Causal Inference
a summary

Contents

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 asspumptions (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 $\mathrm{E}\left[Y^{a=1}\right] \neq \mathrm{E}\left[Y^{a=0}\right]$

$$\mathrm{E}\left[Y^a\right] = \sum_y y p_{Y^a}(y) \qquad \text{(discrete)}$$
$$= \int y f_{Y^a}(y) dy \qquad \text{(continuous)}$$

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: $\mathrm{Pr}\left[Y^{a=1} = 1\right] - \mathrm{Pr}\left[Y^{a=0} = 1\right]$
   risk ratio: $\frac{\mathrm{Pr}\left[Y^{a=1}=1\right]}{\mathrm{Pr}\left[Y^{a=0}=1\right]}$
   odds ratio: $\frac{\mathrm{Pr}\left[Y^{a=1}=1\right]/\mathrm{Pr}\left[Y^{a=1}=0\right]}{\mathrm{Pr}\left[Y^{a=0}=1\right]/\mathrm{Pr}\left[Y^{a=0}=0\right]}$
number needed to treat (NNT) to save 1 life: $-1/$risk difference
sources of random error: sampling variability (use consistent estimators), nondeterministic counterfactuals
association compares $E\left[Y|A = 1\right]$ and $E\left[Y|A = 0\right]$, **causation** compares $E\left[Y^{a=1}\right]$ and $E\left[Y^{a=0}\right]$ (whole population)

Target Trial emulating an ideal randomized experiment
explicitly formulate target trial & show how it is emulated → less vague causal question, helps spot issues
missing data problem unknown counterfactuals
randomized experiments: missing completely at random → exchangeability (= exogeneity as treatment is exogenous)
ideal randomized experiment: no censoring, double-blind, well-defined treatment, & adherence → association is causation
pragmatic trial: no placebo/blindness, realistic monitoring

PICO (population, intervention, comparator, outcome): some components of target trial
three types of causal effects:
   intention-to-treat effect (effect of treatment assignment)
   per-protocol effect (usually dynamic when toxicity arises)
   other intervention effect (strategy changed during follow-up)
controlled direct effects: effect of A on Y not through B
   natural direct effect A on Y if $B^{a=0}$ (cross-world quantity)
   principal stratum effect A on Y for subset with $B^{a=0} = B^{a=1}$
crossover experiment: sequential treatment & outcome $t=0, 1$
individual causal effect $Y_{it}^{a_t=1} - Y_{it}^{a_t=0}$ only identifiable if: no carryover effect, effect ⊥⊥ time, outcome ⊥⊥ time
time zero if eligibility at multiple t (observational data): earliest, random t, all t (adjust variance with bootstrapping)
grace periods: usually treatment starts x months after first eligible, if death before: randomly assign strategy/copy into both

Identifiability Conditions hold in ideal experiments
consistency counterfactuals correspond to data $Y = Y^A$:
if $A = a$, then $Y^a = Y$ for each individual
 • precise definition of Y^a via specifying a (sufficiently well-defined a maybe impossible (effect of DNA before it was discovered), relies on expert consensus)
 • linkage of counterfactuals to data (a must be seen in data)
positivity $\mathrm{Pr}\left[A = a|L = l\right] > 0 \ \forall l$ with $\mathrm{Pr}\left[L = l\right] > 0$;
$$f_L(l) \neq 0 \Rightarrow f_{A|L}(a|l) > 0 \ \forall a, l$$
 • structural violations (inference not on full population)
 • random variability (smooth over with parametric models)
can sometimes be empirically verified (if all is seen in data)
exchangeability unverifiable without randomization
 • *marginal:* $Y^a \perp\!\!\!\perp A \ \widehat{=}$ randomized experiment, counterfactuals are missing completely at random (MCAR)
 • *conditional:* $Y^a \perp\!\!\!\perp A|L \ \widehat{=}$ conditionally randomized, counterfactuals are missing at random (MAR)
alternative definition: $\mathrm{Pr}\left[A = 1|Y^{a=0}, L\right] = \mathrm{Pr}\left[A = 1|L\right]$
additional conditions:
correct measurement mismeasurement of A, Y, L results in bias
correct model specification models $\overset{\text{may}}{\to}$ misspecification bias

Effect Modification A on Y varies across levels of V
null average causal effect ≠ null causal effect per subgroup
population characteristics: causal effect measure is actually "effect in a population with a particular mix of effect modifiers"
transportability: extrapolation of effect to another population (issues: effect modification, versions of treatment, interference)
effects conditional on V may be more transportable
types: additive/multiplicative scale, qualitative (effect in opposite directions)/quantitative, surrogate/causal
calculation:
 • *stratify* by V then standardize/IP weight for L,
 • L as *matching* factor (ensures positivity, difficult if high-dimensional L)
collapsibility: causal risk difference and ratio are weighted averages of stratum-specific risks, can not be done for odds ratio

Interaction   effects of joint interventions A and E

$$\Pr\left[Y^{1,1}{=}1\right] - \Pr\left[Y^{0,1}{=}1\right] \neq \Pr\left[Y^{1,0}{=}1\right] - \Pr\left[Y^{0,0}{=}1\right]$$

A and E have equal status and could also be considered a combined treatment AE, exchangeability for both is needed
additive scale (above): ">" superadditive and "<" subadditive;
multiplicative scale: ">" super- and "<" submultiplicative
difference to effect modification: if E is randomly assigned methods coincide, but V can not be intervened on as E can
monotonicity effect is either nonnegative or nonpositive $\forall i$
sufficient component-cause framework pedagogic model
response types for binary A: helped, immune, hurt, doomed;
for binary A and E: 16 types
(minimal) sufficient causes:
- (minimal) U_1 together with $A = 1$ ensure $Y = 1$
- (minimal) U_2 together with $A = 0$ ensure $Y = 1$
sufficient cause interaction: A and E appear together in a minimal sufficient cause

NPSEM   *non*parametric *s*tructural *e*quation *m*odel
$$V_m = f_m(pa_m, \epsilon_m)$$
counterfactuals are obtained recursively, e.g. $V_3^{v_1} = V_3^{v_1, V_2^{v_1}}$
implies any variable can be intervened on
aka finest causally interpreted structural tree graph (FCISTG)
additional assumption $\cap$ FCISTG $\Rightarrow$ causal Markov condition:
- independent errors (NPSEM-IE): all ϵ_m mutually independent
- fully randomized (FFRCISTG): $V_m^{\bar{v}_{m-1}} \perp\!\!\!\perp V_j^{\bar{v}_{j-1}}$ if $\bar{v}_{j-1}$ subvector of $\bar{v}_{m-1}$

NPSEM-IE $\Rightarrow$ FFRCISTG (assume DAGs represent latter)
NPSEM-IE assume crossworld independencies $\rightarrow$ unverifiable

Causal DAG   draw assumptions before conclusions
rules: arrow means direct causal effect for at least one i, absence means sharp null holds, all common causes are on the graph
neglects: direction of cause (harmful/protective), interactions
convention: time flows from left to right
causal Markov assumption: any variable (v) | its direct causes $(pa_j) \perp\!\!\!\perp$ its non-descendants $(\neg v_j) \Leftrightarrow$ Markov factorization
$$f(v) = \prod_{j=1}^{M} f(v_j | pa_j)$$
d-separation (d for directional): a pathway in a DAG is ...
- blocked if collider or conditioned on non-collider
- opened if conditioned on collider or descendent of collider

2 variables are d-separated if all connecting paths are blocked
under causal Markov: d-separation $\Rightarrow$ independence
under faithfulness: independence $\Rightarrow$ d-separation
faithfulness: effects don't cancel out perfectly
discovery: process of learning the causal structure; requires faithfulness, but even with it is often impossible

Noncausal DAGs   (Hernán and Robins, 2023) Y^a has to be well-defined (identifiability), what about Y^l (if $L \rightarrow Y$)?
if Y^l is not well-defined, but $L \rightarrow Y$, then the graph is not causal
statistical interpretation: only $A \rightarrow Y$ is causal, the rest simply encodes conditional independencies, *but* why should a DAG corresponding to the study variables even exist then?
hidden factor: L is only a surrogate for H, with Y^h well-defined, however, L being a surrogate can introduce bias
pragmatic approach: "cause" as a primary concept which does not need explanation in terms of well-defined interventions (approach is in need of mathematical theory)

SWIGs   *s*ingle *w*orld *i*ntervention *g*raphs
counterfactual graphic approach: A turns into $A|a$, the left (right) side inherits incoming (outgoing) arrows (intervention with $A = a$); all outcomes of A get a superscript a, e.g. Y^a; more than one intervention possible, dynamic strategies require additional arrows from L to a
A and Y^a are d-separated for $L \rightarrow Y^a \perp\!\!\!\perp A|L$ (for FFRCISTG)

Confounding   bias due to common cause of A & Y **not in** L
randomization prevents confounding
backdoor path: noncausal path A to Y with arrow into A
backdoor criterion: all backdoor paths are blocked by L & no descendants of A in $L \Rightarrow$ conditional exchangeability
$Y^a \perp\!\!\!\perp A|L \Rightarrow L$ fulfills backdoor criterion if faithful (FFRCISTG)
confounders in observational studies: occupational factors *(healthy worker bias),* clinical decisions *(confounding by indication/channeling),* lifestyle, genetic factors *(population stratification),* social factors, environmental exposures
given a DAG, confounding is an absolute, confounder is relative
surrogate confounders in L may reduce confounding bias
negative outcome controls: if A and Y share a common cause U: measure effect for Y_0 (before treatment) and Y_1 (after), subtract (assumption of additive equi-confounding)
front door criterion using the full mediator M: $\Pr\left[Y^a = 1\right] =$
$$\sum_m \Pr\left[M = m|A = a\right] \sum_{a'} \Pr\left[Y = 1|M = m, A = a'\right] \Pr\left[A = a'\right]$$

Selection Bias   bias due to common effect of A & Y **in** L
= conditioning on collider (can't be fixed by randomization)
examples: informative censoring, nonresponse bias, healthy worker bias, volunteer bias; often M-bias ($A \leftarrow U_1 \rightarrow L \leftarrow U_2 \rightarrow Y$)
solution: target $Y^{A,C}$, AC fulfills identifiability conditions, if competing events, interventions may not be well-defined
multiplicative survival model: $\Pr\left[Y=0|E=e, A=a\right] = g(e)h(a)$
$\rightarrow$ no interaction between E and A on the multiplicative scale; if $Y = 0$ is conditionally independent, then $Y = 1$ can't be as $\Pr\left[Y{=}1|E{=}e, A{=}a\right] = 1 - g(e)h(a) \rightarrow$ conditioning on a collider could be unbiased if restricted to certain levels ($Y = 0$)

Measurement Bias   aka information bias
measurements X^* of variables X can be included in DAG
independent errors U if $f(U_A, U_Y) = f(U_A)f(U_Y)$
nondifferential A: if $f(U_A|Y) = f(U_A)$; Y: $f(U_Y|A) = f(U_Y)$
mismeasurement $\rightarrow$ bias, if: $A \rightarrow Y$ *or* dependent *or* differantial
reverse causation bias caused by e.g. recall bias: independent but differential A (caused by $Y \rightarrow U_A$)
misclassified treatment: assignment Z does not determine A
exclusion restriction: ensure $Z \not\rightarrow Y$, e.g. via double-blinding
- ***per-protocol effect:*** either as-treated ($\rightarrow$ confounded) or restricted to protocol adhering individuals ($\rightarrow$ selection bias)
- ***intention-to-treat effect*** ($\rightarrow$ measurement bias): advantages: Z is randomized, preserves null (if exclusion restriction holds), = underpowered α-level test of the null (only if monotonicity; underpowered may be problematic if treatment safety is tested)

sometimes mismeasurement doesn't matter as the measurement is of interest (Hernán and Robins, 2023)

Random Variabilty   quantify uncertainty due to small n

CI: e.g. Wald CI $= \hat{\theta} \pm 1.96 \times se(\hat{\theta})$, *calibrated* if it contains 95 % of estimands ($>$: *conservative*, $<$: *anticonservative*)

large sample CI: converge to 95 % vs. *small-sample:* always valid

honest: $\exists n$ where coverage $\geq$ 95 %, *valid:* large-sample & honest

inference: either restrict inference to sample (randomization-based inference) or inference on super-population

super-population: generally a fiction, but $\rightarrow$ simple statistical properties (where does the variability of the distribution come from: assumption population is sampled from super-population)

conditionality principle: inference should be performed conditional on ancillary statistics (e.g. L-A association) as

$$\mathcal{L}(Y) = f(Y|A,L)f(A|L)f(L)$$

exactly ancillary A, L: $f(Y|A,L)$ depends on parameter of interest, but $f(A,L)$ does not share parameters with $f(Y|A,L)$

approximately ancillary: ... does not share **all** parameters ...

continuity principle: also condition on approximate ancillaries

curse of dimensionality: difficult to do conditionality principle

Time-Varying Treatments   compare 2 treatments

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

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

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

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

stochastic strategy: non-deterministic g

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

Sequential Identifiability   sequential versions of

exchangability: $Y^g \perp\!\!\!\perp A_k | \bar{A}_{k-1} \ \forall g, k = 0, 1, ..., K$

conditional exchangeability:

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

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

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

consistency:

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

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

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

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

use SWIGs to visually check d-separation

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

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 $\Rightarrow$ $Y^g = Y \ \forall g \ \Rightarrow \ Y \perp\!\!\!\perp A_0 | L_0 \ \& \ Y \perp\!\!\!\perp A_1 | A_0, L_0, L_1$; therefore: if last two independences don't hold, one assumption is violated

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

2 Models

Modeling   data are a sample from the target population

estimand:	quantity of interest,	e.g. $\mathrm{E}[Y|A=a]$
estimator:	function to use,	e.g. $\widehat{\mathrm{E}}[Y|A=a]$
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 $\to$ true value)
parsimonious model: few parameters estimate many quantities
bias-variance trade-off:
wiggliness $\uparrow$ $\to$ misspecification bias $\downarrow$, CI width $\uparrow$

Variable Selection   can induce bias if L includes:

(decendant 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)

2.1 Traditional Methods

Stratification   calculate risk for each stratum of L
only feasible if enough data per stratum

Outcome Regression   often assume no effect modification
$$\mathrm{E}[Y^{a,c=0}|L] = \beta_0 + \beta_1 a + \beta_2 aL + \beta_3 L = \mathrm{E}[Y|A,C=0,L]$$
faux marginal structural model as no IP weighting/$SW^A(L) = 1$
for ATE only β_1, β_2 of interest, the rest are *nuisance parameters*

Propensity Score Methods $\Pr[A=1|L] =: \pi(L)$
$\Rightarrow A \perp\!\!\!\perp L|\pi(L)$ (definition of a balancing score); can be modelled
- *stratification:* create strata with similar $\pi(L)$ (e.g. deciles), but the average $\pi(L)$ might still be different in some strata
- *standardization:* use $\pi(L)$ instead of L to standardize
- *matching:* find close ($\to$ bias-variance trade-off) values of $\pi(L)$, positivity issues arise often

propensity models don't need to predict well, just ensure exchangeability (good prediction leads to positivity problems)

Instrumental Variable Estimation L unmeasured
surrogate/proxy instruments can be used
instrumental conditions:
1. *relevance condition:* $Z \not\!\perp\!\!\!\perp A$, meaning Z is associated with A (weak association (F-statistic < 10) $\to$ weak instrument)
2. *exclusion restriction:* Z affects Y at most through A
   (a) population level: $\mathrm{E}[Y^{z,a}] = \mathrm{E}[Y^{z',a}]$ (sometimes enough)
   (b) *individual level:* $Y_i^{z,a} = Y_i^{z',a} = Y_i^a$
3. *exchangeability:* Z and Y have no shared causes
   (a) *marginal:* $Y^{a,z} \perp\!\!\!\perp Z$ (typically enough)
   (b) joint: $\{Y^{z,a}; a \in [0,1], z \in [0,1]\} \perp\!\!\!\perp Z$
4. (not needed for an instrument, just the IV estimand below)
   (a) *effect homogeneity:* (i) constant effect $A \to Y$ $\forall i$ (ii) constant average effect $A \to Y$ $\forall A$ (iii) no additive effect modifiers (iv) additive Z-A association is constant across L
   (b) *monotonicity:* $A^{z=1} \geq A^{z=0}$ $\forall i$ (more credible than 4a)

Super Learning   (Van der Laan et al., 2007, 2011)
oracle selector: select best estimator of set of learners Z_i
discrete super learner: select algorithm with smallest cross-validated error (converges to oracle for large sample size)
super learner: improves asymptotically on discrete version
$$\mathrm{logit}(Y=1|Z) = \sum_i \alpha_i Z_i, \text{ with } 0 < \alpha_i < 1 \text{ and } \sum \alpha_i = 1$$
weights α_i are determined inside the cross-validation; for the prediction, Z_i trained on the full data set are used
can be cross-validated itself to check for overfitting (unlikely)

Marginal Structural Models   association is causation
in the IP weighted pseudo-population
$$\text{associational model } \mathrm{E}[Y|A] = \text{ causal model } \mathrm{E}[Y^a]$$
step 1: estimate/model $f[A|L]$ (and $f[A]$) $\to$ get $(S)W^A$
step 2: estimate regression parameters for pseudo-population
effect modification variables V can be included (e.g. $\beta_0 + \beta_1 a + \beta_2 Va + \beta_3 V$; technically not marginal anymore), $SW^A(V) = \frac{f[A|V]}{f[A|L]}$ more efficient than SW^A

common instruments: (physician's) general preference, access to/price of A, genetic factors (Mendelian randomization)
bounds: binary outcome ATE $[-1, 1]$ (width 2) $\overset{data}{\to}$ (width 1)
natural bounds need 2a,3a (width $\Pr[A=1|Z=0] + \Pr[A=0|Z=1]$)
sharp bounds require 2a,3b (narrower than natural bounds)
IV estimand ATE: intention-to-treat $\div$ measure of compliance
(1,2b,3a,4a): ATE; (1,2b,3a,4b): ATE in compliers
binary Z: $\frac{\mathrm{E}[Y|Z=1]-\mathrm{E}[Y|Z=0]}{\mathrm{E}[A|Z=1]-\mathrm{E}[A|Z=0]}$, continuous Z: $\frac{Cov(Y,Z)}{Cov(A,Z)}$;
can be calculated as *two-stage-least-squares estimator:*
1. $\mathrm{E}[A|Z]$ 2. $\mathrm{E}[Y|Z] = \beta_0 + \beta_1\widehat{\mathrm{E}}[A|Z]$ 3. $\hat{\beta}_1$ is IV estimate
disadvantages: often leads to wide CI, small violations of conditions can lead to large biases
regression discontinuity design: if threshold in L exists which determines A perfectly + assumption of continuity in L $\to$ jump in Y at threshold is the causal effect (if no effect modification by L); a fuzzy variant also exists (Hernán and Robins, 2023)

Causal Survival Analysis   time-to-event data
additional censoring due to administrative end of follow-up
competing events (often death): censoring (assume population with death abolished) or not (after death, chance of event is zero, but what is the effect of A?) $\to$ create composite event
survival quantities k is a time point, T is time of event
- *survival* at k: $\Pr[T > k] =: \Pr[D_k = 0]$
- *risk* at k: $1 - \Pr[T > k] = \Pr[T \leq k] = \Pr[D_k = 1]$
- *hazard* at k: $\Pr[T = k|T > k-1] = \Pr[D_k = 1|D_{k-1} = 0]$, *hazard ratio* is paradoxical due to in-built selection bias
modeling: some options
- *Kaplan-Meier* aka product limit formula (nonparametric): $\Pr[D_k = 0] = \prod_{m=1}^{k} \Pr[D_m = 0|D_{m-1} = 0]$
- parametric e.g. log hazards model:
  - use *IP weigths* SW^A in structural marginal model $\mathrm{logit}\Pr\left[D_{k+1}^{a,\bar{c}=\bar{0}} = 0|D_k^{a,\bar{c}=\bar{0}} = 0\right] = \beta_{0,k} + \beta_1 a + \beta_2 ak$
  - *standardize* $(\prod_k 1-)$ parametric hazards model $\Pr[D_{k+1} = 1|D_k = 0, C_k = 0, L, A]$ weighting across L

- **structural nested cumulative failure time model (CFT):**
  $\frac{\Pr[D_k^a=1|L,A]}{\Pr[D_k^{a=0}=1|L,A]} = \exp[\gamma_k(L,A;\psi)]$ (log-linear has no upper limit $1 \to$ rare failure $\uparrow$; if $\downarrow$, use a survival model (CST)), use g-estimation like with AFT
- **accelerated failure time model (AFT)** with g-estimation: $T_i^a/T_i^{a=0} = \exp(-\psi_1 a - \psi_2 a L_i)$, exchangeability for C is guaranteed via artificial censoring (include only individuals who would not have been censored either way)

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 and IP weighting are equivalent, **but** if modeled, different "no misspecification" assumptions: outcome model (standardization), treatment model (IP weighting)

censoring: measure joint effect of A and C with $E\left[Y^{a,c=0}\right]$

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

IP weights $\quad W^{A,C} = W^A \times W^C \qquad$ (uses n) $\qquad$ or

$\qquad\qquad SW^{A,C} = SW^A \times SW^C \quad$ (uses $n^{c=0}$)

g-estimation only adjusts for confounding $\to$ use IP weights

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 $\to$ predict counterfactuals on whole dataset $\to$ average the results ($\to$ CI by bootstrapping)

for discrete L $E[Y|A=a] = \sum_l E[Y|L=l, A=a]\Pr[L=l]$

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 ($L \not\to A$)

FRCISTG *(fully randomized causally interpreted structured graph):* probability tree for $L \to A \to Y$, 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!) $\to$ pseudo-population same size $\to$ CI width $\downarrow$

G-Estimation (additive) structural nested models

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

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

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

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

$Y^{a=0}$ unknown, but because of exchangeability α_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

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

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

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

semi-parametric: agnostic about β_0 and effect of $L \to$ robust $\uparrow$

no time-varying: no nesting; model equals marginal structural models with missing β_0, β_3 (unspecified "no treatment")

sensitivity analysis: unmeasured confounding ($\alpha_1 \neq 0$) can be examined: do procedure for different values of $\alpha_1 \to$ plot α_1 vs. $\psi^\dagger \to$ how sensitive is estimate to unmeasured confounding?

effect modification: add V in both g-estimation equations

doubly robust estimators exist

time-varying nested equations: for each time k

strutural nested mean models separate effect of each a_k

$$\text{E}\left[Y^{\bar{a}_{k-1}, a_k, \underline{0}_{k+1}} - Y^{\bar{a}_{k-1}, \underline{0}_{k+1}}|\bar{L}_{k-1} = \bar{l}_k, \bar{A}_{k-1} = \bar{a}_{k-1}\right] =$$

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

calculations

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

function γ_j can be, e.g. constant (ψ_1), time-varying only ($\psi_1 + \psi_2 k$), or dependent on treatment/covariate history

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

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

find α_1 that is closest to zero

a closed form estimator exists for the linear case

2.3   Doubly Robust Methods

Double-Robustness   (Hernán and Robins, 2023)

g-formula: *either* treatment model $f(L)$ *or* outcome model $b(L)$ *or* appropriately combine both: "two chances to get it right"

all doubly robust estimators

- involve a correction of outcome $\hat{b}(L)$ using the treatment $\hat{f}(L)$
- have a bias depending on a product of the errors $\frac{1}{\pi(l)} - \frac{1}{\hat{\pi}(l)}$ and $b(l) - \hat{b}(l)$ known as second order bias

time-varying: multiple robustness for $k = 0, 1, \ldots K$
$K + 2$ robustness: consistent, if $\hat{f}_0$ to $\hat{f}_l$ and $\hat{b}_{l+1}$ to $\hat{b}_K$ are
2^{K+1} robustness: consistent, if for each k, either $\hat{f}_k$ or $\hat{b}_k$ are

Machine Learning L is high-dimensional

one could use lasso or ML for IP weighting/standardization

but: ML does not guarantee elimination of confounding and has largely unknown statistical properties: how to get CI?

sample splitting: train estimators on training sample T_r, 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 [*alternatively:* split into M samples, use one sample for estimation and $M-1$ for training $\to$ improved finite sample behavior (Hernán and Robins, 2023)]

asymptotic behavior for valid (Wald) CI we need:

- a bias much smaller than $c \cdot 1/\sqrt{n}$, which is how the *se* typically scales (use doubly robust methods for small bias)
- asymptotic normality (for Wald CI)
- for a doubly robust estimator ψ_{dr}, we need sample splitting, otherwise $\hat{b}(l)$ and $\hat{f}(l)$ are correlated with ψ_{dr}

if $\hat{b}(l)$ and $\hat{f}(l)$ are consistent and $\text{E}[\hat{\psi} - \psi|T_r]/se(\hat{\psi})$ converges to $0 \to \hat{\psi}$ with sample splitting is asymptotically normal and unbiased $\to$ CI is calibrated (Hernán and Robins, 2023)

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

Advantages   (Van der Laan et al., 2011)

consistent *if either $\bar{Q}_0$ or g_n are consistent (doubly robust):*

$$\forall \epsilon > 0, P \in \mathcal{M} : \Pr_P\left[|\hat{\theta}_n - \theta(P)| > \epsilon\right] \to 0 \text{ as } n \to \infty$$

collaboratively doubly robust: g_n only needs predictors of Y, as it does not try to fit g_0 well, but improve the fit of $\bar{Q}_n^*$

asymptotic unbiasedness *if either $\bar{Q}_0$ or g_0 are consistent*, super learning makes $\bar{Q}_0$ and g_n max. asymptotically unbiased

asymptotic efficiency *if both $\bar{Q}_0$ and g_n are consistent*: achieves Cramer-Rao bound of minimum possible asymptotic variance (requires asymptotic unbiasedness)

asymptotic linearity *if either $\bar{Q}_0$ or g_n are consistent*: means estimator behaves like empirical mean

- bias converges to zero at rate smaller than $1/\sqrt{n}$
- for large n estimator is approximately normally distributed

Influence Curve   (Hampel, 1974; Van der Laan et al., 2011; Zepeda-Tello et al., 2022) how robust is an estimator?

$$IC_{T,P_n}(O) = \lim_{\epsilon \to 0} \frac{T\left[(1 - \epsilon)P_n + \epsilon \delta_O\right] - T(P_n)}{\epsilon}$$

for estimator T and distribution P_n with $0 < \epsilon < 1$

can also be rewritten as

$$IC_{T,P_n} = \frac{d}{d\epsilon}T\left[(1 - \epsilon)P_n + \epsilon \delta_O\right] = \frac{d}{dP_n}T\left(\delta_O - P_n\right)$$

last part is a ***directional derivative*** at P_n (empirical probability measure that puts mass $1/n$ on O_i) in direction $(\delta_O - P_n)$

$\overline{IC(P_0)} = 0$ and $\text{Var}(IC(P_0))$ is the asymptotic variance of the standard estimator $\sqrt{n}(\psi_n - \psi_0)$, therefore $Var(\hat{\Psi}(P_n)) = \frac{Var_{IC}}{n}$

efficient IC: an estimator is asymptotically efficient $\Leftrightarrow$ its influence curve is the efficient influence curve $IC(O) = D^*(O)$

estimand written as a function of θ, i.e. $\psi := \phi(\theta)$, we know the variance of $\hat{\theta}$, but what is the variance of $\hat{\psi}$

classical delta method: (under regularity conditions) distribution of $\phi(\hat{\theta})$ can be approximated as a normal with a variance proportional to ϕ's rate of change at θ:

$$\sqrt{n}\left(\phi(\hat{\theta}_n) - \phi(\mu)\right) \overset{\text{approx}}{\sim} \text{N}\left(0, \phi'(\theta)\sigma^2\right)$$

which allows for Wald-type CI: $\hat{\theta}_n \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\phi'(\theta)\sigma^2}{n}}$ (very similar to central limit theorem)

Taylor's approximation requirements:

- univariate ϕ: differentiable at θ
- multivariate ϕ: $\exists \, \partial_v\phi(\theta)$ (directional derivative)
- functional ϕ (function of functions): $\exists \, \partial_v\phi(\theta)$ & coincides with one-sided directional (Hadamard) derivatives ($\overset{*}{=} \nabla\phi(\theta)^T v$)

first order Taylor (rearranged†): $\phi(\hat{\theta}_n) \overset{\approx}{\approx} \phi(\theta) \overset{\approx}{+} \partial_{v:=\hat{\theta}-\theta}\phi(\theta)$

classical delta method: if $\{r_n\}_{n=1}^{\infty}$ with $\lim_{n\to\infty} r_n = \infty$, where $r_n(\hat{\theta}_n - \theta)$ converges to $Z \sim N(0,1)$ (e.g. $r_n = \sqrt{n/\sigma^2}$), then

$$r_n\left(\phi(\hat{\theta}_n) - \phi(\theta)\right) \overset{\dagger*}{\approx} \nabla\phi(\theta)^T r_n(\hat{\theta}_n - \theta) \overset{d}{\to} \nabla\phi(\theta)^T Z$$

$$\Rightarrow \mathrm{Var}\left[\phi(\hat{\theta}_n) - \phi(\theta)\right] = \mathrm{Var}\left[\phi(\hat{\theta}_n)\right] \approx \frac{1}{r_n^2}\mathrm{Var}\left[\nabla\phi(\theta)^T Z\right]$$

functional delta: $r_n(\hat{\theta}_n - \theta) \overset{d}{\to} Z \Rightarrow r_n\left(\phi(\hat{\theta}_n) - \phi(\theta)\right) \overset{d}{\to} \partial_Z\phi(\theta)$

influence function: $\psi = \phi(\mathbb{P}_X)$ is a functional estimations rate of change for $\mathbb{P}_X$ to Q, where $Q = \mathbb{1}_{\{Y\}}$

$$\mathrm{IF}_{\phi,\mathbb{P}_X}(Y) := \partial_{Q-\mathbb{P}_X}\phi(\mathbb{P}_X) = \lim_{h\downarrow 0}\frac{\phi((1-h)\mathbb{P}_X + hQ) - \phi(\mathbb{P}_X)}{h},$$

interpretation: rate of change if distribution deviates from $\mathbb{P}_X$ to $Q =$ one observation Y, assigns probability 1 to X taking value Y

use delta: $\phi(\hat{\mathbb{P}}_X) \approx \phi(\mathbb{P}_X) + \mathrm{IF}_{\phi,\mathbb{P}_X}(Y)$, if $(\hat{\theta}_n - \theta) \overset{n\to\infty}{\sim} N(.,.)$

$$\hat{\psi}_n - \psi = \phi(\hat{\theta}_n) - \phi(\theta) \overset{\text{approx}}{\sim} N\left(0, \mathrm{Var}[\mathrm{IF}_{\phi,\mathbb{P}_X}(Y)]\right),$$

where $\widehat{\mathrm{Var}}[\mathrm{IF}_{\phi,\mathbb{P}_X}(Y)] = \frac{1}{n}\sum_{i=1}^n\left(\mathrm{IF}_{\phi,\mathbb{P}_X}(X_i)\right)^2$, which is the classical S^2 estimator since the mean is known $(=0)$

using the delta method (general case)

1. determine asymptotic distribution of $v := r_n(\hat{\theta}_n - \theta)$
2. define ϕ and compute Hadamard derivative
3. multiply asymptotic distribution with Hadamard derivative, then estimate the variance

Simple Plug-In Estimator

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

time-varying $k+2$ robust estimator (related to TMLE)

1. estimate $\hat{f}\left(A_m|\bar{A}_{m-1},\bar{L}_m\right)$ (e.g. logistic model), use it to calculate at each time m: $\widehat{W}^{\bar{A}_m} = \prod_{k=0}^m \frac{1}{\hat{f}\left(A_k|\bar{A}_{k-1},\bar{L}_k\right)}$ and modified IP weights at m: $\widehat{W}^{\bar{A}_{m-1},a_m} = \frac{\widehat{W}^{\bar{A}_{m-1}}}{\hat{f}\left(a_m|\bar{A}_{m-1},\bar{L}_m\right)}$

2. with $\widehat{T}_{K+1} := Y$, recursively for $m = K, K-1, ..., 0$:
   (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 $\widehat{\mathrm{E}}\left[Y^{\bar{a}}\right] = \mathrm{E}\left[\widehat{T}_0\right]$

Augmented IP Weighting (AIPTW) (Hernán and Robins, 2023)

$$\hat{\mathrm{E}}\left[Y^a\right] = \frac{1}{n}\sum_{i=1}^n\left[\frac{\mathbb{1}(A=a)Y}{\hat{f}(A|L)} - \left(\frac{\mathbb{1}(A=a)}{\hat{f}(A|L)} - 1\right)\hat{b}(a,L)\right]$$

Relationship between AIPTW and TMLE for causal effect:

$$\hat{\psi}_{1,AIPTW} - \hat{\psi}_{0,AIPTW} = P_n\left[\hat{b}(1,L)\right] - P_n\left[\hat{b}(0,L)\right]$$

$$-P_n\left[\frac{\{\mathbb{1}(A=1) - \mathbb{1}(A=0)\}\left(Y - \hat{b}(A,L)\right)}{\hat{f}(A|L)}\right]^\dagger$$

using the IRLS estimate for
$b(A,L;\beta,\theta) = \phi\left[m(A,L;\beta) + \theta\left\{\overbrace{\frac{\mathbb{1}(A=1) - \mathbb{1}(A=0)}{\hat{f}(A|L)}}^{\text{clever covariate}}\right\}\right]$ with canonical link ϕ sets the last part† to zero (as the score equation for θ)

TMLE (Van der Laan et al., 2011)

targeted maximum likelihood estimation

$$O = (W, A, Y) \sim P_0$$

target $\Psi(P_0) = \Psi(\bar{Q}_0, Q_{W,0}) = \psi_0$,

often: $\mathrm{E}_{W,0}\left[\mathrm{E}_0(Y|A=1,W) - \mathrm{E}_0(Y|A=0,W)\right]$

first step: outcome model $\bar{Q}_n^0(A,W)$ estimating $\bar{Q}_0$ (part of P_0)

- super learning is often used here, but leads to a biased estimate

- not all of $f(Y|A,W)$ needs to be estimated, just the relevant portion, *typically average outcome* $\mathrm{E}_0(Y|A,W) \to$ efficiency $\uparrow$

second step: update $\bar{Q}_n^0(A,W)$ to $\bar{Q}_n^1(A,W)$ using treatment model g_n estimating $g_0 = P_0(A|W)$

1. model g_n, super learning is a popular choice here, too
2. calculate n clever covariates: $H_n^*(A,W) = \begin{cases} \frac{1}{g_n(1|W)} & \text{if } A_i=1 \\ \frac{1}{g_n(0|W)} & \text{if } A_i=0 \end{cases}$
3. update $\bar{Q}_n^0$, by estimating ϵ_n with offset logistic regression: $\mathrm{logit}\bar{Q}_n^1(A,W) = \mathrm{logit}\bar{Q}_n^0(A,W) + \epsilon_n H_n^*(A,W)$ (converges after first update), then calculate counterfactuals

- goal: bias reduction, get optimal bias-variance trade-off
- removes all asymptotic bias, if consistent estimator is used here

third step: use empirical distribution for $Q_{W,0}$ in a substitution estimator, *e.g.*: $\psi_n^{TMLE} = \frac{1}{n}\sum_{i=1}^n\left[\bar{Q}_n^1(1,W_i) - \bar{Q}_n^1(0,W_i)\right]$

advantages: loss-based (does not only solve efficient influence curve estimating equation, but also uses a loss and working model preserving global constraints), well-defined (as a loss-based learner), substition estimator (respects global constraints $\to$ more robust to outliers and sparsity)

closed form inference based on the influence curve:

$$IC_n^*(O_i) = \overbrace{\left[\frac{\mathbb{1}(A_i=1)}{g_n(1,W_i)} - \frac{\mathbb{1}(A_i=0)}{g_n(0,W_i)}\right]\left[Y - \bar{Q}_n^1(A_i,W_i)\right]}^{a}$$
$$+ \underbrace{\bar{Q}_n^1(1,W_i) - \bar{Q}_n^1(0,W_i) - \psi_{TMLE,n}}_{b}$$

TMLE sets the mean of the IC, $\overline{IC}_n$, to zero (b has already mean zero, see third step, the first part of a is the clever covariate)

sample variance is then: $S^2(IC_n) = \frac{1}{n}\sum_{i=1}^n\left(IC_n(o_i) - \overline{IC}_n\right)^2$

standard error of estimator: $\sigma_n = \sqrt{\frac{S^2(IC_n)}{n}}$

95% CI: $\psi_{TMLE,n} \pm z_{0.975}\frac{\sigma_n}{\sqrt{n}}$; p-value: $2\left[1 - \Phi\left(\left|\frac{\psi_{TMLE,n}}{\sigma_n/\sqrt{n}}\right|\right)\right]$

LTMLE longitudinal

for $t = T, ..., 1$:

1. model $\mathrm{E}(Y_t|\bar{A}_{t-1},\bar{L}_t)$ (fit on individuals that are uncensored and alive at $t-1$

2. plug in $\bar{a}_{t-1} = \bar{d}_{t-1}$; use regression from 1 to predict outcome at time t, ie. $\bar{Y}_t^{\bar{d}_t}$

3. update estimate with
$Y_t = offset(step2resultint) + \epsilon \times clevercovariate$: predict $\bar{Y}_t^{\bar{d}_t}$
(alternatively the clever covariate can be used as a weight)

4. $\hat{\psi}_T =$ mean of $\bar{Y}_1^{\bar{d}_1}$

TMLE advanced (Van der Laan et al., 2011)

targeted minimum loss-based estimation

target parameter $\Psi : \mathcal{M} \to \mathbb{R}$, with $\mathcal{M}$ the statistical model used

1. compute its pathwise derivative at P and corresponding canonical gradient $D^*(P)$ (efficient influence curve: a function of O with mean zero under P)

2. define loss function $L()$ s.t. $P \to E_0 L(P)$ is minimized at true P_0 (or just relevant Q)

3. for a P in model $\mathcal{M}$ define a parametric working model $\{P(\epsilon) : \epsilon\}$ s.t. $P(\epsilon = 0) = P$ and a "score" $\frac{d}{d\epsilon}L(P(\epsilon))$: score (or linear combination of its components) equals $D^*(P)$ at P (or just relevant Q)

4. with initial estimate P_n^0, compute $\epsilon_n^0 = \arg\min_\epsilon \sum_{i=1}^n L(P_n^0(\epsilon))(O_i)$, calculate first iteration $P_n^1 = P_n^0(\epsilon_n^0)$, repeat until $\epsilon_n^k = 0$ (or just relevant Q)

5. get TMLE estimate ψ_0 as the substitution estimator pluggint P_n^* into Ψ

6. TMLE olves the efficient influence curve equation
$$0 = \sum_{i=1}^{n} D^*(P_n^*)(O_i) \rightarrow \text{esymptotic linearity and efficiency}$$

$$\mathcal{L}(O) = \overbrace{\Pr(Y|A,W)}^{Q_Y} \overbrace{\Pr(A|W)}^{g} \overbrace{\Pr(W)}^{Q_W}$$: g itself is not needed as we intervene on treatment, but it can help improving the estimate of Q_Y

$H(A,W)$ depends on target parameter and loss function but is a function of the propensitiy score update initial fit
$$\bar{Q}_n^* = \bar{Q}_n^0 + \hat{\epsilon}H(A,W)$$

valid inference, good finite sample performance,

$H(A,W)$ comes from the influence curve, targeting ensures mean of efficient influence curve $D^*(P)$ is zero

TMLE solves $P_n D^*(P_n^*) = 0$

TMLE is a substitution estimator
$$\psi_n^{TMLE} = \frac{1}{2} \sum_{i=1}^{n} \bar{Q}_n^*(1,W_i) - \frac{1}{2} \sum_{i=1}^{n} \bar{Q}_n^*(0,W_i)$$ therefore mean of b is zero

targeting step makes sure a also has mean zero

MLE solves $\sum_{i=1}^{n} H(A_i,W_i) \left[Y_i - \bar{Q}_n^*(A_i,W_i) \right] = 0$ where $\bar{Q}_n^*(A_i,W_i) = \hat{\epsilon}H(A,W) + \bar{Q}_n^0$ therefore obvious choice:
$$H(A,W) = \frac{A}{g(1,W)} - \frac{1-A}{g(0,W)}$$

influence curve based inference: asymptotic linearity
$$\sqrt{n} \left(\psi_n^{TMLE} - \psi_0 \right) \xrightarrow{D} \text{N}(0,\sigma^2)$$

AIPTW *augmented inverse probability of treatment weighting*

disadvantages (Van der Laan et al., 2011): ignores global constraints $\rightarrow$ often unstable under sparsity, sometimes not well-defined

References

If no citation is given, the information is taken from the book (Hernán and Robins, 2020)

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393.

Hernán, M. A. and Robins, J. M. (2020). *Causal inference: what if.* Boca Raton: Chapman & Hall/CRC.

Hernán, M. A. and Robins, J. M. (2023). *Causal inference: what if.* Boca Raton: Chapman & Hall/CRC.

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.

Schomaker, M., Luque-Fernandez, M. A., Leroy, V., and Davies, M.-A. (2019). Using longitudinal targeted maximum likelihood estimation in complex settings with dynamic interventions. *Statistics in medicine*, 38(24):4888–4911. ISBN: 0277-6715 Publisher: Wiley Online Library.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1). Article 24.

Van der Laan, M. J., Rose, S., et al. (2011). *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer.

Young, J. G., Cain, L. E., Robins, J. M., O'Reilly, E. J., and Hernán, M. A. (2011). Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Statistics in biosciences*, 3:119–143.

Zepeda-Tello, R., Schomaker, M., Maringe, C., Smith, M. J., Belot, A., Rachet, B., Schnitzer, M. E., and Luque-Fernandez, M. A. (2022). The delta-method and influence function in medical statistics: a reproducible tutorial. *arXiv preprint arXiv:2206.15310*.