Causal Learning Handbook

Contents

Welcome!

This site introduces the **Causal Roadmap**, **Targeted Maximum Likelihood Estimation (TMLE)**, and **Super Learner** for modern causal inference in epidemiology, with an emphasis on pharmacoepidemiology and clinical trial analysis.

Chapter 1

Tutorial on Targeted Learning and the Causal Roadmap

1.1 Introduction

This tutorial provides a gentle introduction to the Causal Roadmap and its applications in pharmaco-epidemiologic research. It is designed for a broad audience, including learners from both academia and industry. We systematically walk through each step of the Causal Roadmap—from explicitly formulating a research question, to translating it into a formal causal estimand, to identifying and estimating that estimand from observed data, and finally to drawing valid inferences and interpreting results. Each step is illustrated using a working example from a pharmaco-epidemiology setting, accompanied by interactive, built-in code to facilitate hands-on learning. The structure and content of this tutorial follow, in an analytical way, the Introduction to Causal Inference and the causal Roadmap course (htp://www.ucbbiostat.com/) Petersen and Balzer.

1.2 Why venture down a new path?

Adopting the Causal Roadmap in our approach to research in causal inference enables us to clearly state a scientific question and select an analtyic approach that matches the question being asked while ensuring systematic assessment of our ability/feasibility to answer this question from the data we observe (identifiability). Head to head analysis method comparison lets us select the best approach.

We will now formally introduce the Causal Roadmap but before let us go over some notation!

1.3 Notation

- ***A***: Exposure/Treatement
    - The term treatment is often used in causal inference even with exposures that are not medical treatments. We shall use A=1 for exposed (treated) and A=0 for unexposed (untreated)
- ***Y***: outcome
- ***W***: set of measured confounding variables
- ***U***: set of unmeasured factors
- $\mathbb{E}[Y|A=a]$: expected outcome Y among those who experience exposure A=a in our population. This is a descriptive measure
- $\mathbb{E}[Y_a]$: expected counterfactual outcome Y_a when all experience exposure A=a in our population. This is a causal quantity. Generally $\mathbb{E}[Y|A=a]$ does not equal to $\mathbb{E}[Y_a]$ and this is the fundamental problem of causal inference
- $\mathbb{E}[Y|A=a,W=w]$: expected outcome Y among those who expereince exposure A=a and have covariates W=w, in our population. For example this can be the mean outcome among exposed men. These conditional expectations are often estimated using multivariable regression models.
- $\mathbb{E}[\mathbb{E}[Y|A=a,W=w]]$:expected outcome Y among those who experience exposure A=a and have covariates W=w,averaged across covariate strata in the population. This is a marginal expectation.

1.4 Motivation

- Suppose we are interested in the impact of Drug A vs Drug B on risk of cardiovascular disease among postmenopausal women with osteoporosis.
- Our usual approach would be to collect data on the intervention, outcome (cardiovascular disease) and some covariates. Since the outcome is binary, we would use a logistic regression to estimate the conditional odds ratio by exponentiating the regression coefficient on the intervention (treatment).
- The problem with is approach is that it allows the tool i.e. logistic regression to define the question we answer rather than starting with the question and picking amongst tools that allow us to answer the question.
- To address this problem, we introduce the Causal Roadmap!

1.5 The Causal Roadmap

The Causal Roadmap is a framework that provides a systematic process to move from a research question to estimation and interpretation which guides investigators on how to design and analyse their studies a priori. This framework has the following steps;

- Stating the research question and hypothetical experiment
- Defining the causal model and parameter of interest
- Linking the causal model to the observed data and defining the statistical model
- Assessing identifiability: linking the causal effect to a parameter estimable from the observed data
- Selecting and applying the estimator
- Deriving an estimate of the sampling distribution (statistical uncertainty)
- Making inference (interpreting findings)

We shall now delve into each of these steps in details!

1.6 Step 0: State the question

- This is the very first step of the roadmap. A helpful way to be clear about the scientific question is to explicitly state the experiment that would unambiguously yield estimates of the causal effect of interest.
- For example: What is the effect of a certain medication on the incidence of cardiovascular disease among postmenopausal women who initiated Drug A vs Drug B in the United States?
- We can consider a hypothetical experiment where we ask what would be the the difference in CVD incidence if patients received the intervention drug A vs if all patients received the control drug B (or standard of care).
- To sharply frame our research question, we want to be more specific about;
  - The target population (What age group? where?)
  - The exposure (What dosage? Frequency?)
  - The outcome (over what timeframe?)
  - Ways to change the exposure and their plausibility
- Other interesting hypothetical experiments could include:
  - What would be the difference in CVD incidence if patients were initiated on drug A once they reached a certain risk threshold vs if all patients are initiated on Drug A regardless of their risk profile?
  - What would be the difference in CVD incidence if an additional 10% of patients received the intervention compared to if the intervention uptake remained as observed?

- We note that there is massive flexibility in how we can define our desired hypothetical experiments.

1.6.1 Target Trial Emulation

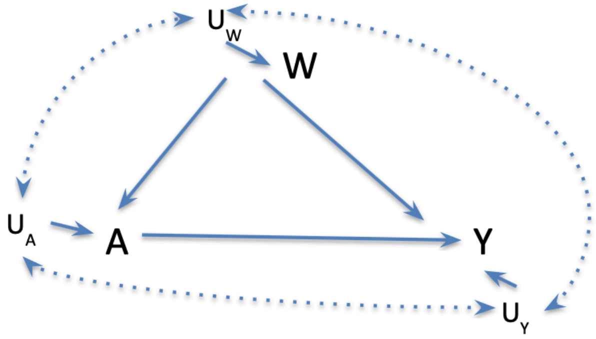
The hypothetical experiment defined in Step 0 can be viewed as a target trial.

Observational studies aim to emulate this trial by aligning eligibility criteria, treatment assignment, follow-up, outcome definitions, and estimands.

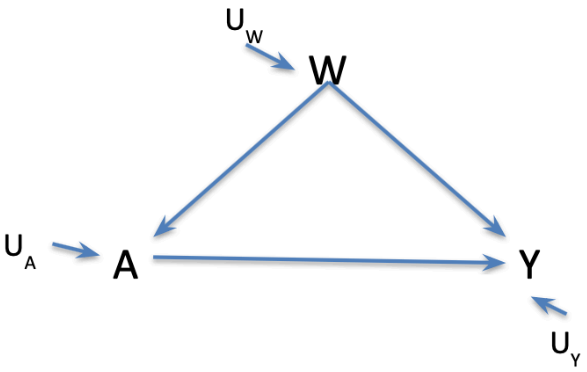
The Causal Roadmap provides the formal structure for conducting such emulations transparently.

1.7 Step 1: Define the causal model

- Causal modeling formalizes our knowledge however limited. We are able to explore which variables affect each other, examine the role of unmeasured factors and the functional form of the relationships between variables.
- In this tutorial, we shall focus on structural causal models and corresponding causal graphs (Pearl 2000). However, we do note that their are many other causal frameworks.
- The figure 1 below corresponds to a simple causal graph with corresponding structural casual model as follows;
  - $W = f_w(U_w)$
  - $A = f_A(W, U_A)$
  - $Y = f_Y(W, A, U_Y)$
- We make no assumptions on the background factors (U_w, U_A, U_Y) or on the functional forms of functions (f_w, f_A, f_Y)

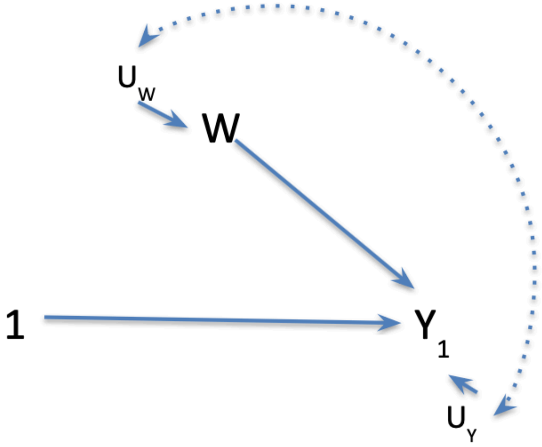


- If you believed no unmeasured confounding, a possible causal model and graph (figure 2) would be;
  - $W = f_w(U_w)$
  - $A = f_A(W, U_A)$
  - $Y = f_Y(W, A, U_Y)$
- Here we assume that the background factors are all independent but still make no assumption on the functional forms of (f_w, f_A, f_Y)
- However, it is important to note that wishing for something does not make it true.

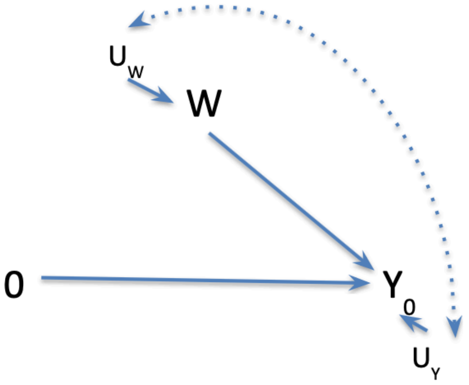


1.8 Step 2: Define the causal parameter of interest

- We now define counterfactuals by intervening on the causal model. We can do this by setting the exposure to a specific level e.g A=1 for all units.
  - $W = f_w(U_w)$
  - $A = 1$
  - $Y_1 = f_Y(W, 1, U_Y)$ where Y_1 is the outcome if possibly-contrary to fact, the unit was exposed (A=1)



- Analogously, we can intervene on the causal model by setting A=0
  - $W = f_w(U_w)$
  - $A = 0$
  - $Y_0 = f_Y(W, 0, U_Y)$ where Y_0 is the outcome if possibly-contrary to fact, the unit was exposed (A=0)



- We use counterfactuals to define the causal parameter;
  - For example, the difference between the expected counterfactual outcomes under these two interventions i.e $\mathbb{E}[Y_1] - \mathbb{E}[Y_0]$ which is known as the average treatment effect(ATE)
  - For a binary outcome, we define the causal risk difference (CRD) as $\mathbb{P}(Y_1 = 1) - \mathbb{P}(Y_0 = 1)$.
- Many other causal parameters are possible!!

1.8.1 Estimand Specification (ICH E9[R1] Framework)

An estimand precisely defines the treatment effect of interest by specifying all attributes of the causal question.

Attribute	Specification
Population	Eligible individuals meeting study inclusion criteria
Treatment Strategies	Intervention A versus comparator B
Endpoint	Binary or time-to-event outcome within a fixed horizon
Intercurrent Events	Addressed via treatment-policy or hypothetical strategy
Summary Measure	Risk difference, risk ratio, or mean difference

Explicit estimand specification ensures alignment between the scientific question, identification assumptions, and estimation strategy.

1.8.2 Treatment-Policy versus Hypothetical Estimands

A treatment-policy estimand contrasts outcomes under initial treatment assignment regardless of subsequent treatment changes (i.e., intention-to-treat estimates as presented in this workshop).

A hypothetical estimand contrasts outcomes under a counterfactual world in which intercurrent events (e.g., switching or discontinuation) do not occur.

The choice between these estimands reflects different scientific questions and determines how intercurrent events are handled during analysis.

1.8.3 Intercurrent Events

Intercurrent events are post-treatment events that affect the interpretation or existence of the outcome, such as treatment switching, discontinuation, or death.

Handling of intercurrent events must be specified at the estimand stage, not deferred to estimation. Common strategies include:

- Treatment-policy: ignore the intercurrent event
- Hypothetical: censor or reweight to eliminate its occurrence
- Composite: redefine the outcome to include the event

This choice determines the causal question being answered.

1.8.4 Time-to-Event Outcomes and Risk-Based Estimands

In many applications, outcomes occur over time and are subject to censoring.

Rather than targeting hazard ratios, the Causal Roadmap naturally accommodates risk-based estimands, such as cumulative incidence at a fixed time horizon.

For example, the causal risk difference at 90 days compares the probability of experiencing the event by day 90 under each treatment strategy.

1.9 Step 3: Link to observed data

- Observed data are denoted O=(W,A,Y) where W reprensents measured covariates, A is the exposure and Y is the outcome.
- We assume that the causal model provides a description of our study under existing conditions(i.e. the real world) and under interventions (i.e.the counterfactual world)
- This provides a link between the causal world and the real (observed) world and therefore our causal model implies our statistical model which is the set of possible distributions of observed data.
- The causal model may but often does not place any restrictions on the statistical model in which case the statistical model is ***non parametric***.
- For example our model says that A is a function of W and U_A but does not specify the form of that function: A= $f_A(W, U_A)$. However, if we know the form, that should be specified in the causal model.

1.9.1 Observed-Data Censoring Rules

The observed data structure must specify which events terminate follow-up and how they relate to the estimand.

Censoring may occur due to administrative end of follow-up, loss to follow-up, or treatment switching.

Whether censoring is causal or administrative depends on the estimand and must be addressed through design or analysis.

1.10 Step 4: Assess Identifiablity

- This process involves linking the causal effect to the parameter estimable from observed data. This requires some assumptions as follows:
  - Temporality: exposure precedes the outcome. This is indicated by an arrow on the causal graph from A to Y