Topics in Machine Learning
Machine Learning for Healthcare

Rahul G. Krishnan
Assistant Professor
Computer science & Laboratory Medicine and Pathobiology

Part I - Randomization

Interventions in physical systems

Equations characterize our understanding of physical systems

Laws of conservation give us equalities that must hold

We can use these equations to answer questions around the effect of interventions

Source: https://www.youtube.com/watch?v=yMdSrTyXuKU

Medicine has progressed a long way

- We've come a long way in our understanding of the human body
- But processes in the body occur at multiple scales and we lack mechanistic models to characterize the effect of every intervention

How should we make decisions in healthcare?

History

Rosini et. al, Vaccines Against Antimicrobial Resistance, 2020

Discovery of early medicines/interventions

- Penicillin
  - Discovered in 1928 by accident when studying the properties of the *Penicillium* mold
  - Hypothesized mechanism was unknown but scientists knew it killed bacteria
  - Reason for use:
    - First used to successfully treat an eye infection in children
- 1945 – First Randomized Control Trial
  - Published "Streptomycin treatment of pulmonary tuberculosis" by Austin Bradford Hill
  - May be likely that the idea of randomization existed long before medical publications
  - Prospective biomedical research study designed to answer specific questions about an intervention

What is a randomized control trial?

Control: Blue pill [the treatment representing the standard of care]

Treatment: The new intervention that is being evaluated

Randomize control trials

- Step 1: Select a cohort of individuals
- Step 2: Randomly split the cohort into two groups
- Step 3: Give one group the control and give the other group the treatment
- Step 4: Observe the patients over time and see who got better

Randomization as a graphical model

Treatment

Control

$$ATE = \sum_{i \in \mathcal{T}} y_t - \sum_{i \in \mathcal{C}} y_c \approxeq \mathbb{E}[Y_t - Y_c]$$

Y is an outcome of interest (positive value good; negative value bad)

Features of a randomized control trial

- Cohort design & randomization
- Study phases
- Controls
- Blinding

Cohort design

Cohort Design & Study plan

- Selection criteria: Who qualifies to participate
- How many people in study?
- Length of study & choice of intervention
- **Discussion: If you were a drug company, and were designing the cohort, what are considerations you might want to incorporate in the design of your cohort.**

Why does randomization work?

- There are many demographic factors that are shared/unique across patients
- Randomization ensures that these are evenly distributed across treatment and control groups
- Ensures that there is no *selection bias* in the cohort

Types of randomization

- Complete randomization
  - Flip a coin
- Stratified randomization
  - Ensuring balance across imbalanced covariates in treatment/control group
  - 100 people but only 10 males in cohort
- Cluster randomization
  - What if you want to test an intervention across “groups” rather than individuals
  - Useful if you suspect the treatment effect across individuals in a group is correlated
  - Examples:
    - Clustering children into classrooms and applying randomization at the classroom level

Phases of a clinical trial

Phases of modern drug design

- Develop a new drug
- Preclinical phase:
  - Initial testing
- Phase 1: Small scale
  - <30 people
  - Goal: Finding a minimum safe dosage for the drug
- Phase 2: Medium scale
  - <100 people
  - Goal: Assessing initial benefit for the drug
- Phase 3: Large scale
  - 100-1000 people
  - Goal: Assessing benefit relative to current standard of care

Pre-Clinical
Preformulation
Synthesis
Toxicology
Initial Formulation
Method Development & Validation
Stability & Degradation Studies
IND
Investigational New Drug Application
Clinical
Formulation Modification
Final Formulation
Method Development & Validation
Stability & Degradation Studies
Extractables/ Leachables
Manufacturer Validation
NDA
New Drug Application
After-Market
Production Risk Mitigation
Counterfeit Analysis
Contamination ID
Failure Analysis
Post Approval Changes
Litigation Support

Controls

Controls

- Controls are the *alternative* that is being considered to the treatment that is proposed
- No treatment
  - Common in new surgical procedures
- Placebo
  - Duplicate the experience of the intervention without its effect
- Standard of care
  - Comparison to current clinical practice – usually in Phase 3

Blinding

What is blinding?

- Blinding:
  - Hiding the identity of who received treatment and who received the control
  - Prevent unintentional bias
- **Discussion: Why do we need to have blinded trials?**

Types of blinding

- Open label
  - Participants and doctors know if they are on treatment/control
- Single
  - Doctors know but participants do not know if they are on treatment/control
- Double
  - Conceal nature of treatment from participants and doctors
  - Only those directing the study know
- Triple
  - Conceal nature of treatment from participants, researchers and administrators/doctors

Challenges in designing and running RCTs

- Cohort selection
  - Need consent and ways to find people who meet the criteria for being part of an RCT
  - Why is it hard: Healthcare infrastructure is disperse
- Cost
  - Multiple stages of the RCT are costly
  - Why is it hard: Require large groups of patients, administrators and coordination between organizations (drug companies and hospitals)
- Time
  - The end-to-end pipeline for drug development takes a *long* time

Discussion: Can we use machine learning to help? If so, where and how?

Machine learning to evaluate eligibility criteria

- Evaluating eligibility criteria of oncology trials using real-world data and AI, Liu et. al, Nature Medicine 2020
- Problem being tackled:
  - Recruitment for trials is challenging
  - Trials can have low enrolment (86% of trials fail to meet recruitment within time)
  - **Why**: Eligibility criteria can be too strict but not all clinicians agree on how to relax them
- **Key idea:** Can we use real-world electronic health record data to emulate clinical trials and use the data to *relax* eligibility criteria.

Dataset

- [a] Flatiron Health EHR-derived database (private database)
  - De-identified data from 280 cancer clinics
  - Cohort: [aSCLC] Advanced small cell lung carcinomas [~61K patients)
- [b] ClinicalTrials.gov – contains information on trials and their associated eligibility criteria
- Took criteria from [b], encoded them as rules in [a] and selected patients who *would have met the criteria for 10 aSCLC trials*

Table 1 | Comparisons of eligibility criteria

Trial name	Original trial criteria		
	No. of criteria	No. of patients	HR
FLAURA	10	2,277	0.81
LUX8	11	129	0.65
Checkmate017	17	523	0.67
Checkmate057	19	792	0.75
Checkmate078	18	1,509	0.74
Keynote010	13	806	0.56
Keynote189	15	4,066	0.88
Keynote407	13	2,031	1.13
BEYOND	12	2,902	1.09
OAK	19	493	0.88
Average	15	1,553	0.82

Study finds that hazard ratios for certain treatments comparable to study criteria in the entire patient cohort

Suggests that there is potential benefit to developing relaxed study design protocols

Pipeline in a slide

a Trial emulation

(1) Implementation and encoding of eligibility criteria

Rule description
Be ≥18 years of age on the day of signing informed consent.

↓

Encoded rule
#inclusion
records[‘diagnosisdate’] >=
demographics[‘birthyear’] + @YEARS(18)

(2) Generation of treatment groups

RWD

	Sex	Drug
1	F	A
2	M	C
3	F	A
4	M	B
5	F	D
6	F	B
7	M	A
8	F	B
…	…	…

Treated arm: drug A

	Sex	Drug
1	F	A
7	M	A
…	…	…

Control arm: drug B

	Sex	Drug
4	M	B
8	F	B
…	…	…

Receive drug A
Receive drug B
Receive drug C or D
Do not meet the eligibility criteria

(3) Emulate blind assignment

Use IPTW to adjust for confounding factors:

- Demographics: age, gender, race, smoking status, and so on.
- Status: group stage, ECOG, and so on.
- Biomarkers: ALK, PDL1, and so on.
- Other factors

(4) Survival analysis

HR = 0.70

b Analysis

Change eligibility criteria in silico:

Input	Output
Criteria: rule 1 + rule 2	*n* = 1,000 patients; HR = 0.70
Criteria: rule 2	*n* = 3,000 patients; HR = 0.80

→

Influence of each rule

Contribution of one rule:
change in the number of patients;
change in HR

→

Criteria relaxation

The set of rules:
large cohort; low HR

We’ll learn about SHAPley values later this month

Key insights - (1)

Table 1 | Comparisons of eligibility criteria

Trial name	Original trial criteria			Fully relaxed criteria		Data-driven criteria		
	No. of criteria	No. of patients	HR	No. of patients	HR	No. of criteria	No. of patients	HR
FLAURA	10	2,277	0.81	3,819	0.82	4	2,546	0.75
LUX8	11	129	0.65	1,350	0.81	5	141	0.58
Checkmate017	17	523	0.67	4,900	0.71	7	4,085	0.71
Checkmate057	19	792	0.75	4,900	0.71	9	2,594	0.66
Checkmate078	18	1,509	0.74	4,900	0.71	9	3,348	0.68
Keynote010	13	806	0.56	1,950	0.51	1	1,948	0.51
Keynote189	15	4,066	0.88	8,818	0.94	7	4,595	0.85
Keynote407	13	2,031	1.13	10,437	1.07	4	9,173	1.04
BEYOND	12	2,902	1.09	9,310	1.14	4	3,043	1.08
OAK	19	493	0.88	1,288	0.87	6	620	0.80
Average	15	1,553	0.82	5,167	0.83	6	3,209	0.77

The number of inclusion/exclusion criteria, the number of eligible patients and the hazard ratio of the overall survival of emulated aNSCLC trials with eligibility criteria under three scenarios: the original criteria used in the trial, fully relaxed criteria and data-driven criteria. The fully relaxed criteria correspond to evaluating the hazard ratio of the overall survival of all of the patients in the Flatiron database who took the treatments in the relevant line of therapy. The data-driven criteria were selected by Shapley values. HR, hazard ratio.

Key insights – (2)

- More inclusive trial resulted in more women being selected into the trial cohort
- Expanded data-driven eligibility included older individuals (without limiting comorbidities being selected into the trial)

Summary

- Hearsay: “Only ~10-15% of interventions in surgical specialties are backed by evidence from randomized control trials”
- Using observational data holds a lot of promise for the design of effective mechanisms
- But there is enormous debate in the community:
  - The Magic of Randomization versus the Myth of Real-World Evidence, Collins et. al, 2020

Part II - Beyond Randomization

Outline – A taste of causal inference

- Recap: Randomized control trials
- Making decisions with predictive models
- Potential outcomes
- Do operator
- Assumptions in causal inference
- Algorithms for estimating ATE and CATE

Last week

Treatment

Control

$$ATE = \sum_{i \in \mathcal{T}} y_t - \sum_{i \in \mathcal{C}} y_c \approxeq \mathbb{E}[Y_t - Y_c]$$

Y is an outcome of interest (positive value good; negative value bad)

Risk stratification

- We studied how supervised machine learning could be used for the early detection of diabetes
- There were several features that were associated with an increased/decreased risk of diabetic onset
- Gastric bypass was the highest-negative weight
  - **Is this a good idea for an intervention?**

Survival analysis

- We can use survival models to predict time of dying/progression of disease in the future
- Trained a CoxPH model and feature associated with high creatinine increased survival time
  - Should we conclude that increasing creatinine improved survival?

Challenges in making decisions from healthcare data

- Doing an RCT is not always ethical
  - e.g. cannot force a control group to smoke to assess its effects on long-cancer
- Can we make causal conclusions from *retrospective* data?
- Key challenges:
  - Unobserved confounding
  - Positive support

What should the graph of observational health data look like?

A simple causal graph for observational data

Doctor might prescribe Metformin if A1C level is high

Patient could have a high BMI making them more likely to suffer from heart failure.

Metformin has an effect on an outcome like death

Causal inference with the Potential Outcomes framework

- Each individual x_i has two potential outcomes:
  - Control outcome [PO had the individual not been treated] $Y_0(x_i)$
  - Treated outcome [PO had the individual been treated] $Y_1(x_i)$
- Conditional average treatment effect [CATE]

$$\mathrm{CATE}(x_i) = \mathbb{E}_{Y_1 \sim p(Y_1|x_i)}[Y_1|x_i] - \mathbb{E}_{Y_0 \sim p(Y_0|x_i)}[Y_0|x_i]$$

- Average treatment effect [ATE]

$$\mathrm{ATE} = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}_{x \sim p(x)}[\mathrm{CATE}(x)]$$

Visualizing notation on potential outcomes

Example – Blood pressure and age

Slide credit: David Sontag

Blood pressure and age

Blood pressure and age

Blood pressure and age

Blood pressure and age

(age, gender, exercise)	Y_0: Sugar levels *had they received* medication A	Y_1: Sugar levels *had they received* medication B	Observed sugar levels
(45, F, 0)	**6**	5.5	6
(45, F, 1)	7	**6.5**	6.5
(55, M, 0)	**7**	6	7
(55, M, 1)	9	**8**	8
(65, F, 0)	8.5	**8**	8
(65,F, 1)	**7.5**	7	7.5
(75,M, 0)	10	**9**	9
(75,M, 1)	**8**	7	8

(Example from Uri Shalit)

For an individual we only observe one of the potential outcomes

- *Fundamental problem of causal inference* (Rubin 1974; Holland 1986)
- ~~*How*~~*When* can we make causal conclusions despite this?

Learning from retrospective data

Learning from interventional data [RCTs]

The do-operator

- The do-operator (Pearl, 2000) is a graphical operator on a causal graph that characterizes the effect of an intervention
- It allows us to mimic the effect of an intervention onto variation in the joint probability distribution

Causal inference

- **What we have:** Data is drawn from the joint distribution on the left
- **What we want:** Samples from the joint distributions on the right
- **Key idea:** Under certain assumptions, we can estimate conditional distributions from the right
- **Strategy:** Write down the causal estimand using quantities estimable from the observed (left) distribution)

Assumptions in causal inference – (1)

Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

T: treatment assignment

We assume:

$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

Also known as Ignorability

The potential outcomes are independent of treatment assignment, conditioned on covariates x

Ignorability

$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

No Ignorability

$$(Y_0, Y_1) \not\perp\!\!\!\perp T \mid x$$

Assumptions in causal inference – (2)

Y_0, Y_1: potential outcomes for control and treated

x: unit covariates (features)

T: treatment assignment

We assume:

$$p(T = t | X = x) > 0 \; \forall t, x$$

Also known as Common Support

Before beginning any causal analysis

- Understand where the data is coming from (are there biases you didn't account for)
- Drawing the causal graph (or even approximations to it) can be a crucial exercise
- What should X, T, Y be? Do they satisfy ignorability and positivity?
- What is the causal query of interest?

Adjustment formula (Hernan and Robins, 2010) (Pearl, 2009)

- Also known as the G-formula, it provides a mechanism to write down a causal estimand using observational data

Average Treatment Effect

The expected causal effect of T on Y:

$$ATE := \mathbb{E}\left[Y_1 - Y_0\right]$$

Average Treatment Effect

The expected causal effect of T on Y:

$$ATE := \mathbb{E}\left[Y_1 - Y_0\right]$$

$$\mathbb{E}\left[Y_1\right] =$$

law of total expectation

$$\mathbb{E}_{x \sim p(x)}\left[\mathbb{E}_{Y_1 \sim p(Y_1|x)}\left[Y_1|x\right]\right] =$$

Average Treatment Effect

The expected causal effect of T on Y:

$$ATE := \mathbb{E}\left[Y_1 - Y_0\right]$$

$$\mathbb{E}\left[Y_1\right] =$$

$$\mathbb{E}_{x \sim p(x)}\left[\mathbb{E}_{Y_1 \sim p(Y_1|x)}\left[Y_1|x\right]\right] =$$

ignorability

$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$

$$\mathbb{E}_{x \sim p(x)}\left[\mathbb{E}_{Y_1 \sim p(Y_1|x)}\left[Y_1|x, T = 1\right]\right] =$$

Average Treatment Effect

The expected causal effect of T on Y:

$$ATE := \mathbb{E}\left[Y_1 - Y_0\right]$$

$$\mathbb{E}\left[Y_1\right] =$$

$$\mathbb{E}_{x \sim p(x)}\left[\mathbb{E}_{Y_1 \sim p(Y_1|x)}\left[Y_1|x\right]\right] =$$

$$\mathbb{E}_{x \sim p(x)}\left[\mathbb{E}_{Y_1 \sim p(Y_1|x)}\left[Y_1|x, T=1\right]\right] =$$

$$\mathbb{E}_{x \sim p(x)}\left[\mathbb{E}\left[Y_1|x, T=1\right]\right]$$

shorter notation

Average Treatment Effect

The expected causal effect of T on Y:

$$ATE := \mathbb{E}\left[Y_1 - Y_0\right]$$

$$\mathbb{E}\left[Y_0\right] =$$

$$\mathbb{E}_{x \sim p(x)}\left[\mathbb{E}_{Y_0 \sim p(Y_0|x)}\left[Y_0|x\right]\right] =$$

$$\mathbb{E}_{x \sim p(x)}\left[\mathbb{E}_{Y_0 \sim p(Y_0|x)}\left[Y_0|x, T = 1\right]\right] =$$

$$\mathbb{E}_{x \sim p(x)}\left[\mathbb{E}\left[Y_0|x, T = 0\right]\right]$$

The adjustment formula

Under the assumption of ignorability, we have that:

$$ATE = \mathbb{E}\left[Y_1 - Y_0\right] =$$

$$\mathbb{E}_{x \sim p(x)}\left[\ \mathbb{E}\left[Y_1 | x, T = 1\right] - \mathbb{E}\left[Y_0 | x, T = 0\right]\ \right]$$

$$\left.\begin{array}{l} \mathbb{E}\left[Y_1 | x, T = 1\right] \\ \mathbb{E}\left[Y_0 | x, T = 0\right] \end{array}\right\}$$

Quantities we can estimate from data

The adjustment formula

Under the assumption of ignorability, we have that:

$$ATE = \mathbb{E}\left[Y_1 - Y_0\right] =$$

$$\mathbb{E}_{x \sim p(x)}\left[\ \mathbb{E}\left[Y_1 | x, T = 1\right] - \mathbb{E}\left[Y_0 | x, T = 0\right]\ \right]$$

$$\left.\begin{array}{l} \mathbb{E}\left[Y_0 | x, T = 1\right] \\ \mathbb{E}\left[Y_1 | x, T = 0\right] \\ \mathbb{E}\left[Y_0 | x\right] \\ \mathbb{E}\left[Y_1 | x\right] \end{array}\right\}$$

Quantities we *cannot* directly estimate from data

The adjustment formula

Under the assumption of ignorability, we have that:

$$ATE = \mathbb{E}\left[Y_1 - Y_0\right] =$$

$$\mathbb{E}_{x \sim p(x)}\left[\ \underbrace{\mathbb{E}\left[Y_1 | x, T = 1\right] - \mathbb{E}\left[Y_0 | x, T = 0\right]}\ \right]$$

$$\left.\begin{array}{l} \mathbb{E}\left[Y_1 | x, T = 1\right] \\ \mathbb{E}\left[Y_0 | x, T = 0\right] \end{array}\right\}$$

Quantities we can estimate from data

Empirically we have samples from $p(x|T = 1)$ or $p(x|T = 0)$.
Extrapolate to $p(x)$

Strategies for Adjustment

- Covariate Adjustment
  - Response surface modeling
  - Use a parametric model of treatments, confounders and outcome

Covariate adjustment (parametric g-formula)

- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y:
  $\mathbb{E}_{x \sim p(x)}\left[\, \mathbb{E}[Y_1 | T = 1, x] - \mathbb{E}[Y_0 | T = 0, x] \right]$
- Fit a model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^{n} f(x_i, 1) - f(x_i, 0)$$

Covariate adjustment (parametric g-formula)

- Explicitly model the relationship between treatment, confounders, and outcome
- Under ignorability, the expected causal effect of T on Y:
  $\mathbb{E}_{x \sim p(x)}\left[\, \mathbb{E}[Y_1 | T = 1, x] - \mathbb{E}[Y_0 | T = 0, x] \right]$
- Fit a model $f(x, t) \approx \mathbb{E}[Y_t | T = t, x]$

$$\widehat{CATE}(x_i) = f(x_i, 1) - f(x_i, 0)$$

Covariate adjustment with linear models

- Assume that:

Blood pressure age medication

$$Y_t(x) = \beta x + \gamma \cdot t + \epsilon_t$$
$$\mathbb{E}[\epsilon_t] = 0$$

$$ATE := \mathbb{E}_{p(x)}[CATE(x)] = \gamma$$

- For causal inference, need to estimate γ well, not $Y_t(x)$ **- Identification, not prediction**
- *Major difference between ML and statistics*

What happens if true model is not linear?

- True data generating process, $x \in \mathbb{R}$:

$$Y_t(x) = \beta x + \gamma \cdot t + \delta \cdot x^2$$

$$ATE = \mathbb{E}[Y_1 - Y_0] = \gamma$$

- Hypothesized model:

$$\widehat{Y}_t(x) = \hat{\beta} x + \hat{\gamma} \cdot t$$

$$\hat{\gamma} = \gamma + \delta \frac{\mathbb{E}[xt]\mathbb{E}[x^2] - \mathbb{E}[t^2]\mathbb{E}[x^2 t]}{\mathbb{E}[xt]^2 - \mathbb{E}[x^2]\mathbb{E}[t^2]}$$

Depending on δ, can be made to be arbitrarily large or small!

Covariate adjustment with non-linear models

- Random forests and Bayesian trees
  Hill (2011), Athey & Imbens (2015), Wager & Athey (2015)
- Gaussian processes
  Hoyer et al. (2009), Zigler et al. (2012)
- Neural networks
  Beck et al. (2000), Johansson et al. (2016), Shalit et al. (2016), Lopez-Paz et al. (2016)

Matching

- Find each individual's nearest counterfactual twin and use their outcome as a proxy for the individual's counterfactual

Identical to covariate adjustment when 1NN classifier is used as a classifier

Effect estimation with matching

$$\forall i \ \mathrm{cfac(i)} = \mathrm{argmin}_{j;\ t_i \neq t_j} d(x_j, x_i)$$

$$\mathrm{CATE}(x_i) = \mathbb{I}[t_i == 1](y_i - y_{\mathrm{cfac(i)}}) + \mathbb{I}[t_i == 0](y_{\mathrm{cfac(i)}} - y_i)$$

$$\mathrm{ATE} = \frac{1}{n} \sum_i \mathrm{CATE}(x_i)$$

Interpretable!

Non-parametric

Sensitive to the choice of metric d

Suffers from all the limitations of K-Nearest Neighbors

Propensity scores

- Reweight samples to turn an observational study into a pseudo-randomized trial

Find w_0, w_1

Propensity score (algorithm)

How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \dots, (x_n, t_n, y_n)$

1. Use any ML method to estimate $\hat{p}(T = t|x)$
2. $$A\hat{T}E = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Propensity score (for an RCT)

How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

Sum over $\sim \frac{n}{2}$ terms

1. Randomized trial $p = 0.5$

2. $$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{0.5} = \frac{2}{n} \sum_{i \text{ s.t. } t_i=1} y_i - \frac{2}{n} \sum_{i \text{ s.t. } t_i=0} y_i$$

Propensity score - derivation

- We want: $\mathbb{E}_{x \sim p(x)}[Y_1(x)]$

- We know that:

$$p(x|T=1) \cdot \frac{p(T=1)}{p(T=1|x)} = p(x)$$

- Thus:

$$\mathbb{E}_{x \sim p(x|T=1)} \left[\frac{p(T=1)}{p(T=1 \mid x)} Y_1(x) \right] = \mathbb{E}_{x \sim p(x)}[Y_1(x)]$$

- We can approximate this empirically as:

$$\frac{1}{n_1} \sum_{i \text{ s.t.} t_i = 1} \left[\frac{n_1/n}{\hat{p}(t_i = 1 \mid x_i)} y_i \right] = \frac{1}{n} \sum_{i \text{ s.t.} t_i = 1} \frac{y_i}{\hat{p}(t_i = 1 \mid x_i)}$$

(similarly for t_i=0)

Propensity score - challenges

- If not much overlap, scores become non-informative
  - Think about what happens if $P(T=1|X) = 1$ or 0 vs 0.5
- When propensity scores are small, the estimators have a large variance

Causal inference is a big field

- Causal inference has been studied in many communities of science including economics, statistics, machine learning
- Different schools of thought – causal graphs vs conditional independence statements
- Beyond the scope to have an in-depth discussion of the all techniques underlying this field in this class
- Lots of research in leveraging ideas from causal inference to improve predictive models

Causal inference - Overview

- The last two lectures give us an overview of two techniques to estimate causal effects:
  - Assumptions necessary for causal inference
    - Positivity
    - Common support
    - No unobserved confounding
  - Covariate adjustment
  - Propensity score matching
  - **Key idea:** Using assumptions to write down estimators of causal effects using observational data
- Many more extensions to more complex regimes where data are time-varying