to the front-door estimator.

6.2.4 Testing for Bias under Asymptotic Normality

Under Assumption E.7.1, each observational estimate $\hat{\tau}_i(k)$ can be compared to the estimate from the randomized trial $\hat{\tau}_i(0)$ for $i \in \mathcal{I}_R$, the groups with common support. Since the observational and randomized datasets are distinct, we can conclude that each $\hat{\tau}_i(k)$ is independent of $\hat{\tau}_i(0)$, and use this to test for the hypothesis that $\tau_i(k) = \tau_i$.

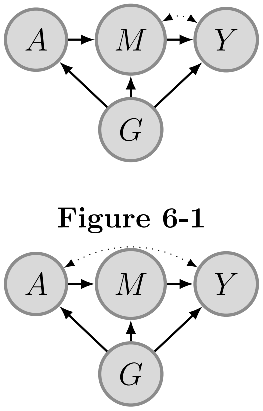


Figure 6-1

Figure 6-2

Figure 6-3: (Ex. 6.2.2) In (6-1), M and Y are confounded by unobservables (bi-directional dotted arrow). In (6-2), A and Y are confounded, but the causal effect is identified via front-door adjustment.

Proposition 6.2.1. *For an observational estimator $\hat{\tau}(k)$, assume Assumptions 6.2.2 and E.7.1 hold. Furthermore, let $N = N_k + N_0$ with fixed proportions, where $N_k = \rho N, N_0 = (1-\rho)N$ for $\rho \in (0,1)$. Define the test statistic*

$$\hat{T}_N(k,i) \coloneqq \frac{\hat{\tau}_i(k) - \hat{\tau}_i(0) - \mu_i(k)}{\hat{s}} \tag{6.3}$$

where $\hat{s}^2 \coloneqq \frac{\hat{\sigma}_i^2(k)}{N_k} + \frac{\hat{\sigma}_i^2(0)}{N_0}$ is the estimated variance, and $\mu_i(k) \coloneqq \tau_i(k) - \tau_i$. This test statistic converges in distribution to a normal distribution as $N \to \infty$, $\hat{T}_N(k,i) \xrightarrow{d} \mathcal{N}(0,1)$.

We present the proof for Proposition 6.2.1 in Appendix D.4. This asymptotic normality allows for the construction of simple hypothesis tests. For instance, one can construct a Wald test for $H_0 : \tau_i(k) = \tau_i$, with asymptotic level α by setting $\mu_i(k) = 0$ in Equation (6.3) and rejecting H_0 whenever, $|\hat{T}_N(k,i)| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the normal CDF. Moreover, the asymptotic power of this

Algorithm 2 Extrapolated Pessimistic Confidence Sets

```
Input: Desired coverage $1-\alpha$. For each $i \in \mathcal{I}_R$, RCT estimate $\hat{\tau}_i(0)$ and variance $\hat{\sigma}_i^2(0)$. For each $i \in \mathcal{I}_R \cup \mathcal{I}_O$, $K$ candidate estimators $\hat{\tau}_i(k)$ and variances $\hat{\sigma}_i^2(k)$. Sample sizes $N_0, \ldots, N_K$.
Initialize: Empty candidate set $\hat{\mathcal{C}} \leftarrow \varnothing$
for $k = 1$ to $K$ do
    Compute $\hat{T}_N(k,i), \forall i \in \mathcal{I}_R$, with $\mu_i(k) = 0$ (Eq. 6.3)
    if $\forall i \in \mathcal{I}_R, \left|\hat{T}_N(k,i)\right| \leq z_{\alpha/4|\mathcal{I}_R|}$, then $\hat{\mathcal{C}} \leftarrow \hat{\mathcal{C}} \cup \{k\}$
end for
for $i \in \mathcal{I}_O$ do
    $\hat{L}_i \leftarrow \min_{k \in \hat{\mathcal{C}}} \hat{L}_i(k)(\alpha/2)$ and $\hat{U}_i \leftarrow \max_{k \in \hat{\mathcal{C}}} \hat{U}_i(k)(\alpha/2)$ (Eq. 6.5)
end for
Return: $\hat{L}_i, \hat{U}_i$ for each $i \in \mathcal{I}_O$.
```

test (the probability of correctly rejecting H_0) is given by

$$1 - \Phi\left(\frac{\mu_i(k)}{\sigma_{k,0}} + z_{\alpha/2}\right) + \Phi\left(\frac{\mu_i(k)}{\sigma_{k,0}} - z_{\alpha/2}\right) \tag{6.4}$$

where $\sigma_{k,0}^2 := \frac{\sigma^2(k)_i}{N_k} + \frac{\sigma_i^2(0)}{N_0}$ [see Theorems 10.4, 10.6 of 284]. Likewise, Assumption E.7.1 implies an asymptotic $1-\alpha$ confidence interval for $\tau_i(k)$ as

$$[\hat{L}_i(k)(\alpha), \hat{U}_i(k)(\alpha)] := \left[\hat{\tau}_i(k) - \frac{z_{\alpha/2} \cdot \hat{\sigma}_i(k)}{\sqrt{N_k}}, \hat{\tau}_i(k) + \frac{z_{\alpha/2} \cdot \hat{\sigma}_i(k)}{\sqrt{N_k}}\right] \tag{6.5}$$

6.3 Meta-Algorithm for Conservative Extrapolation

In this section, we more formally introduce our algorithm (Algorithm 2). There are two primary steps: falsification of estimators, and combination of confidence intervals. First, we attempt to falsify candidate estimators via hypothesis testing, rejecting estimator k whenever we are able to reject the null hypothesis $H_0 : \tau_i(k) = \tau_i, \forall i \in \mathcal{I}_R$. We use Bonferroni correction to control the false positive rate of the test. For the combination of confidence intervals, while we are unlikely to reject the "correct" estimator if one exists (Assumption 6.2.3), we may be unable to reject all "incorrect" (i.e., biased) estimators. This motivates the combination of confidence intervals (for the extrapolated effects) of the accepted estimators by taking the maximum and minimum bounds over all such intervals. Our main result characterizes the properties of our procedure, with proof in Appendix D.4.

Theorem 6.3.1 (Properties of Algorithm 2)**.** *Under Assumptions 7.2.1 and 6.2.2, the output of Algorithm 2 has the following asymptotic properties as $N \to \infty$, where N denotes the total sample size, and the samples used for all estimators are of the same order $N_k = \rho_k N_0, \forall k \geq 1$, for some $\rho_k > 0$.*

1. *Under Assumptions 6.2.3 and E.7.1, for each $i \in \mathcal{I}_O$,*

$$\lim_{N\to\infty} \mathbb{P}(\tau_i \in [\hat{L}_i, \hat{U}_i]) \geq 1 - \alpha \tag{6.6}$$

2. *Under Assumption E.7.1, for each estimator where $\tau_i(k) \neq \tau_i$ for some $i \in \mathcal{I}_R$,*

$$\lim_{N\to\infty} \mathbb{P}(k \in \hat{\mathcal{C}}) = 0 \tag{6.7}$$

The first point says that for each extrapolated effect τ_i, the coverage of the final confidence interval $[\hat{L}_i, \hat{U}_i]$ is at least $1-\alpha$ in the limit. It follows from Assumption 6.2.3 and E.7.1 that at least one estimator provides intervals $[\hat{L}_i(k)(\alpha/2), \hat{U}_i(k)(\alpha/2)]$ that achieve asymptotic coverage of $1-\alpha/2$. The result follows from our choice of threshold for the significance test as well as application of union bounds. The second point says that we will reject estimators that are not consistent for the validation effects, in the limit. Assumption E.7.1 ensures that Proposition 6.2.1 holds for all estimators, so that this rejection is a consequence of the asymptotic power in Equation (6.4), going to 1 for a fixed bias as $N \to \infty$.

Remark 6.3.1. Equations (6.4) and (6.5) are useful for building further intuition. All of the candidate confidence intervals shrink at a rate of $O(1/\sqrt{N})$ as the overall sample size increases. For sufficiently large N, the width of our generated intervals will depend largely on our power to reject biased estimators, which will be higher for observational estimates with larger biases for validation effects.

6.4 Semi-Synthetic Experiments

6.4.1 Setup of Simulation

We generate semi-synthetic RCTs and observational datasets with covariates from the Infant Health and Development Program (IHDP), a randomized experiment on premature infants assessing the effect of home visits from a trained provider on the future cognitive performance [37]. The outcomes are simulated. Our data generation is based on the partial IHDP dataset used in [118], which includes $n_0 = 985$ observation, 28 covariates, and a binary treatment variable. We construct a scenario where there are four subgroups, defined by the infant's birth weight and maternal marital status: (high [$\geq$ 2000g], married), (low [$<$ 2000g], married), (high, single) and (low, single), which we shorthand as HM, LM, HS and LS. We include all subgroups in the observational studies, but exclude the latter two subgroups for the simulated RCT (i.e. only infants with married mothers are in the RCT). This mimics a scenario where only infants with married mothers are recruited into the RCT.

For each simulated dataset, we generate 1 RCT and K observational studies. For the observational studies, we resample the rows of the IHDP dataset to the desired sample size $n = r \cdot n_0$. We performed weighted sampling to induce a different covariate distribution for observational studies, such that male infants, infants whose mothers smoked, and infants whose mothers worked during pregnancy are less prevalent. Then, we introduce confounding in the observational data, generating m_c continuous confounders and m_b binary confounders. Finally, we simulate outcomes in each dataset, modifying the response surface given in Hill [118]. In our experiments, we may choose to conceal some confounders in each observational study to mimic unobserved confounding, denoting the number of concealed variables across the K studies as $\mathbf{c_z} = (c_{z1}, c_{z2}, ..., c_{zK})$. For further details on confounder generation, outcome simulation, and confounder concealment, see Appendix D.6. Data generation parameters include K, r, m_c, m_b, $\mathbf{c_z}$, and the significance level α. By default, we set $K = 5$, $r = 10$, $m_c = 4$, $m_b = 3$, $\mathbf{c_z} = (0, 0, 2, 4, 6)$, and $\alpha = 0.05$. The full hyperparameter search is provided in Appendix D.6, and details of hyperparameter tuning can be found in Appendix D.3.

6.4.2 Implementation and Evaluation of Meta-Algorithm

To implement Algorithm 2, we first obtain GATE estimates for the four subgroups and their estimated variances in each observational study, combining techniques from the DML and trasportability literature [241, 64]. Estimation details are shown in Appendices D.2 and D.3. For the RCT, we stratify the data into the subgroups HM, LM and estimate the GATEs as the difference of mean outcomes between the treated and untreated. The z tests in Algorithm 2 are applied to both GATE estimates in the HM and LM subgroups ($|\mathcal{I}_R| = 2$), and the significance level of the tests is set at $\alpha/4$.

We evaluate performance using two main metrics: (1) the coverage probability of the output confidence intervals (ideally at least $1-\alpha$), and (2) the width of the confidence intervals (narrower is better). In addition to assessing the intervals produced by Algorithm 2, which we call *Extrapolated Pessimistic Confidence Sets (ExPCS)*, we will evaluate intervals produced by a variant of our algorithm, called *Extrapolated Optimistic Confidence Sets (ExOCS)*. In *ExOCS*, after falsifying estimators, we combine confidence intervals using a random-effects meta-analysis on the non-falsified observational studies. We compare *ExPCS* and *ExOCS* against two baselines. *Meta-Analysis* is a random-effects meta-analysis on all observational studies, as described in Section 6.6, with heterogeneity variance estimated via the DerSimonian-Laird moment method [68]. This baseline is the current standard for aggregating observational study results. The second baseline, *Simple Union*, uses the maximum upper bound and minimum lower bound of the $1-\alpha$ confidence intervals across all observational studies, with no falsification procedure.[8]

[8] Note that *Simple Union* combines $1-\alpha$ confidence intervals, while our approach combines $1-\alpha/2$ confidence intervals to account for the probability of rejecting the "correct" estimator, if one exists. As a result, *Simple Union* intervals do not always strictly cover the intervals produced by *ExPCS*.

6.4.3 Results

We perform three semi-synthetic experiments to assess the performance of our proposed meta-algorithm under different scenarios. The first experiment applies our algorithm under the default settings given in Section 6.4.1. In the second experiment, we vary the sample size ratio between the observational studies and the original RCT, r, from 1 to 10. In the third experiment, we vary the proportion of biased observational studies by setting $\mathbf{c_z}$ to be $(0, 0, 0, 0, 0)$, $(0, 0, 0, 0, 3)$, $(0, 0, 0, 3, 3)$ or $(0, 3, 3, 3, 3)$, corresponding to $0, 1, 2, 4$ studies being biased out of a total of 5 observational studies. Results for the latter two experiments are shown over 100 simulations of the datasets. Results for all experiments are shown in Figures 6-4, 6-5, and Figure D-4 in Appendix D.7, respectively. We observe the following:

Meta-algorithm produces confidence intervals that cover the true GATE with nominal probability: We demonstrate in Figure 6-4 the application of our meta-algorithm (*ExPCS*), a variant of it (*ExOCS*), and two other baselines on one dataset. Our goal is to produce narrow confidence intervals that still cover the true GATEs in the extrapolated subgroups. The confidence intervals of *ExPCS* cover the true GATEs in the extrapolated subgroups with reasonable widths. In contrast, intervals produced by *Meta-Analysis* fail to cover the true GATE in both extrapolated subgroups due to the false assumption of unbiasedness across all studies. The *ExOCS* approach produces narrow intervals for the extrapolated effects, though it barely covers the true effect in the HS subgroup. This hints at the need for a conservative combination of non-falsified studies. However, an overly conservative approach (e.g. *Simple Union*) produces wide intervals that may be of little use for meaningful inference.

Although *Meta-Analysis* produces confidence intervals with inadequate coverage, its intervals for the married subgroups still have considerable overlap with the intervals produced by the RCT. This suggests that testing the meta-analyzed GATE estimates against the RCT GATE estimates may not be enough to demonstrate their validity. Compared to our *ExPCS* intervals, the lower bounds of the *Simple Union* intervals are higher in several subgroups, since we use a higher confidence level for the candidate

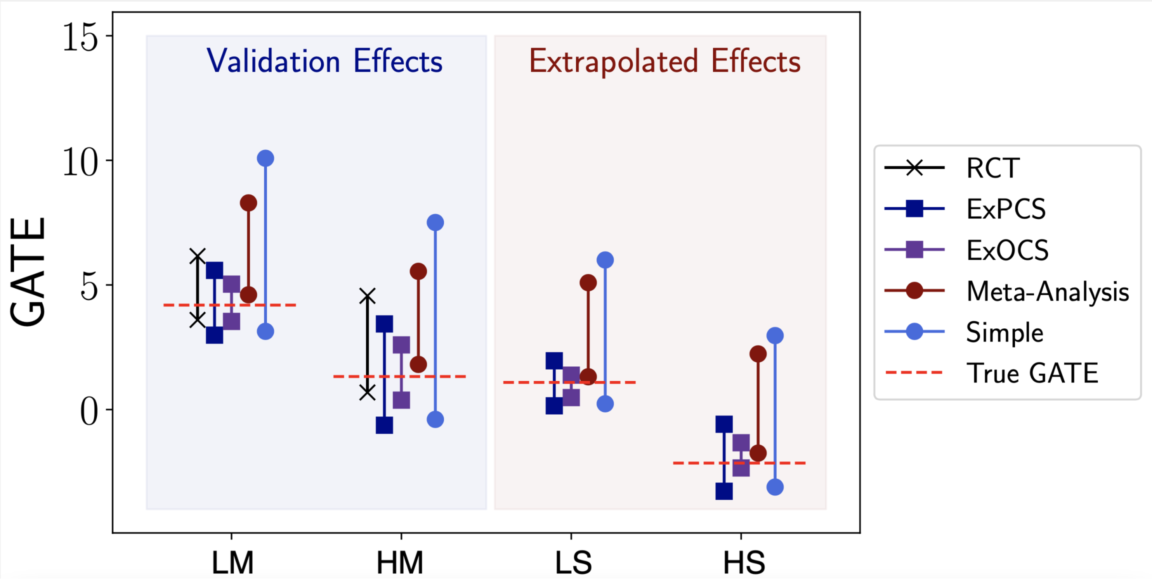


Figure 6-4: The confidence intervals for group average treatment effects (GATE) within the four subgroups output by our algorithm (*ExPCS*), our algorithm variant (*ExOCS*), random-effects meta-analysis on all observational studies (*Meta-Analysis*), simple union bound on all observational studies (*Simple*), and RCT, for one dataset generated using the default parameter settings laid out in Section 6.4.1. *LM, HM, LS, HS* represent four subgroups defined in Section 6.4.1

intervals corresponding to each study to account for probable error in study falsification.

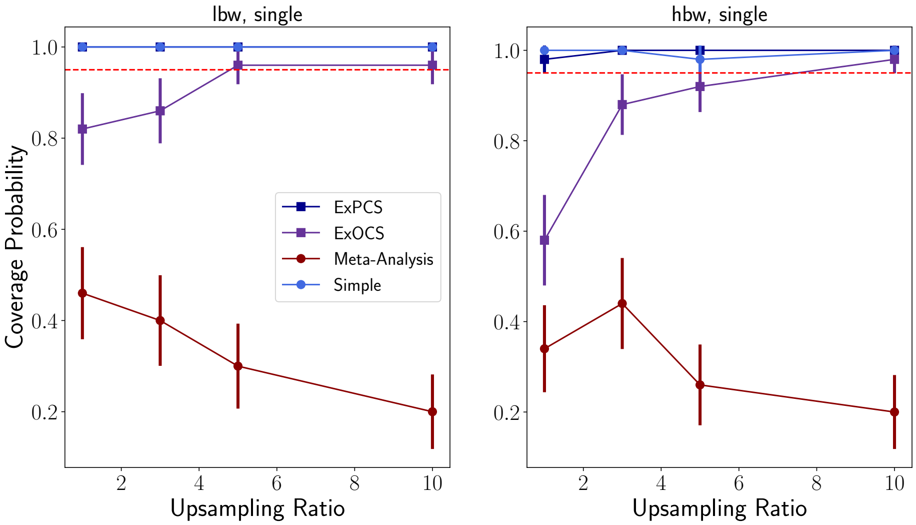


Figure 6-5: Coverage probabilities of confidence intervals shown as a function of the size of the observational studies relative to the RCT. Dotted red lines stand for 95% coverage. Vertical bars are the 95% confidence intervals of the coverage probabilities.

An analysis of increasing observational study size: In Figure 6-5, we find that

Mean width of 95% confidence intervals

		Upsampling ratio			
Group	Estimator	1	3	5	10
LS	*ExPCS*	21.00	9.51	5.53	3.58
	Simple	21.00	10.20	6.95	5.69
HS	*ExPCS*	25.00	8.29	5.36	4.34
	Simple	26.00	9.15	6.90	6.33

Table 6.1: Note that this table should be interpreted in conjunction with Figure 6-5, which shows the coverage probabilities of confidence intervals as a function of the size of the observational studies relative to the RCT. LS / HS stand for groups with low / high birth weight and single mother. Between *ExPCS* and *Simple*, both of which have adequate coverage, *ExPCS* generally has narrower intervals.

the coverage of the *Meta-Analysis* intervals is quite low across all sample sizes and particularly decreases at higher sample sizes. This result is intuitive, as three out of five studies are biased, meaning that meta-analysis will converge to a biased estimate as the amount of data increases. One could attempt to fix this issue through *ExOCS*, which does meta-analysis after falsification. However, *ExOCS* has poor coverage when the sample size of the observational studies is small, since the falsification tests are underpowered (evidenced by the high probability of selecting biased studies in Appendix D.7, Table D.2). Both *ExPCS* and the *Simple Union* intervals have adequate coverage across all sample sizes. However, the widths of the intervals reported at the bottom of Table 6.1 show that *ExPCS* intervals are narrower when there is adequate power, i.e. at higher sample sizes. Ultimately, *ExPCS* will tend to provide intervals that cover the true effect regardless of sample size, and in the case we have sufficient power, these intervals will both have good coverage and narrower width, allowing for more meaningful inference.

6.5 Women's Health Initiative (WHI) Experiments

In order to assess our approach in a real-world setting, we use clinical trial and observational data available from the WHI. Each subgroup is supported in both RCT and observational data, which proves useful for evaluation. At a high level, we "hide" some number of subgroups from the RCT, estimate a confidence interval of the effect estimate using our algorithm on the remaining data, and compare the result to the hidden RCT estimate. We do this over a large set of possible "held-out" subgroups, yielding >2000 different scenarios on which to test our approach. Because the original observational datasets replicate the RCT results fairly well using standard methods, we create additional "biased" datasets by sub-selecting the original observational dataset in a way that induces selection bias. We evaluate each method, for each held-out subgroup, according to the length of the intervals as well as coverage of the RCT point estimates. Below, we describe the specifics of the data, the experimental setup, and the main results of the analysis. For additional details on data preprocessing, setup, and evaluation, see Appendix D.5.

6.5.1 Setup

The Postmenopausal Hormone Therapy (PHT) trial, i.e. the RCT used in this analysis, was run on postmenopausal women aged 50-79 years who had an intact uterus. It studied the effect of hormone combination therapy on several types of cancers, cardiovascular events, and fractures. The observational study (OS) was run in parallel, had a similar follow-up time to the RCT, and tracked similar outcomes. In our analysis, we use a composite outcome, where $Y = 1$ if any of the following events are **observed** to occur in the first 7 years of follow-up, and $Y = 0$ otherwise: coronary heart disease, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, and death due to other causes. This represents a binarization of the "global index" time-to-event outcome from the original study, where $Y = 0$ could also occur due to censoring. We establish treatment and control groups in the OS based on explicit confirmation or denial of usage of both estrogen and progesterone in the first three years. We use only covariates measured in both the RCT and OS to simplify analysis.

6.5.2 Evaluation

Our empirical evaluation consists of several steps. In the first step, we replicate the principal results from the PHT trial, given in Table 2 of [229], by fitting a doubly robust estimator (of the style given in Appendix D.3) on the WHI OS data. Then, while treating the WHI OS dataset as the "unbiased" observational dataset, we simulate additional "biased" observational datasets by inducing selection bias into the WHI OS. The exact mechanism of selection bias and its clinical intuition is given in Appendix D.5. Importantly, this is the only part of the evaluation that involves any simulation.

The second step is to construct a large suite of tasks on which to evaluate our method, by considering different sets of validation-extrapolation subgroups. To construct the subgroups, we consider all pairs of a selected set of binary covariates (see Appendix D.5.6), where each pair defines four subgroups. For example, one covariate pair is ("current smoker", "currently drinks alcohol"). We treat two of the subgroups as validation subgroups and two as extrapolated subgroups. For the latter groups, we apply our algorithm without access to the RCT data, and only use the RCT data for final evaluation. The total number of covariate pairs is 592, leading to 1184 distinct "tasks" (i.e., extrapolated groups). For each task, we evaluate ExPCS (our method), ExOCS, Simple, and Meta-Analysis (described in Section 6.4.2). Additionally, we evaluate an "oracle" method, which is identical to ExPCS, except that it always selects only the original observational study (i.e. the base WHI OS to which we have not added any selection bias). For each method, we compute the following metrics, averaged across all tasks:

- **Length**: length of the confidence interval
- **Coverage**: percentage of tasks where the interval covers the RCT point estimate
- **Unbiased OS Percentage**: the percentage of tasks where the ExPCS approach retains the unbiased study after the falsification step.

	Coverage	**Length**	**OS %**
Simple	0.39	0.416	–
Meta-Analysis	0.03	0.260	–
ExOCS	0.28	0.058	–
ExPCS (ours)	0.45	0.081	0.99
Oracle	0.44	0.068	–

Table 6.2: Coverage, length, and unbiased OS % of ExPCS and baselines. ExPCS achieves comparable coverage to the oracle method with highly efficient intervals. Additionally, we do not reject the unbiased OS in 99% of the tasks.

6.5.3 Results

Table 7.1 reports the metrics above, averaged across all extrapolated subgroups.

Compared to the "simple" baseline, our approach has better coverage with much shorter confidence intervals. Our falsification procedure retains the unbiased observational study 99% of the time, yielding near-oracle coverage rates, but produces substantially shorter intervals than the "simple" baseline. Recall that the simple baseline takes a union over all $1 - \alpha$ intervals estimated from each observational dataset, while ExPCS takes a union of a smaller number of slightly wider $(1 - \alpha/2)$ confidence intervals.

Compared to the Meta-Analysis and ExOCS baselines, we achieve comparable (or much better) length with substantially better coverage. In particular, compared to meta-analysis, we achieve tighter intervals and also cover the RCT estimate with higher frequency. This result is intuitive, since one will get a biased estimate if biased observational studies are included in the meta-analysis. Additionally, conservatively combining the non-falsified estimates (as opposed to *ExOCS*, which does a meta-analysis on the non-falsified estimates) is important to achieve good coverage (0.45 vs 0.28).

We get comparable coverage and interval lengths to the oracle method. Our coverage rate is nearly identical (0.45) to that of the oracle method (0.44), with intervals that are marginally wider (0.081 vs. 0.068). Our slightly improved coverage is possible due to the wider intervals. Note that our measure of "coverage" may be pessimistic, because we track coverage of the RCT point estimate, as opposed to the true causal effect (which is unknown), and the confidence intervals are designed to cover the latter.

Indeed, we report the oracle method precisely as a means of providing a more suitable comparison. Overall, our real-world results suggest that our method of falsification followed by a conservative combination of intervals may be useful for biostatisticians and clinicians when doing meta-analyses.

6.6 Related Work

Meta-analysis for combining observational estimates Among the quantitative approaches for meta-analysis to account for potential bias, our *Meta-Analysis* baseline is standard for meta-analysis of observational data [117] to account for heterogeneity. Allowing for heterogeneity of treatment effects among studies produces wider confidence intervals and thus more conservative inference. If additional study-level covariates are available (e.g. study designs, drop-out rate), several approaches aim to adjust for potential bias, either by modeling the bias magnitude [74, 289, 9, 100], down-weighting studies with higher risk of bias [124, 187], or using Bayesian hierarchical regression to account for difference between subgroups of studies [213, 285]. Our work differs from these approaches, in that (1) we use information from outside the population of interest to assess bias, and (2) we do not place any assumptions on the patterns of bias across studies.

Partial identification and sensitivity analysis These methods seek to place bounds on causal effects when they cannot be point-identified. Our method can be seen as an alternative way of doing so, with a fundamentally different type of assumption. Methods for partial identification rely on having discrete variables and a known causal graph (typically including unobserved confounders) [72, Section 9]. Methods for sensitivity analysis, on the other hand, translate assumptions about the strength and nature of unobserved confounding into bounds on causal effects [224, 225, 295]. In contrast, we do not make any such assumptions, e.g., we allow for continuous variables, and when some candidate estimators are biased due to unmeasured confounding, we do not place any limit a-priori on the bias.

Combining observational and experimental data Prior work has sought to combine RCTs and observational studies for the purpose of more precise estimation of treatment effects [226, 227], or for the purpose of generalizing or transporting estimates from RCTs to observational populations when overlap holds between the two (see Degtiar and Rose [66] for a recent review). In contrast, our work is motivated by settings where there are populations in the observational studies who are not at all represented in trials, e.g., due to a lack of eligibility. Kallus et al. [137] also seek to combine observational and experimental data to extrapolate beyond tfhe support of an RCT. They propose to learn a CATE function on observational data, and then learn a parametric additive correction term on the sample that overlaps between the RCT and observational data. In contrast to this approach, we do not assume that confounding can be corrected for, and instead seek to choose an observational estimate (if one exists) that is already consistent for each sub-population.

Calibrating observational confidence intervals An alternative method for calibration of confidence intervals for observational studies makes use of negative controls [170], such as drug-outcome pairs known to have no causal relationship. Methods range using uses these negative controls to form an empirical null distribution for callibrated p-values[237], and Schuemie et al. [238] extend this approach to calibration of confidence intervals. These techniques have been used in several large-scale observational studies such as the LEGEND-HTN study for comparing antihypertensive drugs [258]. By contrast, our method does not assume the existence of negative controls, but instead uses a form of positive control (i.e., validation effects), one that is based on the same underlying treatment as the extrapolated effect.

6.7 Discussion and Limitations

We have presented a meta-algorithm that constructs conservative confidence intervals for group average treatment effects of subgroups that are not represented in RCTs, but are represented in observational studies. Under the assumption that there exists at least one candidate estimator that is asymptotically normal and consistent for both the validation and extrapolated effects, these intervals will achieve the correct asymptotic coverage of the true effect. However, our method is not without limitations. Most notably, we may fail to reject the null hypothesis due to low power, e.g., when an observational estimate $\hat{\tau}(k)$ has high variance. In practice, we expect that our approach will be most useful when the observational studies in question have large sample sizes, leading to higher-precision estimates of potential bias, and smaller confidence intervals on the extrapolated effects. Second, we may reject the null for reasons that are unrelated to violations of causal assumptions; For instance, it may be the case that X includes all relevant confounders, but that the distribution of X differs between the RCT and an observational study, such that both estimate the ATE correctly, but for different populations. The latter issue could be addressed via a finer partition of subgroups for the construction of validation effects, or via re-weighting approaches for comparing effects from observational and randomized studies, as outlined in [62]. As long as the resulting estimators are asymptotically normal, then our results would still apply under re-weighting. Our hope is that methods such as ours will lead to higher confidence in observational estimates when RCT data is available to falsify observational studies that do not replicate known causal relationships. Finally, great care should be taken to appropriately validate and soundly interpret the results of our method in practice, especially with more sensitive subgroups (e.g. with respect to race or gender).

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 7

Towards More Powerful Falsification via Maximum Moment Restriction-based Hypothesis Tests

Having introduced the paradigm of falsification in the prior chapter, it is worth showing how such a procedure might be applicable in a CDSS. Whenever an oncologist selects more than one treatment in the tool, we may use an ML model trained on observational data to compute a causal effect estimate, e.g. difference in expected survival between treatment A and treatment D, by simply inputing a different treatment in the model. The result of our falsification procedure on these causal estimates can serve as an additional sanity check or contextual information in the CDSS. In this chapter, we build on this theme by exploring more powerful methods of falsification, enabling more meaningful checks.

7.1 Introduction

Observational studies, prevalent in healthcare, economics, and other fields, are an important source of real-world data used to derive granular treatment effect estimates [60, 115, 127, 293]. Indeed, there has been a rich literature in building estimators of heterogeneous treatment effects from observational data, particularly using modern machine learning methods [280, 152, 242]. However, observational studies may lack *internal validity* in that estimates of causal effects (in the observational population) may be biased or inconsistent, e.g., due to unobserved differences between the treatment and control groups, such as in the setting of unobserved confounding. On the other hand, observational studies are representative of more diverse populations, leading to more plausible *external validity*, i.e., ability to generalize estimates across wider populations. In contrast, *Randomized Controlled Trials (RCTs)* have strong internal validity, assuming sound design (e.g. a prospective trial, a-priori definition of hypotheses to be tested) and appropriate randomization. However, RCTs often have restrictive inclusion criteria, which can call their external validity into question [67], and are of limited size, limiting their ability to detect differences in treatment effect for specific sub-populations, or detect differences in adverse event rates (if e.g., the adverse event is rare) [266, 5, 208]. Intuitively, we would like to leverage observational data for estimating treatment effects that cannot be reliably estimated using RCTs, whether due to a lack of statistical power or a lack of patient diversity in the RCT. At the same time, we would like to take advantage of the strong internal validity of the RCT to increase confidence in our observational estimates.

With these considerations in mind, we study the problem of using limited RCT data to "falsify" assumptions of internal and external validity for observational studies. Our method can be applied even when the RCT data does not cover the entire observational population, and hence cannot be used on its own to estimate causal effects. Assuming that the RCT has internal validity, we show that assumptions of internal/external validity of the observational study have a testable implication in the form of a set of conditional moment restrictions (CMRs). We propose a *falsification algorithm* that tests whether or not these CMRs hold, thereby providing an opportunity to reject these assumptions when they fail to hold. This allows us to take advantage of approaches developed in the econometrics literature for testing CMRs, and we use a Maximum Moment Restriction (MMR)-based test [185] for this purpose.

Compared to prior work, the benefits of our approach are two-fold. First, we implicitly check across all subpopulations of covariates X for disagreement between the conditional average treatment effect (CATE) functions as estimated in the RCT versus in the observational study. Second, as an additional benefit, our approach provides an explanation for rejection, in the form of a “witness function”, which describes subpopulations where these estimates diverge.

Importantly, in constructing the test for our problem, we use the insight that not all differences between observational and experimental distributions matter. For instance, there may be differences in unobserved baseline risk factors, which cause estimates of individual potential outcomes to differ, but do not impact the causal effect (the difference between control and treated outcomes). A naive MMR-based test would asymptotically reject in this scenario, but we demonstrate that, by careful construction of the MMR test statistic, we can avoid this failure case.

Our method can be compared to prior approaches to falsification of observational studies. One such approach is to check for a statistically significant difference between estimates of the average treatment effect (ATE) from the RCT and from the observational study [89, 60, 15]. Unfortunately, this approach can lead to false negatives, e.g., if the ATE from the observational study replicates the RCT ATE, despite biases on finer-grained subpopulations. Furthermore, even if an observational study is correctly “rejected”, the approach does not provide an **explanation** for why the observational study was rejected, which is an important practical consideration for both statisticians and policymakers. Another approach is to compare subgroup-level effects instead of the ATE. Hussain et al. [120] adopt this approach in the context of testing (multiple) observational estimates against RCT estimates, essentially testing for differences in group-wise treatment effects. However, this approach requires correction for multiple hypothesis testing across subgroups, and a-priori specification of these subgroups, which can limit its ability to uncover areas of disagreement.

Contributions: We have the following desiderata for our falsification algorithm:

(i) rejecting observational studies when their underlying causal assumptions fail *(high power)*,

(ii) accepting in cases where these causal assumptions hold *(controlled type I error)*, and

(iii) providing an explanation of why an observational study is rejected.

With these desiderata in mind, our main contributions are as follows:

(i) First, we show how to convert causal assumptions on internal and external validity into a set of CMRs, violations of which can be detected using observational and RCT data using existing techniques with theoretical guarantees.

(ii) Second, we demonstrate that our construction of these CMRs avoids a potential failure mode: rejecting observational studies due to differences in unobserved covariates that influence baseline outcomes, but not treatment effects.

(iii) Third, on semi-synthetic and real-world datasets, we show favorable performance of our method with respect to power and type I error, and showcase its ability to produce informative *explanations* of rejections.

7.2 Setup and Motivating Examples

7.2.1 Notation & Assumptions

Let $Y \in \mathcal{Y}$ be the outcome of interest, and $A \in \{0, 1\}$ denote a binary treatment variable. We let Y_a denote the potential outcome under treatment $A = a$, and we use $X \in \mathcal{X}$ to denote the full set of covariates. Note that, in our development, we operate in the setting where there is a single observational study and RCT. We use an indicator variable, $S = \{0, 1\}$, where $S = 1$ denotes data from the observational study and $S = 0$ from the RCT.

To further characterize the observational study and RCT, we let $\mathcal{I}_0$ and $\mathcal{I}_1$ be the observed indices for the RCT and observational study, respectively. Furthermore, we let $\mathcal{I} = \mathcal{I}_0 \cup \mathcal{I}_1$ be the total set of observed indices. We use $|\mathcal{I}|$ to denote the cardinality of a set, and let $|\mathcal{I}_0| = n_0$, $|\mathcal{I}_1| = n_1$, and $|\mathcal{I}| = n$. Finally, $\mathbb{E}[.]$ and $\mathbb{P}[.]$ are expectations and probabilities taken with respect to the joint distribution $\mathbb{P}(Y, A, X, S)$ of the observational study and RCT.

Our goal is to discover violations of causal assumptions that underlie the validity of conditional average treatment effect (CATE) estimates derived from the observational study. To that end, we first state these assumptions formally.

Assumption 7.2.1 (*Internal Validity of Observational Data*)**.** We assume the following in the observational study:

- *Ignorability* — $Y_a \perp\!\!\!\perp A \mid X, (S = 1)$, $\forall a \in \{0, 1\}$.

- *Consistency* — $A = a, S = 1 \Longrightarrow Y_a = Y$, $\forall a \in \{0, 1\}$.

- *Positivity of Treatment* — $\mathbb{P}(X = x, S = 1) > 0 \implies \mathbb{P}(A = a|X = x, S = 1) > 0$, $\forall a \in \{0, 1\}$ and $\forall x \in \mathcal{X}$.

Assumption 7.2.1 gives a standard set of assumptions under which the CATE conditioned on X, $\mathbb{E}[Y_1 - Y_0|X = x, S = 1]$ can be identified. However, this assumption is not testable in isolation. In order to compare observational estimates with those of the RCT to discover flaws, we will first need to assume that the RCT itself provides valid estimates.

Assumption 7.2.2 (*Internal Validity of RCT*)**.** We assume the following in the RCT:

- *Ignorability* — $Y_a \perp\!\!\!\perp A \mid X, (S = 0)$, $\forall a \in \{0, 1\}$.

- *Consistency* — $A = a, S = 0 \Longrightarrow Y_a = Y$, $\forall a \in \{0, 1\}$.

- *Fixed probability of assignment* — $P(A = 1|X = x, S = 0) = p$, for some $p \in (0, 1)$, $\forall x \in \mathcal{X}$.

Assumption 7.2.2 is a generally defensible (and standard) set of assumptions on the validity of the RCT. However, even if both assumptions 7.2.1 and 7.2.2 hold, the corresponding CATE functions are not necessarily comparable. For instance, there may be unmeasured effect modifiers that have different distributions between the RCT and observational study. Under the following additional assumption, the CATE in the RCT (i.e., $\mathbb{E}[Y_1 - Y_0 \mid X = x, S = 0]$) can be identified using observational data.

Assumption 7.2.3 (*External Validity: Observational Study to RCT Transportability of CATE*)**.** We assume the following:

- *Mean Exchangeability of Contrast* — $\mathbb{E}[Y_1 - Y_0|X = x] = \mathbb{E}[Y_1 - Y_0|X = x, S = s]$, $\forall x \in \mathcal{X}$ and $\forall s \in \{0, 1\}$.
- *Positivity of Selection* — $\mathbb{P}(X = x|S = 0) > 0 \implies \mathbb{P}(X = x|S = 1) > 0$, $\forall x \in \mathcal{X}$.

The first part of this assumption is sometimes referred to as "generalizability in effect measure" [63] or "conditional exchangeability in measure" [65]. This assumption is weaker than (and implied by) transportability of counterfactual means (e.g., $\mathbb{E}[Y_a \mid X, S] = \mathbb{E}[Y_a \mid X]$)[1]. It is simple to show that this assumption (along with our other assumptions) is sufficient to identify the CATE in the RCT population using the observational distribution alone.

Proposition 7.2.1. *Under assumptions 7.2.1 and 7.2.3, the CATE of the RCT given* X*,* $\mathbb{E}[Y_1 - Y_0|X, S = 0]$*, is identifiable in the observational data by*

$$\mathbb{E}[Y \mid X, A = 1, S = 1] - \mathbb{E}[Y \mid X, A = 0, S = 1] \tag{7.1}$$

Proposition 7.2.1 follows from substantially the same arguments used by Dahabreh et al. [63] for identification of average treatment effects under similar assumptions, but we include a short proof in Appendix E.1 for completeness, alongside all other proofs for this paper.

Remark 7.2.1. As we demonstrate later on, our statistical test does not distinguish between violations of assumption 7.2.1 or assumption 7.2.3. However, violation of either assumption is a meaningful finding when considering the credibility of causal effects learned from observational data. For instance, even if the observational study is free of unmeasured confounding (i.e., assumption 7.2.1 holds), there may exist unmeasured effect modifiers whose distributions differ substantially across populations, leading to a violation of assumption 7.2.3. In other words, if the true CATE function

[1] Equality relations including random variables are to be understood as "almost sure" (a.s.) relations throughout the manuscript.

varies substantially for individuals with the same covariates X across the observational and RCT populations, then it may not reliably generalize to future patients.

7.2.2 Motivation: Testing for Differences in Causal Contrasts, rather than Counterfactual Means

When it is possible to identify a causal effect from an observational study, we would prefer to avoid rejecting that study unnecessarily. This motivates assumption 7.2.3, which holds even in the scenarios where counterfactual means in the RCT (e.g., the expected outcome under treatment $\mathbb{E}[Y_1 \mid X, S = 0]$) are **not** identifiable from observational data, but where the causal contrast is identified.

This assumption is central to our testing methodology, as we test for a null hypothesis that is satisfied under assumption 7.2.3, even when counterfactual means are not transportable. We build intuition for this assumption in two ways. First, we give a structural causal model that formalizes a sufficient condition for this assumption to hold. Second, we give concrete examples of where this assumption appears to (approximately) hold in practice.

Example 7.2.1. Let U be a set of variables that are unobserved in both the RCT and observational data. Suppose Y is generated according to the following structural equation, with binary treatment A and observed covariates X

$$Y = g(X, U) + \tau(X) \cdot A + \epsilon_0, \tag{7.2}$$

where ϵ_0 is an independent mean-zero random variable ($\mathbb{E}[\epsilon_0] = 0$ and $\epsilon_0 \perp\!\!\!\perp X, U, A, S$) and where $P(U \mid X, S = 1) \neq P(U \mid X, S = 0)$.

In example 7.2.1, $Y_0 = g(X, U) + \epsilon_0$ is influenced by both X and a set of unobserved baseline characteristics U. As a result, the conditional counterfactual mean $\mathbb{E}[Y_0 \mid X, S] = \mathbb{E}[g(X, U) \mid X, S]$ will generally differ across studies, due to the fact that the distribution of U varies across studies. The conditional average treatment effect, on the other hand, is independent of U and S, as $\mathbb{E}[Y_1 - Y_0 \mid X] = \tau(X)$. This quantity is purely a function of X, satisfying our assumption that the CATE does not depend on S.[2]

This scenario is plausible in real-world settings, where the treatment effect is a function of a subset of variables that influence the outcome Y. For a real-world example, consider the SPRINT Trial [253], which studies the impact of intensive blood pressure control (A) on a composite outcome (Y) that includes heart failure and death. Here, previous chronic kidney disease (CKD) is a variable, like U, that has a substantial impact on the outcome Y_0 under no treatment (as reported in Figure 4 of SPRINT Research Group [253]), but does not have a (statistically) significant influence on the treatment effect itself (i.e., $\tau(X)$ in the example above). We discuss this example and other examples of real-world motivation in more detail in Appendix E.2.

7.3 MMR-based Falsification Tests

Next, we observe that assumptions 7.2.1 to 7.2.3 have observable implications on the joint RCT and observational data in the form of a (set of) conditional moment restrictions. As a result, if these restrictions fail to hold, then this implies a violation of the underlying causal assumptions. This suggests a hypothesis-testing approach for detecting violations, which we develop in this section. Notably, the resulting hypothesis test looks for differences between the CATE functions estimated from the RCT and observational studies, but does not test for equality of conditional potential outcomes themselves. This is motivated from our prior discussion, that conditional means of potential outcomes (e.g., $\mathbb{E}[Y_0 \mid X]$) could differ between the RCT and observational data, even when the CATE function itself is identified.

7.3.1 CATE Estimation

The crux of our methodology is to use the CATE estimate from the RCT as a proxy for the true CATE function to falsify or validate an observational estimate. To that end,

[2] The constant treatment effect for individuals with the same X is not necessary, and merely helps simplify notation. One could make a similar observation with $\tau(X, \epsilon_\tau)$ for an additional noise variable ϵ_τ that is independent of U, S.

we first construct an unbiased CATE estimator from RCT data. Since the probability of assignment to each treatment is known by design in RCTs, we can use an IPW-style estimator for the CATE. Similar estimators can be found in standard causal inference textbooks (e.g. Ch. 2 of [114]). A "doubly robust" variant can be used, but if the outcome model is misspecified, this may result in higher variance and a loss of power in our test. Thus, we first define the following "signal" function,

$$\begin{aligned}\psi_0 = &\frac{\mathbf{1}\{S=0\}}{P(S=0 \mid X)} Y \\ &\times \left(\frac{\mathbf{1}\{A=1\}}{P(A=1 \mid S=0)} - \frac{\mathbf{1}\{A=0\}}{P(A=0 \mid S=0)} \right), \end{aligned} \tag{7.3}$$

and then observe that the conditional expectation of this signal $\mathbb{E}[\psi_0 \mid X]$ is equal to the CATE *in the RCT population*, using data from the RCT alone.[3]

Proposition 7.3.1 (*CATE signal from the RCT*). *Under assumption 7.2.2, the instance-wise CATE signal ψ_0 in Equation* (7.3)*, which uses the outcome information from the RCT, is unbiased, i.e.,* $\mathbb{E}[\psi_0|X] = \mathbb{E}[Y_1 - Y_0|X, S=0]$.

Next, we wish to develop a distinct estimate of the CATE in the RCT population, but one which makes use of the observational data. The first step in building such an estimator is to identify, under our causal assumptions, the corresponding *statistical estimand*, i.e. Equation (7.1) in Proposition 7.2.1. Our goal is to check the *validity* of this estimand, which amounts to challenging assumptions 7.2.1 and 7.2.3. Drawing from existing literature [63, 65, 67], we employ the following doubly robust signal,

[3] Note the use of the indicator $\mathbf{1}\{S=0\}$, such that ψ_0 only depends on data from the RCT itself, even though we take the conditional expectation over the combined sample.

which combines response surface modeling and inverse probability weighting (IPW):

$$\psi_1 = \frac{1}{P(S=0 \mid X)} \Bigg[\mathbf{1}\{S=0\} \underbrace{\big(\mu_1(X) - \mu_0(X)\big)}_{\text{Response Surface Signal}} \\ + \mathbf{1}\{S=1\} \frac{P(S=0 \mid X)}{P(S=1 \mid X)} \Bigg(\underbrace{\frac{\mathbf{1}\{A=1\}(Y - \mu_1(X))}{P(A=1 \mid S=1, X)}}_{\text{IPW Signal}} \\ - \underbrace{\frac{\mathbf{1}\{A=0\}(Y - \mu_0(X))}{P(A=0 \mid S=1, X)}}_{\text{IPW Signal}} \Bigg) \Bigg], \tag{7.4}$$

where $\mu_a(X) \coloneqq \mathbb{E}[Y \mid A=a, X, S=1]$.

Proposition 7.3.2 (*CATE signal from the observational data*)**.** *Under Assumptions 7.2.1 and 7.2.3, the instance-wise CATE signal ψ_1 in Eq. 7.4, which uses the outcome information from the observational data, is unbiased for the CATE in the RCT population, i.e., $\mathbb{E}[\psi_1|X] = \mathbb{E}[Y_1 - Y_0|X, S=0]$.*

We are now ready to give the core result of this section, connecting our causal assumptions to the null hypothesis of the statistical test that we will develop in the next section.

Corollary 7.3.1 (*Null Hypothesis on Signal Difference*)**.** *Define $\psi = \psi_1 - \psi_0$ as the instance-wise signal difference between the observational and RCT CATE estimates. Then, under the null hypothesis, i.e. under assumptions 7.2.1 to 7.2.3, we have it that $\mathbb{E}[\psi|X] = 0$.*

Proof. If assumptions 7.2.1 to 7.2.3 hold, then Propositions 7.3.1 and 7.3.2 imply that $\mathbb{E}[\psi_0|X] = \mathbb{E}[\psi_1|X] = \mathbb{E}[Y_1 - Y_0|X, S=0]$. □

Remark 7.3.1. Note that by Corollary 7.3.1, violation of the conditional moment restrictions imply that one or more of our assumptions is incorrect, including the internal validity assumptions on the RCT. If we are willing to independently assume that the RCT is internally valid, then violation of the CMRs implies a violation of one or both of the internal and external validity assumptions *on the observational data.*

7.3.2 Conditional Moment Restriction (CMR) Formulation and Maximum Moment Restriction-based (MMR) Tests

For a practical approach to testing, we leverage the rich literature on conditional moment restriction (CMR) tests. Several examples exist of CMRs being used to express restrictions on functions of the data. One such example is using CMRs to reformulate instrumental variable (IV) regression [298]. However, to our knowledge, using CMRs to compare RCT and observational data as described in this paper has not been previously explored. We present the CMR-version of the null hypothesis in the following proposition:

Proposition 7.3.3 (*Null Hypothesis, CMR*)**.** *Under Assumptions 7.2.1 to 7.2.3, we have a set of conditional moment restrictions (CMRs) on the signal difference, ψ:*

$$H_0 : \mathbb{E}[\psi|X] = 0 \qquad P_X\text{-almost surely,} \tag{7.5}$$

where P_X is the distribution of X on the joint distribution of the RCT and observational study. Equation (7.5) *implies an infinite set of unconditional moment restrictions, $\mathbb{E}[\psi f(X)] = 0, \forall f \in \mathcal{F}$, where $\mathcal{F}$ is the set of measurable functions on $\mathcal{X}$.*

The core part of Proposition 7.3.3 is in showing how we can formulate the CMR given our assumptions, while the second part of the statement is straightforward and follows directly from the law of iterated expectation. Testing CMRs is challenging because an infinite number of equivalent *unconditional* moment restrictions (UMR) must be considered. Thus, we follow a method proposed by Muandet et al. [185], where $\mathcal{F}$ in Proposition 7.3.3 is set to be a reproducing kernel Hilbert space (RKHS). They further show that using the *maximum moment restriction* (MMR) within the unit ball of the RKHS as the test statistic fully captures the original set of CMRs and also has a closed-form expression that can be easily implemented. However, note that here we are directly testing the CMRs, while Muandet et al. [185] consider testing hypotheses *on statistical parameters that imply CMRs*, which leads to a larger set of assumptions on the parameters. Therefore, in the following, we state the hypothesis test with respect to the MMR test statistic and the assumptions required for our use case. A proof showing that these assumptions suffice for the properties of the test to hold is provided in Appendix E.1. This main result will hold for a particular class of kernels, which we define here:

Definition 7.3.1 (*Integrally strictly positive definite (ISPD)*)**.** A kernel $k(\cdot, \cdot) : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ is integrally strictly positive definite if for all $f : \mathcal{W} \to \mathbb{R}$ satisfying $0 < \|f\|_2^2 < \infty$,

$$\int_{\mathcal{W}\times\mathcal{W}} f(w)k(w, w')f(w')dwdw' > 0$$

Now, we are ready to give an MMR-based hypothesis test that tests the null hypothesis given in Proposition 7.3.3:

Theorem 7.3.1 (*Maximum Moment Restriction-based test for CATE function*)**.** *Let $\mathcal{F}$ be a RKHS with reproducing kernel $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is ISPD, continuous and bounded. Suppose $|\mathbb{E}[\psi|X]| < \infty$ almost surely in P_X, and $\mathbb{E}[[\psi k(X, X')\psi']^2] < \infty$ where (ψ', X') is an independent copy of (ψ, X). Let $\mathbb{M}^2 = \sup_{f\in\mathcal{F},||f||\leq 1}(\mathbb{E}[\psi f(X)])^2$. Then,*

1. *The conditional moment testing problem in Eq. 7.5 can be reformulated in terms of the MMR as $H_0^{':\mathbb{M}^2=0}$, $H_1^{':\mathbb{M}^2\neq 0}$.*

Further, let the test statistic be the empirical estimate of $\mathbb{M}^2$,

$$\hat{\mathbb{M}}_n^2 = \frac{1}{n(n-1)} \sum_{i,j\in\mathcal{I},i\neq j} \psi_i k(x_i, x_j)\psi_j$$

2. *Then, under $H_0^{'}$,*

$$n\hat{\mathbb{M}}_n^2 \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1)$$

where Z_j are independent standard normal variables and λ_j are the eigenvalues for $\psi k(x, x')\psi'$.

3. *Under $H_1^{'}$,*

$$\sqrt{n}(\hat{\mathbb{M}}_n^2 - \mathbb{M}^2) \xrightarrow{d} \mathcal{N}(0, 4\sigma^2)$$

where $\sigma^2 = var_{(\psi,X)}[\mathbb{E}_{(\psi',X')}[\psi k(X,X')\psi']]$

Remark 7.3.2. Intuitively, the MMR test statistic is trying to find regions in X where the signal difference (i.e. the difference in the CATE estimates between the observational study and RCT) is maximized. Thus, the larger the signal difference is, the larger the test statistic will be, and the more likely we will be to reject the null hypothesis. This is exactly the behavior that we want from such a test statistic. Note as well that the test statistic is computed using the signal difference directly and *not* separately for each potential outcome mean. This theoretically-grounded choice follows directly from our discussion in Section 7.2.2.

Remark 7.3.3. Note that these asymptotic distributions imply that $n\hat{\mathbb{M}}_n^2$ converges to a distribution with finite variance under the null, but diverges at a rate of $\sqrt{n}$ under the alternative hypothesis, which implies that the MMR test has asymptotic power of one. In addition, since the null distribution does not have a closed form, to obtain the critical value for rejection, we follow Algorithm 1 in [185], which uses bootstrap to simulate the null distribution.

Remark 7.3.4. Proposition 7.3.3 states that the true signal difference, $\psi = \psi_1 - \psi_0$, satisfies a set of CMRs. However, in practice, we perform testing using an estimate of the signal difference, $\widehat{\psi}$, where we plug-in estimates of the underlying nuisance functions, such as the propensity score, $P(A = a|S = 1, X)$. As a result, we might expect our statistical test, all else being equal, to be more likely to reject an observational study, as there are two sources of variation in the test statistic: first, in the signals themselves through the estimated nuisance functions, and second, through variation in the data that exists even when the signals are perfectly estimated. In our semi-synthetic experiments, we find that the difference in the type I error (between using ψ and $\widehat{\psi}$) is minimal for moderately large sample sizes and converges to zero as the sample size increases.

7.3.3 Explainability of MMR-based Falsification Test

Another appealing feature of using the MMR-based approach is that we may express the maximizer,

$$f^* = \arg \sup_{f \in \mathcal{F}, ||f|| \leq 1} (\mathbb{E}[\psi f(X)])^2, \tag{7.6}$$

in closed form (see the proof of corollary 7.3.2 for details). In turn, we may determine where the CATE function estimated by the RCT and the observational study is the most discrepant by looking at regions of $\mathcal{X}$ with large magnitudes of f^*. The function f^*, known as the "witness function", can be found by the following corollary:

Corollary 7.3.2. *The witness function in Equation* (7.6) *can be estimated as*

$$\hat{f}^*(x) = C\frac{1}{n}\sum_i \psi_i k(x_i, x)$$

where C is an unrelated constant so that $\int_{\mathcal{X}} f^{*2}(x)dx = 1$.

Remark 7.3.5. Consider the following example where having a witness function could be beneficial: suppose an endocrinologist wants to determine whether to prescribe SGLT2-inhibitors ($A = 1$) or not ($A = 0$) for diabetic patients. Further assume that there is an RCT and an observational study that studies the effect of SGLT2-inhibitors on HbA1c levels (Y). If our MMR-based approach were to falsify the observational study, the witness function would enable the clinician to understand what types of patients (e.g. people who are $\geq$ 60 years old and have history of heart disease) have conflicting conclusions in the RCT versus observational study with respect to drug benefit.

With this information, they may seek to understand and do follow-up analyses on what violations of the causal assumptions led to the discrepancy in the "older with prior heart disease" patient population. For example, there may be a violation of internal validity of the observational data (e.g. ignorability), where there are still some unmeasured confounders in the observational study for this particular patient group. Alternatively, there could be an external validity violation (e.g. mean exchangeability of contrast), whereby the causal effects themselves are unbiased, but the standard of care of this patient population may be different between the two studies (i.e. unmeasured effect modifiers). Overall, the witness function can provide a window for clinicians to look for possible violations in a specific patient population, allowing for a richer view into observational study results.

Remark 7.3.6. A practitioner may interpret or visualize the witness function in a couple of different ways, which we outline here. A simple method, appropriate for domain experts (e.g. clinicians), is to use domain knowledge to pre-select a subset of covariates on which one can do low-dimensional projections for each pair. We provide an example of this method in our experimental results. Another method is to take the top or bottom 10% of witness function values over X and then look at characteristics of these populations. This approach can guide the choice of low-dimensional parameters to examine (e.g., for major differences across age, the witness function projected onto age can be plotted).

The MMR-based testing framework is useful both because it affords a closed-form expression of the test statistic used to test the CMR in Proposition 7.3.3 *and* gives an explainable view into the rejections of the test via the witness function. We argue that both are crucial for our problem of falsification of causal assumptions in observational studies, specifically internal and external validity. Several other approaches exist in the literature for testing CMRs, and we point the reader to Muandet et al. [185] for an overview. In the following two sections, we will tease out empirically the benefits of our testing approach against several baselines and provide an example of how the witness function can be used in practice on a real-world dataset.

7.4 Semi-Synthetic Experiments

7.4.1 Setup

For this set of experiments, we use covariates from the Infant Health and Development Program (IHDP), an RCT run on premature infants assessing the treatment effect of professional home visits on future cognitive function [37]. We generate an RCT and observational dataset (with simulated outcomes) from the partial IHDP dataset used in Hill [118], which contains 985 observations, 28 covariates, and one binary treatment variable.

A “simulated” dataset in our experiments consists of a single RCT and a single observational dataset. Our simulation strategy for the data draws largely on the approach taken by Hussain et al. [120]. In particular, to generate the RCT, we resample the rows of the IHDP dataset to $n_0 = 2955$. For the observational dataset, we first resample the rows of the IHDP dataset to the desired sample size, $n = s \cdot n_0$. Then, we induce a difference in the covariate distribution between the observational component and the RCT by doing weighted resampling in the observational data, such that male infants, infants whose mothers smoked, and infants with working mothers are less prevalent. To introduce explicit violations of our assumptions in the observational data, we generate m confounders so that we can later conceal some of them to simulate unmeasured confounding. Then, in both the RCT and the observational dataset, we simulate outcomes according to a response surface detailed in Appendix E.3. Finally, we conceal c_z confounders in order of “confounding strength”, which is determined by a vector, $\gamma \in \mathbb{R}^m$. For more information on confounder generation, outcome simulation, and bias simulation via confounder concealment, see Appendix E.3. For parameters m, c_z, and α (significance level), we default to $m = 7$, $c_z = 0$, and $\alpha = 0.05$ unless otherwise specified.

7.4.2 Evaluation

We evaluate our algorithm based on our original desiderata. Namely, we measure *power*, i.e. the rate of rejecting the null hypothesis when the CATE function estimates from the RCT and observational study *do not* converge to the same function and *type I error*: rate of rejecting the null hypothesis given that they *do* converge to the same function.

We use the following two baselines. **Average Treatment Effect (ATE)** – in the RCT, we compute the difference of mean outcomes between the treatment and control groups; in the observational data, we obtain an ATE estimate by leveraging recent techniques in the double machine learning (DML) and transportability literature, akin to the estimator in Dahabreh et al. [65]. **Group Average Treatment Effect (GATE)** – in the RCT, we compute the difference of mean outcomes between the treatment and control groups in pre-specified subgroups defined by the infant's birth weight and maternal marital status[4]; in the observational data, we use a transportable, doubly-robust estimator (see Appendix C in Hussain et al. [120]), to estimate the GATE for each subgroup. Both baselines use hypothesis testing based on asymptotic normality of ATE or subgroup estimates. Note that this approach requires pre-specification of the subgroups.

Both baselines reflect the idea of "falsifying" the observational study by looking at a pre-specified group (subgroups, in the GATE case) to detect differences in the causal effect estimates. Our method, labeled as **MMR-Contrast**, requires no pre-specification and automatically finds "highly-discrepant" regions where the causal effect estimate is different between the RCT and the observational study.

7.4.3 Results

MMR-Contrast largely maintains the desired type I error of 0.05 *while having more power compared to baselines.* As conjectured in remark 7.3.4, MMR-Contrast tends to slightly over-reject, which is reflected in Figure 7-3 by the marginally elevated type I error. Furthermore, MMR-Contrast enjoys greater power than GATE and ATE, particularly in settings where the confounding bias is more subtle. We conjecture that the gain in power is due to MMR-Contrast implicitly checking across all subpopulations of X for disagreements in CATE estimates. Indeed, we see that when the concealed confounder has a weight of 1 (as opposed to 2.75), the difference in power between MMR and ATE is much larger.

When computing MMR-Contrast with ψ versus $\widehat{\psi}$, the empirical gap in type I error shrinks with increasing sample size of the observational study (see Figure 7-4). Reassuringly, we see that the level of the test is maintained at $\alpha = 0.05$ when the true

[4] We specify the following four subgroups: ($\geq$ 2000g, married), ($<$ 2000g, married), ($\geq$ 2000g, single), ($<$ 2000g, single)

Selection Bias	**MMR-Contrast**	**ATE**	**GATE**
$p = 0$	0.29	0.32	**0.17**
$p = 0.05$	**0.67**	0.58	0.40
$p = 0.10$	**0.94**	0.88	0.67
$p = 0.15$	**1.0**	0.98	0.91

Table 7.1: Rejection rate when introducing different amounts of selection bias into the observational data in WHI study. p stands for the strength of selection introduced in the the data (refer to Section 7.5 for details).

signal difference, ψ, is used, which supports our theoretical results. Secondly, using the estimated signal difference, $\widehat{\psi}$, achieves the appropriate type I error when the observational study size at least matches the RCT, i.e. sample size ratio is 1, which one might expect in practice.

Visualizing the witness function in Figure 7-5 demonstrates the covariate regions in which the observational effect estimates are increased or decreased compared to the RCT. We largely see that the witness function yields positive values, implying that the observational study is generally estimating a larger treatment effect (i.e. professional visits benefit child cognitive development) than the RCT. However, there are certain subgroups, e.g. children with high birth order whose mothers do not drink and children with high neonatal health index, for which the observational study estimates lower treatment effects than the RCT. The MMR test is able to discover these subgroup differences, leading to better power than testing for ATE or GATE. Another potential use case of the witness function is for development of treatment guidelines, where subgroups with high witness function values may be “flagged” as having conflicting evidence.

7.5 Women’s Health Initiative (WHI) Experiments

To assess our method in a practical setting, we use observational and clinical trial data from the Women’s Health Initiative (WHI). These studies broadly investigate the impact of hormone therapy and vitamin D supplementation on several clinical outcomes. Conceptually, our analysis consists of first taking B bootstrapped datasets

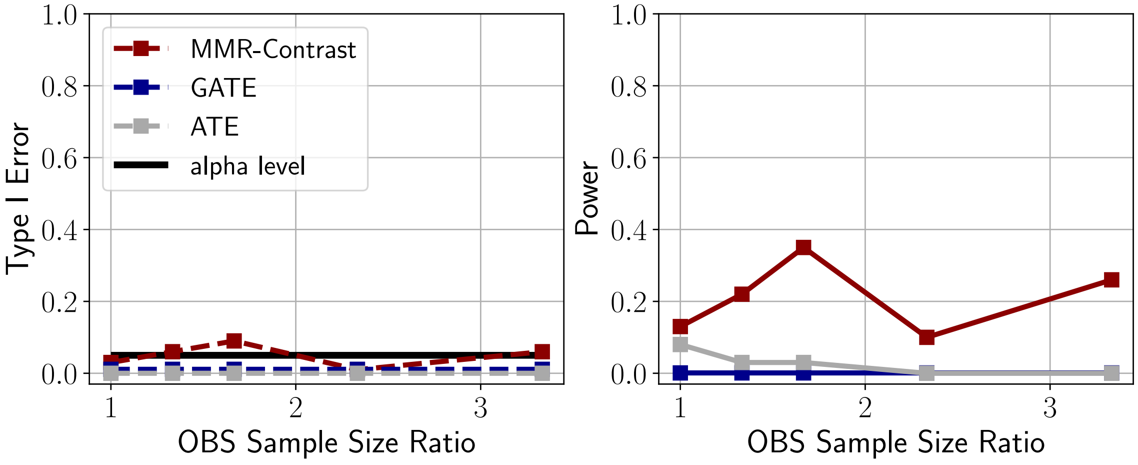


Figure 7-1: Low confounder strength ($\max(\gamma) = 1.$). (left) no unobserved confounders; (right): one confounder concealed

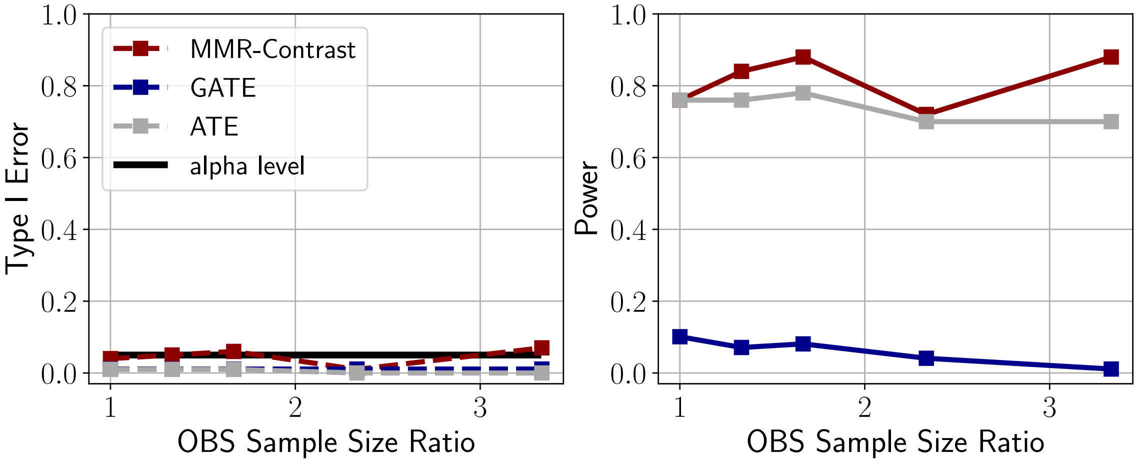


Figure 7-2: High confounder strength ($\max(\gamma) = 2.75$). (left) no unobserved confounders; (right): one confounder concealed

Figure 7-3: Type I error and power of MMR-Contrast, GATE and ATE under different confounder strengths. The left panels in 7-1) and (7-2) show that the level of all three approaches generally retains the nominal level of 0.05. The right panels show the superior power of MMR-Contrast. Particularly, when the confounder strength is lower (as in (7-1)), the difference of CATE estimates between the observational study and the RCT is more difficult to detect, leading to a larger difference of power between MMR-Contrast and ATE. The GATE approach, since it is based on random subgroups, has minimal power, even under the high confounder strength scenario.

from either the original WHI observational study or a "biased" version, then selecting a subset of covariates (to generate subgroups for the GATE baseline), and finally running *GATE*, *ATE*, and *MMR-Contrast* on each bootstrap iteration. We evaluate

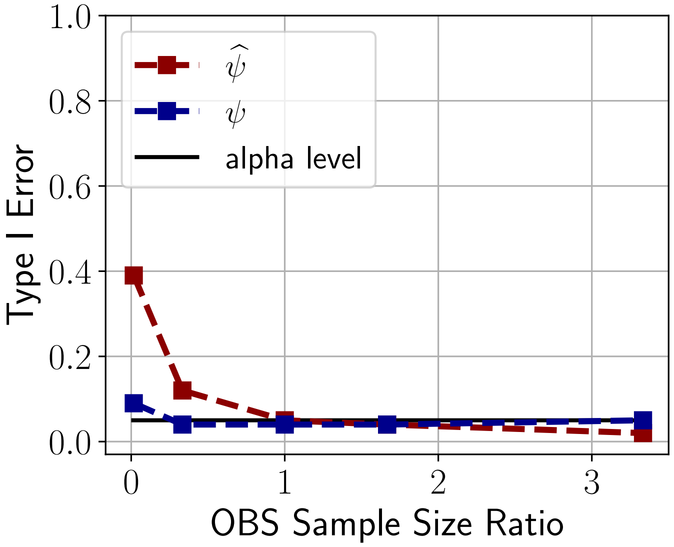


Figure 7-4

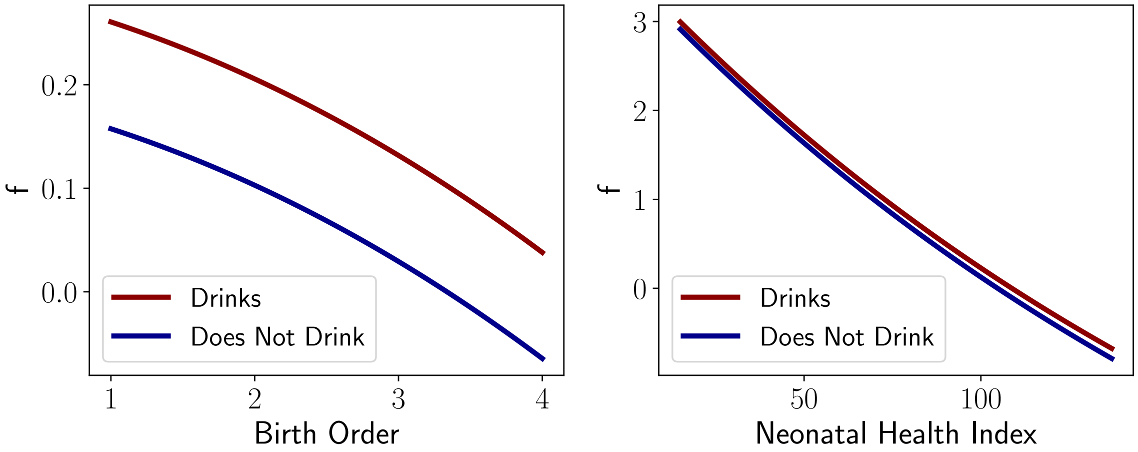


Figure 7-5

Figure 7-6: Panel (7-4) demonstrates the relative performance of tests using test statistics computed with true signals (ψ) and estimated signals ($\widehat{\psi}$). The sample size of the RCT study is fixed ($n_0 = 2955$) and the sample size ratio between the observational study and the original IHDP data ranges from 0.01 to 3.33. The blue line shows that the test using ψ achieves the nominal level (0.05). The red line shows that under small sample sizes of the observational study, the test using $\hat{\psi}$ over-rejects due to errors in nuisance function estimation, which is consistent with our conjecture. Nevertheless, its level promptly converges to 0.05 as the number of samples in the observational study matches or exceeds the RCT. Panel (7-5) demonstrates the witness functions produced as a byproduct of our test, which show mostly positive values and certain negative regions.

the methods by reporting the *rejection rate* in each “bias” setting. To induce different amounts of selection bias into the observational study, we drop patients who were not exposed to the intervention and did not experience the event with some probability p. For further details on data preprocessing, setup, and evaluation, see Appendix E.4.

7.5.1 Setup

We use the Postmenopausal Hormone Therapy (PHT) trial as the RCT in our analysis, which was run on postmenopausal women aged 50-79 years with an intact uterus. The trial investigated the effect of hormone therapy on several types of cancers, cardiovascular events, and fractures, measuring the "time-to-event" for each outcome. In the WHI setup, the observational study component was run in parallel and tracked similar outcomes to the RCT. Our processing of this dataset follows closely to the pre-processing steps taken by Hussain et al. [120]. We binarize a composite outcome, called the "global index", in our analysis, where $Y = 1$ if coronary heart disease, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, or death due to other causes was observed in the first seven years of follow-up, and $Y = 0$ otherwise. Note that $Y = 0$ could also occur from censoring. To establish treatment and control groups in the observational study, we use questionnaire data in which participants confirm or deny usage of combination hormones (i.e. both estrogen and progesterone) in the first three years. For other covariates, we use only those measured in both the RCT and observational study to simplify the analysis.

7.5.2 Results

MMR-Contrast has superior power compared to the baselines in real-world data. As shown in Table 7.1, MMR-Contrast has the best ability to reject studies that have selection bias. Note, as well, that GATE always has lower rejection probability compared to ATE. This result implies that using the GATE approach without prior knowledge on *which* subgroups lead to different effect estimates using observational versus RCT data is highly disadvantageous in terms of statistical power. Though the MMR approach is conceptually similar to GATE, it finds discrepant covariate regions in a data-driven fashion instead of requiring pre-specified groups, thus achieving better power.

7.6 Related Work

Transportability of causal effects: A long line of work gives assumptions under which causal effect estimates can be transported from one population to another. This includes work in statistics on generalizing average effects from one or more RCTs to broader target populations [59, 111, 63, 65], and work in computer science on giving graphical criteria for determining when effects can be transported in more general scenarios [203, 204, 202]. Hartman et al. [111] similarly consider hypothesis testing as a "placebo" test to check assumptions, though their assumptions differ from the ones we consider here. They focus on transporting effects from RCTs to a target population, and do not assume internal validity on observational data. In particular, their assumptions have a testable implication, that the average treated outcome in the target population will match the average treated outcome under a reweighting of the RCT population. Meanwhile, our focus is on testing assumptions of both transportability / external validity, as well as internal validity of observational studies.

Combining experimental and observational data for improved estimation: There is a recent line of work on combining observational and experimental data to yield more precise estimates of causal effects, even when the observational data may be biased [226, 296, 48, 45]. We focus on hypothesis testing as a means of falsification as our primary goal rather than merging data. While Yang et al. [296] use hypothesis testing as a part of their approach, their test depends on the parametric form of the CATE function that they seek to estimate, while our test is nonparametric in nature.

Minimax and variational methods for parameter estimation via CMRs: In addition, there is a growing literature of minimax algorithms that aim to find a well-specified set of model parameters that fulfill a set of CMRs. For example, one line of work looks at constructing a minimax formulation of the generalized method of moments (GMM) framework that aims to estimate highly non-linear model parameters that are also solutions to the moment conditions implied by the problem at hand, e.g.

IV regression [163, 69, 27, 26]. Metzger [180] also gives asymptotic theory for minimax estimators of functionals in CMRs. Another line of work tackles the case where a set of CMRs may only weakly identify the nuisance functions, though the target parameter may still be efficiently and uniquely estimable under some conditions on the estimator [138, 28].This literature focuses on the problem of parameter estimation, while we use the CMR formulation for hypothesis testing of transportability and internal validity assumptions.

7.7 Discussion & Limitations

We have proposed a novel approach for falsifying the assumptions of observational studies using experimental data. These causal assumptions, stating the internal and external validity of observational and experimental data, imply that the conditional average treatment effect is equivalent across all observed subpopulations of X in both the observational and experimental data. This in turn gives rise to testable restrictions on the combined data distribution, implying, as we show, that the difference between two functions of the data is zero-mean for any subset of X. Recent advances in the econometrics literature allow us to test such restrictions. Our approach implicitly searches for regions of X where the CATE estimates disagree between the observational and experimental data, without the need for pre-specifying these subpopulations. Moreover, this approach yields a function that characterizes the regions where disagreement is large. Finally, we design our test to avoid rejecting studies due to differences in baseline factors that do not influence the treatment effect.

However, our approach shares certain limitations with some methods in the literature on testing for violation of causal assumptions. In particular, while violations of the CMRs imply violations of causal assumptions, this does not directly tell us which assumptions are violated (e.g., whether the observational study is subject to hidden confounding, or whether there is simply an unobserved difference between the RCT and observational populations). Finally, due to the fact that policy guidelines can be formed from such RCTs and observational studies in society, it is important for practitioners to consider the biases in the data as well as the aforementioned limitation of our method.

Chapter 8

Evaluating Physician-AI Interaction for Precision Oncology via a Clinical Decision Support System Prototype

Acknowledgement of Co-authors I would like to acknowledge Fernando Acosta Perez, who was integral in the software development for this chapter, and Barbara Lam, who I had weekly brainstorming sessions with to flesh out the ideas in this chapter. Finally, I would like to acknowledge Dr. Andrew Yee who gave insightful clinical guidance throughout the duration of this project.

In this thesis, so far, I have looked at the various methodological and statistical components, both from a predictive and causal perspective, necessary for a CDSS in oncology. Now, in this chapter, I try and understand how physicians might interact with such a system, especially when having to integrate it with types of evidence that they use in their current clinical practice.

8.1 Introduction

Cancer treatment has changed dramatically over the past decade. Historically, the field was driven forward by landmark clinical trials demonstrating the efficacy of various cancer-directed therapies and surgical interventions [92, 85, 151, 150, 73, 12, 52, 13].

Over the past decade, however, advances in molecular sequencing and innovative therapies such as immune checkpoint inhibitors, bispecific antibodies, and CAR T-cells have rapidly evolved the treatment landscape [78, 256, 175]. The highly personalized characterization of each patient's disease coupled with the explosion in treatment options has made it increasingly difficult to capture every patient scenario in a clinical trial setting.

Multiple myeloma (MM) is often highlighted as an area where the rapid development of new therapies has outpaced our ability to conduct clinical trials [216, 262]. The FDA approved over a dozen new agents for MM in the last 10 years [81] and registry studies have demonstrated that approximately 40% of real-world patients with MM do not meet inclusion criteria for the phase 3 trials that led to treatment approvals [262]. While randomized controlled trials (RCTs) are considered to produce the highest caliber of evidence in medicine [39], the reality is that they cannot account for every patient scenario. The gap between clinical trial and real-world settings is exacerbated in the field of MM, where patients are expected to need multiple lines of therapy and clinicians prefer to treat with multi-drug combinations [216].

Clinical decision support systems (CDSSs) offer an alternate path to the realization of precision oncology. Modern CDSSs can leverage machine learning (ML) models that have been trained on vast amounts of data in order to provide personalized treatment recommendations. While several models for MM and other cancers have been proposed and published in the literature [6, 41, 193, 299, 215], few models in general make it to the bedside due to challenges in clinical validation and implementation [265]. Surveys of clinicians have shown that they believe ML has a role to play in clinical care [76, 198, 234]. Studies have gone on to explore how clinicians use ML recommendations for treatment selection in depression [129], how clinicians perceive ML recommendations in terms of trust, interpretability, and diagnostic accuracy [94], and how ML recommendations might influence decisions [1]. However, little work has been done to understand how clinicians might interact with AI or ML-based systems in conjunction with current standard-of-care evidence used in clinical practice, which are often data from RCTs.

Currently, clinicians are tasked with synthesizing all available evidence to maximize patient survival and quality of life while minimizing adverse events. As ML systems proliferate, clinicians will be forced to reconcile findings from RCTs and ML models. If an ML model produces similar outcomes as an RCT, a clinician may feel more confident in their decision. Inevitably, however, an ML model will produce personalized evidence that contradicts RCT data. Understanding how clinicians navigate these complex scenarios when selecting treatments can better prepare us for the implementation of an ML-based CDSS in real-world clinical settings.

In this study, we design a CDSS that displays survival and adverse event data from a synthetic RCT and ML model and evaluate how clinicians synthesize the available data to make treatment decisions for twelve patients diagnosed with multiple myeloma.

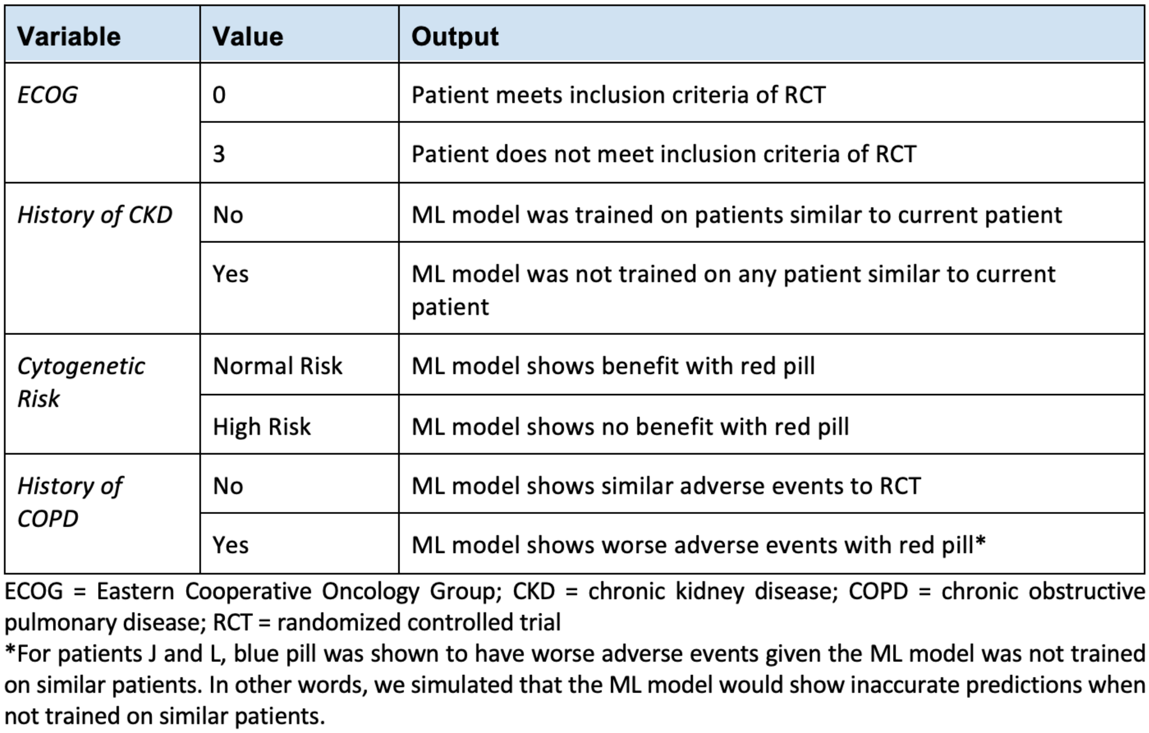
8.2 Methods

8.2.1 Study Design

We created a survey study to evaluate how clinicians synthesize evidence from an RCT and ML model. Participants were presented with survival curves and adverse event outcomes from an RCT and an ML model for 12 different patient scenarios (A-L). They were then asked to select a treatment ("red pill" or "blue pill"), to rate their confidence in their treatment choice on a Likert scale from 1-10, and when ML data was available, to rate their perceived reliability of the model on a Likert scale from 1-10. This study was deemed exempt by the Massachusetts Institute of Technology Institutional Review Board (E-4559, Decision Support Tool for Treatment Selection in Multiple Myeloma).

8.2.2 Patient Scenarios and RCT/ML Combinations

Each patient scenario was based on the same "synthetic" clinical vignette of a patient with multiple myeloma. However, there were four changing variables: history of CKD (or not), history of COPD (or not), cytogenetic risk profile (normal risk versus high-



Variable	Value	Output
ECOG	0	Patient meets inclusion criteria of RCT
	3	Patient does not meet inclusion criteria of RCT
History of CKD	No	ML model was trained on patients similar to current patient
	Yes	ML model was not trained on any patient similar to current patient
Cytogenetic Risk	Normal Risk	ML model shows benefit with red pill
	High Risk	ML model shows no benefit with red pill
History of COPD	No	ML model shows similar adverse events to RCT
	Yes	ML model shows worse adverse events with red pill*

ECOG = Eastern Cooperative Oncology Group; CKD = chronic kidney disease; COPD = chronic obstructive pulmonary disease; RCT = randomized controlled trial

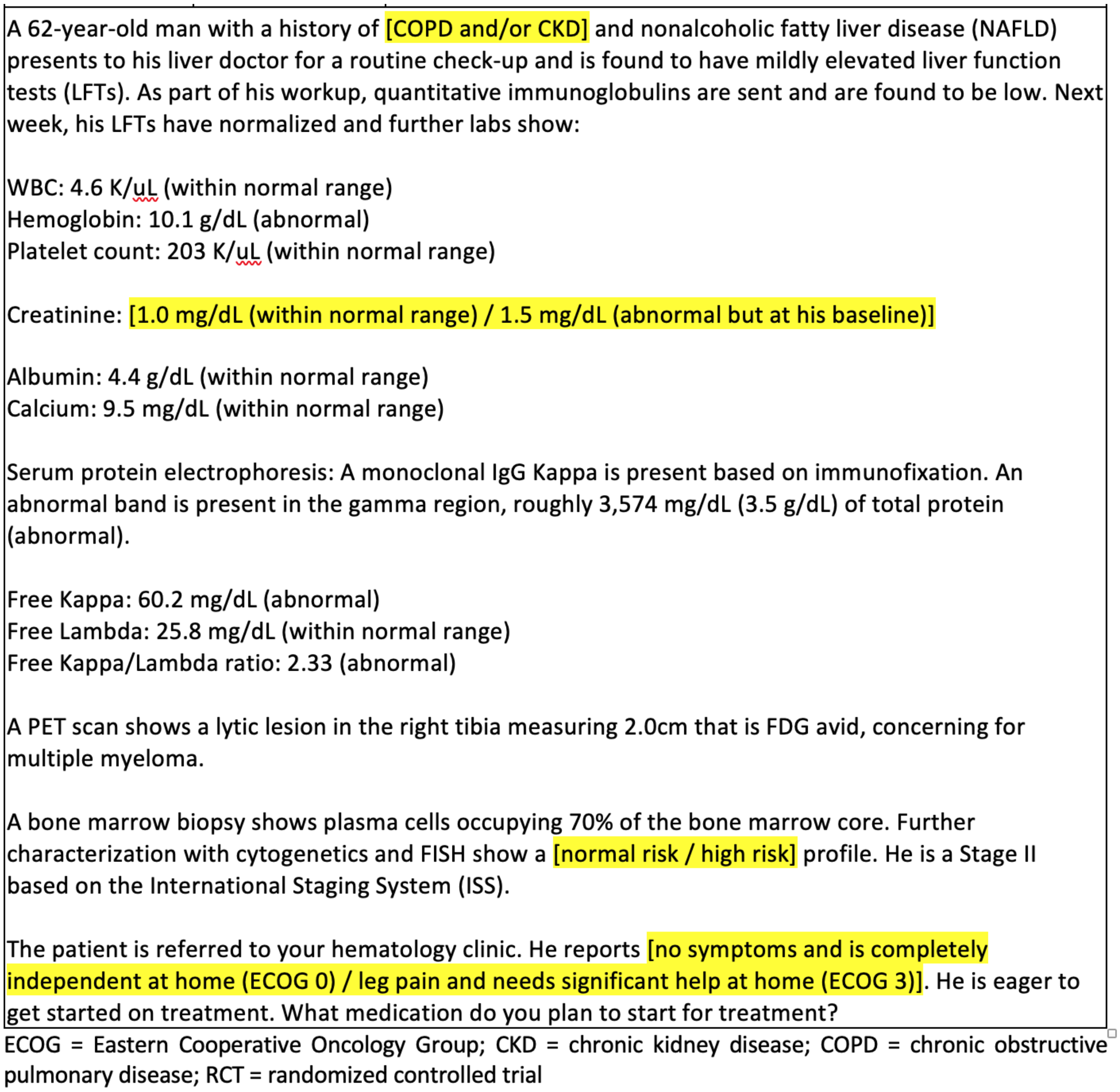
*For patients J and L, blue pill was shown to have worse adverse events given the ML model was not trained on similar patients. In other words, we simulated that the ML model would show inaccurate predictions when not trained on similar patients.

Figure 8-1: Possible values for the baseline variables in the patient scenarios

risk), and ECOG status (0 versus 3). These variables determined whether the patient met the inclusion criteria of the RCT and influenced the ML model's predictions of the patient's survival and adverse event outcomes (Figures 8-1 & 8-2). We generated and presented all possible combinations of variables as patient scenarios. We excluded scenarios in which the patient 1) met RCT inclusion criteria and 2) the ML model was not trained on any similar patients in an attempt to shorten the survey and remove less compelling scenarios. Eight of the twelve scenarios were constructed so that the patient did not meet the inclusion criteria of the RCT in order to mimic this common experience in clinical practice. All RCT and ML combinations are shown in Figure 8-3 and were presented in a randomized fashion to participants.

8.2.3 Tiered Information Approach

Each patient scenario was associated with three to four "tiers" of information (Figure 8-3). With each subsequent tier, the participant received increasing amounts of data to synthesize and interpret. Tier 1 provided RCT (population-level) outcomes only.



A 62-year-old man with a history of [COPD and/or CKD] and nonalcoholic fatty liver disease (NAFLD) presents to his liver doctor for a routine check-up and is found to have mildly elevated liver function tests (LFTs). As part of his workup, quantitative immunoglobulins are sent and are found to be low. Next week, his LFTs have normalized and further labs show:

WBC: 4.6 K/uL (within normal range)
Hemoglobin: 10.1 g/dL (abnormal)
Platelet count: 203 K/uL (within normal range)

Creatinine: [1.0 mg/dL (within normal range) / 1.5 mg/dL (abnormal but at his baseline)]

Albumin: 4.4 g/dL (within normal range)
Calcium: 9.5 mg/dL (within normal range)

Serum protein electrophoresis: A monoclonal IgG Kappa is present based on immunofixation. An abnormal band is present in the gamma region, roughly 3,574 mg/dL (3.5 g/dL) of total protein (abnormal).

Free Kappa: 60.2 mg/dL (abnormal)
Free Lambda: 25.8 mg/dL (within normal range)
Free Kappa/Lambda ratio: 2.33 (abnormal)

A PET scan shows a lytic lesion in the right tibia measuring 2.0cm that is FDG avid, concerning for multiple myeloma.

A bone marrow biopsy shows plasma cells occupying 70% of the bone marrow core. Further characterization with cytogenetics and FISH show a [normal risk / high risk] profile. He is a Stage II based on the International Staging System (ISS).

The patient is referred to your hematology clinic. He reports [no symptoms and is completely independent at home (ECOG 0) / leg pain and needs significant help at home (ECOG 3)]. He is eager to get started on treatment. What medication do you plan to start for treatment?

ECOG = Eastern Cooperative Oncology Group; CKD = chronic kidney disease; COPD = chronic obstructive pulmonary disease; RCT = randomized controlled trial

Figure 8-2: Patient scenario text

Tier 2 provided ML (patient-level outcomes). Tier 3 provided information about how the ML model was trained, i.e. whether the patient was represented in the training cohort, and external validation results (Figure 8-4). Tier 3 is described as "ML model with context" moving forward. Three patient scenarios (C, E, I) included tier 4 data, which described the result of using the ML model to replicate RCT results by running a pseudo-experiment on an observational cohort that matches the RCT cohort (Figure 8-4).

All survival outcomes were presented as progression free survival curves. Adverse event data was presented as the frequency of different symptoms. Cytopenias, upper

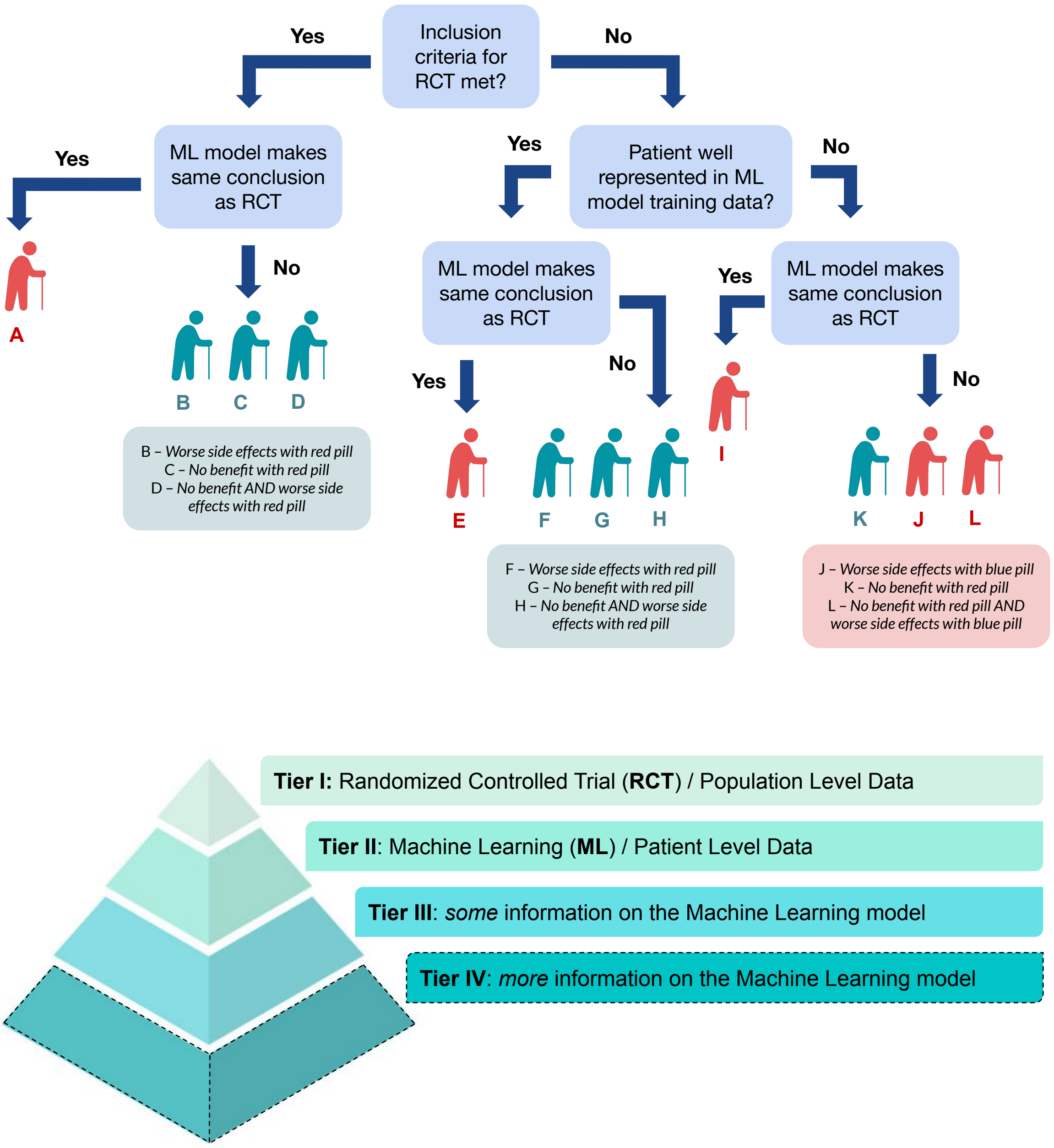
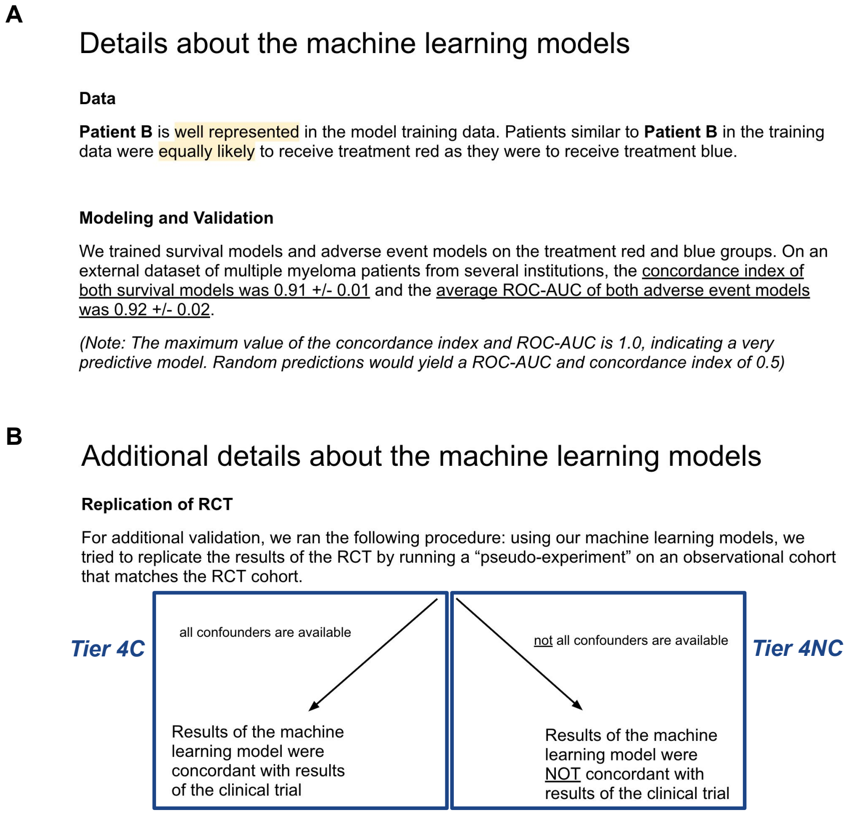

Inclusion criteria for RCT met?
Yes
No
ML model makes same conclusion as RCT
Yes
No
A
B C D
B – Worse side effects with red pill
C – No benefit with red pill
D – No benefit AND worse side effects with red pill
Patient well represented in ML model training data?
Yes
No
ML model makes same conclusion as RCT
Yes
No
E
F G H
F – Worse side effects with red pill
G – No benefit with red pill
H – No benefit AND worse side effects with red pill
ML model makes same conclusion as RCT
Yes
No
I
K J L
J – Worse side effects with blue pill
K – No benefit with red pill
L – No benefit with red pill AND worse side effects with blue pill
Tier I: Randomized Controlled Trial (RCT) / Population Level Data
Tier II: Machine Learning (ML) / Patient Level Data
Tier III: some information on the Machine Learning model
Tier IV: more information on the Machine Learning model

Figure 8-3: Top: Combinations of RCT and ML outcomes presented to survey participants. Note that the RCT results indicate survival benefit with the red pill and largely similar adverse events compared to the blue pill. **Bottom**: Each patient scenario was associated with one ot four tiers of data.

respiratory tract infections, and infusion-related reactions varied most significantly.

At tier 1, participants were told whether their patient met RCT inclusion criteria. Subsequently, they were asked to prescribe a treatment and rate their confidence in treatment choice. At tiers 2 and 3, participants were asked to prescribe a treatment, rate their confidence in treatment choice, and their perceived reliability of the ML model. At tier 4, participants were asked to rate their confidence in treatment choice and their perceived reliability of the model if the ML model results were concordant with the RCT results (“tier 4C”) and if they were not (“tier 4NC”).



Details about the machine learning models

Data

Patient B is well represented in the model training data. Patients similar to **Patient B** in the training data were equally likely to receive treatment red as they were to receive treatment blue.

Modeling and Validation

We trained survival models and adverse event models on the treatment red and blue groups. On an external dataset of multiple myeloma patients from several institutions, the concordance index of both survival models was 0.91 +/- 0.01 and the average ROC-AUC of both adverse event models was 0.92 +/- 0.02.

(Note: The maximum value of the concordance index and ROC-AUC is 1.0, indicating a very predictive model. Random predictions would yield a ROC-AUC and concordance index of 0.5)

Additional details about the machine learning models

Replication of RCT

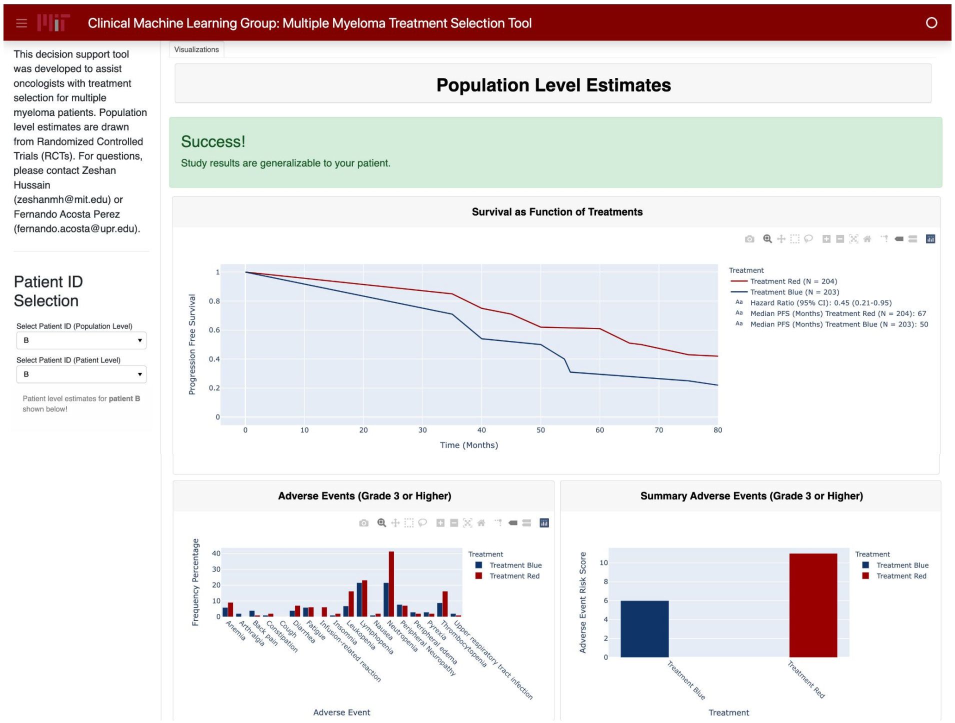
For additional validation, we ran the following procedure: using our machine learning models, we tried to replicate the results of the RCT by running a “pseudo-experiment” on an observational cohort that matches the RCT cohort.

Tier 4C
all confounders are available
Results of the machine learning model were concordant with results of the clinical trial
not all confounders are available
Results of the machine learning model were NOT concordant with results of the clinical trial
Tier 4NC

Figure 8-4: Top: Tier 3 information about the machine learning model, **Bottom**: Tier 4 information about the machine learning model.

8.2.4 Study Tools

Participants used a web-based CDSS we created for the study called the “Multiple Myeloma Decision Support Tool” (MM-DST) to view the RCT and ML outcomes (Figure 8-5). All outcomes were synthetic in nature so as not to bias participants who may have opinions about existing RCTs and ML models. Participants used a Qualtrics survey to review the patient scenarios and rank their confidence and reliability scores (Figure 8-6). Links to both tools are provided in the supplementary materials.



Clinical Machine Learning Group: Multiple Myeloma Treatment Selection Tool

This decision support tool was developed to assist oncologists with treatment selection for multiple myeloma patients. Population level estimates are drawn from Randomized Controlled Trials (RCTs). For questions, please contact Zeshan Hussain (zeshanmh@mit.edu) or Fernando Acosta Perez (fernando.acosta@upr.edu).

Patient ID Selection

Select Patient ID (Population Level): B

Select Patient ID (Patient Level): B

Patient level estimates for **patient B** shown below!

Population Level Estimates

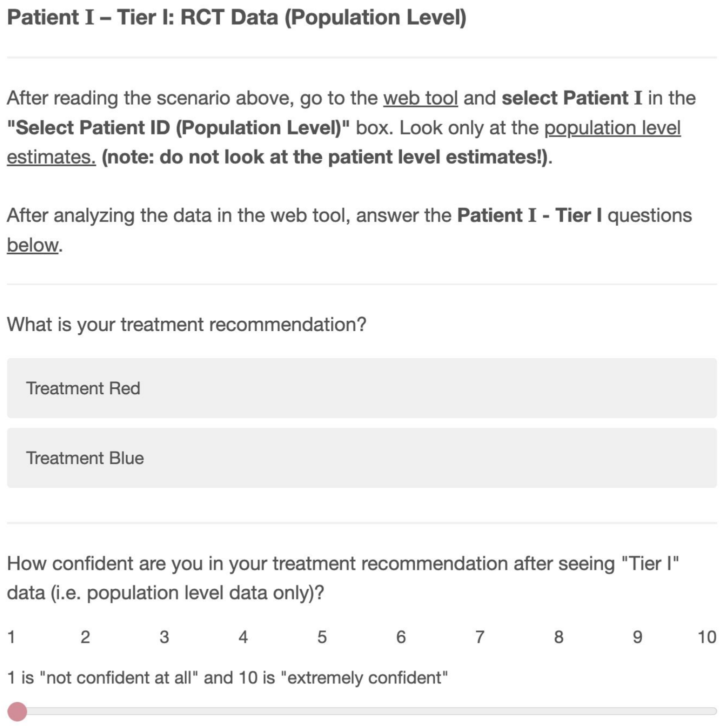
Success!

Study results are generalizable to your patient.

Figure 8-5: View of the CDSS tool. The decision support tool displays survival curves and adverse event estimates from the synthetic RCT study and the ML model.

8.2.5 Data Collection

The survey was piloted with two Hematology-Oncology fellows for clarity, length, and clinical relevance. Participants completed the survey remotely on their own computers. A 5-minute introductory video was provided at the start of the survey explaining how to interact with the survey tools. All participants were asked basic demographic information including age and level of training, and asked to rate on a Likert scale of 1-5 their comfort managing patients with multiple myeloma, interpreting survival curves and adverse event information in RCTs, comfort with machine learning, and with causal inference. At the end of the survey, participants were asked how they evaluated adverse event data, how they approached treatment selection when their patients did not meet inclusion criteria, how they believe ML-driven estimates impact clinical decision making, and how they interpreted the tier 4 replication procedure.



Patient I – Tier I: RCT Data (Population Level)

After reading the scenario above, go to the web tool and **select Patient I** in the **"Select Patient ID (Population Level)"** box. Look only at the population level estimates. **(note: do not look at the patient level estimates!)**.

After analyzing the data in the web tool, answer the **Patient I - Tier I** questions below.

What is your treatment recommendation?

Treatment Red

Treatment Blue

How confident are you in your treatment recommendation after seeing "Tier I" data (i.e. population level data only)?

1 2 3 4 5 6 7 8 9 10

1 is "not confident at all" and 10 is "extremely confident"

Figure 8-6: View of the Qualtrics survey.

For these end-of-survey questions, we provided free text boxes for their responses.

8.2.6 Recruitment

Between January 2023 and April 2023, we recruited physicians in Internal Medicine and Hematology and Oncology from academic institutions via email to participate in our study. Additional participants were recruited through snowball sampling; i.e. participants were asked to recommend additional physicians for recruitment. Participants received up to two email reminders to participate and were offered a $50 Amazon gift card as incentive.

8.2.7 Follow-up Qualitative Interview

A qualitative assessment of one patient scenario, K, was done after survey results were analyzed and the scenario results were found to not fit our initial hypothesis. All participants were invited to participate in a follow up exit interview during which they were asked to "think aloud" as they worked through scenario K. Afterwards, participants were shown how others had responded to scenario K and asked to comment on those results. At least one researcher (BDL or ZH) was present during the interview and took field notes. The interview protocol is provided in the supplementary materials.

Demographic	$N = 32$, n(%)
Age	
21-30 y.o.	16 (50%)
31-40 y.o.	16 (50%)
Sex	
Male	23 (72%)
Female	9 (28%)
Race	
White Caucasian	22 (69%)
Asian	7 (22%)
Hispanic	2 (6%)
Black or African American	1 (3%)
Clinical Role	
Internal Medicine Resident	16 (50%)
Hematology Oncology Fellow	13 (41%)
Hematology Oncology Attending	3 (9%)
Clinical Specialty*	
General Medicine	15 (47%)
Hematology	11 (34%)
Oncology	10 (31%)
Other	2 (6%)

Table 8.1: Descriptive Statistics of the User Study Cohort, *Participants could select more than one, thus totals may sum to greater than 100%

8.2.8 Data Analysis

We used descriptive statistics to analyze respondent characteristics (Table 8.1). For each scenario, we ran two-sample paired t-tests to compare the changes in confidence and reliability between tier 2 versus tier 1 (ML versus RCT data), and tier 3 versus tier 2 (ML model with context versus without). For all t-tests, the null hypothesis was that there would be no difference. Because t-tests have a loose requirement for the data to be normally distributed, we utilized Shapiro-Wilk tests to assess normality of the confidences and the reliability measurements. A power analysis for the paired t-tests with continuous outcomes revealed a requirement of 34 physicians to detect a mean difference of 1 with a standard deviation of 2 (given that the range of the outcomes is 1-10)[1]. Due to the number of statistical tests we ran, all p-values are reported after adjustment via the Holm-Bonferroni correction. We also ran McNemar's tests with Holm-Bonferroni correction to assess the difference in proportions of blue pill selections at different tiers such that the extent of treatment switching can be characterized.

Two researchers (BDL, ZH) analyzed the end-of-survey free text responses and exit interview field notes using Braun and Clarke's methods for thematic analysis [34]. They first reviewed all responses and independently created a list of themes to reflect the data. They then discussed the themes and iteratively developed a final codebook by consensus. The two researchers used this codebook to independently code the responses, assigning up to two codes per responses. The researchers compared their coding assignments and reached consensus for every response.

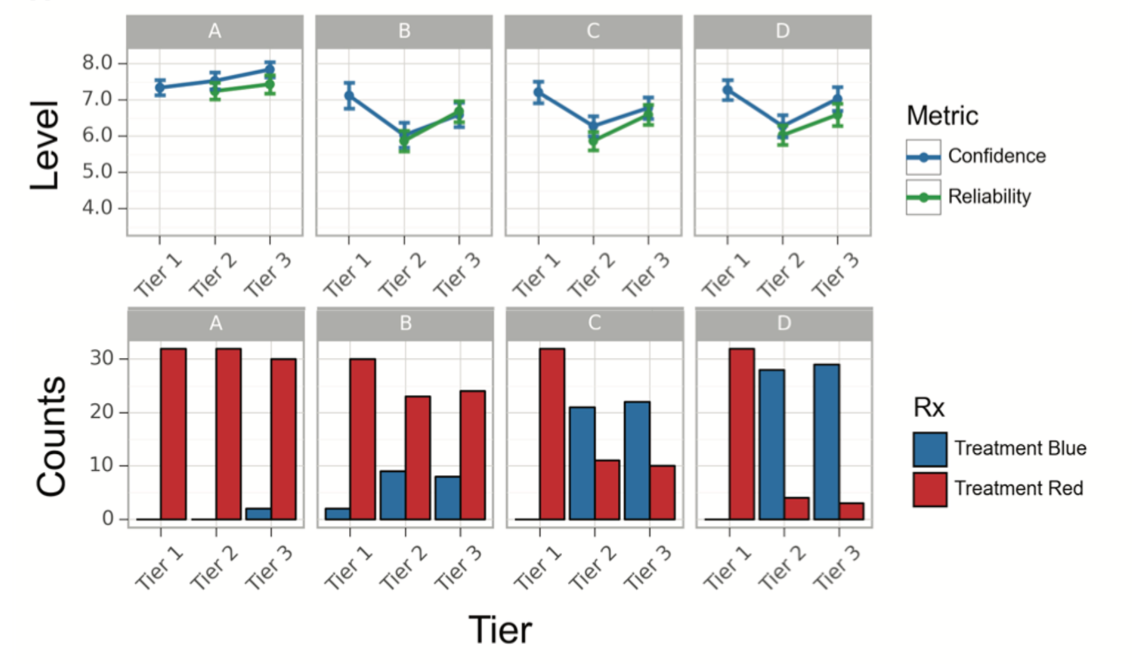
8.3 Results

A total of 284 physicians were invited to participate in the study. 32 ultimately participated, for a response rate of 11.3%. Half were Internal Medicine residents (physicians in training) and half were Hematology and Oncology fellows and attendings (Table 8.1).

8.3.1 Quantitative Analysis

Patient meets inclusion criteria for RCT and ML model (scenarios A, B, C, D) Average confidence at tier 1 across these four scenarios was 7.24, higher than the average confidence at tier 1 across the other eight scenarios (5.99). In scenario A, the ML model results were concordant with the RCT results, showing that the

[1] With 32 physicians recruited, we are able to detect a mean difference of 1 with a standard deviation slightly under 2: ≈ 1.9.

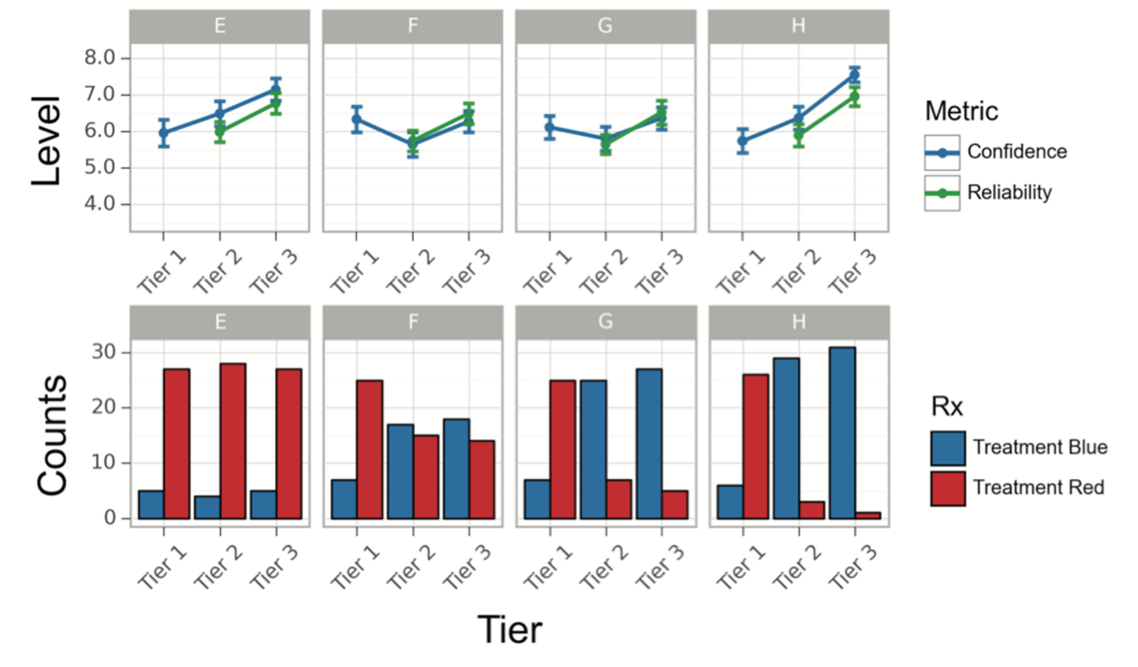


A: ML model shows benefit with red pill and similar adverse events
B: ML model shows benefit with red pill but worse adverse events with red pill
C: ML model shows no benefit with red pill and similar adverse events
D: ML model shows no benefit with red pill and worse adverse events with red pill

Figure 8-7: In these scenarios, the patient meets inclusion criteria for RCT and is well represented in training data of ML model. We measured confidence and perceived reliability at the different tiers as well as selected treatment counts.

red pill results in improved outcomes with similar adverse event outcomes. After being shown tier 3 results (RCT data, ML data, and ML context), physicians had the highest confidence across all scenarios at 7.84 (+/- 1.18) in selecting red pill as their treatment option.

In scenarios B, C, and D, the ML model results were discordant with the RCT results. In scenario B, despite the ML model showing worse adverse events with red pill, the majority of participants chose to treat with red pill though their confidence decreased ($p = .05$). In scenario C, the ML model showed no benefit with red pill and an increasing proportion of participants switched to blue pill (*McNemar's test*: $p = 5.1 \times 10^{-4}$). In scenario D, the ML model showed no benefit with red pill **and** worse adverse events with red pill, with even more participants switching to blue pill (*McNemar's test*: $p = 1.5 \times 10^{-5}$) (Figure 8-7). Across scenarios B, C, and D, confidence tended to decrease after seeing tier 2 ML data (B: $p = 0.05$, C: $p = 0.05$, D: $p = 0.36$), but tended to increase after participants reviewed tier 3 ML with context data (B: $p = 0.002$, C: $p = 0.17$, D: $p = 0.06$), which stated that the ML model had been trained on patients like the one in the scenario.

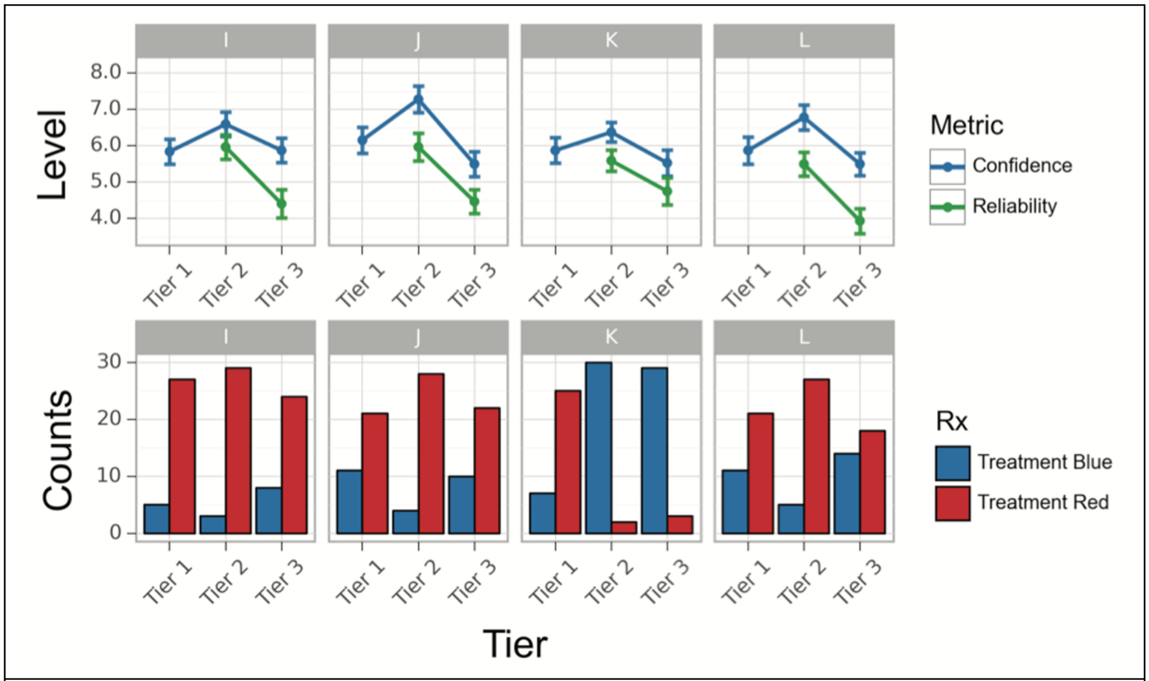


E: ML model shows benefit with red pill and similar adverse events
F: ML model shows benefit with red pill and worse adverse events with red pill
G: ML model shows no benefit with red pill and similar adverse events
H: ML model shows no benefit with red pill and worse adverse events with red pill

Figure 8-8: In these scenarios, the patient does **not** met inclusion criteria for RCT but is well represented in the training data of the ML model. We measured confidence and perceived reliability at the different tiers as well as selected treatment counts.

Patient does not meet inclusion criteria for RCT but ML model was trained on data that represents the patient well (scenarios E, F, G, H) Similar to scenario A, in scenario E when the ML model results were concordant with the RCT results, confidence and perceived reliability increased across tiers. The starting confidence was lower however, at 5.97 (+/- 2.10). In scenario F, when the ML model showed benefit with red pill but worse adverse events, about half of the participants chose to treat with blue pill instead. Confidence and reliability at first remained stable with tier 2 ML data ($p = 0.43$), then increased with tier 3 ML with context data

($p = 0.01$). In scenarios G and H when the ML model showed no benefit with red pill, the majority of participants chose to treat with blue pill, with a statistically significant change in the proportion of participants choosing the blue treatment (*McNemar's test* – G:$p = 0.002$, H:1.9×10^{-4}). As in scenario F, no statistically significant change in confidence was noted at tier 2 (G: $p = 1.$, H: $p = .43$), but there was an increase in confidence and perceived reliability when shown tier 3 ML with context data (G: $p = .18$ [confidence], $p = 0.006$ [reliability]; H: $p = 0.002$ [confidence], $p = 2.2 \times 10^{-5}$ [reliability]). These results are shown in Figure 8-8.



I: ML model shows benefit with red pill and similar adverse events
J: ML model shows benefit with red pill but worse adverse events with blue pill
K: ML model shows no benefit with red pill and similar adverse events
L: ML model shows no benefit with red pill and worse adverse events with blue pill

Figure 8-9: In these scenarios, the patient does **not** met inclusion criteria for RCT and is **not** well represented in the training data of the ML model. We measured confidence and perceived reliability at the different tiers as well as selected treatment counts.

Patient does not meet inclusion criteria for RCT and ML model was not trained on data that represents the patient well (scenarios I, J, K, L) In these scenarios, confidence was low and both confidence and reliability decreased when participants were told that the ML model was not trained on patients like theirs. In scenario I, when the RCT and ML model results were concordant but neither represented the patient well, most participants chose to treat with red pill. In scenario J, the ML model showed worse adverse events with blue pill and most participants maintained their choice of treating with red pill. In scenario K, the ML model showed no benefit with red pill and the majority of participants switched treatment to blue pill (*McNemar's test*, $p = 1.9 \times 10^{-4}$) and maintained their choice even after being told that the ML model had not been trained on patients like theirs. In scenario L, the ML model showed fully discordant results and by tier 3 about half of the participants chose to treat with blue pill and half chose to treat with red pill.

Four of the 32 participants elected to participate in an exit interview where they were asked to "think aloud" as they worked through scenario K. Three participants opted to start with treatment red when they only had tier 1 RCT data. All participants used treatment blue when they were provided with tier 2 ML data. However, when provided tier 3 ML with context data, all participants except one switched back to the red treatment. Four major themes were identified in analysis: (1) there is variability in what clinical factors participants use in their decision-making, (2) there are perceived advantages to an ML model over RCT data, (3) uncertainty and unease around decision-making when participants learned that the ML model had not been trained on similar patients, and (4) the perception that these types of studies are important thought exercises.

Participants described evaluating ECOG status and its potential reversibility, comorbidities and how they might affect drug metabolism, and the perceived impact of adverse events on quality of life to make treatment decisions. They described advantages of an ML model over an RCT including expressivity, training on real-world data, improved external validity, and the ability to give individualized recommendations. When shown tier 3 ML with context data, one participant said, "Now I have no idea!" and another stated, "There's no good, right answer." When shown the finding that most of the 32 participants had remained with treatment blue despite learning that the ML model had not been trained on patients like theirs, none of the participants expressed surprise. Several shared that they also felt conflicted in their treatment choice, and some thought that physicians may be less likely to identify weaknesses in an ML model because they are not trained to critically appraise them, whereas they are more trained to critically think about RCT data. One participant noted, "ML models in general do not seem to do as well extrapolating to things outside their training sets," and explained that they would rather a human extrapolate RCT findings than use an ML model. In contrast, another participant appreciated being told the lack of support in the ML training data: "I thought of it as a disclaimer and not a big deal since [the model] had good performance," and went on to say, "It gives me more confidence because it's being honest about it." All participants reported that the survey study was a helpful thought exercise. In the words of one participant: "It forced me to think through data, how it applies to my patient, and how to maximize benefit."

8.3.2 Qualitative Analysis of Open-ended Questions

When asked about how they evaluate adverse event data, the majority of participants described looking for specific organ side effects and some discussed evaluating that in the context of their patient's comorbidities. One participant reported that the total number of adverse events was important in decision-making. When asked how they approach treatment selection when their patients do not meet inclusion criteria, the majority of participants reported that adverse events became more important. Several also described working to balance the benefits of survival outcomes with adverse event data.

When asked how machine learning-driven results impact clinical decision-making, the majority of participants reported that they find this type of data helpful and that they would incorporate it into their decision-making. "[RCT] results are much harder to generalize for patients who don't nicely meet the inclusion criteria, and we usually have to rely more on anecdote and clinical gestalt. With support from machine-learning datasets, we could feel more confident in making more personalized treatment plans and decisions with our patients," one participant wrote. Several emphasized that concordant results between the RCT and ML model were especially appreciated and could increase confidence: "When they were in agreement, it was a nice confirmation. When there was discrepancy, it was easy to throw out." A few reported that ML model estimates are not helpful with one participant stating, "they may bias towards increasing confidence but [they] rarely changed my mind."

Participants were also asked to describe the replication procedure (tier 4) in their own words. Several participants were not sure or provided an incorrect definition, but the majority were able to do so at a superficial level: "Confirming the ML model was reliable" or "Verifying reproducibility." Several also correctly described the procedure as "simulating" an RCT using observational data.

8.3.3 Replication Experiment

Participants were presented with tier 4 data in three scenarios (C, E, I). When told that the replication experiment was successful (ML model results concordant with RCT results), confidence in treatment increased in scenario E ($p = 5.5 \times 10^{-4}$) and perceived model reliability tended to increase in all three scenarios (C: $p = 0.22$, E: $p = 0.02$, I: $p = 0.17$), though not with statistical significance in scenarios C and I. When told that the replication experiment failed (ML model results discordant with RCT results), confidence in treatment and perceived reliability decreased, particularly for scenario E (confidence: $p = 5.5 \times 10^{-4}$, reliability: $p = 5 \times 10^{-5}$). Participants were presented with tier 4 data in three scenarios (C, E, I). When told that the replication experiment was successful (ML model results concordant with RCT results), confidence in treatment and perceived model reliability increased in scenario E ($p = 5.5 \times 10^{-4}$ [confidence], $p = .02$ [reliability]). In scenarios C and I, neither confidence nor reliability changed with statistical significance ($p > 0.05$). When told that the replication experiment failed (ML model results discordant with RCT results), confidence in treatment and perceived reliability decreased for scenario E ($p = 1.3 \times 10^{-5}$ [confidence], $p = 1.1 \times 10^{-6}$ [reliability]). The full set of results and results of hypothesis testing are shown in Appendix F.

8.4 Discussion

In our study, several key themes were revealed throughout the different scenarios. First, physician confidence is highest when RCT and ML findings are concordant. Second, physicians favor treatments that lead to better survival outcomes, whether that survival benefit is suggested by an RCT or an ML model. Third, participants are likely to follow ML model estimates even before any context around how the model was trained or validated is given.

In scenario A, where the patient meets inclusion criteria for the RCT, is well represented by the ML model training data, and the RCT and ML model estimates are concordant, participant confidence was at its highest. This finding was also reflected in the open-ended answers at the end of the survey when participants described increased confidence with concordant data. In scenarios where the results are discordant, a large proportion of participants switched to the treatment supported by the ML model. This suggests physicians prefer to prioritize survival outcomes, which is consistent with prior reports in the cancer literature [269]. Of note, this prioritization took place whether the survival benefit was estimated by an RCT or an ML model, and before participants received additional information about how the model was trained and validated. It also often occurred with no statistically significant change in confidence ($\alpha = 0.05$).

In scenarios where the patient did not meet inclusion criteria for the RCT but was represented in the ML training data—a clinical situation that may be increasingly common in the future—participants were more likely to adhere to ML model estimates. Interestingly, in scenarios where the patient did not meet inclusion criteria for the RCT and was not represented in the ML training data, the majority of participants continued to rely on ML-generated data to make treatment decisions. In contrast, in our qualitative assessment of scenario K, the majority of participants favored the RCT-supported treatment when they were told that the ML model was not trained on patients like theirs. This contradiction to what we found in the survey suggests there may have been potential biases at play. Participants may make different decisions when being evaluated by a researcher. Our survey was voluntary, and the majority of participants may be more likely to view ML favorably. Regardless, our contradictory finding raises questions of how clinicians will interact with ML models when they are not being observed. The interviewees expressed uncertainty and decreased confidence when their patient was not represented in either the RCT or ML model, though they were "not surprised" that most survey respondents had relied on the ML estimates regardless. They suggested there may be overreliance on ML estimates because clinicians are not trained to interpret ML models the way they are trained to interpret RCTs, *implying a consistent mental model about how to think about RCT data, whereas the same does not exist for ML models*. As validated and unvalidated ML models enter the clinical realm, it may become inevitable that clinicians need to learn to critically appraise them. Indeed, national organizations such as the American Medical Association (AMA) have called for increasing physician education on artificial intelligence (AI) so that physicians can better understand and interpret models [173].

Taken together, the above results suggest physicians may incorporate ML estimates into their clinical decision-making without appropriately evaluating the quality of the model, making it critically important to filter which models make it to the bedside or determine novel ways to calibrate a clinician's mental model about ML. One approach is to run prospective clinical trials assessing the efficacy of ML models in improving clinical outcomes. Though there are calls to make RCTs the gold standard for evaluating algorithms [139], few clinical ML models are evaluated with RCTs [154, 209], due to cost and time required to run them. Similar to what has happened with the proliferation of cancer-directed therapies, the development of ML models will likely outpace our ability to conduct robust trials.

Research evaluating other ways to sanity check ML models is needed, and our study findings suggest that benchmarking ML models trained on observational data against experimental data may be meaningful for clinicians, especially when the RCT does not apply to the patient (e.g. scenario E). Three scenarios had "tier 4" data where participants were provided information about the ability to replicate the results of the RCT using the ML model trained on observational data. In scenario E, while confidence in treatment and perceived reliability of the model did increase when the model results were concordant with the RCT (successful replication), an even more significant change in confidence and perceived reliability occurred when the model results were not concordant (unsuccessful replication). This type of replication procedure may be useful for calibrating a clinician's mental model of the ML data. We based tier 4 information on the work from Chapters 6 and 7, which formalize how one can assess the reliability of estimates from an observational study using RCT data, thereby casting doubt on ML models trained on observational data to make predictions about survival or adverse events. Other work has explored how to emulate a clinical trial from observational data [113, 89, 93].

One suggestion for improving model safety is to make models more explainable so that users can more readily spot errors [71]. On the other hand, one study showed that an explainable model may make it more difficult for users to detect errors, perhaps due to information overload [211]. Methods to improve explainability can hide biases [18] and may be interpreted differently across users, as demonstrated by our exit interviews. After learning that the ML model was not trained on representative patients, one participant reported, "Now I would put less weight on the model," whereas another stated that they appreciated the "disclaimer" and this transparency increased their trust in the model.

How the ML model results are displayed may also influence users. We chose to design a CDSS where the ML model provided outcomes without making recommendations, based on a prior study showing that ML models that give prescriptive advice rather than descriptive evidence are more likely to bias users [1]. Further research to explore how user interfaces and wording of recommendations might influence clinicians is needed before ML models can be safely deployed in the clinical setting.

Calibrating clinicians' mental model of AI As we have alluded to above, our results indicate a need to develop methods and strategies through which a clinician's mental model of ML techniques can be calibrated before they engage with ML-based systems. One approach found in existing literature is to artificially alter the confidence of the ML model based on how miscalibrated the human is on a certain task to maximize overall human-AI performance [274]. In a clinical setting, this approach might be equivalent to increasing the confidence of the ML model's treatment recommendation when the clinician is also very confident about selecting the same treatment, so as not to sway the clinician. However, intentional miscalibration of the AI might not be appropriate in high-stakes settings like healthcare and would require substantial regulatory checks.

Other approaches to improving a clinician's mental model of the AI could be through teaching strategies, which explicitly form a person's mental model through representative case studies and examples. It may be possible to do this teaching in a more optimal, algorithmically-driven way, as done by [184], who give a parameterization of the human's mental model of the AI and propose an algorithm to near-optimally select a set of teaching examples for the human. At a more institutional level, programs such as the Collaborative Institutional Training Initiative (CITI) program [35], which is used for training students and clinicians in human subjects protection and the responsible conduct of research, may also be adapted to training clinicians about proper interpretation and use of ML-based systems. A specific example that could be integrated into the training is one that some participants struggled with in our study, which is how to think about the ML model estimates when their patient is not supported in the model training data. We could further build a clinician's intuition of the generalization of the model by allowing them to change some of the patient's attributes and seeing how that would affect the model predictions. Building out such a training program and assessing its impact on clinician decision-making is a ripe avenue for future work.

Limitations There are several limitations to our study. We had a low response rate, though not dissimilar to other reported web-based survey response rates among physicians [22]. This may be because of the time burden of the survey; we stated in our recruitment email that the study would take 30-60 minutes to complete. Participation was voluntary, which may have introduced self-selection bias into our data. All of our participants were younger than 40 years old and 91% were still in training; in residency or fellowship. Younger physicians may be more likely to view ML favorably [190]. On the other hand, difficulty in interpretation of the ML model estimates that trainees had would also likely occur in attending physicians. Finally, in order to reduce participant burden and maximize participation, we designed the instrument as predominantly multiple-choice. This limited the depth of understanding the physicians' treatment decisions. We tried to partially alleviate this shortcoming by providing options for free-text responses at the end of the survey and conducting a qualitative interview for the patient scenario whose results did not fit our hypothesis, but interview participation was limited. Further qualitative studies that include a more heterogenous group of physicians as well as other clinicians can help us better understand how RCT and ML data are used.

8.5 Conclusion

Our study reveals important patterns of physician-AI interaction when confronted with standard-of-care evidence, i.e. RCT data, and personalized ML evidence. We found that physicians relied more frequently on the ML model when RCT evidence was not applicable and when the ML model revealed a difference in survival benefit compared to the RCT for the patient under consideration. Notably, physicians sometimes trusted ML findings even when the considered patient was not well represented in the model's training data. Follow-up qualitative interviews suggested that this inclination may stem from a lack of an appropriate mental model regarding the limitations of ML models, in contrast to familiarity with RCTs. However, providing additional "sanity checks" that validate the results from an ML-based system against accepted RCT results was helpful in helping clinicians detect problems with the ML model. Our findings suggest that future ML-based CDSS systems do have the potential to change treatment decisions in cancer management, but that meticulous development and validation of these systems before deployment is crucial.

Chapter 9

Conclusion

The promise of precision medicine, particularly in oncology, stems from the increasing availability of observational datasets as well as the multitude of available and potentially effective treatments. Each of these treatments may well result in heterogeneous treatment responses that may not get captured when only looking at the average treatment response. As artificial intelligence becomes more integrated into society more generally, the practice of oncology in the future may be assisted by clinical decision support systems that utilize powerful AI and ML models.

This dissertation has explored the various methodological and statistical components that are necessary for a CDSS in oncology. Chapters 3 & 4 built out the core models that can be used to take in a patient's structured data and output predictions of their survival, proclivity for certain adverse events, as well as their overall disease trajectory. Of course, the reliability of these predictions should be reflected to clinicians in multiple ways, paving the way for Chapter 5, where we explored uncertainty quantification of ML model predictions using conformal inference. Importantly, such predictions can be made for more than one treatment, allowing clinicians to compare and contrast two treatment regimens. This naturally lends itself to a causal framing, which in turn motivated the *falsification* framework introduced and formalized in Chapters 6 & 7. Finally, in Chapter 8, we studied physician interaction with a synthetic version of the envisioned CDSS to begin understanding the effect of such an instrument on physician decision-making.

Although we have made progress in the technical problems underlying the development of ML-based decision support systems for precision oncology, there are several other aspects of this venture that we have not addressed in this thesis. One important element is the use of unstructured data, e.g. clinical notes, imaging reads, pathology reports, etc., that are important sources of information for understanding how to manage a patient's cancer. One path forward is to begin to look at large multimodal models such as foundation models that are trained on large datasets of numerous modalities, e.g. text, structured data, speech, etc, and to fine-tune these models on disease-specific tasks [32]. However, the ideas and framing towards clinically useful training objectives in the beginning chapters of this thesis will be relevant for the fine-tuning of foundation models for downstream disease-specific tasks. Another element to consider is the use of -omic data, such as single-cell RNA sequencing data, proteomics, metabolomics, and cell-free DNA sequencing data, to improve the quality of the signal that can be captured by our models [301, 264, 20]. For example, there is a rich literature of building rich representations of single-cell RNA-seq data using deep learning [43]. These representations can also serve as additional inputs to the models introduced in this thesis. Finally, although we began to explore the interaction and behavior of physicians when engaging with a synthetic AI-based system, there are further societal and legal elements to explore in the deployment of such a system in the clinic.

While there is much excitement about the prospect of moving towards personalized care, there are also concerns about whether the promise of precision medicine is premature. A recent article by Wilkinson et al. in The Lancet Digital Health discusses the limitations of using machine learning for improving accurate diagnosis and tailoring treatments for individual patients [287]. They mention several issues, including lack of clinically useful framing and over-reliance on simple predictive metrics, lack of generalizability to other environments, and the inability of standard machine learning models to do causal inference. Another critique that they and others make is that current attempts at precision medicine too often focus on very high-dimensional genomic data, at the expense of including simple clinical variables that may account for a large amount of the variance [275].

We have attempted to address each of these concerns in this thesis, including building models that are grounded in a clinical problem (e.g. multiple myeloma) and evaluated with respect to clinically relevant management of the disease, assessing the generalizability of our models to patients at a different disease stage, and coming up with "sanity checks" rooted in a causal framing of the problem, where well-specified ML models are important. We have also focused on common clinical variables in lieu of genomic data, leaving incorporation of the latter for future work. Still, this thesis does not attempt to provide complete solutions, and there are several avenues of future work hinted at above and discussed in more detail below. We hope that the ideas in this thesis provide an initial roadmap for building a CDSS in clinical oncology.

Incorporating Genomic Data The heterogeneity in treatment response that makes personalized medicine reasonable is conjectured to be due in large part to the differences in genetic make-up between individuals. The explosion of next generation sequencing technologies enabling collection of rich genomic and proteomic data make it possible to potentially capture the genetic signal leading to differences in treatment response and allow for improved prediction of clinical outcomes. In the past several years, prior work has largely involved learning representations of e.g. single-cell RNA seq data using deep learning and applying clustering methods to discover novel subtypes or associations [264, 301]. However, there has been recent work using transcriptomic data to learn better patient representations and improve prediction of clinical outcomes [291]. These data could also be longitudinal, i.e. captured over time, and may require development of new methods to capture signal found therein. For example, a recent platform aimed to decompose sources of variations within longitudinal bulk and single-cell mutli-omics data as well as identify up- or down-regulated markers across timepoints for individual patients [270]. Other work has applied similar methods on longitudinal -omics data to specific diseases, such as Irritable Bowel Syndrome (IBS), in order to better characterize underlying physiology [178]. Moreover, an important avenue for future work is to continue building methods that best capture the signal in these high-dimensional data, especially if they change over time, and incorporate them into the models presented in this thesis. Doing so would ideally provide better predictions of clinical outcomes and give a more complete picture of the progression of a patient's disease trajectory.

Expanding studies of Physician-AI Interaction in Precision Oncology Our study of physician-AI interaction was limited to a simulated setting, where both the patient scenarios and the machine learning model outputs were synthetic. One avenue for future work would be to make the patient scenarios real and use an ML model trained on real-world data and then conduct the study. In the long term, a prospective clinical trial looking at the effect of an ML-based system on physician decision making as well as studying the nature of physician-AI interaction in these scenarios would be necessary. Existing prospective trials have been predominantly in radiology and do not give a deep view into physician behavior and decision-making [288, 77]. This line of work is crucial for measured and thoughtful translation of ML-based systems into clinical settings.

Leveraging Large Pre-trained Foundation Models An important part of building a decision support system for oncologists is generalizing it to other cancers beyond the ones studied in this thesis. One approach for doing this is through the general paradigm of large-scale pretraining on a set of cancer datasets (e.g. SEER, datasets from the YODA initiative [7, 228]) and then finetuning on downstream tasks, as done for current large language models (e.g. ChatGPT, GPT-4, etc.). Indeed, we take this approach in Chapter 4, where we first pretrain the transformer on a forecasting objective and then finetune on downstream event prediction tasks. One can imagine increasing the scale of this pretraining process many times and also including other data modalities in the training, such as imaging, genomics, and even video data such as echocardiograms.

However, such an endeavor is non-trivial in healthcare and potentially leads to many harms. From a technical perspective, pretraining on the number of data modalities required for generalized precision oncology is difficult, and new “feature-level and semantic-level” fusion strategies in the training of foundation models are necessary [32]. There are also a number of legal and ethical regulations that must be met in order for these technologies to be deployed: patient safety, privacy, and fairness are paramount [263]. Finally, the data sources themselves may be biased due to misrepresentation of certain groups or societal stereotypes being baked into the data [263].

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

Supplementary Material for Chapter 3

The supplementary material for Chapter 3 contains the following sections. For each section, we highlight the key findings about the experiments we conduct.

1. **Learning Algorithms**: This section expands upon the learning algorithm for $\mathbf{SSM}_{\text{PK-PD}}$ in the main paper. We also describe two additional sequential models – a First Order Markov Model (FOMM) and a Gated Recurrent Neural Network (GRU).

2. **Synthetic Dataset**: This section provides an in-depth description of the generative process that underlies the synthetic dataset used in the experimental section.

3. **The Multiple Myeloma Research Foundation CoMMpass Study**: This section provides details on data extraction, pre-processing and construction of the ML-MMRF dataset.

4. **FOMM and GRU Experiments**: We study how incorporating a variant of $\mathbb{PKPD}_{\text{Neural}}$ into a FOMM and GRU improves model generalization. The key take-away from this section, with supporting evidence in Table A.1, is that $\mathbf{SSM}_{\text{PK-PD}}$ improves generalization not just in state space models, but also in other popular choices of neural network based models of sequences such as FOMMs and GRUs.

5. **Semi-synthetic Experiments**: We introduce a semi-synthetic dataset that we use to further evaluate $\textbf{SSM}_{\text{PK-PD}}$. The key take-away from this section, with supporting evidence in Table A.2, is that $\textbf{SSM}_{\text{PK-PD}}$ improves generalization on a new dataset whose sequential patterns mimic real-world multiple myeloma data. These improvements are confirmed in a model misspecification scenario.

6. **Additional Experiments**: This section details additional experiments to interpret the model we develop and understand the relative utility of its various parts.

    i. **Patient Forecasting** - We explore different ways in which $\textbf{SSM}_{\text{PK-PD}}$ may be used to forecast patient trajectories given some initial data. When conditioning on different lengths of patient history and then sampling forward in time, we see a qualitative improvement in samples from $\textbf{SSM}_{\text{PK-PD}}$ compared to one of the best performing baselines.

    ii. **Visualizing Disease Progression** - We extend our analysis of the $\textbf{SSM}_{\text{PK-PD}}$'s latent states to studying how they evolve over the entire disease course. We find that clustering patients based on the latent state reveals subgroups that, due to differences in disease severity, have been assigned different treatment regimens. This result suggests that the latent representation has encoded the patient's underlying disease state.

    iii. **Per-feature Breakdown** - We perform a per-feature analysis of how well $\textbf{SSM}_{\text{PK-PD}}$ and $\textbf{SSM}_{\text{Linear}}$ model different clinical biomarkers, finding that $\textbf{SSM}_{\text{PK-PD}}$ does particularly well for important markers of progression, such as serum IgA.

    iv. **Ablation Analysis** - We study which treatment mechanism function yields the most benefit for modeling the ML-MMRF dataset. Our analysis finds that the Neural Treatment Exponential function provides the most differential gains in NELBO and that the time-varying treatments are crucial for accurately modeling the dynamics of serums IgA, IgG, and Lambda.

A.1 Learning Algorithms

We implement all the models that we experiment with in PyTorch [200].

State Space Models Recall that the generative process is:

$$\begin{aligned}
&p(\mathbf{X}|\mathbf{U}, B) = \int_Z \prod_{t=1}^{T} p(Z_t|Z_{t-1}, U_{t-1}, B; \theta) p(X_t|Z_t; \theta) dZ \\
&Z_t|\cdot \sim \mathcal{N}(\mu_\theta(Z_{t-1}, U_{t-1}, B), \Sigma_\theta^t(Z_{t-1}, U_{t-1}, B)), \\
&X_t|\cdot \sim \mathcal{N}(\kappa_\theta(Z_t), \Sigma_\theta^e(Z_t))
\end{aligned}$$

where the transition function, μ_θ, differs as described in the main paper for $\mathbf{SSM}_{\text{Linear}}$, $\mathbf{SSM}_{\text{NL}}$, $\mathbf{SSM}_{\text{PK-PD}}$, &$\mathbf{SSM}_{\text{MOE}}$.

Maximum Likelihood Estimation of θ: Since the log likelihood $p(\mathbf{X}|\mathbf{U}, B)$ is difficult to evaluate and maximize directly due to the high-dimensional integral, we resort to a variational learning algorithm that instead maximizes a lower bound on the log-likelihood to learn the model parameters, θ. We make use of a structured inference network [148] that amortizes the variational approximation, $q_\phi(\mathbf{Z}|\mathbf{X})$, to the posterior distribution, $p_\theta(\mathbf{Z}|\mathbf{X})$, of each datapoint.

$$\begin{aligned}
&\log p(\mathbf{X}|\mathbf{U}, B; \theta) \geq \mathcal{L}(\mathbf{X}; (\theta, \phi)) \\
&= \sum_{t=1}^{T} \mathbb{E}_{q_\phi(Z_t|\mathbf{X},\mathbf{U},B)}[\log p_\theta(X_t|Z_t)] \\
&- \text{KL}(q_\phi(Z_1|\mathbf{X}, \mathbf{U}, B)||p_\theta(Z_1|B)) \\
&- \sum_{t=2}^{T} \mathbb{E}_{q_\phi(Z_{t-1}|\mathbf{X},\mathbf{U},B)}[\\
&\text{KL}(q_\phi(Z_t|Z_{t-1}, \mathbf{X}, \mathbf{U})||p_\theta(Z_t|Z_{t-1}, U_{t-1}, B))]
\end{aligned} \tag{A.1}$$

The lower bound on the log-likelihood of data, $\mathcal{L}(\mathbf{X}; (\theta, \phi))$, is a differentiable function of the parameters θ, ϕ [148], so we jointly learn them via gradient ascent.

As mentioned in the main paper, if X_t is missing, then it is marginalized out when evaluating the likelihood of the data under the model. Since the inference network also conditions on sequences of observed data to predict the variational parameters, we use forward fill imputation where data are missing.

Hyperparameters: We present the results of the hyperparameter search on the datasets that we study. Please see the evaluation section of the main paper for the specific ranges that we searched over.

- **SSM**$_{\text{Linear}}$

  1. *Synthetic:* State space dimension 48, L2 regularization on all parameters with strength 0.01
  2. *ML-MMRF:* State space dimension 16, L2 regularization on all parameters with strength 0.01

- **SSM**$_{\text{NL}}$

  1. *Synthetic:* State space dimension 48, hidden layer dimension 300, L2 regularization on all parameters with strength 0.1
  2. *ML-MMRF:* State space dimension 48, hidden layer dimension 300, L2 regularization on all parameters with strength 0.1

- **SSM**$_{\text{PK-PD}}$

  1. *Synthetic:* State space dimension 48, L1 regularization on subset of parameters with strength 0.01
  2. *ML-MMRF:* State space dimension 48, L1 regularization on all parameters with strength 0.01

- **SSM**$_{\text{MOE}}$

  1. *Synthetic:* State space dimension 16, hidden layer dimension 300, L1 regularization on all parameters with strength 0.01

2. *ML-MMRF:* State space dimension 48, hidden layer dimension 300, L1 regularization on all parameters with strength 0.01

- **SSM Attn. Hist.**

  1. *Synthetic:* State space dimension 16, hidden layer dimension 100, L1 regularization on all parameters with strength 0.01

  2. *ML-MMRF:* State space dimension 48, hidden layer dimension 300, L1 regularization on all parameters with strength 0.01

SSM Attn. Hist. Baseline: We provide details on the SSM architecture proposed by [3] for disease progression modeling. The generative process of their architecture differs from a normal state space model in that the transition function, μ_θ, assumes that the patient's latent state at time t depends on their entire history of latent states and interventions. Thus, we have,

$$\begin{aligned} & p(\mathbf{X}|\mathbf{U}, B) = \\ & \int_Z \prod_{t=1}^{T} p(Z_t|Z_{1:t-1}, U_{1:t-1}, B; \theta) p(X_t|Z_t; \theta) dZ \\ & Z_t|\cdot \sim \mathcal{N}(\mu_\theta(Z_{1:t-1}, U_{1:t-1}, B), \Sigma_\theta^t(Z_{1:t-1}, U_{1:t-1}, B)), \\ & X_t|\cdot \sim \mathcal{N}(\kappa_\theta(Z_t), \Sigma_\theta^e(Z_t)) \end{aligned} \tag{A.2}$$

Note that we adapt the authors' model to work with a continuous latent state, whereas they utilize a discrete latent state. The crux of their method is to parameterize the transition distribution as an attention-weighted sum of the previous latent states to compute the current latent state. These attention weights are a function of a patient's entire clinical lab and treatment history. Therefore, the transition function that we use to capture their modeling assumptions is as follows:

$$\mu_\theta(Z_{1:t-1}, \boldsymbol{\alpha}_{1:t-1}) = W_h(\sum_{i=1}^{t-1} \boldsymbol{\alpha}_i \odot Z_i) + b_h, \tag{A.3}$$

where $\boldsymbol{\alpha}_{1:t-1} = A_t([X_{1:t-1}, U_{1:t-1}])$ via an attention mechanism, A_t. We use a bi-directional recurrent neural network for the inference network, as opposed to the authors' proposed attentive inference network. We argue that the bi-RNN is just as expressive, since the variational parameters are a function of all past and future observations. Moreover, our goal is to study the effect of altering the generative model in this work.

We also experiment with First Order Markov Models (**FOMM**) and Gated Recurrent Units (**GRU**) [56], which we detail below.

First Order Markov Models FOMMs assume observations are conditionally independent of the past given the previous observation, intervention and baseline covariates. The generative process is:

$$p(\mathbf{X}|\mathbf{U}, B) = \prod_{t=1}^{T} p(X_t|X_{t-1}, U_{t-1}, B);$$
$$X_t|\cdot \sim \mathcal{N}(\mu_\theta(X_{t-1}, U_{t-1}, B), \Sigma_\theta(X_{t-1}, U_{t-1}, B)),$$

where the transition function, μ_θ, differs akin to the transition function of **SSM** models, as described in the main paper. Here, we will experiment with $\mathbf{FOMM}_{\text{Linear}}$, $\mathbf{FOMM}_{\text{NL}}$, $\mathbf{FOMM}_{\text{MOE}}$, &$\mathbf{FOMM}_{\text{PK-PD}}$.

$\mathbb{PKPD}_{\textit{Neural}}$ *for* $\boldsymbol{FOMM}_{\textit{PK-PD}}$: We will use a simpler variant of the $\mathbb{PKPD}_{\text{Neural}}$ formulation introduced in the main paper as a proof of concept. Namely, we have,

$$\mu_\theta(X_{t-1}, U_{t-1}, B) = \sum_{i=1}^{d} \sigma(\boldsymbol{\delta})_i \odot g_i(S_{t-1}, U_{t-1}, B), \tag{A.4}$$

where each $\boldsymbol{\delta}$ is a learned vector of weights and σ refers to a softmax on the weights. Note that the $\mathbb{PKPD}_{\text{Neural}}$ introduced in the main paper is a generalization of Equation A.4; the primary difference is that the attention mechanism allows the weights to be a function of the prior state, which enables the weights to change over time.

Maximum Likelihood Estimation of θ: We learn the model by maximizing $\max_\theta \log p(\mathbf{X}|\mathbf{U}, B)$. Using the factorization structure in the joint distribution of the generative model, we obtain: $\log p(\mathbf{X}|\mathbf{U}, B) = \sum_{t=1}^{T} \log p(X_t|X_{t-1}, U_{t-1}, B)$. Each $\log p(X_t|X_{t-1}, U_{t-1}, B)$ is estimable as the log-likelihood of the observed multi-variate X_t under a Gaussian distribution whose (diagonal) variance is a function $\Sigma_\theta(X_{t-1}, U_{t-1}, B)$ and whose mean is given by the transition function, $\mu_\theta(X_{t-1}, U_{t-1}, B)$. Since each $\log p(X_t|X_{t-1}, U_{t-1}, B)$ is a differentiable function of θ, its sum is differentiable as well, and we may use automatic differentiation to derive gradients of the log-likelihood with respect to θ in order to perform gradient ascent. When any dimension of X_t is missing, that dimension's log-likelihood is ignored (corresponding to marginalization over that random variable) during learning.

Hyperparameters: We present the results of the hyperparameter search on the datasets that we study.

- **FOMM**$_{\text{Linear}}$
  1. *Synthetic:* L1 regularization on all parameters with strength 0.1
  2. *ML-MMRF:* L1 regularization on all parameters with strength 0.1
- **FOMM**$_{\text{NL}}$
  1. *Synthetic:* Hidden layer dimension 200, L1 regularization on all parameters with strength 0.1
  2. *ML-MMRF:* Hidden layer dimension 300, L1 regularization on all parameters with strength 0.1
- **FOMM**$_{\text{PK-PD}}$
  1. *Synthetic:* L1 regularization on subset of parameters with strength 0.1
  2. *ML-MMRF:* L1 regularization on subset of parameters with strength 0.1

Gated Recurrent Neural Network (GRUs): GRUs [54] are auto-regressive models of sequential observations i.e. $p(\mathbf{X}|\mathbf{U}, B) = \prod_{t=1}^{T} p(X_t|X_{<t}, U_{<t}, B))$. GRUs use an intermediate hidden state $h_t \in \mathbb{R}^H$ at each time-step as a proxy for what the

model has inferred about the sequence of data until t. The GRU dynamics govern how h_t evolves via an update gate F_t, and a reset gate R_t:

$$\begin{aligned}
F_t &= \sigma(W_z \cdot [X_t, U_t, B] + V_z h_{t-1} + b_z), \\
R_t &= \sigma(W_r \cdot [X_t, U_t, B] + V_r h_{t-1} + b_r) \\
h_t &= F_t \odot h_{t-1} + (1 - F_t) \odot \\
&\tanh(W_h \cdot [X_t, U_t, B] + V_h(R_t \odot h_{t-1}) + b_h)
\end{aligned} \tag{A.5}$$

$\theta = \{ \ W_z, W_r, W_h \in \mathbb{R}^{H \times (M+L+J)}; V_z, V_r, V_h \in \mathbb{R}^{H \times H}; b_z, b_r, b_h \in \mathbb{R}^H\}$ are learned parameters and σ is the sigmoid function. The effect of interventions may be felt in any of the above time-varying representations and so the "transition function" in the GRU is distributed across the computation of the forget gate, reset gate and the hidden state, i.e. $S_t = [F_t, R_t, h_t]$. We refer to this model as **GRU**.

$\mathbb{PKPD}_{\mathit{Neural}}$ *for **GRU**$_{\mathit{PK\text{-}PD}}$*: We take the output of Equation A.4, $o_t = \mu_\theta(X_{t-1}, U_{t-1}, B)$, and divide it into three equally sized vectors: o_t^f, o_t^r, o_t^h. Then,

$$\begin{aligned}
F_t &= \sigma(o_t^f + V_z h_{t-1} + b_z) \\
R_t &= \sigma(o_t^r + V_r h_{t-1} + b_r) \\
h_t &= F_t \odot h_{t-1} \\
&\quad + (1 - F_t) \odot \tanh(o_t^h + V_h(R_t \odot h_{t-1}) + b_h)
\end{aligned}$$

Maximum Likelihood Estimation of θ: We learn the model by maximizing $\max_\theta \log p(\mathbf{X}|\mathbf{U}, B)$. Using the factorization structure in the joint distribution of the generative model, we obtain: $\log p(\mathbf{X}|\mathbf{U}, B) = \sum_{t=1}^{T} \log p(X_t|X_{<t}, U_{<t}, B)$. At each point in time the hidden state of the GRU, h_t, summarizes $X_{<t}, U_{<t}, B$. Thus, the model assumes $X_t \sim \mathcal{N}(\mu_\theta(h_t), \Sigma_\theta(h_t))$.

At each point in time, $\log p(X_t|X_{<t}, U_{<t}, B)$ is the log-likelihood of a multi-variate Gaussian distribution which depends on θ. As before, we use automatic differentiation to derive gradients of the log-likelihood with respect to θ in order to perform gradient ascent. When any dimension of X_t is missing, that dimension's log-likelihood is ignored

(corresponding to marginalization over that random variable) during learning.

Hyperparameters: We present the results of the hyperameter search on the datasets that we study.

- **GRU**

  1. *Synthetic:* Hidden layer dimension 500, L2 regularization on all parameters with strength 0.1
  2. *ML-MMRF:* Hidden layer dimension 250, L2 regularization on all parameters with strength 0.1

- **GRU**$_{\text{PK-PD}}$

  1. *Synthetic:* Hidden layer dimension 500, L2 regularization on subset of parameters with strength 0.01
  2. *ML-MMRF:* Hidden layer dimension 500, L2 regularization on subset of parameters with strength 0.01

A.2 Synthetic Dataset

Below, we outline the general principles that the synthetic data we design is based on:

- We sample six random baseline values from a standard normal distribution.

- Two of the six baseline values determine the natural (untreated) progression of the two-dimensional longitudinal trajectories. They do so as follows: depending on which quadrant the baseline data lie in, we assume that the patient has one of four subtypes.

- Each of the four subtypes typifies different patterns by which the biomarkers behave such as whether they both go up, both go down, one goes up, one goes down etc. To see a visual example of this, we refer the reader to Figure A-1 (left).

Baseline The generative process for the baseline covariates is $B \sim \mathcal{N}(0; \mathbb{I}); B \in \mathbb{R}^6$.

Treatments (Interventions): There is a single drug (denoted by a binary random variable) that may be withheld (in the first line of therapy) or prescribed in the second line of therapy. For each patient $d_i \sim \text{Unif.}[0, 18]$ denotes when the single drug is administered (and the second line of treatment begins). d_i is the point at which the local clock resets. We can summarize the generative process for the treatments as follows:

$$
\begin{aligned}
& d \sim \text{Unif.}[0, 18] \\
& U_t = 0 \text{ if } t < d \text{ and } 1 \text{ otherwise} \\
& \text{line}_t[0] = 1 \text{ if } t < d \text{ and } 0 \text{ otherwise} \\
& \text{line}_t[1] = 0 \text{ if } t < d \text{ and } 1 \text{ otherwise}
\end{aligned} \tag{A.6}
$$

where $\text{line}_t[0], \text{line}_t[1]$ denote the one-hot encoding for line of therapy. Next, we use the Neural Treatment Exponential (Equation 3.6 in the main paper) to generate a treatment response in the data. The functional form of the Treatment Exponential function (TE) is re-stated below for convenience,

$$
\text{TE}(\text{lc}_t) = \begin{cases} b_0 + \alpha_1/[1 + \exp(-\alpha_2(\text{lc}_t - \frac{\gamma_l}{2}))], \\ \qquad \text{if } 0 \leq \text{lc}_t < \gamma_l \\ b_l + \alpha_0/[1 + \exp(\alpha_3(\text{lc}_t - \frac{3\gamma_l}{2}))], \\ \qquad \text{if } \text{lc}_t \geq \gamma_l \end{cases} \tag{A.7}
$$

The parameters that we use to generate the data are: $\alpha_2 = 0.6, \alpha_3 = 0.6, \gamma_l = 2, b_l = 3$, and $\alpha_1 = [10, 5, -5, -10]$, which we vary based on patient subtype. We set $\alpha_0 = (\alpha_1 + 2b_0 - b_l)/(1 + \exp(-\alpha_3\gamma_l)/2)$ to ensure that the treatment effect peaks at $t = \text{lc}_t + \gamma_l$ and $b_0 = -\alpha_1/(1 + \exp(\alpha_2 \cdot \gamma_l/2))$ for attaining $\text{TE}(0) = 0$.

Biomarkers: We are now ready to describe the full generative process of the longitudinal biomarkers.

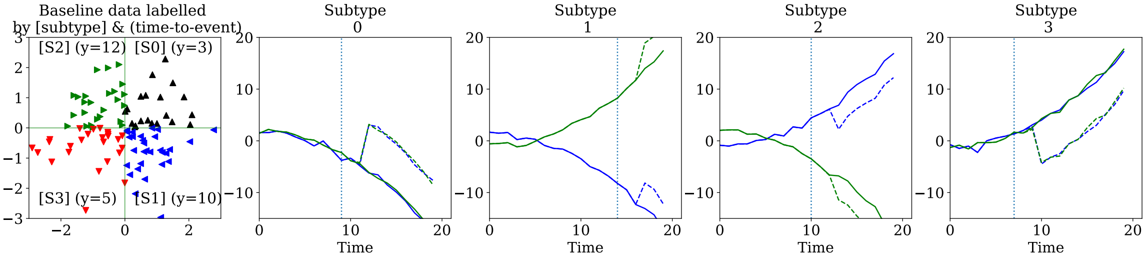


Figure A-1: Visualization of synthetic data: Left: A visualization of "patient"'s baseline data (colored and marked by patient subtype). Right four plots: Examples of patient's longitudinal trajectories along with treatment response. The blue and green longitudinal data denote two different patient biomarkers. The solid blue and green lines are what the trajectories would be with no intervention whereas the dotted blue and green lines are what the trajectories are with an intervention. Gray-dotted line represents an intervention. The subtypes may, optionally, be correlated with patient outcomes as highlighted using the values of y.

$$
\begin{aligned}
&B_{1\ldots,6} \sim \mathcal{N}(0; I), \\
&f_d(t) = 2 - 0.05t - 0.005t^2, && \text{(A.8)} \\
&f_u(t) = -1 + 0.0001t + 0.005t^2, \\
&X_1(t); X_2(t) = && \text{(A.9)} \\
&\begin{cases}
f_d(t) + \mathrm{TE}(\mathrm{lc}_t) + \mathcal{N}(0, 0.25);\ \ f_d(t) + \mathrm{TE}(\mathrm{lc}_t) \\
\qquad + \mathcal{N}(0, 0.25), B_1 \geq 0, B_2 \geq 0 \text{ if subtype 1} \\
f_d(t) + \mathrm{TE}(\mathrm{lc}_t) + \mathcal{N}(0, 0.25);\ \ f_u(t) + \mathrm{TE}(\mathrm{lc}_t) \\
\qquad + \mathcal{N}(0, 0.25), B_1 \geq 0, B_2 < 0 \text{ if subtype 2} \\
f_u(t) + \mathrm{TE}(\mathrm{lc}_t) + \mathcal{N}(0, 0.25);\ \ f_d(t) + \mathrm{TE}(\mathrm{lc}_t) \\
\qquad + \mathcal{N}(0, 0.25), B_1 < 0, B_2 \geq 0 \text{ if subtype 3} \\
f_u(t) + \mathrm{TE}(\mathrm{lc}_t) + \mathcal{N}(0, 0.25);\ \ f_u(t) + \mathrm{TE}(\mathrm{lc}_t) \\
\qquad + \mathcal{N}(0, 0.25), B_1 < 0, B_2 < 0 \text{ if subtype 4},
\end{cases}
\end{aligned}
$$

Intuitively, the above generative process captures the idea that without any effect of treatment, the biomarkers follow the patterns implied by the subtype (encoded in the first two dimensions of the baseline data). However the effect of interventions is felt more prominently after d, the random variable denoting time at which treatment was prescribed.

A.3 The Multiple Myeloma Research Foundation CoMMpass Study

Here, we elaborate upon the data made available by the Multiple Myeloma Research Foundation in the IA13 release of data. We will make code available to go from the files released by the MMRF study to numpy tensors that may be used in any machine learning framework.

Inclusion Criteria: To enroll in the CoMMpass study, patients must be newly diagnosed with symptomatic multiple myeloma, which coincides with the start of treatment. Patients must be eligible for treatment with an immunomodulator or a proteasome inhibitor, two of the most common first line drugs, and they must begin treatment within 30 days of the baseline bone marrow evaluation [188].

A.3.1 Features

Genomic Data: RNA-sequencing of CD38+ bone marrow cells was available for 769 patients. Samples were collected at initiation into the study, pre-treatment. For these patients, we used the Seurat package version 2.3.4 [40] in R to identify variable genes, and we then limit downstream analyses to these genes. We use principal component analysis (PCA) to further reduce the dimensionality of the data. The projection of each patient's gene expression on to the first 40 principal components serves as the genetic features used in our model.

Baseline Data: Baseline data includes PCA scores, lab values at the patient's first visit, gender, age, and the revised ISS stage. The baseline data also includes binary variables detailing the patient's myeloma subtype, including whether or not they have heavy chain myeloma, are IgG type, IgA type, IgM type, kappa type, or lambda type. Additionally, several labs are measured at baseline, as well as longitudinally at subsequent visits. We detail these labs in the next sub-section. The genetic and baseline data jointly comprise B.

Longitudinal Data: Longitudinal data is measured approximately every 2 months and includes lab values and treatment information. The lab values are real-valued numbers whose values evolve over time. They include: absolute neutrophil count (x10^9/l), albumin (g/l), blood urea nitrogen (mmol/l), calcium (mmol/l), serum creatinine (umol/l), glucose (mmol/l), hemoglobin (mmol/l), serum kappa (mg/dl), serum m protein (g/dl), platelet count x10^9/l, total protein (g/dl), white blood count x10^9/l, serum IgA (g/l), serum IgG (g/l), serum IgM (g/l), serum lambda (mg/dl).

Treatment information includes the line of therapy (we group all lines beyond line 3 as line 3+) the patient is on at a given point in time, and the local clock denoting the time elapsed since the last line of therapy. We also include the following treatments as (binary, indicating prescription) features in our model: lenalidomide, dexamethasone, cyclophosphamide, carfilzomib, bortezomib. The aforementioned are the top five drugs by frequency in the MMRF dataset. This dataset has significant missingness, with $\sim 66\%$ of the longitudinal markers missing. In addition, there is right censorship in the dataset, with around 25% of patients getting censored over time.

A.3.2 Data Processing

Longitudinal biomarkers $\mathbf{X}$: Labs are first clipped to five times the median value to correct for outliers or data errors in the registry. They are then normalized to their healthy ranges (obtained via a literature search) as (unnormalized labs - healthy maximum value), and then multiplied by a lab-dependent scaling factor to ensure that most values lie within the range $[-8, 8]$. Missing values are represented as zeros, but a separate mask tensor, where 1 denotes observed and 0 denotes missing, is used to marginalize out missing variables during learning.

Baseline B: The biomarkers in the baseline are clipped to five times their median values. Patients without gene expression data (in the PCA features) are assigned the average normalized PCA score of their five nearest neighbors, using the Minkowski

Dataset	**Held-out Neg Log Likelihood**	**FOMM Linear**	**FOMM Nonlinear**	**FOMM PK-PD**	**FOMM MOE**	**GRU**	**GRU PK-PD**	**SSM PK-PD** (NELBO)
ML-MMRF		92.80	97.53	**90.26**	97.26	**89.89**	99.98	**61.54**

Dataset	**FOMM PK-PD** vs. **FOMM Linear**	**FOMM PK-PD** vs. **FOMM NL**	**FOMM PK-PD** vs. **FOMM MOE**	**GRU PK-PD** vs. **GRU**	—
ML-MMRF	0.792 (0.405)	0.668 (0.457)	0.510 (0.490)	0.406 (0.489)	—
	FOMM Linear	**FOMM NL**	**FOMM PK-PD**	**GRU**	**GRU PK-PD**
Synthetic (50 samples)	71.06 +/- .03	58.80 +/- .03	**56.81 +/- .04**	56.65 +/- .11	**53.49 +/- .04**
Synthetic (1000 samples)	62.93 +/- .03	**57.16 +/- .03**	57.81 +/- .02	31.09 +/- .02	**29.27 +/- .01**

Table A.1: Generalization of FOMM and GRU models on ML-MMRF and Synthetic Data: *Top*: Lower is better. We report negative log-likelihood averaged across five folds of held-out data. *Bottom*: *ML-MMRF*: We report pairwise comparisons of models trained on ML-MMRF. Higher is better. Each number is the fraction (with std. dev.) of held-out patients for which the model that uses $\mathbb{PKPD}_{\text{Neural}}$ has a lower negative log-likelihood than a model in the same family that uses a different transition function. *Synthetic*: Lower is better. We report held-out negative log likelihood with std. dev. on FOMM and GRU to study generalization in the synthetic setting.

distance metric calculated on FISH features, ISS stage, and age.

A.4 FOMM and GRU Experiments

$\mathbb{PKPD}_{Neural}$ *improves generalization in both **FOMM** and **GRU** models on synthetic data:* Table A.1 (bottom) depicts negative log-likelihoods on held-out synthetic data across different models, where a lower number implies better generalization. The nonlinearity of the synthetic data makes unsupervised learning a challenge for $\mathbf{FOMM}_{\text{Linear}}$ at 50 samples, allowing $\mathbf{FOMM}_{\text{PK-PD}}$ to easily outperform it. In contrast, $\mathbf{FOMM}_{\text{NL}}$ can capture non-linearities in the data, making it a strong baseline even at 50 samples. Yet, $\mathbf{FOMM}_{\text{PK-PD}}$ outperforms it, emphasizing the utility of $\mathbb{PKPD}_{\text{Neural}}$ in when data is scarce. At 1000 samples, $\mathbf{FOMM}_{\text{NL}}$ is able to learn enough about the dynamics to improve its performance relative to $\mathbf{FOMM}_{\text{PK-PD}}$. **GRU** is a strong model on this dataset, but at both sample sizes, the $\mathbf{GRU}_{\text{PK-PD}}$ improves generalization.

***FOMM** and **GRU** generalization performance on ML-MMRF :* We observe improvements in generalization across FOMMs with the use of $\mathbb{PKPD}_{\text{Neural}}$. However, we do not see discernible gains from $\mathbf{GRU}_{\text{PK-PD}}$, perhaps due to missingness in the data, which also results in the GRUs generalizing worse than SSM models based on negative log likelihood (Table A.1 (top)). Overall, the GRU, due to the distributed nature of its "transition function", is a harder model to use $\mathbb{PKPD}_{\text{Neural}}$ in, although figuring how to do so effectively is a valuable direction for future work.

A.5 Experiments on Semi-synthetic Dataset

In this section, we cover how to generate the semi-synthetic dataset. We then provide experimental results on generalization performance as well as a result in a model misspecification scenario.

Semi-synthetic data: We train the $\mathbf{SSM}_{\text{PK-PD}}$ model on the ML-MMRF dataset and generate samples from the model. For each sequence of treatments, we generate 30 random samples per training data point resulting in a dataset of size 14000. Then we uniformly at random sample 1000 samples from that pool to form our training set. We perform a similar procedure to generate several held-out sets (size 87000 samples each). This semi-synthetic dataset allows us to ask questions about generalization on data with statistics similar to ML-MMRF.

Generalization and Model Misspecification *$SSM_{PK\text{-}PD}$ generalizes well with fewer samples:* At 1000 samples, we find that the $\mathbf{SSM}_{\text{PK-PD}}$ models generalize better than the baselines, where a lower, more negative number implies better generalization. We see that with few samples, $\mathbf{SSM}_{\text{NL}}$ and $\mathbf{SSM}_{\text{MOE}}$ overfit. However, when sharply increasing the number of samples to 20000, both models recover their performance and even begin to outperform $\mathbf{SSM}_{\text{PK-PD}}$. This result further solidifies the generalization capability of our proposed $\mathbf{SSM}_{\text{PK-PD}}$ model in a data-scarce setting as well as the difficulty of learning a nonlinear model that does not overfit.

***$\boldsymbol{SSM}_{PK\text{-}PD}$* continues to generalize well even when it is mis-specified:** We run a similar experiment to what we ran on ML-MMRF, where we take out the Neural Treatment Exponential mechanism function from μ_θ and instead opt for using a linear function. We see that $\mathbf{SSM}_{\text{PK-PD}}$ **w/o TExp** performs comparably to a $\mathbf{SSM}_{\text{PK-PD}}$ with the Neural Treatment Exponential mechanism, providing further evidence that our architecture can recover the unknown intervention effect in the data via a combination of related mechanism functions.

Table A.2: Held-out NELBO on semi-synthetic data: Trained on 1K or 20K samples and evaluated on 87K samples, we show the mean test NELBO and standard deviation across five test sets.

Sample Size	**SSM Linear**	**SSM NL**	**SSM MOE**
1000	-211.54 +/- 49.61	-192.86 +/- 25.77	-266.37 +/- 10.42
	SSM PK-PD	**SSM PK-PD (w/o TExp)**	——
1000	**-294.00 +/- 10.25**	**-295.44 +/- 8.18**	——
	SSM Linear	**SSM NL**	**SSM MOE**
20000	-322.36 +/- 0.30	-316.09 +/- 0.18	**-322.22 +/- 0.22**
	SSM PK-PD	**SSM PK-PD (w/o TExp)**	——
20000	-319.57 +/- 0.23	-315.60 +/- 0.40	——

Hyperparameters: We present the best hyperparameters for each model at each sample size. We search over the ranges as described in the main paper; however, at 20000 samples, we train for 1000 epochs instead of 15000 epochs, which we found to be a more stable training configuration.

- **SSM$_{\text{Linear}}$**

  1. 1000 samples: State space dimension 64, L2 regularization on all parameters with strength 0.01

  2. 20000 samples: State space dimension 128, L2 regularization on all parameters with strength 0.01

- **SSM$_{\text{NL}}$**

  1. 1000 samples: State space dimension 48, hidden layer dimension on neural network: 300, L2 regularization on all parameters with strength 0.1

  2. 20000 samples: State space dimension 128, hidden layer dimension on neural network: 300, L2 regularization on all parameters with strength 0.01

- **SSM$_{\text{MOE}}$**

  1. 1000 samples: State space dimension 48, hidden layer dimension on each MLP expert: 300, L2 regularization on all parameters with strength 0.01
  2. 20000 samples: State space dimension 128, hidden layer dimension on each MLP expert: 300, L2 regularization on all parameters with strength 0.01

- **SSM$_{\text{PK-PD}}$**

  1. 1000 samples: State space dimension 64, L2 regularization on subset of parameters with strength 0.01
  2. 20000 samples: State space dimension 128, L2 regularization on subset of parameters with strength 0.01

A.6 Additional Analyses

This section presents experimental results that provide an additional qualitative lens onto the $\mathbb{PKPD}_{\text{Neural}}$.

A.6.1 Exploring different strategies of sampling patient data using SSM$_{\textbf{PK-PD}}$ on ML-MMRF

In the main paper (Figure 3-5), we show samples from SSM models trained on ML-MMRF, conditioned on a patient's first two years of data and the sequence of interventions they were prescribed. In each case, we additionally condition on the patient's baseline covariates.

Here, we experiment with different conditioning strategies. Let C denote the point in time until which we condition on patient data and F denote the number of timesteps that we sample forward into the future. We limit our analysis to the subset of patients for which $C + F <= T$ where T is the maximum number of time steps for which we observe patient data.

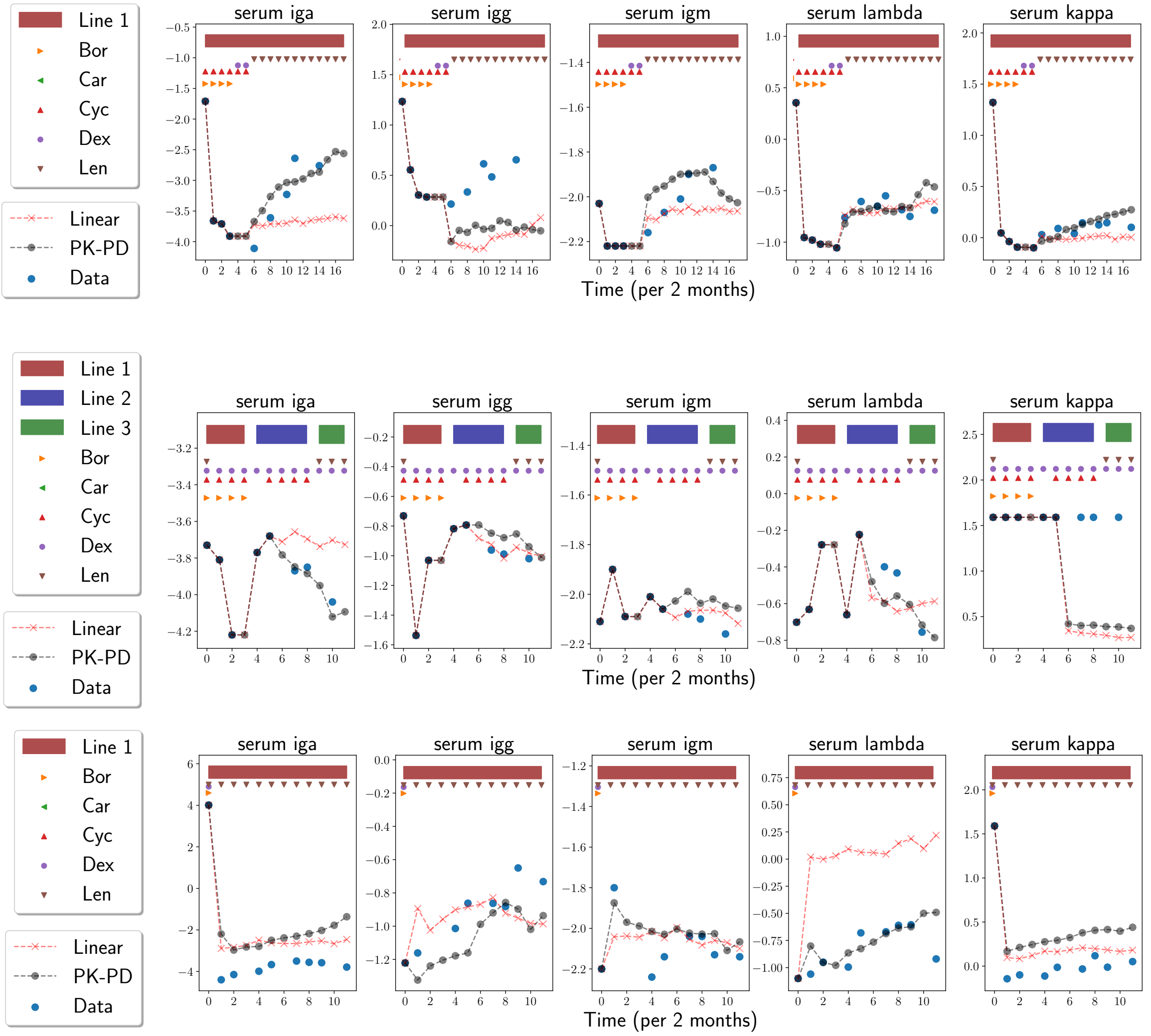


Figure A-2: Forward samples from learned SSM models with differing conditioning strategies: We visualize samples from $\mathbf{SSM}_{\text{PK-PD}}$ (o) and $\mathbf{SSM}_{\text{linear}}$ (x). Each row corresponds to a single patient, whereas each column represents a different biomarker for that patient. **a)**: We condition on 6 months of patient data and forward sample 2 years. **b)**: We condition on 1 year of patient data and forward sample 1 year. **c)**: We condition on a patient's baseline data and forward sample 2 years. As in the main paper, blue circles denote ground truth, and the markers above the trajectories represent treatments prescribed across time.

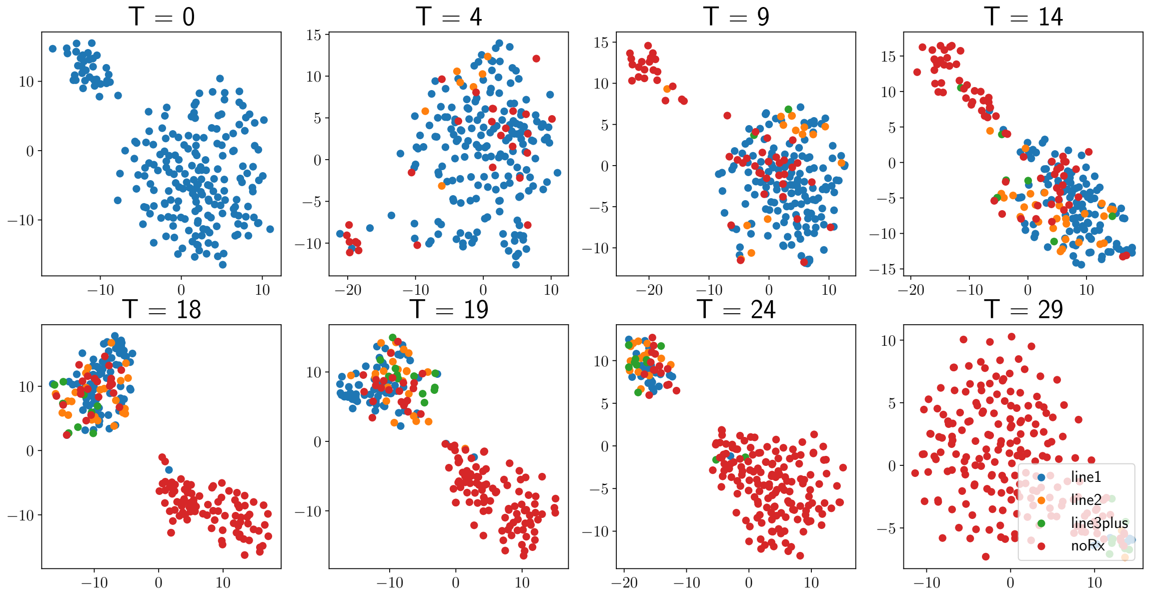


Figure A-3: Z_t Visualizations: We visualize the TSNE representations of each held-out patient's latent state, Z_t, over multiple time points, extending the analysis done in the main paper.

The samples we display are obtained as a consequence of averaging over three different samples, each of which is generated (for the SSM) as follows:

$$\begin{aligned}
&Z \sim q_\phi(Z_C|Z_{C-1}, X_{1:C}, U_{0:C-1}) \\
&Z_k \sim p_\theta(Z_k|Z_{k-1}, U_{k-1}, B) \quad k = C+1, \ldots, C+F \\
&X_k \sim p_\theta(X_k|Z_k) \quad k = C+1, \ldots, C+F
\end{aligned} \tag{A.10}$$

We study the following strategies for simulating patient data from the models.

1. Condition on 6 months of patient data, and then sample forward 2 years,

2. Condition on 1 year of a patient data and then sample forward 1 year,

3. Condition on the baseline data of the patient and then sample forward 2 years.

In Supp. Figure A-2, we show additional samples from **SSM**$_{\text{PK-PD}}$ when conditioning on differing amounts of data. Overall, in all three cases, **SSM**$_{\text{PK-PD}}$ models capture treatment response better than one of the best performing baselines (i.e. **SSM**$_{\text{Linear}}$). For 1. (Figure A-2a)), we see that **SSM**$_{\text{PK-PD}}$ correctly captures that the serum IgA value goes up, while $\mathbf{SSM}_{\text{Linear}}$ predicts that it will stay steady. For 2. (Figure A-2b)), $\mathbf{SSM}_{\text{PK-PD}}$ does well in modeling down-trends, as in serum IgA and serum lambda. For 3. (Figure A-2c)), we similarly see that $\mathbf{SSM}_{\text{PK-PD}}$ captures the up-trending serum IgG and serum lambda.

A.6.2 Analyzing the Latent State learned by $\mathbf{SSM}_{\textbf{PK-PD}}$ over Time

In Supp Figure A-3, we show the latent state of each held-out patient (reduced down to two dimensions via TSNE [176]) over multiple time points, expanding on the two time points that were shown in Figure 3-4 of the main paper. As we saw before, early in the treatment course, the latent representations of the patients have no apparent structure. However, as time goes on, we find that the latent representations separate based on whether treatment is administered or not.

A.6.3 Deep Dive into $\mathbf{SSM}_{\textbf{PK-PD}}$ vs $\mathbf{SSM}_{\textbf{Linear}}$ on ML-MMRF

We are also interested in the absolute negative log likelihood measures and predictive capacity of the models at a per-feature level. In Supp. Figure A-6a), we use importance sampling to estimate the marginal negative log likelihood of $\mathbf{SSM}_{\text{Linear}}$ and $\mathbf{SSM}_{\text{PK-PD}}$ for each covariate across all time points. Namely, we utilize the following estimator,

$$p(\mathbf{X}) \approx \frac{1}{S}\sum_{s=1}^{S}\frac{p(\mathbf{X}|\mathbf{Z}^{(s)})p(\mathbf{Z}^{(s)})}{q(\mathbf{Z}^{(s)}|\mathbf{X})}, \tag{A.11}$$

akin to what is used in [218]. $\mathbf{SSM}_{\text{PK-PD}}$ has lower negative log likelihood compared to $\mathbf{SSM}_{\text{Linear}}$ for several covariates, including neutrophil count, albumin, BUN, calcium, and serum IgA. This result is corroborated with the generated samples in Supp. Figure A-2, which often show that the PK-PD model qualitatively does better at capturing IgA dynamics compared to the Linear model. In general, although there is a some overlap in the estimates of the likelihood under the two models for some features, it is reassuring to see that $\mathbf{SSM}_{\text{PK-PD}}$ does model the probability density of vital markers

Dataset	**Held-out NELBO**	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
ML-MMRF	linear	85.26	73.25	70.61	59.86	71.11
	linear + log-cell	84.42	68.11	66.69	58.83	71.48
	linear + log-cell + te	65.06	57.20	66.73	53.37	58.12

Table A.3: Ablation Experiment on ML-MMRF Dataset: We study the effect of adding each mechanism function to $\mathbf{SSM}_{\text{PK-PD}}$. We report held-out bounds on negative log likelihood.

like serum IgA (which is often used by doctors to measure progression for specific kinds of patients), better than the baseline.

In Supp. Figure A-6b), c), and d), we show the L1 error of $\mathbf{SSM}_{\text{PK-PD}}$ and $\mathbf{SSM}_{\text{Linear}}$ when predicting future values of each covariate. We do so under three different conditioning strategies: 1) condition on 6 months of patient data, and predict 1 year into the future; 2) condition on 6 months of patient data, and predict 2 years into the future; 3) condition on 2 years of patient data, and predict 1 year into the future. Observing 1) and 2) (Supp. Figure A-6b) and c)), we see that prediction quality expectedly degrades when trying to forecast longer into the future. Additionally, we find that when increasing the amount of data we condition on to two years (i.e. forward sampling later on in a patient's disease course) (Supp. Figure A-6d)), the prediction quality is similar to that of conditioning only on six months of data (Supp. Figure A-6b)) [barring serum M-protein and glucose]. This result reflects the ability of our model to generate accurate samples at multiple stages of a patient's disease. Finally, we also report L1 error for forward samples taken from $\mathbf{SSM}_{\text{PK-PD}}$ and $\mathbf{SSM}_{\text{Linear}}$ over 6-month time windows in Supp. Figure A-4. We see that for some biomarkers, such as serum IgA, the PK-PD model has lower L1 error, while for others, such as Calcium, both models do very well.

A.6.4 Ablation Studies for $\mathbf{SSM}_{\text{PK-PD}}$

We report an ablation experiment in Supp. Table A.3, where we assess the effect of adding each mechanism function to $\mathbf{SSM}_{\text{PK-PD}}$ on held-out NELBO. We see that the

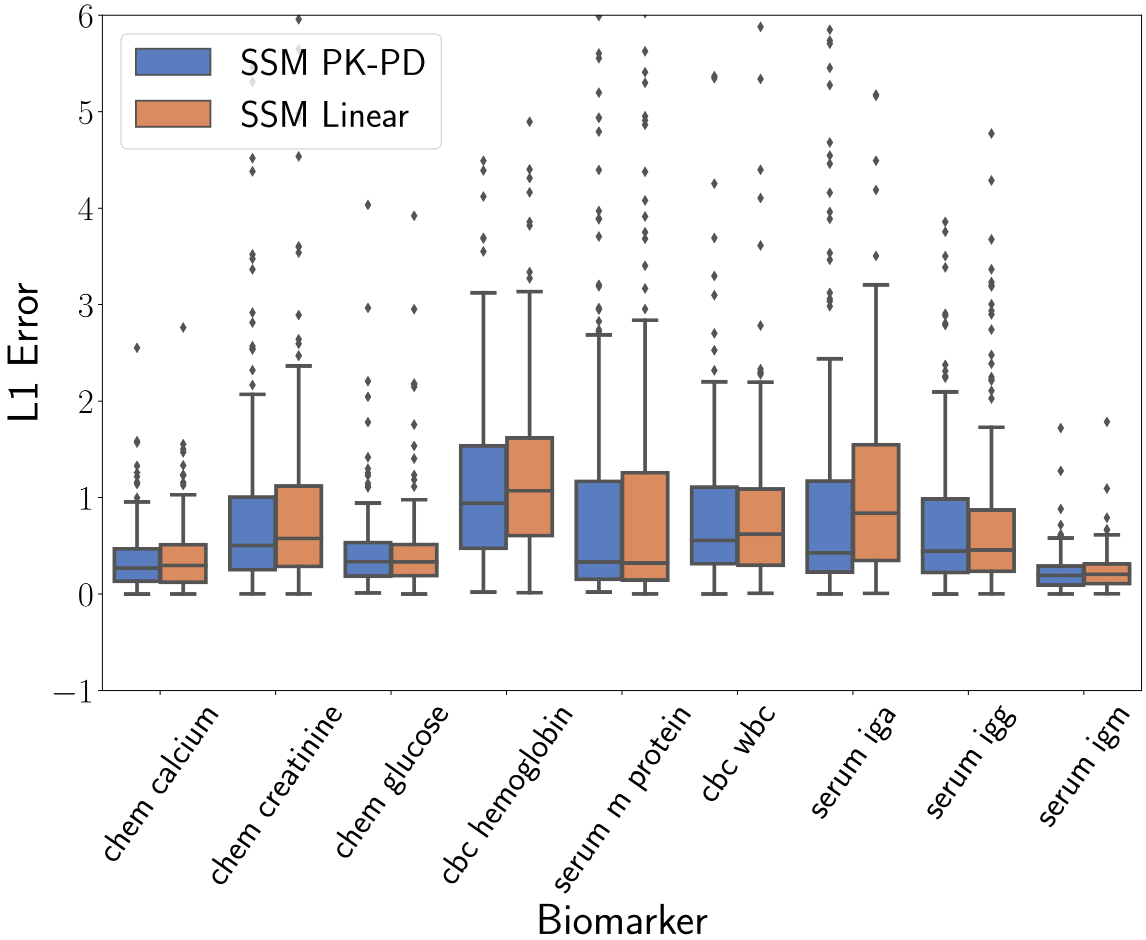


Figure A-4: L1 error for 6-month forward samples from PK-PD and Linear models: We report L1 error for forward samples over a 6-month time window conditioned only on baseline data.

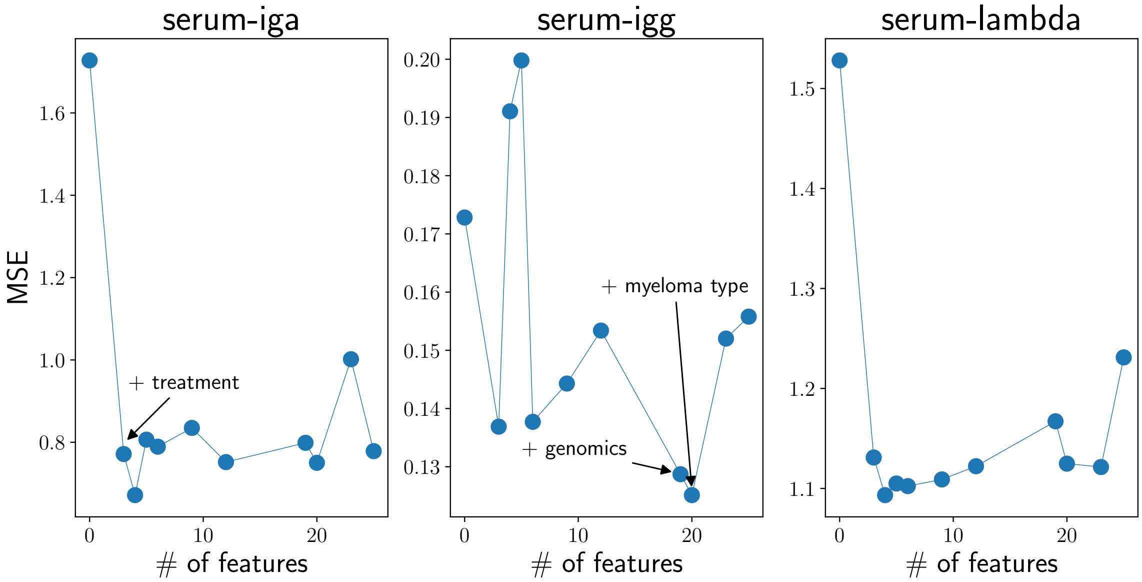


Figure A-5: Feature ablation of conditioning set for subset of biomarkers: We report the per-biomarker mean-squared error (MSE), averaged over all patients and all time points, of $\textbf{SSM}_{\text{PK-PD}}$ models trained on an increasing subset of baseline and treatment features.

Neural Log-Cill Kill function gives a modest improvement, while the addition of the Neural Treatment Exponential function gives most of the improvements.

Secondly, in Figure A-5, we show a feature ablation experiment to determine the importance of baseline and treatment features in forecasting several multiple myeloma markers. We train $\textbf{SSM}_{\text{PK-PD}}$ models on subsets of features, while tuning the latent variable size ($[16, 48, 64, 128]$) on a validation set for each subset. Then, we evaluate the mean-squared error (MSE), averaged over all examples and time points, of each trained model on a separate held-out set. Our results are shown in Figure A-5. We focus on serums IgA, IgG, and lambda, three biomarkers that are commonly tracked in multiple myeloma to evaluate response to treatment and overall progression of disease [156, 102].

We find that for serums IgA and Lambda, adding the treatment signal intuitively leads to a reduction in the MSE. For serum IgG, while the treatment signal helps with predictive performance, the baseline features, such as the genomic and myeloma type features, also seem to play a role.

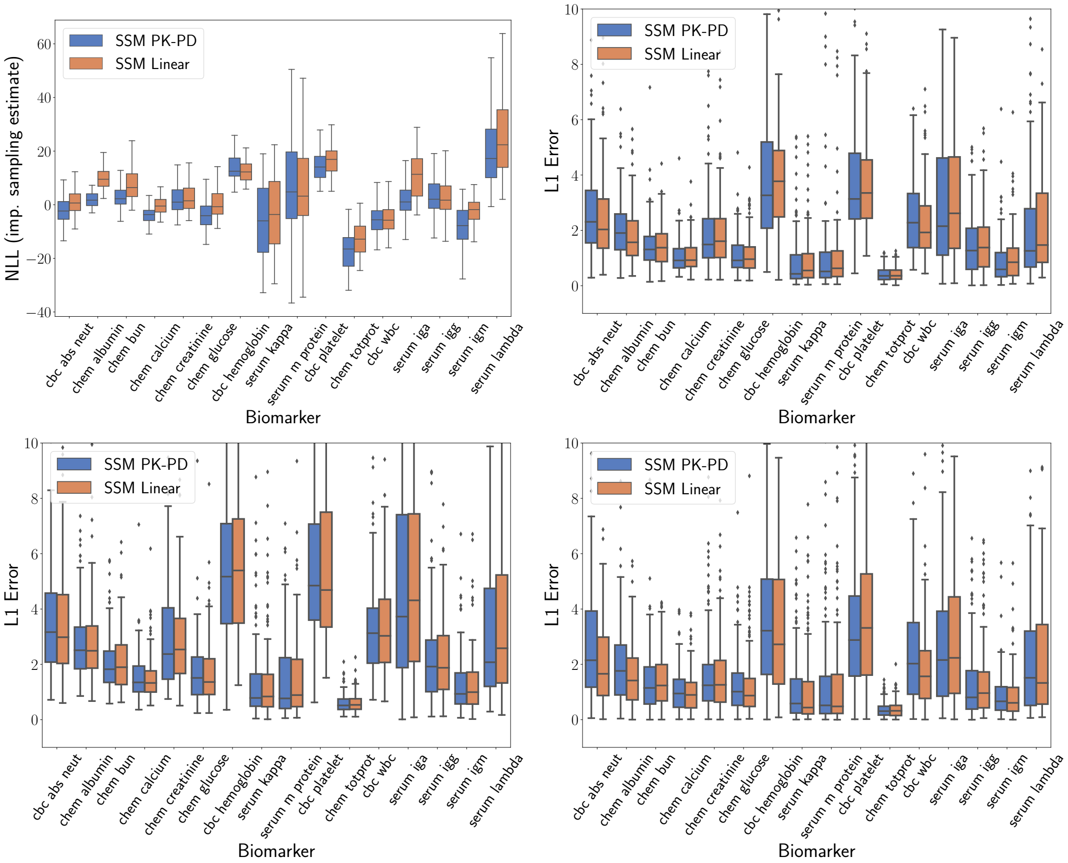


Figure A-6: a) NLL estimates via Importance Sampling [top left]: We estimate the NLL of $\mathbf{SSM}_{\text{PK-PD}}$ and $\mathbf{SSM}_{\text{Linear}}$ for each feature, summed over all time points and averaged over all patients. **b) Condition on 6 months, forward sample 1 year [top right]:** We show L1 Prediction Error for forward samples over a 1 year time window conditioned on 6 months of patient data. At each time point, we compute the L1 error with the observed biomarker and sum these errors (excluding predictions for missing biomarker values) over the prediction window. We employ this procedure for each patient. **c) Condition on 6 months, forward sample 2 years [bottom left]:** We report L1 error for forward samples over a 2 year window conditioned on 6 months of patient data. **d) Condition on 2 years, forward sample 1 year [bottom right]:** Finally, we report L1 error for forward samples over a 1 year time window conditioned on 2 years of patient data.

Appendix B

Supplementary Material for Chapter 4

B.1 Data Preprocessing

B.1.1 Sample Splitting

On the MM2 data, we perform a random 80/20 split of the data into training and test sets. We create five additional random 75/25 splits of the training set into a smaller training set that is used for model training and a validation set used for hyperparameter tuning. For the generalizability experiments, we evaluate the model on all MM1 patients.

B.1.2 Included Variables

We include the following variables for our analysis, filtering out any covariates that were below a missingness threshold of 15% for the baseline covariates and 70% (computed over the entire follow-up time) for the longitudinal covariates.

Baseline Variables, B

- Age, Race, Sex
- Time since diagnosis (months)
- Plasma cell (%), Bone marrow plasma cells (%)

- Baseline involved free light chain level (mg/L), Baseline creatinine (mg/dL), Baseline creatinine clearance (mL/min), Baseline albumin (g/dL)
- Any polymorphisms? (Separate binary covariates for presence of CC, CG, GG specific changes are included as well.)
- Cytogenetics result (abnormal, normal, or indeterminate)
- Baseline plasmacytoma observed
- History of bone lesions
- Extramedullary disease at study entry
- Measurable disease flag (Serum M-protein, Urine M-protein, Both SPEP and UPEP, Serum FLC)
- Liver function based on baseline tests
- ISS at baseline
- Myeloma type
- L-chain type at baseline (kappa, lambda, biclonal)
- Durie-Salmon Stage at initial diagnosis
- Lytic bone disease at initial diagnosis
- ISS stage at initial diagnosis
- ISS stage at study entry
- ECOG
- Beta-2 microglobulin (categorical)
- Free light chain ratio (categorical)
- Cytogenetics abnormality (categorical, high risk vs standard)

- Skeletal survey done at baseline
- Lytic bone lesion present at baseline
- del17 present
- t(4;14) present alone
- 1q amplification status
- Treatment arm

Longitudinal Variables, X

- Albumin (g/L)
- Alkaline Phosphatase (U/L)
- Alanine Aminotransferase (U/L), Aspartate Aminotransferase (U/L), Bilirubin (umol/L), Blood Urea Nitrogen (mmol/L)
- Calcium (mmol/L), Chloride (mmol/L), Corrected Calcium (mmol/L), Glucose (mmol/L), Potassium (mmol/L), Magnesium (mmol/L), Phosphate (mmol/L)
- Carbon Dioxide (mmol/L)
- Creatinine (umol/L), Glomerular Filtration Rate Adj for BSA via CKD-EPI (mL/min/1.73m2)
- Hematocrit, Hemoglobin (g/L), Lymphocytes (10^9/L), Monocytes (10^9/L), Neutrophils (10^9/L), Platelets (10^9/L), Leukocytes (10^9/L)
- Lactate Dehydrogenase (U/L)
- Protein (g/L), Serum Globulin (g/L), Sodium (mmol/L), SPEP Gamma Globulin (g/L), Free SPEP Kappa Light Chain (mg/L), Free SPEP Kappa Lt Chain/Free Lambda Lt Chain, Free SPEP Lambda Light Chain (mg/L), SPEP Monoclonal Protein (g/L), Serum TM Albumin/Globulin, Immunoglobulin A (g/L), Immunoglobulin G (g/L), Immunoglobulin M (g/L)

- Urine Albumin (%), UPEP Monoclonal Protein (mg/day), Urate (umol/L)

Treatment Variables, A

Dosages of lenalidomide, ixazomib, and dexamethasone are included in the longitudinal treatment information if the patient was assigned to the treatment group. Otherwise, if the patient was assigned to the control group, dosages of lenalidomide, dexamethasone, and a placebo pill are given.

B.1.3 Normalization

Our model relies on a wide range of clinical and biological signals that have different scales and variability. We normalize the baseline variables, B, by subtracting the mean of each variable and scaling by the inverse of the standard deviation. For each fold, the mean and standard deviations are computed on the training set alone.

For the longitudinal variables, X, we use a normalization strategy aimed at being clinically informative while still enforcing an informative variance over time. Immunoglobulin markers tend to be very large at baseline and quickly decrease over time. A typical standard scaling of this variable would then tend to mask the variability in the immunoglobulin levels after the first decrease. Yet, to predict progression, the ability to read the future trajectories of immunoglobulins is crucial.

Our normalization strategy uses normal clinical ranges for each variable. Letting (α_j, β_j) be the lower and upper bounds of the normal range of a variable j, we first compute

$$X^{j,*} = \frac{4 * (X^j - \alpha_j)}{\beta_j - \alpha_j} - 2 \tag{B.1}$$

For SPEP light chains, UPEP proteins, and urate, we further scale the variable by a factor $\frac{1}{5}$, to accommodate their values being significantly higher than the normal range. We then incorporate an invertible non-linearity to retain variability of the signal in the lower ranges. The final normalization for a longitudinal variable j is obtained using the following formula,

$$X^{j,\dagger} = \frac{7}{1 + e^{-0.25 \cdot X^{j,*}}} - 3.5. \tag{B.2}$$

B.2 Additional Quantitative Results

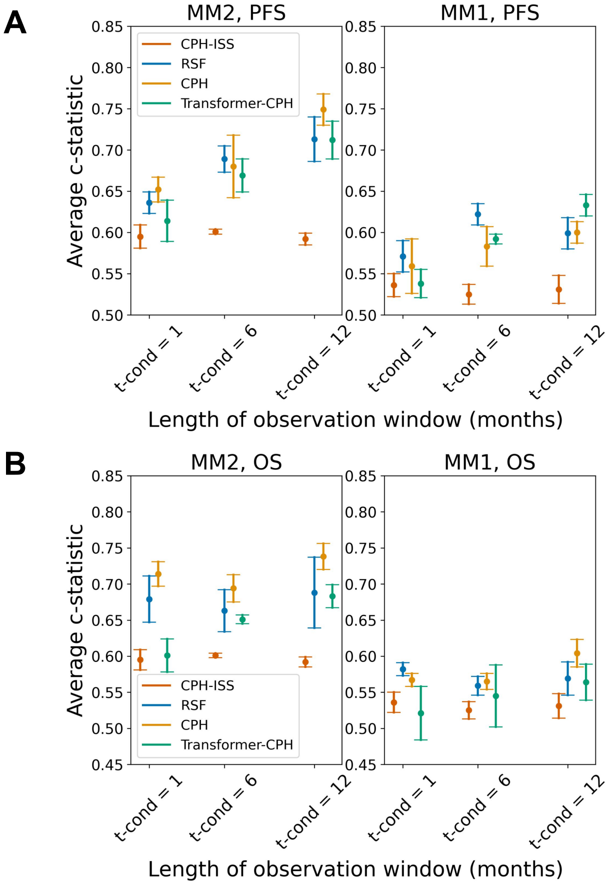


Figure B-1: (A): Results for PFS prediction **(B)**: Results for OS prediction

B.2.1 Additional PFS Prediction Results

We report PFS prediction results for all values of t_{cond} as well as patient subgroups by myeloma type (IgG-dominant and IgA-dominant).

	MM2			MM1		
	1	6	12	1	6	12
RSF	0.636 ± 0.013	0.689 ± 0.016	0.713 ± 0.027	0.571 ± 0.019	0.622 ± 0.013	0.599 ± 0.019
CPH	0.652 ± 0.015	0.68 ± 0.038	0.749 ± 0.019	0.559 ± 0.033	0.583 ± 0.024	0.6 ± 0.013
CPH-ISS	0.595 ± 0.014	0.601 ± 0.003	0.592 ± 0.007	0.536 ± 0.014	0.525 ± 0.012	0.531 ± 0.017
Transformer-CPH	0.614 ± 0.025	0.669 ± 0.02	0.712 ± 0.023	0.538 ± 0.017	0.592 ± 0.006	0.633 ± 0.013

Table B.1: PFS prediction results for all patients, t_{cond} in the columns

	MM2			MM1		
	1	6	12	1	6	12
RSF	0.673 ± 0.05	0.685 ± 0.058	0.676 ± 0.037	0.536 ± 0.014	0.576 ± 0.024	0.627 ± 0.031
CPH	0.632 ± 0.031	0.71 ± 0.041	0.639 ± 0.045	0.569 ± 0.022	0.572 ± 0.022	0.632 ± 0.023
CPH-ISS	0.583 ± 0.027	0.6 ± 0.006	0.592 ± 0.009	0.55 ± 0.024	0.498 ± 0.021	0.561 ± 0.016
Transformer-CPH	0.616 ± 0.029	0.589 ± 0.025	0.654 ± 0.039	0.55 ± 0.026	0.565 ± 0.012	0.681 ± 0.017

Table B.2: PFS prediction results for IgA-dominant myeloma subgroup, t_{cond} in the columns

	MM2			MM1		
	1	6	12	1	6	12
RSF	0.637 ± 0.021	0.699 ± 0.019	0.76 ± 0.033	0.592 ± 0.015	0.629 ± 0.017	0.614 ± 0.018
CPH	0.654 ± 0.017	0.693 ± 0.04	0.801 ± 0.025	0.564 ± 0.056	0.587 ± 0.03	0.608 ± 0.014
CPH-ISS	0.592 ± 0.009	0.596 ± 0.001	0.593 ± 0.007	0.55 ± 0.019	0.55 ± 0.009	0.536 ± 0.015
Transformer-CPH	0.639 ± 0.018	0.709 ± 0.025	0.746 ± 0.028	0.618 ± 0.006	0.621 ± 0.005	0.623 ± 0.018

Table B.3: PFS prediction results for IgG-dominant myeloma subgroup, t_{cond} in the columns

B.2.2 Additional OS Prediction Results

We report OS prediction results for all values of t_{cond} as well as patient subgroups by myeloma type (IgG-dominant and IgA-dominant).

	MM2			MM1		
	1	6	12	1	6	12
RSF	0.679 ± 0.032	0.663 ± 0.029	0.688 ± 0.049	0.582 ± 0.009	0.559 ± 0.013	0.569 ± 0.023
CPH	0.714 ± 0.017	0.694 ± 0.019	0.738 ± 0.018	0.567 ± 0.009	0.565 ± 0.011	0.604 ± 0.019
CPH-ISS	0.595 ± 0.014	0.601 ± 0.003	0.592 ± 0.007	0.536 ± 0.014	0.525 ± 0.012	0.531 ± 0.017
Transformer-CPH	0.601 ± 0.023	0.651 ± 0.006	0.683 ± 0.016	0.521 ± 0.037	0.545 ± 0.043	0.564 ± 0.025

Table B.4: OS prediction results for all patients, t_{cond} in the columns

B.2.3 Additional Adverse Event Prediction Results

We report adverse event prediction results for $t_{\text{cond}} = 1, 6$. We do not report $t_{\text{cond}} = 12$, since the number of events experienced by patients beyond $t = 12$ is too small to allow for meaningful prediction.

	MM2			MM1		
	1	6	12	1	6	12
RSF	0.656 ± 0.062	0.574 ± 0.056	0.664 ± 0.084	0.549 ± 0.024	0.504 ± 0.015	0.577 ± 0.023
CPH	0.674 ± 0.04	0.588 ± 0.06	0.573 ± 0.104	0.561 ± 0.011	0.54 ± 0.015	0.678 ± 0.026
CPH-ISS	0.583 ± 0.027	0.6 ± 0.006	0.592 ± 0.009	0.55 ± 0.024	0.498 ± 0.021	0.561 ± 0.016
Transformer-CPH	0.63 ± 0.028	0.562 ± 0.029	0.615 ± 0.027	0.503 ± 0.042	0.499 ± 0.064	0.604 ± 0.054

Table B.5: OS prediction results for IgA-dominant myeloma subgroup, t_{cond} in the columns

	MM2			MM1		
	1	6	12	1	6	12
RSF	0.641 ± 0.04	0.688 ± 0.029	0.701 ± 0.057	0.579 ± 0.015	0.557 ± 0.015	0.557 ± 0.034
CPH	0.67 ± 0.024	0.681 ± 0.029	0.749 ± 0.014	0.593 ± 0.008	0.57 ± 0.01	0.571 ± 0.021
CPH-ISS	0.592 ± 0.009	0.596 ± 0.001	0.593 ± 0.007	0.55 ± 0.019	0.55 ± 0.009	0.536 ± 0.015
Transformer-CPH	0.58 ± 0.021	0.67 ± 0.005	0.685 ± 0.011	0.54 ± 0.062	0.552 ± 0.05	0.546 ± 0.015

Table B.6: OS prediction results for IgG-dominant myeloma subgroup, t_{cond} in the columns

	AE-0	AE-1	AE-2	AE-3	AE-4	AE-5
RSF	0.59 ± 0.08	0.616 ± 0.048	0.499 ± 0.062	0.484 ± 0.118	0.568 ± 0.092	0.386 ± 0.262
CPH	0.626 ± 0.021	0.505 ± 0.106	0.468 ± 0.036	0.452 ± 0.069	0.631 ± 0.063	0.471 ± 0.149
Transformer-CPH	0.607 ± 0.03	0.533 ± 0.132	0.501 ± 0.054	0.371 ± 0.134	0.55 ± 0.121	0.463 ± 0.223

Table B.7: Prediction of Adverse Events at $t_{\text{cond}} = 1$ month. AE 0-5.

	AE-6	AE-7	AE-8	AE-9	AE-10	AE-11
RSF	0.534 ± 0.078	0.531 ± 0.055	0.729 ± 0.099	0.474 ± 0.047	0.589 ± 0.111	0.454 ± 0.096
CPH	0.616 ± 0.112	0.564 ± 0.022	0.756 ± 0.09	0.495 ± 0.046	0.694 ± 0.092	0.504 ± 0.066
Transformer-CPH	0.658 ± 0.098	0.473 ± 0.097	0.594 ± 0.102	0.521 ± 0.073	0.623 ± 0.114	0.459 ± 0.132

Table B.8: Prediction of Adverse Events at $t_{\text{cond}} = 1$ month. AE 6-11.

	AE-0	AE-1	AE-2	AE-3	AE-4	AE-5
RSF	0.584 ± 0.092	0.586 ± 0.124	0.436 ± 0.047	0.661 ± 0.157	0.483 ± 0.134	0.429 ± 0.169
CPH	0.663 ± 0.042	0.52 ± 0.106	0.472 ± 0.029	0.629 ± 0.111	0.36 ± 0.212	0.31 ± 0.19
Transformer-CPH	0.653 ± 0.092	0.544 ± 0.19	0.468 ± 0.056	0.58 ± 0.106	0.596 ± 0.191	0.369 ± 0.335

Table B.9: Prediction of Adverse Events at $t_{\text{cond}} = 6$ months. AE 0-5.

	AE-6	AE-7	AE-8	AE-9	AE-10	AE-11
RSF	0.471 ± 0.247	0.644 ± 0.036	0.437 ± 0.23	0.454 ± 0.107	0.676 ± 0.109	0.563 ± 0.127
CPH	0.441 ± 0.161	0.636 ± 0.054	0.229 ± 0.197	0.415 ± 0.048	0.929 ± 0.037	0.621 ± 0.121
Transformer-CPH	0.523 ± 0.165	0.57 ± 0.067	0.56 ± 0.367	0.483 ± 0.047	0.861 ± 0.13	0.431 ± 0.208

Table B.10: Prediction of Adverse Events at $t_{\text{cond}} = 6$ months. AE 6-11.

B.2.4 Forecasting Results & Hypothesis Testing

We report forecasting results for all covariates in addition to looking specifically at chemistry lab values (“chem”) as well as serum immunoglobulins (“serum”). We find that the the latter are harder to predict than the former. In both cases, Transformer-CPH does markedly better in forecasting. We verify this result by running pairwise t-tests comparing all methods effdover all observation windows (see Table B.12).

		1	6	12
all	LOCF	1.189 ± 0.0	0.245 ± 0.0	0.22 ± 0.0
	DMM	0.804 ± 0.038	0.614 ± 0.013	0.533 ± 0.019
	RNN	0.429 ± 0.038	0.263 ± 0.007	0.231 ± 0.01
	Tran-CPH	0.341 ± 0.011	0.208 ± 0.002	0.181 ± 0.003
chem	LOCF	0.392 ± 0.0	0.287 ± 0.0	0.26 ± 0.0
	DMM	0.676 ± 0.018	0.649 ± 0.015	0.588 ± 0.006
	RNN	0.338 ± 0.01	0.285 ± 0.004	0.254 ± 0.01
	Tran-CPH	0.278 ± 0.005	0.241 ± 0.003	0.207 ± 0.003
serum	LOCF	2.975 ± 0.0	0.209 ± 0.0	0.124 ± 0.0
	DMM	1.784 ± 0.192	0.949 ± 0.113	0.75 ± 0.067
	RNN	0.879 ± 0.157	0.312 ± 0.027	0.267 ± 0.044
	Tran-CPH	0.66 ± 0.038	0.193 ± 0.004	0.166 ± 0.008

Table B.11: Forecasting results

				1				6				12	
		LOCF	DMM	RNN	Tran-CPH	LOCF	DMM	RNN	Tran-CPH	LOCF	DMM	RNN	Tran-CPH
all	LOCF	1.0	<1e-4	<1e-4	<1e-4	1.0	<1e-4	0.0045	<1e-4	1.0	<1e-4	0.0697	<1e-4
	DMM	<1e-4	1.0	0.0001	<1e-4	<1e-4	1.0	<1e-4	<1e-4	<1e-4	1.0	<1e-4	<1e-4
	RNN	<1e-4	0.0001	1.0	0.0076	0.0045	<1e-4	1.0	0.0001	0.0697	<1e-4	1.0	0.0004
	Tran-CPH	<1e-4	<1e-4	0.0076	1.0	<1e-4	<1e-4	0.0001	1.0	<1e-4	<1e-4	0.0004	1.0
chem	LOCF	1.0	<1e-4	0.0003	<1e-4	1.0	<1e-4	0.3262	<1e-4	1.0	<1e-4	0.2508	<1e-4
	DMM	<1e-4	1.0	<1e-4	<1e-4	<1e-4	1.0	<1e-4	<1e-4	<1e-4	1.0	<1e-4	<1e-4
	RNN	0.0003	<1e-4	1.0	0.0003	0.3262	<1e-4	1.0	<1e-4	0.2508	<1e-4	1.0	0.0005
	Tran-CPH	<1e-4	<1e-4	0.0003	1.0	<1e-4	<1e-4	<1e-4	1.0	<1e-4	<1e-4	0.0005	1.0
serum	LOCF	1.0	0.0002	<1e-4	<1e-4	1.0	0.0001	0.001	0.0009	1.0	<1e-4	0.0019	0.0003
	DMM	0.0002	1.0	0.0012	0.0002	0.0001	1.0	0.0003	0.0001	<1e-4	1.0	0.0002	<1e-4
	RNN	<1e-4	0.0012	1.0	0.0387	0.001	0.0003	1.0	0.0006	0.0019	0.0002	1.0	0.0072
	Tran-CPH	<1e-4	0.0002	0.0387	1.0	0.0009	0.0001	0.0006	1.0	0.0003	<1e-4	0.0072	1.0

Table B.12: Forecasting p-values

Appendix C

Supplementary Material for Chapter 5

C.1 Proof of Proposition 1

Proposition 1. *Consider a sequence of plug-in estimates $\{\widehat{Q}(\tilde{\alpha}, .)\}_{\tilde{\alpha}}$ obtained from a calibration sample $\mathcal{D}_{c,1}$, and the corresponding conformity scores $\mathcal{V} = \{V(X_i, Y_i) : i \in \mathcal{D}_{c,2}\}$ obtained from another sample $\mathcal{D}_{c,2}$, where $\mathcal{D}_{c,1}$ and $\mathcal{D}_{c,2}$ are two disjoint subsets of $\mathcal{D}_c$. If $\{(X_i, Y_i) : 1 \leq i \leq n+1\}$ are exchangeable, then the interval in (5.11) satisfies*

$$\mathbb{P}(Y_{n+1} \in \widehat{C}_n(X_{n+1})) \geq 1 - \alpha.$$

Proof. The construction of the interval in (5.11) implies that, conditioned on $\mathcal{D}_{c,1}$, we have

$$Y_{n+1} \in \widehat{C}_n(X_{n+1}) \iff V(X_{n+1}, Y_{n+1}) \leq Q_{\mathcal{V}}(1-\alpha),$$

hence it follows that $\mathbb{P}(Y_{n+1} \in \widehat{C}_n(X_{n+1}) \,|\, \mathcal{D}_{c,1})) = \mathbb{P}(V(X_{n+1}, Y_{n+1} \leq Q_{\mathcal{V}}(1-\alpha) \,|\, \mathcal{D}_{c,1}))$. Since all training and calibration samples (X_i, Y_i) are exchangeable, then conditioned on $\mathcal{D}_{c,1}$, the calibration conformity scores in $\mathcal{V}$ and the conformity score on the test point $V(X_{n+1}, Y_{n+1})$ are exchangeable. By exchangeability of calibration and testing conformity scores, it follows from Lemma 2 in [223] that:

$$\mathbb{P}(V(X_{n+1}, Y_{n+1}) \leq Q_{\mathcal{V}}(1-\alpha) \,|\, \mathcal{D}_{c,1}) \geq 1-\alpha,$$

which concludes the statement after marginalizing over the randomness of $\mathcal{D}_{c,1}$. □

C.2 Influence function of the quantile functional

Recall that $Q(\alpha, x)$ is the level-α quantile of $Y|X = x$, i.e.,

$$Q(\alpha, x) = F^{-1}(\alpha) \coloneqq \inf\{y \in \mathbb{R} : F(y \,|\, X = x) \geq \alpha\},$$

where $F(.)$ is the conditional cumulative density function (CDF), $F(y \,|\, X = x) \coloneqq \mathbb{P}(Y \leq y \,|\, X = x)$. Throughout this Section, we assume that $\alpha = F(F^{-1}(\alpha))$. Recall that the influence function of the functional $Q(\alpha)$ at the distribution P for a given point (x, y) in the direction of the localized distribution G_x is defined as follows:

$$\mathrm{IF}(y; Q(\alpha), P) = \lim_{\varepsilon \to 0} \frac{Q_{P^y}(\alpha) - Q_P(\alpha)}{\varepsilon}, \tag{C.1}$$

where $P_\varepsilon = (1-\varepsilon)P + \epsilon\, \delta_y$. Let $F_\varepsilon = (1-\varepsilon)F + \varepsilon \cdot \delta_y$ be the CDF of the perturbed distribution, then we have

$$\begin{aligned} \alpha &= F_\varepsilon \circ F_\varepsilon^{-1}(\alpha) \\ &= (1-\varepsilon) \cdot F(F_\varepsilon^{-1}(\alpha)) + \varepsilon \cdot \delta_y(F_\varepsilon^{-1}(\alpha)). \end{aligned} \tag{C.2}$$

By differentiating both sides with respect to ε, we have the following

$$\begin{aligned} \frac{\partial \alpha}{\partial \varepsilon} = 0 &= \frac{\partial}{\partial \varepsilon}\left(F_\varepsilon \circ F_\varepsilon^{-1}(\alpha)\right) \\ &= \frac{\partial}{\partial \varepsilon}\left((1-\varepsilon) \cdot F(F_\varepsilon^{-1}(\alpha)) + \varepsilon \cdot \delta_y(F_\varepsilon^{-1}(\alpha))\right) \\ &= -F(F_\varepsilon^{-1}(\alpha)) + (1-\varepsilon) f(F_\varepsilon^{-1}(\alpha)) \cdot \frac{\partial F_\varepsilon^{-1}(\alpha)}{\partial \varepsilon} + \delta_y(F_\varepsilon^{-1}(\alpha)) + \varepsilon \cdot \frac{\partial}{\partial \varepsilon}\delta_y(F_\varepsilon^{-1}(\alpha)). \end{aligned} \tag{C.3}$$

Note that the influence function definition in (C.1) can be rewritten as follows:

$$\mathrm{IF}(y; Q(\alpha), P) = \lim_{\varepsilon \to 0} \frac{Q_{P^y}(\alpha) - Q_P(\alpha)}{\varepsilon} = \left(\frac{\partial}{\partial \varepsilon} F_\varepsilon^{-1}(\alpha)\right)_{\varepsilon=0}. \tag{C.4}$$

By setting $\varepsilon = 0$ in (C.3), we have the following:

$$\begin{aligned} 0 &= \left(-F(F_\varepsilon^{-1}(\alpha)) + (1-\varepsilon) f(F_\varepsilon^{-1}(\alpha)) \cdot \frac{\partial F_\varepsilon^{-1}(\alpha)}{\partial \varepsilon} + \delta_y(F_\varepsilon^{-1}(\alpha)) + \varepsilon \cdot \frac{\partial}{\partial \varepsilon} \delta_y(F_\varepsilon^{-1}(\alpha))\right)_{\varepsilon=0} \\ &= -F(F^{-1}(\alpha)) + f(F^{-1}(\alpha)) \cdot \left(\frac{\partial F_\varepsilon^{-1}(\alpha)}{\partial \varepsilon}\right)_{\varepsilon=0} + \delta_y(F^{-1}(\alpha)). \end{aligned} \tag{C.5}$$

By rearranging the terms in (C.5), we have

$$\left(\frac{\partial F_\varepsilon^{-1}(\alpha)}{\partial \varepsilon}\right)_{\varepsilon=0} = \frac{F(F^{-1}(\alpha)) - \delta_y(F^{-1}(\alpha))}{f(F^{-1}(\alpha))}. \tag{C.6}$$

Hence, the influence function of the quantile is given by:

$$\mathrm{IF}(y; Q(\alpha), P) = \left(\frac{\partial F_\varepsilon^{-1}(\alpha)}{\partial \varepsilon}\right)_{\varepsilon=0} = \frac{\alpha - \mathbf{1}_{\{F^{-1}(\alpha) \geq y\}}}{f(F^{-1}(\alpha))}. \tag{C.7}$$

C.3 Additional Experiments

We test our method on 9 datasets: **MEPS-19**, **MEPS-20**, **MEPS-21**, **Facebook-1**, **Facebook-2**, **Bio**, **Kin8nm**, **Naval** and **Blog**. The MEPS (medical expenditure panel survey) datasets comprise surveys of families and individuals, their medical providers, and employers across the United States collected over three years [223, 82, 57]. The facebook datasets contain features extracted from facebook posts, and the task is to predict how many comments the post will receive. The difference between **Facebook-1** and **Facebook-2** are the features that are included. The **Bio** dataset consists of features describing the physiochemical properties of protein tertiary structure, including fractional areas of different regions of the protein, distances between particular amino acids, and other spacial distribution constraints. The task is to predict the size of the residue. The remaining data sets are available in the UCI repository.

Evaluating Transparency In Table C.1, we show the marginal coverage, worst-

	MEPS-19			MEPS-20			MEPS-21		
	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$
QR methods									
QR-RF	0.90	1.00	0.54	0.91	1.06	0.55	0.93	1.29	0.65
QR-NN	0.79	0.54	0.67	0.81	0.56	0.64	0.79	0.53	0.64
CP methods									
CP	0.89	1.28	0.19	0.90	1.24	0.15	0.90	1.29	0.16
LACP	0.89	0.61	0.20	0.89	0.59	0.20	0.90	0.62	0.26
CQR	0.89	1.12	0.46	0.89	1.05	0.50	0.89	1.27	0.62
CCH	0.96	5.37	0.79	0.97	5.35	0.75	0.97	5.35	0.78
CQ	0.87	2.02	0.76	0.88	2.01	0.87	0.88	2.10	0.88
CUQR	0.89	1.25	0.73	0.89	1.21	0.76	0.90	1.26	0.81

	Facebook-1			Facebook-2			Bio		
	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$
QR methods									
QR-RF	0.93	0.85	0.78	0.92	0.81	0.86	0.92	1.33	0.84
QR-NN	0.81	0.55	0.68	0.80	0.52	0.64	0.81	0.94	0.56
CP methods									
CP	0.90	1.39	0.72	0.90	1.39	0.82	0.90	2.42	0.85
LACP	0.90	0.69	0.76	0.90	0.69	0.84	0.90	1.15	0.83
CQR	0.90	0.83	0.77	0.90	0.83	0.87	0.90	1.36	0.83
CCH	0.89	0.72	0.65	0.89	0.64	0.67	0.90	1.07	0.85
CQ	0.89	1.34	0.79	0.89	1.35	0.87	0.90	2.41	0.80
CUQR	0.90	1.36	0.87	0.89	1.36	0.87	0.90	2.40	0.88

	Kin8nm			Naval			Blog		
	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$	C_{av}	L_{av}	$C_G^{w.c.}$
QR methods									
QR-RF	0.93	1.36	0.89	0.90	0.63	0.87	0.79	0.73	0.76
QR-NN	0.79	0.94	0.74	0.78	0.55	0.73	0.79	0.73	0.76
CP methods									
CP	0.90	2.17	0.83	0.89	1.31	0.78	0.89	1.89	0.57
LACP	0.90	1.09	0.84	0.89	0.60	0.85	0.89	1.06	0.63
CQR	0.90	1.33	0.85	0.89	0.70	0.85	0.90	1.34	0.82
CCH	0.89	1.14	0.86	0.89	0.48	0.83	0.98	5.58	0.96
CQ	0.89	2.16	0.85	0.87	1.27	0.87	0.87	1.81	0.76
CUQR	0.89	2.19	0.85	0.86	1.26	0.85	0.87	1.82	0.67

Table C.1: Marginal coverage, efficiency and conditional coverage of all baselines on benchmark data sets.

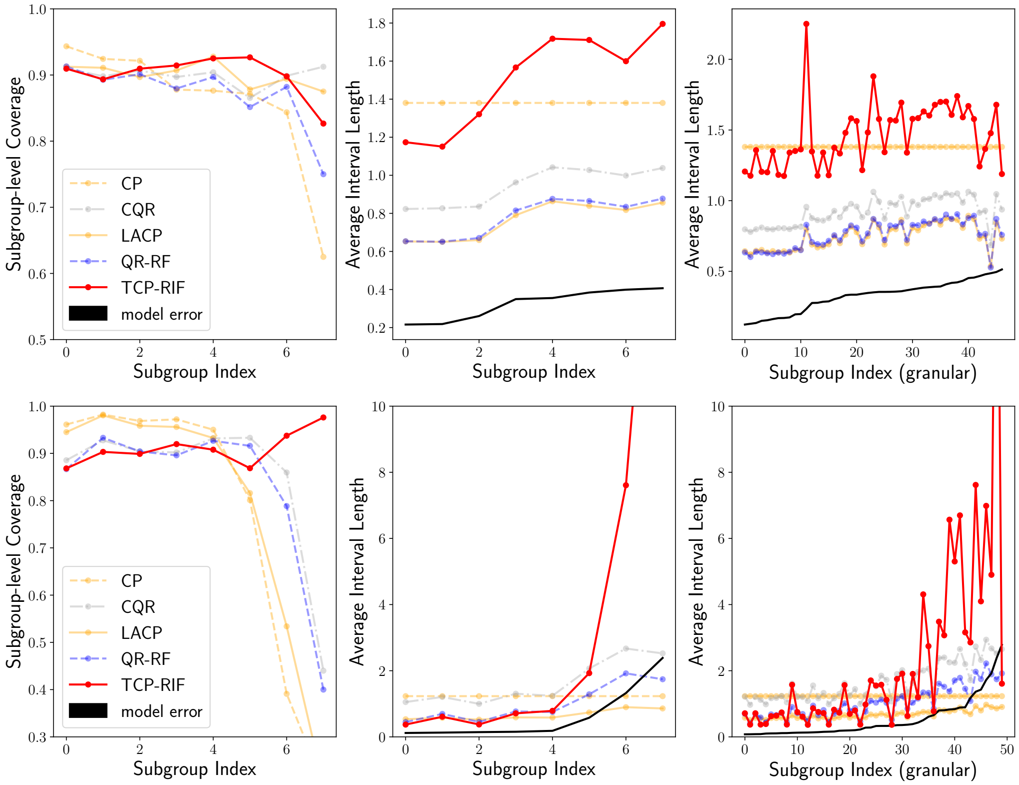


Figure C-1: Adaptivity of predictive inference baselines on the **Facebook-1** and **MEPS-20** dataset.

case subgroup coverage, and average interval length for all baselines as well as our CUQR approach on the three additional datasets. In general, we recapitulate the observations that we discuss in the main paper, though we will take care to provide additional analysis for some of the more competitive baselines. Regarding the QR-RF baseline, we find that although it is superior in terms of efficiency, it does not have theoretical guarantees in terms of marginal or conditional coverage. Indeed, the conditional coverage is worse than the worst-case conditional coverage achieved by CUQR in all three datasets. With respect to CQR, although it is competitive in the **Facebook-2** dataset, it does not have any theoretical guarantees for coverage at the subgroup level (only marginal coverage is guaranteed). Consequently, we see for **Facebook-1** and **Bio** inferior conditional coverage compared to CUQR.

Evaluating Adaptivity Similar to the experiment described in the main paper, we continue to evaluate the extent to which baselines are adaptive, i.e. how congruent the lengths of the intervals are with the true uncertainty of the base model, as reflected by the model error. We run this experiment on the **Facebook-1** and **MEPS-20** datasets. We see, as before, that either the baseline is not adaptive (e.g. CP or LACP) or has limited adaptiveness, but *not* enough to maintain the target coverage in the subgroup (e.g. QR-RF, CQR). QR-RF and CP, in particular, have a steep dropoff in coverage for the subgroups in which the model is more uncertain (i.e. higher indexed subgroups). On the other hand, we see CUQR both maintaining coverage in all subgroups and adapting the interval lengths given the uncertainty in the subgroup.

Appendix D

Supplementary Material for Chapter 6

D.1 When can biased estimators be falsified?

As discussed in Examples 6.2.1 and 6.2.2, we imagine that observational estimators differ in a few possible ways. They may represent the same identification strategy applied to different datasets, different identification strategies applied to the same dataset (e.g., different choices of confounders), or some combination of the two.

Assumption 6.2.3 states that there exists a consistent and asymptotically normal observational estimator for τ, as defined in Def. E.17. This is a fundamental assumption in our work, and so we build additional intuition for when we might expect this condition to hold, and when we might be able to falsify this assumption. In this section, we give basic intuition regarding patterns of confounding, and in Section D.2, we discuss issues of transportability.

In Example D.1.1, we give a simple example where the causal graph is consistent across two subgroups, and where an estimator must control for all confounders to get consistent estimates of the GATE in either subgroup. In this setting, falsification is possible. On the other hand, in Example D.1.2, we give a counterexample, where there are multiple estimators that can deliver consistent estimates of the GATE on the RCT subpopulation, but only one provides consistent estimates across all subpopulations.

Example D.1.1 (Consistent confounding across subgroups)**.** In the causal graph

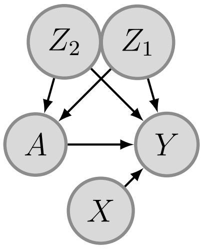


Figure D-1

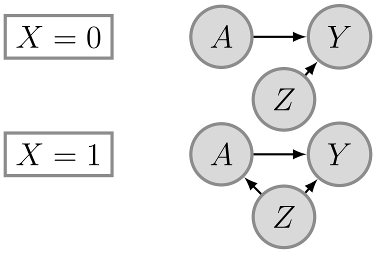


Figure D-2

Figure D-3: Example D.1.1 is depicted in (D-1), and Example D.1.2 in (D-2)

shown in Figure D-1, there are two sets of confounders, Z_1, Z_2, a binary treatment variable A, a binary subgroup variable X, and the outcome Y. We assume a linear outcome model, whereby $E[Y|X, Z_1, Z_2, A] = \alpha + \beta X + \gamma_1 AX + \gamma_2 A(1 - X) + \delta_1 Z_1 + \delta_2 Z_2$. Note that the true group average treatment effect (GATE) for the two subgroups are, $\text{GATE}(X = 0) = \gamma_2; \text{GATE}(X = 1) = \gamma_1$. It is straightforward to show that not conditioning on the full set of confounders will lead to biased GATE estimates for both subgroups, whereas conditioning on both Z_1 and Z_2 will lead to consistent estimates for both subgroups.

Example D.1.2 (Selective confounding by subgroup)**.** Let there be two subgroups, $X = 0$ and $X = 1$, with the former having support in both RCT and observational studies and the latter having support in only observational data. Now, suppose we had the following treatment assignment mechanism, $p(A = 1|X, Z) = f(Z) \cdot \mathbf{1}(X = 1) + c \cdot \mathbf{1}(X = 0)$, where Z is a set of confounders, f is a nonlinear function of Z, and c is a constant. A candidate estimator that does not condition on Z would be able to get consistent estimates for the validation effect but not the extrapolated effect. On the other hand, conditioning on Z would allow for consistent estimates on both validation and extrapolated effects.

D.2 Conditions for valid observational / randomized comparisons

Recall that we had defined the group average treatment effect (GATE) as follows in Equation (E.17)

$$\tau_i := \begin{cases} \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 0], & \text{if } i \in \mathcal{I}_R \\ \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 1], & \text{if } i \in \mathcal{I}_O \end{cases}, \tag{D.1}$$

and refer to τ_i for $i \in \mathcal{I}_R$ as a validation effect, and τ_i for $i \in \mathcal{I}_O$ as an extrapolated effect. In this section, we discuss sufficient conditions under which these causal effects are identifiable from observational data drawn from a distribution $D = k$, and give examples of doubly-robust estimators of these quantities. These assumptions cover both comparisons of the observational studies to the randomized trial (used for validation), as well as the normalization of observational estimates (used for confidence intervals on the extrapolated effects).

Our goal in presenting these results is to build intuition in this setting for when we might expect a consistent observational estimator to exist across all groups. This is a well-studied topic, often in the context of generalizing effect estimates from randomized trials to other supported populations (e.g., all trial-eligible individuals). We primarily make use of results in that literature to build intuition here, pointing the reader to Degtiar and Rose [66] for a recent review whose presentation we largely mirror, with modifications to account for our notation.

D.2.1 Identification

First, we state standard assumptions under which the GATE in the observational population for $D = k$,

$$\mathbb{E}[Y_1 - Y_0 \mid G = i, D = k], \tag{D.2}$$

is identifiable from data in the dataset $D = k$, with notation adapted to our setting.

Assumption D.2.1. The following conditions hold for the distribution $\mathbb{P}(\cdot \mid D = k)$:

1. *Conditional Exchangeability over A*: $Y_a \perp\!\!\!\perp A \mid X, D = k$ for all treatments a.
2. *Positivity of Treatment Assignment*: $\mathbb{P}(X = x \mid D = k) > 0 \implies \mathbb{P}(A = a \mid X = x, D = k) > 0$ for all a.
3. *Consistency*: $A = a \implies Y_a = Y$

These causal assumptions ensure that the ATE and CATE can be identified from observational data for the observational population and are standard in the causal inference literature [126]. In order to transport these estimates to the RCT population (or from one observational dataset to another), we require additional assumptions. Next, we give assumptions under which these estimates can be transported to another population $D = k'$, where in our case $k' \in \{0, 1\}$.

Assumption D.2.2. Let k correspond to a source population, and k' correspond to the target population. Conditioned on the event $D \in \{k, k'\}$, define the random variable $S = 1$ if $D = k$ and $S = 0$ otherwise. Then let the following hold, on the distribution $\mathbb{P}(\cdot \mid D \in \{k, k'\})$.

1. *Conditional Exchangeability over S*: $Y_a \perp\!\!\!\perp S \mid X$ for all treatments a.
2. *Positivity of Selection*: $\mathbb{P}(X = x) > 0 \implies \mathbb{P}(S = 1 \mid X = x) > 0$ almost surely over X for all a.
3. *Consistency*: $S = s$ and $A = a \implies Y_a = Y$

Here, we note that this introduces non-trivial additional assumptions. Most notably, we require that the potential outcomes are independent of the dataset, given X. This would be violated, for instance, if the distribution of unobservable effect modifiers differs between different observational studies. As a result, we note that it is possible for an observational study to fail to replicate the RCT results due to failures of transportability (failure of Assumption D.2.2) even if it has "internal validity", allowing for identification of the causal effect in the population $D = k$. There also exists a large body of work on identifying transportable causal effects via causal graphs [203, 204, 202].

D.2.2 Estimation of the ATE in the target population

Regarding estimation, Dahabreh et al. [64] consider the problem of transporting average treatment effects from randomized trials to observational studies, under Assumption D.2.1 with $k = 0$ and Assumption D.2.2 with $k = 0, k' = 1$. These assumptions admit identification of the potential outcomes means as follows (see Section 4.2 of Dahabreh et al. [64])

$$\mathbb{E}[Y_a \mid S = 0] = \mathbb{E}[\mathbb{E}[Y \mid X, S = 1, A = a] \mid S = 0] \tag{D.3}$$

where the outer expectations are over $\mathbb{P}(X \mid S = 0)$, i.e., the covariate distribution of the target population. Dahabreh et al. [61] give a doubly robust estimator for the statistical quantity on the right-hand side as the empirical expectation of the following pseudo-outcome (see Equation A.13 of Dahabreh et al. [64])

$$\hat{\mu}(a) = \frac{1}{n}\sum_{i=1}^{n} Y_i^a(\hat{\eta}, \hat{\pi}) \tag{D.4}$$

where n is the total samples in both the source $S = 1$ and target $S = 0$ samples, and where

$$Y_i^a(\hat{\eta}, \hat{\pi}) \coloneqq \frac{1}{\hat{\pi}}\left(\mathbf{1}\left\{S_i = 1, A_i = a\right\} \cdot \frac{1 - \hat{p}(X_i)}{\hat{p}(X_i)\hat{e}_a(X_i)} \cdot \{Y_i - \hat{g}_a(X_i)\} + (1 - S_i)\hat{g}_a(X_i)\right). \tag{D.5}$$

In Equation (D.5), $\hat{\eta} \coloneqq (\hat{g}_a, \hat{e}_a, \hat{p})$, and $\hat{\pi} \coloneqq n^{-1}\sum_{i=1}^{n}\mathbf{1}\left\{S_i = 0\right\}$ is an estimate of $\mathbb{P}(S = 0)$, $\hat{g}_a(X)$ is an estimate of the mean conditional outcome $\mathbb{E}[Y \mid A = a, S = 1, X]$, $\hat{p}(X)$ is an estimate of the selection probability $\mathbb{P}(S = 1 \mid X)$, and $\hat{e}_a(X)$ is an estimate of the propensity score $\mathbb{P}(A = a \mid S = 1, X)$. Dahabreh et al. [61] derives precise asymptotic properties of this estimator, which is asymptotically normal and consistent for the observational quantity on the right-hand side of Equation (D.3). In particular, this estimator is doubly-robust in the sense that it is consistent if either $\hat{p}(X)$ or $\hat{g}_a(X)$ is consistent, but requires consistency of $\hat{e}_a(X)$. It also enjoys the rate double-robustness property, retaining consistency and asymptotic normality even if the estimators for $\hat{p}, \hat{g}$ converge at slower than parametric rates, and allows for the same cross-fitting schemes used in the Double ML [50] literature for relaxing Donsker conditions.

Note that the average treatment effect in this setting can be estimated by the following contrast, which is similarly an empirical expectation of a pseudo-outcome

$$\hat{\mu}(1) - \hat{\mu}(0) = \frac{1}{n}\sum_{i=1}^{n} Y_i^1(\hat{\eta}, \hat{\pi}) - Y_i^0(\hat{\eta}, \hat{\pi}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\hat{\eta}, \hat{\pi}), \tag{D.6}$$

where $Y_i(\hat{\eta}, \hat{\pi}) \coloneqq Y_i^1(\hat{\eta}, \hat{\pi}) - Y_i^0(\hat{\eta}, \hat{\pi})$. Furthermore, the variance of these estimates can be estimated using either sandwich estimators from M-estimation theory [255], or via bootstrap methods. We refer the reader to Sections 5.3, 5.4 and Appendix A.4 of [64] for more details.

D.3 Estimation and comparison of GATE in semi-synthetic experiments

In Sections 6.2.2 and D.2, we discuss several estimators for average treatment effects (ATEs) that are known to be asymptotically normal, such as the double ML estimator discussed in Example 6.2.1 or the doubly-robust estimator in Section D.2.

Given a fixed set of discrete subgroups, one could analyze each subgroup independently and apply such estimators directly, since the ATE in each subgroup is precisely the GATE. This would be a straightforward way to ensure that the same formal guarantees hold regarding asymptotic normality. While this approach would be feasible in our experimental setting, due to the small number of groups, it is less practical in general, especially with a larger number of groups, since information cannot be shared across nuisance models such as $\hat{g}_a, \hat{e}_a, \hat{p}$ discussed in Section D.2.

In an effort to emulate a more realistic setting, we take a slightly different approach in the semi-synthetic experiments. We draw inspiration from the double ML approach given in Semenova and Chernozhukov [241] for GATE estimation, while taking into consideration the transportation of causal effects in the sense of Section D.2. Note that in Semenova and Chernozhukov [241], the required assumptions and proofs for asymptotic normality of estimators are provided on a case-by-case basis, which does not include our case with transportation. Therefore, in the following we will briefly describe their approach, then show how we construct our GATE estimators and provide the required assumptions for their asymptotic normality.

Semenova and Chernozhukov [241] focuses on the setting where there exists some pseudo-outcome / signal, $Y(\eta)$, and where one is interested in summarizing the function, $\tau(x) = \mathbb{E}[Y(\eta) \mid X = x]$, with a linear regression function (in the simplest case, a set of group indicators). When $Y(\eta)$ is the doubly-robust score [222, 221] (see Equation (D.11)), $\tau(x)$ is equal to the CATE function, and the best approximation by group indicators gives the GATE.

Our general procedure is as follows: for estimation of $\hat{\tau}(k)$ and the respective variances, we construct a score function / pseudo-outcome, $\tilde{Y}$, whose empirical conditional expectation (in each group) provides an estimate of the GATE, and whose empirical variance we use as an estimate of the variance. We describe this procedure in more detail below. Throughout, X should be taken to refer to the covariates that are observed in a given observational study.

Comparing Validation Effect Estimates In our simulation setup, all of the observational datasets are drawn from a common distribution, which differs from the RCT distribution, requiring the use of the techniques and assumptions discussed in Section D.2 to estimate the GATE, $\tau_i = \mathbb{E}[Y_1 - Y_0 \mid G = i, D = 0]$, using data from the observational distributions.

To generate the observational estimates $\hat{\tau}_i(k), \hat{\sigma}_i^2(k)$ in this setting, we cannot simply take empirical conditional expectation / variance of the score function given in Equation D.6. Rather, the GATE is identified under Assumptions D.2.1 and D.2.2 as a conditional expectation of the score times a correction factor, as discussed in the following proposition.

Proposition D.3.1. *In the setting of Section D.2, under Assumptions D.2.1 and D.2.2, the conditional mean potential outcome in the target distribution is identified as*

$$\mathbb{E}[Y_a \mid S = 0, G = i] = \frac{\mathbb{P}(S = 0)}{\mathbb{P}(S = 0 \mid G = i)} \mathbb{E}[Y^a(\eta_0, \pi_0) \mid G = i], \tag{D.7}$$

where $Y^a(\eta, \pi)$ is defined as in Equation D.8.

$$Y^a(\eta, \pi) \coloneqq \frac{1}{\pi} \left(\mathbf{1}\{S = 1, A = a\} \cdot \frac{1 - p(X)}{p(X)e_a(X)} \cdot \{Y - g_a(X)\} + (1 - S)g_a(X) \right) \tag{D.8}$$

where $\eta \coloneqq (g_a, e_a, p)$ with true underlying parameters $\eta_0 = (g_{a0}, e_{a0}, p_0)$, $\pi \coloneqq \mathbb{P}(S = 0)$ with true value π_0, $g_a(X) \coloneqq \mathbb{E}[Y \mid A = a, S = 1, X]$, $p(X) \coloneqq \mathbb{P}(S = 1 \mid X)$, and $e_a(X) \coloneqq \mathbb{P}(A = a \mid S = 1, X]$.

A proof is provided in Appendix D.4. Note that this is equivalent to replacing the estimate of $1/\mathbb{P}(S = 0)$ in the score with an estimate of $1/\mathbb{P}(S = 0 \mid G = i)$, before computing the empirical conditional expectations of the score.

Now, for each observational dataset, we construct estimates $\hat{\tau}_i(k), \hat{\sigma}_i^2(k)$ for $i \in \mathcal{I}_R$ as follows:

1. We collect observational samples from the two validation groups {lbw, married} and {hbw, married}, which we denote as $G = 0, G = 1$ respectively. We combine these observational samples with the samples from the RCT, using $S = 0$ to denote RCT samples (the target distribution) and $S = 1$ to denote observational samples.

2. We define our signal for each sample as

$$\tilde{Y}_i(\hat{\eta}, \hat{\pi}_g) \coloneqq \tilde{Y}_i^1(\hat{\eta}, \hat{\pi}_g) - \tilde{Y}_i^0(\hat{\eta}, \hat{\pi}_g) \tag{D.9}$$

where we define the modified score $\tilde{Y}_i^a$, in light of Proposition D.3.1, as

$$\tilde{Y}_i^a(\hat{\eta}, \hat{\pi}_g) := \frac{1}{\hat{\pi}_g(G_i)} \Bigg(\mathbf{1}\left\{S_i = 1, A_i = a\right\} \cdot \frac{1 - \hat{p}(X_i)}{\hat{p}(X_i)\hat{e}_a(X_i)} \cdot \{Y_i - \hat{g}_a(X_i)\} \\ + (1 - S_i)\hat{g}_a(X_i) \Bigg), \tag{D.10}$$

where $\hat{\pi}_g(G_i)$ is defined as an estimate of $\pi_g(G_i) := \mathbb{P}(S = 0 \mid G_i)$, computed using empirical averages.

3. We use 3-fold cross-fitting as described in Semenova and Chernozhukov [241] to generate the signals for each sample, such that for the i-th datapoint, the score $\tilde{Y}_i(\hat{\eta}, \hat{\pi})$ uses plug-in estimates $\hat{\eta} = (\hat{g}_1, \hat{g}_0, \hat{e}_1, \hat{p})$ that are learned on the folds that do not include the i-th datapoint, and $\hat{\pi}$ is estimated using empirical averages. In practice, we use a multi-layer perceptron (MLP) regressor for estimating $\hat{g}_a$, and ℓ_2-regularized logistic regression for estimating $\hat{e}_1, \hat{p}$, with hyperparameters described in Section D.6. For each model, we reserve 20% of the current fold in the cross fitting procedure as a validation set to do hyperparameter selection.

4. Finally, we estimate $\hat{\tau}_i(k)$ as the empirical average $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g) \mid G = i]$, and we use the empirical conditional variance of this score to estimate the variance $\hat{\sigma}_i^2(k)$.

We construct the RCT estimate $\hat{\tau}_i(0)$ (using the RCT sample alone) as the difference of the empirical conditional means $\mathbb{E}_{N_0}[Y \left(\frac{\mathbf{1}\{A=1\}}{\hat{P}(A=1)} - \frac{\mathbf{1}\{A=0\}}{1-\hat{P}(A=1)} \right) \mid G = i]$, where $\hat{P}(A = 1)$ is an empirical average. We compute $\hat{\sigma}_i^2(0)$ as the empirical conditional variance of this quantity. We then conduct testing, as described in Algorithm 2.

Asymptotic normality of transported estimators We herein provide sufficient assumptions that guarantee the asymptotic normality of our transported GATE estimators, i.e. the empirical average $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g) | G = i]$:

Assumption D.3.1 (Observational dataset covers the whole support of covariates)**.**

$$\inf_{x \in \mathcal{X}} p_0(x) = \varepsilon_p > 0$$

Note that Assumption D.3.1 is implied by Assumption 7.2.1.

Assumption D.3.2 (Bounded within-subgroup variance of conditional treatment effects in the RCT)**.**

$$\sup_{x \in \mathcal{X}} var[g_{10}(x) - g_{00}(x) | G = i, S = 0] = \sigma^2_{\tau i} < \infty$$

Assumption D.3.3 (Overlap between treatments in the observational dataset)**.**

$$\inf_{x \in \mathcal{X}} \min(e_{00}(x), e_{10}(x)) = \varepsilon_e > 0$$

Assumption D.3.4 (Finite outcome conditional variance in the observational dataset)**.**

$$\max_{a \in \{0,1\}} \sup_{x \in \mathcal{X}} \mathbb{E}[(Y - g_{a0}(x))^2 | X = x, S = 1, A = a] = \bar{\sigma}^2 < \infty$$

Assumption D.3.5 (Properties of the nuisance function estimators)**.** Let $\hat{\eta}_{(n)}$ be a sequence of estimators for η indexed by the size of the cross-fitting training fold n. We assume that there exists

- $\epsilon_n = o_P(1)$, a sequence of positive numbers

- $\mathcal{T}_n$, a sequence of nuisance function vector sets in the neighborhood of $\eta_0 = (g_{10}, g_{00}, e_{10}, p_0)$ satisfying $\mathbb{P}(\hat{\eta}_{(n)} \in \mathcal{T}_n) \geq 1 - \epsilon_n$

- $\mathbf{g}_n, \mathbf{e}_n, \mathbf{p}_n$, sequences of worst root mean square errors for the nuisance functions

g_1, g_0, e_1, p, defined as follows:

$$\mathbf{g}_n \coloneqq \max_{a\in\{0,1\}} \sup_{\eta\in\mathcal{T}_n} \sqrt{\mathbb{E}[g_a(X) - g_{a0}(X)]^2}$$
$$\mathbf{e}_n \coloneqq \sup_{\eta\in\mathcal{T}_n} \sqrt{\mathbb{E}[e_1(X) - e_{10}(X)]^2}$$
$$\mathbf{p}_n \coloneqq \sup_{\eta\in\mathcal{T}_n} \sqrt{\mathbb{E}[p(X) - p_0(X)]^2}$$

so that the following assumptions hold:

Assumption A: (Rate of nuisance error)

$$\mathbf{g}_n \vee \mathbf{e}_n \vee \mathbf{p}_n = o_P(1)$$

Assumption B: (Rate of nuisance error product)

$$\sqrt{n}\mathbf{g}_n(\mathbf{e}_n \vee \mathbf{p}_n) = o_P(1)$$

Assumption C: (Bounded nuisance estimates)

$$\sup_{\eta\in\cup_{n=1}^{\infty}\mathcal{T}_n} \left(\max_{a\in\{0,1\}} \sup_{x\in\mathcal{X}} |g_a(x)| \vee \sup_{x\in\mathcal{X}} \left|\frac{1}{p(x)}\right| \vee \sup_{x\in\mathcal{X}} \left|\frac{1}{e_1(x)}\right| \vee \sup_{x\in\mathcal{X}} \left|\frac{1}{e_0(x)}\right| \right) = \bar{\mathcal{C}} < \infty$$

Theorem D.3.1. *Suppose Assumptions D.3.1 to D.3.5 hold. Then, the empirical average, $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$, where $\tilde{Y}$ is defined in Equation D.9 and $\hat{\eta}$ is estimated with cross-fitting, is asymptotically normal.*

Remark: As we will prove later in section D.4.2, Assumptions D.3.1 to D.3.4 guarantee that when the nuisance function vector $(g_{10}, g_{00}, e_{10}, p_0)^\top$ is known (i.e. need not be estimated), the transported GATE estimator is asymptotically normal. In practice, $(g_{10}, g_{00}, e_{10}, p_0)^\top$ is not known and has to be estimated, so Assumption D.3.5 lays out sufficient properties the nuisance function vector estimator needs to satisfy. In particular, Assumptions D.3.5.A and D.3.5.B permit that the convergence rate of estimators can be slower than $o_P(n^{-1/2})$, which is useful when X is high-dimensional and machine learning models are required to estimate the nuisance functions. To date, a variety of commonly-used machine learning models have been shown to enjoy a convergence rate of at least $o_P(n^{-1/4})$, e.g. [38, 25, 24] for certain ℓ_1 penalized models, [281] for a class of regression trees and random forests, and [47] for a class of neural nets. This implies when these models are applied to the estimation of $(g_{10}, g_{00}, e_{10}, p_0)^\top$, Assumptions D.3.5.A and D.3.5.B hold, so our transported GATE estimator is asymptotically normal and Assumption E.7.1 is satisfied.

Constructing Confidence Intervals for the Extrapolated Effects In our experimental setup, the data generating distribution for all observational studies is identical, so no transportation of effects is required, which enables the application of existing results. We use the doubly-robust score [222, 221] as the signal for the conditional average treatment effect,

$$Y(\eta) = \mu(1, X) - \mu(0, X) + \frac{A(Y - \mu(1, X))}{s(X)} - \frac{(1 - A)(Y - \mu(0, X))}{1 - s(X)}, \tag{D.11}$$

where $\eta \coloneqq (\mu, s)$, and $\mu(A, X) \coloneqq \mathbb{E}[Y|A, X]$, and $s(X) \coloneqq \mathbb{P}(A = 1 \mid X)$. We use a multi-layer perceptron (MLP) regressor as a plug-in estimate $\hat{\mu}$ of μ, and ℓ_2-regularized logistic regression as a plug-in estimate $\hat{s}$ of s, with hyperparameters described in Section D.6.

Following example 2.2 from Semenova and Chernozhukov [241], we approximate the conditional treatment effect with a linear combination of subgroup dummy variables $G = (G_0, G_1, G_2, G_3)^\top$, so the combination weights correspond to the GATEs $\tau(k) = (\tau(k)_0, \tau(k)_1, \tau(k)_2, \tau(k)_3)$. This amounts to regressing the estimated signal $\hat{Y}_i(\hat{\eta})$ with G. As long as the propensity score is bounded above and below away from 0 and 1 (Assumption 4.10(a) of Semenova and Chernozhukov [241]), and the convergence rates of the response surface and propensity score estimates are sufficiently fast (Assumption 4.11), Corollary 4.1 and a set of mild technical conditions justify Theorem 3.1 in Semenova and Chernozhukov [241], which gives a result on pointwise asymptotic normality for the regression coeffcients $\hat{\tau}(k) = (\hat{\tau}(k)_0, \hat{\tau}(k)_1, \hat{\tau}(k)_2, \hat{\tau}(k)_3) \in \mathbb{R}^4$, so that for any unit vector $\gamma \in \mathbb{R}^4$ where $\|\gamma\| = 1$,

$$\lim_{N_k \to \infty} \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{N_k}\gamma^\top(\hat{\tau}(k) - \tau(k))}{\sqrt{\gamma^\top \Omega \gamma}} < t \right) - \Phi(t) \right| = 0$$

where Ω can be consistently estimated with Equation 2.5 in Semenova and Chernozhukov [241]

$$\hat{\Omega} = \left(\frac{1}{N_k} \sum_j G_j G_j^\top \right)^{-1} \left(\frac{1}{N_k} \sum_j G_j G_j^\top (\hat{Y}_j(\hat{\eta}) - G_j^\top \hat{\tau}(k))^2 \right) \left(\frac{1}{N_k} \sum_j G_j G_j^\top \right)^{-1}$$

Setting γ as 1 in the $(i+1)$th element and 0 elsewhere thus yields

$$\lim_{N_k \to \infty} \sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{N_k}(\hat{\tau}_i(k) - \tau_i(k))}{\sqrt{\Omega_{ii}}} < t \right) - \Phi(t) \right| = 0$$

We therefore estimate $\hat{\sigma}_i^2(k)$, the variance of $\hat{\tau}_i(k)$, with $\hat{\Omega}_{ii}$, and as this converges in probability to Ω_{ii}, the asymptotic normality of the above follows via Slutsky's theorem.

D.4 Proofs

D.4.1 Proofs for propositions and theorems

Proposition 6.2.1. *For an observational estimator $\hat{\tau}(k)$, assume Assumptions 6.2.2 and E.7.1 hold. Furthermore, let $N = N_k + N_0$ with fixed proportions, where $N_k = \rho N, N_0 = (1-\rho)N$ for $\rho \in (0,1)$. Define the test statistic*

$$\hat{T}_N(k,i) := \frac{\hat{\tau}_i(k) - \hat{\tau}_i(0) - \mu_i(k)}{\hat{s}} \tag{6.3}$$

where $\hat{s}^2 := \frac{\hat{\sigma}_i^2(k)}{N_k} + \frac{\hat{\sigma}_i^2(0)}{N_0}$ is the estimated variance, and $\mu_i(k) := \tau_i(k) - \tau_i$. This test statistic converges in distribution to a normal distribution as $N \to \infty$, $\hat{T}_N(k,i) \xrightarrow{d} \mathcal{N}(0,1)$.

Proof. As $N \to \infty$, we have it that

$$\sqrt{\rho N}(\hat{\tau}_i(k) - \tau_i(k)) \xrightarrow{d} \mathcal{N}(0, \sigma_i^2(k))$$
$$\sqrt{(1-\rho)N}(\hat{\tau}_i(0) - \tau_i) \xrightarrow{d} \mathcal{N}(0, \sigma_i^2(0))$$

where we have written ρN in place of N_k, and similarly for N_0. By Slutsky's theorem, we can multiply by the constants $\rho^{-1/2}$ and $(1-\rho)^{-1/2}$ to get both results in terms of $\sqrt{N}$. We can then use independence of $\hat{\tau}(k), \hat{\tau}(0)$ to write that

$$\underbrace{\sqrt{N}\begin{pmatrix} \hat{\tau}_i(k) - \tau_i(k) \\ \hat{\tau}_i(0) - \tau_i \end{pmatrix}}_{Z-\theta} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_i^2(k)/\rho & 0 \\ 0 & \sigma_i^2(0)/(1-\rho) \end{bmatrix} \right).$$

We now apply the Delta method. Let $Z = (\hat{\tau}_i(k), \hat{\tau}_i(0))$ denote the (column) vector of estimates, and similarly let $\theta = (\tau_i(k), \tau_i)$. Letting $f(X) = X_1 - X_2$, we can argue that

$$\sqrt{N}(Z - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma) \implies \sqrt{N}(f(Z) - f(\theta)) \xrightarrow{d} \mathcal{N}\left(0, \nabla f(\theta)^\top \Sigma \; \nabla f(\theta)\right),$$

where the resulting variance is given by

$$\nabla f(\theta)^\top \Sigma \; \nabla f(\theta) = \frac{\sigma_i^2(k)}{\rho} + \frac{\sigma_i^2(0)}{1-\rho},$$

and $f(Z) - f(\theta) = \tau_i(k) - \tau_i - \mu_i(k)$.

$$\sqrt{N}(\tau_i(k) - \tau_i - \mu_i(k)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_i^2(k)}{\rho} + \frac{\sigma_i^2(0)}{1-\rho}\right),$$

and accordingly that

$$\frac{\tau(k)_i - \tau_i - \mu_i(k)}{\sqrt{\frac{\sigma_i^2(k)}{N_k} + \frac{\sigma_i^2(0)}{N_0}}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where this also holds (by Slutsky's theorem) with $\sigma^2(k)_i$ and $\sigma_i^2(0)$ replaced by their empirical estimates, which converge in probability. □

Theorem 6.3.1 (Properties of Algorithm 2)**.** *Under Assumptions 7.2.1 and 6.2.2, the output of Algorithm 2 has the following asymptotic properties as $N \to \infty$, where N denotes the total sample size, and the samples used for all estimators are of the same order $N_k = \rho_k N_0, \forall k \geq 1$, for some $\rho_k > 0$.*

1. Under Assumptions 6.2.3 and E.7.1, for each $i \in \mathcal{I}_O$,

$$\lim_{N\to\infty} \mathbb{P}(\tau_i \in [\hat{L}_i, \hat{U}_i]) \geq 1 - \alpha \tag{6.6}$$

2. Under Assumption E.7.1, for each estimator where $\tau_i(k) \neq \tau_i$ for some $i \in \mathcal{I}_R$,

$$\lim_{N\to\infty} \mathbb{P}(k \in \hat{\mathcal{C}}) = 0 \tag{6.7}$$

Proof. **(1)** By asymptotic normality and consistency of each dimension of $\tau(k)$, the test statistic $\hat{T}_N(k, i)$ converges in distribution to $\mathcal{N}(0, 1)$. As a result, for each $i \in \mathcal{I}_R$, the probability that $\left|\hat{T}_N(k, i)\right| > z_{\alpha/(4|\mathcal{I}_R|)}$ converges to $\alpha/(2\,|\mathcal{I}_R|)$. By an application of the union bound, the probability that this occurs for any $i \in \mathcal{I}_R$ is bounded by $\alpha/2$. Similarly, by the assumed properties of $\tau(k)$, the probability that the confidence interval $[\hat{L}_i(k)(\alpha/2), \hat{U}_i(k)(\alpha/2)]$ fails to capture the true value of τ_i converges to $\alpha/2$. By another application of the union bound, for each $i \in \mathcal{I}_O$, the probability that either $\tau(k)$ is not selected or τ_i is not contained in the interval is upper bounded by α. The result follows.

(2) By asymptotic normality of each $\tau(k)$, the power calculation in Equation (6.4) holds, and as $N \to \infty$, the probability of rejecting the null hypothesis converges to zero as $\sigma^2_{k,0}$ becomes arbitrarily large, which occurs as both $N_k, N_0 \to \infty$. □

Proposition D.3.1. *In the setting of Section D.2, under Assumptions D.2.1 and D.2.2, the conditional mean potential outcome in the target distribution is identified as*

$$\mathbb{E}[Y_a \mid S = 0, G = i] = \frac{\mathbb{P}(S = 0)}{\mathbb{P}(S = 0 \mid G = i)} \mathbb{E}[Y^a(\eta_0, \pi_0) \mid G = i], \tag{D.7}$$

where $Y^a(\eta, \pi)$ is defined as in Equation D.8.

$$Y^a(\eta, \pi) := \frac{1}{\pi}\left(\mathbf{1}\left\{S=1, A=a\right\} \cdot \frac{1-p(X)}{p(X)e_a(X)} \cdot \{Y - g_a(X)\} + (1-S)g_a(X)\right) \tag{D.8}$$

where $\eta := (g_a, e_a, p)$ with true underlying parameters $\eta_0 = (g_{a0}, e_{a0}, p_0)$, $\pi := \mathbb{P}(S=0)$ with true value π_0, $g_a(X) := \mathbb{E}[Y \mid A=a, S=1, X]$, $p(X) := \mathbb{P}(S=1 \mid X)$, and $e_a(X) := \mathbb{P}(A=a \mid S=1, X]$.

Proof. First, we can observe by standard arguments that the conditional expectation of $Y^a(\eta_0, \pi_0)$ given X is given by the following

$$\mathbb{E}[Y^a(\eta_0, \pi_0) \mid X=x] = \mathbb{E}\left[\frac{1-S}{\mathbb{P}(S=0)} g_{a0}(X) \middle| X=x\right],$$

because the first term in Equation (D.8) is mean-zero conditioned on $X=x$. This follows by the law of total expectation: for any event where $S=1, A=a$ does not hold, the first term is zero due to the indicator, and for any other event $S=1, A=a, X=x$, the first term is mean-zero, since the first term becomes a constant (determined by $S=1, A=a, X=x$) times a mean-zero random variable $Y - \mathbb{E}[Y \mid A=a, S=1, X=x]$.

As a result, we can write that

$$\begin{aligned}
&\mathbb{E}[Y^a(\eta_0, \pi_0) \mid G=i] \\
&= \mathbb{E}\left[\frac{1-S}{\mathbb{P}(S=0)}\mathbb{E}[Y \mid A=a, S=1, X] \middle| G=i\right] \\
&= \frac{1}{\mathbb{P}(S=0)}\int_x \sum_s \mathbf{1}\left\{s=0\right\} \mathbb{E}[Y \mid A=a, S=1, x] p(s, x \mid G=i) dx \\
&= \frac{1}{\mathbb{P}(S=0)}\int_x \sum_s \mathbf{1}\left\{s=0\right\} \mathbb{E}[Y_a \mid S=1, x] p(s, x \mid G=i) dx \\
&\qquad \text{Assumption D.2.1}, Y_a \perp\!\!\!\perp A \mid X, S=1 \\
&= \frac{1}{\mathbb{P}(S=0)}\int_x \sum_s \mathbf{1}\left\{s=0\right\} \mathbb{E}[Y_a \mid S=0, x] p(s, x \mid G=i) dx \\
&\qquad \text{Assumption D.2.2}, Y_a \perp\!\!\!\perp S \mid X
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0] p(S=0, x \mid G=i) dx \\
&= \frac{1}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0] p(x \mid S=0, G=i) \mathbb{P}(S=0 \mid G=i) dx \\
&= \frac{\mathbb{P}(S=0 \mid G=i)}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0] p(x \mid S=0, G=i) dx \\
&= \frac{\mathbb{P}(S=0 \mid G=i)}{\mathbb{P}(S=0)} \int_x \mathbb{E}[Y_a \mid x, S=0, G=i] p(x \mid S=0, G=i) dx
\end{aligned}
$$

$$X = x \Rightarrow G = i, \forall x : p(x \mid G=i) > 0$$

$$= \frac{\mathbb{P}(S=0 \mid G=i)}{\mathbb{P}(S=0)} \mathbb{E}[Y_a \mid S=0, G=i]$$

and the result follows from dividing both sides by the first term on the right-hand side, which we can observe is equivalent to multiplying both sides by

$$\frac{\mathbb{P}(S=0)}{\mathbb{P}(S=0 \mid G=i)} = \frac{\mathbb{P}(S=0)\mathbb{P}(G=i)}{\mathbb{P}(S=0, G=i)} = \frac{\mathbb{P}(G=i)}{\mathbb{P}(G=i \mid S=0)} \tag{D.12}$$

□

D.4.2 Asymptotic normality of cross-fitted transported GATE estimators

Theorem D.3.1. *Suppose Assumptions D.3.1 to D.3.5 hold. Then, the empirical average, $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$, where $\tilde{Y}$ is defined in Equation D.9 and $\hat{\eta}$ is estimated with cross-fitting, is asymptotically normal.*

Proof sketch: Our strategy for the proof consists of two stages. First, we show that if the nuisance function is known to be η_0 and plugged into the estimator as $\mathbb{E}[\tilde{Y}(\eta_0, \hat{\pi}_g)|G = i]$, the resulting estimator, which we later refer to as the oracle estimator, is asymptotically normal. Second, we show that even if the true nuisance function is not known, as long as we have an estimator, $\hat{\eta}$, of the nuisance function that follows certain properties, the resulting estimator $\mathbb{E}[\tilde{Y}(\hat{\eta}, \hat{\pi}_g)|G = i]$ converges to the oracle estimator in probability. Then, by Slutsky's Theorem, the resulting estimator is also asymptotically normal.

Before diving into the first stage of the proof, we introduce additional notation to reflect the cross-fitting nature of our GATE estimator. Let the combined sample size of the observational study and RCT be N with sample indices $[N] := \{1, 2, ..., N\}$. We denote $(I_m)_{m=1}^M$ as a M-fold random partition of $[N]$, so that each fold has size $N_M = N/M$. The plug-in nuisance function estimate for the m^{th} fold, $\hat{\eta}_m$, is then estimated from the rest of the folds $I_m^c := [N] \backslash I_m$. For brevity, we denote the size of the rest of the folds as $N_M^c = N - N/M$.

We now restate the definition of the treatment effect signal $\tilde{Y}_j(\eta, \pi_g) = \tilde{Y}_j((g_1, g_0, e_1, p)^\top, \pi_g)$:

$$\begin{aligned}
\tilde{Y}_j(\eta, \pi_g) &:= \tilde{Y}_j^1(\eta, \pi_g) - \tilde{Y}_j^0(\eta, \pi_g) \\
\tilde{Y}_j^a(\eta, \pi_g) &:= \frac{1}{\pi_g(G_i)} \left(\mathbf{1}\{S_j = 1, A_j = a\} \cdot \frac{1 - p(X_j)}{p(X_j) e_a(X_j)} \cdot \{Y_j - g_a(X_j)\} + (1 - S_j) g_a(X_j) \right)
\end{aligned}$$

In the remainder of the development, we will drop the subscript j, which represents one of the N samples, for conciseness.

Stage 1 — *Proving the asymptotic normality of the oracle estimator*

For brevity, we define the following unweighted signal:

Definition D.4.1 (Unweighted signal functional)**.**

$$\begin{aligned}
\mathcal{Y}(\eta) &= \pi_g(G) \tilde{Y}(\eta, \pi_g) \\
&= \pi_g(G)(\tilde{Y}^1(\eta, \pi_g) - \tilde{Y}^0(\eta, \pi_g)) \\
&= (1 - S)(g_1(X) - g_0(X)) + S \frac{1 - p(X)}{p(X)} \frac{(A - e_1(X))(Y - g_A(X))}{e_1(X) e_0(X)}
\end{aligned}$$

From the proof of Proposition D.3.1, we have the following identities for the unweighted signals:

Lemma D.4.1 (Conditional mean of unweighted (oracle) signal)**.** *The conditional mean of the unweighted (oracle) signal is equivalent to the following:*

$$\begin{aligned}
\mathbb{E}[\mathcal{Y}(\eta_0) | G = i] &= \tau_i \pi_g(i) \\
\mathbb{E}[\mathcal{Y}(\eta_0) | G = i, S = 0] &= \tau_i
\end{aligned}$$

.

Proof. First, we have,

$$\begin{aligned}\mathbb{E}[\mathcal{Y}(\eta_0)|G=i] &= \mathbb{E}[\pi_g(G)\tilde{Y}(\eta_0,\pi_g)|G=i] \\ &= \mathbb{E}[\pi_g(i)\tilde{Y}(\eta_0,\pi_g)|G=i] \\ &= \pi_g(i)\mathbb{E}[\tilde{Y}(\eta_0,\pi_g)|G=i] \\ &= \tau_i\pi_g(i)\end{aligned}$$

Next, using Definition D.1 of the unweighted signal functional and the fact that we condition on $S=0$, we have,

$$\mathbb{E}[\mathcal{Y}(\eta_0)|G=i,S=0] = \mathbb{E}\left[g_{10}(X)-g_{00}(X)|G=i,S=0\right],$$

which is τ_i as desired. □

In addition, we can rewrite our estimator $\hat{\tau}_i := \mathbb{E}[\tilde{Y}(\hat{\eta},\hat{\pi}_g)|G=i]$ with the unweighted signals:

$$\begin{aligned}\hat{\tau}_i = \mathbb{E}[\tilde{Y}(\hat{\eta},\hat{\pi}_g)|G=i] &= \frac{\sum_m\sum_{j\in I_m}\mathbf{1}(G_j=i)\tilde{Y}(\hat{\eta}_m,\hat{\pi}_g)}{\sum_j\mathbf{1}(G_j=i)} \\ &= \frac{\sum_m\sum_{j\in I_m}\mathbf{1}(G_j=i)\frac{1}{\hat{\pi}_g(G_j)}\mathcal{Y}(\hat{\eta}_m)}{\sum_j\mathbf{1}(G_j=i)}, \quad \text{from } \textbf{Def. D.1} \\ &= \frac{\sum_m\sum_{j\in I_m}\mathbf{1}(G_j=i)\frac{1}{\hat{\pi}_g(i)}\mathcal{Y}(\hat{\eta}_m)}{\sum_j\mathbf{1}(G_j=i)} \\ &= \frac{1}{\hat{\pi}_g(i)}\frac{\sum_m\sum_{j\in I_m}\mathbf{1}(G_j=i)\mathcal{Y}(\hat{\eta}_m)}{\sum_j\mathbf{1}(G_j=i)} \\ &= \frac{\sum_m\sum_{j\in I_m}\mathbf{1}(G_j=i)\mathcal{Y}(\hat{\eta}_m)}{\frac{\sum_j\mathbf{1}(G_j=1,S_j=0)}{\sum_j\mathbf{1}(G_j=i)}\sum_j\mathbf{1}(G_j=i)} \\ &= \frac{\sum_m\sum_{j\in I_m}\mathbf{1}(G_j=i)\mathcal{Y}(\hat{\eta}_m)}{\sum_j\mathbf{1}(G_j=1,S_j=0)}\end{aligned}$$

Now, using the above expression, we can define the oracle estimator, where we *know* the true value of η, which is η_0:

Definition D.4.2 (Oracle GATE Estimator)**.**

$$\hat{\tau}_{i0} := \frac{\sum_j \mathbf{1}(G_j = i)\mathcal{Y}_j(\eta_0)}{\sum_j \mathbf{1}(G_j = i, S_j = 0)}$$

To show the asymptotic distribution of the oracle GATE estimator, we restate several assumptions:

Assumption D.3.1 (Observational dataset covers the whole support of covariates)**.**

$$\inf_{x \in \mathcal{X}} p_0(x) = \varepsilon_p > 0$$

Assumption D.3.2 (Bounded within-subgroup variance of conditional treatment effects in the RCT)**.**

$$\sup_{x \in \mathcal{X}} var[g_{10}(x) - g_{00}(x)|G = i, S = 0] = \sigma_{\tau i}^2 < \infty$$

Assumption D.3.3 (Overlap between treatments in the observational dataset)**.**

$$\inf_{x \in \mathcal{X}} \min(e_{00}(x), e_{10}(x)) = \varepsilon_e > 0$$

Assumption D.3.4 (Finite outcome conditional variance in the observational dataset)**.**

$$\max_{a \in \{0,1\}} \sup_{x \in \mathcal{X}} \mathbb{E}[(Y - g_{a0}(x))^2 | X = x, S = 1, A = a] = \bar{\sigma}^2 < \infty$$

These assumptions ensure that the oracle signals have finite conditional variance, which we prove in the following lemma.

Lemma D.4.2 (Finite conditional variance of unweighted oracle signal)**.** *Under Assumptions D.3.1 - D.3.4, we have that,*

$$var[\mathcal{Y}(\eta_0)|G = i] := \sigma_i^2 < \infty, \forall i \in [d]$$

Proof.

$$
\begin{aligned}
& var[\mathcal{Y}(\eta_0)|G=i] \\
=& \mathbb{E}[\mathcal{Y}^2(\eta_0)|G=i] - [\mathbb{E}[\mathcal{Y}(\eta_0)|G=i]]^2 \\
=& \mathbb{E}\Bigg[\Bigg((1-S)(g_{10}(X)-g_{00}(X)) + \\
& \qquad S\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\Bigg)^2\Bigg|G=i\Bigg] - \pi_g(i)^2\tau_i^2 \\
& \qquad\qquad \text{Lem. } D.4.1 \\
=& \Bigg\{\mathbb{E}\left[(1-S)(g_{10}(X)-g_{00}(X))^2\middle|G=i\right] + \\
& \qquad \mathbb{E}\left[S\left(\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\right)^2\middle|G=i\right]\Bigg\} - \pi_g(i)^2\tau_i^2 \\
& \qquad\qquad \text{since } S\in\{0,1\} \\
=& \Bigg\{\mathbb{E}\left[(g_{10}(X)-g_{00}(X))^2\middle|G=i,S=0\right]\pi_g(i) + \\
& \qquad \mathbb{E}\left[\left(\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\right)^2\middle|G=i,S=1\right](1-\pi_g(i))\Bigg\} - \pi_g(i)^2\tau_i^2 \\
=& \Bigg\{\left[var\left[g_{10}(X)-g_{00}(X)|G=i,S=0\right]+\tau_i^2\right]\pi_g(i) + \\
& \qquad \mathbb{E}\left[\left(\frac{1-p_0(X)}{p_0(X)}\frac{(A-e_{10}(X))(Y-g_{A0}(X))}{e_{10}(X)e_{00}(X)}\right)^2\middle|G=i,S=1\right](1-\pi_g(i))\Bigg\} - \pi_g(i)^2\tau_i^2 \\
<& \left\{\left[\sigma_{\tau i}^2+\tau_i^2\right]\pi_g(i) + \frac{\mathbb{E}[(Y-g_{A0}(X))^2|G=i,S=1]}{\varepsilon_\pi^2\varepsilon_e^2(1-\varepsilon_e)^2}(1-\pi_g(i))\right\} - \pi_g(i)^2\tau_i^2 \\
& \qquad\qquad \text{Asmp. } D.3.1 - D.3.3 \\
\leq& \left\{\left[\sigma_{\tau i}^2+\tau_i^2\right]\pi_g(i) + \frac{\bar{\sigma}^2}{\varepsilon_\pi^2\varepsilon_e^2(1-\varepsilon_e)^2}(1-\pi_g(i))\right\} - \pi_g(i)^2\tau_i^2 < \infty \\
& \qquad\qquad \text{Asmp. } D.3.4
\end{aligned}
$$

□

Now, using the above lemmas, we are ready to prove the main result of stage 1 of the proof, stated below.

Proposition D.4.1 (Asymptotic normality of oracle GATE estimator)**.** *Under Assumptions D.3.1 - D.3.4,*

$$\sqrt{N}(\hat{\tau}_{i0}-\tau_i) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_i^2-\pi_g(i)(1-\pi_g(i))\tau_i^2}{\pi_g(i)^2\mathbb{P}(G=i)}\right)$$

Proof. We have that,

$$\begin{aligned}
&\sqrt{N}(\hat{\tau}_{i0}-\tau_i)\\
=&\sqrt{N}\left(\frac{\sum_j \mathbf{1}(G_j=i)\mathcal{Y}_j(\eta_0)}{\sum_j \mathbf{1}(G_j=i,S_j=0)}-\tau_i\right)\\
=&\sqrt{N}\left(\frac{\sum_j \mathbf{1}(G_j=i)\mathcal{Y}_j(\eta_0)}{\sum_j \mathbf{1}(G_j=i,S_j=0)}-\frac{\sum_j \mathbf{1}(G_j=i,S_j=0)\tau_i}{\sum_j \mathbf{1}(G_j=i,S_j=0)}\right)\\
=&\sqrt{N}\frac{\sum_j \mathbf{1}(G_j=i)(\mathcal{Y}_j(\eta_0)-\tau_i\mathbf{1}(S_j=0))}{\sum_j \mathbf{1}(G_j=i,S_j=0)}\\
=&\frac{\sqrt{N}\frac{1}{N}\sum_j \mathbf{1}(G_j=i)(\mathcal{Y}_j(\eta_0)-\tau_i\mathbf{1}(S_j=0))}{\frac{1}{N}\sum_j \mathbf{1}(G_j=i,S_j=0)} \quad \begin{array}{l}\xrightarrow{d} \mathcal{N}(0,(\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i)))\mathbb{P}(G=i))\\ \xrightarrow{p} \mathbb{P}(G=i,S=0)=\mathbb{P}(G=i)\pi_g(i)\end{array}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \begin{array}{c}\text{Proven below}\\ \text{From WLLN}\end{array}\\
\xrightarrow{d}&\mathcal{N}\left(0,\frac{\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i))}{\pi_g(i)^2\mathbb{P}(G=i)}\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{Slutsky's lemma}
\end{aligned}$$

In the above, we used the fact that,

$$\sqrt{N}\left[\frac{1}{N}\sum_j \mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))\right] \xrightarrow{d} \mathcal{N}(0,(\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i)))\mathbb{P}(G=i))$$

To show this, we observe that

$$\begin{aligned}
&\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))]\\
=&\mathbb{E}[\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0)|G=i]\mathbb{P}(G=i)\\
=&(\mathbb{E}[\mathcal{Y}(\eta_0)|G=i]-\tau_i\mathbb{E}[\mathbf{1}(S=0)|G=i])\mathbb{P}(G=i)\\
=&(\mathbb{E}[\mathcal{Y}(\eta_0)|G=i]-\tau_i\pi_g(i))\mathbb{P}(G=i)=0\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{Lem. } D.4.1\\
&var[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))]\\
=&\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))]^2-(\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))])^2\\
=&\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))]^2\\
=&\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))^2]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{since } \mathbf{1}(G=i)\in\{0,1\}\\
=&\mathbb{E}[(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))^2|G=i]\mathbb{P}(G=i)\\
=&\left\{\mathbb{E}[\mathcal{Y}^2(\eta_0)|G=i]+\tau_i^2\mathbb{E}[\mathbf{1}(S=0)|G=i]-2\tau_i\mathbb{E}[\mathcal{Y}(\eta_0)\mathbf{1}(S=0)|G=i]\right\}\mathbb{P}(G=i)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{since } \mathbf{1}(S=0)\in\{0,1\}\\
=&\left\{var[\mathcal{Y}(\eta_0)|G=i]+(\mathbb{E}[\mathcal{Y}(\eta_0)|G=i])^2+\tau_i^2\pi_g(i)\right.\\
&\qquad\left.-2\tau_i\pi_g(i)\mathbb{E}[\mathcal{Y}(\eta_0)|G=i,S=0]\right\}\mathbb{P}(G=i)\\
=&\left\{\sigma_i^2+\tau_i^2\pi_g(i)^2+\tau_i^2\pi_g(i)-2\tau_i^2\pi_g(i)\right\}\mathbb{P}(G=i)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\text{Asmp. } D.4.2\text{, Lem. } D.4.1\\
=&(\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i)))\mathbb{P}(G=i)<\infty
\end{aligned}$$

Therefore, from central limit theorem,

$$\sqrt{N}\left[\frac{1}{N}\sum_j\mathbf{1}(G=i)(\mathcal{Y}(\eta_0)-\tau_i\mathbf{1}(S=0))\right]\xrightarrow{d}\mathcal{N}(0,(\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i)))\mathbb{P}(G=i))$$

□

Stage 2 — *Proving the asymptotic normality of the cross-fitted estimator,* $\tilde{Y}(\hat{\eta},\hat{\pi}_g)$

With asymptotic normality of the oracle estimator shown above in Stage 1, we can show the asymptotic normality of the cross-fitted estimator (i.e. our estimator) by decomposing its error into the error of the oracle estimator and the difference between our estimator and the oracle estimator:

$$\begin{aligned}\sqrt{N}(\hat{\tau}_i - \tau_i) &= \sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \sqrt{N}(\hat{\tau}_i - \hat{\tau}_{i0}) \\ &= \sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \sqrt{N}\frac{\sum_m \sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))}{\sum_j \mathbf{1}(G_j = i, S_j = 0)} \\ &= \sqrt{N}(\hat{\tau}_{i0} - \tau_i) + \frac{\sum_m \frac{1}{\sqrt{N}} \sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))}{\frac{1}{N}\sum_j \mathbf{1}(G_j = i, S_j = 0)}\end{aligned}$$

The asymptotic distribution of the cross-fitted estimator therefore hinges on the asymptotic property of $\frac{1}{\sqrt{N}} \sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))$, which in turn depends on the convergence property of the nuisance function estimate $\hat{\eta}_m$ and its influence on the signal $\mathcal{Y}$. We therefore restate the last required assumption governing the convergence properties of $\hat{\eta}_m$:

Assumption D.3.5 (Properties of the nuisance function estimators)**.** Let $\hat{\eta}_{(n)}$ be a sequence of estimators for η indexed by the size of the cross-fitting training fold n. We assume that there exists

- $\epsilon_n = o_P(1)$, a sequence of positive numbers
- $\mathcal{T}_n$, a sequence of nuisance function vector sets in the neighborhood of $\eta_0 = (g_{10}, g_{00}, e_{10}, p_0)$ satisfying $\mathbb{P}(\hat{\eta}_{(n)} \in \mathcal{T}_n) \geq 1 - \epsilon_n$
- $\mathbf{g}_n, \mathbf{e}_n, \mathbf{p}_n$, sequences of worst root mean square errors for the nuisance functions g_1, g_0, e_1, p, defined as follows:

$$\begin{aligned}\mathbf{g}_n &\coloneqq \max_{a \in \{0,1\}} \sup_{\eta \in \mathcal{T}_n} \sqrt{\mathbb{E}[g_a(X) - g_{a0}(X)]^2} \\ \mathbf{e}_n &\coloneqq \sup_{\eta \in \mathcal{T}_n} \sqrt{\mathbb{E}[e_1(X) - e_{10}(X)]^2} \\ \mathbf{p}_n &\coloneqq \sup_{\eta \in \mathcal{T}_n} \sqrt{\mathbb{E}[p(X) - p_0(X)]^2}\end{aligned}$$

so that the following assumptions hold:

Assumption A: (Rate of nuisance error)

$$\mathbf{g}_n \vee \mathbf{e}_n \vee \mathbf{p}_n = o_P(1)$$

Assumption B: (Rate of nuisance error product)

$$\sqrt{n}\mathbf{g}_n(\mathbf{e}_n \vee \mathbf{p}_n) = o_P(1)$$

Assumption C: (Bounded nuisance estimates)

$$\sup_{\eta \in \cup_{n=1}^{\infty} \mathcal{T}_n} \left(\max_{a\in\{0,1\}} \sup_{x\in\mathcal{X}} |g_a(x)| \vee \sup_{x\in\mathcal{X}} \left| \frac{1}{p(x)} \right| \vee \sup_{x\in\mathcal{X}} \left| \frac{1}{e_1(x)} \right| \vee \sup_{x\in\mathcal{X}} \left| \frac{1}{e_0(x)} \right| \right) = \bar{\mathcal{C}} < \infty$$

Based on the assumptions above, we have the following bounds on the convergence rate of the signals when the nuisance function estimates are in the high-probability neighborhood, $\mathcal{T}_n$:

Lemma D.4.3 (Bounds on bias of signal)**.** *Under Assumptions D.3.5.B and D.3.5.C*

$$\sqrt{n} \sup_{\eta\in\mathcal{T}_n} |\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))]| = o_P(1)$$

Lemma D.4.4 (Bounds on MSE of signal)**.** *Under Assumptions D.3.1, D.3.3, D.3.4, D.3.5.A and D.3.5.C*

$$\sup_{\eta\in\mathcal{T}_n} \mathbb{E} \left|\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))\right|^2 = o_P(1)$$

which in turn implies that $\frac{1}{\sqrt{N}} \sum_{j\in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0))$ converges to zero in probability:

Lemma D.4.5 (Numerator of difference is $o_P(1)$)**.** *Under Assumptions D.3.1, D.3.3,* *D.3.4 and D.3.5,*

$$\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))=o_P(1),\quad \forall m\in\{1,2,...,M\}$$

The proofs for Lemmas D.4.3 to D.4.5 are more labor-intensive and we defer these proofs to later subsections. Based on these lemmas, we arrive at the main result of Stage 2.

Theorem D.4.1 (Asymptotic normality of the cross-fitted transported GATE estimator)**.** *Under Assumptions D.3.1 - D.3.5,*

$$\sqrt{N}(\hat{\tau}_i-\tau_i)\xrightarrow{d}\mathcal{N}\left(0,\frac{\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i))}{\pi_g(i)^2\mathbb{P}(G=i)}\right)$$

Proof.

$$\begin{aligned}
&\sqrt{N}(\hat{\tau}_i-\tau_i)\\
=&\sqrt{N}(\hat{\tau}_{i0}-\tau_i)+\sqrt{N}(\hat{\tau}_i-\hat{\tau}_{i0})\\
=&\sqrt{N}(\hat{\tau}_{i0}-\tau_i)+\sqrt{N}\frac{\sum_m\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))}{\sum_j\mathbf{1}(G_j=i,S_j=0)}\\
=&\sqrt{N}(\hat{\tau}_{i0}-\tau_i)+\frac{\sum_m\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))\xrightarrow{p}0}{\frac{1}{N}\sum_j\mathbf{1}(G_j=i,S_j=0)\xrightarrow{p}\mathbb{P}(G=i,S=0)}\\
&\qquad\qquad\qquad\qquad\text{Lem. }D.4.5\\
&\qquad\qquad\qquad\qquad\quad\text{WLLN}\\
\xrightarrow{d}&\mathcal{N}\left(0,\frac{\sigma_i^2-\tau_i^2\pi_g(i)(1-\pi_g(i))}{\pi_g(i)^2\mathbb{P}(G=i)}\right)\\
&\qquad\qquad\qquad\qquad\text{Prop. }D.4.1\text{, Slutsky's lemma}
\end{aligned}$$

□

Note that Theorem D.4.1 is simply Theorem D.3.1, which is the primary result of this section, with the variance explicitly stated. Thus, Theorem D.3.1 is proven.

D.4.3 Proof for Lemmas D.4.3 and D.4.4

First, we prove Lemmas D.4.3 and D.4.4, which will be necessary for Lemma D.4.5. Recall that Lemma D.4.5 was essential for the proof of the asymptotic normality result in Theorem D.4.1.

Lemma D.4.3 (Bounds on bias of signal). *Under Assumptions D.3.5.B and D.3.5.C*

$$\sqrt{n} \sup_{\eta \in \mathcal{T}_n} |\mathbb{E}[\mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))]| = o_P(1)$$

Lemma D.4.4 (Bounds on MSE of signal). *Under Assumptions D.3.1, D.3.3, D.3.4, D.3.5.A and D.3.5.C*

$$\sup_{\eta \in \mathcal{T}_n} \mathbb{E} \left| \mathbf{1}(G = i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0)) \right|^2 = o_P(1)$$

Proof. We first define partial unweighted signal functionals for the two counterfactual outcomes

Definition D.4.3 (Partial unweighted signal functionals).

$$\begin{aligned}
\mathcal{Y}^1(\eta) &:= \left[(1 - S)g_1(X) + S \frac{1 - p(X)}{p(X)} \frac{A(Y - g_1(X))}{e_1(X)} \right] \\
\mathcal{Y}^0(\eta) &:= \left[(1 - S)g_0(X) + S \frac{1 - p(X)}{p(X)} \frac{(1 - A)(Y - g_0(X))}{e_0(X)} \right] \\
\Rightarrow \mathcal{Y}(\eta) &= \mathcal{Y}^1(\eta) - \mathcal{Y}^0(\eta)
\end{aligned}$$

At a high level, we will prove the above lemmas by decomposing the errors of signal functionals into simpler terms that can be bounded by standard concentration inequalities. This idea will be repeated for both the bias and MSE of the signals. To simplify the analysis, we can split up the unweighted signal into “partial signals” (for the treatment and control groups). Therefore, we set out to show the following lemmas:

Lemma D.4.6 (Bounds on bias of partial signal). *Under Assumptions D.3.5.B and*

D.3.5.C

$$\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))]\right| = o_P(1)$$
$$\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))]\right| = o_P(1)$$

Lemma D.4.7 (Bounds on MSE of partial signal)**.** *Under Assumptions D.3.1, D.3.3, D.3.4, D.3.5.A and D.3.5.C*

$$\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2 = o_P(1)$$
$$\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2 = o_P(1)$$

In the following subsections, we prove the $\mathcal{Y}_1$ part of Lemmas D.4.6 and D.4.7. The $\mathcal{Y}_0$ part will follow by symmetry. First, we further define $\eta(X) = \eta_0(X) + \delta_\eta(X)$, in detail:

$$g_1(X) = g_{10}(X) + \delta_{g_1}(X)$$
$$p(X) = p_0(X) + \delta_p(X)$$
$$e(Z) = e_0(X) + \delta_e(X)$$

so that (omitting the parameter X for brevity),

$$\begin{aligned}
&\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0) \\
=&\left[(1-S)(g_{10}+\delta_{g_1}) + \frac{1-p_0-\delta_p}{p_0+\delta_p}\frac{SA(Y-g_{10}-\delta_{g_1})}{e_0+\delta_e}\right] - \left[(1-S)g_{10} + \frac{1-p_0}{p_0}\frac{SA(Y-g_{10})}{e_0}\right] \\
=&(1-S)\delta_{g_1} + \frac{1-p_0-\delta_p}{p_0+\delta_p}\frac{SA(Y-g_{10}-\delta_{g_1})}{e_0+\delta_e} - \frac{1-p_0}{p_0}\frac{SA(Y-g_{10}-\delta_{g_1})}{e_0} - \frac{1-p_0}{p_0}\frac{SA}{e_0}\delta_{g_1} \\
=&\left((1-S) - \frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1} + \left(\frac{1-p_0-\delta_p}{(p_0+\delta_p)(e_0+\delta_e)} - \frac{1-p_0}{p_0e_0}\right)SA(Y-g_{10}-\delta_{g_1}) \\
=&\left((1-S) - \frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1} - \frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA(Y-g_{10}-\delta_{g_1}) \\
=&\underbrace{\left((1-S) - \frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1}}_{S_1} - \underbrace{\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA(Y-g_{10})}_{S_2} \\
&\qquad\qquad\qquad\qquad + \underbrace{\frac{e_0\delta_p + (1-p_0)p_0\delta_e + (1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA\delta_{g_1}}_{S_3} \\
\coloneqq& S_1 - S_2 + S_3
\end{aligned}$$

Proof for Lemma D.4.6

For Lemma D.4.6 we want to bound

$$\begin{aligned}
\left|\mathbb{E}\left[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0))\right]\right| &= \left|\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}(G=i)\left(\mathcal{Y}^1(\eta) - \mathcal{Y}^1(\eta_0)\right)\middle|X\right]\right]\right| \\
&= \left|\mathbb{E}\left[\mathbf{1}(G=i)\mathbb{E}\left[S_1 - S_2 + S_3|X\right]\right]\right| \\
&\qquad\qquad\qquad G \text{ is a function of } X \\
&= \left|\mathbb{E}\left[\mathbf{1}(G=i)\left(\mathbb{E}\left[S_1|X\right] - \mathbb{E}\left[S_2|X\right] + \mathbb{E}\left[S_3|X\right]\right)\right]\right|
\end{aligned}$$

For the term $\mathbb{E}[S_1|X]$,

$$\begin{aligned}
&\mathbb{E}[S_1|X]\\
=&\mathbb{E}\left[\left((1-S)-\frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1}\middle|X\right]\\
=&\left((1-\mathbb{E}[S|X])-\frac{1-p_0}{p_0}\frac{\mathbb{E}[SA|X]}{e_0}\right)\delta_{g_1} && \eta_0,\delta_\eta \text{ are functions of } X\\
=&\left((1-p_0)-\frac{1-p_0}{p_0}\frac{p_0e_0}{e_0}\right)\delta_{g_1}=0 && p_0(X)=\mathbb{P}[S=1|X]\\
& && e_0(X)=\mathbb{P}[A=1|S=1,X]
\end{aligned}$$

For the term $\mathbb{E}[S_2|X]$

$$\begin{aligned}
&\mathbb{E}[S_2|X]\\
=&\mathbb{E}\left[\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA(Y-g_{10})\middle|X\right]\\
=&\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\mathbb{E}[SA(Y-g_{10})\mid X] && \eta_0,\delta_\eta \text{ are functions of } X\\
=&\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\cdot 0=0 && g_{10}(X)=\mathbb{E}[Y|S=1,A=1,X]
\end{aligned}$$

For the term $\mathbb{E}[S_3|X]$,

$$\begin{aligned}
&\mathbb{E}[S_3|X]\\
=&\mathbb{E}\left[\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA\delta_{g_1}\middle|X\right]\\
=&\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\mathbb{E}\left[SA|X\right]\delta_{g_1} && \eta_0,\delta_\eta \text{ are functions of } X\\
=&\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}p_0e_0\delta_{g_1} && p_0(X)=\mathbb{P}[S=1|X]\\
& && e_0(X)=\mathbb{P}[A=1|S=1,X]\\
=&\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{pe}\delta_{g_1}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathbf{1}(G=i)[\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0)]\right]\right|^2 \\
=&\left|\mathbb{E}\left[\mathbf{1}(G=i)\left(\mathbb{E}[S_1|Z]-\mathbb{E}[S_2|Z]+\mathbb{E}[S_3|Z]\right)\right]\right|^2 \\
=&\left|\mathbb{E}\left[\mathbf{1}(G=i)\mathbb{E}[S_3|Z]\right]\right|^2 \\
\leq&\left(\mathbb{E}\left|\mathbf{1}(G=i)\mathbb{E}[S_3|Z]\right|\right)^2 && |EA|\leq E|A| \\
\leq&\left(\mathbb{E}\left|\mathbb{E}[S_3|Z]\right|\right)^2 \\
=&\left(\mathbb{E}\left|\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{pe}\delta_{g_1}\right|\right)^2 \\
\leq&\left(\mathbb{E}\left|\frac{e_0\delta_p}{pe}\delta_{g_1}\right|+\mathbb{E}\left|\frac{(1-p_0)p_0\delta_e}{pe}\delta_{g_1}\right|+\mathbb{E}\left|\frac{(1-p_0)\delta_p\delta_e}{pe}\delta_{g_1}\right|\right)^2 && \text{Triangular ineq.} \\
\leq&\bar{\mathcal{C}}^4\left(\mathbb{E}\left|\delta_p\delta_{g_1}\right|+\mathbb{E}\left|\delta_e\delta_{g_1}\right|+\mathbb{E}\left|\delta_p\delta_e\delta_{g_1}\right|\right)^2 && \text{Assmp. } D.3.5.C \\
=&\bar{\mathcal{C}}^4\left(\mathbb{E}|\delta_p\delta_{g_1}|+\mathbb{E}|\delta_e\delta_{g_1}|+\frac{1}{2}\mathbb{E}|\delta_e||\delta_p\delta_{g_1}|+\frac{1}{2}\mathbb{E}|\delta_p||\delta_e\delta_{g_1}|\right)^2 \\
\leq&\bar{\mathcal{C}}^4\left(\mathbb{E}|\delta_p\delta_{g_1}|+\mathbb{E}|\delta_e\delta_{g_1}|+\frac{1}{2}\mathbb{E}|\delta_p\delta_{g_1}|+\frac{1}{2}\mathbb{E}|\delta_e\delta_{g_1}|\right)^2 && |\delta_p|,|\delta_e|\leq 1 \\
=&\frac{9}{4}\bar{\mathcal{C}}^4\left(\mathbb{E}|\delta_p\delta_{g_1}|+\mathbb{E}|\delta_e\delta_{g_1}|\right)^2 \\
\leq&\frac{9}{4}\bar{\mathcal{C}}^4\left(\sqrt{\mathbb{E}\delta_p^2}\sqrt{\mathbb{E}\delta_{g_1}^2}+\sqrt{\mathbb{E}\delta_e^2}\sqrt{\mathbb{E}\delta_{g_1}^2}\right)^2 && \text{Hölder's ineq.}
\end{aligned}$$

So we have,

$$\begin{aligned}
&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}\mathbf{1}(G=i)[Y^1(\eta)-Y^1(\eta_0)]\right| \\
\leq&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\frac{3}{2}\bar{\mathcal{C}}^2\left(\sqrt{\mathbb{E}\delta_p^2}\sqrt{\mathbb{E}\delta_{g_1}^2}+\sqrt{\mathbb{E}\delta_e^2}\sqrt{\mathbb{E}\delta_{g_1}^2}\right) \\
\leq&\frac{3}{2}\bar{\mathcal{C}}^2\sqrt{n}\mathbf{g}_N\left(\mathbf{p}_N+\mathbf{e}_N\right) && \text{Assump. } D.3.5 \\
=&o_P(1) && \text{Assump. } D.3.5.B
\end{aligned}$$

Proof for Lemma D.4.7

Here we first place bounds on $\mathbb{E}S_1^2, \mathbb{E}S_2^2$ and $\mathbb{E}S_3^2$ for future use. For the term $\mathbb{E}S_1^2$, we have,

$$\begin{aligned}
&\mathbb{E}S_1^2 \\
=&\mathbb{E}\left[\mathbb{E}\left[\left(\left((1-S)-\frac{1-p_0}{p_0}\frac{SA}{e_0}\right)\delta_{g_1}\right)^2\middle|X\right]\right] \\
=&\mathbb{E}\left[(1-p_0)^2\,\mathbb{E}\left[\left(\frac{1-S}{1-p_0}-\frac{SA}{p_0e_0}\right)^2\middle|X\right]\delta_{g_1}^2\right] && \eta_0, \delta_\eta \text{ are functions of } X \\
=&\mathbb{E}\left[(1-p_0)^2\,\mathbb{E}\left[\frac{(1-S)^2}{(1-p_0)^2}+\frac{(SA)^2}{(p_0e_0)^2}\middle|X\right]\delta_{g_1}^2\right] && S\in\{0,1\} \\
=&\mathbb{E}\left[(1-p_0)^2\,\mathbb{E}\left[\frac{1-S}{(1-p_0)^2}+\frac{SA}{(p_0e_0)^2}\middle|X\right]\delta_{g_1}^2\right] && 1-S, SA\in\{0,1\} \\
=&\mathbb{E}\left[(1-p_0)^2\left(\frac{1}{1-p_0}+\frac{1}{p_0e_0}\right)\delta_{g_1}^2\right] && p_0(X)=\mathbb{P}[S=1|X] \\
& && e_0(X)=\mathbb{P}[A=1|S=1,X] \\
\leq&\frac{2}{\varepsilon_p\varepsilon_e}\mathbb{E}\delta_{g_1}^2 && \text{Assmp. } D.3.1, D.3.3
\end{aligned}$$

We can similarly bound $\mathbb{E}S_2^2$,

$$\begin{aligned}
&\mathbb{E}S_2^2 \\
=&\mathbb{E}\left[\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA(Y-g_{10})\right)^2\middle|X\right]\right] \\
=&\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\right)^2\mathbb{E}\left[(SA(Y-g_{10}))^2\middle|X\right]\right] \\
&\qquad\qquad\qquad\qquad \eta_0, \delta_\eta \text{ are functions of } X \\
=&\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\right)^2\right. \\
&\qquad\qquad \left.\mathbb{E}\left[(Y-g_{10})^2\middle|X,S=1,A=1\right]\mathbb{P}(S=1,A=1|X)\right] \\
&\qquad\qquad\qquad\qquad S,A\in\{0,1\} \\
\leq&E\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\right)^2\bar{\sigma}^2\mathbb{P}(S=1,A=1|X)\right] \\
&\qquad\qquad\qquad\qquad \text{Assmp. } D.3.4
\end{aligned}$$

$$=E\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{pep_0e_0}\right)^2\bar{\sigma}^2p_0e_0\right]$$

$$p_0(X)=\mathbb{P}[S=1|X]$$
$$e_0(X)=\mathbb{P}[A=1|S=1,X]$$

$$\leq\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left|e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e\right|^2$$

Assmp. $D.3.1, D.3.3, D.3.5.C$

$$\leq\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[e_0|\delta_p|+(1-p_0)p_0|\delta_e|+(1-p_0)|\delta_p\delta_e|\right]^2$$

Triangular ineq.

$$\leq\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[|\delta_p|+|\delta_e|+|\delta_p\delta_e|\right]^2$$
$$=\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[|\delta_p|+|\delta_e|+\frac{1}{2}|\delta_p||\delta_e|+\frac{1}{2}|\delta_p||\delta_e|\right]^2$$
$$\leq\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[|\delta_p|+|\delta_e|+\frac{1}{2}|\delta_p|+\frac{1}{2}|\delta_e|\right]^2$$

$$|\delta_p|,|\delta_e|\leq 1$$

$$=\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[|\delta_p|+|\delta_e|\right]^2$$
$$=\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\left[\mathbb{E}\delta_p^2+\mathbb{E}\delta_e^2+2\mathbb{E}|\delta_p||\delta_e|\right]$$
$$\leq\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\left[\mathbb{E}\delta_p^2+\mathbb{E}\delta_e^2+2\sqrt{\mathbb{E}\delta_p^2}\sqrt{\mathbb{E}\delta_e^2}\right]$$

Cauchy-Schwartz

$$=\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}\left(\sqrt{\mathbb{E}\delta_p^2}+\sqrt{\mathbb{E}\delta_e^2}\right)^2$$

Finally, we bound $\mathbb{E}S_3^2$,

$$\mathbb{E}S_3^2$$
$$=\mathbb{E}\left[\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}SA\delta_{g_1}\right)^2\middle|X\right]\right]$$
$$=\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\delta_{g_1}\right)^2\mathbb{E}\left[(SA)^2|X\right]\right] \qquad \eta_0,\delta_\eta \text{ are functions of } X$$

$$=\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{(p_0+\delta_p)(e_0+\delta_e)p_0e_0}\delta_{g_1}\right)^2\mathbb{E}\left[SA|X\right]\right] \qquad SA\in\{0,1\}$$

$$=\mathbb{E}\left[\left(\frac{e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e}{pep_0e_0}\delta_{g_1}\right)^2 p_0e_0\right] \qquad \begin{matrix} p_0(X)=\mathbb{P}[S=1|X] \\ e_0(X)=\mathbb{P}[A=1|S=1,X]\end{matrix}$$

$$\leq\frac{\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left|\left(e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e\right)\delta_{g_1}\right|^2 \qquad \text{Assmp. } D.3.1, D.3.3, D.3.5.C$$

$$=\frac{\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[|e_0\delta_p+(1-p_0)p_0\delta_e+(1-p_0)\delta_p\delta_e|^2|\delta_{g_1}|^2\right]$$

$$\leq\frac{\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\left[\left(|e_0\delta_p|+|(1-p_0)p_0\delta_e|+|(1-p_0)\delta_p\delta_e|\right)^2|\delta_{g_1}|^2\right] \qquad \text{Triangular ineq.}$$

$$\leq\frac{9\bar{C}^4}{\varepsilon_p\varepsilon_e}\mathbb{E}\delta_{g_1}^2 \qquad 0\leq p_0,e_0,|\delta_p|,|\delta_e|\leq 1$$

From the above, we have

$$\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2$$

$$=\mathbb{E}|\mathbf{1}(G=i)||S_1-S_2+S_3|^2$$

$$\leq\mathbb{E}|S_1-S_2+S_3|^2$$

$$\leq\mathbb{E}(|S_1|+|S_2|+|S_3|)^2$$

$$\text{Triangular ineq.}$$

$$=\left[\mathbb{E}S_1^2+\mathbb{E}S_2^2+\mathbb{E}S_3^2+2\mathbb{E}|S_1S_2|+2\mathbb{E}|S_1S_3|+2\mathbb{E}|S_2S_3|\right]$$

$$\leq\Big[\mathbb{E}S_1^2+\mathbb{E}S_2^2+\mathbb{E}S_3^2+ \\ 2\sqrt{\mathbb{E}|S_1|^2}\sqrt{\mathbb{E}|S_2|^2}+2\sqrt{\mathbb{E}|S_1|^2}\sqrt{\mathbb{E}|S_3|^2}+2\sqrt{\mathbb{E}|S_2|^2}\sqrt{\mathbb{E}|S_3|^2}\Big]$$

$$\text{Cauchy-Schwartz}$$

$$=\left[\sqrt{\mathbb{E}S_1^2}+\sqrt{\mathbb{E}S_1^2}+\sqrt{\mathbb{E}S_1^2}\right]^2$$

$$\leq\left[\sqrt{\frac{2}{\varepsilon_p\varepsilon_e}}\sqrt{\mathbb{E}\delta_{g_1}^2}+\sqrt{\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}}\left(\sqrt{\mathbb{E}\delta_p^2}+\sqrt{\mathbb{E}\delta_e^2}\right)+\sqrt{\frac{9\bar{C}^4}{\varepsilon_p\varepsilon_e}}\sqrt{\mathbb{E}\delta_{g_1}^2}\right]^2$$

$$\leq C\left[\sqrt{\mathbb{E}\delta_{g_1}^2}+\sqrt{\mathbb{E}\delta_p^2}+\sqrt{\mathbb{E}\delta_e^2}\right]^2$$

$$\text{letting } C:=\frac{\left(3\bar{C}^2+\sqrt{2}\right)^2}{\varepsilon_p\varepsilon_e}\vee\frac{9}{4}\frac{\bar{\sigma}^2\bar{C}^4}{\varepsilon_p\varepsilon_e}$$

So we have,

$$
\begin{aligned}
&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2\\
\leq& C\sup_{\eta\in\mathcal{T}_n}\left[\sqrt{\mathbb{E}\delta_{g_1}^2}+\sqrt{\mathbb{E}\delta_p^2}+\sqrt{\mathbb{E}\delta_e^2}\right]^2\\
\leq& C\left[\sup_{\eta\in\mathcal{T}_n}\sqrt{\mathbb{E}\delta_{g_1}^2}+\sup_{\eta\in\mathcal{T}_n}\sqrt{\mathbb{E}\delta_p^2}+\sup_{\eta\in\mathcal{T}_n}\sqrt{\mathbb{E}\delta_e^2}\right]^2\\
\leq& C\left[\mathbf{g}_n+\mathbf{p}_n+\mathbf{e}_n\right]^2 && \text{Assmp. } D.3.5\\
\leq& 9C\left[\mathbf{g}_n\vee\mathbf{p}_n\vee\mathbf{e}_n\right]^2=o_P(1) && \text{Assmp. } D.3.5.A
\end{aligned}
$$

Assembling the proofs for Lemmas D.4.3 and D.4.4

For Lemma D.4.3:

$$
\begin{aligned}
&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0))]\right|\\
=&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))]-\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))]\right|\\
\leq&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left\{\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))]\right|+\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))]\right|\right\}\\
&\qquad\qquad\text{Triangular ineq.}\\
\leq&\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))]\right|+\sqrt{n}\sup_{\eta\in\mathcal{T}_n}\left|\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))]\right|=o_P(1)\\
&\qquad\qquad\text{Lem. } D.4.6
\end{aligned}
$$

For Lemma D.4.4:

$$
\begin{aligned}
&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0))\right|^2\\
=&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))-\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2\\
\leq&\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left\{\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|+\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|\right\}^2\\
&\qquad\qquad\text{Triangular ineq.}\\
=&\sup_{\eta\in\mathcal{T}_n}\Big\{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2+\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2+
\end{aligned}
$$

$$
\begin{aligned}
&\qquad 2\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|\}\\
\leq &\sup_{\eta\in\mathcal{T}_n}\Big\{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2+\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2+\\
&\qquad 2\sqrt{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2}\sqrt{\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2}\Big\}\\
&\qquad\qquad\qquad\qquad\text{Cauchy-Schwartz}\\
\leq &\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2+\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2+\\
&\qquad 2\sqrt{\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^1(\eta)-\mathcal{Y}^1(\eta_0))\right|^2}\sqrt{\sup_{\eta\in\mathcal{T}_n}\mathbb{E}\left|\mathbf{1}(G=i)(\mathcal{Y}^0(\eta)-\mathcal{Y}^0(\eta_0))\right|^2}\\
=&o_P(1)\text{ Lem. } D.4.7
\end{aligned}
$$

□

D.4.4 Proof for Lemma D.4.5

Now that we have shown Lemmas D.4.3 and D.4.4, it remains to show Lemma D.4.5.

Lemma D.4.5 (Numerator of difference is $o_P(1)$)**.** *Under Assumptions D.3.1, D.3.3, D.3.4 and D.3.5,*

$$
\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))=o_P(1),\quad\forall m\in\{1,2,...,M\}
$$

Proof. First we observe

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))\\
=&\frac{1}{\sqrt{N}}\left[\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))-\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]\right]+\\
&\qquad\qquad\frac{1}{\sqrt{N}}\sum_{j\in I_m}\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]\\
=&\underbrace{\frac{N_M}{\sqrt{N}}\left[\left(\frac{1}{N_M}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))\right)-\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]\right]}_{R_1(m)}+
\end{aligned}
$$

$$\underbrace{\frac{N_M}{\sqrt{N}}\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]}_{R_2(m)}$$

$$:= R_1(m) + R_2(m)$$

We define the event $\mathcal{E}_N$ as $\cap_k \left(\hat{\eta}_m \in \mathcal{T}_{N_M^c}\right)$, i.e. all M nuisance function estimates falling into the high-probability neighborhood where Lemmas D.4.3 and D.4.4 apply. From union bound,

$$1-\mathbb{P}(\mathcal{E}_N) \leq \sum_k \mathbb{P}(\hat{\eta}_m \notin \mathcal{T}_{N_M^c}) \ \leq K\epsilon_{N_M^c} = o_P(1) \qquad \because \epsilon_n = o_P(1)$$

Conditional on $\mathcal{E}_N$ and the data complementary to fold m, which we denote as D_m, we have for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}(|R_1(m)| \geq \epsilon|\mathcal{E}_N, D_m)\\
=&\mathbb{P}\left(\left|\left(\frac{1}{N_M}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))\right)-\right.\right.\\
&\qquad\qquad\left.\left.\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))]\right| \geq \frac{\sqrt{N}}{N_M}\epsilon\right|\mathcal{E}_N, D_m\right)\\
\leq&\frac{N_M^2}{N\epsilon^2}var\left[\frac{1}{N_M}\sum_{j\in I_m}\mathbf{1}(G_j=i)(\mathcal{Y}_j(\hat{\eta}_m)-\mathcal{Y}_j(\eta_0))\middle|\mathcal{E}_N, D_m\right]\\
&\qquad\qquad\text{Chebyshev ineq.}\\
=&\frac{N_M}{N\epsilon^2}var\left[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0))|\mathcal{E}_N, D_m\right]\\
\leq&\frac{1}{M\epsilon^2}\mathbb{E}\left[(\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m)-\mathcal{Y}(\eta_0)))^2\middle|\mathcal{E}_N, D_m\right]\\
\leq&\frac{1}{M\epsilon^2}\sup_{\eta\in\mathcal{T}_{N_M}}\mathbb{E}\left[(\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0)))^2\middle|D_m\right]\\
&\qquad\qquad\hat{\eta}_m \in \mathcal{T}_{N_M^c}\text{ under }\mathcal{E}_N\\
\leq&\frac{1}{M\epsilon^2}\sup_{\eta\in\mathcal{T}_{N_M}}\mathbb{E}\left(\mathbf{1}(G=i)(\mathcal{Y}(\eta)-\mathcal{Y}(\eta_0))\right)^2\\
&\qquad\qquad\text{Fold } m \text{ is independent of } D_m\\
=&\frac{1}{M\epsilon^2}o_P(1) = o_P(1) \text{ Lem. } D.4.4
\end{aligned}$$

Also, conditional on $\mathcal{E}_N$ and D_m

$$\begin{aligned}
|R_2(m)| =& \left| \frac{N_M}{\sqrt{N}\sqrt{N_M^c}} \sqrt{N_M^c} \mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m) - \mathcal{Y}(\eta_0))] \middle| \mathcal{E}_N, D_m \right| \\
=& \frac{1}{\sqrt{M-1}} \sqrt{N_M^c} \, |\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\hat{\eta}_m) - \mathcal{Y}(\eta_0))]| \mathcal{E}_N, D_m| \\
\leq& \frac{1}{\sqrt{M-1}} \sqrt{N_M^c} \sup_{\eta \in \mathcal{T}_{N_M^c}} |\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))|D_m]| \\
& \qquad\qquad\qquad\qquad \hat{\eta}_m \in \mathcal{T}_{N_M^c} \text{ under } \mathcal{E}_N \\
=& \frac{1}{\sqrt{M-1}} \sqrt{N_M^c} \sup_{\eta \in \mathcal{T}_{N_M^c}} |\mathbb{E}[\mathbf{1}(G=i)(\mathcal{Y}(\eta) - \mathcal{Y}(\eta_0))]| \\
& \qquad\qquad\qquad\qquad \text{Fold } m \text{ is independent of } D_m \\
=& \frac{1}{\sqrt{M-1}} o_P(1) = o_P(1) \text{ Lem. } D.4.3
\end{aligned}$$

Which implies that, conditional on $\mathcal{E}_N$ and D_m, $R_1(m) + R_2(m) = o_P(1)$. So for any $\epsilon > 0$:

$$\begin{aligned}
& \mathbb{P}\left(\left| \frac{1}{\sqrt{N}} \sum_{j \in I_m} \mathbf{1}(G_j = i)(\mathcal{Y}_j(\hat{\eta}_m) - \mathcal{Y}_j(\eta_0)) \right| > \epsilon \right) \\
=& \mathbb{P}(|R_1(m) + R_2(m)| \geq \epsilon) \\
=& \mathbb{P}(|R_1(m) + R_2(m)| \geq \epsilon | \mathcal{E}_N) \mathbb{P}(\mathcal{E}_N) + \mathbb{P}(|R_1(m) + R_2(m)| \geq \epsilon | \bar{\mathcal{E}}_N)(1 - \mathbb{P}(\mathcal{E}_N)) \\
\leq& \mathbb{P}(|R_1(m) + R_2(m)| \geq \epsilon | \mathcal{E}_N) + (1 - \mathbb{P}(\mathcal{E}_N)) \\
=& \int \mathbb{P}(|R_1(m) + R_2(m)| \geq \epsilon | \mathcal{E}_N, D_m) d\mathbb{P}(D_m | \mathcal{E}_N) + (1 - \mathbb{P}(\mathcal{E}_N)) \\
=& o_P(1) + o_P(1) = o_P(1)
\end{aligned}$$

and the lemma is proven. □

D.5 Details on WHI Experiments

We assess our algorithm on clinical trial data and observational data available from the Women's Health Initiative (WHI). The RCTs were run by the WHI via 40 US clinical centers from 1993-2005 (1993-1998: enrollment + randomization; 2005: end of follow-up) on postmenopausal women aged 50-79 years, and the observational dataset was designed and run in parallel on a similar population. Note that this data is publicly available to researchers and requires only an application on BIOLINCC (https://biolincc.nhlbi.nih.gov/studies/whi_ctos/).

D.5.1 Data

WHI RCT – There are three clinical trials associated with the WHI. The RCT that we will be leveraging in this set of experiments is the Postmenopausal Hormone Therapy (PHT) trial, which was run on postmenopausal women aged 50-79 years who had an intact uterus. This trial included a total of $N_{HT} = 16608$ patients. The intervention of interest was a hormone combination therapy of estrogen and progesterone. Specifically, post-randomization, the treatment group was given 2.5 mg of medroxyprogesterone as well as 0.625 mg of estrogen a day. The control group was given a placebo. Finally, there are several outcomes that were tracked and studied in the principal analysis done on this trial [229]. These outcomes are of three broad categories: a) cardiovascular events, including coronary heart disease, which served as a primary endpoint b) cancer (e.g. endometrial, breast, colorectal, etc.), and c) fractures.

WHI OS – The observational study component of the WHI tracked the medical events and health habits of $N = 93676$ women. Recruitment for the study began in 1994 and participants were followed until 2005, i.e. a similar follow-up to the RCT. Follow-up was done in a similar fashion as in the RCT (i.e. patients would have annual visits, in addition to a "screening" visit, where they would be given survey forms to fill out to track any events/outcomes). Thus, the same outcomes, including cancers, fractures, and cardiovascular events, are tracked in the observational study.

D.5.2 Outcome

The outcome of interest in our analysis is a "global index", which is a summary statistic of several outcomes, including coronary heart disease, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, and death due to other causes. Events or outcomes are tracked for each patient, and are recorded as “day of event/outcome” in the data, where the initial time-point for follow-up is the same for both the RCT and OS. At a high level, the “global index” is essentially the minimum “event day” when considering all the previously mentioned events.

We binarize the “global index,” by choosing a time point, t, before the end of follow-up and letting $Y = 1$ if the observed event day is before t and $Y = 0$ otherwise. Thus, we are looking at whether the patient will experience the event within some particular period of time or not. We set $t = 7$ years. Note that we sidestep censorship of a patient before the threshold by defining the outcomes in the following way: $Y = 1$ indicates that a patient is observed to have the event before the threshold, and $Y = 0$ indicates that a patient is not observed to have the event before the threshold. We apply this binarization in the same way for both the RCT and OS. Extending our method to a survival analysis framing is beyond the scope of this paper, but an interesting direction for future work.

D.5.3 Intervention

Recall from above that the intervention studied in the RCT was 2.5mg of medroxyprogesterone + 0.625 mg of estrogen and the control was a placebo pill. The RCT was run as an “intention-to-treat” trial. To establish “treatment” and “control” groups in the OS, we leverage the annual survey data collected from patients and assign a patient to the treatment group if they confirm usage of both estrogen and progesterone in the first three years. A patient is assigned to the control group if they deny usage of both estrogen and progesterone in the first three years. We exclude a patient from the analysis if she confirms usage of one and not the other OR if the field in the survey is missing OR if they take some other hormone therapy. We end up with a total of $N_{obs} = 33511$ patients.

D.5.4 Data Processing + Covariates

We use only covariates that are measured both in the RCT and OS to simplify the analysis. Because this information is gathered via the same set of questionnaires, they each indicate the same type of covariate. In other words, there is consistency of meaning across the same covariates across the RCT and the OS. We end up with a total of 1576 covariates.

D.5.5 Details of Experimental Setup

We give a more detailed exposition of the steps in our experimental workflow, which were described in brief in the main paper.

- **Step 0**: *Replicate the principal results from the PHT trial, given in Table 2 of [229], using the WHI OS data.* In this step, we fit a doubly robust estimator of the style given in Appendix D.3.

- **Step 1**: *While treating the WHI OS dataset as the "unbiased" observational dataset (hence the need for Step 0), simulate additional "biased" observational datasets by inducing bias into the WHI OS.* We construct four additional "biased" datasets (for a total of five observational datasets, including the WHI OS dataset), where we use the following procedure to induce selection bias – of the people who were not exposed to the treatment and did not end up getting the event, we drop each person with some probability, p. We set $p = [0.1, 0.3, 0.5, 0.7]$ to get the four additional observational datasets.

  This type of selection bias may reflect the following clinical scenario: consider a patient who is relatively healthy who does not end up taking any hormone therapy. This patient might enroll initially in the OS, but may drop out or stop responding to the surveys. If the committee running the study does not explicitly account for this drop-out rate, then the resultant study will suffer from selection bias. [19] detail additional examples of selection bias that can occur in observational studies. Importantly, this part is the only part of our setup that involves any simulation. However, in order to properly evaluate our method, we need to know which datasets are biased and unbiased in our set. Thus, we opt to simulate the bias.

- **Step 2**: *Run our procedure over "multiple tasks," generating confidence intervals on the treatment effect for different subgroups.* To do so, we compile a list of covariates, taking both from [236] as well as covariates with high feature importance in both the propensity score model and response surface model from the estimator in Step 0. We generate all pairs from this list and use each pair to generate four subgroups. We treat two of the subgroups as validation subgroups and two of them as extrapolated subgroups in that we "hide" the RCT data in those subgroups when fitting our doubly robust transported estimator. (This gives us the benefit of knowing the RCT result for the extrapolated subgroups, which is useful in evaluation). Pairs that do not have enough support (threshold of 400 observations) in each group are removed. The total number of "tasks" (or covariate pairs) that we have is 592 (and therefore 2368 subgroups).

- **Step 3**: *Evaluate ExPCS (our method), ExOCS, Simple, and Meta-Analysis for each of the covariate pairs.* Additionally, we evaluate an "oracle" method, which always selects only the original observational study (i.e. the base WHI OS to which we have not added any selection bias) and reports the interval estimate computed on this study. To evaluate these methods, we will treat the RCT point estimates as "correct." For each, we compute the following metrics: **Length** – length of the confidence interval for the subgroup; **Coverage** – percentage of tasks for which the method's interval covers the RCT point estimate; **Unbiased OS Percentage** – across all tasks, the percentage at which our approach retains the unbiased study after the falsification step.

Note that we utilize sample splitting when running the above procedure. Namely, we use 50% of the data as a "training" set, where we experiment with different classes of covariates and different types of bias, and then reserve 50% of the data as a "testing" set, on which we do the final run of the analysis and report results. All nuisance functions in the doubly robust estimator are fit with a Gradient Boosting Classifier with significant regularization. In practice, we found that any highly-regularized tree-based model works well.

D.5.6 Covariate List for Task Generation

Below is the list of covariates used to generate the tasks in **Step 2** of our experiment:

- ALCNOW (current alcohol user)
- BMI ≤ 30
- BLACK
- SMOKING (current smoker)
- DIAB (diabetes ever)
- HYPT (hypertension ever)
- BRSTFEED (breastfeeding)
- MSMINWK ≤ 106 (minutes of moderate to strenuous activity per week)
- BRSTBIOP (breast biopsy done)
- RETIRED
- EMPLOYED
- OC (oral contraceptive use ever)
- LIVPRT (live with husband or partner)
- MOMALIVE (natural mother still alive)
- LATREGION-Northern $>$ 40 degrees north
- BKBONE (broke bone ever)

- NUMFALLS
- GRAVID (gravidity)
- AGE ≤ 64
- ANYMENSA ≤ 51 (age at last bleeding)
- MENOPSEA ≤ 50 (age at last regular period)
- MENO ≤ 51 (age at menopause)
- LSTPAPDY (days from randomization to last pap smear)
- BMI ≤ 27.7
- TMINWK ≤ 191 (minutes of recreational exercise per week)
- HEIGHT ≤ 161
- WEIGHT ≤ 72
- WAIST ≤ 86
- HIP ≤ 105
- WHR ≤ 0.81 (waist to hip ratio)
- TOTHCAT (HRT duration by category)
- MEDICARE (on medicare)
- HEMOGLBN ≤ 13
- PLATELET ≤ 244
- WBC ≤ 6
- HEMATOCR ≤ 40

D.6 Details on Semi-Synthetic Experiment (Data Generation and Model Hyperparameters)

D.6.1 Data Generation

For each simulated dataset, we generate 1 RCT and K observational studies. The RCT is assumed to have covariate values identical to the IHDP dataset but is restricted to infants with married mothers. For the observational studies, we resample the rows of the IHDP dataset to the desired sample size $n = rn_0$. The covariate distribution of the observational studies are made different from the RCT by weighted sampling, with the relative weights set as

$$w = 0.8^{\mathbf{1}(\text{male infant})+\mathbf{1}(\text{mother smoked})+\mathbf{1}(\text{mother worked during pregnancy})}$$

Then, to introduce confounding (in the observational data), we generate m_c continuous confounders and m_b binary confounders. Each continuous confounder is drawn from a mixture of $0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(3,1)$ in the RCT and $(0.25 + 0.5A)\mathcal{N}(3,1) + (0.75 - 0.5A)\mathcal{N}(0,1)$ in the observational studies, where A is the treatment indicator. Similarly, each binary confounder is drawn from $\text{Bern}(0.5)$ in the RCT and $\text{Bern}(0.25 + 0.5A)$ in the observational studies. In the following, we denote the covariate vector as $X \in \mathbb{R}^{m_x}$ where $m_x = 28$ is the number of covariates in the IHDP dataset, and the generated confounder vector as $Z \in \mathbb{R}^{(m_c+m_b)}$. For brevity, we also denote the vector $(A, X^\top)^\top$ as $\tilde{X}$.

For outcome simulation in the datasets, we modify *response surface B* from Hill [118] to account for additional confounding variables. We set the following counterfactual outcome distributions:

$$\begin{aligned} Y_0 &\sim \mathcal{N}\left(\exp\left[\left(\tilde{X} + \frac{1}{2}\mathbf{1}\right)^\top \beta\right] + Z^\top\gamma, 1\right) \\ Y_1 &\sim \mathcal{N}(\tilde{X}^\top\beta + Z^\top\gamma + \omega, 1), \end{aligned}$$

where $\mathbf{1} \in \mathbb{R}^{(m_x+1)}$ is vector of ones, $\beta \in \mathbb{R}^{(m_x+1)}$ is a vector where each element is randomly sampled from $(0, 0.1, 0.2, 0.3, 0.4)$ with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$, $\gamma \in \mathbb{R}^{(m_c+m_b)}$ is a vector where each element is randomly sampled from $(0.1, 0.2, 0.5, 0.75, 1)$ with uniform probability, and $\omega = 23$ is a constant chosen to limit the size of the GATEs. The observed outcome is then $Y \coloneqq AY_1 + (1 - A)Y_0$. We then conceal a number of confounders, chosen in order from the highest to lowest weighted, from each observational study to mimic the scenario of unobserved confounding. The number of concealed confounders in each observational study is denoted as $\mathbf{c_z} = (c_{z1}, c_{z2}, ..., c_{zK})$.

D.6.2 Hyperparameters

Logistic regression

Hyperparametes	Value set
Penalty type	ℓ_2
Penalty coefficient	$\{1, 0.1, 0.01, 0.001\}$

Multilayer perceptron regression

Hyperparametes	Value set
# of hidden layers and # of perceptrons	$[1, (100)], [2, (50, 50)], [2, (25, 25)]$
Activation function	ReLU, tanh
Solver	Adam
Alpha	$(1, 0.1, 0.01, 0.001, 0.0001)$
Learning rate	0.001
# of epochs	$(250, 500)$

D.7 Additional Semi-Synthetic Experimental Results

An analysis of including biased observational studies: In Figure D-4 and Table D.1, we study coverage probability and width of confidence intervals in the presence of biased studies. *Meta-Analysis* intervals approach zero coverage probability

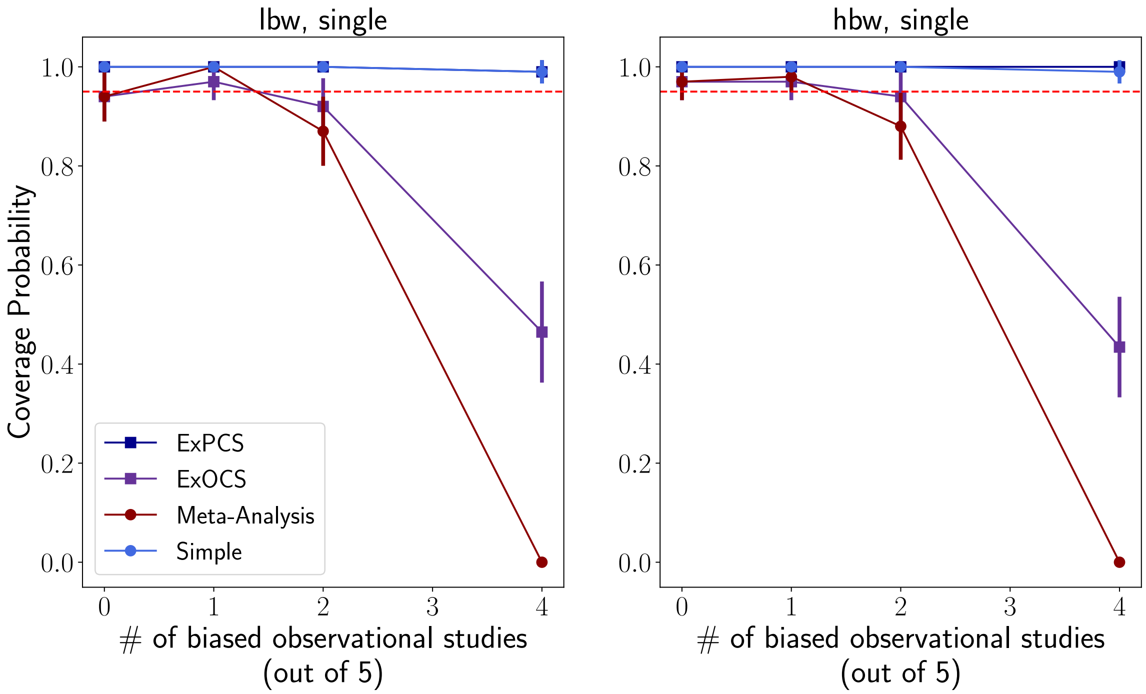


Figure D-4: Coverage probabilities of confidence intervals shown as a function of the number of biased observational datasets (out of five). In the 4/5 biased studies case, the average interval widths for each approach is shown for two subgroups. We observe that *ExPCS* achieves the best balance of interval width and coverage.

Mean width of 95% confidence intervals (4 biased studies)				
	ExPCS	ExOCS	Meta	Simple
LS	5.29	3.55	2.34	5.51
HS	5.84	3.76	2.48	6.15

Table D.1: This table should be interpreted in conjunction with Figure D-4. In the 4/5 biased studies case, the average interval widths for each approach is shown for two subgroups. We observe that *ExPCS* achieves the best balance of interval width and coverage.

as the number of biased studies increases. Indeed, a fundamental assumption of this approach is that differences between estimates are only due to random variation, leading to poorer coverage probability when there are more biased studies. *ExOCS* allows for elimination of biased studies in principle through falsification, resulting in improved coverage. However, it does not maintain the desired threshold of coverage (95%), since biased estimators may still be included after falsification either due to chance or by being underpowered. Finally, *ExPCS* and *Simple Union* intervals have

Upsampling ratio	1.0	3.0	5.0	10.0
P(selecting biased study)	0.98	0.80	0.68	0.60

Table D.2: P(selecting biased study) as a function of upsampling ratio

good coverage across the board, but as before, *ExPCS* results in narrower intervals.

Overall, we find that our method is robust to biased studies, yielding a good balance between coverage and width. In the case where one has adequate power, *ExOCS* could be a reasonable alternative to get narrower intervals for a sacrifice in coverage (even in the presence of biased studies). However, this implicitly assumes that an estimator consistent for the validation effects will be consistent for the extrapolated effects. If this assumption does not hold, then *ExOCS* will have poor coverage.

Biased estimator selection: In Table D.2, we see that the probability of selecting the biased estimator goes down with increasing sample size of the observational studies, reflected by the increasing sample size ratio, r. This result validates our intuition that our method is more useful and results in more precise estimates of bias as we obtain more observational samples.

Appendix E

Supplementary Material for Chapter 7

E.1 Proofs

E.1.1 Proof of Proposition 7.2.1

Proposition 7.2.1. *Under assumptions 7.2.1 and 7.2.3, the CATE of the RCT given X, $\mathbb{E}[Y_1 - Y_0|X, S = 0]$, is identifiable in the observational data by*

$$\mathbb{E}[Y \mid X, A = 1, S = 1] - \mathbb{E}[Y \mid X, A = 0, S = 1] \tag{7.1}$$

Proof.

$$\begin{aligned}
&\mathbb{E}[Y_1 - Y_0 \mid X, S = 0] \\
&= \mathbb{E}[Y_1 - Y_0 \mid X, S = 1] \\
&= \mathbb{E}[Y_1 \mid X, S = 1] - \mathbb{E}[Y_0 \mid X, S = 1] \\
&= \mathbb{E}[Y \mid X, A = 1, S = 1] - \mathbb{E}[Y \mid X, A = 0, S = 1]
\end{aligned}$$

The first equality follows from the mean exchangeability of the contrast (Assumption 7.2.3) and the second from the linearity of the expectation operator. The final equality follows from ignorability and consistency (Assumption 7.2.1). □

E.1.2 Proof of Proposition 7.3.1

Proposition 7.3.1 (*CATE signal from the RCT*)**.** *Under assumption 7.2.2, the instance-wise CATE signal ψ_0 in Equation* (7.3)*, which uses the outcome information from the RCT, is unbiased, i.e.,* $\mathbb{E}[\psi_0|X] = \mathbb{E}[Y_1 - Y_0|X, S = 0]$.

Proof. Here, we show that the signal

$$\psi_0 = \left(\frac{\mathbf{1}\{A=1\}}{P(A=1 \mid S=0)} - \frac{\mathbf{1}\{A=0\}}{P(A=0 \mid S=0)}\right) \cdot \frac{\mathbf{1}\{S=0\}}{P(S=0 \mid X)} Y$$

is an unbiased estimator of the CATE in RCT population under consistency and fully randomized treatment assignment (i.e., $P(A \mid X, S=0) = P(A \mid S=0)$, and $Y_a \perp\!\!\!\perp A$ as in Assumption 7.2.2). In particular, we can observe that

$$\begin{aligned}\mathbb{E}[\psi_0(A,Y,S,X) \mid X] &= \sum_{A,Y,S} \psi_0(A,Y,S,X) \cdot P(A,Y,S \mid X) \\ &= \sum_{A,Y,S} \frac{\mathbf{1}\{A=1\}}{P(A=1 \mid S=0)} \cdot \frac{\mathbf{1}\{S=0\}}{P(S=0 \mid X)} Y \cdot P(A,Y,S \mid X) \\ &\quad - \sum_{A,Y,S} \frac{\mathbf{1}\{A=0\}}{P(A=0 \mid S=0)} \cdot \frac{\mathbf{1}\{S=0\}}{P(S=0 \mid X)} Y \cdot P(A,Y,S \mid X)\end{aligned} \tag{E.1}$$

We focus on the first term, observing that the second term can be handled similarly. We first re-write the first term as

$$\begin{aligned}&\sum_{A,Y,S} \frac{P(Y \mid S,A,X)P(A \mid S,X)P(S \mid X)}{P(A=1 \mid S=0)P(S=0 \mid X)} \cdot \mathbf{1}\{A=1, S=0\} \cdot Y \\ &= \sum_{Y} Y \cdot \frac{P(Y \mid S=0, A=1, X)\cancel{P(A=1 \mid S=0, X)}\cancel{P(S=0 \mid X)}}{\cancel{P(A=1 \mid S=0)}\cancel{P(S=0 \mid X)}} \\ &= \mathbb{E}[Y \mid S=0, A=1, X] \\ &= \mathbb{E}[Y_1 \mid S=0, A=1, X] \\ &= \mathbb{E}[Y_1 \mid X, S=0]\end{aligned} \tag{E.2}$$

Repeating the similar arguments for the second term in Eq. E.1, we have $\mathbb{E}[\psi_0(A,Y,S,X) \mid$

$X] = \mathbb{E}[Y_1 - Y_0 \mid X, S = 0]$, which completes the proof. □

E.1.3 Proof of Proposition 7.3.2

Proposition 7.3.2 (*CATE signal from the observational data*)**.** *Under Assumptions 7.2.1 and 7.2.3, the instance-wise CATE signal ψ_1 in Eq. 7.4, which uses the outcome information from the observational data, is unbiased for the CATE in the RCT population, i.e.,* $\mathbb{E}[\psi_1|X] = \mathbb{E}[Y_1 - Y_0|X, S = 0]$.

Proof. We have,

$$\begin{aligned}\psi_1 = \frac{1}{P(S=0\mid X)}\Bigg[&\mathbf{1}\{S=0\}\left(\mu_1(X)-\mu_0(X)\right)\\ &+\mathbf{1}\{S=1\}\,\frac{P(S=0\mid X)}{P(S=1\mid X)}\left(\frac{\mathbf{1}\{A=1\}\left(Y-\mu_1(X)\right)}{P(A=1\mid S=1,X)}\right.\\ &\left.-\frac{\mathbf{1}\{A=0\}\left(Y-\mu_0(X)\right)}{P(A=0\mid S=1,X)}\right)\Bigg]\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\psi_1(A,Y,S,X)\mid X] = \frac{1}{\cancel{P(S=0\mid X)}}\Bigg[&\sum_{A,Y}\left(\mu_1(X)-\mu_0(X)\right)\\ &P(Y\mid S=0,A,X)P(A\mid S=0,X)\cancel{P(S=0\mid X)}\\ &+\frac{\cancel{P(S=0\mid X)}}{\cancel{P(S=1\mid X)}}\left(\sum_{Y}\frac{Y-\mu_1(X)}{\cancel{P(A=1\mid S=1,X)}}P(Y\mid S=1,A=1,X)\right.\\ &\cancel{P(A=1\mid S=1,X)}\cancel{P(S=1\mid X)}\\ &-\sum_{Y}\frac{Y-\mu_0(X)}{\cancel{P(A=0\mid S=1,X)}}P(Y\mid S=1,A=0,X)\\ &\left.\cancel{P(A=0\mid S=1,X)}\cancel{P(S=1\mid X)}\right)\Bigg]\\ &=\sum_{A,Y}\left(\mu_1(X)-\mu_0(X)\right)P(Y\mid S=0,A,X)P(A\mid S=0,X)\\ &\quad+\sum_{Y}\left(Y-\mu_1(X)\right)P(Y\mid S=1,A=1,X)\\ &\quad-\sum_{Y}\left(Y-\mu_0(X)\right)P(Y\mid S=1,A=0,X)\end{aligned}$$

$$
\begin{aligned}
&= (\mu_1(X) - \mu_0(X)) \underbrace{\sum_{A,Y} P(Y, A \mid S = 0, X)}_{=1} \\
&\quad + \mathbb{E}[Y \mid S = 1, A = 1, X] - \mu_1(X) \cdot \underbrace{\sum_{Y} P(Y \mid S = 1, A = 1, X)}_{=1} \\
&\quad - \mathbb{E}[Y \mid S = 1, A = 0, X] - \mu_0(X) \cdot \underbrace{\sum_{Y} P(Y \mid S = 1, A = 0, X)}_{=1}
\end{aligned} \tag{E.3}
$$

$$
= \mathbb{E}[Y \mid S = 1, A = 1, X] - \mathbb{E}[Y \mid S = 1, A = 0, X] \tag{E.4}
$$

$$
= \mathbb{E}[Y_1 - Y_0 \mid X, S = 0] \tag{E.5}
$$

Note that we have Eq. E.3 and Eq. E.4 since $\mu_a(X) = \mathbb{E}[Y \mid S = 1, A = a, X]$. Eq. E.5 follows from Proposition 7.2.1. □

E.1.4 Proof of Proposition 7.3.3

We restate Proposition 7.3.3 here for convenience.

Proposition 7.3.3 (*Null Hypothesis, CMR*)**.** *Under Assumptions 7.2.1 to 7.2.3, we have a set of conditional moment restrictions (CMRs) on the signal difference, ψ:*

$$
H_0 : \mathbb{E}[\psi | X] = 0 \qquad P_X\text{-almost surely,} \tag{7.5}
$$

where P_X is the distribution of X on the joint distribution of the RCT and observational study. Equation (7.5) *implies an infinite set of unconditional moment restrictions, $\mathbb{E}[\psi f(X)] = 0, \forall f \in \mathcal{F}$, where $\mathcal{F}$ is the set of measurable functions on $\mathcal{X}$.*

Proof. Under Assumptions 7.2.1 to 7.2.3, we have $\mathbb{E}[\psi_0|X] = \mathbb{E}[Y_1 - Y_0|X, S = 0]$ and $\mathbb{E}[\psi_1|X] = \mathbb{E}[Y_1 - Y_0|X, S = 0]$ by Propositions 7.3.1 and 7.3.2 as discussed in Section 7.3.1. That is, $\mathbb{E}[\psi|X] = 0$ where $\psi = \psi_1 = \psi_0$ is the signal difference given two CATE signals. Let $\mathcal{F}$ be the set of measurable functions on $\mathcal{X}$. Then, by the Law of Iterated Expectations we have

$$\mathbb{E}[\psi f(X)] = \mathbb{E}_X[\mathbb{E}[\psi f(X)|X]] = \mathbb{E}_X[\mathbb{E}[\psi|X]f(X)], \qquad P_X\text{-a.s.}, \forall f \in \mathcal{F}$$

We see that Eq. 7.5 implies the following infinite set of unconditional moment restrictions,

$$\mathbb{E}[\psi f(X)] = 0, \qquad P_X\text{-a.s.}, \forall f \in \mathcal{F}$$

□

Proof. First, to show how to express our null hypothesis as a CMR, we have that, under **??**, $\mathbb{E}[\psi_0|X_\tau] = \mathbb{E}[Y_1 - Y_0|X_\tau, S = 0]$, and under Assumptions 7.2.1 and 7.2.3 and **??**, we have $\mathbb{E}[\psi_1|X_\tau] = \mathbb{E}[Y_1 - Y_0|X_\tau, S = 0]$. Under the null hypothesis, we have it that the CATE inferred from the RCT and the CATE inferred from the observational study transported to the RCT population are equivalent, i.e. $\mathbb{E}[\psi_0|X_\tau] = \mathbb{E}[\psi_1|X_\tau]$. Rearranging terms and from our definition of ψ, we arrive at the CMR formulation of the null hypothesis.

Now, suppose we have a measurable set of functions, $\mathcal{F}$, on $\mathcal{X}_\tau$. Then, we have, $\mathbb{E}[\psi f(X_\tau)] = \mathbb{E}_{X_\tau}[\mathbb{E}[\psi f(X_\tau)|X_\tau]] = \mathbb{E}_{X_\tau}[\mathbb{E}[\psi|X_\tau]f(X_\tau)]$ for any $f \in \mathcal{F}$, which follows from the Law of Iterated Expectation. We see that Eq. 7.5 implies the following infinite set of unconditional moment restrictions,

$$\mathbb{E}[\psi f(X_\tau)] = 0, \quad \forall f \in \mathcal{F}$$

□

E.1.5 Proofs for Theorem 7.3.1 and Corollary 7.3.2

We restate the theorem and corollary here for convenience.

Theorem 7.3.1 (*Maximum Moment Restriction-based test for CATE function*)**.** *Let $\mathcal{F}$ be a RKHS with reproducing kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is ISPD, continuous and bounded. Suppose $|\mathbb{E}[\psi|X]| < \infty$ almost surely in P_X, and $\mathbb{E}[[\psi k(X, X')\psi']^2] < \infty$*

where (ψ', X') *is an independent copy of* (ψ, X)*. Let* $\mathbb{M}^2 = \sup_{f\in\mathcal{F},||f||\leq 1}(\mathbb{E}[\psi f(X)])^2$*. Then,*

1. *The conditional moment testing problem in Eq. 7.5 can be reformulated in terms of the MMR as* $H_0^{':\mathbb{M}^2=0}$*,* $H_1^{':\mathbb{M}^2\neq 0}$*.*

Further, let the test statistic be the empirical estimate of $\mathbb{M}^2$*,*

$$\hat{\mathbb{M}}_n^2 = \frac{1}{n(n-1)} \sum_{i,j\in\mathcal{I},i\neq j} \psi_i k(x_i, x_j)\psi_j$$

2. *Then, under* $H_0^{'}$*,*
$$n\hat{\mathbb{M}}_n^2 \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1)$$
where Z_j *are independent standard normal variables and* λ_j *are the eigenvalues for* $\psi k(x, x')\psi'$*.*

3. *Under* $H_1^{'}$*,*
$$\sqrt{n}(\hat{\mathbb{M}}_n^2 - \mathbb{M}^2) \xrightarrow{d} \mathcal{N}(0, 4\sigma^2)$$
where $\sigma^2 = var_{(\psi,X)}[\mathbb{E}_{(\psi',X')}[\psi k(X, X')\psi']]$

Corollary 7.3.2. *The witness function in Equation* (7.6) *can be estimated as*

$$\hat{f}^*(x) = C\frac{1}{n}\sum_i \psi_i k(x_i, x)$$

where C *is an unrelated constant so that* $\int_{\mathcal{X}} f^{*2}(x)dx = 1$*.*

The following proof follows [185]. Let us define the following operator,

$$Mf = \mathbb{E}[\psi f(X)] \tag{E.6}$$

where $f \in \mathcal{F}$. Since $|\mathbb{E}[\psi|X]| < \infty$ almost surely in P_X, M is a bounded linear operator. By Riesz representation theorem, there exists a unique $g \in \mathcal{F}$ such that

$$Mf = \langle f, g\rangle$$

where

$$g = \mathbb{E}[\psi k(X, \cdot)].$$

g is called the conditional moment embedding (CMME) of the CMR, $\mathbb{E}[\psi|X]$, in $\mathcal{F}$ w.r.t. P_X. Therefore, it follows that

$$\mathbb{M}^2 = \sup_{f\in\mathcal{F},\|f\|\leq 1} (\mathbb{E}[\psi f(X)])^2 = \sup_{f\in\mathcal{F},\|f\|\leq 1} \langle f, g\rangle^2 = \left\langle \frac{g}{\|g\|}, g \right\rangle^2 = \|g\|^2$$

Note that the above implies that the witness function $f^* = \arg\sup_{f\in\mathcal{F},\|f\|\leq 1}(\mathbb{E}[\psi f(X)])^2 = \frac{g}{\|g\|}$. Since g is defined as $\mathbb{E}[\psi k(X, .)]$, it can be empirically estimated as $\frac{1}{n}\sum_{i=1}^{n} \psi_i k(x_i, .)$, which leads to Corollary 7.3.2.

Since $\mathbb{M}^2 = \|g\|^2$, the first statement in Theorem 7.3.1 is essentially

$$\mathbb{E}[\psi|X] = 0, P_X\text{-almost surely} \Leftrightarrow \|g\|^2 = 0$$

That is, $g \in \mathcal{F}$ fully captures the information of the CMR for all $x \in \mathcal{X}$. This equivalence, which we will now prove, is crucial since our statistical test is based on $\|g\|^2$ and its estimates, while Proposition 7.3.3 is directed to the CMR:

($\Rightarrow$) We note that since $\mathcal{F}$ is a Hilbert space, it follows that $g \in \mathcal{F}$, and $\forall f \in \mathcal{F}, \langle f, g\rangle = \mathbb{E}[\psi f(X)] = 0$ (Proposition 7.3.3). g can now only be a zero vector. Therefore, $\|g\|^2 = 0$.

($\Leftarrow$)

$$\begin{aligned}
& \|g\|^2 = 0 \\
\Rightarrow\ & \|\mathbb{E}[\psi k(X, .)]\|^2 = 0 \\
\Rightarrow\ & \|\mathbb{E}[\mathbb{E}[\psi|X]k(X, .)]\|^2 = 0 \\
\Rightarrow\ & \left\| \int_{\mathcal{X}} k(x, .)\mathbb{E}[\psi|x]p_X(x)dx \right\|^2 = 0 \\
\Rightarrow\ & \iint_{\mathcal{X}\times\mathcal{X}} p_X(x)\mathbb{E}[\psi|x]k(x, x^{')\mathbb{E}[\psi|x']p_X(x^{')dxdx'=0}} \\
\Rightarrow\ & \|\mathbb{E}[\psi|x]p_X(x)\|^2 = 0 \qquad (\because k(\cdot, \cdot) \text{ is ISPD})
\end{aligned}$$

$$\Rightarrow \ \mathbb{E}[\psi|x] = 0, P_X\text{-almost surely}$$

Finally, we move to the second and third statements of Theorem 7.3.1, which define the estimator and its statistical properties. Since $\mathbb{M}^2 = \|g\|^2 = \|\mathbb{E}[\psi k(X,.)]\|^2 = \mathbb{E}[\mathbb{E}[\psi k(X,X')\psi']]$ where (X,ψ) and (X',ψ') are independently and identically distributed, we may use a U-statistic to estimate $\mathbb{M}^2$, which is exactly

$$\hat{\mathbb{M}}_n^2 = \frac{1}{n(n-1)} \sum_{i,j\in\mathcal{I},i\neq j} \psi_i k(x_i,x_j)\psi_j$$

The asymptotic distribution of U-statistics has been investigated intensively in the literature. Specifically, from Section 5.5 of Serfling [244], taking the special case of kernels with two inputs, we have the following lemma:

Lemma E.1.1 ((Serfling))**.** *Given a kernel $h(.,.) : \mathcal{W}\times\mathcal{W} \to \mathbb{R}$ where $\mathbb{E}_{(W,W')}[h(W,W')] = \theta$ and $\mathbb{E}_{(W,W')}[h^2(W,W')] < \infty$, the asymptotic distribution of the U-statistic $U_n = \frac{1}{n(n-1)}\sum_{i\neq j} h(w_i,w_j)$ can be categorized into two cases based on $\zeta_1 = var_W(\mathbb{E}_{W'}[h(W,W')])$:* $\begin{cases} \sqrt{n}(U_n-\theta) \xrightarrow{d} N(0,4\zeta_1), & \zeta_1 > 0 \\ n(U_n-\theta) \xrightarrow{d} \sum_{j=1}^{\infty}\lambda_j(Z_j^2-1), & \zeta_1 = 0 \end{cases}$ *where Z_j are independent standard normal variables and λ_j are the eigenvalues of h, i.e. the solutions for $\mathbb{E}_{W'}[h(W',w)v(w)] - \lambda v(w) = 0$*

Note that if we set $W = (\psi,X)$, $h(W,W') = \psi k(X,X')\psi'$, $\theta = \mathbb{M}^2$, $\zeta_1 = \sigma^2$, the second and third statements of Theorem 7.3.1 holds as long as $\mathbb{M}^2 = 0 \Leftrightarrow \sigma^2 = var_{(\psi,X)}[\mathbb{E}_{(\psi',X')}[\psi k(X,X')\psi']] = 0$, which we will now show:

$(\Rightarrow)$

$$\mathbb{E}_{(\psi',X')}[\psi k(X,X')\psi'] = \langle \psi k(X,.), \mathbb{E}_{(\psi',X')}[\psi' k(X',.)]\rangle = \|\psi k(X,.)\| \left\langle \frac{\psi k(X,.)}{\|\psi k(X,.)\|}, g \right\rangle$$

Now since

$$\frac{\psi k(X,.)}{\|\psi k(X,.)\|} \in \mathcal{F}, \left\| \frac{\psi k(X,.)}{\|\psi k(X,.)\|} \right\| = 1$$

and

$$\mathbb{M}^2 = 0 \Rightarrow \sup_{f\in\mathcal{F},\|f\|\leq 1} \langle f, g\rangle = 0 \Rightarrow \langle f, g\rangle = 0, \forall f \in \mathcal{F}, \|f\| \leq 1$$

We conclude,

$$\mathbb{M}^2 = 0 \Rightarrow \left\langle \frac{\psi k(X,.)}{\|\psi k(X,.)\|}, g \right\rangle = 0 \Rightarrow \mathbb{E}_{(\psi',X')}[\psi k(X,X')\psi'] = 0$$
$$\Rightarrow var_{(\psi,X)}[\mathbb{E}_{(\psi',X')}[\psi k(X,X')\psi']] = 0$$

(⇐)

We first note that $var_{(\psi,X)}(\mathbb{E}_{(\psi',X')}[\psi k(X,X')\psi']) = 0$ implies that $\mathbb{E}_{(\psi',X')}[\psi k(X,X')\psi']$ is a constant $P_{(\psi,X)}$-almost surely. We denote this constant as c so we have

$$\mathbb{E}_{(\psi',X')}[\psi k(X,X')\psi'] = c, P_{(\psi,X)}\text{-almost surely} \tag{E.7}$$

From the definition of ψ, let $X = x^*$ be in the support of the observational study, then

$$\begin{aligned}\mathbb{E}[\psi|S=1,X=x^*] &= \frac{1}{P(S=1|X=x^*)}\mathbb{E}\Bigg[\frac{\mathbf{1}(A=1)(Y-\mu_1(x^*))}{P(A=1|S=1,X=x^*)} \\ &\qquad - \frac{\mathbf{1}(A=0)(Y-\mu_0(x^*))}{P(A=0|S=1,X=x^*)}|S=1,X=x^*\Bigg] \\ &= \frac{1}{P(S=1|X=x^*)}\Bigg[\mathbb{E}[Y-\mu_1(x^*)|A=1,S=1,X=x^*] \\ &\qquad - \mathbb{E}[Y-\mu_0(x^*)|A=0,S=1,X=x^*]\Bigg] \\ &= 0,\end{aligned}$$

where the last equality stems from the definition of μ_1 and μ_0. Now note that

$$\begin{aligned}\mathbb{E}_\psi[\mathbb{E}_{(\psi',X')}[\psi k(X,X')\psi'|S=1,X=x^*]] &= \mathbb{E}_\psi[\mathbb{E}_{(\psi',X')}[\psi k(x^*,X')\psi'|S=1,X=x^*]] \\ &= \mathbb{E}_\psi[\psi|S=1,X=x^*]\mathbb{E}_{(\psi',X')}[k(x^*,X')\psi'] \\ &= 0\cdot\mathbb{E}_{(\psi',X')}[k(x^*,X')\psi'] = 0\end{aligned}$$

But also we have, from (E.7),

$$\mathbb{E}_{\psi}[\mathbb{E}_{(\psi',X')}[\psi k(X,X')\psi' | S=1, X=x^*]] = \mathbb{E}_{\psi}[c] = c$$

Therefore, we have $c = 0$ and thus

$$\mathbb{M}^2 = \mathbb{E}_{(\psi,X)}[\mathbb{E}_{(\psi',X')}[\psi k(X,X')\psi']] = \mathbb{E}_{(\psi,X)}[0] = 0$$

so this side of the arrow is also proven.

E.2 Motivating Empirical Examples of Generalization of Treatment Effects rather than Counterfactual Means

In Figure E-1, we plot data that is publicly available in SPRINT Research Group [253] and Franklin et al. [89]. The former is a randomized trial that reports on outcomes across subgroups, where we observe that subgroups often have larger differences in their baseline outcomes than in their treatment effects. The latter is a study that attempts to replicate ten RCTs using observational data. For each observational study and trial, they report on not only the resulting differences in rates (between treatment and control), but also the marginal rates under each of treatment and control. This is done for both the observational studies and the original RCTs. We can observe that the estimated "treatment effects" tend to be closer together (between the observational studies and RCTs) than the estimated "counterfactual means", such as the marginal rate under control. We can view this empirical example as one where Assumption 7.2.3 approximately holds in practice, i.e. the treatment effect appears to generalize across observational and RCT populations, but the counterfactual means do not.

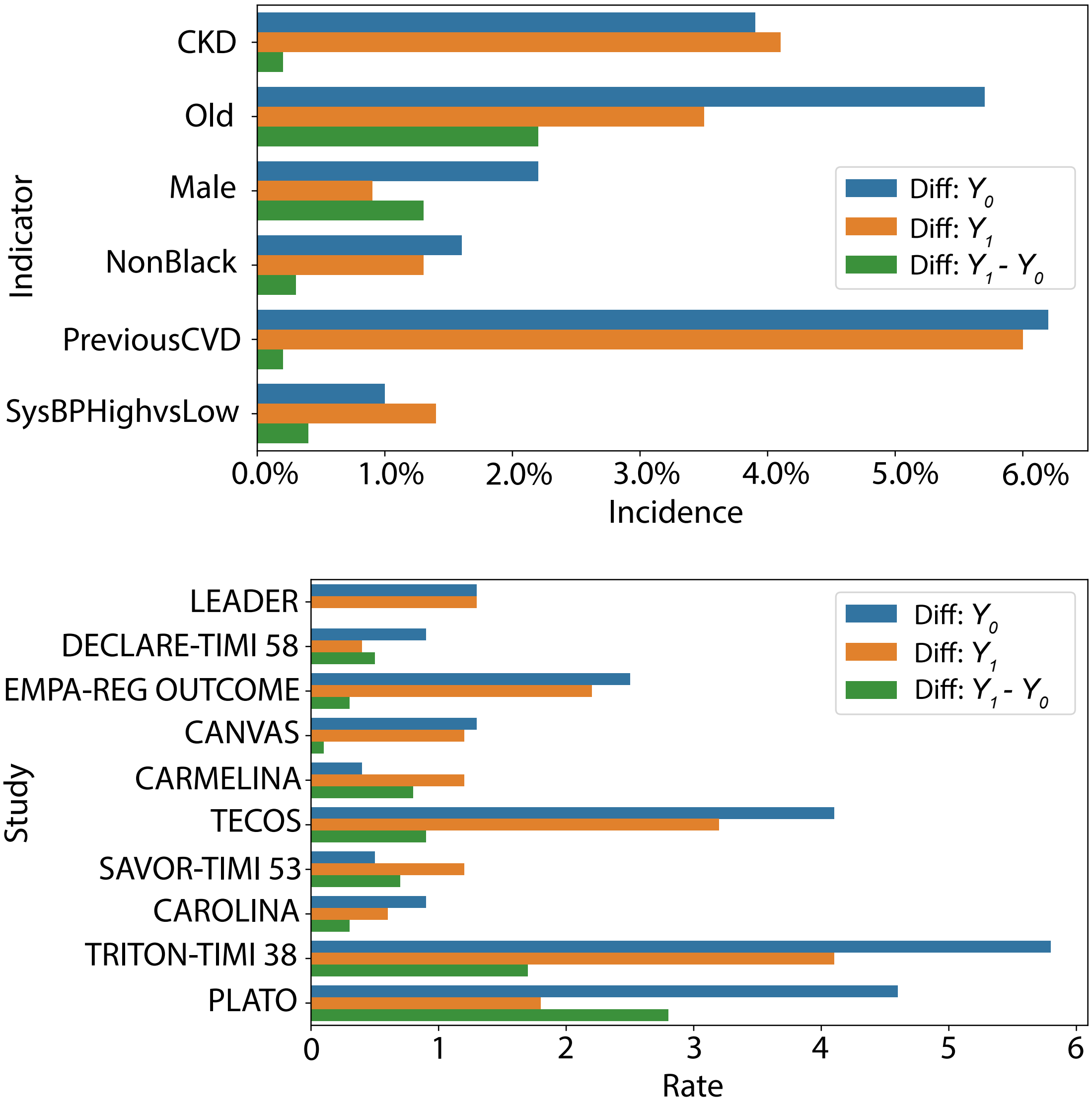


Figure E-1: *Top:* For each binary group indicator I in the SPRINT Trial, we compare the absolute difference between $\mathbb{E}[Y_0 \mid I = 1]$ versus $\mathbb{E}[Y_0 \mid I = 0]$, and similarly for Y_1 and $Y_1 - Y_0$, where Y is a binary variable indicating the observation of the primary composite outcome. The data supporting this plot is taken from Figure 4 of SPRINT Research Group [253]. Generally, the latter difference is smaller (sometimes by an order of magnitude) than the differences for individual potential outcomes. *Bottom:* For each attempted replication of an RCT by an Observational study, we compare the differences in the reported incidence rates under treatment, under the control/comparator, and the difference between the two (analogous to the treatment effect). The latter tends to be smaller than both of the differences in counterfactual means in 6 / 10 replications, and smaller than at least one of the differences in counterfactual means in all 10 replications. This data is taken from Table 2 of Franklin et al. [89], where we use the reported statistics in Table 2.

E.3 IHDP Experiment Details

For both the semi-synthetic and real-world experiments, we follow closely the setup proposed by Hussain et al. [120], with a few differences highlighted below.

E.3.1 Confounder Generation & Outcome Simulation

We generate one RCT and one observational study in each of our 100 simulations, with the randomness appearing in our confounder generation, simulation of the potential outcomes, and the amount of noise in each. In the RCT, we retain the original IHDP data, i.e. the covariates and the binary treatment variable, but resample the dataset with equal probability to generate a final dataset of size, $n_0 = 2955$. For generation of the observational dataset, we first resample the rows of the IHDP dataset to the desired sample size, $n = s \cdot n_0$, but do the resampling in a weighted fashion, such that male infants, infants whose mothers smoked, and infants with working mothers are less prevalent. The weights are set as,

$$w = \frac{1}{1 + \exp(-0.2(\mathbf{1}(\text{male infant}) + \mathbf{1}(\text{mother smoked}) + \mathbf{1}(\text{mother worked during pregnancy})))}$$

Note that this differs from the reweighting scheme used in Hussain et al. [120] in that we use a non-linearity in the reweighting, since we wish for the covariates used in the reweighting (i.e. sex, smoking status, working status) to be effect modifiers.

Next, we generate confounders for the observational dataset. Each confounder, z, is a function of a subset of the covariates, X_s, and the treatment, A:

$$z = X_s^\top \xi + X_s^\top \delta \odot A + \mathcal{N}(0, 1),$$

where $X_s \in \mathbb{R}^4$ and the coefficients ξ and δ are set as: $\xi = (0.1, -0.1, 0.2, -0.3, 0.4)$ and $\delta = (1., -.1, .5, -3, 4)$. X_s consists of the following covariates — ("neonatal health index", "birth order of infant", "drinks alcohol or not", "mother finished high school"). Confounder generation for the RCT is similar but does not include any dependency on the treatment: $X_s^\top \xi + \mathcal{N}(0, 1)$. We repeat this procedure m times to yield m

confounders.

To detail the outcome simulation, we borrow notation from Hussain et al. [120], where we let $Z \in \mathbb{R}^m$ denote the generated confounder vector and $X \in \mathbb{R}^{m_x}$ denote the covariate vector, where $m_x = 28$ is the number of covariates in the original IHDP dataset. Similarly, we let $\tilde{X} = (A, X^\top)^\top$. Then, we set the following counterfactual outcome distributions:

$$\begin{aligned} Y_0 &\sim \mathcal{N}\left(\left(\tilde{X} + \frac{1}{2}\mathbf{1}\right)^\top \beta + Z^\top \gamma, 1\right) \\ Y_1 &\sim \mathcal{N}(\tilde{X}^\top \beta + Z^\top \delta + \omega, 1), \end{aligned}$$

where $\mathbf{1} \in \mathbb{R}^{m_x+1}$ is a vector of ones, $\beta \in \mathbb{R}^{m_x+1}$ is a vector where each element is randomly sampled from $(0, 0.1, 0.2, 0.3, 0.4)$ with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$, and $\gamma \in \mathbb{R}^m$ is a vector where each element is randomly sampled from one of two vectors with uniform probability depending on the strength of confounding desired: $(0.1, 0.2, .5, .75, 1.)$ or $(1., 1.75, 2., 2.25, 2.75)$. In Figure 7-1, for example, we sample from the first vector to generate the confounders, and in Figure 7-2, we sample from the second vector. The observed outcome is then set as, $Y := AY_1 + (1-A)Y_0$. Finally, $\omega = 23$ to bound the magnitude of the counterfactual outcome under treatment. We conceal confounders in order to simulate unobserved confounding, letting c_z be the number of confounders concealed. As alluded to in the main paper, the order of how we conceal confounders is determined by their "confounding strength", i.e. from highest to lowest weighted.

E.4 Women's Health Initiative (WHI) Experiment Details

We follow substantially the same setup as in Hussain et al. [120], which we recounted partially in the main paper (Section 7.5) but do so fully in this section. The WHI conducted several clinical trials as well as an observational study in parallel to study the effect of various hormonal and dietary interventions on the health and quality of life of postmenopausal women. As mentioned in the main paper, of the three clinical trials run by WHI, we use the Postmenopausal Hormone Therapy (PHT) trial for our analysis, which looked at the effect of combination hormone therapy on postmenopausal women aged 50-79 years who had not undergone a hysterectomy [229]. The data used in our analysis is publicly available on BIOLINCC (`https://biolincc.nhlbi.nih.gov/studies/whi_ctos`).

E.4.1 Data

We briefly review the core characteristics of both the RCT and observational study components of the WHI study. The RCT studies the effect of a combination of 2.5mg of medoxyprogesterone and 0.625mg of estrogen on a population of $N_{HT} = 16608$ postmenopausal women. Each patient is randomly assigned to either the treatment group (i.e. estrogen + progesterone combination is given) or the control group, in which the placebo is given. The outcomes tracked in the RCT are of three categories: 1) cardiovascular events, including coronary heart disease, 2) cancers (endometrial, breast, etc.), and 3) fractures (e.g. hip, bone, etc.). The observational study component studies similar outcomes in a cohort of $N = 93676$ women. Women were recruited for this component in 1996, and follow-up was done until 2005, which is a similar timeframe as the RCT. Information about therapies that the patients were taking across the follow-up were tracked via questionnaires, which were taken on a yearly basis.

E.4.2 Outcome and Intervention

As mentioned in Section 7.5 of the main paper, we define a binary outcome based on the "global index" score given to each patient, which is a composite index derived from whether or not a patient experiences any one of the following events: coronary heart disease, stroke, pulmonary embolism, endometrial cancer, colorectal cancer, hip fracture, or death due to other causes. Furthermore, the we let $Y = 1$ if any one of these events is observed in the first seven years of follow-up and $Y = 0$ otherwise. Notably, $Y = 0$ may also occur due to censoring.

In terms of the intervention, the RCT is run as an "intention-to-treat" trial. For the observational study component, we determine treatment and control groups based on explicit affirmation or denial of the use of estrogen and progesterone combination therapy in the first three years, which we glean from the annual survey data. Using this procedure, we end up with a total of $N_{OS} = 33511$ patients. Finally, we restrict the set of covariates used to those that are measured in both the RCT and the observational study. Each covariate indicates the same meaning, since the same set of questionnaires are used to gather them. The resulting number of covariates is 1576.

E.4.3 Experimental Workflow

We detail our experimental setup in this section. We note the following algorithm applies to one row of Table 7.1 in the main paper. Indeed, to get the remaining results, we re-apply this algorithm after "introducing" some selection bias into the observational dataset. We have the following experimental workflow:

- Step 1: Generate B bootstrapped datasets of the base WHI observational dataset.

- Step 2: Set list of r covariate pairs $X_1, \ldots, X_r$. We use the same set of covariates as used by Hussain et al. [120], which can be found in Appendix E of their paper, to generate the r covariate pairs.

- Step 3: For $i = 1 \to B$

  - Apply **MMR-Contrast** (see Appendix E.5 for implementation details). Set $\Lambda_{MMR}[i] = 1$ if p-value is < 0.05, else let $\Lambda_{MMR}[i] = 0$.

  - Apply **ATE**. We use the same estimator as Hussain et al. [120], but average over the entire population to get the ATE from the observational study and RCT, respectively (i.e. set each patient to be part of the same group).

Set $\Lambda_{ATE}[i] = 1$ if the test rejects the null hypothesis and $\Lambda_{ATE}[i] = 0$ otherwise.

- For $j = 1 \to r$
- Apply **GATE** using the four subgroups derived from X_j, as in Hussain et al. [120]. Set $\Lambda_{GATE}[i][j] = 1$ if the test rejects the null hypothesis and $\Lambda_{GATE}[i][j] = 0$ otherwise.

Thus, the rejection rates for **MMR-Contrast**, **ATE**, **GATE** are $\frac{1}{B}\sum_i \Lambda_{MMR}[i]$, $\frac{1}{B}\sum_i \Lambda_{ATE}[i]$, $\frac{1}{B \cdot r}\sum_i \sum_j \Lambda_{GATE}[i][j]$, respectively. We repeat the above workflow to the WHI dataset that has induced selection bias. To add selection bias to the data, with probability p, we drop patients who were not exposed to the intervention and did not experience the event. To obtain the results for Table 7.1, we run our experimental procedure for $p = (0., 0.05, 0.10, 0.15)$.

E.5 Details on Implementation of the MMR-Contrast Method

The implementation of the MMR-Contrast method references the workflow illustrated in [120] and [185]: the former for signal calculation and the latter for significance testing.

E.5.1 Calculation of signal difference

As elaborated in the main text, in the combined data (combining the RCT and observational study), for each observation i producing data (y_i, s_i, a_i, x_i), we define the true signal difference as,

$$\psi_i = \frac{1}{P(S=0|X=x_i)}\{\mathbf{1}(s_i=0)\left[[\mu_1(x_i) - \mu_0(x_i)] - \left[\frac{\mathbf{1}(a_i=1)}{P(A=1|S=0)} - \frac{\mathbf{1}(a_i=0)}{P(A=0|S=0)}\right] y_i\right] + \mathbf{1}(s_i=1)\frac{P(S=0|X=x_i)}{P(S=1|X=x_i)}\left[\frac{\mathbf{1}(a_i=1)(y_i-\mu_1(x_i))}{P(A=1|S=1,X=x_i)} - \frac{\mathbf{1}(a_i=0)(y_i-\mu_0(x_i))}{P(A=0|S=1,X=x_i)}\right]\right\}$$

where $\mu_1(x_i) = \mathbb{E}[Y|S = 0, A = 1, X = x_i]$ and $\mu_0(x_i) = \mathbb{E}[Y|S = 0, A = 0, X = x_i]$. Note that the true signal difference includes several unknown nuisance functions that need to be estimated:

- Response surface: $\mu_1(X), \mu_0(X)$
- Selection propensity: $P(S = 1|X), P(S = 0|X)$
- Treatment propensity in the observational study: $P(A = 1|S = 1, X), P(A = 0|S = 1, X)$
- Treatment propensity in the RCT: $P(A = 1|S = 0), P(A = 0|S = 0)$

The treatment propensity in the RCT is estimated with the empirical probability of treatment within the RCT data. The response surface, selection propensity and treatment propensity in the observational study are estimated using cross-fitting: the combined data is randomly split into $K = 3$ folds, and the nuisance functions used in each fold are estimated with data out of that fold, using the following models with grid search for hyperparameters. Default hyperparameters in *scikit-learn* for the linear regression model were used. The best hyperparameters found for the gradient boosting classifier, also in *scikit-learn*, were as follows: "learning-rate": 0.01, "n-estimators": 50, "max-depth": 2, "min-samples-leaf": 50, "min-samples-split": 50, "max-features": "sqrt" [206].

	Response Surface	Selection Propensity	Treatment Propensity (observational)
IDHP	Linear regression	Gradient Boosting Classifier	Gradient Boosting Classifier
WHI	Gradient Boosting Classifier	Gradient Boosting Classifier	Gradient Boosting Classifier

As an aside, in Figure 7-6(a) where we compare the performance of statistics using estimated signals and true signals, we plug in the response surface and selection propensity model implied by our simulation settings into ψ_i to get the true signal difference.

E.5.2 Hypothesis testing

After obtaining the estimated signal difference $\hat{\psi}_i$ by plugging in the estimated nuisance functions into ψ_i, the test statistic is calculated as,

$$n\hat{\mathbb{M}}_n^2 = \frac{1}{(n-1)} \sum_{i,j \in \mathcal{I}, i \neq j} \hat{\psi}_i k(x_i, x_j) \hat{\psi}_j$$

where $k(.,.)$ is set as a polynomial kernel of order 3. One may also use a laplacian kernel or RBF kernel, although we found the polynomial and laplacian kernels to work best in practice. To obtain the p-value for the test, we follow [185] and generate $B = 100$ samples of multinomials $\mathbf{w}_k = (w_{k1}, w_{k2}, \ldots, w_{kn})^\top \sim \text{Multinom}(n, (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})), k = 1, 2, \ldots, B$. For each k, we define the bootstrap sample of the null distribution:

$$n\hat{\mathbb{M}}_{n(k)}^2 = n \sum_{i,j \in \mathcal{I}, i \neq j} \frac{w_{ki} - 1}{n} \hat{\psi}_i k(x_i, x_j) \hat{\psi}_j \frac{w_{kj} - 1}{n}$$

The p-value is then calculated as

$$\frac{\left[\sum_{k=1}^{B} \mathbf{1}(n\hat{\mathbb{M}}_n^2 \leq n\hat{\mathbb{M}}_{n(k)}^2)\right] + 1}{B + 1}$$

Note that we do not re-estimate the propensity score function in each bootstrap iteration.

E.6 Beyond Testing CATE Signals

In this section, we provide a different formulation of our falsification procedure that tests the potential outcome signals for $\mathbb{E}[Y_a|X], a \in \{0, 1\}$, individually, instead of the signal for the contrast $\mathbb{E}[Y_1 - Y_0|X]$. This demonstrates that our formulation can be adapted to testing other functions of the potential outcome distribution, other than the one we originally considered.

First, we modify our external validity assumption to accommodate testing individual potential outcomes:

Assumption E.6.1 (*External Validity: Observational Study to RCT Transportability of Potential Outcomes*)**.** We assume the following:

- *Mean Exchangeability* — $\mathbb{E}[Y_a|X=x] = \mathbb{E}[Y_a|X=x, S=s]$, $\forall x \in \mathcal{X}$, $\forall s \in \{0,1\}$, and $\forall a \in \{0,1\}$.
- *Positivity of Selection* — $\mathbb{P}(X=x|S=0) > 0 \implies \mathbb{P}(X=x|S=1) > 0$, $\forall x \in \mathcal{X}$.

Now, we will introduce additional notation for our signal functions. Namely, we have the outcome signal from the RCT as follows,

$$\begin{aligned}\psi_0^a &= \frac{\mathbf{1}\{S=0\}}{P(S=0|X)} \cdot \frac{Y\mathbf{1}\{A=a\}}{P(A=a|S=0)}, a \in \{0,1\} \\ \boldsymbol{\psi}_0 &= (\psi_0^0, \psi_0^1)^\top \end{aligned} \tag{E.8}$$

Similarly, we have the following outcome signal in the RCT population, but estimated from observational data, as developed in the main paper,

$$\begin{aligned}\psi_1^a = \frac{1}{P(S=0|X)}\left[\mathbf{1}\{S=0\}\mu_a(X) + \mathbf{1}\{S=1\}\frac{P(S=0|X)}{P(S=1|X)}\frac{\mathbf{1}\{A=a\}(Y-\mu_a(X))}{P(A=a|S=1,X)}\right], a \in \{0,1\} \\ \boldsymbol{\psi}_1 = (\psi_1^0, \psi_1^1)^\top \end{aligned} \tag{E.9}$$

Note that the main difference here compared to the main paper is that we define signal functions individually for each potential outcome and then let $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ be a vector of signals. Now, we have the following proposition, which shows that the vector signals are unbiased for the potential outcomes in the RCT population:

Proposition E.6.1 (*Potential Outcome Signals from the RCT and Observational Data*)**.** *Under assumption 7.2.2 (internal validity of the RCT), the instance-wise potential outcome vector* $\boldsymbol{\psi}_0$ *in Equation* (E.8)*, which uses the outcome information from the RCT, is unbiased, i.e.,* $\mathbb{E}[\boldsymbol{\psi}_0|X] = \mathbb{E}[\mathbf{Y}|X, S=0] = \mathbb{E}[(Y_0, Y_1)^\top|X, S=0]$*. Furthermore, under assumption 7.2.1 and assumption E.6.1, the instance-wise potential outcome vector* $\boldsymbol{\psi}_1$ *in Equation* (E.9)*, which uses the outcome information from the observational data, is unbiased for the potential outcomes in the RCT population, i.e.* $\mathbb{E}[\boldsymbol{\psi}_1|X] = \mathbb{E}[\mathbf{Y}|X, S=0] = \mathbb{E}[(Y_0, Y_1)^\top|X, S=0]$.

Proof. We first show $\mathbb{E}[\boldsymbol{\psi}_0|X] = \mathbb{E}[(Y_0, Y_1)^\top|X, S = 0]$, i.e. $\mathbb{E}[\psi_0^a|X] = \mathbb{E}[Y_a|X, S = 0], a \in \{0, 1\}$

$$\begin{aligned}
\mathbb{E}[\psi_0^a|X] &= \mathbb{E}\Big[\frac{\mathbf{1}\{S=0\}}{P(S=0|X)}\frac{Y\mathbf{1}\{A=a\}}{P(A=a|S=0)}\Big|X\Big] \\
&= \frac{1}{P(S=0|X)P(A=a|S=0)}\mathbb{E}[\mathbf{1}\{S=0, A=a\}Y|X] \\
&= \frac{1}{P(S=0|X)P(A=a|S=0,X)}\mathbb{E}[\mathbf{1}\{S=0, A=a\}Y|X] \qquad \text{(E.10)} \\
&= \frac{P(S=0, A=a|X)}{P(S=0|X)P(A=a|S=0,X)}\mathbb{E}[Y|X, S=0, A=a] \\
&= \mathbb{E}[Y|X, S=0, A=a] \\
&= \mathbb{E}[Y_a|X, S=0], \qquad \text{(E.11)}
\end{aligned}$$

where (E.10) is from fixed probability of assignment in Assumption 7.2.2, and (E.11) is from consistency and ignorability in Assumption 7.2.2.

We then show $\mathbb{E}[\boldsymbol{\psi}_1|X] = \mathbb{E}[(Y_0, Y_1)^\top|X, S = 1]$, i.e. $\mathbb{E}[\psi_1^a|X] = \mathbb{E}[Y_a|X, S = 1], a \in \{0, 1\}$

$$\begin{aligned}
\mathbb{E}[\psi_1^a|X] &= \mathbb{E}\left[\frac{1}{P(S=0|X)}\left[\mathbf{1}\{S=0\}\mu_a(X) + \mathbf{1}\{S=1\}\frac{P(S=0|X)}{P(S=1|X)}\frac{\mathbf{1}\{A=a\}(Y-\mu_a(X))}{P(A=a|S=1,X)}\right]\middle|X\right] \\
&= \frac{\mathbb{E}[\mathbf{1}\{S=0\}|X]\mu_a(X)}{P(S=0|X)} + \frac{\mathbb{E}[\mathbf{1}\{S=1, A=a\}(Y-\mu_a(X))|X]}{P(S=1|X)P(A=a|S=1,X)} \\
&= \frac{P(S=0|X)\mu_a(X)}{P(S=0|X)} + \frac{P(S=1, A=a|X)\mathbb{E}[(Y-\mu_a(X))|X, S=1, A=a]}{P(S=1|X)P(A=a|S=1,X)} \\
&= \mu_a(X) + \mathbb{E}[(Y-\mu_a(X))|X, S=1, A=a] \\
&= \mu_a(X) + \mathbb{E}[Y|X, S=1, A=a] - \mu_a(X) \\
&= \mu_a(X) + \mathbb{E}[Y_a|X, S=1] - \mu_a(X) \qquad \text{(E.12)} \\
&= \mathbb{E}[Y_a|X, S=1],
\end{aligned}$$

where (E.12) is from consistency and ignorability in Assumption 7.2.1. The proposition is now proven. □

We can show a similar corollary to corollary 7.3.1 in the main paper, where we developed the null hypothesis on the CATE signals. Now, we do so for the potential outcome vector signals. Namely, we have,

Corollary E.6.1 (*Null Hypothesis on Potential Outcome Difference*)**.** *Define $\boldsymbol{\psi} = \boldsymbol{\psi}_1 - \boldsymbol{\psi}_0$ as the instance-wise signal difference between the observational and RCT potential outcome estimates. Then, under the null hypothesis, i.e. under assumptions 7.2.1 and 7.2.2 and assumption E.6.1, we have it that $\mathbb{E}[\boldsymbol{\psi}|X] = \mathbf{0}$.*

Proof. If assumptions 7.2.1 and 7.2.2 and assumption E.6.1 hold, then Proposition E.6.1 implies that $\mathbb{E}[\boldsymbol{\psi}_0|X] = \mathbb{E}[\boldsymbol{\psi}_1|X] = \mathbb{E}[\mathbf{Y}|X, S = 0] = \mathbb{E}[(Y_0, Y_1)^\top|X, S = 0]$. □

Our assumptions, i.e. assumption 7.2.2, assumption 7.2.1, and assumption E.6.1, give us a set of conditional moment restrictions (CMRs) on the signal difference, $\boldsymbol{\psi}$, which is a difference of vector signals:

$$H_0 : \mathbb{E}[\boldsymbol{\psi}|X] = \mathbf{0} \quad P_X\text{-almost surely} \tag{E.13}$$

As before, P_X is the distribution of X on the joint distribution of the RCT and observational study. By the law of iterated expectations, akin to the development in Proposition 7.3.3, Equation (E.13) implies an infinite set of unconditional moment restrictions,

$$\mathbb{E}[\boldsymbol{\psi}^\top \boldsymbol{f}(X)] = 0, \forall \boldsymbol{f} \in \mathcal{F} \times \mathcal{F}, \tag{E.14}$$

where $\mathcal{F}$ is the set of measurable functions on $\mathcal{X}$. Note that now, $\boldsymbol{f}$ is a vector-valued function, where $\boldsymbol{f}(X) = (f_0(X), f_1(X))^\top$. Now, as in the main paper, we follow the CMR testing procedure presented in Muandet et al. [185], where we let $\mathcal{F}$ be a RKHS and use the maximum moment restriction (MMR) within the unit ball of the RKHS as our test statistic. Following this, we present the following theorem, which is a modified version of Theorem 7.3.1.

Theorem E.6.1 (*Maximum Moment Restriction-based test for Potential Outcomes*)**.** *Let $\mathcal{F}$ be a RKHS with reproducing kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is ISPD, continuous and bounded, equipped with inner product $\langle .,. \rangle_{\mathcal{F}}$. Denote $\mathcal{F}^2$ as the product RKHS $\mathcal{F} \times \mathcal{F}$ equipped with inner product $\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{F}^2} = \langle (f_1, f_2)^\top, (g_1, g_2)^\top \rangle_{\mathcal{F}^2} :=$*

$\langle f_1, g_1\rangle_{\mathcal{F}} + \langle f_2, g_2\rangle_{\mathcal{F}}$. Suppose the elements of $|\mathbb{E}[\boldsymbol{\psi}|X]| < \infty$ almost surely in P_X, and $\mathbb{E}[[k(X,X')\boldsymbol{\psi}^\top\boldsymbol{\psi}']^2] < \infty$ where $(\boldsymbol{\psi}', X')$ is an independent copy of $(\boldsymbol{\psi}, X)$. Let $\mathbb{M}^2 = \sup_{\boldsymbol{f}\in\mathcal{F}^2, ||\boldsymbol{f}||\leq 1}(\mathbb{E}[\boldsymbol{\psi}^\top\boldsymbol{f}(X)])^2$. Then,

1. *The conditional moment testing problem in Eq. E.13 can be reformulated in terms of the MMR as $H_0^{':\mathbb{M}^2=0}$, $H_1^{':\mathbb{M}^2\neq 0}$.*

Further, let the test statistic be the empirical estimate of $\mathbb{M}^2$,

$$\hat{\mathbb{M}}_n^2 = \frac{1}{n(n-1)}\sum_{i,j\in\mathcal{I}, i\neq j} k(x_i, x_j)\boldsymbol{\psi}_i^\top\boldsymbol{\psi}_j$$

2. *Then, under $H_0^{'}$,*

   $$n\hat{\mathbb{M}}_n^2 \xrightarrow{d} \sum_{j=1}^{\infty}\lambda_j(Z_j^2 - 1)$$

   where Z_j are independent standard normal variables and λ_j are the eigenvalues for $k(x,x')\boldsymbol{\psi}^\top\boldsymbol{\psi}'$.

3. *Under $H_1^{'}$,*

   $$\sqrt{n}(\hat{\mathbb{M}}_n^2 - \mathbb{M}^2) \xrightarrow{d} \mathcal{N}(0, 4\sigma^2)$$

   where $\sigma^2 = var_{(\boldsymbol{\psi},X)}[\mathbb{E}_{(\boldsymbol{\psi}',X')}[k(X,X')\boldsymbol{\psi}^\top\boldsymbol{\psi}']]$

Proof. The proof is very similar to how we proved Therorem 7.3.1. Let us define the following operator,

$$M\boldsymbol{f} = \mathbb{E}[\boldsymbol{\psi}^\top\boldsymbol{f}(X)] \tag{E.15}$$

where $\boldsymbol{f}\in\mathcal{F}^2$. Since the elements of $|\mathbb{E}[\boldsymbol{\psi}|X]| < \infty$ almost surely in P_X, M is a bounded linear operator. By Riesz representation theorem, there exists a unique $\boldsymbol{g}\in\mathcal{F}^2$ such that

$$M\boldsymbol{f} = \langle\boldsymbol{f}, \boldsymbol{g}\rangle_{\mathcal{F}^2}$$

where

$$\boldsymbol{g} = \mathbb{E}[\boldsymbol{\psi}k(X,\cdot)].$$

Therefore, it follows that

$$\mathbb{M}^2 = \sup_{\boldsymbol{f}\in\mathcal{F}^2, \|\boldsymbol{f}\|\leq 1} \left(\mathbb{E}[\boldsymbol{\psi}^\top \boldsymbol{f}(X)]\right)^2 = \sup_{\boldsymbol{f}\in\mathcal{F}^2, \|\boldsymbol{f}\|\leq 1} \langle \boldsymbol{f}, \boldsymbol{g}\rangle^2_{\mathcal{F}^2} = \left\langle \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|}, \boldsymbol{g}\right\rangle^2_{\mathcal{F}^2} = \|\boldsymbol{g}\|^2$$

Since $\mathbb{M}^2 = \|\boldsymbol{g}\|^2$, the first statement in Theorem E.6.1 is essentially

$$\mathbb{E}[\boldsymbol{\psi}|X] = \mathbf{0}, P_X\text{-almost surely} \Leftrightarrow \|\boldsymbol{g}\|^2 = 0$$

That is, $\boldsymbol{g} \in \mathcal{F}^2$ fully captures the information of the CMR for all $x \in \mathcal{X}$. This equivalence, which we will now prove, is crucial since our statistical test is based on $\|\boldsymbol{g}\|^2$ and its estimates, while Corollary E.6.1 is directed to the CMR:

($\Rightarrow$) We note that since $\mathcal{F}^2$ is a Hilbert space, it follows that $\boldsymbol{g} \in \mathcal{F}^2$, and from (E.14), $\forall \boldsymbol{f} \in \mathcal{F}^2, \langle \boldsymbol{f}, \boldsymbol{g}\rangle_{\mathcal{F}^2} = \mathbb{E}[\boldsymbol{\psi}^\top \boldsymbol{f}(X)] = 0$. $\boldsymbol{g}$ can now only be a zero vector. Therefore, $\|\boldsymbol{g}\|^2 = 0$.

($\Leftarrow$)

$$\begin{aligned}
& \|\boldsymbol{g}\|^2 = 0 \\
\Rightarrow \; & \|\mathbb{E}[\boldsymbol{\psi} k(X, .)]\|^2 = 0 \\
\Rightarrow \; & \|\mathbb{E}[\mathbb{E}[\boldsymbol{\psi}|X] k(X, .)]\|^2 = 0 \\
\Rightarrow \; & \left\| \int_{\mathcal{X}} k(x, .)\mathbb{E}[\boldsymbol{\psi}|x] p_X(x) dx \right\|^2 = 0 \\
\Rightarrow \; & \iint_{\mathcal{X}\times\mathcal{X}} p_X(x)\mathbb{E}[\boldsymbol{\psi}^\top|x] k(x, x^{')\mathbb{E}[\boldsymbol{\psi}|x'] p_X(x^{'}) dx dx' = 0} \\
\Rightarrow \; & \|\mathbb{E}[\boldsymbol{\psi}|x] p_X(x)\|^2 = 0 \qquad (\because k(\cdot,\cdot) \text{ is ISPD}) \\
\Rightarrow \; & \mathbb{E}[\boldsymbol{\psi}|x] = \mathbf{0}, P_X\text{-almost surely}
\end{aligned}$$

Finally, we move to the second and third statements of Theorem E.6.1, which define the estimator and its statistical properties. Since $\mathbb{M}^2 = \|\boldsymbol{g}\|^2 = \|\mathbb{E}[\boldsymbol{\psi} k(X, .)]\|^2 = \mathbb{E}[\mathbb{E}[\boldsymbol{\psi}^\top k(X, X')\boldsymbol{\psi}']]$ where $(X, \boldsymbol{\psi})$ and $(X', \boldsymbol{\psi}')$ are independently and identically distributed, we may use a U-statistic to estimate $\mathbb{M}^2$, which is exactly

$$\hat{\mathbb{M}}_n^2 = \frac{1}{n(n-1)} \sum_{i,j \in \mathcal{I}, i \neq j} \boldsymbol{\psi}_i^\top k(x_i, x_j) \boldsymbol{\psi}_j$$

Now note that in Lemma E.1.1, if we set $W = (\boldsymbol{\psi}, X)$, $h(W, W') = \boldsymbol{\psi}^\top k(X, X')\boldsymbol{\psi}'$, $\theta = \mathbb{M}^2$, $\zeta_1 = \sigma^2$, the second and third statements of Theorem 7.3.1 holds as long as $\mathbb{M}^2 = 0 \Leftrightarrow \sigma^2 = var_{(\boldsymbol{\psi},X)}[\mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}^\top k(X,X')\boldsymbol{\psi}']] = 0$, which we will now show:

$(\Rightarrow)$

$$\mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}^\top k(X,X')\boldsymbol{\psi}'] = \langle \boldsymbol{\psi} k(X,.), \mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}' k(X',.)]\rangle_{\mathcal{F}^2} = \|\boldsymbol{\psi} k(X,.)\| \left\langle \frac{\boldsymbol{\psi} k(X,.)}{\|\boldsymbol{\psi} k(X,.)\|}, \boldsymbol{g} \right\rangle_{\mathcal{F}^2}$$

Now since

$$\frac{\boldsymbol{\psi} k(X,.)}{\|\boldsymbol{\psi} k(X,.)\|} \in \mathcal{F}^2, \left\| \frac{\boldsymbol{\psi} k(X,.)}{\|\boldsymbol{\psi} k(X,.)\|} \right\| = 1$$

and

$$\mathbb{M}^2 = 0 \Rightarrow \sup_{\boldsymbol{f} \in \mathcal{F}^2, \|\boldsymbol{f}\| \leq 1} \langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{F}^2} = 0 \Rightarrow \langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{F}^2} = 0, \forall \boldsymbol{f} \in \mathcal{F}^2, \|\boldsymbol{f}\| \leq 1$$

We conclude

$$\begin{aligned} \mathbb{M}^2 = 0 \Rightarrow \left\langle \frac{\boldsymbol{\psi} k(X,.)}{\|\boldsymbol{\psi} k(X,.)\|}, \boldsymbol{g} \right\rangle_{\mathcal{F}^2} &= 0 \Rightarrow \mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}^\top k(X,X')\boldsymbol{\psi}'] = 0 \\ \Rightarrow var_{(\boldsymbol{\psi},X)}[\mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}^\top k(X,X')\boldsymbol{\psi}']] &= 0 \end{aligned}$$

$(\Leftarrow)$

We first note that $var_{(\boldsymbol{\psi},X)}(\mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}^\top k(X,X')\boldsymbol{\psi}']) = 0$ implies that $\mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}^\top k(X,X')\boldsymbol{\psi}']$ is a constant $P_{(\boldsymbol{\psi},X)}$-almost surely. We denote this constant as c so we have

$$\mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}^\top k(X,X')\boldsymbol{\psi}'] = c, P_{(\boldsymbol{\psi},X)}\text{-almost surely} \tag{E.16}$$

From the definition of $\boldsymbol{\psi}$, let $X = x^*$ be in the support of the observational study, then

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\psi}|S=1,X=x^*] &= \frac{1}{P(S=1|X=x^*)}\mathbb{E}\Bigg[\Bigg(\frac{\mathbf{1}(A=1)(Y-\mu_1(x^*))}{P(A=1|S=1,X=x^*)}, \\
&\qquad \frac{\mathbf{1}(A=0)(Y-\mu_0(x^*))}{P(A=0|S=1,X=x^*)}\Bigg)^\top \Bigg| S=1,X=x^*\Bigg] \\
&= \frac{1}{P(S=1|X=x^*)}(\mathbb{E}[Y-\mu_1(x^*)|A=1,S=1,X=x^*], \\
&\qquad \mathbb{E}[Y-\mu_0(x^*)|A=0,S=1,X=x^*)])^\top \\
&= \mathbf{0},
\end{aligned}$$

where the last equality stems from the definition of μ_1 and μ_0. Now note that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\psi}}[\mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}^\top k(X,X')\boldsymbol{\psi}'|S=1,X=x^*]] &= \mathbb{E}_{\boldsymbol{\psi}}[\mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}^\top k(x^*,X')\boldsymbol{\psi}'|S=1,X=x^*]] \\
&= (\mathbb{E}_{\boldsymbol{\psi}}[\boldsymbol{\psi}|S=1,X=x^*])^\top \mathbb{E}_{(\boldsymbol{\psi}',X')}[k(x^*,X')\boldsymbol{\psi}'] \\
&= \mathbf{0}^\top \mathbb{E}_{(\boldsymbol{\psi}',X')}[k(x^*,X')\boldsymbol{\psi}'] = 0
\end{aligned}$$

But also we have, from (E.16),

$$\mathbb{E}_{\boldsymbol{\psi}}[\mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}^\top k(X,X')\boldsymbol{\psi}'|S=1,X=x^*]] = \mathbb{E}_{\boldsymbol{\psi}}[c] = c$$

Therefore, we have $c=0$ and thus

$$\mathbb{M}^2 = \mathbb{E}_{(\boldsymbol{\psi},X)}[\mathbb{E}_{(\boldsymbol{\psi}',X')}[\boldsymbol{\psi}^\top k(X,X')\boldsymbol{\psi}']] = \mathbb{E}_{(\boldsymbol{\psi},X)}[0] = 0$$

so this side of the arrow is also proven.

□

We label this alternate formulation, where we test on the potential outcomes directly instead of the contrast, as **MMR-Absolute**. We give the rejection rate of **MMR-Absolute** under different amounts of selection bias induced in the WHI dataset in Table E.1. We find that the **MMR-Absolute** approach vastly over-rejects, indicating the utility of testing the causal contrast as opposed to the absolute potential outcomes.

Selection Bias	**MMR-Contrast**	**MMR-Absolute**	**ATE**	**GATE**
$p = 0$	0.29	1.0	0.32	**0.17**
$p = 0.05$	**0.67**	1.0	0.58	0.40
$p = 0.10$	**0.94**	1.0	0.88	0.67
$p = 0.15$	**1.0**	1.0	0.98	0.91

Table E.1: Rejection rate when introducing different amounts of selection bias into the observational data in WHI study. p stands for the strength of selection introduced in the the data (refer to Section 7.5 for details).

E.7 When does testing for bias across subgroups improve power?

In our experimental results, we find that the GATE approach has limited power compared to the ATE approach. Indeed, the performance of GATE versus ATE depends in part on the choice of subgroups used for GATE. In the extreme, if the difference in effect is identical across all subgroups, testing for differences in ATE may have higher power once multiple-testing corrections are applied. To build intuition, we will provide a simple example for when a GATE-based test might have higher power compared to an ATE-based test. We will then formalize this example and provably show under what conditions a GATE-based test would have higher asymptotic power compared to an ATE-based test. When referring to the test that tests differences of GATEs or ATEs, we will use the bold form: **GATE** and **ATE**. When referring to the causal quantity itself, we will simply use GATE and ATE.

Toy Example: To build intuition, we will use a toy example to construct three scenarios in which the asymptotic power between **GATE** and **ATE** may differ. Consider testing whether there is bias in a population, where the null hypothesis is that the population mean is zero. Let there be two subgroups in the population, **G1** and **G2**. Finally, let δ be a term denoting the asymptotic bias. Figure E-2 shows three separate scenarios:

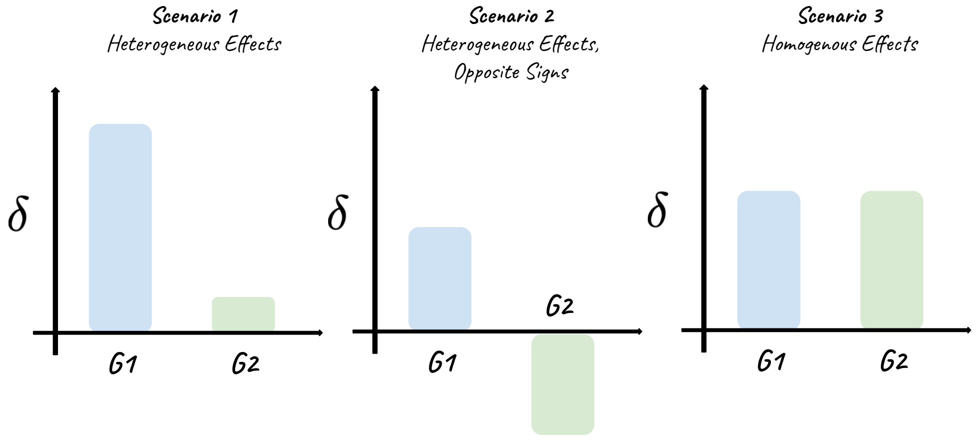


Figure E-2: Barplot depiction of three toy scenarios, where we plot the asymptotic bias, denoted by δ (see Equations (E.25) and (E.26)), of the observational estimator in each subgroup. Our goal is to detect, from finite samples, whether or not this asymptotic bias is non-zero for any subgroup. In scenario 3, pooling the data and testing for the overall bias (the **ATE** approach) yields better power than testing for differences across subgroups. Explicitly testing the bias in each subgroup (the **GATE** approach) is beneficial in scenarios like 1 and 2 where heterogeneity exists. The x-axis contains the group name, and the y-axis indicates the magnitude of δ.

- In scenario 1, the bias in **G1** is significantly higher than the bias in **G2**. As we formalize below, **GATE** will have higher power than **ATE** as $|\delta|$ gets larger and the sample size of **G1** is reasonable. See below for precise conditions.

- In scenario 2, the bias in the two subgroups have the same magnitude but are in opposite directions. Below, we show that **GATE** has better power than **ATE** in this scenario, given a large enough $|\delta|$ to overcome the penalty of multiple hypothesis testing. This result is intuitive since testing differences in ATE would fail to reject the null since the average effect over the entire population would be close to zero.

- In scenario 3, the bias is the same magnitude and direction in both subgroups. We show below that the **ATE** has better power than **GATE** regardless of what the magnitude of δ is. Intuitively, pooling together the two subgroups would yield a larger sample to detect the bias.

In the subsequent paragraphs, we will formalize these three scenarios in the context of our setting, where we have estimates from observational and RCT data. Note that the theoretical framework that we introduce below covers these three scenarios as well as others.

E.7.1 Notation and Assumptions

We recall some notation and definitions from [120].

Definition E.7.1 (GATE, Hussain et al. [120])**.** We define the group average treatment effect (GATE) as

$$\tau_i \coloneqq \mathbb{E}[Y_1 - Y_0 \mid G = i, S = 0] \tag{E.17}$$

where G is the group indicator variable taking values $\{1, 2\}$, and $S = 0$ indicates the RCT population.

The GATE estimator for subgroup i using RCT data will be denoted, $\hat{\tau}_i(0)$, while the estimator using observational data will be denoted, $\hat{\tau}_i(1)$.

Definition E.7.2 (ATE)**.** We define the average treatment effect (ATE) as

$$\tau \coloneqq \mathbb{E}[Y_1 - Y_0 \mid S = 0] \tag{E.18}$$

where $S = 0$ indicates the RCT population.

Akin to the GATE estimators, the ATE estimator using RCT data will be denoted, $\hat{\tau}(0)$, while the estimator using observational data will be denoted, $\hat{\tau}(1)$. Writing ρ_{i0} (ρ_{i1}) as the proportion of observations in the RCT (the obserational study) that belongs to subgroup i, we then modify Assumption 2.4 from Hussain et al. [120] as follows, :

Assumption E.7.1. All GATE estimators are pointwise asymptotically normally distributed and independent

$$\sqrt{\rho_{i0} N_0}(\hat{\tau}_i(0) - \tau_i(0))/\hat{\sigma}_i(0) \xrightarrow{d} \mathcal{N}(0, 1) \tag{E.19}$$

$$\sqrt{\rho_{i1} N_1}(\hat{\tau}_i(1) - \tau_i(1))/\hat{\sigma}_i(1) \xrightarrow{d} \mathcal{N}(0, 1) \tag{E.20}$$

Here, $\xrightarrow{d}$ denotes convergence in distribution, and $\hat{\sigma}_i^2(k)$ is an estimate of the variance that converges in probability to $\sigma_i^2(k)$, the asymptotic variance of $\sqrt{\rho_{ik}N_k}(\hat{\tau}_i(k) - \tau_i(k))$, for $k = 0$ and $k = 1$.

In addition to assumptions on the GATE estimators, we also have assumptions on the asymptotic distributions of the ATE estimators for both studies:

Assumption E.7.2. Both ATE estimators are asymptotically normally distributed and independent

$$\sqrt{N_0}(\hat{\tau}(0) - \tau(0))/\hat{\sigma}(0) \xrightarrow{d} \mathcal{N}(0, 1) \tag{E.21}$$

$$\sqrt{N_1}(\hat{\tau}(1) - \tau(1))/\hat{\sigma}(1) \xrightarrow{d} \mathcal{N}(0, 1) \tag{E.22}$$

where $\hat{\sigma}^2(k)$ is an estimate of the variance that converges in probability to $\sigma^2(k)$, the asymptotic variance of $\sqrt{N_k}(\hat{\tau}(k) - \tau(k))$, for $k = 0$ and $k = 1$.

E.7.2 Theoretical Example

Given the assumptions and definitions, we present a formal example:

Example E.7.1. Suppose there are two subgroups in the RCT and observational study. To reflect the consistency of the RCT GATE estimators and quantify the bias of the GATE estimators from the observational study, we define

$$(RCT, \text{group } 1) \quad \tau_1(0) = \tau_1 \tag{E.23}$$

$$(RCT, \text{group } 2) \quad \tau_2(0) = \tau_2 \tag{E.24}$$

$$(OBS, \text{group } 1) \quad \tau_1(1) = \tau_1 + \delta_1 \tag{E.25}$$

$$(OBS, \text{group } 2) \quad \tau_2(1) = \tau_2 + \delta_2 \tag{E.26}$$

For simplicity, we assume that in both the RCT and observational study, half of the population is in group 1 and half of the population is in group 2, i.e. $\rho_{i0} = \rho_{i1} = 1/2$

for $i = 1, 2$. Then we have

$$\tau(0) = \frac{\tau_1(0) + \tau_2(0)}{2} = \frac{\tau_1 + \tau_2}{2} \tag{E.27}$$

$$\tau(1) = \frac{\tau_1(1) + \tau_2(1)}{2} = \frac{\tau_1 + \delta_1 + \tau_2 + \delta_2}{2} \tag{E.28}$$

Lastly, we introduce the following shorthand notations, writing the total sample size $N = N_0 + N_1$ and letting $N_0 = \rho N$, $N_1 = (1-\rho)N$:

$$\boldsymbol{\sigma} = \sqrt{N_0 + N_1}\sqrt{\frac{\sigma^2(0)}{N_0} + \frac{\sigma^2(1)}{N_1}} = \sqrt{\frac{\sigma^2(0)}{\rho} + \frac{\sigma^2(1)}{1-\rho}} \tag{E.29}$$

$$\boldsymbol{\sigma_1} = \sqrt{\rho_{10}N_0 + \rho_{11}N_1}\sqrt{\frac{\sigma_1^2(0)}{\rho_{10}N_0} + \frac{\sigma_1^2(1)}{\rho_{11}N_1}} = \sqrt{\frac{\sigma_1^2(0)}{\rho} + \frac{\sigma_1^2(1)}{1-\rho}} \tag{E.30}$$

$$\boldsymbol{\sigma_2} = \sqrt{\rho_{20}N_0 + \rho_{21}N_1}\sqrt{\frac{\sigma_2^2(0)}{\rho_{20}N_0} + \frac{\sigma_2^2(1)}{\rho_{21}N_1}} = \sqrt{\frac{\sigma_2^2(0)}{\rho} + \frac{\sigma_2^2(1)}{1-\rho}} \tag{E.31}$$

To simplify the development, we will make the following assumption for this example:

Assumption E.7.3. Assume that $\sigma^2(0) = \sigma_1^2(0) = \sigma_2^2(0)$ and $\sigma^2(1) = \sigma_1^2(1) = \sigma_2^2(1)$, so that we can write Equations (E.29) to (E.31) as,

$$\boldsymbol{\sigma} = \boldsymbol{\sigma_1} = \boldsymbol{\sigma_2} = \sqrt{\frac{\sigma^2(0)}{\rho} + \frac{\sigma^2(1)}{1-\rho}} \tag{E.32}$$

The asymptotic power of the **ATE** and **GATE** can then be given by the following propositions:

Proposition E.7.1 (*Asymptotic power of* ***ATE***)**.** *Under Assumption E.7.3, the asymptotic power of* ***ATE*** *as* $N \to \infty$ *(holding* ρ *as constant) is given by*

$$1 - \left[\Phi\left(\frac{|\frac{\delta_1+\delta_2}{2}|}{\boldsymbol{\sigma}/\sqrt{N}} + z_{\alpha/2}\right) - \Phi\left(\frac{|\frac{\delta_1+\delta_2}{2}|}{\boldsymbol{\sigma}/\sqrt{N}} - z_{\alpha/2}\right)\right]$$

Proof. From Proposition 2.1 of [120], given the asymptotic distributions from Assumption E.7.2 and Equations (E.27), (E.28), we have $\tau(1) - \tau(0) = \frac{\delta_1+\delta_2}{2}$ and thus

$$\frac{\hat{\tau}(1) - \hat{\tau}(0) - \frac{\delta_1+\delta_2}{2}}{\hat{\sigma}/\sqrt{N}} \xrightarrow{d} \mathcal{N}(0,1) \tag{E.33}$$

which allows us to construct a Z-test on the null hypothesis $H_0 : \frac{\delta_1+\delta_2}{2} = 0$ based on the rejection region

$$\left| \frac{\hat{\tau}(1) - \hat{\tau}(0)}{\hat{\sigma}/\sqrt{N}} \right| > z_{\alpha/2}$$

The asymptotic power of the Z-test under the alternative hypothesis distribution shown in (E.33) is then, from Theorems 10.4, 10.6 in [284]

$$1-\Phi\left(\frac{|\frac{\delta_1+\delta_2}{2}|}{\boldsymbol{\sigma}/\sqrt{N}}+z_{\alpha/2}\right)+\Phi\left(\frac{|\frac{\delta_1+\delta_2}{2}|}{\boldsymbol{\sigma}/\sqrt{N}}-z_{\alpha/2}\right) = 1-\left[\Phi\left(\frac{|\frac{\delta_1+\delta_2}{2}|}{\boldsymbol{\sigma}/\sqrt{N}} + z_{\alpha/2}\right) - \Phi\left(\frac{|\frac{\delta_1+\delta_2}{2}|}{\boldsymbol{\sigma}/\sqrt{N}} - z_{\alpha/2}\right)\right]$$

□

Proposition E.7.2 (*Asymptotic power of* ***GATE***)**.** *Under Assumption E.7.3, the asymptotic power of* ***GATE*** *is given by*

$$\begin{aligned} 1 - &\left[\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta_1|}{\boldsymbol{\sigma}/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta_1|}{\boldsymbol{\sigma}/\sqrt{N}} - z_{\alpha/4}\right)\right] \\ &\left[\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta_2|}{\boldsymbol{\sigma}/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta_2|}{\boldsymbol{\sigma}/\sqrt{N}} - z_{\alpha/4}\right)\right] \end{aligned}$$

Proof. With arguments similar to Proposition E.7.1, since the total sample size for subgroup i is $\rho_{i0}N_0 + \rho_{i1}N_1 = N/2$, the asymptotic power of the Z-test comparing the GATE estimates for group i would be ($i \in \{1,2\}$)

$$\begin{aligned} \xi_i &= 1 - \left[\Phi\left(\frac{|\delta_i|}{\boldsymbol{\sigma_i}/\sqrt{N/2}} + z_{\alpha/4}\right) - \Phi\left(\frac{|\delta_i|}{\boldsymbol{\sigma_i}/\sqrt{N/2}} - z_{\alpha/4}\right)\right] \\ &= 1 - \left[\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta_i|}{\boldsymbol{\sigma}/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta_i|}{\boldsymbol{\sigma}/\sqrt{N}} - z_{\alpha/4}\right)\right] \end{aligned}$$

where the last equality stems from Assumption E.7.3. Since we are rejecting the null hypothesis of $H_0 : \delta_1 = 0$ and $\delta_0 = 0$ when the test in either subgroup shows significance, and the two tests are independent, the power of **GATE** is then

$$
\begin{aligned}
&1-(1-\xi_1)(1-\xi_2)\\
=&1-\left[\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta_1|}{\boldsymbol{\sigma}/\sqrt{N}}+z_{\alpha/4}\right)-\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta_1|}{\boldsymbol{\sigma}/\sqrt{N}}-z_{\alpha/4}\right)\right]\\
&\left[\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta_2|}{\boldsymbol{\sigma}/\sqrt{N}}+z_{\alpha/4}\right)-\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta_2|}{\boldsymbol{\sigma}/\sqrt{N}}-z_{\alpha/4}\right)\right]
\end{aligned}
$$

□

We now investigate three scenarios regarding the pattern of bias for the GATE estimators from the observational study:

Scenario 1: Only the GATE estimator for subgroup 1 is biased

This scenario can be depicted by letting $\delta_1=\delta\neq 0$ and $\delta_2=0$, so that we have $\frac{\delta_1+\delta_2}{2}=\frac{\delta}{2}$. The power of **ATE** and **GATE** in this scenario can be given by, based on Propositions E.7.1 and E.7.2:

$$
\begin{aligned}
\xi_{\mathbf{ATE}}&=1-\left[\Phi\left(\frac{|\delta/2|}{\boldsymbol{\sigma}/\sqrt{N}}+z_{\alpha/2}\right)-\Phi\left(\frac{|\delta/2|}{\boldsymbol{\sigma}/\sqrt{N}}-z_{\alpha/2}\right)\right] &\text{(E.34)}\\
\xi_{\mathbf{GATE}}&=1-[\Phi(z_{\alpha/4})-\Phi(-z_{\alpha/4})]\left[\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}}+z_{\alpha/4}\right)-\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}}-z_{\alpha/4}\right)\right]\\
&=1-\left(1-\frac{\alpha}{2}\right)\left[\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}}+z_{\alpha/4}\right)-\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}}-z_{\alpha/4}\right)\right]\\
&=1-\left(1-\frac{\alpha}{2}\right)\left[\Phi\left(\sqrt{2}\frac{|\delta/2|}{\boldsymbol{\sigma}/\sqrt{N}}+z_{\alpha/4}\right)-\Phi\left(\sqrt{2}\frac{|\delta/2|}{\boldsymbol{\sigma}/\sqrt{N}}-z_{\alpha/4}\right)\right] &\text{(E.35)}
\end{aligned}
$$

Denoting $\delta^*:=\frac{|\delta/2|}{\boldsymbol{\sigma}/\sqrt{N}}\geq 0$, we may simplify the expressions as,

$$
\begin{aligned}
\xi_{\mathbf{ATE}}&=1-\left[\Phi(\delta^*+z_{\alpha/2})-\Phi(\delta^*-z_{\alpha/2})\right] &\text{(E.36)}\\
\xi_{\mathbf{GATE}}&=1-\left(1-\frac{\alpha}{2}\right)\left[\Phi\big(\sqrt{2}\delta^*+z_{\alpha/4}\big)-\Phi\big(\sqrt{2}\delta^*-z_{\alpha/4}\big)\right] &\text{(E.37)}
\end{aligned}
$$

Before we derive sufficient conditions for $\xi_{\mathbf{GATE}} > \xi_{\mathbf{ATE}}$, we state the following lemma on the properties of $\Phi(.)$ and $\Phi^{-1}(.)$:

Lemma E.7.1. $\forall \alpha \in (0,1), \frac{z_{\alpha/4}}{z_{\alpha/2}} < \frac{z_{1/4}}{z_{1/2}} \approx 1.1185.$

Lemma E.7.2. *$\forall a > 1$, $\Phi(ax) - \Phi(x)$ is a strictly decreasing function in x as $x > \sqrt{\frac{2\log a}{a^2-1}}$.*

Proof. Taking the derivative of $\Phi(ax) - \Phi(x)$ with respect to x, we have

$$\frac{\partial}{\partial x}\big[\Phi(ax) - \Phi(x)\big] = a\phi(ax) - \phi(x) = a\frac{1}{\sqrt{2\pi}}e^{-\frac{a^2x^2}{2}} - \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\Big[ae^{-\frac{a^2-1}{2}x^2} - 1\Big]$$

When $x > \sqrt{\frac{2\log a}{a^2-1}}$, we have, since $a > 1$,

$$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\Big[ae^{-\frac{a^2-1}{2}x^2} - 1\Big] < \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\Big[ae^{-\frac{a^2-1}{2}\left(\frac{2\log a}{a^2-1}\right)} - 1\Big] = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\Big[a \cdot \frac{1}{a} - 1\Big] = 0$$

Therefore, at $x > \sqrt{\frac{2\log a}{a^2-1}}$, $\Phi(ax) - \Phi(x)$ has strictly negative derivatives which implies it is strictly decreasing. □

Now we may derive the sufficient condition for $\xi_{\mathbf{GATE}} > \xi_{\mathbf{ATE}}$,

$$\begin{aligned}
& \xi_{\mathbf{GATE}} > \xi_{\mathbf{ATE}} \\
\Leftrightarrow\quad & \Phi(\delta^* + z_{\alpha/2}) - \Phi(\delta^* - z_{\alpha/2}) > \Big(1 - \frac{\alpha}{2}\Big)\Big[\Phi\big(\sqrt{2}\delta^* + z_{\alpha/4}\big) - \Phi\big(\sqrt{2}\delta^* - z_{\alpha/4}\big)\Big] \\
\Leftarrow\quad & \Phi(\delta^* + z_{\alpha/2}) - \Phi(\delta^* - z_{\alpha/2}) > \Phi\big(\sqrt{2}\delta^* + z_{\alpha/4}\big) - \Phi\big(\sqrt{2}\delta^* - z_{\alpha/4}\big) \\
\Leftrightarrow\quad & \Phi(\sqrt{2}\delta^* - z_{\alpha/4}) - \Phi(\delta^* - z_{\alpha/2}) > \Phi\big(\sqrt{2}\delta^* + z_{\alpha/4}\big) - \Phi\big(\delta^* + z_{\alpha/2}\big) \\
\Leftarrow\quad & \Phi(\sqrt{2}\delta^* - \sqrt{2}z_{\alpha/2}) - \Phi(\delta^* - z_{\alpha/2}) > \Phi\big(\sqrt{2}\delta^* + \sqrt{2}z_{\alpha/2}\big) - \Phi\big(\delta^* + z_{\alpha/2}\big) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (Lemma\ E.7.1) \\
\Leftrightarrow\quad & \Phi[\sqrt{2}(\delta^* - z_{\alpha/2})] - \Phi[\delta^* - z_{\alpha/2}] > \Phi\big[\sqrt{2}(\delta^* + z_{\alpha/2})\big] - \Phi\big[\delta^* + z_{\alpha/2}\big]
\end{aligned}$$

Since $\delta^* - z_{\alpha/2} < \delta^* + z_{\alpha/2}$, from Lemma E.7.2, the last inequality holds as long as

$\delta^* - z_{\alpha/2} > \sqrt{\frac{2\log(\sqrt{2})}{(\sqrt{2})^2-1}} = \sqrt{\log 2}$. That is, a sufficient condition for $\xi_{\mathbf{GATE}} > \xi_{\mathbf{ATE}}$ is

$$\delta^* > \sqrt{\log 2} + z_{\alpha/2}$$

or, equivalently,

$$|\delta| > \frac{2\boldsymbol{\sigma}}{\sqrt{N}}\big(\sqrt{\log 2} + z_{\alpha/2}\big) \tag{E.38}$$

Intuitively, we see from the above condition that as the magnitude of the bias in subgroup 1 increases or the sample size N increases, **GATE** will eventually have greater power than **ATE**.

Scenario 2: The GATE estimators for both subgroups are biased by the same magnitude but opposite direction

This scenario can be depicted by letting $\delta_1 = \delta$ and $\delta_2 = -\delta$, $\delta \neq 0$, so that we have $\frac{\delta_1+\delta_2}{2} = 0$. Under which the power of **ATE** and **GATE** can be given by, based on Propositions E.7.1 and E.7.2:

$$\xi_{\mathbf{ATE}} = 1 - \big[\Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2})\big] = \alpha \tag{E.39}$$

$$\xi_{\mathbf{GATE}} = 1 - \left[\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}} - z_{\alpha/4}\right)\right]^2 \tag{E.40}$$

We may give a lower bound for $\xi_{\mathbf{GATE}}$:

$$\begin{aligned}\xi_{\mathbf{GATE}} &= 1 - \left[\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}} - z_{\alpha/4}\right)\right]^2 \\ &> 1 - \left[1 - \Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}} - z_{\alpha/4}\right)\right]^2\end{aligned}$$

Therefore, a sufficient condition for $\xi_{\mathbf{GATE}} > \xi_{\mathbf{ATE}}$, i.e. the power of **GATE** to be greater than **ATE** is,

$$|\delta| > \frac{\boldsymbol{\sigma}}{\sqrt{N/2}}(z_{\alpha/4} + \Phi^{-1}(1 - \sqrt{1-\alpha})) \tag{E.41}$$

which can be attained with a large enough bias magnitude $|\delta|$ or large enough sample size N that overcomes the penalty of multiple testing.

Scenario 3: The GATE estimators for both subgroups are biased by the same magnitude and direction

This scenario can be depicted by letting $\delta_1 = \delta_2 = \delta \neq 0$, so that we have $\frac{\delta_1+\delta_2}{2} = \delta$. Under which the power of **ATE** and **GATE** can be given by, based on Propositions E.7.1 and E.7.2:

$$\xi_{\mathbf{ATE}} = 1 - \left[\Phi\left(\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}} + z_{\alpha/2}\right) - \Phi\left(\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}} - z_{\alpha/2}\right)\right] \tag{E.42}$$

$$\xi_{\mathbf{GATE}} = 1 - \left[\Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}} + z_{\alpha/4}\right) - \Phi\left(\frac{1}{\sqrt{2}}\frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}} - z_{\alpha/4}\right)\right]^2 \tag{E.43}$$

Denoting $\delta^* := \frac{|\delta|}{\boldsymbol{\sigma}/\sqrt{N}} \geq 0$, we may simplify the expressions as,

$$\xi_{\mathbf{ATE}} = 1 - \left[\Phi\Big(\delta^* + z_{\alpha/2}\Big) - \Phi\Big(\delta^* - z_{\alpha/2}\Big)\right] \tag{E.44}$$

$$\xi_{\mathbf{GATE}} = 1 - \left[\Phi\Big(\frac{1}{\sqrt{2}}\delta^* + z_{\alpha/4}\Big) - \Phi\Big(\frac{1}{\sqrt{2}}\delta^* - z_{\alpha/4}\Big)\right]^2 \tag{E.45}$$

Therefore, the condition for $\xi_{\mathbf{ATE}} > \xi_{\mathbf{GATE}}$ is equivalent to

$$g(\delta^*) := \left[\Phi\Big(\frac{1}{\sqrt{2}}\delta^* + z_{\alpha/4}\Big) - \Phi\Big(\frac{1}{\sqrt{2}}\delta^* - z_{\alpha/4}\Big)\right]^2 - \left[\Phi\Big(\delta^* + z_{\alpha/2}\Big) - \Phi\Big(\delta^* - z_{\alpha/2}\Big)\right] > 0 \tag{E.46}$$

A graph for $g(\delta^*)$ with $\alpha = 0.005, 0.01, 0.05, 0.1$ is shown in Figure E-3, which demonstrates that $g(\delta^*) > 0$ is satisfied for any $\delta^* > 0$. Therefore, under the scenario where a common bias is shared across subgroups, the power of **ATE** is greater than **GATE** irrespective of the magnitude of bias.

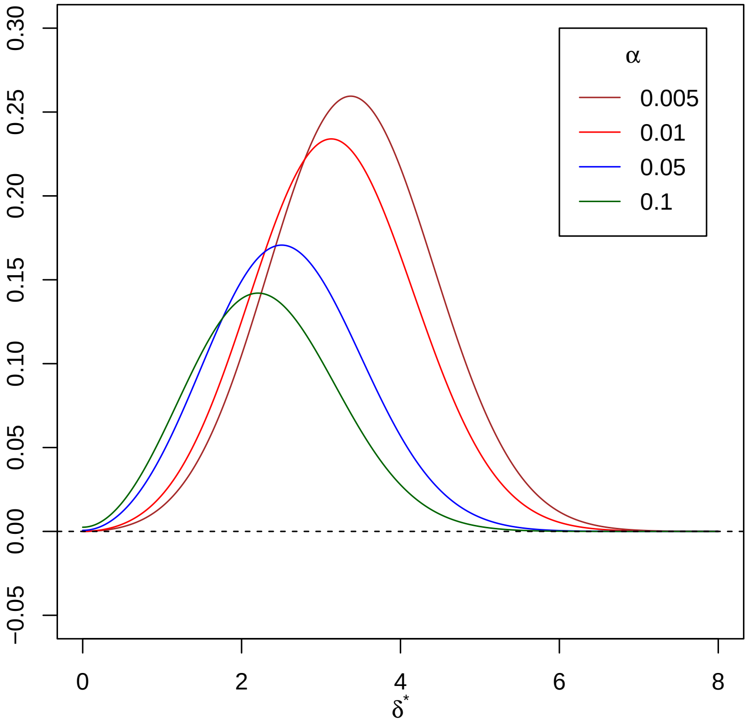


Figure E-3: Plot of function $g(.)$ under $\alpha = 0.005, 0.01, 0.05$ and 0.1

Overall, we find that the relative asymptotic power of **ATE** and **GATE** depends on the homogeneity of bias amongst the subgroups and the magnitude of the bias, and should be analyzed on a case-by-case basis.

Appendix F

Supplementary Material for Chapter 8

Full results for all scenarios are shown below. The statistical tests in the "red pill" row are McNemar's tests done to assess for significant change in proportion of treatment blue selections. All other tests to detect statistically significant changes in confidence and perceived reliability are two-sample paired t-tests. p-values are shown adjusted for multiple hypothesis testing via the Holm-Bonferonni correction. Results of the Shapiro-Wilks tests (done before conducting the 2-sample t-tests) for normality were all non-significant after adjusting for multiple hypothesis testing via the Bonferroni correction.

	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
A	Yes	Yes	Similar to RCT
Red pill	32	32	
Blue pill	0	0	
Confidence	7.34 (+/- 1.19)	7.53 (+/- 1.53) (Tier 1 vs Tier 2: $p = 1.$)	
Reliability	N/A	7.25 (+/- 1.30)	
	Tier 3	**Tier 4**	
	ML model trained on similar patients	**Reproducibility experiment**	
A	Yes	N/A	
Red pill	30 (Tier 1 vs Tier 3: $p = 1$)		
Blue pill	2		
Confidence	7.84 (+/- 1.18) (Tier 2 vs Tier 3: $p = .46$)		
Reliability	7.44 (+/- 1.43) (Tier 2 vs Tier 3: $p = 1.$)		
	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
B	Yes	Yes	Worse with red pill
Red pill	30	23 (Tier 1 vs Tier 2: $p = 0.83$)	
Blue pill	2	9	
Confidence	7.12 (+/- 2.06)	6.03 (+/- 1.94) (Tier 1 vs Tier 2: $p = 0.05$)	
Reliability	N/A	5.88 (+/- 1.60)	
	Tier 3	**Tier 4**	
	ML model trained on similar patients	**Reproducibility experiment**	
B	Yes	N/A	
Red pill	24 (Tier 1 vs Tier 3: $p = 1$)		
Blue pill	8		
Confidence	6.59 (+/- 1.89) (Tier 2 vs Tier 3: $p = .002$)		
Reliability	6.69 (+/- 1.63) (Tier 2 vs Tier 3: $p = .003$)		

Table F.1

	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
C	Yes	No	Similar to RCT
Red pill	32	11 (Tier 1 vs Tier 2: $p = 5e^{-4}$)	
Blue pill	0	21	
Confidence	7.22 (+/- 1.69)	6.28 (+/- 1.55) (Tier 1 vs Tier 2: $p = 0.05$)	
Reliability	N/A	5.88 (+/- 1.45)	
	Tier 3	**Tier 4**	
	ML model trained on similar patients	**Reproducibility experiment**	
C	Yes	Concordant	N/A
Red pill	10 (Tier 1 vs Tier 3: $p = 3e^{-4}$)	9 (Tier 3 vs Tier 4C: $p = 1.$)	
Blue pill	22	23	
Confidence	6.78 (+/- 1.71) (Tier 2 vs Tier 3: $p = .17$)	7.00 (+/- 1.87) (Tier 3 vs Tier 4C: $p = 1.$)	
Reliability	6.59 (+/- 1.56) (Tier 2 vs Tier 3: $p = .004$)	7.09 (+/- 1.81) (Tier 3 vs Tier 4C: $p = .22$)	
	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
D	Yes	No	Worse with red pill
Red pill	32	4 (Tier 1 vs Tier 2: $p = 1.5e^{-5}$)	
Blue pill	0	28	
Confidence	7.28 (+/- 1.57)	6.28 (+/- 1.70) (Tier 1 vs Tier 2: $p = 0.36$)	
Reliability	N/A	6.03 (+/- 1.51)	
	Tier 3	**Tier 4**	
	ML model trained on similar patients	**Reproducibility experiment**	
D	Yes	N/A	
Red pill	3 (Tier 1 vs Tier 3: $p = 9e^{-6}$)		
Blue pill	29		
Confidence	7.03 (+/- 1.86) (Tier 2 vs Tier 3: $p = .06$)		
Reliability	6.59 (+/- 1.75) (Tier 2 vs Tier 3: $p = .43$)		

Table F.2

	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
E	No	Yes	Similar to RCT
Red pill	27	28 (Tier 1 vs Tier 2: $p = 1.$)	
Blue pill	5	4	
Confidence	5.97 (+/- 2.10)	6.50 (+/- 1.95) (Tier 1 vs Tier 2: $p = 0.17$)	
Reliability	N/A	6.00 (+/- 1.56)	
	Tier 3	**Tier 4**	
	ML model trained on similar patients	**Reproducibility experiment**	
E	Yes	Concordant	Non-Concordant
Red pill	27 (Tier 1 vs Tier 3: $p = 1$)	27 (Tier 3 vs Tier 4C: $p = 1.$)	26 (Tier 3 vs Tier 4NC: $p = 1.$)
Blue pill	5	5	6
Confidence	7.16 (+/- 1.73) (Tier 2 vs Tier 3: $p = .008$)	7.56 (+/- 1.82) (Tier 3 vs Tier 4C: $p = 6e^{-4}$)	5.81 (+/- 1.74) (Tier 3 vs Tier 4NC: $p = 6e^{-4}$
Reliability	6.78 (+/- 1.63) (Tier 2 vs Tier 3: $p = .001$)	7.28 (+/- 1.72) (Tier 3 vs Tier 4C: $p = 0.02$)	5.06 (+/- 1.71) (Tier 3 vs Tier 4NC: $p = 5e^{-5}$)
	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
F	No	Yes	Worse with red pill
Red pill	25	15 (Tier 1 vs Tier 2: $p = .27$)	
Blue pill	7	17	
Confidence	6.34 (+/- 2.01)	5.66 (+/- 1.90) (Tier 1 vs Tier 2: $p = 0.04$)	
Reliability	N/A	5.75 (+/- 1.62)	
	Tier 3	**Tier 4**	
	ML model trained on similar patients	**Reproducibility experiment**	
F	Yes	N/A	
Red pill	14 (Tier 1 vs Tier 3: $p = .27$)		
Blue pill	28		
Confidence	6.28 (+/- 1.64) (Tier 2 vs Tier 3: $p = .01$)		
Reliability	6.50 (+/- 1.58) (Tier 2 vs Tier 3: $p = .05$)		

Table F.3

	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
G	No	No	Similar to RCT
Red pill	25	7 (Tier 1 vs Tier 2: $p = .002$)	
Blue pill	7	25	
Confidence	6.12 (+/- 1.75)	5.81 (+/- 1.84) (Tier 1 vs Tier 2: $p = 1.$)	
Reliability	N/A	5.66 (+/- 1.45)	
	Tier 3	**Tier 4**	
	ML model trained on similar patients	**Reproducibility experiment**	
G	Yes	N/A	
Red pill	5 (Tier 1 vs Tier 3: $p = 7e^{-4}$)		
Blue pill	27		
Confidence	6.38 (+/- 1.73) (Tier 2 vs Tier 3: $p = .18$)		
Reliability	6.53 (+/- 1.85) (Tier 2 vs Tier 3: $p = .006$)		
	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
H	No	No	Worse with red pill
Red pill	26	3 (Tier 1 vs Tier 2: $p = 2e^{-4}$)	
Blue pill	6	29	
Confidence	5.75 (+/- 1.85)	6.38 (+/- 1.76) (Tier 1 vs Tier 2: $p = 0.43$)	
Reliability	N/A	5.91 (+/- 1.74)	
	Tier 3	**Tier 4**	
	ML model trained on similar patients	**Reproducibility experiment**	
H	Yes	N/A	
Red pill	1 (Tier 1 vs Tier 3: $p = 7e^{-5}$)		
Blue pill	31		
Confidence	7.56 (+/- 1.14) (Tier 2 vs Tier 3: $p = 2e^{-3}$)		
Reliability	6.97 (+/- 1.47) (Tier 2 vs Tier 3: $p = 2e^{-5}$)		

Table F.4

	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
I	No	Yes	Similar to RCT
Red pill	27	29 (Tier 1 vs Tier 2: $p = 1$)	
Blue pill	5	3	
Confidence	5.84 (+/- 1.97)	6.59 (+/- 1.87) (Tier 1 vs Tier 2: $p = .06$)	
Reliability	N/A	5.97 (+/- 1.86)	

	Tier 3	**Tier 4**	
	ML model trained on similar patients	**Reproducibility experiment**	
I	No	Concordant	Non-Concordant
Red pill	24 (Tier 1 vs Tier 3: $p = 1.$)	25 (Tier 3 vs Tier 4C: $p = 1.$)	21 (Tier 3 vs Tier 4NC: $p = 1.$)
Blue pill	8	7	11
Confidence	5.88 (+/- 1.90) (Tier 2 vs Tier 3: $p = .18$)	5.97 (+/- 2.01) (Tier 3 vs Tier 4C: $p = 1.$)	4.94 (+/- 1.94) (Tier 3 vs Tier 4NC: $p = .05$)
Reliability	4.41 (+/- 2.18) (Tier 2 vs Tier 3: $p = 4e^{-4}$)	4.94 (+/- 2.28) (Tier 3 vs Tier 4C: $p = .17$)	3.91 (+/- 2.08) (Tier 3 vs Tier 4NC: $p = .46$)

	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
J	No	Yes	Worse with blue pill
Red pill	21	28 (Tier 1 vs Tier 2: $p = .44$)	
Blue pill	11	4	
Confidence	6.16 (+/- 2.03)	7.28 (+/- 2.07) (Tier 1 vs Tier 2: $p = 2e^{-3}$)	
Reliability	N/A	5.97 (+/- 2.11)	

	Tier 3	**Tier 4**
	ML model trained on similar patients	**Reproducibility experiment**
J	No	N/A
Red pill	22 (Tier 1 vs Tier 3: $p = 1.$)	
Blue pill	10	
Confidence	5.50 (+/- 1.97) (Tier 2 vs Tier 3: $p = 3e^{-4}$)	
Reliability	4.47 (+/- 1.89) (Tier 2 vs Tier 3: $p = 6e^{-4}$)	

Table F.5

	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
K	No	No	Similar to RCT
Red pill	25	2 (Tier 1 vs Tier 2: $p = 2e - 4$)	
Blue pill	7	30	
Confidence	5.88 (+/- 1.98)	6.38 (+/- 1.54) (Tier 1 vs Tier 2: $p = 1.$)	
Reliability	N/A	5.59 (+/- 1.67)	
	Tier 3	**Tier 4**	
	ML model trained on similar patients	**Reproducibility experiment**	
K	No	N/A	
Red pill	3 (Tier 1 vs Tier 3: $p = 3e^{-4}$)		
Blue pill	29		
Confidence	5.53 (+/- 2.05) (Tier 2 vs Tier 3: $p = .05$)		
Reliability	4.75 (+/- 2.12) (Tier 2 vs Tier 3: $p = .26$)		
	Tier 1	**Tier 2**	
Scenario	**Inclusion criteria for RCT met**	**ML model predicts benefit with red pill**	**ML model's predicted adverse events**
L	No	No	Worse with blue pill
Red pill	21	27 (Tier 1 vs Tier 2: $p = 1.$)	
Blue pill	11	5	
Confidence	5.88 (+/- 2.09)	6.78 (+/- 1.96) (Tier 1 vs Tier 2: $p = 0.03$)	
Reliability	N/A	5.50 (+/- 1.87)	
	Tier 3	**Tier 4**	
	ML model trained on similar patients	**Reproducibility experiment**	
L	No	N/A	
Red pill	18 (Tier 1 vs Tier 3: $p = 1.$)		
Blue pill	14		
Confidence	5.50 (+/- 1.75) (Tier 2 vs Tier 3: $p = .009$)		
Reliability	3.94 (+/- 1.94) (Tier 2 vs Tier 3: $p = 7e^{-4}$)		

Table F.6

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[1] Hammaad Adam, Aparna Balagopalan, Emily Alsentzer, Fotini Christia, and Marzyeh Ghassemi. Mitigating the impact of biased artificial intelligence in emergency decision-making. *Communications Medicine*, 2(1):149, 2022.

[2] Daniel E Adkins. Machine learning and electronic health records: a paradigm shift, 2017.

[3] Ahmed M Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. *Advances in neural information processing systems*, 32, 2019.

[4] Ahmed M Alaa, Zeshan Hussain, and David Sontag. Conformalized unconditional quantile regression. In *International Conference on Artificial Intelligence and Statistics*, pages 10690–10702. PMLR, 2023.

[5] Usama Ahmed Ali, R Joren, Beata M Reiber, Pieter C van der Sluis, and Marc G Besselink. Sample size of surgical randomized controlled trials: a lack of improvement over time. *Journal of Surgical Research*, 228:1–7, 2018.

[6] Alessandro Allegra, Alessandro Tonacci, Raffaele Sciaccotta, Sara Genovese, Caterina Musolino, Giovanni Pioggia, and Sebastiano Gangemi. Machine learning and deep learning applications in multiple myeloma diagnosis, prognosis, and treatment selection. *Cancers*, 14(3):606, 2022.

[7] Anita Ambs, Joan L Warren, Keith M Bellizzi, Marie Topor, Samuel C Chris Haffer, and Steven B Clauser. Overview of the seer—medicare health outcomes survey linked dataset. *Health care financing review*, 29(4):5, 2008.

[8] Brandon Anderson, Truong Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. In *Advances in Neural Information Processing Systems*, pages 14537–14546, 2019.

[9] Andrew Anglemyer, Hacsi T Horvath, and Lisa Bero. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane database of systematic reviews*, (4):MR000034, April 2014.

[10] Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: De-Biased inference of average treatment effects in high dimensions. *arXiv preprint (1604.07125)*, April 2016.

[11] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, April 2019.

[12] Michel Attal, Jean-Luc Harousseau, Anne-Marie Stoppa, Jean-Jacques Sotto, Jean-Gabriel Fuzibet, Jean-François Rossi, Philippe Casassus, Hervé Maisonneuve, Thierry Facon, Norbert Ifrah, et al. A prospective, randomized trial of autologous bone marrow transplantation and chemotherapy in multiple myeloma. *New England Journal of Medicine*, 335(2):91–97, 1996.

[13] Michel Attal, Valerie Lauwers-Cances, Cyrille Hulin, Xavier Leleu, Denis Caillot, Martine Escoffre, Bertrand Arnulf, Margaret Macro, Karim Belhadj, Laurent Garderet, et al. Lenalidomide, bortezomib, and dexamethasone with transplantation for myeloma. *New England Journal of Medicine*, 376(14):1311–1320, 2017.

[14] Hervé Avet-Loiseau, Rafael Fonseca, David Siegel, Meletios A Dimopoulos, Ivan Špička, Tamás Masszi, Roman Hájek, Laura Rosiñol, Vesselina Goranova-Marinova, Georgi Mihaylov, et al. Carfilzomib significantly improves the progression-free survival of high-risk patients in multiple myeloma. *Blood, The Journal of the American Society of Hematology*, 128(9):1174–1180, 2016.

[15] Lindsey R Baden, Hana M El Sahly, Brandon Essink, Karen Kotloff, Sharon Frey, Rick Novak, David Diemert, Stephen A Spector, Nadine Rouphael, C Buddy Creech, et al. Efficacy and safety of the mrna-1273 sars-cov-2 vaccine. *New England journal of medicine*, 2020.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[17] Yu Bai, Song Mei, Huan Wang, Yingbo Zhou, and Caiming Xiong. Efficient and differentiable conformal prediction with general function classes. *arXiv preprint arXiv:2202.11091*, 2022.

[18] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. The road to explainability is paved with bias: Measuring the fairness of explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1194–1206, 2022.

[19] Hailey R Banack, Jay S Kaufman, Jean Wactawski-Wende, Bruce R Troen, and Steven D Stovitz. Investigating and remediating selection bias in geriatrics research: the selection bias toolkit. *Journal of the American Geriatrics Society*, 67(9):1970–1976, 2019.

[20] Xuanwen Bao, Run Shi, Tianyu Zhao, Yanfang Wang, Natasa Anastasov, Michael Rosemann, and Weijia Fang. Integrated analysis of single-cell rna-seq and bulk rna-seq unravels tumour heterogeneity plus m2-like tumour-associated macrophage infiltration and aggressiveness in tnbc. *Cancer Immunology, Immunotherapy*, 70:189–202, 2021.

[21] Henry Barlow, Shunqi Mao, and Matloob Khushi. Predicting high-risk prostate cancer using machine learning methods. *Data*, 4(3):129, 2019.

[22] Brendan J Barnhart, Siddharta G Reddy, and Gerald K Arnold. Remind me again: physician response to web surveys: the effect of email reminders across 11 opinion survey efforts at the american board of internal medicine from 2017 to 2019. *Evaluation & the Health Professions*, 44(3):245–259, 2021.

[23] M Jésus Bayarri and James O Berger. The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

[24] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*, 2011.

[25] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4): 791–806, 2011.

[26] Andrew Bennett and Nathan Kallus. The variational method of moments. *arXiv preprint arXiv:2012.09422*, 2020.

[27] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019.

[28] Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Inference on strongly identified functionals of weakly identified functions. *arXiv e-prints*, pages arXiv–2208, 2022.

[29] Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint (2003.12659)*, March 2020.

[30] Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *ICLR*, 2020.

[31] Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela van der Schaar. From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 2020.

[32] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[33] Katie R Bradwell, Jacob T Wooldridge, Benjamin Amor, Tellen D Bennett, Adit Anand, Carolyn Bremer, Yun Jae Yoo, Zhenglong Qian, Steven G Johnson, Emily R Pfaff, et al. Harmonizing units and values of quantitative data elements in a very large nationally pooled electronic health record (ehr) dataset. *Journal of the American Medical Informatics Association*, 29(7):1172–1182, 2022.

[34] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[35] Paul Braunschweiger and Kenneth W Goodman. The citi program: an international online resource for education in human subjects protection and the responsible conduct of research. *Academic Medicine*, 82(9):861–864, 2007.

[36] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[37] Jeanne Brooks-Gunn, Fong-ruey Liaw, and Pamela Kato Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3):350–359, 1992.

[38] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, 2011.

[39] Patricia B Burns, Rod J Rohrich, and Kevin C Chung. The levels of evidence and their role in evidence-based medicine. *Plastic and reconstructive surgery*, 128(1):305, 2011.

[40] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[41] Giovanni Cammarota, Gianluca Ianiro, Anna Ahern, Carmine Carbone, Andriy Temko, Marcus J Claesson, Antonio Gasbarrini, and Giampaolo Tortora. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature reviews gastroenterology & hepatology*, 17(10):635–648, 2020.

[42] Emmanuel Candes, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1): 24–45, 2023.

[43] Yue Cao, Shila Ghazanfar, Pengyi Yang, and Jean Yang. Benchmarking of analytical combinations for covid-19 outcome prediction using single-cell rna sequencing data. *bioRxiv*, pages 2023–01, 2023.

[44] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):1–12, 2018.

[45] Shuxiao Chen, Bo Zhang, and Ting Ye. Minimax rates and adaptivity in combining experimental and observational data. *arXiv preprint (2109.10522)*, September 2021.

[46] Weiqi Chen, Wenwei Wang, Bingqing Peng, Qingsong Wen, Tian Zhou, and Liang Sun. Learning to rotate: Quaternion transformer for complicated periodical time series forecasting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 146–156, 2022.

[47] Xiaohong Chen and Halbert White. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.

[48] David Cheng and Tianxi Cai. Adaptive combination of randomized and observational data. *arXiv preprint (2111.15012)*, November 2021.

[49] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. *arXiv preprint (1712.04802)*, December 2017.

[50] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The econometrics journal*, 21(1): C1–C68, January 2018.

[51] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48), 2021.

[52] J Anthony Child, Gareth J Morgan, Faith E Davies, Roger G Owen, Susan E Bell, Kim Hawkins, Julia Brown, Mark T Drayson, and Peter J Selby. High-dose chemotherapy with hematopoietic stem-cell rescue for multiple myeloma. *New England Journal of Medicine*, 348(19):1875–1883, 2003.

[53] Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

[54] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[55] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[56] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[57] Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff Schneider. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *Advances in Neural Information Processing Systems*, 34, 2021.

[58] Jordan M Cloyd, Victor Heh, Timothy M Pawlik, Aslam Ejaz, Mary Dillhoff, Allan Tsung, Terence Williams, Laith Abushahin, John FP Bridges, and Heena Santry. Neoadjuvant therapy for resectable and borderline resectable pancreatic cancer: a meta-analysis of randomized controlled trials. *Journal of clinical medicine*, 9(4):1129, 2020.

[59] Stephen R Cole and Elizabeth A Stuart. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American journal of epidemiology*, 172(1):107–115, July 2010.

[60] Noa Dagan, Noam Barda, Eldad Kepten, Oren Miron, Shay Perchik, Mark A Katz, Miguel A Hernán, Marc Lipsitch, Ben Reis, and Ran D Balicer. Bnt162b2 mrna covid-19 vaccine in a nationwide mass vaccination setting. *New England Journal of Medicine*, 2021.

[61] Issa J Dahabreh, Sarah E Robertson, Lucia C Petito, Miguel A Hernán, and Jon A Steingrimsson. Efficient and robust methods for causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a target population. *arXiv preprint (1908.09230)*, August 2019.

[62] Issa J Dahabreh, Sarah E Robertson, Eric J Tchetgen, Elizabeth A Stuart, and Miguel A Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, June 2019.

[63] Issa J Dahabreh, Sarah E Robertson, Eric J Tchetgen, Elizabeth A Stuart, and Miguel A Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, June 2019.

[64] Issa J Dahabreh, Sarah E Robertson, Jon A Steingrimsson, Elizabeth A Stuart, and Miguel A Hernán. Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39(14):1999–2014, June 2020.

[65] Issa J Dahabreh, Sarah E Robertson, Jon A Steingrimsson, Elizabeth A Stuart, and Miguel A Hernan. Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39(14):1999–2014, 2020.

[66] Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *arXiv preprint (2102.11904)*, February 2021.

[67] Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *arXiv preprint arXiv:2102.11904*, 2021.

[68] Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

[69] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.

[70] Meletios A Dimopoulos, Pieter Sonneveld, Nelson Leung, Giampaolo Merlini, Heinz Ludwig, Efstathios Kastritis, Hartmut Goldschmidt, Douglas Joshua, Robert Z Orlowski, Raymond Powles, et al. International myeloma working group recommendations for the diagnosis and management of myeloma-related renal impairment. *Journal of Clinical Oncology*, 34(13):1544–1557, 2016.

[71] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[72] Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. September 2021.

[73] Brian GM Durie, Antje Hoering, Muneer H Abidi, S Vincent Rajkumar, Joshua Epstein, Stephen P Kahanic, Mohan Thakuri, Frederic Reu, Christopher M Reynolds, Rachael Sexton, et al. Bortezomib with lenalidomide and dexamethasone versus lenalidomide and dexamethasone alone in patients with newly diagnosed myeloma without intent for immediate autologous stem-cell transplant (swog s0777): a randomised, open-label, phase 3 trial. *The Lancet*, 389(10068): 519–527, 2017.

[74] David M Eddy, Vic Hasselblad, and Ross Shachter. An introduction to a bayesian method for meta-analysis: the confidence profile method, 1990.

[75] Huseyin Melih Elibol, Vincent Nguyen, Scott Linderman, Matthew Johnson, Amna Hashmi, and Finale Doshi-Velez. Cross-corpora unsupervised learning of trajectories in autism spectrum disorders. *The Journal of Machine Learning Research*, 17(1):4597–4634, 2016.

[76] Srinivas Emani, Angela Rui, Hermano Alexandre Lima Rocha, Rubina F Rizvi, Sergio Ferreira Juaçaba, Gretchen Purcell Jackson, and David W Bates. Physicians' perceptions of and satisfaction with artificial intelligence in cancer treatment: A clinical decision support system experience and implications for low-middle–income countries. *JMIR cancer*, 8(2):e31461, 2022.

[77] David K Eng, Nishith B Khandwala, Jin Long, Nancy R Fefferman, Shailee V Lala, Naomi A Strubel, Sarah S Milla, Ross W Filice, Susan E Sharp, Alexander J Towbin, et al. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: a prospective multicenter randomized controlled trial. *Radiology*, 301(3):692–699, 2021.

[78] K Esfahani, L Roudaia, NA Buhlaiga, SV Del Rincon, N Papneja, and WH Miller. A review of cancer immunotherapy: from the past, to the present, to the future. *Current Oncology*, 27(s2):87–97, 2020.

[79] Thierry Facon, Christopher P Venner, Nizar J Bahlis, Fritz Offner, Darrell J White, Lionel Karlin, Lotfi Benboubker, Sophie Rigaudeau, Philippe Rodon, Eric Voog, et al. Oral ixazomib, lenalidomide, and dexamethasone for transplant-ineligible patients with newly diagnosed multiple myeloma. *Blood*, 137(26): 3616–3628, 2021.

[80] Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *arXiv preprint (1309.4686)*, February 2018.

[81] US FDA et al. Oncology (cancer)/hematologic malignancies approval notifications, 2021.

[82] Shai Feldman, Stephen Bates, and Yaniv Romano. Improving conditional coverage via orthogonal quantile regression. *Advances in Neural Information Processing Systems*, 34, 2021.

[83] Luisa Turrin Fernholz. *Von Mises calculus for statistical functionals*, volume 19. Springer Science & Business Media, 2012.

[84] Sergio Firpo, Nicole M Fortin, and Thomas Lemieux. Unconditional quantile regressions. *Econometrica*, 77(3):953–973, 2009.

[85] Carla S Fisher, Julie A Margenthaler, Kelly K Hunt, and Theresa Schwartz. The landmark series: axillary management in breast cancer. *Annals of surgical oncology*, 27:724–729, 2020.

[86] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[87] Despina Fotiou, Maria Gavriatopoulou, and Evangelos Terpos. Multiple myeloma and thrombosis: prophylaxis and risk prediction tools. *Cancers*, 12(1):191, 2020.

[88] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

[89] Jessica M Franklin, Elisabetta Patorno, Rishi J Desai, Robert J Glynn, David Martin, Kenneth Quinto, Ajinkya Pawar, Lily G Bessette, Hemin Lee, Elizabeth M Garry, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the rct duplicate initiative. *Circulation*, 143(10):1002–1013, 2021.

[90] Joseph Futoma, Mark Sendak, Blake Cameron, and Katherine Heller. Predicting disease progression with a model for multivariate longitudinal clinical data. In *Machine Learning for Healthcare Conference*, pages 42–54, 2016.

[91] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[92] Adriana C Gamboa and Shishir K Maithel. The landmark series: gallbladder cancer. *Annals of Surgical Oncology*, 27:2846–2858, 2020.

[93] Xabier García-Albéniz, John Hsu, and Miguel A Hernán. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *European journal of epidemiology*, 32:495–500, 2017.

[94] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as ai say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1):31, 2021.

[95] Jennifer S Gewandter, Michael P McDermott, Hua He, Shan Gao, Xueya Cai, John T Farrar, Nathaniel P Katz, John D Markman, Stephen Senn, Dennis C Turk, et al. Demonstrating heterogeneity of treatment effects among patients: An overlooked but important step toward precision medicine. *Clinical Pharmacology & Therapeutics*, 106(1):204–210, 2019.

[96] Indranil Ghosh, Rabin K Jana, and Manas K Sanyal. Analysis of temporal pattern, causal interaction and predictive modeling of financial markets using nonlinear dynamics, econometric models and machine learning algorithms. *Applied Soft Computing*, 82:105553, 2019.

[97] Rahul Gopalkrishnan. *Advances in deep generative modeling for clinical data*. PhD thesis, Massachusetts Institute of Technology, 2020.

[98] Devendra Goyal, Donna Tjandra, Raymond Q Migrino, Bruno Giordani, Zeeshan Syed, Jenna Wiens, Alzheimer's Disease Neuroimaging Initiative, et al. Characterizing heterogeneity in the progression of alzheimer's disease using longitudinal clinical and neuroimaging biomarkers. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:629–637, 2018.

[99] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[100] Sander Greenland. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2): 267–306, 2005.

[101] Philip R. Greipp, Jesus San Miguel, Brian G.M. Durie, John J. Crowley, Bart Barlogie, Joan Bladé, Mario Boccadoro, J. Anthony Chile, Hervé Avet-Loiseau, Robert A. Kyle, Juan J. Lahuerta, Heinz Ludwig, Gareth Morgan, Raymond Powles, Kazuyuki Shimizu, Chaim Shustik, Pieter Sonneveld, Patrizia Tosi, Ingemar Turesson, and Jan Westin. International staging system for multiple myeloma. *Journal of Clinical Oncology*, 2005.

[102] International Myeloma Working Group. Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the international myeloma working group. *British journal of haematology*, 121(5): 749–757, 2003.

[103] Leying Guan. Conformal prediction with localization. *arXiv preprint arXiv:1908.08558*, 2019.

[104] Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *arXiv preprint arXiv:2106.08460*, 2021.

[105] Camila Guerrero, Noemi Puig, Maria-Teresa Cedena, Ibai Goicoechea, Cristina Perez, Juan-José Garcés, Cirino Botta, Maria-Jose Calasanz, Norma C Gutierrez, Maria-Luisa Martin-Ramos, et al. A machine learning model based on tumor and immune biomarkers to predict undetectable mrd and survival outcomes in multiple myeloma. *Clinical Cancer Research*, 28(12):2598–2609, 2022.

[106] Yu Gui, Rohan Hore, Zhimei Ren, and Rina Foygel Barber. Conformalized survival analysis with adaptive cutoffs. *arXiv preprint arXiv:2211.01227*, 2022.

[107] Chirag Gupta, Arun K Kuchibhotla, and Aaditya Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, page 108496, 2021.

[108] Gordon H Guyatt, Andrew D Oxman, Regina Kunz, Gunn E Vist, Yngve Falck-Ytter, and Holger J Schünemann. What is "quality of evidence" and why is it important to clinicians? *Bmj*, 336(7651):995–998, 2008.

[109] Gordon H Guyatt, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, and Holger J Schünemann. Grade: an emerging consensus on rating quality of evidence and strength of recommendations. *Bmj*, 336(7650):924–926, 2008.

[110] P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[111] Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society. Series A,*, 178(3):757–778, June 2015.

[112] MA Hernan and J Robins. Causal inference: What if. boca raton: Chapman & hill/crc. 2020.

[113] Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183 (8):758–764, 2016.

[114] Miguel A Hernan and James M Robins. *Causal Inference.* CRC Press, Boca Raton, FL, February 2021.

[115] Miguel A Hernán, Alvaro Alonso, Roger Logan, Francine Grodstein, Karin B Michels, Meir J Stampfer, Walter C Willett, JoAnn E Manson, and James M Robins. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology (Cambridge, Mass.)*, 19(6):766, 2008.

[116] José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pages 1861–1869. PMLR, 2015.

[117] Julian PT Higgins, James Thomas, Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and Vivian A Welch. *Cochrane handbook for systematic reviews of interventions.* John Wiley & Sons, 2019.

[118] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[119] Jiri Hron, Alexander G de G Matthews, and Zoubin Ghahramani. Variational gaussian dropout is not bayesian. *arXiv preprint arXiv:1711.02989*, 2017.

[120] Zeshan Hussain, Michael Oberst, Ming-Chieh Shih, and David Sontag. Falsification before extrapolation in causal effect estimation. *arxiv preprint arXiv:2209.13708*, 2022.

[121] Zeshan Hussain, Ming-Chieh Shih, Michael Oberst, Ilker Demirel, and David Sontag. Falsification of internal and external validity in observational studies via conditional moment restrictions. In *International Conference on Artificial Intelligence and Statistics*, pages 5869–5898. PMLR, 2023.

[122] Zeshan M Hussain, Rahul G Krishnan, and David Sontag. Neural pharmacodynamic state space modeling. In *International Conference on Machine Learning*, pages 4500–4510. PMLR, 2021.

[123] Lucy Hutchinson, Bernhard Steiert, Antoine Soubret, Jonathan Wagg, Alex Phipps, Richard Peck, Jean-Eric Charoin, and Benjamin Ribba. Models and machines: How deep learning will take clinical pharmacology to the next level. *CPT: pharmacometrics & systems pharmacology*, 8(3):131–134, 2019.

[124] Joseph G Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, pages 46–60, 2000.

[125] Joseph G Ibrahim, Haitao Chu, and Liddy M Chen. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796, 2010.

[126] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, April 2015.

[127] Guido W Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

[128] Daniel Jacob. Group average treatment effects for observational studies. *arXiv preprint (1911.02688)*, November 2019.

[129] Maia Jacobs, Melanie F Pradier, Thomas H McCoy Jr, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1):108, 2021.

[130] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[131] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015.

[132] Bastien Jamet, Ludivine Morvan, Cristina Nanni, Anne-Victoire Michaud, Clément Bailly, Stéphane Chauvie, Philippe Moreau, Cyrille Touzeau, Elena Zamagni, Caroline Bodet-Milin, et al. Random survival forest to predict transplant-eligible newly diagnosed multiple myeloma outcome including fdg-pet radiomics: a combined analysis of two independent prospective european trials. *European journal of nuclear medicine and molecular imaging*, 48:1005–1015, 2021.

[133] Kevin B Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L Snowdon. Precision medicine, ai, and the future of personalized health care. *Clinical and translational science*, 14(1): 86–93, 2021.

[134] Gavin S Jones and David R Baldwin. Recent advances in the management of lung cancer. *Clinical Medicine*, 18(Suppl 2):s41, 2018.

[135] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[136] Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects through double machine learning. In *AAAI*. aaai.org, 2021.

[137] Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In S. Bengio and H. Wallach and H. Larochelle and K. Grauman and N. Cesa-Bianchi and R. Garnett, editor, *Advances in Neural Information Processing Systems*, 31. Curran Associates, Inc., 2018.

[138] Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

[139] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9, 2019.

[140] N Keum, DH Lee, DC Greenwood, JE Manson, and E Giovannucci. Vitamin d supplementation and total cancer incidence and mortality: a meta-analysis of randomized controlled trials. *Annals of Oncology*, 30(5):733–743, 2019.

[141] Sepehr Golriz Khatami, Christine Robinson, Colin Birkenbihl, Daniel Domingo-Fernández, Charles Tapley Hoyt, and Martin Hofmann-Apitius. Challenges of integrative disease modeling in alzheimer's disease. *Frontiers in Molecular Biosciences*, 6, 2019.

[142] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[143] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

[144] Christoph A Klein. Parallel progression of primary tumours and metastases. *Nature Reviews Cancer*, 9(4):302–312, 2009.

[145] David G Kleinbaum and Mitchel Klein. *Survival analysis a self-learning text*. Springer, 1996.

[146] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[147] Jan Koutnik, Klaus Greff, Faustino Gomez, and Juergen Schmidhuber. A clockwork rnn. In *International Conference on Machine Learning*, pages 1863–1871, 2014.

[148] Rahul Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[149] Rowan Kuiper, Mark van Duin, Martin H van Vliet, Annemiek Broijl, Bronno van der Holt, Laila El Jarari, Erik H van Beers, George Mulligan, Hervé Avet-Loiseau, Walter M Gregory, et al. Prediction of high-and low-risk multiple myeloma based on gene expression and the international staging system. *Blood, The Journal of the American Society of Hematology*, 126(17):1996–2004, 2015.

[150] Shaji K Kumar, Angela Dispenzieri, Martha Q Lacy, Morie A Gertz, Francis K Buadi, Shivlal Pandey, Prashant Kapoor, David Dingli, Suzanne R Hayman, Nelson Leung, et al. Continued improvement in survival in multiple myeloma: changes in early mortality and outcomes in older patients. *Leukemia*, 28(5): 1122–1128, 2014.

[151] Shaji K Kumar, Natalie S Callander, Melissa Alsina, Djordje Atanackovic, J Sybil Biermann, Jason C Chandler, Caitlin Costello, Matthew Faiman, Henry C Fung, Cristina Gasparetto, et al. Multiple myeloma, version 3.2017, nccn clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, 15(2):230–269, 2017.

[152] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

[153] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[154] Thomas YT Lam, Max FK Cheung, Yasmin L Munro, Kong Meng Lim, Dennis Shung, and Joseph JY Sung. Randomized controlled trials of artificial intelligence in clinical practice: Systematic review. *Journal of Medical Internet Research*, 24 (8):e37188, 2022.

[155] Ola Landgren, David S Siegel, Daniel Auclair, Ajai Chari, Michael Boedigheimer, Tim Welliver, Khalid Mezzi, Karim Iskander, and Andrzej Jakubowiak. Carfilzomib-lenalidomide-dexamethasone versus bortezomib-lenalidomide-dexamethasone in patients with newly diagnosed multiple myeloma: results from the prospective, longitudinal, observational commpass study. *Blood*, 132:799, 2018.

[156] Dirk Larson, Robert A Kyle, and S Vincent Rajkumar. Prevalence and monitoring of oligosecretory myeloma. *The New England journal of medicine*, 367(6): 580–581, 2012.

[157] Yann LeCun. Learning invariant feature hierarchies. In *European conference on computer vision*, pages 496–505. Springer, 2012.

[158] Choong Ho Lee and Hyung-Jin Yoon. Medical big data: promise and challenges. *Kidney research and clinical practice*, 36(1):3, 2017.

[159] Seung-Yeoun Lee and Robert A Wolfe. A simple test for independent censoring under the proportional hazards model. *Biometrics*, pages 1176–1182, 1998.

[160] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 71–96, 2014.

[161] Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

[162] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[163] Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.

[164] Bin Li, Kaili Ren, Lei Shen, Peijie Hou, Zhenqiang Su, Alessandra Di Bacco, Jin-Liern Hong, Aaron Galaznik, Ajeeta B Dash, Victoria Crossland, et al. Comparing bortezomib-lenalidomide-dexamethasone (vrd) with carfilzomib-lenalidomide-dexamethasone (krd) in the patients with newly diagnosed multiple myeloma (ndmm) in two observational studies. *Blood*, 132:3298, 2018.

[165] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32, 2019.

[166] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.

[167] Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 2022.

[168] Bryan Lim, Ahmed Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems*, pages 7483–7493, 2018.

[169] Julia Ling, Andrew Kurzawski, and Jeremy Templeton. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics*, 807:155–166, 2016.

[170] Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, May 2010.

[171] Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*, 2021.

[172] Yu-Ying Liu, Shuang Li, Fuxin Li, Le Song, and James M Rehg. Efficient learning of continuous-time hidden markov models for disease progression. In *Advances in neural information processing systems*, pages 3600–3608, 2015.

[173] Kimberly Lomis, Pamela Jeffries, Anthony Palatta, Melanie Sage, Javaid Sheikh, Carl Sheperis, and Alison Whelan. Artificial intelligence for health professions educators. *NAM perspectives*, 2021, 2021.

[174] H Ludwig, JS Miguel, MA Dimopoulos, Antonio Palumbo, R Garcia Sanz, R Powles, S Lentzsch, W Ming Chen, J Hou, A Jurczyszyn, et al. International myeloma working group recommendations for global myeloma care. *Leukemia*, 28(5):981–992, 2014.

[175] Jiabing Ma, Yicheng Mo, Menglin Tang, Junjie Shen, Yanan Qi, Wenxu Zhao, Yi Huang, Yanmin Xu, and Cheng Qian. Bispecific antibodies: from research to clinical application. *Frontiers in Immunology*, page 1555, 2021.

[176] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[177] Răzvan V Marinescu, Arman Eshaghi, Marco Lorenzi, Alexandra L Young, Neil P Oxtoby, Sara Garbarino, Sebastian J Crutch, Daniel C Alexander, Alzheimer's Disease Neuroimaging Initiative, et al. Dive: A spatiotemporal progression model of brain pathology in neurodegenerative disorders. *NeuroImage*, 192: 166–177, 2019.

[178] Ruben AT Mars, Yi Yang, Tonya Ward, Mo Houtti, Sambhawa Priya, Heather R Lekatz, Xiaojia Tang, Zhifu Sun, Krishna R Kalari, Tal Korem, et al. Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell*, 182(6):1460–1473, 2020.

[179] Elizabeth S McDonald, Amy S Clark, Julia Tchou, Paul Zhang, and Gary M Freedman. Clinical diagnosis and management of breast cancer. *Journal of Nuclear Medicine*, 57(Supplement 1):9S–16S, 2016.

[180] Jonas Metzger. Adversarial estimators. *arXiv preprint arXiv:2204.10495*, 2022.

[181] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

[182] Adrián Mosquera Orgueira, Marta Sonia González Pérez, José Ángel Díaz Arias, Beatriz Antelo Rodríguez, Natalia Alonso Vence, Ángeles Bendaña López, Aitor Abuín Blanco, Laura Bao Pérez, Andrés Peleteiro Raíndo, Miguel Cid López, et al. Survival prediction and treatment optimization of multiple myeloma patients using machine-learning models based on clinical and gene expression data. *Leukemia*, 35(10):2924–2935, 2021.

[183] DR Mould, NG Denman, and S Duffull. Using disease progression models as a tool to detect drug effect. *Clinical Pharmacology & Therapeutics*, 82(1):81–86, 2007.

[184] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5323–5331, 2022.

[185] Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. In *Conference on Uncertainty in Artificial Intelligence*, pages 41–50. PMLR, 2020.

[186] Brady Neal. Introduction to causal inference. 2015.

[187] Beat Neuenschwander, Michael Branson, and David J Spiegelhalter. A note on the power prior. *Statistics in medicine*, 28(28):3562–3566, 2009.

[188] NIH. Relating clinical outcomes in multiple myeloma to personal assessment of genetic profile (com-mpass). *Clinical Trials website. https://clinicaltrials.gov/ct2/show/NCT01454297*, 2016.

[189] Larry Norton. Cancer log-kill revisited. *American Society of Clinical Oncology Educational Book*, 34(1):3–7, 2014.

[190] Songhee Oh, Jae Heon Kim, Sung-Woo Choi, Hee Jeong Lee, Jungrak Hong, and Soon Hyo Kwon. Physician confidence in artificial intelligence: an online mobile survey. *Journal of medical Internet research*, 21(3):e12422, 2019.

[191] Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4932–4941. PMLR, 2019.

[192] Ian Osband. Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. In *NIPS workshop on bayesian deep learning*, volume 192, 2016.

[193] Mustafa Erhan Ozer, Pemra Ozbek Sarica, and Kazim Yalcin Arga. New machine learning applications to accelerate personalized medicine in breast cancer: rise of the support vector machines. *Omics: a journal of integrative biology*, 24(5): 241–246, 2020.

[194] Antonio Palumbo, Hervé Avet-Loiseau, Stefania Oliva, Henk M Lokhorst, Hartmut Goldschmidt, Laura Rosinol, Paul Richardson, Simona Caltagirone, Juan José Lahuerta, Thierry Facon, et al. Revised international staging system for multiple myeloma: a report from international myeloma working group. *Journal of clinical oncology*, 33(26):2863, 2015.

[195] Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pages 239–260. PMLR, 2021.

[196] Harris Papadopoulos. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka, 2008.

[197] Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

[198] Ravi B Parikh, Christopher R Manz, Maria N Nelson, Chalanda N Evans, Susan H Regli, Nina O'Connor, Lynn M Schuchter, Lawrence N Shulman, Mitesh S Patel, Joanna Paladino, et al. Clinician perspectives on machine learning prognostic algorithms in the routine care of patients with cancer: a qualitative study. *Supportive Care in Cancer*, 30(5):4363–4372, 2022.

[199] Chan Park and Hyunseung Kang. A groupwise approach for inferring heterogeneous treatment effects in causal inference. *arXiv preprint (1908.04427)*, August 2019.

[200] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019.

[201] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[202] Judea Pearl. Generalizing experimental findings. *Journal of Causal Inference*, 3 (2):259–266, September 2015.

[203] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1):247–254, August 2011.

[204] Judea Pearl and Elias Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579–595, November 2014.

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

[206] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[207] Adler Perotte, Rajesh Ranganath, Jamie S Hirsch, David Blei, and Noémie Elhadad. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. *Journal of the American Medical Informatics Association*, 22(4):872–880, 2015.

[208] Rachel Phillips, Lorna Hazell, Odile Sauzet, and Victoria Cornelius. Analysis and reporting of adverse events in randomised controlled trials: a review. *BMJ open*, 9(2):e024537, 2019.

[209] Deborah Plana, Dennis L Shung, Alyssa A Grimshaw, Anurag Saraf, Joseph JY Sung, and Benjamin H Kann. Randomized clinical trials of machine learning interventions in health care: a systematic review. *JAMA Network Open*, 5(9): e2233946–e2233946, 2022.

[210] Stephen R Porter. Quantile regression: Analyzing changes in distributions instead of means. *Higher education: Handbook of theory and research*, pages 335–381, 2015.

[211] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

[212] Francesco Paolo Prete, Angela Pezzolla, Fernando Prete, Mario Testini, Rinaldo Marzaioli, Alberto Patriti, Rosa Maria Jimenez-Rodriguez, Angela Gurrado, and Giovanni FM Strippoli. Robotic versus laparoscopic minimally invasive surgery for rectal cancer: a systematic review and meta-analysis of randomized controlled trials. *Annals of surgery*, 267(6):1034–1046, 2018.

[213] Teresa C Prevost, Keith R Abrams, and David R Jones. Hierarchical models in generalized synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in medicine*, 19(24):3359–3376, 2000.

[214] Cécile Proust-Lima and Jeremy MG Taylor. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. *Biostatistics*, 10(3):535–549, 2009.

[215] Rizwan Qureshi, Syed Abdullah Basit, Jawwad A Shamsi, Xinqi Fan, Mehmood Nawaz, Hong Yan, and Tanvir Alam. Machine learning based personalized drug response prediction for lung cancer patients. *Scientific Reports*, 12(1):18935, 2022.

[216] S Vincent Rajkumar. Multiple myeloma: 2022 update on diagnosis, risk stratification, and management. *American journal of hematology*, 97(8):1086–1107, 2022.

[217] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

[218] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[219] Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, volume 6. International Conference on Representation Learning, 2018.

[220] Dimitris Rizopoulos. *Joint models for longitudinal and time-to-event data: With applications in R*. CRC press, 2012.

[221] James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, March 1995.

[222] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[223] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32:3543–3553, 2019.

[224] P R Rosenbaum and D B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 45(2):212–218, 1983.

[225] Paul R Rosenbaum, PR Rosenbaum, and Briskman. *Design of observational studies*, volume 10. Springer, 2010.

[226] Evan Rosenman, Guillaume Basse, Art Owen, and Michael Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *arXiv preprint arXiv:2002.06708*, 2020.

[227] Evan TR Rosenman, Art B Owen, Mike Baiocchi, and Hailey R Banack. Propensity score methods for merging observational and experimental datasets. *Statistics in Medicine*, 2021.

[228] Joseph S Ross, Joanne Waldstreicher, Stephen Bamford, Jesse A Berlin, Karla Childers, Nihar R Desai, Ginger Gamble, Cary P Gross, Richard Kuntz, Richard Lehman, et al. Overview and experience of the yoda project with clinical trial data sharing after 5 years. *Scientific data*, 5(1):1–14, 2018.

[229] Jacques E Rossouw, Garnet L Anderson, Ross L Prentice, Andrea Z LaCroix, Charles Kooperberg, Marcia L Stefanick, Rebecca D Jackson, Shirley AA Beresford, Barbara V Howard, Karen C Johnson, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *Jama*, 288(3):321–333, 2002.

[230] Peter M. Rothwell. External validity of randomised controlled trials: "to whom do the results of this trial apply?", 2005. ISSN 01406736.

[231] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[232] Donald B Rubin. Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26, 1977.

[233] Charumathi Sabanayagam, Dejiang Xu, Daniel SW Ting, Simon Nusinovici, Riswana Banu, Haslina Hamzah, Cynthia Lim, Yih-Chung Tham, Carol Y Cheung, E Shyong Tai, et al. A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *The Lancet Digital Health*, 2(6):e295–e302, 2020.

[234] Jane Scheetz, Philip Rothschild, Myra McGuinness, Xavier Hadoux, H Peter Soyer, Monika Janda, James JJ Condon, Luke Oakden-Rayner, Lyle J Palmer, Stuart Keel, et al. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Scientific reports*, 11(1):1–10, 2021.

[235] Douglas Scherr, Peter W Swindle, and Peter T Scardino. National comprehensive cancer network guidelines for the management of prostate cancer. *Urology*, 61 (2):14–24, 2003.

[236] Peter F Schnatz, Xuezhi Jiang, Aaron K Aragaki, Matthew Nudy, David M O'Sullivan, Mark Williams, Erin S LeBlanc, Lisa W Martin, JoAnn E Manson, James M Shikany, et al. Effects of calcium, vitamin d, and hormone therapy on cardiovascular disease risk factors in the women's health initiative: a randomized controlled trial. *Obstetrics and gynecology*, 129(1):121, 2017.

[237] Martijn J Schuemie, Patrick B Ryan, William DuMouchel, Marc A Suchard, and David Madigan. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in medicine*, 33(2):209–218, January 2014.

[238] Martijn J Schuemie, George Hripcsak, Patrick B Ryan, David Madigan, and Marc A Suchard. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11):2571–2577, March 2018.

[239] Peter Schulam and Suchi Saria. Integrative analysis using coupled latent variable models for individualizing prognoses. *The Journal of Machine Learning Research*, 17(1):8244–8278, 2016.

[240] Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems*, pages 1697–1708, 2017.

[241] Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The econometrics journal*, 24(2):264–289, 2021.

[242] Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

[243] Sungyong Seo and Yan Liu. Differentiable physics-informed graph networks. *arXiv preprint arXiv:1902.02950*, 2019.

[244] Robert J Serfling. *Approximation theorems of mathematical statistics*. Wiley Series in Probability and Statistics. Wiley-Interscience, Newy York, September 2009.

[245] Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34, 2021.

[246] Kristen A. Severson, Lana M. Chahine, Luba A. Smolensky, Kenney Ng, Jianying Hu, and Soumya Ghosh. Personalized input-output hidden markov models for disease progression modeling. *medRxiv*, 2020.

[247] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[248] Ricardo Silva. Observational-interventional priors for dose-response learning. In *Advances in Neural Information Processing Systems*, pages 1561–1569, 2016.

[249] Ole-Johan Skrede, Sepp De Raedt, Andreas Kleppe, Tarjei S Hveem, Knut Liestøl, John Maddison, Hanne A Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet*, 395(10221):350–360, 2020.

[250] Suzanne Sniekers and Aad van der Vaart. Adaptive bayesian credible sets in regression with a gaussian process prior. *Electronic Journal of Statistics*, 9(2): 2475–2527, 2015.

[251] Hossein Soleimani, Adarsh Subbaswamy, and Suchi Saria. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. In *33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017*. AUAI Press Corvallis, 2017.

[252] Pieter Sonneveld, Hervé Avet-Loiseau, Sagar Lonial, Saad Usmani, David Siegel, Kenneth C Anderson, Wee-Joo Chng, Philippe Moreau, Michel Attal, Robert A Kyle, et al. Treatment of multiple myeloma with high-risk cytogenetics: a consensus of the international myeloma working group. *Blood, The Journal of the American Society of Hematology*, 127(24):2955–2962, 2016.

[253] SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. *New England Journal of Medicine*, 373(22):2103–2116, 2015.

[254] SPRINT Research Group, Jackson T Wright, Jr, Jeff D Williamson, Paul K Whelton, Joni K Snyder, Kaycee M Sink, Michael V Rocco, David M Reboussin, Mahboob Rahman, Suzanne Oparil, Cora E Lewis, Paul L Kimmel, Karen C Johnson, David C Goff, Jr, Lawrence J Fine, Jeffrey A Cutler, William C Cushman, Alfred K Cheung, and Walter T Ambrosius. A randomized trial of intensive versus standard Blood-Pressure control. *The New England journal of medicine*, 373(22):2103–2116, November 2015.

[255] Leonard A Stefanski and Dennis D Boos. The calculus of M-Estimation. *The American statistician*, 56(1):29–38, 2002.

[256] Robert C Sterner and Rosalie M Sterner. Car-t cell therapy: current limitations and potential strategies. *Blood cancer journal*, 11(4):69, 2021.

[257] A Keith Stewart, S Vincent Rajkumar, Meletios A Dimopoulos, Tamás Masszi, Ivan Špička, Albert Oriol, Roman Hájek, Laura Rosiñol, David S Siegel, Georgi G Mihaylov, et al. Carfilzomib, lenalidomide, and dexamethasone for relapsed multiple myeloma. *New England Journal of Medicine*, 372(2):142–152, 2015.

[258] Marc A Suchard, Martijn J Schuemie, Harlan M Krumholz, Seng Chan You, Ruijun Chen, Nicole Pratt, Christian G Reich, Jon Duke, David Madigan, George Hripcsak, and Patrick B Ryan. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *The Lancet*, 394(10211):1816–1826, November 2019.

[259] Zhaonan Sun, Soumya Ghosh, Ying Li, Yu Cheng, Amrita Mohan, Cristina Sampaio, and Jianying Hu. A probabilistic disease progression modeling approach and its application to integrated huntington's disease observational data. *JAMIA Open*, 2(1):123–130, 2019.

[260] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In *Machine Learning for Healthcare Conference*, pages 322–337. PMLR, 2017.

[261] S Taghipour, D Banjevic, AB Miller, N Montgomery, AKS Jardine, and BJ Harvey. Parameter estimates for invasive breast cancer progression in the canadian national breast screening study. *British journal of cancer*, 108(3):542–548, 2013.

[262] Evangelos Terpos, Joseph Mikhael, Roman Hajek, Ajai Chari, Sonja Zweegman, Hans C Lee, María-Victoria Mateos, Alessandra Larocca, Karthik Ramasamy, Martin Kaiser, et al. Management of patients with multiple myeloma beyond the clinical-trial setting: understanding the balance between efficacy, safety and tolerability, and quality of life. *Blood cancer journal*, 11(2):40, 2021.

[263] Anja Thieme, Aditya Nori, Marzyeh Ghassemi, Rishi Bommasani, Tariq Osman Andersen, and Ewa Luger. Foundation models in healthcare: Opportunities, risks & strategies forward. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2023.

[264] Tian Tian, Ji Wan, Qi Song, and Zhi Wei. Clustering single-cell rna-seq data with a model-based deep learning approach. *Nature Machine Intelligence*, 1(4): 191–198, 2019.

[265] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[266] Ruth Tsang, Lindsey Colley, and Larry D Lynd. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *Journal of clinical epidemiology*, 62(6):609–616, 2009.

[267] Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

[268] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

[269] V Valentí, J Ramos, C Pérez, L Capdevila, I Ruiz, Lidia Tikhomirova, M Sánchez, I Juez, M Llobera, E Sopena, et al. Increased survival time or better quality of life? trade-off between benefits and adverse events in the systemic treatment of cancer. *Clinical and Translational Oncology*, 22:935–942, 2020.

[270] Suhas V Vasaikar, Adam K Savage, Qiuyu Gong, Elliott Swanson, Aarthi Talla, Cara Lord, Alexander T Heubeck, Julian Reading, Lucas T Graybuck, Paul Meijer, et al. A comprehensive platform for analyzing longitudinal multi-omics data. *Nature Communications*, 14(1):1684, 2023.

[271] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[272] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018.

[273] Charles S Venuto, Nicholas B Potter, E Ray Dorsey, and Karl Kieburtz. A review of disease progression models of parkinson's disease and applications in clinical trials. *Movement Disorders*, 31(7):947–956, 2016.

[274] Kailas Vodrahalli, Tobias Gerstenberg, and James Y Zou. Uncalibrated models can improve human-ai collaboration. *Advances in Neural Information Processing Systems*, 35:4004–4016, 2022.

[275] Alexander Volkmann, Riccardo De Bin, Willi Sauerbrei, and Anne-Laure Boulesteix. A plea for taking all available clinical information into account when assessing the predictive value of omics data. *BMC medical research methodology*, 19:1–15, 2019.

[276] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

[277] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[278] Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. 1999.

[279] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, July 2018.

[280] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[281] Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

[282] Rui Wang, Robin Walters, and Rose Yu. Incorporating symmetry into deep dynamics models for improved generalization. *arXiv preprint arXiv:2002.03061*, 2020.

[283] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94, 2014.

[284] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, NY, 2004.

[285] Nicky J Welton, Anthony E Ades, JB Carlin, DG Altman, and JAC Sterne. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):119–136, 2009.

[286] Jeffrey West and Paul K Newton. Chemotherapeutic dose scheduling based on tumor growth rates provides a case for low-dose metronomic high-entropy therapies. *Cancer research*, 77(23):6717–6728, 2017.

[287] Jack Wilkinson, Kellyn F Arnold, Eleanor J Murray, Maarten van Smeden, Kareem Carr, Rachel Sippy, Marc de Kamps, Andrew Beam, Stefan Konigorski, Christoph Lippert, et al. Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health*, 2(12):e677–e680, 2020.

[288] Axel Wismüller and Larry Stockmaster. A prospective randomized clinical trial for measuring radiology study reporting time on artificial intelligence-based detection of intracranial hemorrhage in emergent care head ct. In *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 11317, pages 144–150. SPIE, 2020.

[289] Robert L Wolpert and Kerrie L Mengersen. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science*, 19(3):450–471, 2004.

[290] Sifan Wu, Xi Xiao, Qianggang Ding, Peilin Zhao, Ying Wei, and Junzhou Huang. Adversarial sparse transformer for time series forecasting. *Advances in neural information processing systems*, 33:17105–17115, 2020.

[291] Zhenqin Wu, Alexandro E Trevino, Eric Wu, Kyle Swanson, Honesty J Kim, H Blaize D'Angio, Ryan Preska, Gregory W Charville, Piero D Dalerba, Ann Marie Egloff, et al. Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nature Biomedical Engineering*, pages 1–14, 2022.

[292] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10): 1419–1428, 2018.

[293] Yu Xie, Christopher Near, Hongwei Xu, and Xi Song. Heterogeneous treatment effects on children's cognitive/non-cognitive skills: A reevaluation of an influential early childhood intervention. *Social science research*, 86:102389, 2020.

[294] Yanbo Xu, Yanxun Xu, and Suchi Saria. A bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine learning for healthcare conference*, pages 282–300. PMLR, 2016.

[295] Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. August 2018.

[296] Shu Yang, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomized trial and Real-World data for treatment heterogeneity estimation. *arXiv preprint (2005.10579)*, May 2020.

[297] Liangliang Zhang, Chae Young Lim, Tapabrata Maiti, Yingjie Li, Jongeun Choi, Andrea Bozoki, David C Zhu, Key Laboratory of Applied Statistics of the Ministry of Education (KLAS), and for the Alzheimer's Disease Neuroimaging Initiative. Analysis of conversion of alzheimer's disease using a multi-state markov model. *Statistical methods in medical research*, 28(9):2801–2819, 2019.

[298] Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Maximum moment restriction for instrumental variable regression. *arXiv preprint arXiv:2010.07684*, 2020.

[299] Sushen Zhang, Seyed Mojtaba Hosseini Bamakan, Qiang Qu, and Sha Li. Learning for personalized medicine: a comprehensive review from a deep learning perspective. *IEEE reviews in biomedical engineering*, 12:194–208, 2018.

[300] Xi Zhang, Jingyuan Chou, Jian Liang, Cao Xiao, Yize Zhao, Harini Sarva, Claire Henchcliffe, and Fei Wang. Data-driven subtyping of parkinson's disease using longitudinal clinical records: a cohort study. *Scientific reports*, 9(1):1–12, 2019.

[301] Jie Zheng and Ke Wang. Emerging deep learning methods for single-cell rna-seq data analysis. *Quantitative Biology*, 7:247–254, 2019.

[302] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023.

[303] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

[304] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pages 27268–27286. PMLR, 2022.