

---

# Causal Inference

a summary

---

# Contents

<b>1</b>	<b>General</b>	<b>3</b>	2.2 G-Methods . . . . .	4
<b>2</b>	<b>Models</b>	<b>4</b>	2.2.1 Time-varying A . . . . .	5
2.1	Traditional Methods . . . . .	4	<b>3</b>	<b>Longitudinal Data</b>
				<b>6</b>

# 1 General

## Causal Roadmap (Petersen and van der Laan, 2014)

systematic approach linking causality to statistical procedures

**1. Specifying Knowledge.** structural causal model (unifying counterfactual language, structural equations, & causal graphs): a set of possible data-generating processes, expresses background knowledge and its limits

**2. Linking Data.** specifying measured variables and sampling specifics (latter can be incorporated into the model)

**3. Specifying Target.** define hypothetical experiment: decide

1. variables to intervene on: one (point treatment), multiple (longitudinal, censoring/missing, (in)direct effects)
2. intervention scheme: static, dynamic, stochastic
3. counterfactual summary of interest: absolute or relative, marginal structural models, interaction, effect modification
4. population of interest: whole, subset, different population

**4. Assessing Identifiability.** are knowledge and data sufficient to derive estimand and if not, what else is needed?

**5. Select Estimand.** current best answer: knowledge-based assumptions + which minimal convenience-based assumptions (transparency) gets as close as possible

**6. Estimate.** choose estimator by statistical properties, nothing causal here

**7. Interpret.** hierarchy: statistical, counterfactual, feasible intervention, randomized trial

## Average Causal Effect $E[Y^{a=1}] \neq E[Y^{a=0}]$

$$E[Y^a] = \sum_y yp_{Y^a}(y) \quad (\text{discrete})$$

$$= \int yf_{Y^a}(y)dy \quad (\text{continuous})$$

individual causal effect  $Y_i^{a=1} \neq Y_i^{a=0}$  generally unidentifiable

*null hypothesis:* no average causal effect

*sharp null hypothesis:* no causal effect for any individual

**notation**  $A, Y$ : random variables (differ for individuals);  $a, y$ : particular values; counterfactual  $Y^{a=1}$ :  $Y$  under treatment  $a = 1$

**stable unit treatment value assumption (SUTVA)**  $Y_i^a$  is well-defined: no interference between individuals, no multiple versions of treatment (weaker: treatment variation irrelevance)

**causal effect measures** typically based on means

*risk difference:*  $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$

*risk ratio:*  $\frac{\Pr[Y^{a=1}=1]}{\Pr[Y^{a=0}=1]}$

*odds ratio:*  $\frac{\Pr[Y^{a=1}=1]/\Pr[Y^{a=1}=0]}{\Pr[Y^{a=0}=1]/\Pr[Y^{a=0}=0]}$

*number needed to treat (NNT)* to save 1 life:  $-1/\text{risk difference}$

**sources of random error:** sampling variability (use consistent estimators), nondeterministic counterfactuals

**association** compares  $E[Y|A = 1]$  and  $E[Y|A = 0]$ , **causation** compares  $E[Y^{a=1}]$  and  $E[Y^{a=0}]$  (whole population)

**randomized experiments (target trial)** 2.1 and 2.2; 3.6; maybe 22

**identifiability conditions** most of 3

positivity: p. 155, p. 162

additional conditions: chapter 13.5

exchangeability: p 172f

positivity:  $f_L(l) \neq 0 \Rightarrow f_{A|L}(a|l) > 0 \forall a, l$

consistency: if  $A = a$ , then  $Y^a = Y$  for each individual  $Y = Y^A$

**effect modification** chapter 4

**interaction** chapter 5

**causal diagrams** chapter 6, include swigs from 7.5 and that one technical point  
more on SWIGS p 242ff

**confounding** chapter 7

**selection bias** chapter 8

**measurement bias** chapter 9

**random variability** chapter 10

## 2 Models

**Modeling** data are a sample from the target population

*estimand*: quantity of interest, e. g.  $E[Y|A=a]$   
*estimator*: function to use, e. g.  $\hat{E}[Y|A=a]$   
*estimate*: apply function to data, e. g. 4.1

**model**: a priori restriction of joint distribution/dose-response curve; *assumption*: no model misspecification (usually wrong)

**non-parametric estimator**: no restriction (saturated model) = *Fisher consistent estimator* (entire population data  $\rightarrow$  true value)

**parsimonious model**: few parameters estimate many quantities

**bias-variance trade-off**:

wiggleness  $\uparrow \rightarrow$  misspecification bias  $\downarrow$ , CI width  $\uparrow$

### 2.1 Traditional Methods

**Outcome regression** chapter 15

**instrumental variable estimation** chapter 16

**causal survival analysis** chapter 17

**Variable Selection** can induce bias if  $L$  includes:

(descendant of) collider: *selection bias under the null*  
 noncollider effect of  $A$ : *selection bias under the alternative*  
 mediator: *overadjustment for mediators*

temporal ordering is not enough to conclude anything

**bias amplification**: e.g. by adjusting for an instrument  $Z$  (can also reduce bias)

**Machine Learning**  $L$  is high-dimensional

use lasso or ML for IP weighting/standardization

**but**: ML does not guarantee elimination of confounding and has largely unknown statistical properties

$\rightarrow$  **doubly robust estimator**: consistent if bias  $< \frac{1}{\sqrt{n}}$

*sample splitting*: train estimators on training sample, use resulting estimators for doubly robust method on estimation sample (CIs on estimation sample are valid, but  $n$  halved)

*cross-fitting*: do again the other way round, average the two estimates, get CI via bootstrapping

**problems**: unclear choice of algorithm, is bias small enough?

### 2.2 G-Methods

**G-Methods** generalized treatment contrasts: adjust for (surrogate) confounders  $L$

- **standardization** two types of g-formula
- **IP weighting** also g-formula
- **g-estimation**: not needed unless longitudinal

**Standardization** plug-in (or parametric if so) g-formula

$$E[Y^a] = \overbrace{E[E[Y|A=a, L=l]]}^{\text{conditional expectation}} = \overbrace{\int E[Y|L=l, A=a] f_L[l] dl}^{\text{joint density estimator}}$$

weighted average of stratum-specific risks; unknowns can be estimated non-parametrically or modeled

**no need to estimate  $f_L[l]$ /integrate** as empirical distribution can be used: estimate outcome model  $\rightarrow$  predict counterfactuals on whole dataset  $\rightarrow$  average the results ( $\rightarrow$  CI by bootstrapping)  
**for discrete  $L$**   $E[Y|A=a] = \sum_l E[Y|L=l, A=a] \Pr[L=l]$

**time-varying** standardize over all possible  $\bar{L}$ -histories  
 simulates joint distribution of counterfactuals  $(Y^{\bar{a}}, \bar{L}^{\bar{a}})$  for  $\bar{a}$   
**joint density estimator (jde)**

$$\text{discrete: } E[Y^{\bar{a}}] = \sum_{\bar{l}} E[Y|\bar{A}=\bar{a}, \bar{L}=\bar{l}] \prod_{k=0}^K f(l_k|\bar{a}_{k-1}, \bar{l}_{k-1})$$

$$\text{continuous: } \int f(y|\bar{a}, \bar{l}) \prod_{k=0}^K f(l_k|\bar{a}_{k-1}, \bar{l}_{k-1}) dl$$

for *stochastic strategies* multiply with  $\prod_{k=0}^K f^{int}(a_k|\bar{a}_{k-1}, \bar{l}_k)$

**iterated conditional expectation (ice)**

$$E[Y_T^{\bar{a}}] = E[E[E[...E[Y_T|\bar{A}_{T-1}=\bar{a}_{T-1}, \bar{L}_T]...\bar{A}_0=a_0, L_1]|L_0]]$$

**g-null paradox**

Proof: for  $L_0 \rightarrow A_0 \rightarrow Y_0 \rightarrow L_1 \rightarrow A_1 \rightarrow Y_1$ ,  $\bar{a} = (a_0, a_1)$

$$E[Y_1^{\bar{a}}] \stackrel{\text{CE}}{=} E[E[Y_1^{\bar{a}}|A_0=a_0, L_0]]$$

$$(\text{ice}) \stackrel{\text{CE}^*}{=} E[E[E[Y_1|\bar{L}, \bar{A}=\bar{a}, Y_0]|A_0=a_0, L_0]]$$

$$\stackrel{\text{LTP}}{=} E\left[\sum_{l_1} E[Y_1|A_0=a_0, \bar{L}, Y_0] \Pr[l_1|a_0, l_0, y_0]\right]$$

$$\stackrel{\text{LTP}}{=} \sum_{l_0} \left[\sum_{l_1} E[Y_1|A_0=a_0, \bar{L}, Y_0] \Pr[l_1|a_0, l_0, y_0]\right] \Pr[l_0]$$

$$(\text{ide}) \stackrel{\text{sum}}{=} \sum_{\bar{l}} E[Y_1|A_0=a_0, \bar{L}, Y_0] \Pr[l_1|a_0, l_0] \Pr[l_0]$$

CE: conditional expectation; \*: exchangeability;

LTP: law of total probability

**Marginal Structural Models** association is causation in the IP weighted pseudo-population

associational model  $E[Y|A] =$  causal model  $E[Y^a]$

*step 1*: estimate/model  $f[A|L]$  (and  $f[A]$ )  $\rightarrow$  get  $(S)W^A$

*step 2*: estimate regression parameters for pseudo-population

**effect modification** variables  $V$  can be included (e.g.

$\beta_0 + \beta_1 a + \beta_2 V a + \beta_3 V$ ; technically not marginal anymore),  
 $SW^A(V) = \frac{f[A|V]}{f[A|L]}$  more efficient than  $SW^A$

**Censoring** measuring joint effect of  $A$  and  $C$

$E[Y^{a,c=0}]$  is of interest

**standardization**  $E[Y|A=a] = \int E[Y|L=l, A=a, C=0] dF_L[l]$

**IP weights**  $W^{A,C} = W^A \times W^C$  (uses  $n$ ) or  
 $SW^{A,C} = SW^A \times SW^C$  (uses  $n^{c=0}$ )

**g-estimation** can only adjust for confounding, not selection bias  
→ use IP weights

**G-Estimation** (additive) structural nested models

$$\text{logit Pr} [A = 1 | H(\psi^\dagger), L] = \alpha_0 + \alpha_1 H(\psi^\dagger) + \alpha_2 L$$

$$H(\psi^\dagger) = Y - \psi_1 A$$

find  $\psi^\dagger$  which renders  $\alpha_1 = 0$ ; 95 %-CI: all  $\psi^\dagger$  for which  $p > 0.05$   
closed-form solution for linear models

**derivation:**  $H(\psi^\dagger) = Y^{a=0}$

$$\text{logit Pr} [A = 1 | Y^{a=0}, L] = \alpha_0 + \alpha_1 Y^{a=0} + \alpha_2 L$$

$Y^{a=0}$  unknown, but because of exchangeability  $\alpha_1$  should be zero

$$Y^{a=0} = Y^a - \psi_1 a$$

equivalent to  $Y^{a=0} = Y^{a=1} - \psi_1$ , but using no counterfactuals

**structural nested mean model**

$$\text{additive: } E[Y^a - Y^{a=0} | A = a, L] = \beta_1 a (+ \beta_2 a L)$$

$$\text{multiplicative: } \log \left( \frac{E[Y^a | A = a, L]}{E[Y^{a=0} | A = a, L]} \right) = \beta_1 a (+ \beta_2 a L)$$

multiplicative is preferred if  $Y$  always positive, but does not extend to longitudinal case

semi-parametric: agnostic about  $\beta_0$  and effect of  $L \rightarrow$  robust  $\uparrow$

**no time-varying:** no nesting; model equals marginal structural models with missing  $\beta_0, \beta_3$  (unspecified “no treatment”)

**sensitivity analysis:** unmeasured confounding ( $\alpha_1 \neq 0$ ) can be examined: do procedure for different values of  $\alpha_1 \rightarrow$  plot  $\alpha_1$  vs.  $\psi^\dagger \rightarrow$  how sensitive is estimate to unmeasured confounding?

**effect modification:** add  $V$  in both g-estimation equations

**doubly robust estimators** exist

**IP Weighting** inverse probability of treatment (g-formula)

$$E[Y^a] = E \left[ \frac{I(A=a)Y}{f[A|L]} \right]; W^A = \frac{1}{f[A|L]}; SW^A = \frac{f(A)}{f[A|L]}$$

unknowns can be estimated non-parametrically or modeled

**pseudo-population:** everyone is treated & untreated ( $L \not\rightarrow A$ )

**FRCISTG** (fully randomized causally interpreted structured

graph): probability tree for  $L \rightarrow A \rightarrow Y$ , can be used to

calculate/visualize simulation of values for  $A$

**for discrete  $A, L$**   $f[a|l] = \Pr[A = a, L = l]$

**estimators:** Horvitz-Thompson; Hajek (modified version)

**stabilized weights  $SW^A$**  should have an average of 1 (check!)

$\rightarrow$  pseudo-population same size  $\rightarrow$  CI width  $\downarrow$

**Standardization and IP Weighting** are equivalent,

**but** if modeled, different “no misspecification” assumptions:

standardization: outcome model

IP weighting: treatment model

**doubly robust estimators:** reduce model misspecification bias, consistent if either model is correct; *e. g.:*

1. fit outcome regression with variable  $R = \begin{cases} +W^A & \text{if } A=1 \\ -W^A & \text{if } A=0 \end{cases}$
2. standardize by averaging

## 2.2.1 Time-varying A

**IP Weighting**

$$W^{\bar{A}} = \prod_{k=0}^K \frac{1}{f(A_k | \bar{A}_{k-1}, \bar{L}_k)}$$

$$SW^{\bar{A}} = \prod_{k=0}^K \frac{f(A_k | \bar{A}_{k-1})}{f(A_k | \bar{A}_{k-1}, \bar{L}_k)}$$

**Doubly Robust Estimator** sequential estimation

1. estimate  $\hat{f}(A_m | \bar{A}_{m-1}, \bar{L}_m)$  (e. g. logistic model), use it to calculate at each time  $m$ :  $\widehat{W}^{\bar{A}}_m = \prod_{k=0}^m \frac{1}{\hat{f}(A_k | \bar{A}_{k-1}, \bar{L}_k)}$  and modified IP weights at  $m$ :  $\widehat{W}^{\bar{A}_{m-1}, a_m} = \frac{\widehat{W}^{\bar{A}_{m-1}}}{\hat{f}(a_m | \bar{A}_{m-1}, \bar{L}_m)}$
2. with  $\widehat{T}_{K+1} := Y$ , recursively for  $m = K, K-1, \dots, 0$ :
  - (a) fit outcome regression on  $\widehat{T}_{m+1}$  with variable  $\widehat{W}^{\bar{A}}_m$
  - (b) calculate  $\widehat{T}_m$  using the outcome model with  $\widehat{W}^{\bar{A}_{m-1}, a_m}$
3. calculate standardized mean outcome  $\widehat{E}[Y^{\bar{a}}] = E[\widehat{T}_0]$

**valid, if** treatment or outcome model correct, or treatment correct until  $k$  and outcome otherwise ( $k+1$  robustness)

**G-Estimation** nested equations: for each time  $k$

**structural nested mean models** separate effect of each  $a_k$

$$E[Y^{\bar{a}_{k-1}, a_k, \bar{a}_{k+1}} - Y^{\bar{a}_{k-1}, \bar{a}_{k+1}} | \bar{L}^{\bar{a}_{k-1}} = \bar{L}_k, \bar{A}_{k-1} = \bar{a}_{k-1}] =$$

$$a_k \gamma_k(\bar{a}_{k-1}, \bar{L}_k, \beta)$$

**calculations**

$$H_k(\psi^\dagger) = Y - \sum_{j=k}^K A_j \gamma_j(\bar{A}_{j-1}, \bar{L}_j, \psi^\dagger)$$

function  $\gamma_j$  can be, e. g. constant ( $\psi_1$ ), time-varying only ( $\psi_1 + \psi_2 k$ ), or dependent on treatment/covariate history

$$\text{logit Pr} [A_k = 1 | H_k(\psi^\dagger), \bar{L}_k, \bar{A}_{k-1}] =$$

$$\alpha_0 + \alpha_1 H_k(\psi^\dagger) + \alpha_2 w_k(\bar{L}_k, \bar{A}_{k-1})$$

find  $\alpha_1$  that is closest to zero

closed form estimator exists for the linear case

**Censoring**  $\bar{C}$ : monotonic type of missing data

**standardization:**

$$\int f(y | \bar{a}, \bar{c} = \bar{0}, \bar{l}) \prod_{k=0}^K dF(l_k | \bar{a}_{k-1}, c_{k-1} = 0, \bar{l}_{k-1})$$

**IP weighting:**

$$SW^{\bar{C}} = \prod_{k=1}^{K+1} \frac{1 \cdot \Pr(C_k = 0 | \bar{A}_{k-1}, C_{k-1} = 0)}{\Pr(C_k = 0 | \bar{A}_{k-1}, C_{k-1} = 0, \bar{L}_k)}$$

### 3 Longitudinal Data

**Time-Varying Treatments** compare 2 treatments

treatment history up to  $k$ :  $\bar{A}_k = (A_0, A_1, \dots, A_k)$

shorthand: always treated  $\bar{A} = \bar{1}$ , never treated  $\bar{A} = \bar{0}$

**static strategy:**  $g = [g_0(\bar{a}_{-1}), \dots, g_K(\bar{a}_{K-1})]$

**dynamic strategy:**  $g = [g_0(\bar{l}_0), \dots, g_K(\bar{l}_K)]$

**stochastic strategy:** non-deterministic  $g$

optimal strategy is where  $E[Y^g]$  is maximized (if high is good)

**Sequential Identifiability** sequential versions of

**exchangeability:**  $Y^g \perp\!\!\!\perp A_k | \bar{A}_{k-1} \quad \forall g, k = 0, 1, \dots, K$

*conditional exchangeability:*

$$(Y^g, L_{k+1}^g) \perp\!\!\!\perp A_k | \bar{A}_{k-1} = g(\bar{L}_k), \bar{L}^k \quad \forall g, k = 0, 1, \dots, K$$

**positivity:**  $f_{\bar{A}_{k-1}, \bar{L}_k}(\bar{a}_{k-1}, \bar{l}_k) \neq 0 \Rightarrow$

$$f_{A_k | \bar{A}_{k-1}, \bar{L}_k}(a_k | \bar{a}_{k-1}, \bar{l}_k) > 0 \quad \forall (\bar{a}_{k-1}, \bar{l}_k)$$

**consistency:**

$$Y^{\bar{a}} = Y^{\bar{a}^*} \quad \text{if } \bar{a} = \bar{a}^*; \quad Y^{\bar{a}} = Y \quad \text{if } \bar{A} = \bar{a};$$

$$\bar{L}_k^{\bar{a}} = \bar{L}_k^{\bar{a}^*} \quad \text{if } \bar{a}_{k-1} = \bar{a}_{k-1}^*; \quad \bar{L}_k^{\bar{a}} = \bar{L}_k \quad \text{if } \bar{A}_{k-1} = \bar{a}_{k-1}$$

**generalized backdoor criterion** (static strategy): all backdoors into  $A_k$  (except through future treatments) are blocked  $\forall k$

**static sequential exchangeability for  $Y^{\bar{a}}$**

$$Y^{\bar{a}} \perp\!\!\!\perp A_k | \bar{A}_{k-1}, \bar{L}_k \quad \text{for } k = 0, 1, \dots, K$$

use SWIGs to visually check d-separation

**time-varying confounding**  $E[Y^{\bar{a}} | L_0] \neq E[Y | A = \bar{a}, L_0]$

**Treatment-Confounder Feedback**  $A_0 \rightarrow L_1 \rightarrow A_1$ :

an unmeasured  $U$  influencing  $L_1$  and  $Y$  turns  $L_1$  into a collider;

traditional adjustment (e.g. stratification) biased: use g-methods

**g-null test** sequential exchangeability & sharp null true  $\Rightarrow$

$$Y^g = Y \quad \forall g \quad \Rightarrow \quad Y \perp\!\!\!\perp A_0 | L_0 \quad \& \quad Y \perp\!\!\!\perp A_1 | A_0, L_0, L_1; \text{ therefore:}$$

if last two independences don't hold, one assumption is violated

**g-null theorem:**  $E[Y^g] = E[Y]$ , if the two independences hold ( $\Rightarrow$  sharp null: only if strong faithfulness (no effect cancelling))

**target trial** chapter 22 (does that even really fit in here, maybe push to 3rd paragraph in without models)

# References

*If no citation is given, the source is (?)*

Petersen, M. L. and van der Laan, M. J. (2014). Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology (Cambridge, Mass.)*, 25(3):418–426.

