# Causal Inference in Healthcare: Approaches to Causal Modeling and Reasoning through Graphical Causal Models

Riddhiman Adib
*Marquette University*

# CAUSAL INFERENCE IN HEALTHCARE: APPROACHES TO CAUSAL MODELING AND REASONING THROUGH GRAPHICAL CAUSAL MODELS

by

Riddhiman Adib, M.S.

A Dissertation Submitted to the Faculty of the
Graduate School, Marquette University,
in Partial Fulfillment of the Requirements for
the Degree of Doctor of Philosophy

Milwaukee, Wisconsin

May 2022

**ABSTRACT**

CAUSAL INFERENCE IN HEALTHCARE: APPROACHES TO CAUSAL MODELING AND REASONING THROUGH GRAPHICAL CAUSAL MODELS

Riddhiman Adib, M.S.

Marquette University, 2022

In the era of big data, researchers have access to large healthcare datasets collected over a long period. These datasets hold valuable information, frequently investigated using traditional Machine Learning algorithms or Neural Networks. These algorithms perform great in finding patterns out of datasets (as a predictive machine); however, the models lack extensive interpretability to be used in the healthcare sector (as an explainable machine). Without exploring underlying causal relationships, the algorithms fail to explain their reasoning. Causal Inference, a relatively newer branch of Artificial Intelligence, deals with interpretability and portrays causal relationships in data through graphical models. It explores the issue of causality and works towards an explainability of underlying causal models deeply buried in data.

For this dissertation work, the research goal is to use Causal Inference to build an applied framework that lets researchers leverage observational datasets in understanding causal relationships between features. To achieve that, we focus on specific objectives such as (a) the addition of background knowledge to causal structure learning algorithms, (b) the proposal of new causal inference methodologies, (c) generation of theories connecting causality to standard statistical analyses (e.g., Odds Ratio, Survival Analysis), and (d) application of proposed approaches in real-world healthcare problems. This dissertation encapsulates the tasks mentioned above, through various new methodologies and experiments under the rubric of Structural Theory of Causation. We discuss the common research theme in causal inference, historical development, the structural theory of causation, and underlying assumptions. Finally, we explore the impact of these proposed methodologies in real-world treatment controversy of Delirium patients, by examining the efficacy of antipsychotic drugs prescribed in treating Delirium in the ICU, from a curated observational healthcare dataset.

# ACKNOWLEDGMENTS

Riddhiman Adib, M.S.

I am thrilled and ecstatic today, expressing my acknowledgment for the most remarkable achievement of my life. My Ph.D. in Computer Science is a hard-earned one, coming from a background in Electrical Engineering and studying in a foreign land. And now that I am here, I would love to mention and pay respect to all the people who have supported me in my journey to this day.

Firstly, I would like to express my sincerest gratitude to my Ph.D. advisor, Dr. Sheikh Iqbal Ahamed, for his support throughout my doctoral studies. Graduate school is a demanding time for the students; without constant encouragement, guidance, and comfort of a caring supervisor, this is an impossible feat to complete.

I sincerely believe that every star and moon aligned perfectly until this moment in my life to let me have this incredible moment and earn this great honor. Every person, who contributed somehow to my life, deserves recognition in my gratitude. I would like to start with my wonderful childhood at Banasree, Dhaka, my childhood and school friends, and my buddies at Ideal School and College. My teachers at Ideal School and a wonderful year (grade 6) at MSC Cadet Coaching built the base of my future; they shaped me into who I am now.

From grade 7 to 12 *(middle school & high school)*, I went through a fantastic journey at Jhenidah Cadet College. Everything I am right now and will be in the future is because of my cadet college. This military school taught me perseverance, discipline, and the value of hard work above everything. I got a lifetime of amazing friends and peers from my cadet college days. After that, I got a fantastic opportunity to learn from and with the best in Bangladesh during my undergrad in EEE at BUET. BUET took us through a rigorous curriculum; however, it also let me have my bit of fun and independence on the campus. And I got another set of awesome friends here as well.

I am fortunate to have renowned mentors in my life. I have learned a lot at my job at Harriken.com and BanglaTrac Technologies from my mentors and colleagues. In my grad school, I had a prominent opportunity to learn from and work with Dr. Adibuzzaman. He has guided me relentlessly and helped me find my passion in research. This dissertation is an outcome of our continuous brainstorming and ideations over the last four years. With that, I would also like to thank my

professors at Marquette and my external collaborators: Dr. Osmani, Dr. Paul, Dr. Madiraju, Dr. Stacee, and Dr. Norah.

Milwaukee has been my home for the last 5 years, for my whole grad life. This is a fantastic city with a stunning lake beside us; this is a blessing for someone who loves living beside water bodies. I have met great friends and peers here. I want to thank my dearest friends who have been with us through thick and thin, my department buddies at Cudahy Hall, Marquette, and my fellow lab members *(past & present)*.

I am lucky enough to have a pair of wonderful and caring parents, Jessmin Akter & Md. Abdur Rahman. They have supported me with whatever whimsical wishes I wanted to pursue in my life; they have also taught me to be caring, respectful, cautious, and inquisitive. My younger brother, Taskin Sabit Rudra, has always been one of my greatest advocates and well-wisher. He has always been by my side and one of the exemplary guys one could have as a brother. I also want to thank my cousin, Bellah, who has literally taught me everything I know. He has been a mentor from my early childhood for long nine years (2000-2009), and he is the best teacher I have ever witnessed in my life. His two most significant gifts were: (1) understanding and relying on 'logic' in life, and, (2) teaching me how to learn topics by myself. I would also like to thank my constant well-wishers: my cousins *(puchkis)*, aunts *(khalas-mamas-chachus)*, grandparents *(nanu & dadu-dadi)*, and my extended family.

Last but not least, I want to acknowledge the person without any of this would not happen *(or even if it had happened, it would not have any value)*, my constant support, my rock, and the best buddy a guy could ask for; my wife, Jannatul Ferdause Tumpa. I am blessed to have her on this crazy journey, and we both completed our Ph.D. together at Marquette. Since my early undergrad years, I have had her on my side; and together, we have ridden through extraordinary times, sometimes rough but mostly fun. I want to thank Tumpa for being my partner in this incredible ride and supporting me to this day. As I am writing this now and wrapping up my Ph.D. works *(and acknowledgment)*, in front of me is our little bundle of joy, our firstborn, Audriyana Adib. She makes me feel that all of this: relocation to a foreign land, staying apart from our family, the constant stress, studying for the last 30 years, this is all worth it. Finally, before wrapping it up, I want to mention all others in my life; I admire you all for contributing to my life and supporting me to this point.

# TABLE OF CONTENTS

**ACKNOWLEDGMENTS** **i**

**LIST OF TABLES** **vii**

**LIST OF FIGURES** **viii**

**CHAPTER 1 Causal Inference and it's Impact in Healthcare Research** **1**

1.1 Introduction . . . . . 1

1.2 Big Data, Machine Learning, Causal Inference . . . . . 3

1.3 Background and Related Works . . . . . 5

1.3.1 Structural Theory of Causation . . . . . 5

1.3.2 Causal Structure Learning Algorithms . . . . . 6

1.3.3 Transportability and Causal Queries of Interest . . . . . 7

1.3.4 Applied Causal Inference . . . . . 9

1.4 Motivation . . . . . 10

1.4.1 Computational Significance . . . . . 11

1.4.2 Clinical Importance . . . . . 12

1.5 Research Theme . . . . . 13

1.6 Research Questions . . . . . 13

1.7 Organization of this Thesis . . . . . 15

**CHAPTER 2 CKH: Causal Knowledge Hierarchy for Estimating Structural Causal Models from Data and Priors** **17**

2.1 Introduction . . . . . 17

2.2 Related Work . . . . . 19

2.2.1 Structural Causal Models . . . . . 20

2.2.2 Structure Learning Algorithms . . . . . 22

2.2.3 Inter-rater Agreement . . . . . 24
2.2.4 Hierarchy of Evidence . . . . . 26
2.2.5 State of the Art . . . . . 27
2.3 Expert Augmented Causal Model with Knowledge Hierarchy . . . . . 29
2.3.1 Causal Knowledge Hierarchy . . . . . 29
2.3.2 Algorithm . . . . . 35
2.4 Experimental Results . . . . . 41
2.5 Discussion . . . . . 48
**CHAPTER 3 A Causally Formulated Hazard Ratio Estimation through Backdoor Adjustment on Structural Causal Model 53**
3.1 Introduction . . . . . 53
3.1.1 Clinical Relevance . . . . . 57
3.1.2 Technical Significance . . . . . 57
3.1.3 Generalizable Insights . . . . . 58
3.2 Related Work . . . . . 59
3.3 Background . . . . . 63
3.3.1 Hazard Ratio . . . . . 63
3.3.2 Structural Causal Models . . . . . 65
3.3.3 Problem Definition . . . . . 67
3.4 Methods . . . . . 67
3.4.1 Assumptions . . . . . 68
3.4.2 Approach . . . . . 70
3.5 Experiments and Applications . . . . . 77
3.5.1 Experimental Data . . . . . 77
3.5.2 Ewing's Sarcoma Data . . . . . 80
3.6 Discussion and Conclusion . . . . . 82

**CHAPTER 4 Pragmatic Clinical Trials in the Rubric of Structural Causal Models** **86**

4.1 Introduction . . . . . 86

4.2 Background . . . . . 87

4.2.1 Experimental Studies . . . . . 88

4.2.2 Pragmatic Clinical Trials . . . . . 89

4.2.3 Structural Causal Models . . . . . 92

4.2.4 SCM for Scientific Studies . . . . . 93

4.2.5 Problem Definition . . . . . 94

4.2.6 Related Works . . . . . 94

4.3 Structural Causal Model for Pragmatic Clinical Trials . . . . . 95

4.3.1 Defining PCT for SCM . . . . . 96

4.3.2 Features of PCT . . . . . 97

4.3.3 Outcome Analysis for PCT . . . . . 99

4.4 Example of PCT with SCM . . . . . 102

4.4.1 SCM for PCT . . . . . 104

4.4.2 Outcome Analysis . . . . . 104

4.5 Discussion and Conclusion . . . . . 105

**CHAPTER 5 Causal Discovery on the Effect of Antipsychotic Drugs on Delirium Patients in the ICU using Large EHR Dataset** **110**

5.1 Background and Problem Statement . . . . . 110

5.2 Method . . . . . 116

5.3 Results . . . . . 117

5.3.1 Covariate Selection . . . . . 117

5.3.2 Data Curation Process . . . . . 119

5.3.3 Data Overview & Exploratory Insights . . . . . 119

5.3.4 Predictive Analysis on MIMIC-Delirium dataset . . . . . 121

5.3.5 Causal Analysis on MIMIC-Delirium dataset . . . . . 122
5.4 Discussion . . . . . 127
**CHAPTER 6 Conclusion and Future Work** **130**
6.1 Broader Impact and Summary Contribution of this dissertation . . . 130
6.2 Future Work . . . . . 132
**BIBLIOGRAPHY** **135**

# LIST OF TABLES

2.1 Summary of different SLAs and their outputs . . . . . 24

2.2 Alteration in causal edges in the simulation . . . . . 47

2.3 Iterations of simulations with false information injected in each tier . 47

3.1 Hazard Ratio for simulated dataset, calculated using existing model and our proposed approach . . . . . 79

3.2 Hazard Ratio for Ewing dataset, calculated using existing model and our proposed approach . . . . . 82

4.1 Population distribution for different values of treatment prescribed $X$, treatment received $X'$ and outcome $Y$ . . . . . 105

4.2 Outcome metrics for the PCT . . . . . 105

5.1 Features in MIMIC-Delirium . . . . . 118

5.2 Outcomes estimation in Average Treatment Effects (ATE) . . . . . 126

# LIST OF FIGURES


1.1 An example of a causal DAG . . . . . 6

1.2 Two example Causal structure generated in two different studies. *(Left)* Causal structure generated in the study of modelling air pollution, climate and health data. *(Right)* Causal structure generated in the study of ct honeycombing with increased mortality for interstitial lung diseases 8

1.3 An example of a selection diagram . . . . . 9

2.1 Simplest graphical model representing observational study with three variables, . . . . . 21

2.2 Evidence-Based Medicine (EBM) Resources . . . . . 27

2.3 General overview of causal structure generation pipeline using CKH . 30

2.4 Tiers of Causal Knowledge Source . . . . . 31

2.5 Ground-truth Causal Model versus Structural Causal Models with edge confidences at individual tiers of CKH . . . . . 42

2.6 Structural Causal Models as outputs of Tier 2 in CKH . . . . . 46

3.1 An example survival curve, collected from Girard et al. ([34]) . . . . . 55

3.2 Schematic overview of the proposed approach . . . . . 68

3.3 Simple observational study (treatment $X$, outcome in survival-time $T$ and single confounder $Z$) . . . . . 69

3.4 Converted Causal DAGs with survival time converted to binary outcome of survival at different timepoints . . . . . 73

3.5 Unadjusted survival curve for simulated data *(left)* and, survival curve generated after applying proposed approach *(right)* . . . . . 79

3.6 Unadjusted survival curve for Ewing dataset *(left)* and, survival curve generated after applying proposed approach *(right)* . . . . . 81

3.7 Two example graphs where the backdoor adjustment will produce different results compared to an approach based on the ignorability assumption . . . . . 83

4.1 Visualization of PRECIS (PRagmatic Explanatory Continuum Indicator Summary) . . . . . 91

4.2 SCM representation of scientific studies . . . . . 94

4.3 Graphical representation of the proposed structural causal model for pragmatic clinical trials . . . . . 96

4.4 Graphical representation of the structural causal model of a $RCT_{PCT}$, *(left)* with population $\Pi_s$ & treatment $X$ ($= X'$), and *(right)* with population $\Pi_s$ as a selection bias through node $S$ on population $Z$ . . 97

4.5 Hypothetical PCT in patients with cardiovascular disease. Intervention, A = medical management + surgery, vs. control, B = medical management only. Collected from McCoy et al. [76] . . . . . 102

4.6 Graphical overview of SCM representation of the example PCT . . . 104

5.1 RCT for Antipsychotics-based treatment for Delirium . . . . . 112

5.2 Target observational study from large EHR data . . . . . 115

5.3 Data mining protocol (simplified) . . . . . 120

5.4 Data distribution on age in years *(left)* and length-of-stay in days *(right)* 120

5.5 Correlation heatmap of MIMIC-Delirium . . . . . 122

5.6 Combined Causal Graph for Delirium in the ICU *(blue: treatment, red: primary and secondary outcomes)* . . . . . 124

# CHAPTER 1
# Causal Inference and it's Impact in Healthcare Research

## 1.1 Introduction

Causal inference methodologies, specially the structural theory of causation [91], has shown potential to extract causal relationships from observational datasets based on certain assumptions, and is now thoroughly explored and developed by scientists. Structural theory of causation depicts experiments through causal directed acyclic graphs, mitigates biases, and proposes theorems that aid in drawing causal inferences from observational as well as experimental studies. However, since it is a relatively newer subgroup of data science and artificial intelligence, we theorize that Structural theory of causation has yet a lot more unexplored potential.

Experimental studies with varying designs and research goals, such as Randomized Controlled Trials (RCT), Pragmatic Clinical Trials (PCT), are frequently conducted in health sciences [116]. The general intent is to draw an inference on treatment intervention (i.e., drug efficacy) on a specific population group (i.e., patients under critical care, people over age 65) as well as the more general population (i.e., hospital-admitted patients, elder-care facilities) under usual care. However, due to differences in the experiment settings (i.e., adherence to drug prescription, presence of control group), transfer of knowledge from one study to another is not trivial [66], and therefore, making a general inference of intervention

efficacy becomes unclear. It also hinders the usage of large-scale healthcare data and merging of causal information from these observational datasets (Obs) [87]. Thus, there is a need for structural methodologies, assumptions, inferences and information, to draw unbiased causal inferences from observational data, or from one experiment to another, or in general, combining all of them (RCT+PCT+Obs), leveraging their unique design attributes.

The objective of this research proposal is to scheme a set of methodologies that can define studies with diverse experimental settings (RCT, PCT, Obs), make inferences based on assumptions, and extract (and transfer) causal effects of treatment interventions. Our in-depth understanding of the experiments and underlying assumptions, along with ongoing research work on survival analysis from observational studies using the structural theory of causation makes this research proposal uniquely strong and plausible. Availability of publicly available large healthcare datasets (MIMIC-III), extraction of dataset on focused experimental scenario (e.g., delirium) with related covariates, and access to a real-world pragmatic clinical trial dataset on 351 number of Intensive Care Unit (ICU) patients make this proposal more credible in validating the generated hypotheses.

The general specific aims of this research are:

- to aid in new novel/make amends for improvement of existing causal structure learning algorithms based on observational dataset and assumptions,
- to generate a methdology to incorporate background knowledge in building the causal structure,
- to propose novel methodolgies to extract casusal effect using source dataset, target dataset and Causal model, and,
- to validate the proposed hypotheses using experimental and observational dataset from antipsychotic drug usage on delirium-induced patients in the ICU.

## 1.2 Big Data, Machine Learning, Causal Inference

In the era of 'Big Data,' the world is transmitting an enormous amount of information, their collection, storage, analysis has become a standard task in every part of our life. The proposition of new statistical tools and models have enabled machine learning processes to be applied to them, improvement of computation power (GPU and parallel computing) have made designing neural network possible.

Researchers now have access to an enormous amount of observational data, most of which are still unexplored and contains a great potential to possess causal effects of particular interventions (drugs/practices/actions), similar to a clinical trial [8]. The causal exploration is an ever-growing research problem for statisticians,

mathematicians, computer scientists, and epidemiologists, extracting causal information from observational data, as we do with experimental data.

Causal inference is the science of drawing conclusions on causal relationships, depending on the conditions of the occurrence of the effects. Finding and establishing a causal relationship is not as trivial as finding correlations between variables; however, the process relies on them, along with certain assumptions *(discussed in later sections)*. The prime difference between causal inference and association-driven inference is that the former explores the response of the effects when the cause is manipulated/intervened. Causal inference lets scientists start from association, move up to performing interventions, and finally rise to analyze counterfactuals [13, 92].

Although Causal Inference shares some actions with machine learning protocols, there are individual differences between them as well. The similarities are in some of the model search and feature selection processes, overlapping of models, scoring of models, etc. The dissimilarities are more prominent; the major one is, machine learning focuses on prediction through curve-fitting [121]. It mostly does not care or consider the cause of an outcome. Causal inference, on the other hand, tries to define the underlying causal mechanism between variables. Another significant difference is, causal inference contains the knowledge of the intervention, which traditional machine learning models lack. Finally, causal inference requires a

precise standard procedure of validation of results, since there is a lack of process to gather ground-truth causal mechanisms. This lacking is one of the severe impediments in the development of the science of causal inference.

## 1.3 Background and Related Works

### 1.3.1 Structural Theory of Causation

In Causal Inference, researchers build their works in two distinct directions: 1) using Rubin Causal Model (RCM) [56, 107], based on potential outcomes framework, and, 2) using the structural theory of causation [89, 91], based on probabilistic graphical models. We rely on the structural theory of causation since the graphical representation of causal models opens up more opportunities for better visualization, understanding, and clearer definitions (i.e., backdoor criterion) [6].

A structural causal model [91] is a 4-tuple $\{U, V, f, P(u)\}$ where U is a set of background (exogenous) variables that are determined by factors outside of the model, V is a set $\{V1, V2, ..., Vn\}$ of observable (endogenous) variables that are determined by variables in the model, F is a set of functions $\{f1, f2, ..., fn\}$ such that each f is a mapping from the respective domains of U to V, and P(u) is a probability distribution over the exogenous variables. This is commonly represented through a directed acyclic graph (Causal DAG) (Figure 1.1).

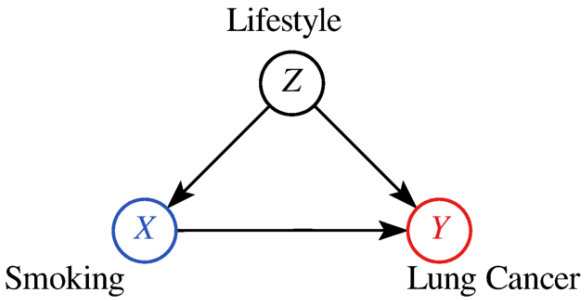



Figure 1.1: An example of a causal DAG

### 1.3.2 Causal Structure Learning Algorithms

One of the critical research areas in Causal inference is Causal structure learning algorithms. Causal structure learning algorithms (SLA) are specific graph search algorithms that detect causal structure expressed as graphs with nodes and edges, from the conditional dependencies of in-between dataset variables and additional assumptions [121]. Based on their approach and associated assumptions, causal structure learning algorithms have been broadly categorized into five (5) categories [44], such as constraint-based methods (e.g., Peter-Clark (PC), fast causal inference (FCI)), score-based methods (e.g., greedy equivalence search (GES), greedy interventional equivalence search (GIES), hybrid methods (e.g., max-min hill climbing (MMHC)), structural equation models with additional restrictions (e.g., linear non-Gaussian acyclic models (LINGAM)), and, exploiting invariance properties (e.g., backShift). Building and improving causal SLA is a high potential research area in causal inference and has been applied many times to generate an underlying causal structure given a specific scenario and dataset.

There are multiple issues associated with causal SLAs, which are open research problems. Most of the times, the algorithms are not very computationally efficient (high runtime complexity), and fails to detect a causal structure. Lack of a ground-truth makes them very hard to compare and diagnose. Change in assumptions, missing data, lack of appropriate covariate data also makes them very vulnerable to a false-positive result. A more robust approach is through the ensembling of their findings. Epidemiologists have been continuously involving causal structure learning algorithms for identifying underlying causal structures. Different studies have taken different paths, few studies [128, 2] have used specific SLA algorithms to detect a causal DAG applicable for a targeted research question, whereas others [110, 10] have assumed the causal structure from literature, and validated them using datasets available. Albeit their success in estimating causal effects, due to strong assumptions, missing data, and general variation in experiments, there are potential areas to explore and apply causal SLA in epidemiology.

### 1.3.3 Transportability and Causal Queries of Interest

There are many causal queries in experiments researchers aim to explore. One area of importance is transportability, which is a crucial feature in the structural theory of causation, and still being explored in recent research works [13]. Transportability presents research methodologies in Causal Inference, where experiments from the

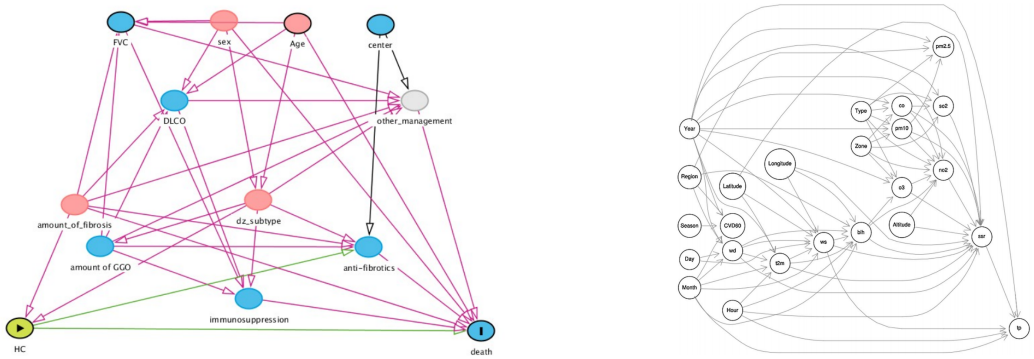



Figure 1.2: Two example Causal structure generated in two different studies. *(Left)* Causal structure generated in the study of modelling air pollution, climate and health data. *(Right)* Causal structure generated in the study of ct honeycombing with increased mortality for interstitial lung diseases

source domain can be leveraged to answer a query in the target domain. Although we focus on epidemiology in our proposal (under the keyword generalizability), this research has an impact in other fields of science as well, under the keywords, external validity in psychology, meta-analysis in statistics, transfer learning in machine learning, etc. We represent causal mechanisms under transportability through Selection diagrams (Figure 1.3), which is also a directed acyclic graph and represents an overlap to causal diagrams of different studies along with selection criteria as a node. Since our research aims to translate causal queries from one study to another, theories of transportability are the most helpful tool. There are additional other directions of causal inquiry in epidemiology: modeling of target trial from observational data and related assumptions, causal modeling of survival analysis, etc. We plan to aid in these explorations through causal inference.

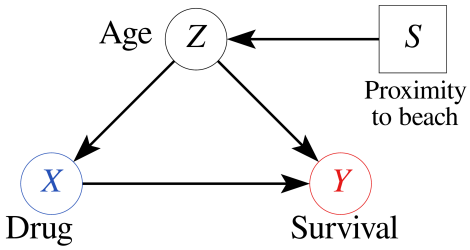



Figure 1.3: An example of a selection diagram

### 1.3.4 Applied Causal Inference

Theories of causal inference rely on observational data. Without applying our theories, we cannot effectively claim the efficacy and validity of our proposed approaches. During research works of prior summer practicum for research in Computational sciences program, we worked on extracting an observational dataset for delirium-induced patients in the Intensive Care Unit (ICU) from large observational data sets along with various covariates correlated with delirium. We utilized the MIMIC III database [58], extensive electronic health records dataset with 53,423 distinct hospital admissions. We defined a target trial, and based on the target trial, we extracted information of the subject group diagnosed with delirium (ICD9 code 293.0) from the MIMIC III database and conducted a retrospective cohort analysis. We categorized them into three groups based on commonly prescribed antipsychotic drugs (APD); patients prescribed Haloperidol, other APD, and no APD. Primary outcomes are death in hospital and death timeline (death in

30 days / 90 days / a year). Secondary outcomes contain length-of-stay in the ICU, and time-in-mechanical ventilation.

### 1.4 Motivation

The motivation of this research is to generate standardized methodologies aiding in the transfer of knowledge from or between experimental studies and observational studies. We aim to build methods that can define studies with diverse experimental settings (RCT, PCT, Observational) and extract causal effects of treatment interventions, along with validation through the incorporation of observational and experimental datasets, exploring the efficacy of antipsychotic drugs (APD) on delirium-induced patients in the ICU.

The proposed research activities make notable contributions directly to the field of Statistical Science, Machine Learning, and Epidemiology. However, the underlying causal theories are not only limited to a specific branch of science, rather discuss scientific inquiries overall [69]. Our proposed methodology will: 1) build a bridge between experimental studies and observational studies, and thus open the opportunity to use a large amount of observational data we have access to, 2) showcase a procedure for researchers to use to draw causal inference from datasets (experimental or observational) in consideration to their design features, and 3) provide an insight on the mental causal model scientists consider while designing an experiment.

### 1.4.1 Computational Significance

The ability to translate scientific findings in one experimental setting and transfer to a different one is a vital process of the scientific investigation. Inference techniques are highly sought in Epidemiology, Sociology, Finance, Psychology, etc. and theories of causal inference have shown great promises in quantifying the causal effects in these fields.

The research area of emulating experiments (RCT/PCT) using an observational dataset is highly expected but relatively newer, and still contains many controversies. Researchers have shown the possibility of using observational datasets to emulate RCTs for antiretroviral therapy (Lodi et al. 2019) and ARDS (Bikak et al. 2018). We plan to build our approach on top of these existing research works and find novel ways to address the shortcomings. Our proposed research plan will be focused on: 1) to propose a standard statistical and computational framework to build (emulate) experiments (RCT/PCT) from Observational dataset with minimal bias, 2) to propose Causal Inference approach to existing statistical methodologies, like transportability, Survival Analysis, 3) to showcase a new way to explore large datasets and interpret their underlying causal structure.

Our proposed methodology aims to lower the gaps between experimental and observational studies and to generate an unbiased causal effect estimation from dissimilar datasets with varying experimental settings. The method and pipeline

can provide the initial results for more extensive studies and a complete framework for the application of causal inference methods in other diseases and interventions.

### 1.4.2 Clinical Importance

While RCTs are the gold standard to identify causal effects of interventions [41], it is time-consuming and costly. On the other hand, the data collected during routine care (i.e., electronic health records (EHR)) might also be valuable in generating insight, identifying the disease pathway, and estimating interventions' effect using methods for causal inference.

The motivation behind applying the proposed methodology, specifically in Delirium-induced patients in the ICU, comes from antipsychotic drugs' use in its treatment and the controversies around it. Delirium frequently occurs in the Intensive Care Unit (ICU) (up to 80% cases [35]) and is associated with longer hospital stay (10% increased mortality for each additional day [96]). Delirium is commonly treated with antipsychotic drugs (APD) such as haloperidol and ziprasidone. However, multiple randomized controlled trials (RCTs) have shown either conflicting or inconclusive results about the efficacy of APD in the treatment of delirium [82, 42]. This calls for an efficacy analysis from a separate standpoint, resulting from observational datasets collected over a longer time-period and wider population demography.

## 1.5 Research Theme

The general research theme of this proposal is to develop application methodologies of Causal inference in extracting causal information from an observational dataset. We hypothesize that, compared to traditional statistical methodologies and machine learning models, causal inference is more suitable for manipulating observational and experimental data. Our proposed methodologies will contribute to building foundations for that.

The specific aims are planned to distribute the research plans into four distinct and significant parts, specific aim 1 addresses theoretical prospects of using the structural theory of causation in defining scientific studies, specific aim 2 explores methods to apply background knowledge in generating causal structures, specific aim 3 focuses on specific theories, assumptions, methodologies of extracting causal information (through transportability, causal survival analysis, etc.) under the framework, and specific aim 4 explores application and validation of our proposed methodology in a real-world research problem scenario.

## 1.6 Research Questions

Driven by the research motivations and based on the research theme, the specific aims for our research proposal are:

- **Research Aim 1:** We aim to propose a generalized process of learning simplified Causal Model from observational datasets.
  - Causal structure learning (Proposal of new or results/performance of old ones)
  - Assumptions for causal structure learning
  - Ensembling results from structure learning algorithms
  - Reduced-order causal graph
- **Research Aim 2:** We plan to collaborate with information from other sources (domain knowledge, literature, experts' opinion) to the causal model generated from observational data.
  - Incorporation of background knowledge in the causal graph
  - Idea of Knowledge gathering and collaborating from multiple sources: from experts' experience, from literature, from observational datasets available
  - Building of a standardized dictionary dataset of causal connections between variables under scenarios
  - Possibility of contribution to reduced-order causal graph generation
- **Research Aim 3:** We aim exploration of ways to generate optimum

responses to causal queries for a specific population and extract that information using source dataset, target dataset, and a causal model.

  - Specific methodologies/ algorithms to extract causal information from causal graph and data
  - Issues involving transportability
  - Definition and assumptions for experiments (e.g., Pragmatic trial, RCT)
  - Causal survival analysis

- **Research Aim 4:** We plan to apply the idea of extracting causal information from observational data, as described above, in real-world clinical problems.
  - Specific clinical/epidemiological application through Causal inference
  - Efficacy of Antipsychotic drug on delirium-induced patients in the ICU

### 1.7 Organization of this Thesis

Elaborating upon the research theme explored and aims defined, the remainder of this thesis describes various methodologies and experiments aiding to the plan. In chapter 2, we propose a Causal Knowledge Hierarchy to build robust causal structure from data and other sources of information. In chapter 3, we will present a use of do-calculus on structural causal models, to estimate causally formulated hazard ratio for survival analysis. In chapter 4, we will discuss an ideation of

structural causal models for specific trials conducted in healthcare research, specifically pragmatic clinical trials. In chapter 5, we will apply the previously discussed causal inference methodologies to a real-world problem, estimating efficacy of Antipsychotics in treatment of Delirium in the ICU, through use of large healthcare dataset. Finally, we will conclude with research summary, contributions and future works in chapter 6.

# CHAPTER 2
# CKH: Causal Knowledge Hierarchy for Estimating Structural Causal Models from Data and Priors

## 2.1 Introduction

Causal knowledge discovery or identifying cause-and-effect relationships between variables is a fundamental objective in various domains such as epidemiology and medicine [102], sociology [33], and economics [37]. Without understanding causal relationships, scientists rely on correlations, which do not allow for estimation of intervention effect (i.e., doing) of a variable on a model outcome. While randomized controlled trials remain the gold standard for exploring causation [65], they are often infeasible because of cost, time, and/or ethical reasons [36]. Thus, causal discovery from observational data that is complementary to experimental studies is of significant interest [46, 73, 84].

Recent developments in the theory of causal inference under the Pearl causal hierarchy (PCH) [95, 94, 13], also known as structural theory of causation (within the potential outcome framework) provides the methodologies to estimate causal effects from observational data. Within this, a causal model is expressed through structural causal models (SCMs) [90]. SCMs represent variables of interest (exogenous and endogenous), causal relationships between the variables, and underlying probability distributions. SCMs use a graphical representation of the

causal model, formalize the identification of causal effects from observational and experimental data, estimate the interventional distribution ($P(y|do(x))$) through do-calculus [93], and assess hypothetical scenarios from available data and model with necessary assumptions explicitly encoded into the model.

PCH is grounded in the three layers of causation: ($L_1$) seeing, ($L_2$) doing, and ($L_3$) imagining. Recent work on the causal hierarchy theorem (CHT) proved that discovery in a higher layer of causation using only information from a lower layer is not feasible [11]. In other words, estimating the effect of experimentation (i.e., "doing ($L_2$)") is not feasible based only on observational data (i.e., "seeing ($L_1$)" ) [11]. Hence it is critical to augment observational data ($L_1$) with other sources of information such as expert knowledge to derive the effect of intervention. Expert knowledge can come in different forms such as expert opinions, established causal relationships and, peer-reviewed literature [26]. Within each of these forms, confidence in knowledge can vary. However, no methodological framework exists for incorporating domain expertise with data-driven causal discovery from observational data in a systematic way [122].

We develop a methodological framework to augment data-driven causal discovery tools with human in the loop (HTL) models. The additional causal knowledge sources include different tiers of knowledge, but is not limited to: background knowledge, expert opinion, and literature. For this purpose, we have

broadly categorized possible causal knowledge sources into three tiers and proposed a causal knowledge hierarchy (CKH) between the tiers. Using this causal knowledge hierarchy, we develop an associated standardized methodology to curate necessary causal information and merge them to derive the structural causal model (SCM). We also provide both theoretical and simulated results of the framework along with algorithmic pseudo code detailing the implementation.

We make the following specific contributions in this work:

1. We propose a causal knowledge hierarchy (CKH) on the foundation of levels of evidence in medicine based on the confidence in the causal information.
2. We present a standard methodological pipeline, based on CKH, to capture causal knowledge from different sources and combine them to derive the SCM.
3. We show the effectiveness of our proposed method in a simulated experiment, detailing the implementation, along with evaluation.

## 2.2 Related Work

Before presenting our proposed methodology we briefly discuss the relevant scientific concepts needed to explain the individual steps as well as the rationale behind them. The objective of our method is to generate an expert augmented **Structural Causal Model (SCM)** for a specific hypothesis, where part of the causal information used to generate the graphical representation comes from the

application of **Structure learning algorithms** on datasets. The building blocks of our proposed methodology includes structural causal models (SCM), **Inter-rater agreement functions** for generating aggregated information and a Causal Knowledge Hierarchy inspired from **Hierarchy of Evidence** in evidence-based health research. These concepts are introduced in this section with details.

### 2.2.1 Structural Causal Models

Developed on the foundations of probabilistic graphical models, SCMs are graphical representations of the causal relationships between variables, and are used to draw causal inferences. An SCM is often expressed by a causal graph $G$. Each node $V$ in $G$ represents an observed or unobserved variable, and each directed edge $E$ represents the causal relationships between them.

An SCM $M$ is a 4-tuple $\langle U, V, f, P(u) \rangle$ [89] where,

- $U$ is a set of background (exogenous) variables that are determined by factors outside of the model,
- $V$ is a set of observable (endogenous) variables that are determined by variables in the model,
- $F$ is a set of functions such that each $f_i \subseteq F$ is a mapping from the respective domains of $U_i \cup PA_i$ to $V_i$, where $U_i \subseteq U$ and

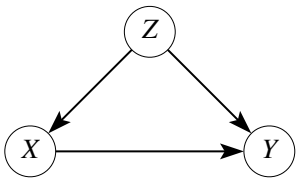



Figure 2.1: Simplest graphical model representing observational study with three variables,

- $PA_i \subseteq V \setminus V_i$ and the entire set $F$ forms a mapping from $U$ to $V$, and $P(u)$ is a probability distribution over the exogenous variables.

A simple structural causal model, with treatment $X$, outcome $Y$ and confounder $Z$, is expressed using causal directed acyclic graph *(nodes are the variables, edges portray causal relationships between variables)* in Figure 2.1.

To find the causal effect of variable $X$ on variable $Y$, do-calculus is introduced [95]. Do-calculus comes with its own set of strong mathematical tools, such as, rules of do-calculus, backdoor criterion, that is used to map the observational reality to the corresponding experimental reality with the identifiability equation by adjusting for different kinds of biases (e.g., confounding bias), if it exists.

Figure 2.1 represents a simple graphical model for an SCM. Although an SCM is used to represent the underlying causal model, in reality, the ground truth causal model in social sciences or medical sciences is never fully known [30]. The causal graph usually represents a set of assumptions explicitly in the problem

domain of interest. Given a data set, an SCM can be any of the causal model for a Markov equivalence class [119], meaning multiple causal models can be true within a Markov equivalence class for a given data set. Consequently, the validation of a causal model is one of the fundamental challenges in causal inference research. The state of the art generates the most-fitting SCM from datasets using structure learning algorithms *(described in next subsection)* using the properties of conditional probability distributions.

### 2.2.2 Structure Learning Algorithms

Other than domain expertise, observational or experimental data can be used to generate a causal graph. Data are the result or snapshot of the underlying causal mechanisms between variables. To recover the causal relationships from data, a rich set of algorithms have been developed over the past thirty years [118, 127, 115]. Causal structured learning is where we try to learn the causal graph or aspects of the causal mechanism. The problem is fundamentally a model selection problem, and these algorithms are called structured learning algorithms (SLA) [44, 25], where a graph is learned or estimated that best describes the dependence structure in a given data set. The learning process includes relying on necessary assumptions (i.e., causal sufficiency, causal faithfulness, linearity), finding conditional dependencies between variables (i.e., Bayes' theorem) and differentiating between different causal structures (i.e., chains, forks, colliders).

Specifically, learning an SCM (or, Bayesian Networks) with a directed acyclic graph (DAG) $G$ and parameters $\theta$ from a dataset $D$ with $n$ observations is completed in two steps [111]: (1) finding the DAG $G$ which encodes the dependence structure of data $D$, called structured learning, and (2) estimating the parameters $\theta$, given the obtained $G$ from structured learning, called parameter learning:

$$P(G, \theta|D) = P(G|D) \cdot P(\theta|G, D)$$

Consequently, SLAs are a key component in estimating causal effects within a dataset.

Several algorithms have been proposed in the literature for SLAs [44], however they differ in their approaches, assumptions, and graphical objects generated. This makes their outcomes varying (even based on the same data source) and difficult to compare.

The main classes of existing SLAs [44] are:

1. **Constraint-based methods:** Peter-Clark (PC), rankPC , fast causal inference (FCI), and rankFCI
2. **Score-based methods:** greedy equivalence search (GES), rankGES, greedy interventional equivalence search (GIES), and rankGIES
3. **Hybrid methods:** Max-min hill climbing (MMHC)

| | PC | FCI | GES | GIES | MMHC | LINGAM |
|---|---|---|---|---|---|---|
| Causal sufficiency | Yes | *No* | Yes | Yes | Yes | Yes |
| Causal faithfulness | Yes | Yes | Yes | Yes | Yes | *No* |
| Acyclicity | Yes | Yes | Yes | Yes | Yes | Yes |
| Non-gaussian errors | *No* | *No* | *No* | *No* | *No* | Yes |
| Known do-intervention | *No* | *No* | *No* | Yes | *No* | Yes |
| Output | CPDAG | PAG | CPDAG | PDAG | DAG | DAG |

Table 2.1: Summary of different SLAs and their outputs

4. **Structural equation models with additional restrictions:** linear non-Gaussian acyclic models(LINGAM)

An overview of their generated graphical models and assumptions required are summarised in Table 2.1.

Since different SLAs can generate different SCMs from the same datasets, there is a need for a principled approach for combining information, which is also correlated with the agreement between them. For this purpose, we leverage inter-rater agreement functions *(described in next section)* to generate an aggregated graphical model that best represents the data along with other sources of causal information (e.g., output SCMs of SLAs, expert opinion or peer-reviewed literature).

### 2.2.3 Inter-rater Agreement

Inter-rater agreement [77] is the degree of agreement among raters, which generates a score on homogeneity, or consensus, in the ratings given by judges or raters. In

causal inference we frequently arrive at multiple ratings on causal relationships (by experts' opinion, or from outputs of SLAs), and this is a mechanism to mitigate the discrepancy. To the best of our knowledge, this mechanism has not been previously used in the context of causal graph generation.

Inter-rater agreement function relies on three operational definitions of agreement [108]:

1. Reliable raters agree with the "official" rating of a performance.
2. Reliable raters agree with each other about the exact ratings to be awarded.
3. Reliable raters agree about which performance is better and which is worse.

In addition, reliable raters are assumed to behave as independent witnesses to the model where they express their independence by disagreeing slightly. In our proposed methodology, we assume the expert opinion, literature, or SLAs are independent raters of causal relationships who capture and express their judgements based on their individual knowledge sources.

Different types of inter-rater agreement functions and scores have been proposed, each with their unique features and strengths. We present a brief overview [77] of a few of them here.

1. Percent agreement

2. Cohen's kappa coefficient
3. Fleiss kappa *(adaptation of Cohen's kappa for 3 or more raters)*
4. Joint probability of agreement
5. Pearson r coefficient
6. Krippendorff's alpha *(useful when there are multiple raters and multiple possible ratings)*

Along with inter-rater agreement function applied on rated causal relationships between variables by raters or algorithms, we propose to incorporate the well established **Hierarchy of Evidence** in evidence-based health research in our methodology.

### 2.2.4 Hierarchy of Evidence

A hierarchy of evidence is needed when there are multiple results or inferences from similar scientific studies (sometimes even contradictory) and one has to choose one or combine them. Hierarchy of evidence (or, levels of evidence) is a scoring that quantifies the rank or strength of the results or outcomes from scientific and experimental studies. The hierarchy relies on the study design, validity and applicability to patient care, and quality of data [1]. When choosing between multiple findings from experimental studies, hierarchy of evidence is critically important. For example, in healthcare professionals are required to decide on

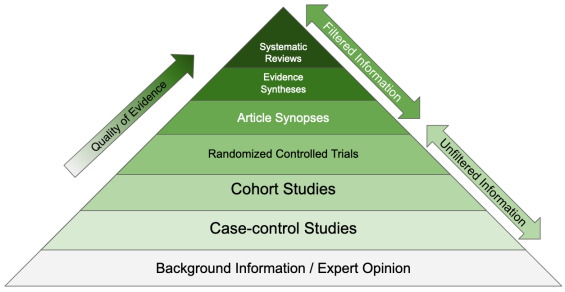



Figure 2.2: Evidence-Based Medicine (EBM) Resources

clinical actions based on the best evidence available. One of the most significant reasons behind using a hierarchy of evidence is to upgrade quality of care, by identifying and promoting practice that is effective and by eliminating those who are ineffective or harmful [5].

Different hierarchy of evidence have been proposed in the literature, based on design of studies and the endpoints measured. A commonly accepted level of effectiveness rating scheme [1] is presented in Figure 2.2.

### 2.2.5 State of the Art

The need for causal inference from sources other than data has been explored in the literature [122]. From a theoretical perspective, research has been conducted on integrating causal information from varying sources. Lee et al. [70] proposed GID-PO that identifies the causal effect from partially-observed distributions. In another result [71], an algorithmic approach that combines data collected under multiple, disparate regimes (observational and interventional) to identify specific

causal effects was presented. However, the focus of these experiments was not on prior knowledge or varying knowledge sources. Borboudakis et al. [16] used path-constraints to incorporate prior causal knowledge, without explicitly discussing the impact of causal knowledge sources. For tiered knowledge, Andrews et al. [7] proposed tiered background knowledge where each tier consists of sets of variables with causal relationships, preceding another set of variables (aka, tier), and demonstrated that FCI (Fast Causal Inference) is a sound and complete causal structure elarning algorithm with and incorporation of this knowledge. From an empirical viewpoint, [86] developed a causal model from medical literature and electronic medical record (EMR) data, by generating two independent graphs, one based on the literature and one from the EMR data, and merged them. The method did not consider other sources of knowledge and did not compare knowledge sources as well as their confidence of information. In a related result, a prior-knowledge-based causal discovery algorithm [129] has been proposed to discover the underlying causal mechanism between bone mineral density and its factors from the clinical data, where prior knowledge was handpicked manually and added to the algorithm as a whitelist between edges. Finally, from a software application perspective, a graphically similar software application to help researchers navigate published findings has been proposed by the software "ResearchMaps" [67]. However, this is primarily a visualization tool to illustrate interconnected features with a graph. In summary, although previous research has

attempted to resolve causal information from varying sources, a unified and principled approach to build a generic SCM is still needed.

## 2.3 Expert Augmented Causal Model with Knowledge Hierarchy

We propose the concept of tiers of causal knowledge and the generalized algorithm for causal model learning through knowledge hierarchy. Specifically, we propose *Causal Knowledge Hierarchy (CKH)* that uses three tiers of knowledge analogous to the "hierarchy of evidence" [1] and the associated weight for each tier. We discuss the assumptions within the CKH required for our proposed methodology. Finally, we establish step-by-step actions for the method. Our approach starts with the inputs: problem statement (PS), defined keywords (K), (empty) structural causal model (SCM), pre-defined tier weights (W), and Inter-rater agreement function (IRR) (*discussed in supplementary document*). For each tier, a general series of steps is described. Finally, the SCM goes through an edge orientation phase to produce the fully specified SCM with individual edge weights ($\langle U, V, F, P(u), C_E \rangle$). An overview of our method is presented in Figure 2.3.

### 2.3.1 Causal Knowledge Hierarchy

Levels of evidence [1], is a well established knowledge hierarchy based on the study design, data collection, and sample size. Based on this concept we propose a "Causal Knowledge Hierarchy (CKH)" to incorporate causal information from different sources. CKH is a multi-level descriptor between types of knowledge and

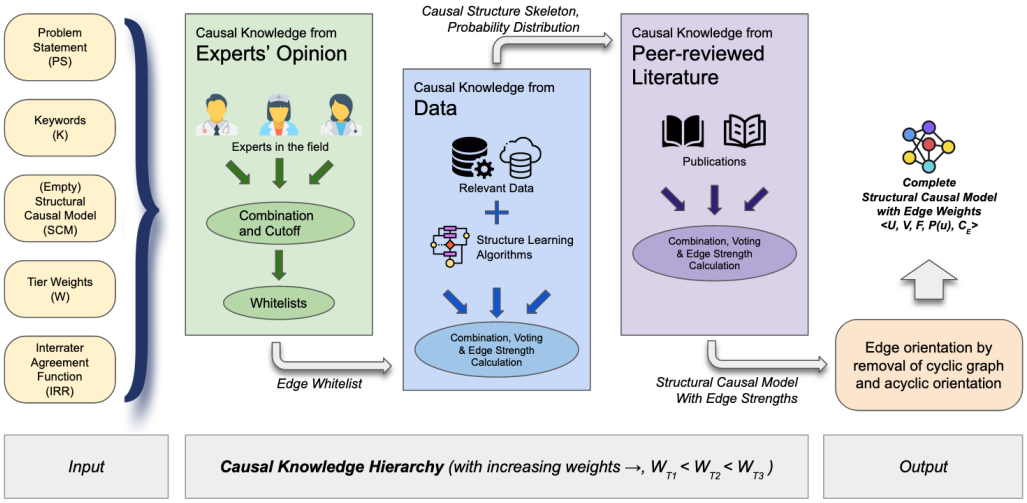



Figure 2.3: General overview of causal structure generation pipeline using CKH

their contribution to the overall causal structure in a problem domain. We initially define three common sources of causal information and propose a hierarchy between them. We define necessary assumptions to make our proposed framework effective and generalizable. We categorize sources of causal knowledge for scientific studies into three distinct classes and define a hierarchy (**CKH**) based on the statistical confidence in the causal information they hold.

**Definition 1** *The three tiers of the CKH are: (1)* ***Tier 1: Causal knowledge from expert opinion / expertise*** *(**CK**$_{\mathbf{E}}$), (2)* ***Tier 2: Causal knowledge from data*** *(**CK**$_{\mathbf{D}}$), and (3)* ***Tier 3: Causal knowledge from peer-reviewed literature*** *(**CK**$_{\mathbf{L}}$). The target structural causal model is a function of convex*

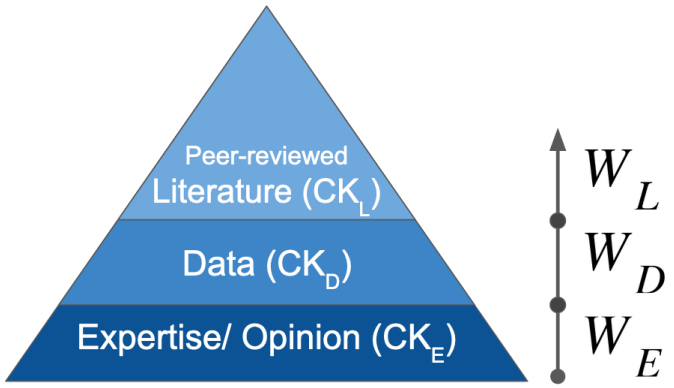



Figure 2.4: Tiers of Causal Knowledge Source

*combination of causal knowledges from the three tiers of sources,*

$$SCM \leftarrow f(CK_E, CK_D, CK_L)$$

**Tier 1: Causal Knowledge from Expert Opinion**

Tier 1 of CKH incorporates causal knowledge based on the expertise and opinions ($\mathbf{CK_E}$) from researchers, scientists, and, subject matter experts (SME). This includes, but is not limited to, inputs from physicians, discussion with application users and intervention participants and, by researchers working in a specific problem domain; and excludes any knowledge directly from peer-reviewed literature. Causal knowledge is generally captured through surveys or structured communications' methods-driven group discussions (e.g., Delphi method [72]). This collaborative knowledge requires further validitation through scientific studies, and is prone to high levels of bias due to variation in the expert's training and experience. We

classify this causal knowledge as Tier 1 ($\mathbf{CK_E}$), and assign a lower weight ($\mathbf{W_E}$) since it contributes diverse information with lesser confidence.

### Tier 2: Causal Knowledge from Data

Tier 2 encodes causal knowledge generated from data sources ($\mathbf{CK_D}$). Data can be from various study designs such as an experimental study (e.g., data from randomized controlled trials), an observational study (e.g., text mining data from social media), or in between (e.g., data from pragmatic clinical trial). Depending on the data generation mechanism, different structural causal models are used to explain the causal relationships between the variables. However, there may be bias from selection, confounding, or other experimental design features. We associate the causal knowledge gathered from data ($\mathbf{CK_D}$) at Tier 2, with a relatively higher weight ($\mathbf{W_D}$) than that of Tier 1 ($\mathbf{CK_E}$). The rationale for this is: (a) data can be collected from different study designs, locations, and corroborated over time, (b) data can be analyzed further with newer methodologies and models, and (c) data can convey the effect of causal relationships between covariates for scientific studies. Because of the higher weights, it contributes more to that conjoined causal model, and can even alter directions of certain causal relationships defined from Tier 1.

### Tier 3: Causal Knowledge from Literature

Tier 3 is causal information from peer-reviewed literature ($CK_L$) and has the highest weight in the CKH. It excludes any knowledge from opinions of experts, without references. Examples of Tier 3 include causal knowledge from peer-reviewed and published literature, systematic reviews, meta-analyses, evidence syntheses, article synopses, and causal effect of interventions published as studies. Within Tier 3, there may be different levels of evidence. The data extraction process additionally falls under the domain of text-mining and natural language processing (NLP). Causal knowledge from literature ($\mathbf{CK_L}$) may have its own biases such as selection bias for inclusion exclusion criteria or transportability bias for differences in population.

### Tier Weights for Causal Knowledge Hierarchy

**Axiom 1** *Each tier of causal knowledge hierarchy (CKH) has individual weights* $(W_E, W_D, W_L)$*, signifying the confidence of the causal information. A higher tier holds a higher weight and provides more robust causal information compared to that from lower tiers. By definition,*

$$CK_E \propto W_E,\ CK_D \propto W_D,\ CK_L \propto W_L$$

Based on the causal information hierarchy proposed, we define three weights

$W = \{W_E, W_D, W_L\}$ for each tier of the causal knowledge hierarchy (*refer to Axiom 1*). The weights are defined such that:

1. $\sum_{\mathbf{i}} \mathbf{W_i} = \mathbf{1}$: By definition of convex combination, the sum of all three tier weights ($\sum W = W_E + W_D + W_L$) is 1.0. A full agreement for a specific edge connection and direction from all three tiers of the CKH results in maximum edge confidence of 1.0.

2. $\mathbf{W_E} < \mathbf{W_D} < \mathbf{W_L}$: The weights are defined in an increasing order. At any time, causal information from one tier can only contribute a maximum of their tier weight. Thus, this increasing score ensures a hierarchy between each tier.

The weights are not fixed values and depend on the specific research question as well as availability of causal knowledge for the specific research question. Practically, the weights are hyper-parameters to the proposition and need to be agreed upon by researchers while generating the structural causal model.

**Assumptions**

For our proposed methodological framework, there are two associated assumptions.

**Assumption 1** *Knowledge within the same tier of CKH does not override one another.*

For conflict resolutions with contradictory causal information within the same tier (such as, $A \rightarrow B$ from dataset 1 and $A \leftarrow B$ from dataset 2, both from $CK_D$),we find the strength of the causal relationship based on all the information within the same tier. Selection of knowledge sources within a tier is subjective and depends on the experimenter. Consequently, we do not propose any hierarchy within a tier, rather, we compute the conjoined strength of the causal connections. A similar direction in causal connections and edges increase the confidence, whereas contradictory causal connections and edges reduce the confidence of the edge.

**Assumption 2** *Within CKH, knowledge from upper tier (or, in special case, tier with more weight) can reverse/ override knowledge from lower tier.*

Unlike the earlier assumption, when we have contradictory causal information from different tiers (such as, $A \rightarrow B$ from Tier 1 ($CK_E$) and $A \leftarrow B$ from Tier 2 ($CK_D$)), the direction of causal relationship from an upper tier can override that from a lower tier.

### 2.3.2 Algorithm

Our algorithm works with the following inputs: problem statement (**PS**), defined keywords (**K**), (empty) structural causal model (**SCM**), pre-defined tier weights (**W**), and inter-rater agreement function (**IRR**). For each of the three tiers of CKH,

we encode specific and relevant causal knowledge within the tier, and systematically update the knowledge base to derive the causal structure.

We start with an empty SCM, with no values assigned to $\mathbf{U}$, $\mathbf{V}$, $\mathbf{F}$ and $\mathbf{P}(\mathbf{u})$ (*introduced and discussed in the supplementary document*). For each individual tier of causal knowledge hierarchy, we follow the steps described below:

1. **Encode tier-specific information:** For each tier, we encode all the information in the causal graph specific to that tier. Specifically, we encode:

   (a) experts' opinion and background knowledge as individual edges, and their confidence score (between 0 and 1.0) in those edges (with directions), in Tier 1 ($\mathbf{CK_E}$),

   (b) edges generated by the causal structure learning algorithms run on the data sets, in Tier 2 ($\mathbf{CK_D}$), and,

   (c) causal relationships extracted from literature as directional edges in Tier 3 ($\mathbf{CK_L}$).

2. **Develop a scoring matrix from encoded information:** From the information encoded for each tier, we build a causal information-based scoring matrix $P$ for the specific tier, with a dimension of $m \times n$. Here $m$ is the number of rows equal to the count of unique pairs of nodes (variables). For $q$ number of total variables ($U \cup V$) in an individual tier, we have

$m = \binom{q}{2} = \frac{q(q-1)}{2}$. $n$ is the number of columns in the scoring matrix $P$. For a specific row in $P$ with the nodes (or variables) $A$ and $B$, we have four columns signifying the type of causal connection between them: i) $A \rightarrow B$, ii) $A \leftarrow B$, iii) no causal connection between $A$ and $B$ and, iv) no causal information available between $A$ and $B$. The complete matrix represents the causal knowledge in the specific tier.

3. **Compute individual edge confidence based on agreement from scoring matrix**: Next we calculate individual edge confidence from each row in the scoring matrix $P$, through plurality voting. For each pair of nodes $A, B$ in $P$, we iterate through rows $i$ of $P$ and use the equation of edge confidence of causal connections between variables $A$ and $B$:

$$q_i = [P_i(n)] \text{ for } n = 0, 1, 2 \tag{2.1}$$

$$e_{A,B} = \frac{max(q_i)}{\sum q_i} \tag{2.2}$$

Here, $e_{A,B}$ signifies the confidence of the causal connection or on the directional edge between variables $A$ & $B$, $q_i$ represents the first three values of row $i$ of scoring matrix $P$ with variables $A$ and $B$.

4. **Estimate agreement score from the scoring matrix:** Using the

inter-rater agreement function ($IRR$), we calculate the agreement score ($\alpha$) (i.e., Fleiss' kappa, varies from 0 to 1) from the generated scoring matrix $P$ (*explained further under background in supplementary document*).

5. **Compute individual weighted edge confidence:** Next, we calculate the weighted edge strength for all edges within a tier, using the equation:

$$weighted\_e_{A,B} = e_{A,B} \times \alpha \times W_i \tag{2.3}$$

Here, $weighted\_e_{A,B}$ signifies the weighted edge confidence of the causal connection between variables $A$ & $B$, $\alpha$ is the agreement score calculated previously, and $W_i$ represents the weight of the specific tier ($W_E$ / $W_D$ / $W_L$). Within a specific tier, only $e_{A,B}$ is different for individual edges, whereas $\alpha$ and $W_i$ remains the same. In the best case, where edge weight for a specific edge and agreement score are both 1.0, the weighted edge strength can be the maximum (the weight of the Tier).

6. **Extract tier-specific insights:** From the generated weighted edge confidences $weighted\_e_{A,B}$, we extract tier-specific causal insights and carry them forward to the next tier. Specifically, in Tier 1 ($\mathbf{CK_E}$), we set a predefined confidence threshold to select edge whitelist. Any edge with edge confidence over the threshold is put in the whitelist and used in the structure learning algorithms in Tier 2 ($\mathbf{CK_D}$). Similarly, in Tier 2 ($\mathbf{CK_D}$), we extract

an incomplete causal structure skeleton $\langle U, V, F \rangle$ and probability distribution $\langle P(u) \rangle$, to carry over to the next tier. Finally, in Tier 3 ($\mathbf{CK_L}$), we extract the complete SCM $\langle U, V, F, P(u) \rangle$, with encoded information from all tiers of the CKH.

7. **Move extracted insights to next step:** Next we get the output of the specific tier, the extracted insights and use them in the next tier as inputs.

**Edge Orientation Step:** In this step, we check for any potential cycles between variables, and re-orient them prioritizing the weighted edge strengths, and generate a complete directed acyclic graph. For this task, we implement the edge orientation process from PC algorithm [111]. For each triplet of nodes, $A - B - C$, we recursively set edge directions using the two rules: (a) if $A$ is adjacent to $B$ and there is a strictly directed path from $A$ to $B$, we then replace $A - B$ with $A \rightarrow B$ (to avoid introducing cycle), and (b) if $A$ and $C$ are not adjacent and we have $A \rightarrow B$ & $B - C$, we replace $B - C$ with $B \rightarrow C$ (to avoid introducing new v-structures). The complete directed acyclic graph, along with the computed $\langle U, V, F, P(u) \rangle$ is the resultant SCM for the problem domain. The algorithmic pseudocode is presented in Algorithm 12.

**Algorithm 1** Structural Causal Model Estimation using Causal Knowledge Hierarchy (CKH)

**procedure** ESTIMATION-BY-CKH($PS$, $K$, $SCM$, $W$, $IRR$)
  Initialize empty confidence for all edges: $CONF \leftarrow \phi$
  Update modified $SCM$ for output: $SCM_m \leftarrow SCM + CONF$
  Select group of experts: $Exp \leftarrow [exp_1, exp_2, exp_3, ...]$ ▷ (a) Tier 1: $CK_E$
  **while** expert $exp_i$ in $Exp$ **do**
    $CR'_{CK_E}[i] \leftarrow extract\_causal\_relationships(exp_i, PS, K, U \cup V = \phi)$
  $U, V \leftarrow get\_vars(CR'_{CK_E})$
  **while** expert $exp_i$ in $Exp$ **do**
    $CR''_{CK_E}[i] \leftarrow extract\_causal\_relationships(exp_i, PS, K, U \cup V)$
  $U, V \leftarrow get\_vars(CR'_{CK_E} \cup CR''_{CK_E})$
  $P \leftarrow create\_grading\_tuple(CR'_{CK_E} \cup CR''_{CK_E})$
  $F, CONF \leftarrow compute\_weighted\_confidence(W_E, P, IRR, SCM_m)$
  $SCM_m \leftarrow update\_scm(SCM_m, [U, V, F, CONF])$
  Using $U \cup V$ and $K$, gather relelvant datasets: $D = [d_1, d_2, d_3, ...]$ ▷ (b) Tier 2: $CK_D$
  Select different causal structure learning algorithms: $SLA = [sla_1, sla_2, sla_3]$
  **while** output $model$ in $SLA \times D$ **do**
    $CR'_{CK_D}[i] \leftarrow extract\_causal\_relationships(model)$
  $U, V \leftarrow get\_vars(CR'_{CK_D})$
  $P(u) \leftarrow get\_probability\_distribution(D)$
  $P \leftarrow create\_grading\_tuple(CR'_{CK_D})$
  $F, CONF \leftarrow compute\_weighted\_confidence(W_D, P, IRR, SCM_m)$
  $SCM_m \leftarrow update\_scm(SCM_m, [U, V, F, P(u), CONF])$
  Using $PS$ and $K$, gather relevant literature: $L = [l_1, l_2, l_3, ...]$ ▷ (c) Tier 3: $CK_L$
  $U, V \leftarrow get\_vars(L)$
  **while** literature $l$ in $L$ **do**
    $CR'_{CK_L}[i] \leftarrow extract\_causal\_relationships(L)$
  $P \leftarrow create\_grading\_tuple(CR'_{CK_L})$
  $F, CONF \leftarrow compute\_weighted\_confidence(W_L, P, IRR, SCM_m)$
  $SCM_m \leftarrow update\_scm(SCM_m, [U, V, F, CONF])$
  $SCM_m \leftarrow orient\_edges(SCM_m)$ ▷ (d) Edge orientation
  **return** $SCM_m = \langle U, V, F, P(u), CONF \rangle$

**Algorithm 2** Computation of Weighted Confidence for Individual Tier

**procedure** COMPUTE_WEIGHTED_CONFIDENCE($W, P, IRR, SCM$)
    Measure agreement score: $\alpha \leftarrow IRR(P)$
    $CONF \leftarrow \phi$
    **while** edge in $SCM$ **do**
        Equation 1 and 2 to find edge confidence: $C_{edge}$
        Equation 3 to find weighted edge confidence: $weighted\_C_{edge}$
        $CONF.append(weighted\_C_{edge})$
    $F \leftarrow extract\_edge\_connections(P)$
    **return** $F, CONF$

## 2.4 Experimental Results

For our experiments, we rely on *a simulation with a standard causal model*. We use pre-defined default values for hyper-parameters, and validate our results with the initial ground truth causal model and provide sensitivity analysis of the SCM.

**Ground Truth Causal Model** We start the simulation with a causal model with a ground truth SCM (Figure 2.5), (a). For this, we refer to the "clgaussian" dataset from 'bnlearn' library [64]. The dataset has 5000 data-points and is generated from a causal model with one normal (Gaussian) variable, four discrete variables and three conditional Gaussian variables. For validation, we assume this initially defined causal model to be the Ground Truth Directed Acyclic Graph (GTDAG).

**Optimization Function** Keywords are defined with a complete set of variables $\{A, B, C, D, E, F, G, H\}$. We define the optimization problem *to identify the causal*

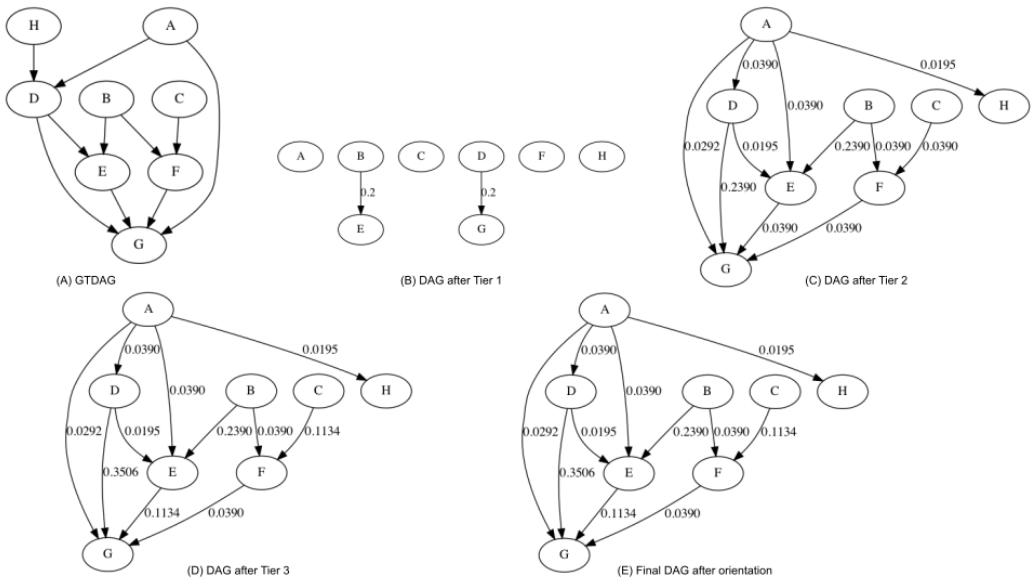



Figure 2.5: Ground-truth Causal Model versus Structural Causal Models with edge confidences at individual tiers of CKH

*effect and relations between variables $D$ and $G$*, along with all associated variables $\{A, B, C, E, F\}$ with the best-fitting GTDAG for the SCM. We also hypothesize that the domain is well-explored *(with sufficient experts, data and literature)* and set the values of weights as $W_{T1} = 0.2, W_{T2} = 0.3, W_{T3} = 0.5$, as well as confidence threshold for Tier 1 $threshold = 0.8$.

**Tier 1: Causal Knowledge from Experts' Opinion** For Tier 1, we consider only one expert $Exp = \{exp_1\}$, and use their knowledge, expressed through causal graphs *(although in reality, we might have multiple experts)*. We go through this step in two phases, the first causal graphs are encoded from each of the experts by describing the problem statements ($PS$) and keywords ($K$). We integrate all the

variables and create a super-set of nodes, $U \cup V$. The second causal graph is derived after providing $U \cup V$, along with both $PS$ and $K$ to the experts. This generates $2x$ causal graphs from experts' opinions based on the problem statement. From these graphs, we generate simplified causal connections:

- $A \rightarrow D$, confidence: 0.5
- $D \rightarrow G$, confidence: 1.0
- $B \rightarrow E$, confidence: 1.0

Based on the pre-defined confidence threshold for Tier 1, we derive a combined causal graph from them. In the combined causal graph in Tier 1, we have $M =< U, V, F, P(u) = \phi, CONF = \phi >$, where $U$ and $V$ comes from the super-set of variables suggested by the experts, and $F$ is defined from the causal relationships suggested by the experts. We calculate the agreement score (using Kappam' fleiss), and measure total confidence values using the agreement score and $W_{T1} = 0.2$. Since we use one expert in this experiment, the agreement score for this tier is 1.0, and we use this value in Equation 3 to compute the weighted edge confidence. We build our scoring matrix $P$ based on the weighted edge confidence higher than the confidence threshold. The resulting causal graph after Tier 1 is presented in Figure 2.5, (b).

**Tier 2: Causal Knowledge from Data** For Tier 2, we generate three separate datasets, $D = \{d_1, d_2, d_3\}$ based on the GTDAG. To simulate a varying number of datasets in real-world, we generate and use three different datasets for which the underlying causal relations are invariant. $d_1$, is generated from variables $\{A, D, G\}$, $d_2$, is generated from only variables $\{B, D, E, G, H\}$, and $d_3$, is generated from all the variables. We use two specific structure learning algorithms, $SLA = \{sla_1, sla_2\}$, with $sla_1$ being **PC algorithm** and $sla_2$ being **MMHC algorithm**. Additionally, we use edge whitelists from Tier 1 for structure learning. We show outputs of all SLAs on datasets in Figure 2.6.

We update the scoring matrix $P$ for this tier and compute weighted edge confidence. We resolve contradictory edges between Tier 1 and Tier 2 by selecting the highest weighted edge confidence of the two. We sum up the weighted edge confidences and the resulting causal graph after Tier 2 is shown in Figure 2.5, (c).

**Tier 3: Causal Knowledge from Literature** In Tier 3, we go through peer-reviewed literature for the problem domain. For this simulation, we assume three causal information sets, $L = \{l1, l2, l3\}$, each of which is extracted from individually published literature. The summary of causal relationships extracted from the literature $L$ is:

1. Literature 1 ($l1$): $D \rightarrow G$, $A \rightarrow D$

2. Literature 2 ($l2$): $F \rightarrow G$, $E \rightarrow G$, $D \rightarrow G$, $C \rightarrow F$

3. Literature 3 ($l3$): $B \rightarrow E$, $C \rightarrow F$, $D \rightarrow E$, $D \rightarrow G$, $E \rightarrow G$

Similar to the previous tiers, we update the scoring matrix $P$ and compute weighted edge confidence as well. We also resolve conflicting edges, *(if any)*, between new causal graph in Tier 3 and original causal graph after Tier 2, depending on edge confidence. Finally, we sum up the weighted edge confidences with previous tiers and the resulting causal graph after Tier 3 is shown in Figure 2.5, (d). A summary of the edges with edge orientations and combined weighted edge confidence is shown in the supplementary document, Table 3.

**Edge Orientation** In the last stage of edge orientation, we see whether any cycles were created in the process. In case one is found, we follow edge orientation process *(as described in method section)* and derive the updated SCM. We present the eventual output in Figure 2.5, (e).

Here we present each individual structural causal models generated through application of structure learning algorithms on the datasets. We have applied $SLA = \{sla_1, sla_2\}$ on dataset $D = \{d_1, d_2, d_3\}$.

**Evaluation** For evaluating our proposed method, we compare the output SCM of the algorithm with the GTDAG. Specifically, we compare edges with directions from

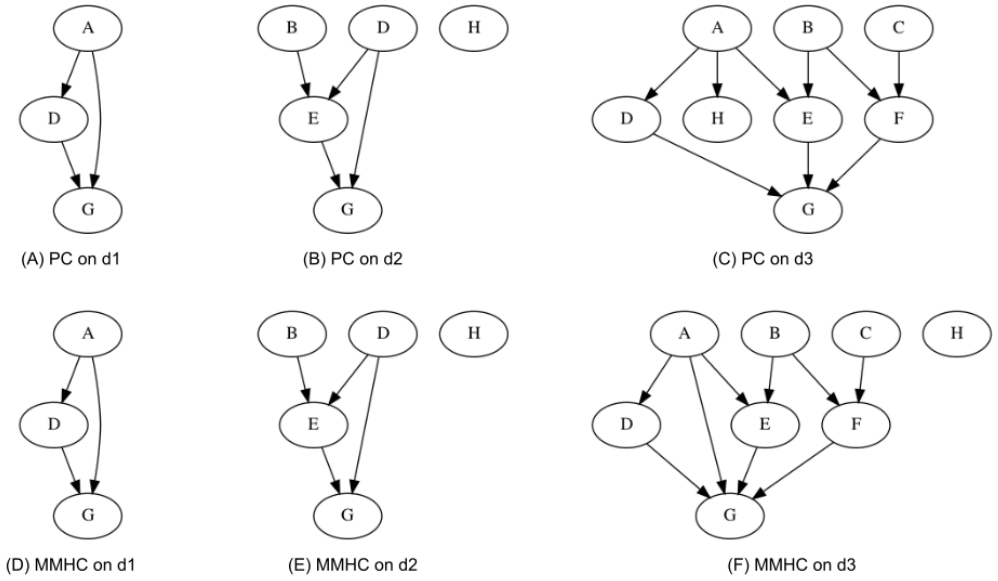



Figure 2.6: Structural Causal Models as outputs of Tier 2 in CKH

our proposed method with that of the GTDAG, and report the average accuracy, along with precision, recall and F1 score. An edge-by-edge comparison of generated output DAG with that of GTDAG is considered as a classification problem [127, 111]. For the simulation, with a node number of $q = 8$, we check for $\binom{n}{2} = 28$ edges' causal directions. On average, our proposed method achieved an average accuracy of **89.29**% (precision: **86.80**%, recall: **89.29**%, f1-score: **87.70**%).

It is possible to incorporate incorrect causal knowledge due to biased opinion, dataset, or publication. For this, we additionally perform sensitivity analysis with perturbed edges within individual tiers. We randomly select three edges and alter their directions. Specifically, we add the following information:

We run the simulation initially without any alteration, and then with

| id | Change in simulation | Ground truth | Altered information |
|---|---|---|---|
| A1 | Add false causal edge | $C \dots D$ | $C \rightarrow D$ |
| A2 | Alter true causal edge | $E \rightarrow G$ | $E \leftarrow G$ |
| A3 | Remove true causal edge | $B \rightarrow F$ | $B \dots F$ |

Table 2.2: Alteration in causal edges in the simulation

| | | Agreement scores | | | Metrics | | | |
|---|---|---|---|---|---|---|---|---|
| Change in edge | | Tier 1 | Tier 2 | Tier 3 | Accuracy | Precision | Recall | f1-score |
| No alteration | | 1.0 | 0.13 | 0.372 | 0.8929 | 0.8680 | 0.8929 | 0.8770 |
| Tier 1 | A1 | 1.0 | 0.13 | 0.372 | 0.8571 | 0.8438 | 0.8571 | 0.8427 |
| | A1+A2 | 1.0 | 0.13 | 0.372 | 0.8214 | 0.8364 | 0.8214 | 0.8244 |
| | A1+A2+A3 | 1.0 | 0.13 | 0.372 | 0.8214 | 0.8364 | 0.8214 | 0.8244 |
| Tier 2 | A1 | 1.0 | 0.124 | 0.28 | 0.8929 | 0.8680 | 0.8929 | 0.8770 |
| | A1+A2 | 1.0 | 0.115 | 0.28 | 0.8929 | 0.8680 | 0.8929 | 0.8770 |
| | A1+A2+A3 | 1.0 | 0.107 | 0.28 | 0.8929 | 0.8680 | 0.8929 | 0.8770 |
| Tier 3 | A1 | 1.0 | 0.13 | 0.319 | 0.8929 | 0.8680 | 0.8929 | 0.8770 |
| | A1+A2 | 1.0 | 0.13 | 0.28 | 0.8929 | 0.8680 | 0.8929 | 0.8770 |
| | A1+A2+A3 | 1.0 | 0.13 | 0.28 | 0.8929 | 0.8680 | 0.8929 | 0.8770 |

Table 2.3: Iterations of simulations with false information injected in each tier

multiple perturbed edge directions. For each tier, we perturb one edge (A1), two edges (A1+A2), and three edges (A1+A2+A3), and report the general accuracy, precision, recall, and F1-score, along with the change in agreement scores in each case. Table 2.3 shows the reported outcomes for each simulations.

In general, with gradual perturbation, performance metrics as well as agreement scores decrease, however this decrease is not drastic, due to the weights of tiers of CKH. For each tier, agreement scores decrease, however the decrease in agreement score does not necessarily alter the outcome. In Tier 1, because of using only one expert, performance metrics decrease more. In other tiers, since there are

multiple knowledge source, wrong information from one source reduces the agreement scores but does not change performance metrics much.

## 2.5 Discussion

Identifying causality is a critical part of many analyses, specifically in clinical research where trust in models is low, and safety and efficacy of clinical decisions is essential. In that context, SCM provides the theoretical foundation for identifying causation from large datasets. However, the lack of methodologies to derive an SCM for estimating causal effect is a fundamental research problem. We have proposed a novel methodology to combine causal knowledge from various sources such as experts' opinion, data, and literature to derive domain-specific accurate SCMs. We discuss the importance of causal information from sources other than just data, and present the rationale behind using hierarchy of causal knowledge. As demonstrated by our experiments, our proposed method (CKH) effectively identifies the most compatible causal models, with higher accuracy and F1-score, from opinions of experts working in the field, outputs of SLAs on existing data, and reported information in peer-reviewed publications. Further discussion is addressed in the appendix.

The CKH-driven algorithm relies on availability and abundance of causal knowledge sources, making it unreliable when there is a lack of causal knowledge from multiple tiers of sources (e.g., CKH-generated causal model from only data will

have low confidence compared to that generated from all three sources). Similar to other open research problems in causal inference, it cannot verify a ground-truth SCM. An alternate to CKH is to not make strong inferences about causal directions to build a DAG and rather derive a Markov equivalence class. Within individual tiers under CKH for a specific problem domain, there is a challenge in finding and selecting experts. Similar difficulty arises within Tier 3 (literature, $CK_L$), since extraction of causal knowledge from literature is itself a research problem under Natural Language Processing (NLP) [131, 62] and is currently being tackled by NLP researchers.

**Rationale with the Theory of SCM** Finding the right and the most-compatible causal model with the underlying data generating mechanism is critical for estimating causal effects through rules of do-calculus. This is a challenging research problem since data itself cannot differentiate between SCM within a Markov equivalence class. It has been proven in a recent seminal work [13] that additional complementary causal knowledge is needed along with data to derive a SCM and the proposed methodology can be a strong tool to aid in that process. Our proposed methodology uses a systematic approach to incrementally derive the SCM with appropriated scoring for different levels of evidence and can generate high accuracy even in the presence of contradictory causal connections. The algorithm would be beneficial for applied causal inference researchers, specially

in epidemiology, medicine and social sciences. Proposed CKH-driven algorithm effectively estimates all necessary components of a computed SCM, $U, V, F$ from all tiers of causal knowledge and $P(u)$ from Tier 2 (data) only. However, it also produces weighted edge confidences $CONF$, which is a key contribution of this algorithm. The edge confidence signifies the strength or confidence of information we have on the specific edge, however it does not state anything about the strength of causal relationships between variables or parameters of functions $F$.

**Completeness of the Derived Causal Model** The causal model estimated through CKH algorithm is built through encoding of causal knowledge sources iteratively and thus holds a summary of causal information from all possible and relevant sources. The rationale for its completeness is that, $\langle U, V, F, P(u) \rangle$, values of four (4) components of the structural causal model derived with the proposed CKH-driven algorithm, is curated from all the tiers of CKH *(and thus contains all necessary and relevant information needed to generate a complete outcome)*. This collaborated information is also weighted appropriately based on the significance and the impact of causal knowledge sources. Another key argument for completeness of the derived causal model comes from the tier weights and their usage. Till now, we have used an increasing weight for the three tiers ($W_E < W_D < W_L$) for a well-researched problem domain, with sufficient experts, data and publications available on the domain specific problem of interest. For an

evolving problem domain (e.g., COVID-19 crisis), where we do not have an abundance of well-established peer-reviewed literature, we can alter and adjust the tier weights as fit for the problem at hand. For example, in estimating the effect of a specific old drug in treatment of COVID-19 patients, we would have more weight on Tier 2 (data, $CK_D$) compared to Tier 3 (literature, $CK_L$), simply because of lack of strong evidence from literature and might use an alternative variation of tier weights hierarchy ($W_E \leq W_D > W_L$). For such reasons, we conjecture that CKH provides a fundamental, necessary, and sufficient building mechanism for constructing structural causal models for a problem domain, given causal knowledge from a varying sources. A rigorous proof for the completeness is still under investigation.

**Limitations of CKH and challenges of individual tiers** The CKH-driven algorithm relies on availability and abundance of causal knowledge sources, making it unreliable when there is a lack of causal knowledge from multiple tiers of sources (e.g., CKH-generated causal model from only data will have low confidence compared to that generated from all three sources). Similar to other open research problems in causal inference, it cannot verify a ground-truth SCM. An alternate to CKH is to not make strong inference about causal directions to build a DAG and rather derive a Markov equivalence class. Within individual tiers under CKH for a specific problem domain, there is a challenge in finding and selecting experts. Similar difficulty arises within Tier 3 (literature, $CK_L$), since extraction of causal

knowledge from literature is itself a research problem under Natural Language Processing (NLP) [131, 62] and is currently being tackled by NLP researchers.

**Application in specific fields of Science** Proposed CKH-driven causal model generation has high impact for specific fields of science, and especially in health science. Identifying the cause for an outcome and quantifying the causal effect is of high importance in health science and epidemiology. An ongoing work is aiming to derive a SCM for the treatment of delirium patients in the ICU [15] based on the CKH. Other than that, CKH has implications in other branches of science, where the notion of causality is critical, such as, sociology and finance.

# CHAPTER 3
# A Causally Formulated Hazard Ratio Estimation through Backdoor Adjustment on Structural Causal Model

## 3.1 Introduction

Experimental studies such as randomized controlled trials (RCT) are considered the gold-standard in hypothesis testing. For safety and efficacy reasons and regulatory purposes, most new drugs or treatments are studied through RCTs [38]. RCTs provide the best mechanism to identify the causal effect of treatments or interventions, by adjusting for observed and unobserved confounders under the rubric of a potential outcome framework [29]. Despite clear advantages of RCTs in drug-trials, in practice, they are expensive, time-consuming, and not feasible in many cases due to ethical reasons. Other issues with RCTs include low recruitment rate, loss to follow-up, insufficient sample size, and being prone to selection bias [83, 31]. While RCTs remain the best way to establish causation, large amounts of data captured with new technologies during routine healthcare (e.g., electronic health records (EHR) or wearable devices), colloquially termed *big health data*, has the potential to discover causal effects from observational studies to complement RCTs. With proper methodological considerations, observational studies can provide a way to *emulate* RCTs and go beyond statistical correlation [47, 49].

In the 1970s, the potential outcome framework was extended to observational

studies to identify causal relationships from observational data through the Rubin Causal Model (RCM) [105, 101, 54]. Recent advances in structural causal model (SCM) provides the methodological framework under the potential outcome framework for graphically formalizing the identification of causal effects from observational and experimental data [89, 95]. SCMs can be used to *emulate* RCTs from observational data in many cases if the graphical model is identifiable [14], which signifies the capability of estimating the interventional distribution ($P(y|do(x))$) from the available data with the assumptions incorporated in the model.

Experimental studies (including RCTs) frequently explore and report survival analysis measures. Survival analysis is the branch of statistics that analyzes the expected duration of time-to-event with outcome statistics such as hazard ratio, odds ratio, and risk ratio. Survival analysis has been well-studied under the potential outcome framework with experimental studies and with RCM for observational studies [20, 45]. Recent research has also studied survival analysis with RCM for observational studies considering the data generating mechanism or the study designs to estimate outcome statistics such as hazard ratio, odds ratio, risk ratio, and risk difference [24, 51].

Commonly reported outcomes from survival analysis in experimental clinical studies include the survival curve and hazard ratio (HR). The survival curve

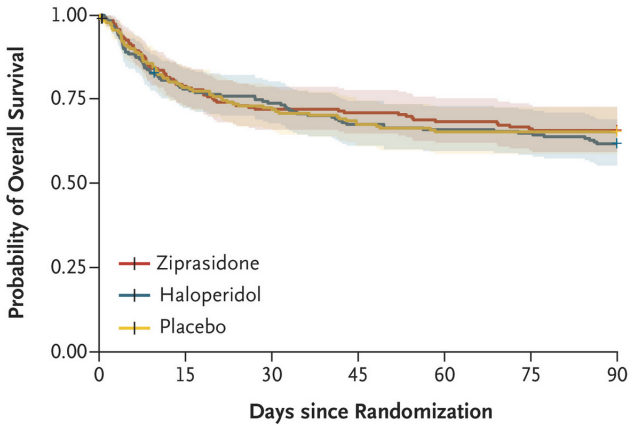



Figure 3.1: An example survival curve, collected from Girard et al. ([34])

graphically reports the hazard in a population and represents the fraction of the population that survived in the treatment and the control group over time. HR describes the comparative hazard between the treatment and the control group. Hazard function, or simply *hazard* signifies the rate of events-of-interest (e.g., a death) at time $t$, conditional on survival until time $t$ and beyond [123]. For example, we present a survival curve (Figure 3.1) as reported in [34], where probability of overall survival of patients in drug groups (starting at 100%) is presented with time passed, and the probability declines with time.

Even though HR is widely used in practice as a standard tool for comparative evaluation of the outcome between treatment and control groups, it depends on the length of the study and, by definition, has an inherent selection bias (since only the survived population at time $t$ are *selected* at time $t+1$) [45]. In addition, both the survival curve and HR do not consider the study design, that is

RCT versus observational study, in their formalization. Consequently, it is difficult to interpret the results of an intervention from only the reported hazard ratio [45] and compare different studies with varying study designs and time lengths. The researcher has to consider the design of the study, length of the study as well has the hazard ratio to understand the effectiveness of the treatment. Structural causal models (SCMs) provide a framework to explicitly define the design of the study, the assumptions for the study, as well as the length of the study. However, to the best of our knowledge, a framework to compute the hazard ratio with SCMs does not exist.

Previous approaches for adjusted survival curves under the rubric of RCM used inverse probability weighting (IPW) to adjust for confounders in the estimand [20]. However, this approach has a strong assumption, namely *ignorability* [106, 9]. The ignorability assumption states that there are no unobserved confounders in the model, and the variables considered for IPW satisfy the backdoor criterion. Although an approach with instrumental variable can be used when the treatment assignment is non-ignorable [9], in practice, this is rather a strong assumption and a variable can be a mediator, a collider, an M-bias, or a confounder [68]. In this paper, we formulate the estimation of the hazard ratio from observational studies under the rubric of SCMs that does not depend on the ignorability assumption. We provide a principled approach to define observational studies using SCMs, redefine with time-specific survival as outcomes (instead of survival time as the only

outcome), and therefore mathematically transform observational studies to the corresponding experimental studies by adjusting for confounders with the backdoor criterion and then, sample from the experimental studies to estimate hazard ratios. We provide the mathematical formalization of the approach with a simple causal graph and with detailed mathematical derivation, and validate the results with a simulated data set and a benchmark data set on Ewing's sarcoma.

### 3.1.1 Clinical Relevance

Most clinical research reports HR with survival analysis. However, the reported HR and its process of calculation do not take into account the study design (e.g., RCT vs. observational study) and corresponding assumptions (e.g., ignorability). This makes it harder to compare the results of different studies with different study designs, sample populations, study lengths and assumptions. Our proposed method with SCMs estimates HR by explicitly describing the study designs and assumptions for a better clinical understanding of the effect of the treatment of interest.

### 3.1.2 Technical Significance

We propose a novel approach to estimate the HR from observational studies with SCM, taking the causal relationship between treatment and outcome into account. In HR calculation for survival analysis of observational studies, our review of the literature identifies a lack of causal interpretation. Our proposed approach first

develops a time-invariant causal model and estimates the survival time after adjusting for the confounders in the SCM using backdoor adjustment and do-calculus. The development of an SCM enables us to identify the confounding variables, unlike with the ignorability assumption where we adjust for every variable available (except treatment and outcome), as well as properly adjust using the minimal set, thus reducing computational requirements. The computed survival times are considered "as-if" they were sampled from an RCT. The newly adjusted survival times are capable of expressing the true causal effect of treatment on the outcome through the survival curve and HR. We validate the proposed method in both simulated experiments and with observational data.

### 3.1.3 Generalizable Insights

We propose a novel method of estimating the HR for observational studies under the rubric of SCMs. The method can be used for any observational studies with survival data, after defining the SCM. Our method of estimating the HR through SCMs clearly defines the study-design and assumptions in the model. All the source code for this study is shared with the research community through a GIT repository. A Python-based library has been released that takes the data, the graph, and length of the study as input and provides the adjusted survival curve with backdoor adjustment and the hazard ratio as the output. Our approach is limited in the cases

when i) the SCM is not defined and ii) the SCM is not identifiable through the adjustment formula or backdoor adjustment (i.e., there is no backdoor set).

## 3.2 Related Work

Survival analysis [63] is a methodological approach for modeling and comparing the time-to-event between two populations. The event is called a hazard, which can be death, an adverse clinical event, or a mechanical failure for physical systems. It compares the condition of survival in the treatment versus control group, and reports outcomes with statistical measures such as the HR. Frequently reported approaches in survival analysis include Kaplan Meier survival curve, Cox proportional hazards model, life tables, and survival trees,. We review a non-parametric approach of the Kaplan Meier survival curve and the semi-parametric approach of the Cox proportional hazards model.

The Kaplan Meier survival curve [61] is a non-parametric statistic representing the survival function and HR in the treatment and the control group. It provides a visual comparison between survival functions in different treatment or control groups; it does not differentiate between RCTs or observational studies. Data from both of the approaches can be plotted as the Kaplan Meier curve. It is up to the individual researcher to interpret and explain the Kaplan Meier curve based on the study design. Cox PH model, on the other hand, is computationally complex. However, it is a commonly used approach for survival analysis, and is

widely used to compute the HR in epidemiological studies. The key aspect of it is the underlying proportional hazards assumption [22], stating that the HRs of the treatment and control group are proportional and is a function of the covariates. It is a semi-parametric model since no assumption is made about the baseline hazard function (i.e., hazard function with no covariates). In general, it is effective in estimating both regression coefficients ($\beta_i$) and the HR [63]. Futher, it is unbiased *(when estimated considering all possible covariates).*

We review existing approaches to compute the HR for observational and experimental studies. Previous work on survival analysis for observational data with RCM under potential outcome used IPW to adjust for confounders [20]. However, RCM requires the ignorability assumption that all variables considered for adjustment with IPW satisfy the backdoor criterion. In reality, a variable can also be a mediator, and in those cases adjusting for the mediators will result in inaccurate analysis. It has also been shown that the HR estimation approach has an inherent selection bias [48, 45] as only the patients who survived at time $t$ were sampled at time $t + 1$ to be considered for the estimation. SCMs provide the mathematical machinery to identify the backdoor variables given a causal graph. We used the same Ewing's sarcoma data set as studied in [20] with the same assumptions (i.e., all the covariates satisfy the backdoor criterion) to arrive at the same result as a validation strategy for our approach.

For survival analysis, it was shown that in some cases the Kaplan Meier curve may show no difference between the treatment and control groups when in reality there is a statistically significant difference in the HR, if it is adjusted properly [75]. The rationale behind this phenomenon is that a non-parametric approach is used to plot the survival curve, whereas a semi-parametric method is used to calculate the HR. The authors [75] presented an approach to construct a plot of the survival curve consistent with the HR calculated. In this work, the adjusted survival curve for a specific treatment group was estimated by calculating a mixture of the estimated survival functions for separate strata, and weighted based on the distribution of the covariate in the sample dataset. However, the approach does not consider the design of the study in the survival analysis.

To extend the existing definition of the Cox PH Model, the Marginal Structural Cox PH Model has been introduced and used to find the effect of Zidovudine on the survival of HIV-positive men [52]. Statistical analysis in the presence of time-dependent confounders is commonly done through a standard Cox PH model. However, Robin [99] has previously shown that this approach cannot adjust for all biases. Similar to previous work under the RCM, the authors used the conditional ignorability assumption. This is a much stronger assumption compared to using the SCM to identify confounding variables opening the backdoor. Several

other researchers [109, 103] have used the IPW approach, although without using SCMs.

The existing literature to compute the HR does not consider the study design and might lead to misinterpretation if the data were not sampled correctly or adjusted for the right confounding variables. While previous research alludes to this problem, they do not provide the mathematical machinery for survival analysis. Although the traditional Cox PH Model can minimize the effects of biases, it is not the same as "adjustment" of confounding variables. The bias is reduced by fitting the Cox PH regression model until convergence [22], it does not consider the study design or the data generating mechanism. The model fitting approach does not generate a causally meaningful interpretation despite reduction in biases. Our goal is to formulate an approach that estimates the HR through a causal formulation considering the data generating mechanism with SCM, that portrays the direct causal effect of treatment on outcome, in terms of the HR metric. The assumption of variables opening the backdoor path in the SCM as confounders and adjustment on the dataset based on that enables a more causally interpretable estimation of the HR.

## 3.3 Background

### 3.3.1 Hazard Ratio

To define the HR, we use the hazard function [123] in the Cox proportional hazard model:

$$h(t, \mathbf{X}) = h_0(t) \exp(\sum_{i=1}^{p} \beta_i X_i) \tag{3.1}$$

Based on this, the Hazard Ratio (HR) is defined [63] as:

$$HR = \frac{h(t, \mathbf{X}_{x=1})}{h(t, \mathbf{X}_{x=0})} \tag{3.2}$$

Here, $h(t, \mathbf{X})$ represents the hazard function at time $t$ and the vector with the covariates of the model $\mathbf{X}$. $\mathbf{X}$ can also be written as $[w_0, w_1, ..., w_m, z_0, z_1, ..., z_n, x]$, where $x$ is the treatment, $z_i$ are the confounders, and $w_i$ are the other associated covariates. $\mathbf{X}_{x=1}$ represents the value of the covariate vector $\mathbf{X}$ with value of the treatment set as 1 ($x = 1$), making $\mathbf{X}_{x=1} = [w_0, w_1, ..., w_m, z_0, z_1, ..., z_n, 1]$. $\beta$ represents the *maximum likelihood estimates (MLE)* for each covariate. In other words, $\beta$ is the corresponding coefficient for each covariate that fits the data into a converging model for the Cox regression.

As expressed in Equation 3.1, the proportional hazard assumption defines

the hazard function $h(t, \mathbf{X})$ to be composed of the baseline hazard function $h_0(t)$ (i.e., hazard when all covariates are set to 0), multiplied with the exponential of the sum of $\beta$ multiplied by the corresponding covariate.

Since we have defined the HR and hazard function, we can simplify the equation of the HR to:

$$\begin{aligned} HR &= \frac{h(t, \mathbf{X}_{x=1})}{h(t, \mathbf{X}_{x=0})} \\ &= \frac{h_0(t) \exp(\beta_x 1 + \beta_z Z + ...)}{h_0(t) \exp(\beta_x 0 + \beta_z Z + ...)} \\ &= \exp(\beta_x) \end{aligned} \tag{3.3}$$

In other words, the HR is equivalent to the exponential of the regression coefficient $\beta$. However, computing $\beta$ is non-trivial since, in practice, one does not know the baseline hazard function *($h_0(t)$)*. We can only estimate the HR using the maximum likelihood function, and iterating until the model converges to a pre-defined error bound [63].

Although the HR is an important outcome, it has limitations in explaining causal relationships. No causal mechanism is understood from the HR. This is because the HR is calculated from the convergence of regression models and, confounding and other such bias is handled by simply including the covariates to the model. It is then up to the individual researcher to make sure that the right data are used to measure the HR and interpret accordingly. For example, an HR

calculated from an RCT provides the casually linked hazard for the intervention, whereas the same HR calculated from an observational study simply provides a correlated hazard. This existing approach simplifies the calculation and reduces the burden on the researcher. However, we frequently find differences between the survival curve and the HR. This difference, or bias, arises because of the inherent definitions of the survival curve and Cox PH model.

### 3.3.2 Structural Causal Models

Structural causal models (SCMs), developed on the foundations of probabilistic graphical models, draw inferences that explain the causal relationship between variables. With an SCM, a causal model is defined first and is expressed with a graphical representation. Definition 1 gives the formal description of an SCM: [14, 89].

**Definition 2 (Structural Causal Model)** *A structural causal model $M$ is a 4-tuple $\langle U, V, f, P(u)\rangle$ where:*

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

2. *$V$ is a set $\{V_1, V_2, ..., V_n\}$ of observable (endogenous) variables that are determined by variables in the model (i.e., determined by variables in $U \cup V$ ),*

3. *$F$ is a set of functions $\{f_1, f_2, ..., f_n\}$ such that each $f_i$ is a mapping from the respective domains of $U_i \cup PA_i$ to $V_i$, where $U_i \subseteq U$ and $PA_i \subseteq V$ $V_i$ and the entire set $F$ forms a mapping from $U$ to $V$. In other words, each $f_i$ in $v_i \leftarrow f_i(pa_i, u_i), i = 1, ..., n$, assigns a value to $V_i$ that depends on the values of the select set of variables $(U_i \cup PA_i)$, and*

4. *$P(u)$ is a probability distribution over the exogenous variables.*

An SCM is often expressed by a causal graph $G$. Each node $V$ in $G$ represents an observed or unobserved variable, and each directed edge $E$ represents the causal relationships between them. To find the causal effect of variable $X$ on variable $Y$, do-calculus is introduced [95]. Do-calculus is used to map the observational reality to the corresponding experimental reality with the identifiability equation by adjusting for different kinds of biases (e.g., confounding bias), if it exists. The backdoor criterion provides a powerful tool to identify the variables that need to be adjusted for this transformation (in other words, adjust for confounding bias) and is defined in definitions 2 and 3.

**Definition 3 (Backdoor Criterion)** *Given an ordered pair of variables $(X, Y)$ in a directed acyclic graph $G$, a set of variables $Z$ satisfies the backdoor criterion relative to $(X, Y)$ if no node in $Z$ is a descendant of $X$, and $Z$ blocks every path between $X$ and $Y$ that contains an arrow into $X$.*

**Definition 4 (Backdoor Adjustment)** *If a set of variables $Z$ satisfies the back-door criterion relative to $(X, Y)$, then the causal effect of $X$ on $Y$ is identifiable and is given by the formula: $P(y|do(x)) = \sum_z P(y|x, z)P(z)$*

### 3.3.3 Problem Definition

Our research problem is to develop a method to compute the HR for observational studies by leveraging the SCM by explicitly declaring our assumptions and adjusting for the right confounders. The goal is to acknowledge the defined roles of variables in the SCM, and use a minimum set of confounders to adjust for backdoor, thus building a computationally-efficient and more accurate model for objective estimation and comparison. The algorithm will take three sets/inputs, (1) observational dataset $D$ consisting of treatment, outcome in survival-time and other covariates, (2) SCM supporting the causal mechanism and dataset, $G$, and, (3) length-of-trial $T$. At the completion of the algorithm, the output will be: (1) adjusted survival curve $S$ *(non-parametric estimation)*, and (2) hazard ratio of treatment, $HR$ *(semi-parametric estimation)* (Figure 3.2). The assumption in our approach is that the observational data are available, and the SCM is fully specified.

## 3.4 Methods

In this section, we formalize our approach to mathematically transform the time-dependent observational data to the corresponding experimental data by

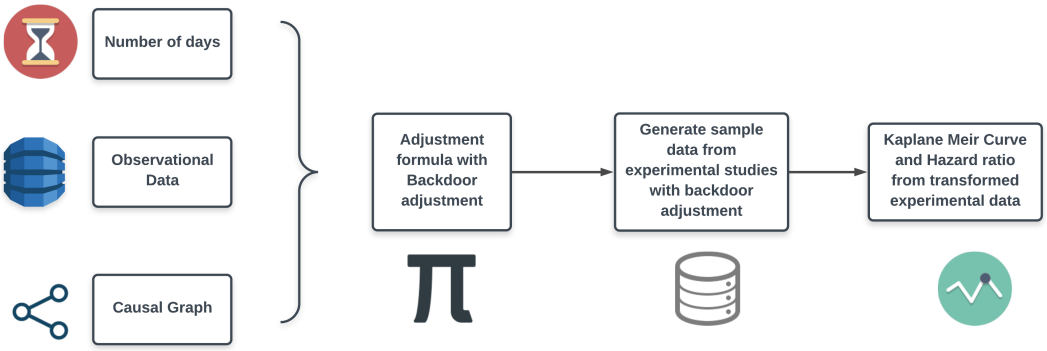



Figure 3.2: Schematic overview of the proposed approach

leveraging the SCM. We then use the adjusted dataset for estimating HR using Cox PH Model. Our proposed approach focuses on causal effect of treatment on outcome to measure HR. We start with an observational study scenario and define all related assumptions. The schematic diagram for the proposed approach is shown in Figure 3.2, with observational data, corresponding causal diagram and the length of study is provided as input. The approach first uses backdoor adjustment to create sample data from experimental study, and then computes the hazard ratio from the sampled experimental data.

### 3.4.1 Assumptions

We assume a simple observational study for a population, consisting of treatment $X$, confounding variable $Z$, and outcome in survival time $T$. In this scenario, treatment $X$ is a dichotomous variable ($X = 1$ signifying treatment and $X = 0$ signifying control). Outcome $T$ is the survival time from the beginning of the study

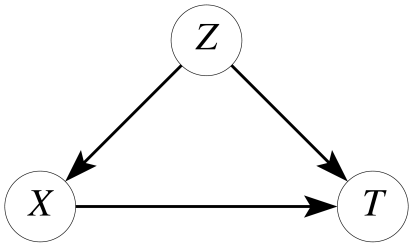



Figure 3.3: Simple observational study (treatment $X$, outcome in survival-time $T$ and single confounder $Z$)

and is a continuous variable in time units. Although the confounding variables can be a categorical or continuous variable, for simplicity, we assume the confounder $Z$ to be a dichotomous variable. This observational study can be expressed as an SCM and with a graphical form $G$ through causal directed acyclic graph (causal DAG) in Figure 3.3, where treatment, confounder, and outcome is expressed by the nodes $X$, $Z$ and $T$ respectively.

From the definition of the SCM, we can express the underlying functions defining the causal relationships between variables by: $Z \leftarrow f_z(U_z)$, $X \leftarrow f_x(Z, U_x)$, $T \leftarrow f_t(Z, X, U_t, h_0(t))$. Here, $U = \{U_x, U_t, U_z\}$ is the set of exogenous variables, $V = \{Z, X, T\}$ is the set of endogenous variables, $f = \{f_z, f_x, f_t\}$ is the set of structural functions.

- $f_z(U_z)$ shows that confounder $Z$ is independent of any other endogenous variables.
- $f_x(Z, U_x)$ expresses the dependency of $X$ on $Z$. Since $Z$ is parameter for both

functions $f_x$ and $f_y$, $Z$ imposes a bias on the model ($P(X|Z=0) \neq P(X|Z=1)$), and the function $f_x$ defines whether the bias is strong or weak.

- $f_t(Z, X, U_t, h_0(t))$ defines the effect of $X$ and $Z$ on the survival time $T$. This function also depends on the baseline hazard function $h_0(t, \mathbf{X})$ since this defines the rate of decline in survival.

We also assume to know the sample size of population $n$ and a maximum length of survival time $t_{max}$.

### 3.4.2 Approach

**Transformation of single study to multiple studies**

Experimental studies commonly have different study time-lengths, e.g., different number of days as the outcome endpoints (e.g., 30-day survival, 90-day survival, etc.). This variable is a dichotomous variable and describes a patients' status of survival at the end of the study. While analyzing a study similar to these, we do not take into account survival at each day, or survival after end-of-trial day, since we do not have the opportunity to do so. In our problem definition, we only have the survival time of individuals; however there is no defined end time for the trial. From the individual survival time, We can easily get the $i$-th day survival of every individual in the dataset, $i$ being the number of days from the beginning of the

study. We use *days* as smallest unit of time, since we assume the dataset reports survival in units of days. However, it could be any other units of times (*e.g. minutes, or weeks*) depending on the problem domain and dataset.

Since our observational study has a maximum survival time of all individuals $t_{max}$, we assume we calculate the variables $Y_i$, signifying the $i$-th day survival, $i$ ranging from 0 to $t_{max}$. Conversion of continuous variable $T$ describing survival time into multiple variables $Y_i$, each describing survival at the $i$-th day, essentially breaks down the single observational study into $t_{max}$ number of observational studies with variables $X$, $Z$ and $Y_i$, each of which is now a dichotomous variable.

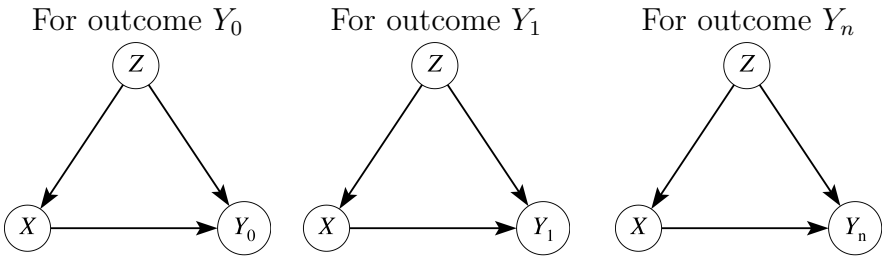
Through the transformation, from a single SCM $G$, we end up with $n$ different SCMs, each with the same treatment $X$ and the confounder $Z$, but different outcome (survival at $i$-th day). Note that, in our assumption, the causal graph is time-invariant, i.e., the functional relationship between the variables does not change over time. This conversion is represented by $n$ different SCMs (Transformed graphs A, Figure 3.4 (a)), where $n \geq t_{max}$.

An important point to note here is that, the single confounder $Z$ and treatment $X$ from the original observational study is not being transformed, only the outcome is distributed into multiple variables. In other words, we assume a point intervention and the confounding variables are invariant in time. And since we are transforming from a single trial to multiple trials, the outcomes $Y_i$s of these

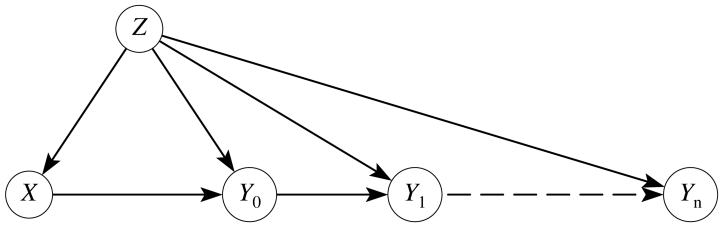
separate trials are not conditionally dependent on each other *(e.g. a RCT with 30-day survival as outcome does not analyze about whether any patient died at 20th day or 29th day.)*.

However, in extracting information from obs. data, there is dependency between them. Specifically, $Y_i$ has causal effect on $Y_j$ (where $j > i$), since if $Y_i$ is 0 *(e.g. patient died at i-th day)*, all $Y_j$ (where $j > i$) is 0 *(e.g. patient remains dead for all consecutive days)*. Also, $Y_i$ only has direct causal effect on $Y_{i+1}$, every other corresponding effect is mediated through. If $X$ has causal effect on $Y_i$, it is mediated through $Y_{i-1}$. For example, $X \perp\!\!\!\perp Y_1|Y_0$, in absence of any backdoor variables. Reasoning behind this assumption is that, without having any underlying effect of treatment on outcome at $i$-th day, subject is suddenly prone to hazard on $i + 1$-th day. For example, this is unlikely that, if a subject is advised a treatment (drug), the subject has no hazardous effect until 10-th day and suddenly finds a hazardous effect on 11-th day. It is possible that the subject does not show any symptom on 10-th day, or we cannot measure the internal hazardous effect of the drug on 10-th day (due to lack of symptoms).

The relationship between $Y_i$s is reflected through a single transformed SCM (Transformed graph B, Figure 3.4 (b)), where $n \geq t_{max}$. The similarities between transformed graphs A and transformed graph B is that they both have same $Z$ and $X$, and the dissimilarities are:




(a) Converted Causal DAGs with no dependency between $Y_i$s




(b) Converted single Causal DAG with dependencies between $Y_i$s

Figure 3.4: Converted Causal DAGs with survival time converted to binary outcome of survival at different timepoints

1. Transformed graphs A portrays $n$ different trials with different outcomes, whereas transformed graph B is a single trial.

2. For transformed graphs A, $Y_i \perp\!\!\!\perp Y_j$ (where $j \neq i$), however for transformed graph B, $Y_i \not\perp\!\!\!\perp Y_j$ (where $j \neq i$).

3. Since two causal DAGs are different, transformed graphs A and transformed graph B have two different equations for $P(Y_i|do(X))$.

In summary, we transform the single observational study into multiple different trials expressed through two different transformations (transformed graphs A and transformed graph B, Figure 3.4), each with the same treatment $X$ and confounding $Z$, but with different survival time as the outcomes, as the death (or

failure) increases over time. These outcomes are the status of survival (or death) at $i$-th day, where $i$ is 0 to $n$ ($n \geq t_{max}$).

**Generation of Survival Curve**

Applying Backdoor adjustment formula in transformed graphs A, the causal effect of $X$ on $Y_i$ (for all $n$ causal graphs) is:

$$P(Y_i|do(X)) = \sum_{Z} P(Y_i|X,Z)P(Z)$$

In transformed graph B, the causal effect of $X$ on $Y_i$ is:

$$P(Y_i|do(X)) = \sum_{Z,Y_{i-1},...Y_1,Y_0} \left( \prod_{k=0}^{n} P(Y_k|Y_{k-1},...,Y_0,X,Z) \right) \cdot P(Z)$$

Since $P(A|B,C)P(B|C) = P(A,B|C)$ *(using rules of conditional probabilities)*, we can reduce this equation to,

$$P(Y_i|do(X)) = \sum_{Z,Y_{i-1},...Y_1,Y_0} P(Y_i,Y_{i-1},...,Y_0|X,Z)P(Z)$$

Finally, for $j <= i$, $P(Y_j = 1|Y_i = 1) = 1$ *(e.g. if a person is alive at 30th day, he has been alive for the last 29 days as well)*,

$P(Y_i = 1, Y_{i-1} = 1) = P(Y_{i-1} = 1|Y_i = 1)P(Y_i = 1) = P(Y_i = 1)$, which reduces our

equation down to the same as that of transformed graphs A:

$$P(Y_i = 1|do(X)) = \sum_Z P(Y_i = 1|X, Z)P(Z)$$

This signifies whether we use transformed graphs A or transformed graph B, we end up with same adjustment formula.

For each of the newly transformed causal DAGs, we can now adjust for the confounder using the backdoor adjustment formula. We calculate adjusted probabilities $P_{adj}$ and thus adjusted counts $C_{adj}$ for each of the $n$ causal graphs. Using the values of $P_{adj}$, we generate survival curve with Kaplan Meier fitter.

**Calculation of Hazard Ratio**

Since we calculated $C_{adj}$ for each of $n$ causal graphs, we know number of adjusted individuals alive at each unit (day) of time. This helps us build back the adjusted survival time $T_{adj}$ for individuals, as it was in the original dataset. The newly calculated survival time $T_{adj}$ is adjusted for the confounding bias, as if they were sampled from an RCT. We measure the HR using Cox PH model with the adjusted survival time $T_{adj}$ as outcome.

**Algorithm**

Algorithm 1 generates adjusted Kaplan Meier curve as well as the HR from Cox PH model on the adjusted dataset. The input for the algorithm is the dataset, specifically, confounder $Z$, treatment $X$, survival time $T$, and event status $S$. In the algorithm, variables in uppercase letters signify vectors, and variables in lowercase signify single variables. Internal procedures *convert_single_to_multiple_trials* (Algorithm 2) are shown separately.

**Algorithm 3** Causally Formulated Hazard Ratio Estimation

**procedure** CFHRE($Z, X, T, S$)
    global $n \leftarrow length(T)$
    global $t_{max} \leftarrow max(T)$
    $Y_i \leftarrow convert\_single\_to\_multiple\_trials(T, S)$
    **while** $i \leftarrow 0$ to $t_{max}$ **do**
        $adj\_p_i \leftarrow \sum_Z P(Y_i = 1 \mid X, Z)P(Z)$
        $adj\_c_i \leftarrow adj\_p_i * count(X)$
    $survival\_curve \leftarrow plot(time, cumulative(adj\_p_i))$
    $adj\_X, adj\_T \leftarrow convert\_multiple\_to\_single\_trial(adj\_c_i)$
    $model \leftarrow cox\_ph\_model(adj\_X, adj\_T)$
    $HR_{drug} \leftarrow exp(model.\beta_{drug})$
    **return** $survival\_curve, HR_{drug}$

**Algorithm 4** Conversion of single trial to multiple trials

**procedure** CONVERT_SINGLE_TO_MULTIPLE_TRIALS($T, S$)
    **while** $i \leftarrow 0$ to $t_{max}$ **do**
        **while** $j \leftarrow 0$ to $n$ **do**
            $Y_i[j] \leftarrow ((T[j] <= i)\ \&\ (S[j] = 1))\ ?\ 0 : 1$
    **return** $Y_i$

## 3.5 Experiments and Applications

We evaluated the proposed approach for computing HR and visualizing survival curves with an experimental and observational dataset: (1) a synthetic dataset derived from a linear acyclic model with Gaussian noise; (2) a real-world dataset on disease-free survival in patients with Ewing's Sarcoma [75]. The rationale to consider these two datasets are: (1) both of the underlying causal model has a backdoor path through confounders, and, (2) both these datasets have treatment and control group that satisfy the proportionality hazarads assumption.

### 3.5.1 Experimental Data

We simulate an observational study with $n = 200$ patients. A subgroup of the patients received a treatment ($X = 1$), and the remaining patients did not ($X = 0$). We generate data on survival time $T$ (in days) defined as the outcome. The treatment assignment is confounded by sex (e.g. $Z = 1$ for female, $Z = 0$ for others). The scenario has a causal model as depicted in Figure 3.3.

For the simulation, we generated outcome variable survival-time through defining a baseline hazard function. We defined survival time to be exponentially varying with time, in the form of: $T \leftarrow a.exp((b + cZ + dX + eZX) * i) + E$, with $Z$ being confounder, $X$ being treatment, $E$ being the noise/error and $i$ being the index of patient. The other parameters were set to

$a = 5, b = 0.025, c = 0.005, d = -0.015, e = 0.075, E = U(-0.5, 0.5)$, they were selected such that the HR remains close to 1, however injection of bias through $Z$ portrays different outcome in survival curve.

We simulate the study with a strong biased effect from confounder $Z$. We define the strength of bias by an imbalance of conditional probabilities in each stratum of $Z$ through the function $f_x(Z, U_x)$. For the defined strong bias case, $P(X = 1|Z = 0) = 0.75$ and $P(X = 1|Z = 1) = 0.25$. It translates to, if $Z = 0$ stands for females in this trial, 75% received the drug, whereas, in $Z = 1$ or others, only 25% received the drug. In a randomized controlled trial, under a no-confounding-bias scenario, we should have
$P(X = 1|Z = 0) = P(X = 1|Z = 1) = 0.5$.

After we generate the experimental data, we applied Algorithm 1. We compared the existing approach of survival curve and survival curve from the adjusted dataset side-by-side in Figure 3.5. Figure on left shows significant difference in survival curve between treatment and control group. The treatment population ($X = 1$) seems to be more prone to hazard compared to the control population ($X = 0$). Figure on right shows adjusted survival curve to be overlapping, signifying no significant difference in hazard rate in both the treatment and the control population.

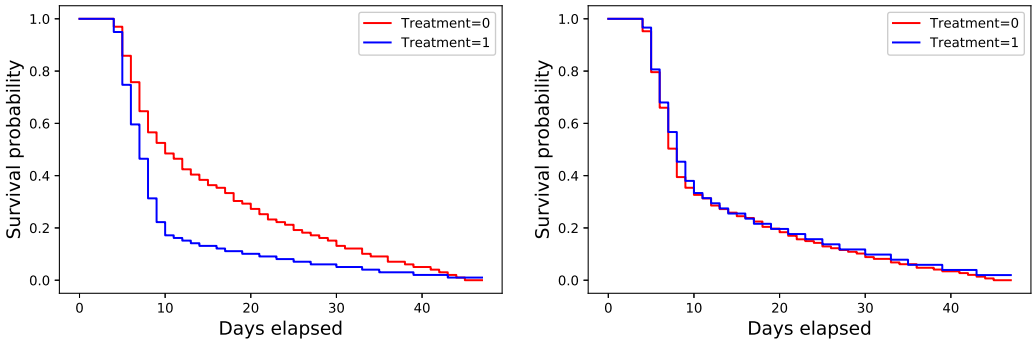



Figure 3.5: Unadjusted survival curve for simulated data *(left)* and, survival curve generated after applying proposed approach *(right)*

| | Existing model | | Proposed model |
|---|---|---|---|
| | Observational data excluding confounding variable (biased estimate) | Observational data including confounding variable (traditional approach) | Transformed and adjusted data |
| Hazard Ratio | **1.66** | **0.80** | **1.00** |
| *(95% Confidence Interval)* | ***(1.25-2.20)*** | ***(0.57-1.12)*** | ***(0.76-1.33)*** |

Table 3.1: Hazard Ratio for simulated dataset, calculated using existing model and our proposed approach

Table 3.1 presents the HR found in the fitted Cox PH model in three different processes:

1. using only the treatment and outcome from the original dataset,
2. using data of all covariates (treatment, outcome and confounder) from the original dataset, and
3. using only the treatment and outcome from the adjusted dataset following our proposed approach.

First column reports a biased estimate of HR, by using only treatment and outcome (excluding confounder) in Cox PH model. The second column reports a

standard estimate of HR, by including all known variables (including confounder). The third column reports HR calculated in our proposed approach, using only treatment and outcome (excluding confounder). The first approach represents scenarios where: 1) we ignore confounding, assuming it does not impact the treatment, or, 2) we do not possess data on the confounding variable (unmeasured confounding). This approach, however, results in an incorrect approximation of the HR. The second approach represents the existing approach to calculate HR. The third one shows our approach, and it eliminates the need for using confounding in model fitting since we are already adjusting for that.

Here, in Figure 3.5, the difference in unadjusted survival curve is similar to fitting Cox PH model with only $X$ and $T$, leaving out $Z$, as found following the first approach generating HR=1.66. On the other hand, the overlapping adjusted survival curve is validated by calculated HR, following both the existing approach with the Cox PH model (HR=0.8) and our algorithm (HR=1.0).

### 3.5.2 Ewing's Sarcoma Data

We also applied the proposed method to a real-world dataset of patients with Ewing's Sarcoma [75]. The dataset was selected based on its survival data and known causal DAG consisting of confounders. The dataset consists of a total of 76 Ewing's sarcoma patients with disease-free survival days as the outcome. 47 of the

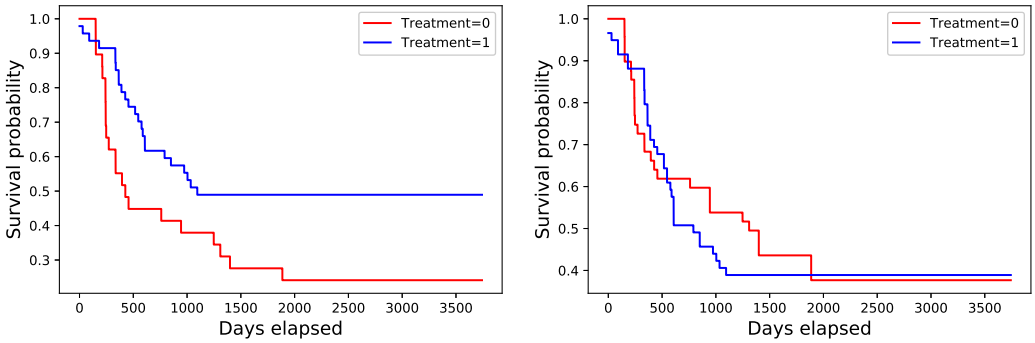



Figure 3.6: Unadjusted survival curve for Ewing dataset *(left)* and, survival curve generated after applying proposed approach *(right)*

patients received a novel treatment (S4), and 29 received (one of) three (S1—S3) standard treatments.

The level of Serum lactic acid dehydrogenase (LDH) acted as the confounder, since high LDH levels indicated a lesser likelihood of treatment assignment along with an impact on survival time. In our analysis, we marked patients receiving S4 as the treatment group ($X = 1$) and patients receiving S1-S3 as the control group ($X = 0$). We applied our algorithm on this data set and the survival curve with the existing approach. Results of our algorithm is shown in Figure 3.6. Figure on left presents treatment group ($X = 1$) to be less hazardous than control group ($X = 0$). Figure on right is the adjusted survival curve with mostly overlapping survival curves of two groups, although treatment group ($X = 1$) seems slightly more prone to hazards. The adjusted survival curve shows similar results, as demonstrated in Makuch et al. [75]. We also present the calculated HR following the three processes described in the earlier subsection. In Table 3.2, the HR calculated by our approach

| | Existing model | | Proposed model |
|---|---|---|---|
| | Observational data excluding confounding variable (biased estimate) | Observational data including confounding variable (traditional approach) | Transformed and adjusted data |
| Hazard Ratio | **0.53** | **1.12** | **1.04** |
| *(95% Confidence Interval)* | ***(0.30-0.96)*** | ***(0.59-2.11)*** | ***(0.57-1.87)*** |

Table 3.2: Hazard Ratio for Ewing dataset, calculated using existing model and our proposed approach

(HR=1.04) differs from the HR calculated in the traditional way (HR=1.12), presenting the drug to be a little less hazardous. However, the 95% confidence interval for both of these coincide, signifying that the true value lies within this range. The first column reports a biased estimate of the HR based on only the treatment and outcome (excluding confounder) in Cox PH model. The second column reports a standard estimate of the HR that includes all known variables (including the confounder(s)). The third column reports the HR calculated in our proposed approach, using only the treatment and outcome (excluding the confounder). The reason for getting an accurate estimate of the HR even when excluding the confounder is because we adjusted the dataset beforehand using a minimum set of confounders from the SCM, thus focusing on the true causal effect of treatment on outcome.

## 3.6 Discussion and Conclusion

We propose a novel method to estimate the HR using the Cox PH Model through the transformation of observational data to corresponding experimental data leveraging an underlying SCM. The transformed data are mathematically

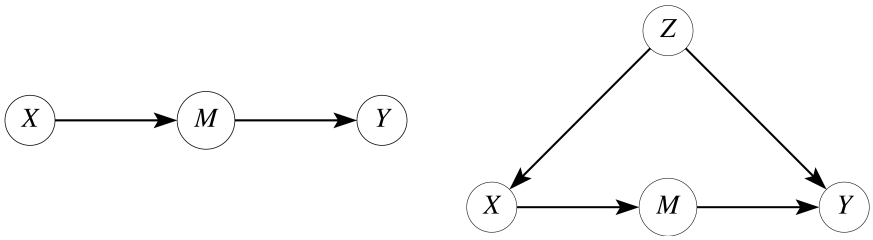



Figure 3.7: Two example graphs where the backdoor adjustment will produce different results compared to an approach based on the ignorability assumption

guaranteed to be adjusted for the confounding bias with the assumption that the SCM represents the data generating mechanism. Previous approaches under RCM that estimate the survival curve use the ignorability assumption, and will not work when the variables selected do not satisfy the backdoor criterion. Ignorability assumption states that, distribution of the potential outcomes $(Y(0), Y(1))$ is independent of the treatment variable by randomly assigning treatment: $\{Y(0), Y(1)\} \perp\!\!\!\perp X$. An extension of the idea, conditional ignorability states, distribution of the potential outcomes $(Y(0), Y(1))$ is independent of the treatment variable $(X)$, conditional on the covariates $(Z)$: $\{Y(0), Y(1)\} \perp\!\!\!\perp X|Z$. Using conditional ignorability for adjustment on covariates allowed researchers to draw inferences from observational studies as well; however, adjusting all covariates irrespective of their causal relationship with treatment and outcome can contribute more bias to the model and incorrect estimation of causal effects.

We present two scenarios as example in Figure 3.7 *(in the left hand side, $X$ is the treatment, $Y$ is the mediator, and $M$ is a mediator. For the second graph, $Z$*

*acts as a confounder as well. The left hand side is the example of a mediator and the right hand side graph is known as the front-door setting.).* In the first scenario (Figure 3.7, left), we show an SCM with treatment $X$ and outcome $Y$ with a third covariate $M$. Here $M$ acts as a mediator in between $X$ and $Y$, thus the backdoor adjustment gives a null set, meaning no adjustment is needed. The do-calculus formula would be: $P(Y|do(X)) = P(Y|X)$. Adjusting on $M$ based on conditional ignorability will produce an incorrect estimation of causal effect. In the second scenario (Figure 3.7, right), we discuss a setting called front-door adjustment where we identify the variables to be adjusted with two applications of backdoor adjustment [89]. In an SCM with a mediator (shown in Figure 3.7 (right)), the covariate $M$ does not satisfy the backdoor criterion and acts as a mediator between treatment $X$ and outcome $T$. Thus, adjusting with $M$ irrespective of its role as mediator will produce a biased estimate of the HR. The accurate backdoor adjustment formula (with $M$ as mediator) is $P(Y|do(X)) = \sum_M P(M|X) \sum_X P(Y|X, M)P(X)$. However, adjustment by assuming $M$ (and $Z$) as confounder gives an incorrect adjustment formula: $P(Y|do(X)) = \sum_{M,Z} P(Y|M, X, Z)P(Z)P(M)$. Our approach (with backdoor criterion) can correctly identify the variables to be adjusted for estimating HR using SCM and do-calculus.

Both the survival curve and the HR help to build a strong interpretation of

the survival analysis of an experiment. The HR is most frequently reported as it summarizes the overall effect of treatment. However, survival curve encodes information on changes in survival over time [45], which, in certain cases gives us better insight. Our proposed method is capable of generating both the survival curve and the HR, along with proper backdoor adjustment based on the underlying SCM. The HR calculated from the adjusted dataset requires only the treatment and outcome variables, and thus relies on direct causal relationships of treatment and outcome. For this purpose, we assume knowledge of the true causal model, an absence of unmeasured confounders, functional relationship in the SCM being time-invariant, and, proportionality of the HR in the outcome. In reality, defining the causal graph with SCM, that is, causal structure learning, requires a principled approach. The development of statistical and computational algorithms for causal structure learning is an active research area [44, 104], and, is not well-established in the current literature. We are currently working on a methodological framework to develop the causal graph with structure learning algorithm and domain expertise.

# CHAPTER 4
# Pragmatic Clinical Trials in the Rubric of Structural Causal Models

## 4.1 Introduction

Experimental studies with varying designs and research goals, such as Randomized controlled trials (RCT), Pragmatic clinical trials (PCT), are frequently conducted in many branches of science to derive the causal effect of interventions. Conversely, observational studies (OBS) capture the outcome of an incident without any alteration of the independent variable. Due to differences in the experiment settings (e.g., goal, population group, treatment protocol), the causal findings of the experiments are harder to compare, and the transfer of knowledge from one study population to another is not very trivial. Thus, there is a need for generalizability or structural methodology to draw unbiased causal inferences from experiments (RCT+PCT+OBS), leveraging their unique design attributes.

In recent times, through the advancement of machine learning and artificial intelligence, finding newer ways of causal explorations from datasets available, i.e., data-driven causal inference, is of high interest. Structural Theory of Causation (SCM), proposed by Judea Pearl [95] and extended by many other researchers [14], holds the potential to define scientific studies for causal inference, express them through graphical causal models and transfer knowledge in between them.

SCMs allow researchers to represent scientific studies systematically. SCM

representation of an experimental study helps portray underlying causal mechanisms, express causal effects of interventions and answer hypothetical questions. However, core differences of PCT from RCT and OBS [32], in terms of (a) population, (b) setting, (c) comparison arm (treatment), and (d) outcome, makes it challenging for objective evaluation of interventions and their effect. This paper illustrates a causal representation of PCT and relevant mathematical formulations to aid causal effect estimations and objective evaluations in a target population and interpret existing analysis techniques through a causal lens.

Contrary to RCT and OBS frequently being formulated through SCM [95], representation of PCT with SCM is still an ongoing research problem [50]. Additionally, novel ways of utilizing priors (background knowledge) to build a comprehensive causal model from data is also under exploration [86]. In summary, a standardized way to represent PCT through SCM is not yet fully grounded on the theories of recent advances in causal inference.

## 4.2 Background

This section describes the relevant background concepts, such as various scientific studies, including pragmatic clinical trials and their unique attributes. We then discuss the structural theory of causal within causal inference and structural

causal models for various scientific studies. Finally, we present our problem formulation, followed by related works.

### 4.2.1 Experimental Studies

In broader terms, based on design factors, scientific studies follow two routines: experimental studies and observational studies. Experimental studies are at the core of most scientific investigations. In experimental studies, experimenters introduce a dependent variable (e.g., treatment or procedure) and consecutively observe an outcome [19]. Most commonly, the underlying research question is uncovering the effect of an outcome compared to an intervention or factor.

The design of experimental studies is a well-explored research area [28, 27]. The most popular and effective experimental study, especially to find causal relations, is the randomized controlled trial (RCT) [39]. In RCTs, researchers explore the effects of treatment on outcome in a narrower population (with clear and strict inclusion-exclusion criteria) with randomization (to control for both known and unknown confounding) [32]. RCTs are harder to implement and cost more; however, they unquestionably justify the causal effect of treatment by comparing treatment arms.

Since experimental studies require significant resources (in time and expenses) and are sometimes unethical or infeasible for a certain population,

researchers occasionally conduct studies through exploring existing real-world data (e.g., EHR dataset) collected without any intervention. These studies are called observational studies (OBS) (or, natural experiments) [100]. RCT and OBS are inherently different from each other, the two prime differences being (a) presence of intervention, (b) de-confounding through randomization. In general, RCTs are considered as a higher level of evidence compared to OBS [21].

One other type of experimental study is pragmatic clinical trials (PCT). By nature, PCTs are more fluid and have characteristics floating between an RCT and an OBS.

### 4.2.2 Pragmatic Clinical Trials

#### Definition

Pragmatic clinical trials (PCT) are a variety of experimental studies that aim to explore correlations between treatment and outcome in a real-world health system, contrary to focusing on causal explorations [81]. Uncovering causal effects through experimental studies requires extreme deconfounding and strict inclusion-exclusion criteria, sometimes making the study result irrelevant to real-world practice. The goal is to define clinical decision-making rather than regulatory approval. Two significant challenges of PCT are: (1) missing data and (2) non-adherence to protocol.

### Features

Due to its pragmatic nature, features of PCT have drawn much discussion from the scientific community. [32] defines PCT as a variation of RCT, with four critical *pragmatic* design elements: (a) real-world population (recruitment extended to fit all potentially eligible individuals receiving care in participating setting), (b) real-world setting (commonly takes place in a flexible setting closer to patients' usual clinical care, avoiding the need for specially trained research staff for data collection), (c) appropriate comparison arm (sometimes combining multiple drugs or multiple doses of the same drug), and (d) relevant outcome (goal is to understand the real-world implications of the intervention). [74, 98] have laid out nine features of PCT, depicted as a wheel in Figure 4.1 (lower score signifies explanatory and higher axis signifies pragmatism in nature).

### Analysis Methods

Since the treatment population group varies based on adherence and loss-to-follow-up in PCTs, various analysis protocol is followed in the data investigation of PCTs. The three most common analysis protocols for PCT are (1) Intention-to-treat (ITT), (2) As-treated (AT), and (3) Per-protocol (PP).

In Intention-to-treat (ITT) analysis, all randomized patients are included, regardless of whether they adhered to the treatment prescribed or subsequent

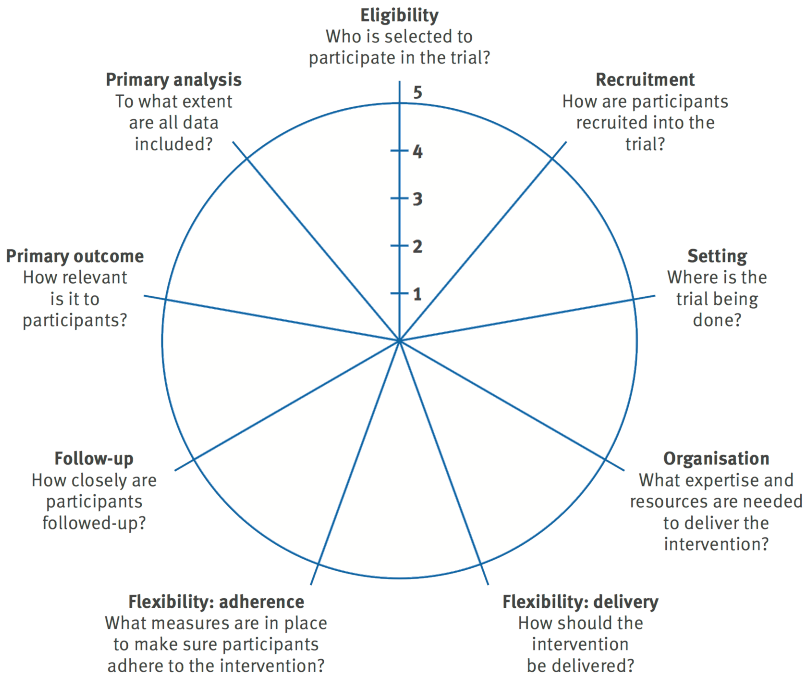



Figure 4.1: Visualization of PRECIS (PRagmatic Explanatory Continuum Indicator Summary)

withdrawal [40, 79]. It essentially ignores anything after randomization (e.g., withdrawal, protocol non-compliance), and in general, avoids overoptimistic estimates of the intervention efficacy. For this reason, ITT is the most recommended method in PCTs [40, 17].

In As-treated (AT) analysis, patients are incorporated based on the treatment they received, irrespective of their randomization status [117]. Likewise, in per-protocol (PP) analysis, only those patients are included who genuinely adhered to the study prescribed, i.e., for whom the treatment prescribed and treatment received are same [112]. PP analysis represents a 'best-case' scenario in trial results since it represents patients who completed the treatment initially

allocated and thus ignores protocol deviation or non-adherence. Both AT and PP analyses give a biased estimate of intervention efficacy; however, they are essential for the report since they reflect the impact of non-compliance and non-adherence.

### 4.2.3 Structural Causal Models

The structural theory of causation was proposed and established on the foundations of probabilistic graphical models by Judea Pearl [95] and many other researchers [14, 119]. Under this theory, structural causal models (SCM) are a structured definition of a causal model, often portrayed through graphs. We present the formal description of an SCM [89] in Definition 5:

**Definition 5 (Structural Causal Model)** *A structural causal model $M$ is a 4-tuple*

$\langle U, V, f, P(u) \rangle$ *where:*

1. *$U$ is a set of background (exogenous) variables that are determined by factors outside of the model,*
2. *$V$ is a set $\{V_1, V_2, ..., V_n\}$ of observable (endogenous) variables that are determined by variables in the model (i.e., determined by variables in $U \cup V$ ),*
3. *$F$ is a set of functions $\{f_1, f_2, ..., f_n\}$ such that each $f_i$ is a mapping from the respective domains of $U_i \cup PA_i$ to $V_i$, where $U_i \subseteq U$ and $PA_i \subseteq V$ $V_i$ and the*

*entire set $F$ forms a mapping from $U$ to $V$. In other words, each $f_i$ in $v_i \leftarrow f_i(pa_i, u_i), i = 1, ..., n$, assigns a value to $V_i$ that depends on the values of the select set of variables $(U_i \cup PA_i)$, and*

4. *$P(u)$ is a probability distribution over the exogenous variables.*

Causal directed acyclic graphs (DAG) are commonly portrayed to express an SCM. In a causal DAG $G$, node $V$ represents an observed or unobserved variable, and directed edge $E$ represents the causal relationships between two nodes. With the purpose of investigating the causal effect of one variable on another, do-calculus was developed [95]. Do-calculus is a multi-functional tool (mathematical formulation) to map the observational truth to the corresponding experimental reality by adjusting for different kinds of biases, such as confounding (if it exists).

### 4.2.4 SCM for Scientific Studies

SCM and causal DAG have been frequently used in the literature to represent various scientific studies [125]. Figure 4.2 shows two graphical structures of SCM, one for observational study (left) and the other for randomized controlled trial (right). Both of them have treatment $X$, outcome $Y$, and confounder $Z$; the only distinction being a lack of arrow (causal connection) from $Z$ to $X$, thus representing the randomization done prior to the study. Representation through SCM helps

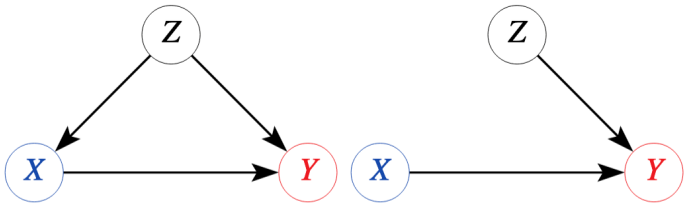



Figure 4.2: SCM representation of scientific studies

provide a structural definition to distinct trials and allows application of do-calculus for causal effect estimation and counterfactual evaluation.

### 4.2.5 Problem Definition

Since the strength of SCM in representing different studies and exploring the underlying causal mechanisms is well-established in the literature, researchers are looking for ways to represent PCT using the rubric of SCM. For this work, we focus on the two following research questions:

1. how can we represent PCT through SCM?
2. how can we represent the analysis techniques commonly deployed in PCT using SCM and do-calculus?

### 4.2.6 Related Works

For causal exploration on PCT, different general guidelines have been proposed in the literature; however, a unified approach is severely lacking. [50] discussed issues involving pragmatic trials in general, along with a general causal graphical structure

and adherence as a node in the graph. Without using any underlying causal structure, [81] have presented an elaborated guideline for a causal understanding of diverse, unique features of PCT qualitatively and figuratively to estimate the ITT and PP effects (of both point and sustained intervention). Later, in continuation to the previous two works, [80] discussed a wide variety of graphical representations of PCTs, but without employing any do-calculus for ITT or PP effect estimations. Although all the works used causal graphical structures for representing PCTs, they did not decide on a single definition or discuss its use with do-calculus for ITT, AT, or PP analysis.

## 4.3 Structural Causal Model for Pragmatic Clinical Trials

In this section, we introduce the notion of structural causal models (SCM) for pragmatic clinical trials (PCT). We iterate through the unique features of PCT, such as eligibility criteria, non-adherence, and loss-to-follow-up, and examine their potential interpretations in structural causal models. Following that, using the notations proposed, we discuss the frequently used analysis methods for PCT, such as intention-to-treat, as-treated, and per-protocol analysis. For simplicity, we assume a point intervention with no time-varying components (both in treatments or outcomes).

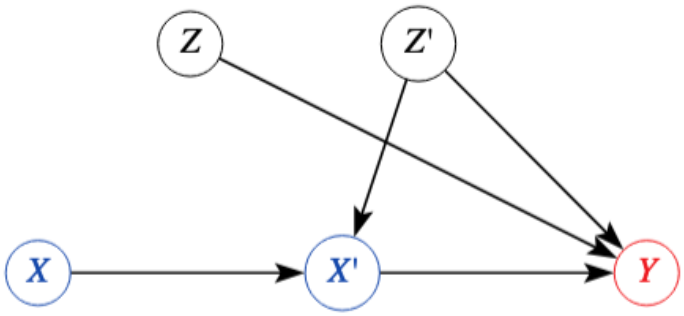



Figure 4.3: Graphical representation of the proposed structural causal model for pragmatic clinical trials

### 4.3.1 Defining PCT for SCM

To express a PCT through SCM, we start with defining the PCT. We assume, we are working with a $PCT$ with population group $\Pi$, where the query of interest is finding the effect of a treatment protocol $X$ (*not the same as 'causal' effect of treatment $X$, explained in section 4.3.1*) on outcome $Y$. Different arms of treatment protocol $X$ might have overlapping components, such as the same drug (or software feature) with a different dosage (or color palette). We propose that the target $PCT$ for the population $\Pi$ can be expressed through a structural causal model $M =< U, V, F, P_u >$, with a graphical representation through graph $G$, with two versions of treatment $X$ and $X'$.

#### Equivalent RCT

For reference and comparison, if the query of interest for the researchers were, in fact, finding the 'causal' effect of treatment $X$ on outcome $Y$, the standard

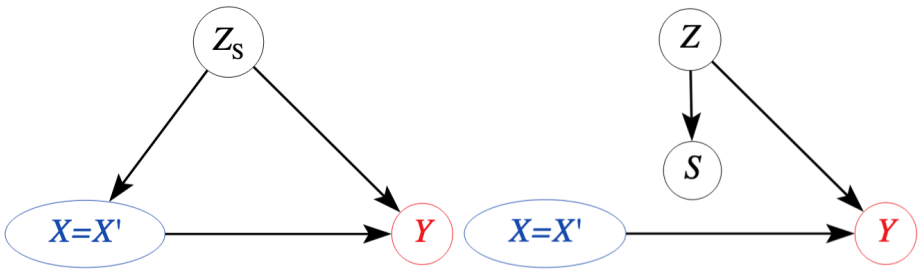



Figure 4.4: Graphical representation of the structural causal model of a $RCT_{PCT}$, *(left)* with population $\Pi_s$ & treatment $X$ $(= X')$, and *(right)* with population $\Pi_s$ as a selection bias through node $S$ on population $Z$

procedure would be to conduct a randomized controlled trial (RCT) on a stricter population group $\Pi_s$. In that case, the causal graph for the RCT would be similar to Figure 4.4 (left), where treatment would be $X(= X')$. The reasoning behind having different population groups for PCT ($\Pi$) and RCT ($\Pi_s$) comes from their definitions; $\Pi_s$ would be a narrower focused group of $\Pi$ with minimal possible confounding to outcomes. We will refer to this equivalent RCT as $RCT_{PCT}$.

### 4.3.2 Features of PCT

#### Treatment, Outcome and Covariates

A general graphical representation $G$ of the proposed SCM $M$ for PCT is presented in Figure 4.3. Here, the independent variable, or treatment, is represented by $X$ *(and $X'$, explained in next subsection)*, and $Y$ represents the dependent variable or outcome. Covariates $Z$ and $Z'$ represents all other relevant variables; however, $Z$ do

not have any causal effect in adherence to the trial (i.e., $X'$), whereas $Z'$ are the covariates that affect adherence (e.g., affects treatment received $X'$).

Since $X$ is provided through randomization, there is no causal relationship ($\rightarrow$) between $Z$ or $Z'$ and $X$. However, as $Z'$ are indicators of adherence, there is a causal relationship between $Z'$ and $X'$.

**Non-adherence**

Since non-adherence to treatment is a core component in $PCT$, they are depicted through two separate nodes $X$ and $X'$ in the proposed causal graph. $X$ represents the treatment prescribed (through randomization), and the treatment received (or followed by trial participants) is represented by $X'$. $X'$ is different from $X$ due to non-adherence; however, it is still influenced by $X$. The relationship between $X$ and $X'$ has previously been expressed [80] through adherence to the trial, as a percentage of adherence to the treatment prescribed.

**Eligibility criteria**

Compared to RCTs, PCTs are more liberal in including patients from varying demographics. As previously discussed in section 4.2, eligibility criteria are the key reason behind this population demographic difference between a PCT and a similar RCT, and thus, between $\Pi$ and $\Pi_s$. This difference can also be viewed as a selection [12] through node $S$, where $S = 1$ defines being eligible for the RCT, equivalent to

the target PCT. However, selection through node $S$ in a study does not always trigger selection bias to that study.

**Loss-to-follow-up**

In most PCTs, population lost-to-follow-up is a concern for the scientists [32]. Since trial participants tend to show lesser adherence to the protocol, some generally do not follow through with the treatment prescribed or disconnect with the research team and end up being the population lost-to-follow-up. During data analysis, this population data lost-to-follow-up are generally censored [50]. The conditioning of censored data can also be viewed as survivorship bias [18], through a node $C$. For $C = 1$, we select a population group who completed the trial and were not lost-to-follow-up, thus looking at a population who 'survived' the study.

### 4.3.3 Outcome Analysis for PCT

**Query of Interest**

By definition, the query of interest in a PCT is finding the 'effectiveness' of a treatment protocol, not the 'efficacy' of specific treatment [97]. Based on that, [80] have described the vital causal interests in a $PCT$: intention-to-treat effect, the per-protocol effect of continuous adherence to treatment versus placebo, and in general, the effect of good adherence to trial protocol versus poor adherence in the placebo arm.

**Intention-to-treat Analysis**

In the intention-to-treat analysis, we explore the effects on outcomes based on randomized or prescribed treatment. Since all the participants (in some cases, excluding loss-to-follow-up) are included in this analysis, we express the concern by:

$$P(Y|X) \tag{4.1}$$

**As-treated Analysis**

For as-treated analysis, by definition, we look for participants who indeed took the treatment rather than prescribed, so the "true" treatment intervention would be $X'$, not $X$. For that, we express the concern by:

$$P(Y|X') \tag{4.2}$$

**Per-protocol Analysis**

Finally, we include the population who followed through treatment prescribed for per-protocol analysis. We exclude the population who have taken a different treatment than what was prescribed; that is, for whom $X$ and $X'$ did not match.

With this, we express the concern by:

$$P(Y|X = a, X' = a) \tag{4.3}$$

**Additional Study Metrics**

Pragmatic clinical trials additionally report other relevant study metrics, such as Odds Ratio (OR), Risk Ratio (RR), and Hazard Ratio (HR) [23]. These metrics are used to detect the association of treatment with the outcome and provide additional insight into treatment effects. We present equations to calculate their values based on conditional probability below. In Equation 4.6, $h(t, \mathbf{X}_{x=a})$ represents hazard function with time $t$ and the vector with the covariates of the model $X$ with the value $a$ ($X = [z, x]$, where $x$ is the treatment and $z$ are the confounders).

$$OR = \frac{\frac{P(Y=0|X=1)}{P(Y=1|X=1)}}{\frac{P(Y=0|X=0)}{P(Y=1|X=0)}} \tag{4.4}$$

$$RR = \frac{\frac{P(Y=0|X=1)}{P(Y=0|X=1)+P(Y=1|X=1)}}{\frac{P(Y=0|X=0)}{P(Y=0|X=0))+P(Y=1|X=0)}} \tag{4.5}$$

$$HR = \frac{h(t, \mathbf{X}_{x=1})}{h(t, \mathbf{X}_{x=0})} \tag{4.6}$$

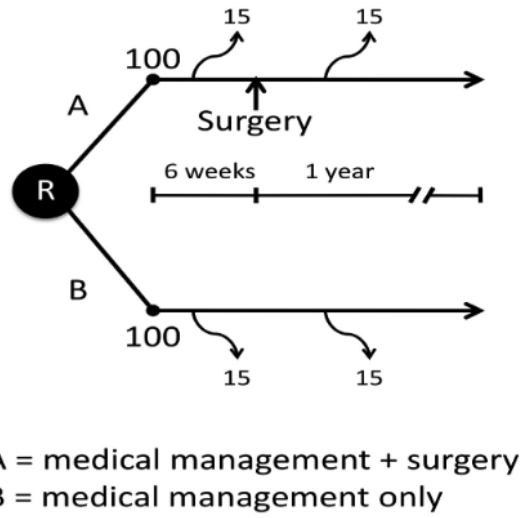



Figure 4.5: Hypothetical PCT in patients with cardiovascular disease. Intervention, A = medical management + surgery, vs. control, B = medical management only. Collected from McCoy et al. [76]

Given a known structural causal model, interpreting their causal equivalent is established in the literature: causal odds ratio & causal risk ratio [88] and causally formulated hazard ratio [3].

## 4.4 Example of PCT with SCM

In this section, we apply definitions and assumptions from section 4.3 to represent a hypothetical PCT through SCM, and leverage Equation 4.1, Equation 4.2, Equation 4.3, Equation 4.4, Equation 4.5, and Equation 4.6 on the dataset to find relevant treatment effects and outcome metrics.

For this purpose, we leverage a hypothetical pragmatic clinical trial, discussed in [76] and presented in Figure 4.5. In this PCT, an investigator conducted a study to evaluate whether the addition of surgery to a conventional

medical therapy would benefit the patients (e.g., effective in controlling death in patients with cardiovascular disease). A total of two hundred (200) patients were enrolled, and half of them were allocated the new treatment protocol. The intervention treatment group received a combination of medical management and surgery, whereas the control group received only medical management.

**Assumption of Ground Truth** With the usage of this dataset, we are also assuming the 'ground truth' that the surgical intervention does not affect outcomes. Researchers are searching for this 'ground truth'; one of the ways to do that is to conduct this hypothetical PCT.

**Study Timeline Overview** As shown in Figure 4.5, after randomization, both arms of intervention contained a total of 100 patients. The medical management continues from randomization, and there is a timeline gap or waiting period of six (6) weeks from randomization to surgery. In treatment group A, 15 patients died before the six-week waiting period, and an additional 15 died between six weeks and 12 months. Similarly, 15 patients died before six weeks in treatment group B, and another 15 died between six weeks and 12 months.

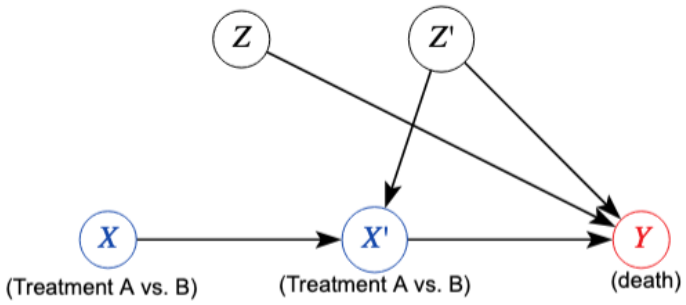



Figure 4.6: Graphical overview of SCM representation of the example PCT

### 4.4.1 SCM for PCT

Using definitions from section 4.3, we represent the PCT through a SCM, as represented graphically in Figure 4.6. $X$ is the treatment prescribed after randomization, where the population was divided equally between two treatment protocols. $X'$ is the treatment received, different from $X$ due to patients (count of 15) not going through surgery within six weeks. $Y$ is the outcome, death in a year for this trial. Although our graph shows $Z$ and $Z'$, we do not have any data on record on these two for this specific PCT.

### 4.4.2 Outcome Analysis

We reorganize the trial dataset to count patient outcomes for each value of $X$, $X'$, and $Y$.

Application of equations from section 4.3 are presented in Table 4.2. The calculated results match with the results reported in [76] and show that the equations discussed hold their originality, with the addition of SCM for a better

| $X$ | $X'$ | $Y$ | $Count$ |
|---|---|---|---|
| Treatment A | Treatment A | No death | 70 |
| Treatment A | Treatment A | Death | 15 |
| Treatment A | Treatment B | No death | 0 |
| Treatment A | Treatment B | Death | 15 |
| Treatment B | Treatment A | No death | 0 |
| Treatment B | Treatment A | Death | 0 |
| Treatment B | Treatment B | No death | 70 |
| Treatment B | Treatment B | Death | 30 |

Table 4.1: Population distribution for different values of treatment prescribed $X$, treatment received $X'$ and outcome $Y$

| | **ITT** | **AT** | **PP** |
|---|---|---|---|
| **RR** | $\frac{P(Y=0\|X=1)}{P(Y=0\|X=1)+P(Y=1\|X=1)}$ / $\frac{P(Y=0\|X=0)}{P(Y=0\|X=0)+P(Y=1\|X=0)}$ $= \frac{P(Y=0\|X=1)}{P(Y=0\|X=0)}$ $= \frac{\frac{15+15}{70+15+0+15}}{\frac{0+30}{0+0+70+30}}$ $= \frac{0.3}{0.3}$ $= 1.00$ | $\frac{P(Y=0\|X'=1)}{P(Y=0\|X'=1)+P(Y=1\|X'=1)}$ / $\frac{P(Y=0\|X'=0)}{P(Y=0\|X'=0)+P(Y=1\|X'=0)}$ $= \frac{P(Y=0\|X'=1)}{P(Y=0\|X'=0)}$ $= \frac{\frac{15+0}{70+15+0+0}}{\frac{15+30}{0+15+70+30}}$ $= \frac{0.18}{0.39}$ $= 0.46$ | $\frac{P(Y=0\|X=1,X'=1)}{P(Y=0\|X=1,X'=1)+P(Y=1\|X=1,X'=1)}$ / $\frac{P(Y=0\|X=0,X'=0)}{P(Y=0\|X=0,X'=0)+P(Y=1\|X=0,X'=0)}$ $= \frac{P(Y=0\|X=1,X'=1)}{P(Y=0\|X=0,X'=0)}$ $= \frac{\frac{15}{70+15}}{\frac{30}{70+30}}$ $= \frac{0.18}{0.3}$ $= 0.60$ |
| **OR** | $\frac{\frac{P(Y=0\|X=1)}{P(Y=1\|X=1)}}{\frac{P(Y=0\|X=0)}{P(Y=1\|X=0)}}$ $= \frac{\frac{0.3}{1-0.3}}{\frac{0.3}{1-0.3}}$ $= 1.00$ | $\frac{\frac{P(Y=0\|X'=1)}{P(Y=1\|X'=1)}}{\frac{P(Y=0\|X'=0)}{P(Y=1\|X'=0)}}$ $= \frac{\frac{0.18}{1-0.18}}{\frac{0.39}{1-0.39}}$ $= 0.34$ | $\frac{\frac{P(Y=0\|X=1,X'=1)}{P(Y=1\|X=1,X'=1)}}{\frac{P(Y=0\|X=0,X'=0)}{P(Y=1\|X=0,X'=0)}}$ $= \frac{\frac{0.18}{1-0.18}}{\frac{0.3}{1-0.3}}$ $= 0.51$ |

Table 4.2: Outcome metrics for the PCT

understanding of the trial. Similar results can also be estimated through the equations provided from datasets used in other similar studies for PCT [79, 59].

## 4.5 Discussion and Conclusion

In this work, we have discussed the notion of leveraging structural causal models within causal inference to represent pragmatic clinical trials. Our proposition, along with relevant data analysis on the simulated PCT dataset, shows a prospective path

of exploring PCTs for treatment effect estimations, counterfactual analyses, and transportability methods explorations.

**Strengths** The essential contribution of this proposition is the notion of leveraging SCM for expressing PCTs. SCM and relevant causal inference methodologies have already been highly beneficial in estimating causal effects for different experimental and observational studies [124]. PCTs are highly meaningful for decision-makers as they are easier to conduct and convey treatment efficacy in a standard-setting. Since PCTs are more fluid in their nature than other experimental studies, the need to draw causal estimations from PCT is also higher than others. The uniqueness of this proposition is defined by the usage of $X$ and $X'$ representing treatments as two causally connected yet different variables.

**Causal Equivalent of Guidelines for PCT** The four key design elements of PCT, by definition [32], are real-world population, real-world setting, appropriate comparison arm, and relevant outcome. Excluding only real-world settings, the SCM definition for PCT can utilize all the other elements. The concept also reflects and pairs perfectly with the guidelines provided by [81].

**Causal Interpretation of Analysis Equations** Given an OBS with $X$ as treatment, $Y$ as an outcome, and $Z$ as confounders, we easily find the conditional

probability of outcome $Y$ given $X$ $[P(Y|X)]$. To find the equivalent causal effect, we either conduct a similar RCT with treatment randomized (aka de-confounded) and look at $P(Y|X)$ or simulate the RCT from the OBS using do-calculus $(P(Y|do(X)))$. Resembling to that conversion of $P(Y|X)$ to $P(Y|do(X))$, we explore causal effects from the equations Equation 4.1, Equation 4.2, and Equation 4.3 by applying do-calculus on these. It results in:

$$P(Y|do(X)) = P(Y|X) \tag{4.7}$$

$$P(Y|do(X')) = \sum_{Z'} P(Y|X', Z')P(Z') \tag{4.8}$$

$$P(Y|do(X = a), do(X' = a)) = \sum_{Z'} P(Y|X = a, X' = a, Z')P(Z') \tag{4.9}$$

Equation 4.7, Equation 4.8, and Equation 4.9 provides two interesting insights to the notion proposed.

**(1)** Since $X$ is randomized, Equation 4.8 is equal to its equivalent conditional probability equation. This estimation is the most standard (unbiased) estimation in providing treatment effect, which also aligns with [79]. Nevertheless, it still cannot

minimize bias introduced by loss to follow-up, as $X'$ is not considered in this equation. Equation 4.8 does not use $X$ but uses $X'$, and is deconfounded by using $Z'$. The effect estimation is helpful, but the causal estimation requires a knowledge of measured confounders $Z'$, which is hard to find in the real world. This equation is also valuable since it shows the effect of non-adherence on the trial participants (through $X'$). Equation 4.9 uses both $X$ and $X'$ in estimating the effect, by which it captures the essence of the population who strictly adhered to the protocol.

**(2)** Although $X$ and $X'$ represent treatment in different population percentages, they still fundamentally represent the same treatment for the study. While conducting a real-life PCT, with patients lost to follow-up, the ITT analysis results do not match with AT analysis results. Under normal conditions, $P(Y|X)$ and $P(Y|X')$ would never be equal. However, with do-calculus, it is expected that $P(Y|do(X))$ and $P(Y|do(X'))$ would be the same since they both indicate the causal effect of treatment on outcome. It raises the idea that, if we can identify a true set of confounders $Z'$ (that affects adherence), we can estimate the true causal effect of treatment on outcome from a PCT, and in those cases, Equation 4.7, Equation 4.8, and Equation 4.9 will all produce the same effect estimate.

**Limitations** The prime challenge is defining the relevant causal structure for the SCMs representing the PCT. RCTs (and Obs) are frequently expressed through

SCMs; however, that does not happen with PCT due to their pragmatic nature by definition. Researchers continuously explore ways to build causal structure through data and priors (background knowledge, peer-reviewed literature). Another critical challenge in this research is to find an appropriate set of confounders $Z'$. Confounding variables, in most cases, are not observed, measured, or even found. Finally, in PCTs, the treatment prescribed generally differs from the treatment received. Thus, adherence to the trial is vital, and causal effect estimation becomes complex when the information is unavailable or hard to determine.

**Future Works** Our future work will include instrumental variable analysis [12], by using treatment $X$ in Figure 4.4 as an instrumental variable for the proposed causal graph. We will additionally explore time-series intervention with the definition proposed, in place of point intervention, by altering the SCM and related transportability equations.

# CHAPTER 5
# Causal Discovery on the Effect of Antipsychotic Drugs on Delirium Patients in the ICU using Large EHR Dataset

## 5.1 Background and Problem Statement

With a focus on the theoretical development of causal inference methodologies in the previous three chapters, this chapter aims to discuss a practical, real-world application of the causal inference framework to untangle unknown healthcare information. For this purpose, we look into Delirium patients in the ICU.

Delirium (or acute brain failure) [35] is a disorder or disruption of consciousness, presenting with a reduced capacity to focus, sustain, or shift concentration. Delirium occurs in about 80% cases in the Intensive Care Unit (ICU) and is associated with a more extended hospital stay, increased mortality for each additional day with Delirium in the ICU [96] and other clinical complications such as self-extubation and removal of catheters. Two of the significant issues in diagnosing and treating delirium patients are:

- Currently, no biomarker exists to diagnose Delirium; rather, Delirium is diagnosed with subjective assessment tools such as the confusion assessment method (CAM) [57, 130]. This diagnosis requires the presence of a physician active in the medical center and makes the diagnosis and detection of Delirium patients in the real-world challenging.

- Delirium is commonly treated with antipsychotic drugs (APD) [35] such as Haloperidol, Ziprasidone, Olanzapine, etc. However, multiple randomized controlled trials (RCTs) have shown either conflicting or inconclusive results about the efficacy of APD in the treatment of delirium [82, 42]. This has created a controversy over the efficacy or safety of APD in treating Delirium.

RCTs have been considered the gold standard since the 1960s [38]. The goal was to identify the causation of diseases and understand the causal effect of drugs by the regulatory bodies such as the FDA and clinical communities. The key ideas behind RCT are:

- By random assignment of treatment or interventions, the confounding bias, i.e., the bias due to the assignment of treatment or presence of other variables, can be removed from the estimand, including the unobserved confounders.
- By comparing similar population groups of treatment and control arm, an estimation can be made about treatment efficacy in the target population group.

However, RCTs have their own set of challenges as well. RCTs have become increasingly time-consuming, costly, and are often infeasible for safety and efficacy reasons [31]. Thus there is a need to find alternatives to RCTs, possibly to detect

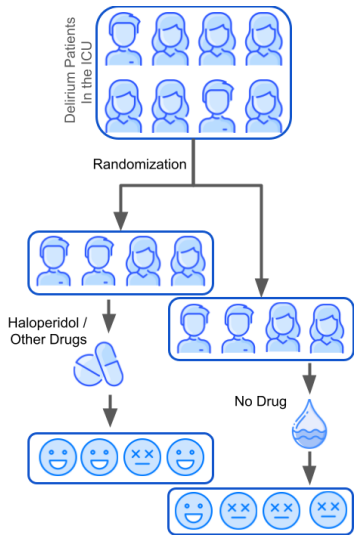



Figure 5.1: RCT for Antipsychotics-based treatment for Delirium

causal effects from other sources of information and aid in removing controversies of treatments in the field.

Recent advances in technology and the adoption of computerized systems in routine healthcare have enabled the collection and curation of large volumes of data during routine healthcare, albeit with confounding biases. At the same time, recent advances in the theory of causal model, more specifically structure causal models (SCM), provides the framework for adjusting for these confounding biases in many cases. This removal of biases can be done (sometimes even if the confounders are unobserved) from observational data using adjustment formulas such as backdoor/front door criterion [14, 95, 120]. However, this approach requires developing a graphical representation of the problem domain with meticulous scrutiny of the variables' relationship, structure learning algorithms, clinical experience, and existing literature.

With the availability (collection, storage, maintenance) of large-scale data in different branches of science (healthcare, finance, sociology) and technology (connected health, smart home), opportunities exist to extract necessary information from it. Big data has aided in the revolution of neural networks, advanced reinforcement learning, and improved statistical machine learning methodologies. Most research advancements integrating big data revolve around curve fitting and correlation. However, without causal relationships, scientists lack the power of intervention or to even explore hypothetical scenarios (counterfactuals).

Big data is responsible for many breakthroughs and advancements in healthcare, contributing to improved treatment policy solutions and collaborated information from multiple sources. Although most breakthroughs are based on predictive models, causal relationships are more crucial for healthcare. This has led to countless experimental trials (randomized controlled trials, case-control studies) on finding the efficacy or impact of an intervention on target outcomes. Causal inference leverages big data and contributes to finding causal information, sometimes even without experiments. One of the strengths of causal inference methodologies is to draw conclusions on causal effects from observational data. Causal inference and its potential with big data are not limited to healthcare only; it has shown great potential in other fields (finance, sociology, law) as well

[53, 85, 113]. Artificial intelligence is iteratively improved with research work and is getting better at decision making and predictive modeling.

Since RCTs cost a lot in terms of money and time, emulation of RCTs from the observational dataset can help reduce them. It also would aid in using datasets from all over the world to find causation in other diseases and health complexities. While RCTs are the gold standard for identifying causal effects of interventions, it is time-consuming and costly. On the other hand, the data collected during routine care, such as electronic health records (EHR), might also be valuable to generate insight, identify the disease pathway and estimate the effect of interventions using recent advances in methods for causal inference.

We aim to study the efficacy of APD in the treatment of Delirium using retrospective cohort analysis. We plan to use the Causal inference framework to look for the underlying causal structure model, leveraging the availability of large observational data on ICU patients. It will help us to untangle the causal relationship between variables and look into the counterfactual world (what-if). To explore safety outcomes associated with APD, our research work targets to develop a causal model for Delirium in the ICU using large observational data sets. We aim to utilize the MIMIC III database, an extensive electronic health records (EHR) dataset with 53,423 distinct hospital admissions [4]. Our null hypothesis is: that there is no significant difference in outcomes for delirium patients under different

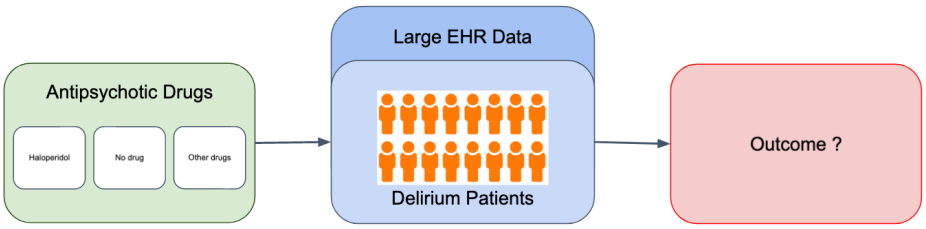



Figure 5.2: Target observational study from large EHR data

drug-group in the ICU. If successful, our research work should help clear the common controversy over prescribing APDs as well as shed light on the underlying causal mechanism triggering Delirium in ICU patients. In other words, we propose the following specific aims.

1. Create and curate three cohorts for patients with Delirium in the ICU from MIMIC EHR data.
2. Develop structural causal models (SCM) with the domain expertise to integrate clinical knowledge and probabilistic information from the data to estimate the causal effect of interventions.
3. Validate the models with statistical methods and independent data sets.

Epidemiologists have continuously involved causal inference tools, such as causal structure learning algorithms, in identifying underlying causal structures. The process is impactful since it generates a causal model based on the information available (data, literature, expertise), leading to a better understanding of the

disease ecosystem; and estimates causal effects based on that. This creates a potential to explore Delirium-treatment-related controversies through observational datasets. Different studies have taken different paths; few studies [128, 2] have used specific SLA algorithms to detect a causal DAG applicable for a targeted research question, whereas others [110, 10] have assumed the causal structure from literature, and validated them using datasets available. We plan to create a similar computational pipeline for Delirium patients in the ICU inspired by these.

## 5.2 Method

To create a data cohort on Delirium patients in the ICU, along with relevant covariates, we seek help from MIMIC-III [58], a publicly available large electronic healthcare dataset. MIMIC-III is curated for twelve (12) years (2001-2012) and holds information on around 53k distinct hospital admissions with around 40k distinct patient histories. The database is well-maintained, de-identified, and open for researchers (with necessary and relevant access protocol) to explore and investigate.

The general process starts with appropriate data mining and data preparation process. We plan to extract information regarding Delirium patients (based on relevant ICD-9 code) and related covariates (decided upon exploring literature). We then move forward with the data analysis protocol, which consists of three (3) types of analysis:

- Exploratory analysis: to explore data distribution and dataset properties
- Machine learning-driven analysis: to infer primary point of interest (i.e., primary outcome) based on all available covariates, as a standard approach to prediction
- Causal analysis:
  - Causal structure generation: to regenerate underlying causal model through various structure learning algorithms
  - Causal effect estimation: to evaluate the 'true' causal effect of treatment on our defined points of interests

## 5.3 Results

This section describes our data curation protocol in detail, along with data exploration and analysis. We present our general findings based on those steps taken.

### 5.3.1 Covariate Selection

We start the process by defining the research questions *(Is Haloperidol better at treating Delirium patients in the ICU, compared to no antipsychotics or other antipsychotics, such as Ziprasidone, Olanzapine, etc.?).* We formulate this question based on controversies present in existing literature *(described in the background*

| sex | age | race | icd9 codes |
|---|---|---|---|
| sofa | apsiii | surgery | pneumonia |
| sepsis | dementia | alzheimers | depression |
| anxiety | met. acidosis | airway obs. | copd |
| liver disease | heart disease | mechvent. | mechvent. count |
| time to mechvent. | **drug group** | drug categories count | drug timelength |
| **death in hospital** | **death timeline** | **length of stay** | **time in mechvent.** |

Table 5.1: Features in MIMIC-Delirium

*section)*. Our null hypothesis is that there is no significant difference in target outcomes for Delirium patients under different antipsychotics treatment groups in the ICU. We define the treatment as the antipsychotics prescribed after being diagnosed with Delirium in the ICU, with three different arms (Haloperidol, no antipsychotics, and other antipsychotics). Our primary outcomes are (1) patient death in hospital and (2) patient death timeline (death in 30 days / 90 days / a year / survived more than a year). Our secondary outcomes are (1) length of stay in the ICU and (2) time put in mechanical ventilation. A total of fifty (50) relevant covariates are explored and marked, which are closely correlated with our points of interests (primary and secondary outcomes) for Delirium patients in the ICU. However, due to the lack of availability of all covariates in the observational dataset, we opt for the most significant twenty-eight (28) covariates, as listed in Table 5.1. Here, the drug group (Haloperidol, no drug, other drugs) is the treatment provided. Primary outcomes are death in hospital & death timeline, and secondary outcomes are the length of stay & time in mechvent.

These all together defined a target trial that we plan to emulate. The target trial is inspired by existing RCTs done on delirium patients to find the effects of antipsychotics and is designed to minimize the effect of confounding variables and (selection) bias. Based on these, we start our data curation process from the MIMIC-III dataset.

### 5.3.2 Data Curation Process

To determine eligible Delirium patients, we look into patients with ICD-9 code 293.0 (Delirium due to conditions classified elsewhere) [55]. We extract relevant information about the patients from admissions, icu_stays, and diagnoses_icd table to form the base dataset. We then infuse it with information from cptevents, d_icd_diagnoses and prescriptions tables, and other views presented in the public repository of the database (sofa, apsiii, ventdurations) [78]. We merge all information together to create our target dataset of 1398 patients. We name this curated dataset as *MIMIC-Delirium* for future references.

### 5.3.3 Data Overview & Exploratory Insights

After our data curation to create MIMIC-Delirium dataset, we successfully extract 1671 ICU stays with 1445 hospital admission counts on 1398 unique patients and their relevant 28 covariate information. In terms of treatment provided in the ICU, we found 681 *(40.75%)* were given Haloperidol, 528 *(31.60%)* were given other

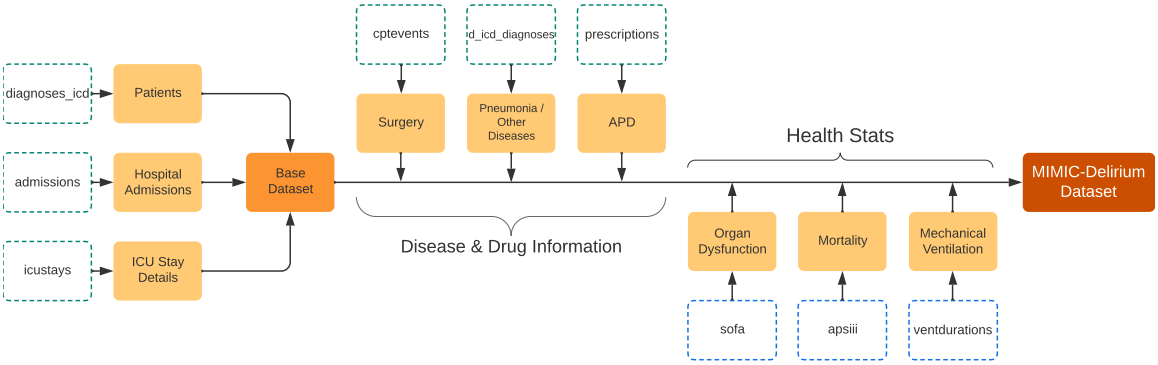



Figure 5.3: Data mining protocol (simplified)

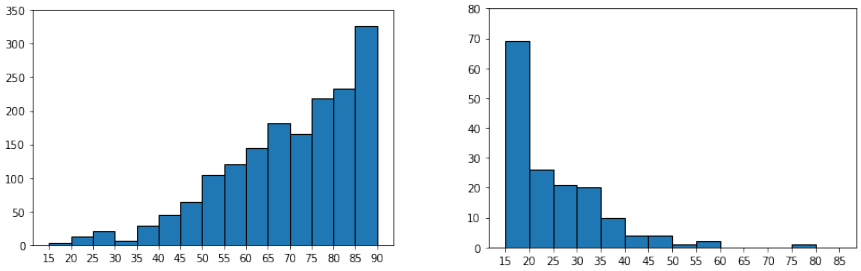



Figure 5.4: Data distribution on age in years *(left)* and length-of-stay in days *(right)*

antipsychotics and 462 *(27.65%)* were given no antipsychotics. In terms of outcome, 311 *(18.61%)* had death in 30 days, 108 *(6.46%)* had death in 90 days, 175 *(10.47%)* had death in a year, and 253 *(15.14%)* survived at least a year (information on 821 *(49.13%)* were unknown). Among the common associated diseases in the ICU, 375 *(22.44%)* had Sepsis, 484 *(28.96%)* had Pneumonia, 1035 *(61.94%)* had (a variation of) heart diseases, and 97 *(5.80%)* had (a variation of) liver diseases. Figure 5.4 shows the general data distribution on age in years *skewed to right since Delirium is frequent in elderly population* and length-of-stay in days *(skewed to left since higher number of ICU stay is severe and rare).*

Additionally, we had findings such as mean length-of-stay and max length-of-stay is higher for patients in the Haloperidol drug group, most patients, who were given multiple APD, were given Haloperidol, The Haloperidol group has a higher death rate in a year than the other two groups, etc.

For the statistical analyses, we conducted a one-way between-subjects ANOVA to compare the effect of the drug group on length of stay in Haloperidol, no drug, and other drugs group. With $p < 0.05$, we found a significant effect of the drug group on the length-of-stay. Post hoc comparisons by the Tukey HSD test indicate that the mean score for the Haloperidol group (mean: 7.47, deviation: 8.55) was significantly higher compared to no drug group (mean: 4.12, deviation: 5.66) and other drugs group (mean: 5.44, deviation: 6.14).

### 5.3.4 Predictive Analysis on MIMIC-Delirium dataset

Before our deep dive into causal exploration, we briefly explored the MIMIC-Delirium dataset for predictive analysis. We employed standard supervised classification algorithms on the complete dataset, with all 24 covariates (discarding the output features) as features and death in hospital as the label. We deployed 10-fold cross-validation with Logistic Regression, Support Vector Machine, and XGBoost algorithm. Mean accuracy with Logistic Regression is 89.71%, mean accuracy with SVM is 89.11%, and test-mlogloss-mean for XGBoost (with 50 rounds

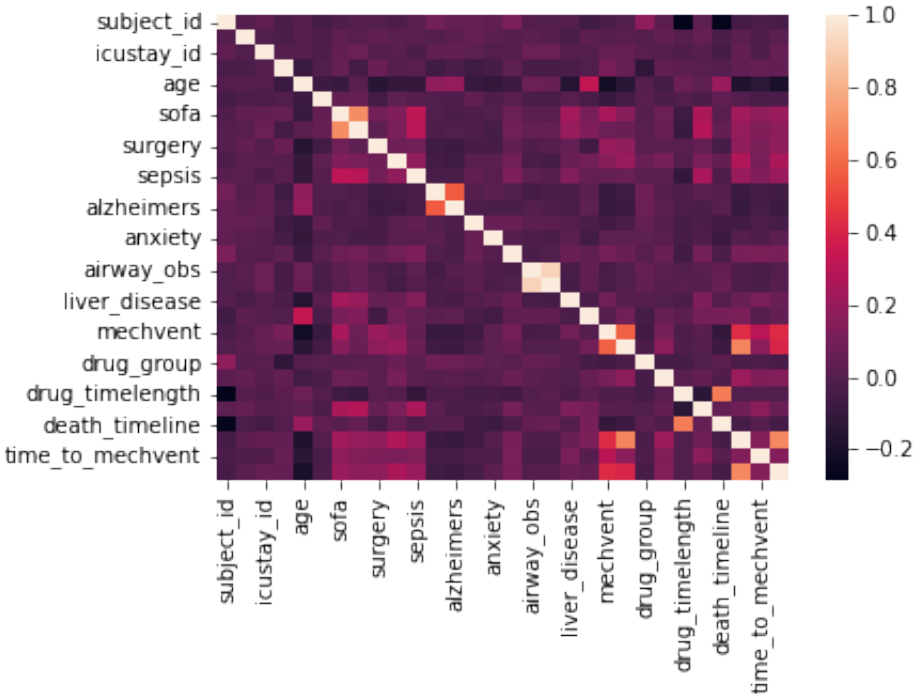



Figure 5.5: Correlation heatmap of MIMIC-Delirium

of boosts) is 0.2724. For XGBoost, we also find that length-of-stay and age have the highest impact in predicting outcome death in this case, which is self-explanatory. Figure 5.5 shows the general correlation between features as a heatmap.

### 5.3.5 Causal Analysis on MIMIC-Delirium dataset

Our causal analysis is built upon two steps: (1) causal structure generation and (2) causal effect estimation (based on causal structure generated).

#### Causal Structure Generation

To generate the most feasible underlying causal structure from the MIMIC-Delirium dataset, we rely on causal structure learning algorithms (SLA), with assumptions of

causal sufficiency and faithfulness. Specifically, we apply eight (8) causal structure learning algorithms: (1) PC, (2) FCI, (3) GES, (4) GIES, (5) GDS, (6) LINGAM, (7) MMHC, and (8) MMTABU, with help from existing R libraries: (1) pcalg [60, 43] and (2) pchc [126]. With the application of these SLAs, we have eight (8) individual causal graphs. However, we apply majority voting to each edge to merge all this information together. This merging defines an edge as being present in the final graph if it is present in more than 50% cases (more than four graphs). Although this is a straightforward and naive solution to merge multiple causal graphs, we employ this ensembling method since no standard has been established in the literature yet. Figure 5.6 shows the final merged causal graph generated.

**Causal Effect Estimation**

With the causal structure generated, we now focus on causal effect estimation. For this purpose, we employ the pipeline proposed by Microsoft **Do-Why** library [114]:

- Modeling
- Identification
- Estimation
- Refutation

With modeling completed as part of the causal structure generation step, we

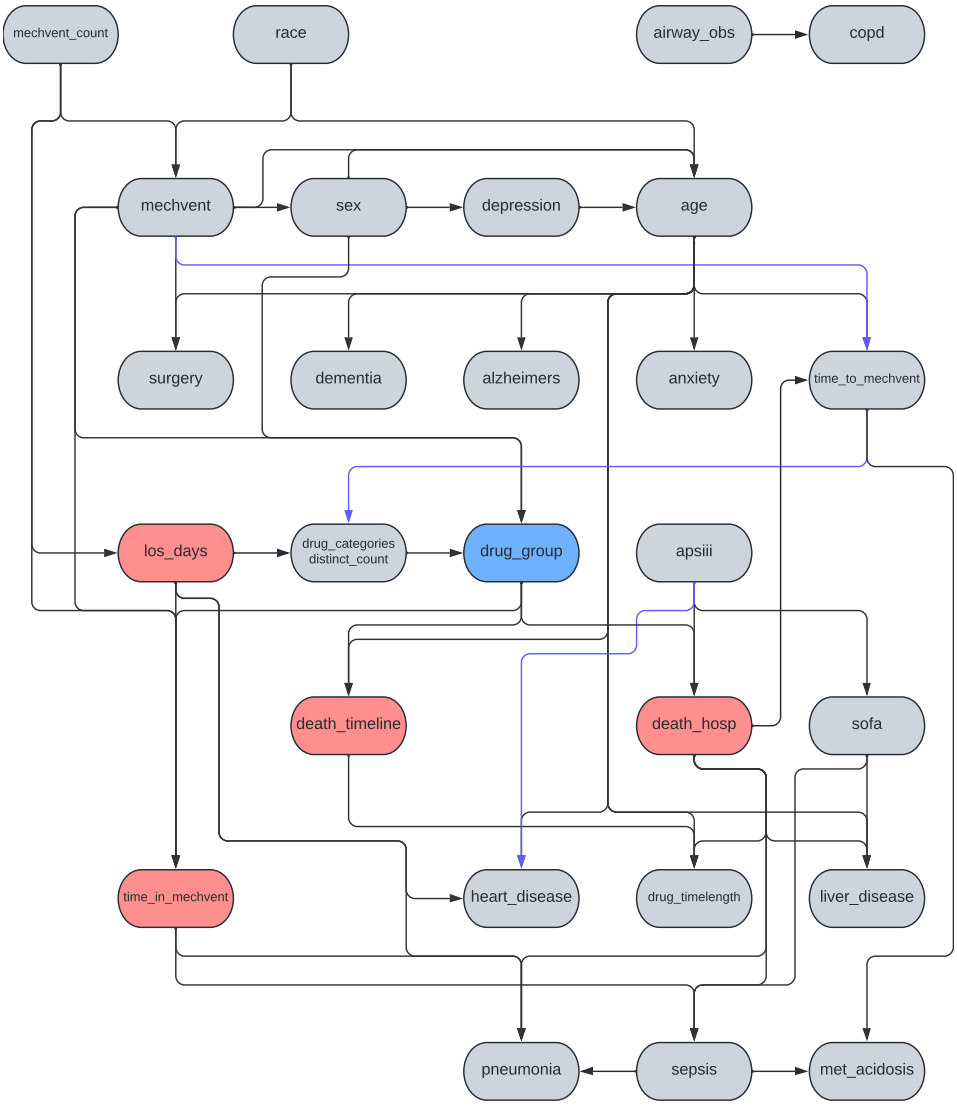



Figure 5.6: Combined Causal Graph for Delirium in the ICU *(blue: treatment, red: primary and secondary outcomes)*

now focus on causal effect identification and estimation. Based on the causal structure generated, we identify the conditional probability equation for the four target outcomes. Specifically, we express the do-calculus operations [13] in order to 'virtually' manipulate the outcomes. The do-calculus equations are presented below:

- $P(death\_in\_hosp|do(drug\_group)) = \sum_{age} P(death\_in\_hosp|drug\_group, age)P(age)$
- $P(death\_timeline|do(drug\_group)) = \sum_{age} P(death\_timeline|drug\_group, age)P(age)$
- $P(los\_days|do(drug\_group)) =$
  $\sum_{heart\_disease,mechvent} P(los\_days|drug\_group, heart\_disease, mechvent)P(heart\_disease, mechvent)$
- $P(time\_in\_mechvent|do(drug\_group)) =$
  $\sum_{age,mechvent} P(death\_in\_hosp|drug\_group, age, mechvent)P(age, mechvent)$

We now find the causal effect estimates based on these causal expressions identified. In Table 5.2, we present the causal effect estimations, as Average Treatment Effects (ATE), for treatment, aka, drug group on the four target outcomes. As shown in the table, the causal effect of treatment on death in Delirium and death timeline is very close. However, any drug, Haloperidol (1.8372) and other drugs (1.6102), does much better in reducing hospital length of stay compared to the no drug patient group (-0.0533). In addition to that, any drug performs better (8.1912) in reducing time in mechanical ventilation compared to no drug (4.4827), and Haloperidol does better (12.3007) than any other drugs (8.1912).

We now move to the final stage of causal effect estimation, which is the

| | Causal effect of drug group on: | | | |
|---|---|---|---|---|
| | **death in hospital** | **death timeline** | **length of stay in days** | **time in mech. vent.** |
| **Hal. vs. No Drug** | 0.0310 | -0.1291 | 1.8372 | 12.3007 |
| **Other Drug vs. No Drug** | 0.0216 | 0.0373 | 1.6102 | 8.1912 |
| **Hal. vs. Other Drug** | 0.0113 | -0.1386 | -0.0533 | 4.4827 |

Table 5.2: Outcomes estimation in Average Treatment Effects (ATE)

refutation of the estimated effect. We do so in four different steps: by adding a random common cause to the causal model, adding an unobserved common cause to the causal model, using a Placebo treatment, and using a subset of data. The expectation for these four is that:

- **Adding a random common cause**: should not change the estimated outcome from before since this should be adjusted by use of do-calculus expressions
- **Adding an unobserved common cause**: should change the estimated outcome from before since the unobserved confounder induces non-removable biases in the system
- **Using a placebo treatment**: should be close to zero since placebo treatment should not have any impact on the outcome
- **Using a subset of data**: should not change the estimated outcome from before since underlying data distribution did not change

Application of these four steps results in the following values, which also align with our expectations for a stable causal model and estimated effect:

- Estimated effect: 0.0309
- Add a random common cause: 0.0310
- Add an Unobserved Common Cause: 0.0262
- Use a Placebo Treatment: 0.0003
- Use a subset of data: 0.0319

## 5.4 Discussion

We have explored a potential observational study on Delirium patients in the ICU in this study. Our curated dataset is analyzed through two lenses: regular observational analysis and 'simulated' randomized controlled trial through the structural theory of causation. We have multiple novel contributions to this research work:

- Our observational study creates a prospective data cohort *(MIMIC-Delirium)* for Delirium patients
- Data properties for *MIMIC-Delirium* provides insight into the general patient demography in the ICU
- Machine learning-driven analysis on *MIMIC-Delirium* presents usage of

prediction-based computational modeling and these algorithm's general strengths

- Causal analysis on *MIMIC-Delirium* found:
  - No significant impact (X) of Antipsychotics choice in one of the primary outcomes, death in hospital
  - No significant impact (X) of Antipsychotics choice in length of stay in the ICU; however, usage of any drug shows better outcome (X) compared to that with no drugs
  - Haloperidol performs better (X) in affecting time in mechanical ventilation, compared to the similar impact of usage of other drugs or no drugs,

Our study relies on a few underlying assumptions. We assume that the Delirium patients in the ICU represent general Delirium demography since it occurs more frequently ( 80% cases in ICU) in the ICU compared to other traditional medical settings. Additionally, in generating the causal structure, we did not incorporate any background knowledge from peer-reviewed literature because of the existing controversies over the usage and benefits of Antipsychotics on the Delirium population *(discussed in the background section)*. One of the critical limitations of our study is the lack of involvement of experienced physicians actively working in

the ICU. Their involvement can aid in disputing general confusion in different parts of the study; however, bias from their understanding needs to be handled by involving multiple physicians. This limitation can be mitigated in future work.

In summary, our proposed analysis and pipeline create pathways for similar studies, especially in the healthcare research domain. The abundance of curated large electronic healthcare data presents a potential to find unexplored insights in a specific population group. Causal inference, especially the structural theory of causation, holds the potential to handle such research questions, look for causal insights, and report them appropriately.

# CHAPTER 6
# Conclusion and Future Work

## 6.1 Broader Impact and Summary Contribution of this dissertation

Search for causality is one of the core research questions in the healthcare research domain. Causal Inference is a great tool, built upon statistics and curated heavily for data science. Although many researchers are poking at exploring controversial research questions through Causal Inference, this dissertation primarily focused on the unexplored paradigms connected with various kinds of studies (RCT, Obs., PCT) conducted in the healthcare research domain. Our motivation was to improve the current shortcomings of healthcare research through the eyes of an ever-expanding data science arena.

The dissertation proposes novel methodologies on various critical points of Causal Inference methodologies directed at aiding Healthcare research. We have summarized the contributions in the following segments.

Our first study (CKH for SCM) proposes a novel methodology to compare and combine causal knowledge from multi-dimensional sources, such as experts' opinions, data, and literature, to derive domain-specific accurate SCMs. The methodology is incredibly beneficial for applied causal inference researchers, especially in the scientific fields of epidemiology, medicine, and social sciences, where insight into causal mechanisms is highly sought after. Additionally, our proposed

methodology allows adjustments of tier weights as fitted to work with shifting and evolving problem domains, such as the COVID-19 crisis. The process relies on the availability and abundance of causal knowledge sources, which faces challenges of curating experts in the field or extracting causal knowledge from literature (NLP).

Our second study (Causally Formulated HR) uses do-calculus to estimate causally formulated Hazard Ratio on survival dataset. Our proposed approach alters the original SCM into multiple SCMs with different endpoints. Doing so enables us to calculate conditional probabilities and thus backdoor adjustment on SCM. Our approach does not alter the original definition of HR; however, it formulates HR through alteration of SCM, which in effect uses only the causal effect of treatment on outcome. The notion is highly impactful since, through the transformation of SCM and backdoor adjustment, we get rid of biases from confounders and look at the causal survival effect of treatment on the outcome through hazard ratio.

Our third study (PCT through SCM) contributes to the ideation and use of structural causal models (SCM) for pragmatic clinical trials (PCT), commonly conducted in healthcare research. Our goal behind this representation of PCT through SCM is that, we expect PCT to be holding hidden causal information. Our idea shows the interaction and comparison between treatment provided and treatment received (which differs from treatment provided due to low adherence to PCT) through do-calculus equations. Our study has addressed vital design elements

of PCT: real-world population, real-world setting, appropriate comparison arm, and relevant outcome. The ideation of PCT through SCM would enable healthcare researchers to analyze more varieties of PCTs and other relevant trials. However, the prime challenge still remains in finding a suitable causal structure.

Our fourth and final study (MIMIC-Delirium) presents an application of Causal Inference methodologies in a specially curated dataset from a large EHR dataset. Our observational study has created a prospective data cohort for Delirium patients in the ICU. Analyzed data properties for MIMIC-Delirium have provided insight into the general patient demography. Our machine learning-driven analysis on MIMIC-Delirium has presented the strengths of prediction algorithms. Our causal analysis of MIMIC-Delirium has found:

- No significant impact of Antipsychotics choice in death in hospital,
- No significant impact of Antipsychotics choice in length of stay; however, any drug does better than no drugs, and,
- Haloperidol performs much better than other Antipsychotics or no drugs in affecting time in mechanical ventilation.

## 6.2 Future Work

The dissertation project has multiple directions it can be extended to. The causal knowledge hierarchy for causal structure estimation is a theoretical proposition; it

has the potential to be applied in specific problem domains. Although our research focus specializes in the healthcare domain, it can be significantly appropriate in other branches of science also, such as sociology, finance, agriculture, etc. On top of that, each tier of CKH has its category of knowledge sources. This categorization creates the potential to further extend within the tiers and investigate accordingly.

Our estimation of hazard ratio through the adjustment to the structural causal model is a pioneer in bridging traditional statistical methodologies with the newer concept of the structural causal model. This work can be extended to exploring time-varying interventions for various studies. Calculation of causally formulated hazard ratios for specific applications, such as the real-world trial of the effect of Antipsychotics in the Delirium patient group or the effect of COVID-19 vaccines in target population groups, is also a possible applied outcome of this work.

In expressing pragmatic clinical trials through structural causal models, further research can be extended by exploring instrumental variable analysis by using treatment $X$ as an instrumental variable. It can also be expanded to analyze time-series interventions with the definition proposed in place of point interventions.

Finally, in our applied work on Delirium patients in the ICU and the efficacy of Antipsychotics on them, this dissertation can be expanded in multidimensions. Our proposed framework can be used to analyze other relevant procedures, such as survival analysis or Cox regression analysis. In terms of Delirium, additional

correlated variables can be extracted to ensure a better fitted causal model and better prediction efficiency. A collaboration with ICU physicians can be done to generate an external validation dataset or create new trials. Regarding similar healthcare problem domains, the proposed framework can be recreated for other controversial research questions, such as the causal model for sepsis.

# BIBLIOGRAPHY


[1] B. Ackley, G. Ladwig, B. Swan, S. Tucker *et al.*, "Evidence based nursing care guidelines," *Medical Surgical Interventions. Mosby Elsevier, syf*, vol. 15, 2008.

[2] A. Adegunsoye, J. M. Oldham, S. K. Bellam, S. Montner, M. M. Churpek, I. Noth, R. Vij, M. E. Strek, and J. H. Chung, "Computed tomography honeycombing identifies a progressive fibrotic phenotype with increased mortality across diverse interstitial lung diseases," *Annals of the American Thoracic Society*, vol. 16, no. 5, pp. 580–588, 2019.

[3] R. Adib, P. Griffin, S. I. Ahamed, and M. Adibuzzaman, "A causally formulated hazard ratio estimation through backdoor adjustment on structural causal model," in *Machine Learning for Healthcare Conference*. PMLR, 2020, pp. 376–396.

[4] M. Adibuzzaman, K. Musselman, A. Johnson, P. Brown, Z. Pitluk, and A. Grama, "Closing the data loop: An integrated open access analysis platform for the mimic database," in *2016 Computing in Cardiology Conference (CinC)*. IEEE, 2016, pp. 137–140.

[5] A. K. Akobeng, "Principles of evidence based medicine," *Archives of disease in childhood*, vol. 90, no. 8, pp. 837–840, 2005.

[6] D. Aliprantis, "A distinction between causal effects in structural and rubin causal models," *FRB of Cleveland Working Paper No. 15-05*, 2015.

[7] B. Andrews, P. Spirtes, and G. F. Cooper, "On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 4002–4011.

[8] A. Anglemyer, H. T. Horvath, and L. Bero, "Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials," *Cochrane Database of Systematic Reviews*, no. 4, 2014.

[9] J. D. Angrist, G. W. Imbens, and D. B. Rubin, "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, vol. 91, no. 434, pp. 444–455, 1996.

[10] L. H. Arendt, C. H. Ramlau-Hansen, A. J. Wilcox, T. B. Henriksen, J. Olsen, and M. S. Lindhard, "Placental weight and male genital anomalies: a nationwide danish cohort study," *American journal of epidemiology*, vol. 183, no. 12, pp. 1122–1128, 2016.

[11] E. Bareinboim, J. Correa, D. Ibeling, and T. Icard, "On pearl's hierarchy and the foundations of causal inference," *ACM Special Volume in Honor of Judea Pearl (provisional title)*, 2020.

[12] E. Bareinboim and J. Pearl, "Controlling selection bias in causal inference," in *Artificial Intelligence and Statistics*. PMLR, 2012, pp. 100–108.

[13] ——, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7345–7352, 2016.

[14] ——, "Causal inference and the data-fusion problem," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7345–7352, 2016.

[15] M. Bikak, R. Adib, W. Ingram, P. Griffin, and M. Adibuzzaman, "Outcomes of use of antipsychotic for delirium in the icu: A big data approach," in *D50. CRITICAL CARE: THE METAMORPHOSIS-PAIN, SEDATION, DELIRIUM, ICU-ACQUIRED WEAKNESS, AND PALLIATIVE CARE*. American Thoracic Society, 2019, pp. A6672–A6672.

[16] G. Borboudakis and I. Tsamardinos, "Incorporating causal prior knowledge as path-constraints in bayesian networks and maximal ancestral graphs," *arXiv preprint arXiv:1206.6390*, 2012.

[17] K. Boutis and A. Willan, "Intention-to-treat and per-protocol analysis," *Cmaj*, vol. 183, no. 6, pp. 696–696, 2011.

[18] J. N. Carpenter and A. W. Lynch, "Survivorship bias and attrition effects in measures of performance persistence," *Journal of financial economics*, vol. 54, no. 3, pp. 337–374, 1999.

[19] D. Coggon, D. Barker, and G. Rose, *Chapter 9. Experimental studies*. John Wiley & Sons, 2009.

[20] S. R. Cole and M. A. Hernán, "Adjusted survival curves with inverse probability weights," *Computer methods and programs in biomedicine*, vol. 75, no. 1, pp. 45–49, 2004.

[21] J. Concato, "Observational versus experimental studies: what's the evidence for a hierarchy?" *NeuroRx*, vol. 1, no. 3, pp. 341–347, 2004.

[22] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[23] P. Cummings, "The relative merits of risk ratios and odds ratios," *Archives of pediatrics & adolescent medicine*, vol. 163, no. 5, pp. 438–445, 2009.

[24] V. Didelez and N. Sheehan, "Mendelian randomization as an instrumental variable approach to causal inference," *Statistical methods in medical research*, vol. 16, no. 4, pp. 309–330, 2007.

[25] M. Drton and M. H. Maathuis, "Structure learning in graphical modeling," *Annual Review of Statistics and Its Application*, vol. 4, pp. 365–393, 2017.

[26] M. J. Druzdzel and F. J. Díez, "Combining knowledge from different sources in causal probabilistic models," *The Journal of Machine Learning Research*, vol. 4, pp. 295–316, 2003.

[27] L. Eriksson, E. Johansson, N. Kettaneh-Wold, C. Wikström, and S. Wold, "Design of experiments," *Principles and Applications, Learn ways AB, Stockholm*, 2000.

[28] R. A. Fisher, "Design of experiments," *Br Med J*, vol. 1, no. 3923, pp. 554–554, 1936.

[29] R. A. Fisher *et al.*, "The design of experiments." *The design of experiments.*, no. 7th Ed, 1960.

[30] D. Freedman and P. Humphreys, "Are there algorithms that discover causal structure?" *Synthese*, vol. 121, no. 1, pp. 29–54, 1999.

[31] T. R. Frieden, "Evidence for health decision making—beyond randomized, controlled trials," *New England Journal of Medicine*, vol. 377, no. 5, pp. 465–475, 2017.

[32] V. Gamerman, T. Cai, and A. Elsäßer, "Pragmatic randomized clinical trials: best practices and statistical guidance," *Health Services and Outcomes Research Methodology*, vol. 19, no. 1, pp. 23–35, 2019.

[33] M. Gangl, "Causal inference in sociological research," *Annual review of sociology*, vol. 36, 2010.

[34] T. D. Girard, M. C. Exline, S. S. Carson, C. L. Hough, P. Rock, M. N. Gong, I. S. Douglas, A. Malhotra, R. L. Owens, D. J. Feinstein *et al.*, "Haloperidol and ziprasidone for treatment of delirium in critical illness," *New England Journal of Medicine*, vol. 379, no. 26, pp. 2506–2516, 2018.

[35] T. D. Girard, P. P. Pandharipande, and E. W. Ely, "Delirium in the intensive care unit," *Critical care*, vol. 12, no. S3, p. S3, 2008.

[36] T. A. Glass, S. N. Goodman, M. A. Hernán, and J. M. Samet, "Causal inference in public health," *Annual review of public health*, vol. 34, pp. 61–75, 2013.

[37] I. D. Gow, D. F. Larcker, and P. C. Reiss, "Causal inference in accounting research," *Journal of Accounting Research*, vol. 54, no. 2, pp. 477–523, 2016.

[38] J. A. Greene and S. H. Podolsky, "Reform, regulation, and pharmaceuticals—the kefauver–harris amendments at 50," *New England Journal of Medicine*, vol. 367, no. 16, pp. 1481–1483, 2012.

[39] J. Grossman and F. J. Mackenzie, "The randomized controlled trial: gold standard, or merely standard?" *Perspectives in biology and medicine*, vol. 48, no. 4, pp. 516–534, 2005.

[40] S. K. Gupta, "Intention-to-treat concept: a review," *Perspectives in clinical research*, vol. 2, no. 3, p. 109, 2011.

[41] E. Hariton and J. J. Locascio, "Randomised controlled trials—the gold standard for effectiveness research," *BJOG: an international journal of obstetrics and gynaecology*, vol. 125, no. 13, p. 1716, 2018.

[42] K. Hatta, Y. Kishi, K. Wada, T. Odawara, T. Takeuchi, T. Shiganami, K. Tsuchida, Y. Oshima, N. Uchimura, R. Akaho *et al.*, "Antipsychotics for delirium in the general hospital setting in consecutive 2453 inpatients: a prospective observational study," *International journal of geriatric psychiatry*, vol. 29, no. 3, pp. 253–262, 2014.

[43] A. Hauser and P. Bühlmann, "Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2409–2464, 2012.

[44] C. Heinze-Deml, M. H. Maathuis, and N. Meinshausen, "Causal structure learning," *Annual Review of Statistics and Its Application*, vol. 5, pp. 371–391, 2018.

[45] M. A. Hernán, "The hazards of hazard ratios," *Epidemiology (Cambridge, Mass.)*, vol. 21, no. 1, p. 13, 2010.

[46] ——, "The c-word: scientific euphemisms do not improve causal inference from observational data," *American journal of public health*, vol. 108, no. 5, pp. 616–619, 2018.

[47] M. A. Hernán, A. Alonso, R. Logan, F. Grodstein, K. B. Michels, M. J. Stampfer, W. C. Willett, J. E. Manson, and J. M. Robins, "Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease," *Epidemiology (Cambridge, Mass.)*, vol. 19, no. 6, p. 766, 2008.

[48] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins, "A structural approach to selection bias," *Epidemiology*, pp. 615–625, 2004.

[49] M. A. Hernán and J. M. Robins, "Observational studies analyzed like randomized trials and vice versa," in *Methods in Comparative Effectiveness Research*. Chapman and Hall/CRC, 2017, pp. 127–148.

[50] M. A. Hernán, J. M. Robins *et al.*, "Per-protocol analyses of pragmatic trials," *N Engl J Med*, vol. 377, no. 14, pp. 1391–1398, 2017.

[51] M. A. Hernán, "A definition of causal effect for epidemiological research," *Journal of Epidemiology & Community Health*, vol. 58, no. 4, pp. 265–271, 2004.

[52] M. Á. Hernán, B. Brumback, and J. M. Robins, "Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men," *Epidemiology*, pp. 561–570, 2000.

[53] S. Hofer, C. Bopp, C. Hoerner, K. Plaschke, R. M. Faden, E. Martin, H. J. Bardenheuer, and M. A. Weigand, "Injury of the blood brain barrier and up-regulation of icam-1 in polymicrobial sepsis," *Journal of Surgical Research*, vol. 146, no. 2, pp. 276–281, 2008.

[54] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[55] ICD9Data.com. Icd-9-cm diagnosis code 293.0 : Delirium due to conditions classified elsewhere. [Online]. Available: http://www.icd9data.com/2015/Volume1/290-319/290-294/293/293.0.htm

[56] G. W. Imbens and D. B. Rubin, "Rubin causal model," in *Microeconometrics*. Springer, 2010, pp. 229–241.

[57] S. K. Inouye, C. H. van Dyck, C. A. Alessi, S. Balkin, A. P. Siegal, and R. I. Horwitz, "Clarifying confusion: the confusion assessment method: a new method for detection of delirium," *Annals of internal medicine*, vol. 113, no. 12, pp. 941–948, 1990.

[58] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, p. 160035, 2016.

[59] B. C. Johnston and G. H. Guyatt, "Best (but oft-forgotten) practices: intention-to-treat, treatment adherence, and missing participant outcome data in the nutrition literature," *The American journal of clinical nutrition*, vol. 104, no. 5, pp. 1197–1201, 2016.

[60] M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann, "Causal inference using graphical models with the r package pcalg," *Journal of statistical software*, vol. 47, pp. 1–26, 2012.

[61] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[62] K. A. Keith, D. Jensen, and B. O'Connor, "Text and causal inference: A review of using text to remove confounding from causal estimates," *arXiv preprint arXiv:2005.00649*, 2020.

[63] D. G. Kleinbaum and M. Klein, *Survival analysis*. Springer, 2010, vol. 3.

[64] D. Knuth. Clgaussian-test. [Online]. Available: https://www.bnlearn.com/documentation/man/clgaussian-test.html

[65] C. P. Kovesdy and K. Kalantar-Zadeh, "Observational studies versus randomized controlled trials: avenues to causal inference in nephrology," *Advances in chronic kidney disease*, vol. 19, no. 1, pp. 11–18, 2012.

[66] W. A. Kukull and M. Ganguli, "Generalizability: the trees, the forest, and the low-hanging fruit," *Neurology*, vol. 78, no. 23, pp. 1886–1891, 2012.

[67] A. Landreth and A. J. Silva, "The need for research maps to navigate published work and inform experiment planning," *Neuron*, vol. 79, no. 3, pp. 411–415, 2013.

[68] D. J. Lederer, S. C. Bell, R. D. Branson, J. D. Chalmers, R. Marshall, D. M. Maslove, D. E. Ost, N. M. Punjabi, M. Schatz, A. R. Smyth *et al.*, "Control of confounding and reporting of results in causal inference studies. guidance

for authors from editors of respiratory, sleep, and critical care journals," *Annals of the American Thoracic Society*, vol. 16, no. 1, pp. 22–28, 2019.

[69] S. Lee, J. Correa, and E. Bareinboim, "Generalized transportability: Synthesis of experiments from heterogeneous domains," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York, NY: AAAI Press, 2020.

[70] S. Lee and E. Bareinboim, "Causal effect identifiability under partial-observability," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5692–5701.

[71] S. Lee, J. D. Correa, and E. Bareinboim, "General identifiability with arbitrary surrogate experiments," in *Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 389–398.

[72] H. A. Linstone, M. Turoff *et al.*, *The delphi method*. Addison-Wesley Reading, MA, 1975.

[73] S. Listl, H. Jürges, and R. G. Watt, "Causal inference from observational data," *Community dentistry and oral epidemiology*, vol. 44, no. 5, pp. 409–415, 2016.

[74] K. Loudon, S. Treweek, F. Sullivan, P. Donnan, K. E. Thorpe, and M. Zwarenstein, "The precis-2 tool: designing trials that are fit for purpose," *bmj*, vol. 350, 2015.

[75] R. W. Makuch, "Adjusted survival curve estimation using covariates," *Journal of chronic diseases*, vol. 35, no. 6, pp. 437–443, 1982.

[76] C. E. McCoy, "Understanding the intention-to-treat principle in randomized controlled trials," *Western Journal of Emergency Medicine*, vol. 18, no. 6, p. 1075, 2017.

[77] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[78] MIT-LCP. mimic-code/sofa.sql at main · mit-lcp/mimic-code. [Online]. Available: https://github.com/MIT-LCP/mimic-code/blob/main/mimic-iii/concepts/severityscores/sofa.sql

[79] V. M. Montori and G. H. Guyatt, "Intention-to-treat principle," *Cmaj*, vol. 165, no. 10, pp. 1339–1341, 2001.

[80] E. J. Murray, E. C. Caniglia, and L. C. Petito, "Causal survival analysis: A guide to estimating intention-to-treat and per-protocol effects from randomized clinical trials with non-adherence," *Research Methods in Medicine & Health Sciences*, vol. 2, no. 1, pp. 39–49, 2021.

[81] E. J. Murray, S. A. Swanson, and M. A. Hernán, "Guidelines for estimating causal effects in pragmatic randomized trials," *arXiv preprint arXiv:1911.06030*, 2019.

[82] K. J. Neufeld, J. Yue, T. N. Robinson, S. K. Inouye, and D. M. Needham, "Antipsychotic medication for prevention and treatment of delirium in hospitalized adults: a systematic review and meta-analysis," *Journal of the American Geriatrics Society*, vol. 64, no. 4, pp. 705–714, 2016.

[83] A. Nichol, M. Bailey, D. Cooper, O. behalf of the POLAR *et al.*, "Challenging issues in randomised controlled trials," *Injury*, vol. 41, pp. S20–S23, 2010.

[84] A. Nichols, "Causal inference with observational data," *The Stata Journal*, vol. 7, no. 4, pp. 507–541, 2007.

[85] T. Nishioku, S. Dohgu, F. Takata, T. Eto, N. Ishikawa, K. B. Kodama, S. Nakagawa, A. Yamauchi, and Y. Kataoka, "Detachment of brain pericytes from the basal lamina is involved in disruption of the blood–brain barrier caused by lipopolysaccharide-induced sepsis in mice," *Cellular and molecular neurobiology*, vol. 29, no. 3, pp. 309–316, 2009.

[86] G. Nordon, G. Koren, V. Shalev, B. Kimelfeld, U. Shalit, and K. Radinsky, "Building causal graphs from medical literature and electronic medical records," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1102–1109.

[87] S. Norris, D. Atkins, W. Bruening, S. Fox, E. Johnson, R. Kane, S. C. Morton, M. Oremus, M. Ospina, G. Randhawa *et al.*, "Selecting observational studies for comparing medical interventions," in *Methods guide for effectiveness and comparative effectiveness reviews [Internet]*. Agency for Healthcare Research and Quality (US), 2010.

[88] T. M. Palmer, J. A. Sterne, R. M. Harbord, D. A. Lawlor, N. A. Sheehan, S. Meng, R. Granell, G. D. Smith, and V. Didelez, "Instrumental variable estimation of causal risk ratios and causal odds ratios in mendelian randomization analyses," *American journal of epidemiology*, vol. 173, no. 12, pp. 1392–1403, 2011.

[89] J. Pearl, *Causality*. Cambridge university press, 2009.

[90] ——, "Causal inference," *Causality: Objectives and Assessment*, pp. 39–58, 2010.

[91] ——, "An introduction to causal inference," *The international journal of biostatistics*, vol. 6, no. 2, 2010.

[92] ——, "Theoretical impediments to machine learning with seven sparks from the causal revolution," *arXiv preprint arXiv:1801.04016*, 2018.

[93] ——, "The seven tools of causal inference, with reflections on machine learning," *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[94] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics surveys*, vol. 3, pp. 96–146, 2009.

[95] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[96] M. A. Pisani, S. Y. J. Kong, S. V. Kasl, T. E. Murphy, K. L. Araujo, and P. H. Van Ness, "Days of delirium are associated with 1-year mortality in an older intensive care unit population," *American journal of respiratory and critical care medicine*, vol. 180, no. 11, pp. 1092–1097, 2009.

[97] F. Porzsolt, N. G. Rocha, A. C. Toledo-Arruda, T. G. Thomaz, C. Moraes, T. R. Bessa-Guerra, M. Leão, A. Migowski, A. R. A. da Silva, and C. Weiss, "Efficacy and effectiveness trials have different goals, use different tools, and generate different messages," *Pragmatic and observational research*, vol. 6, p. 47, 2015.

[98] M. Purgato, C. Barbui, S. Stroup, and C. Adams, "Pragmatic design in randomized controlled trials," *Psychological medicine*, vol. 45, no. 2, pp. 225–230, 2015.

[99] J. M. Robins, "Causal inference from complex longitudinal data," in *Latent variable modeling and applications to causality*. Springer, 1997, pp. 69–117.

[100] P. R. Rosenbaum, "Observational study," *Encyclopedia of statistics in behavioral science*, 2005.

[101] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[102] K. J. Rothman and S. Greenland, "Causation and causal inference in epidemiology," *American journal of public health*, vol. 95, no. S1, pp. S144–S150, 2005.

[103] A. Rotnitzky and J. M. Robins, "Inverse probability weighting in survival analysis," *Wiley StatsRef: Statistics Reference Online*, 2014.

[104] B. M. Rottman and F. C. Keil, "Causal structure learning over time: Observations and interventions," *Cognitive psychology*, vol. 64, no. 1-2, pp. 93–125, 2012.

[105] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[106] ——, "Bayesian inference for causal effects: The role of randomization," *The Annals of statistics*, pp. 34–58, 1978.

[107] ——, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[108] F. E. Saal, R. G. Downey, and M. A. Lahey, "Rating the ratings: Assessing

the psychometric quality of rating data." *Psychological bulletin*, vol. 88, no. 2, p. 413, 1980.

[109] G. A. Satten and S. Datta, "The kaplan–meier estimator as an inverse-probability-of-censoring weighted average," *The American Statistician*, vol. 55, no. 3, pp. 207–210, 2001.

[110] D. K. Schlüter, R. Griffiths, A. Adam, A. Akbari, M. L. Heaven, S. Paranjothy, A.-M. N. Andersen, S. B. Carr, T. Pressler, P. J. Diggle *et al.*, "Impact of cystic fibrosis on birthweight: a population based study of children in denmark and wales," *Thorax*, vol. 74, no. 5, pp. 447–454, 2019.

[111] M. Scutari, C. E. Graafland, and J. M. Gutiérrez, "Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms," *International Journal of Approximate Reasoning*, vol. 115, pp. 235–253, 2019.

[112] P. Sedgwick, "Intention to treat analysis versus per protocol analysis of trial data," *Bmj*, vol. 350, 2015.

[113] A. Semmler, T. Okulla, M. Sastre, L. Dumitrescu-Ozimek, and M. T. Heneka, "Systemic inflammation induces apoptosis with variable vulnerability of different brain regions," *Journal of chemical neuroanatomy*, vol. 30, no. 2-3, pp. 144–157, 2005.

[114] A. Sharma, E. Kiciman *et al.*, "DoWhy: A Python package for causal inference," https://github.com/microsoft/dowhy, 2019.

[115] S. Shimizu, "Lingam: Non-gaussian methods for estimating causal structures," *Behaviormetrika*, vol. 41, no. 1, pp. 65–98, 2014.

[116] B. Sibbald and M. Roland, "Understanding controlled trials. why are randomised controlled trials important?" *BMJ: British Medical Journal*, vol. 316, no. 7126, p. 201, 1998.

[117] V. A. Smith, C. J. Coffman, and M. G. Hudgens, "Interpreting the results of intention-to-treat, per-protocol, and as-treated analyses of clinical trials," *JAMA*, vol. 326, no. 5, pp. 433–434, 2021.

[118] P. Spirtes, C. Glymour, R. Scheines, S. Kauffman, V. Aimale, and F. Wimberly, "Constructing bayesian network models of gene expression networks from microarray data," 2000.

[119] P. Spirtes, "Introduction to causal inference." *Journal of Machine Learning Research*, vol. 11, no. 5, 2010.

[120] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[121] P. Spirtes and K. Zhang, "Causal discovery and inference: concepts and recent methodological advances," in *Applied informatics*, vol. 3, no. 1. Springer, 2016, p. 3.

[122] ——, "Causal discovery and inference: concepts and recent methodological advances," in *Applied informatics*, vol. 3. SpringerOpen, 2016, pp. 1–28.

[123] S. L. Spruance, J. E. Reid, M. Grace, and M. Samore, "Hazard ratio in clinical trials," *Antimicrobial agents and chemotherapy*, vol. 48, no. 8, pp. 2787–2792, 2004.

[124] S. D. Stovitz and I. Shrier, "Causal inference for clinicians," *BMJ evidence-based medicine*, 2019.

[125] P. W. Tennant, E. J. Murray, K. F. Arnold, L. Berrie, M. P. Fox, S. C. Gadd, W. J. Harrison, C. Keeble, L. R. Ranker, J. Textor *et al.*, "Use of directed acyclic graphs (dags) to identify confounders in applied health research: review and recommendations," *International journal of epidemiology*, vol. 50, no. 2, pp. 620–632, 2021.

[126] M. Tsagris, *pchc: Bayesian Network Learning with the PCHC and Related Algorithms*, 2021, r package version 0.6. [Online]. Available: https://CRAN.R-project.org/package=pchc

[127] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine learning*, vol. 65, no. 1, pp. 31–78, 2006.

[128] C. Vitolo, M. Scutari, M. Ghalaieny, A. Tucker, and A. Russell, "Modeling air pollution, climate, and health data using bayesian networks: A case study of the english regions," *Earth and Space Science*, vol. 5, no. 4, pp. 76–88, 2018.

[129] W. Wang, G. Hu, B. Yuan, S. Ye, C. Chen, Y. Cui, X. Zhang, and L. Qian, "Prior-knowledge-driven local causal structure learning and its application on causal discovery between type 2 diabetes and bone mineral density," *IEEE Access*, vol. 8, pp. 108 798–108 810, 2020.

[130] A. Wassenaar, M. van den Boogaard, T. van Achterberg, A. Slooter, M. Kuiper, M. Hoogendoorn, K. Simons, E. Maseda, N. Pinto, C. Jones *et al.*, "Multinational development and validation of an early prediction model for delirium in icu patients," *Intensive care medicine*, vol. 41, no. 6, pp. 1048–1056, 2015.

[131] Z. Wood-Doughty, I. Shpitser, and M. Dredze, "Challenges of using text classifiers for causal inference," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2018. NIH Public Access, 2018, p. 4586.