Modern Statistical Learning Methods for Observational Biomedical Data

# Chapter 2: Basic identification and estimation of an average treatment effect

**David Benkeser**
Emory Univ.

**Marco Carone**
Univ. of Washington

**Larry Kessler**
Univ. of Washington

MODULE 4

**Summer Institute in Statistics for Clinical and Epidemiological Research**

July 2019

# Contents of this chapter

# When is identification possible?

To compute the average treatment effect, it suffices to compute both counterfactual means $\psi_1 := E[Y(1)]$ and $\psi_0 := E[Y(0)]$ since

$$\gamma := ATE = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)] .$$

The observed data consist of $O_1, O_2, \ldots, O_n \overset{iid}{\sim} P_0$, with $O_i := (W_i, A_i, Y_i)$ and

$W_i$ = the vector of baseline patient characteristics (i.e., potential confounders);
$A_i$ = the treatment/intervention received;
$Y_i$ = the outcome of interest.

**Two fundamental questions:**

1. When is $E[Y(a)]$ identifiable (i.e., estimable) from the observed data?
2. When it is so, how can we identify it?

# When is identification possible?

First key condition: **the randomization** (or ignorability) **condition**

In a clinical trial, it is true that $(Y(0), Y(1)) \perp A$ since the treatment allocation mechanism provides no information on the counterfactual outcomes.

In an observational study, this is generally not true.

e.g.: treatment allocation $\longleftrightarrow$ disease severity $\longrightarrow$ counterfactual survival

The randomization condition is said to hold provided

$$(Y(0), Y(1)) \perp A \mid W$$

implying that the treatment is randomized within strata of the recorded covariates.

This will generally hold if:

- the study guarantees it by design (e.g., stratified randomized trial);
- all potential confounders have been recorded.

This condition is generally not verifiable empirically – prior knowledge is key.

# When is identification possible?

Second key condition: **the positivity** (or experimental treatment assignment) **condition**

In a clinical trial, all patients may potentially be assigned to each treatment group.

In an observational study, this is generally not true.

e.g.: no patient with mild disease assigned to (risky) experimental treatment

The positivity condition holds provided, for each $a$,

$$P(A = a \mid W = w) > 0 \text{ for every plausible value } w$$

implying that each patient may potentially be assigned to any treatment group.

In some cases, the plausibility of this condition can be assessed empirically.

# When is identification possible?

If both randomization and positivity conditions hold, the counterfactual mean $E[Y(a)]$ can generally be identified from the observed data.

It can be calculated as a summary of the distribution $P_0$ of the observed data unit $O$.

We will focus on the two most important identification formulas:

- the G-computation formula; (Robins, 1986)
- the inverse-probability-of-treatment weighting (IPTW) formula. (Horvitz & Thompson, 1952; Robins, Hernan & Brumback, 2000)

# The G-computation identification formula

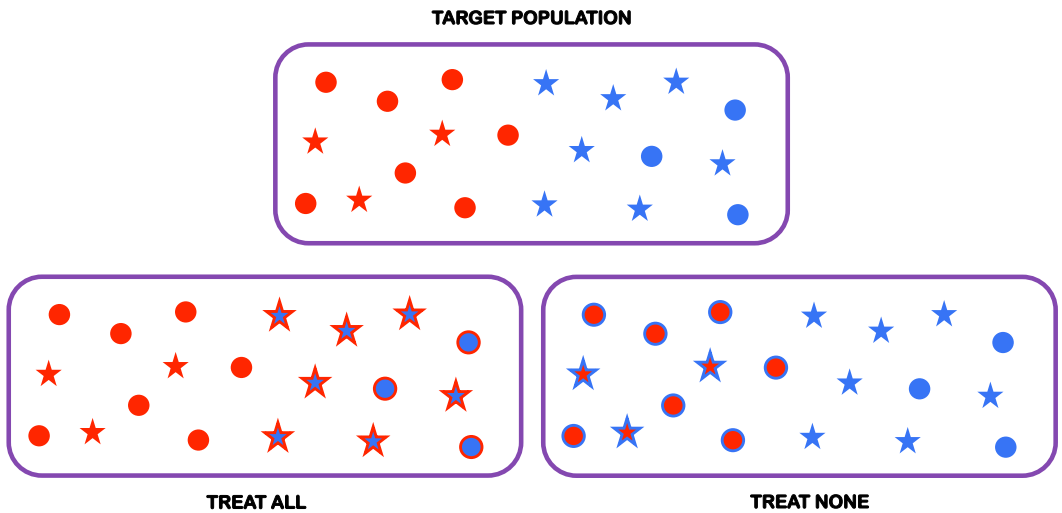
If the randomization and positivity conditions hold, then by the G-computation formula

$$E[Y(a)] \;=\; E\left[E(Y \mid A = a, W)\right] \;=\; \sum_{w} E(Y \mid A = a, W = w) P(W = w) \;.$$

---

Heuristically, what does this amount to doing?

1. Find the expected outcome under treatment $A = a$ for each type of patient.
2. Average these out according to the composition of the target population.



# The G-computation identification formula

If the randomization and positivity conditions hold, then by the G-computation formula

$$E[Y(a)] \;=\; E\left[E(Y \mid A=a, W)\right] \;=\; \sum_w E(Y \mid A=a, W=w)P(W=w)\ .$$

---

**Key observation:** averaging is performed relative to the marginal distribution of $W$!

G-computation pools subgroup-specific treatment effects across target population:

$$\begin{aligned} ATE \;&=\; E[Y(1)] - E[Y(0)] \\ &=\; \sum_w \{E(Y \mid A=1, W=w) - E(Y \mid A=0, W=w)\}P(W=w) \end{aligned}$$

Contrast this with the naive difference in means between treatment groups:

$$\begin{aligned} \text{naive difference} \;&=\; E(Y \mid A=1) - E(Y \mid A=0) \\ &=\; \sum_w E(Y \mid A=1, W=w)P(W=w \mid A=1) \\ &\qquad\qquad - \sum_w E(Y \mid A=0, W=w)P(W=w \mid A=0) \end{aligned}$$

# The G-computation identification formula

If the randomization and positivity conditions hold, then by the G-computation formula

$$E[Y(a)] \;=\; E\left[E(Y \mid A = a, W)\right] \;=\; \sum_w E(Y \mid A = a, W = w) P(W = w) \;.$$

---

**Key observation:** averaging is performed relative to the marginal distribution of $W$!

This allows us to easily extend the idea to other (related) causal estimands:

**average treatment effect among the treated**:

$$ATT \;=\; \sum_w \{E(Y \mid A = 1, W = w) - E(Y \mid A = 0, W = w)\} P(W = w \mid A = 1)$$

**average treatment effect among controls**:

$$ATC \;=\; \sum_w \{E(Y \mid A = 1, W = w) - E(Y \mid A = 0, W = w)\} P(W = w \mid A = 0)$$

When would such causal estimands be preferred?

# The G-computation identification formula

The G-computation formula can be derived as follows.

$$
\begin{aligned}
E[Y(a)] &= \sum_w E[Y(a) \mid W = w] P(W = w) && \text{(law of total expectation)} \\
&= \sum_w E[Y(a) \mid A = a, W = w] P(W = w) && \text{(randomization property)} \\
&= \sum_w E(Y \mid A = a, W = w) P(W = w) && \text{(consistency)}
\end{aligned}
$$

For $E(Y \mid A = a, W = w)$ to be defined, the positivity assumption must hold.

# The IPTW identification formula

The **inverse-probability-of-treatment weighting (IPTW)** identification formula gives an alternative means of expressing the ATE in terms of the observed data distribution.

If the randomization and positivity conditions hold, then by the IPTW formula

$$E[Y(a)] \;=\; E\left[\frac{I(A=a)Y}{P(A=a \mid W)}\right].$$

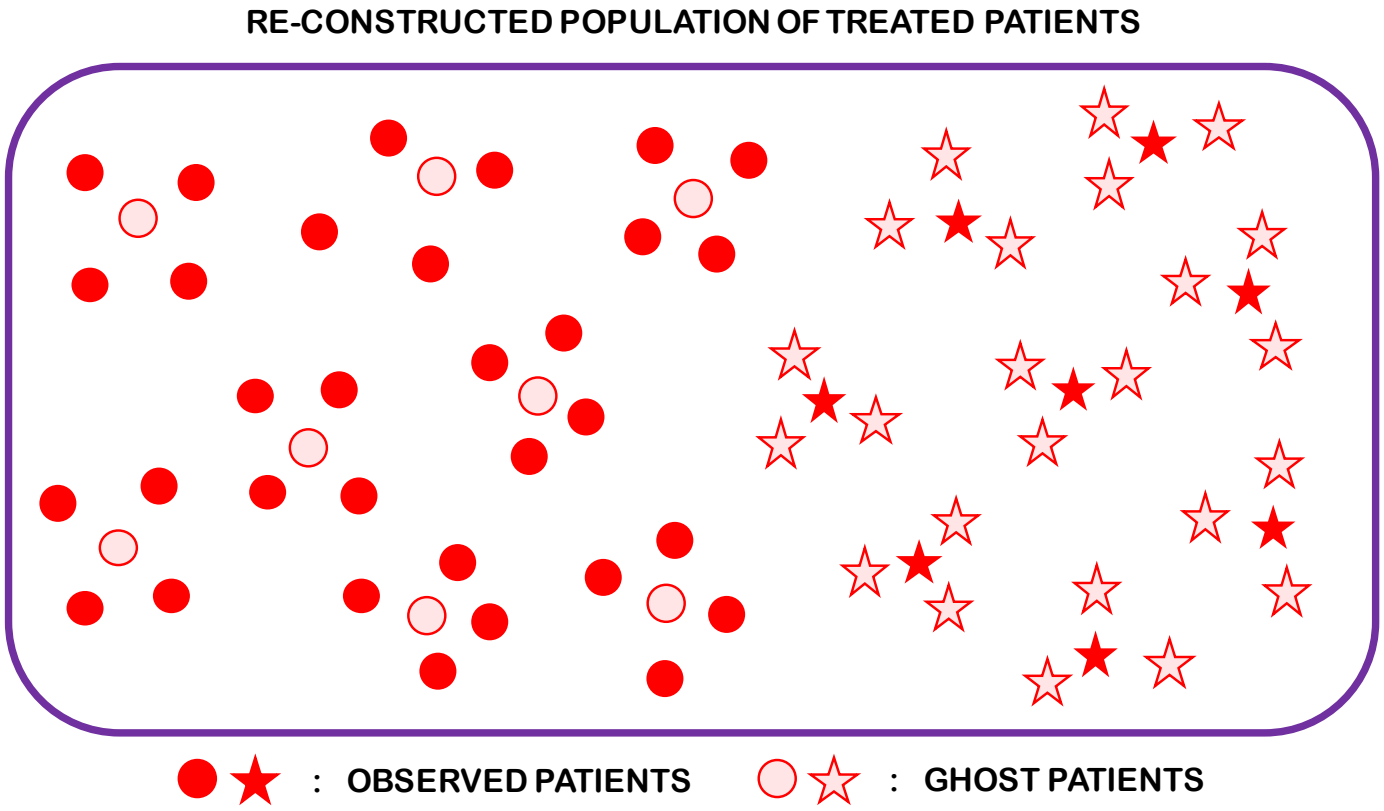
This is simply a weighted average of the outcome of treated patients, reweighted according to their propensity of having been treated in the first place.

If $P(A=1 \mid W=w) = .05$, a patient with $W=w$ had a 5% chance of being treated.

For each such patient treated, approximately 19 similar patients were not. Each treated patient with $W=w$ must stand in for the other 19. This patient has weight

$$\frac{1}{P(A=1 \mid W=w)} \;=\; \frac{1}{0.05} \;=\; 20\,.$$

# The IPTW identification formula




$$P(A = 1 \mid W = \bigstar) = 0.25 \quad P(A = 1 \mid W = \bullet) = 0.80$$

# The IPTW identification formula

The IPTW formula is equivalent to the G-computation formula.

By repeated use of the law of total expectation, we have that

$$
\begin{aligned}
E\left[\frac{I(A=1)Y}{P(A=1 \mid W)}\right] &= E\left[E\left[\frac{I(A=1)Y}{P(A=1 \mid W)}\middle| A, W\right]\right] \\
&= E\left[\frac{I(A=1)}{P(A=1 \mid W)} E\left(Y \mid A, W\right)\right] \\
&= E\left[\frac{I(A=1)}{P(A=1 \mid W)} E\left(Y \mid A=1, W\right)\right] \\
&= E\left[E\left[\frac{I(A=1)}{P(A=1 \mid W)} E\left(Y \mid A=1, W\right)\middle| W\right]\right] \\
&= E\left[\frac{E(Y \mid A=1, W)}{P(A=1 \mid W)} P(A=1 \mid W)\right] = E\left[E(Y \mid A=1, W)\right] .
\end{aligned}
$$

# Estimation based on the G-computation and IPTW formulas

Via the identification formulas, we express quantities we care about in the counterfactual world as quantities defined in the observed data world.

This required certain **causal assumptions**.

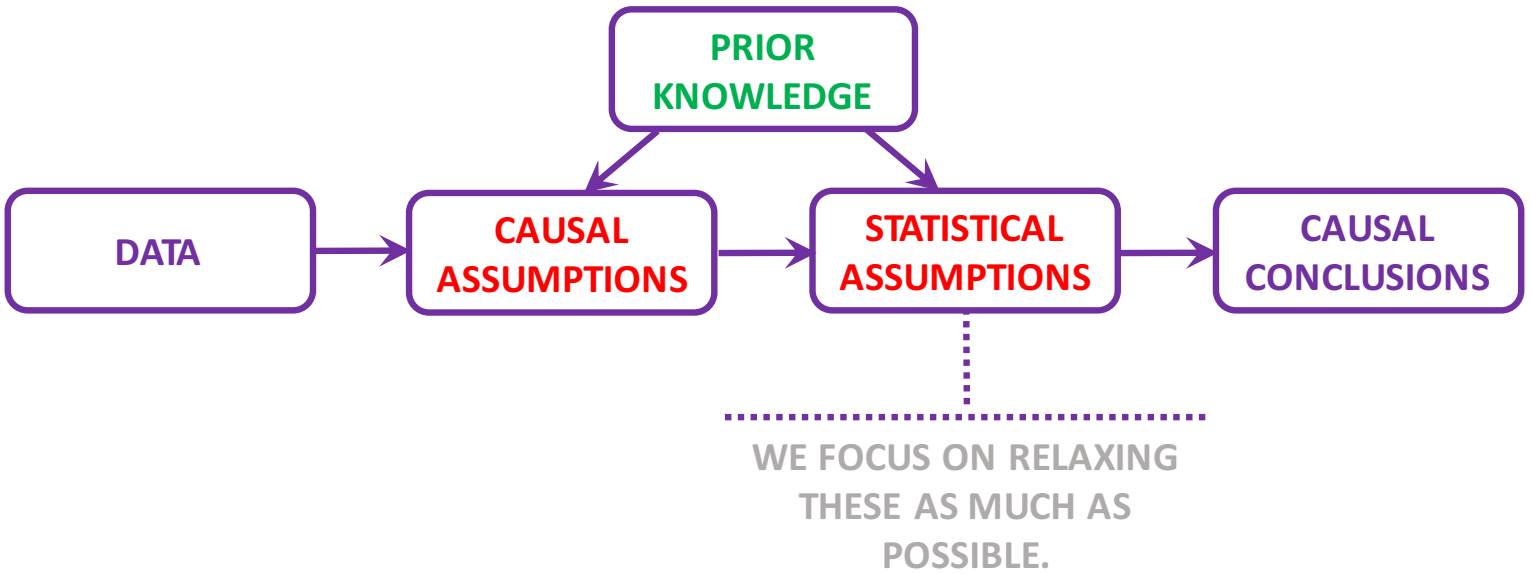
- Many of these are empirically unverifiable, and so cannot be relaxed for free.
- Alternative assumptions exist. Otherwise, partial identification is possible under weaker assumptions. (More on this in Chapter 6.)

This is certainly progress since we can estimate quantities in the observed data world!

Practitioners make **statistical assumptions** of varying degrees to tackle the resulting estimation/inference problem.

- Most of these are verifiable and thus unnecessary (except for convenience).
- The approach we advocate for uses modern statistical learning to reduce the risk of misleading conclusions due to inappropriate statistical assumptions.

# Estimation based on the G-computation and IPTW formulas



# Estimation based on the G-computation and IPTW formulas

Two quantities (defined in the observed data world) play a critical role in nearly all methods for causal inference:

$$\begin{aligned} \text{the outcome regression :} \quad & \bar{Q}(a,w) := E(Y \mid A=a, W=w) \\ \text{the propensity score :} \quad & g(w) := P(A=1 \mid W=w) \quad . \end{aligned}$$

The various methods we will discuss explicitly require estimates of $\bar{Q}$ and/or $g$.

In the following, we will denote by $\bar{Q}_n$ and $g_n$ estimators of $\bar{Q}$ and $g$, respectively.

# Estimation based on the G-computation and IPTW formulas

**Estimation via the G-computation formula $E[\bar{Q}(a, W)]$**

$$\begin{aligned}
\psi_{n,G,1} &:= \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n(1, W_i) \\
\psi_{n,G,0} &:= \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n(0, W_i) \\
\gamma_{n,G} &:= \psi_{n,G,1} - \psi_{n,G,0} = \frac{1}{n}\sum_{i=1}^{n} \left[\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)\right]
\end{aligned}$$

**Estimation via the IPTW formula $E\left[\frac{I(A=a)}{P(A=a \mid W)} Y\right]$**

$$\begin{aligned}
\psi_{n,IPTW,1} &:= \frac{1}{n}\sum_{i=1}^{n} \left[\frac{I(A_i = 1)}{g_n(W_i)}\right] Y_i \\
\psi_{n,IPTW,0} &:= \frac{1}{n}\sum_{i=1}^{n} \left[\frac{I(A_i = 0)}{1 - g_n(W_i)}\right] Y_i \\
\gamma_{n,IPTW} &:= \psi_{n,IPTW,1} - \psi_{n,IPTW,0} = \frac{1}{n}\sum_{i=1}^{n} \left[\frac{I(A_i = 1)}{g_n(W_i)} - \frac{I(A_i = 0)}{1 - g_n(W_i)}\right] Y_i
\end{aligned}$$

# Estimation based on the G-computation and IPTW formulas

**In practice, which of these two approaches should we adopt?**

If outcome regression $\bar{Q}$ is easier to estimate well, G-computation seems like a good bet. If instead propensity score is easier to estimate well, IPTW approach is sensible.

(Note: In reality, we will be able to improve upon both! See Chapter 4.)

In any case, we need to estimate at least one of $\bar{Q}$ or $g$.

There are many approaches possible for estimating a regression function, ranging from very flexible (e.g., nonparametric methods) to rather rigid (e.g., parametric methods).

- (nonparametric) empirical moment, kernel regression, neural networks, random forests;
- (semiparametric) generalized additive models, partially linear additive models;
- (parametric) linear regression, logistic regression, spline regression.

For reasons that will be made clear soon, in this chapter, we will only explicitly mention empirical moment estimators and parametric methods.

# Estimation based on the G-computation and IPTW formulas

When the covariate vector $W$ can only take a few values, the most flexible approach possible consists of using an empirical moment estimator.

$$\begin{aligned}\bar{Q}_n(a,w) &:= \frac{\sum_{i=1}^n Y_i I(A_i = a, W_i = w)}{\sum_{i=1}^n I(A_i = a, W_i = w)} \\ &= \text{observed mean outcome among patients with } A = a \text{ and } W = w\end{aligned}$$

$$\begin{aligned}\bar{g}_n(w) &:= \frac{\sum_{i=1}^n A_i I(W_i = w)}{\sum_{i=1}^n I(W_i = w)} \\ &= \text{observed proportion treated among patients with } W = w\end{aligned}$$

This approach makes no assumption about the regression curves, and does not borrow information across patient types. What is the implication on its **bias** and **variance**?

# Estimation based on the G-computation and IPTW formulas

An example with binary outcome $Y \in \{0, 1\}$ and binary covariate $W$:

| | $W = 1$ | $W = 0$ | |
|---|---|---|---|
| $A = 1$ | $20 + 20$ | $25 + 50$ | 115 |
| $A = 0$ | $6 + 4$ | $14 + 56$ | 80 |
| | 50 | 145 | 195 |

$$\bar{Q}_n(1,1) = \frac{20}{40} = 0.5 \quad \bar{Q}_n(1,0) = \frac{25}{75} = 0.33 \quad \bar{Q}_n(0,1) = \frac{6}{10} = 0.6 \quad \bar{Q}_n(0,0) = \frac{14}{70} = 0.2$$

$$g_n(1) = \frac{40}{50} = 0.8 \quad g_n(0) = \frac{75}{145} = 0.52$$

$$P_n(W = 1) = \frac{50}{195} = 0.256 \quad P_n(W = 0) = \frac{145}{195} = 0.744$$

# Estimation based on the G-computation and IPTW formulas

$$\bar{Q}_n(1,1) = \frac{20}{40} = 0.5 \quad \Big| \quad \bar{Q}_n(1,0) = \frac{25}{75} = 0.33 \quad \Big| \quad \bar{Q}_n(0,1) = \frac{6}{10} = 0.6 \quad \Big| \quad \bar{Q}_n(0,0) = \frac{14}{70} = 0.2$$

$$g_n(1) = \frac{40}{50} = 0.8 \quad \Big| \quad g_n(0) = \frac{75}{145} = 0.52$$

$$P_n(W=1) = \frac{50}{195} = 0.256 \quad \Big| \quad P_n(W=0) = \frac{145}{195} = 0.744$$

---

$$\begin{aligned}
\gamma_{n,G} &= \left[\bar{Q}_n(1,0)P_n(W=0) + \bar{Q}_n(1,1)P_n(W=1)\right] - \left[\bar{Q}_n(0,0)P_n(W=0) + \bar{Q}_n(0,1)P_n(W=1)\right] \\
&= \left[\frac{25}{75}\cdot\frac{145}{195} + \frac{20}{40}\cdot\frac{50}{195}\right] - \left[\frac{14}{70}\cdot\frac{145}{195} + \frac{6}{10}\cdot\frac{50}{195}\right] = 0.376 - 0.303 = 0.073
\end{aligned}$$

$$\begin{aligned}
\gamma_{n,IPTW} &= \left[\frac{P_n(Y=1,A=1,W=0)}{g_n(0)} + \frac{P_n(Y=1,A=1,W=1)}{g_n(1)}\right] \\
&\qquad - \left[\frac{P_n(Y=1,A=0,W=0)}{1-g_n(0)} + \frac{P_n(Y=1,A=0,W=1)}{1-g_n(1)}\right] \\
&= \left[\frac{25}{195}\Big/\frac{75}{145} + \frac{20}{195}\Big/\frac{40}{50}\right] - \left[\frac{14}{195}\Big/\left(1-\frac{75}{145}\right) + \frac{6}{195}\Big/\left(1-\frac{40}{50}\right)\right] \\
&= 0.376 - 0.303 = 0.073
\end{aligned}$$

# Estimation based on the G-computation and IPTW formulas

If $W$ has many discrete components (e.g., medical history variables), a large sample is required to make this empirical approach perform well. If $W$ has a continuous component (e.g., BMI), this approach cannot be used at all.

Information must be borrowed across patient types using regression techniques.

It is common (but not necessarily good) practice to use linear regression for estimating $\bar{Q}$ and logistic regression for estimating $g$.

For example, suppose we are willing to assume that $\bar{Q}(a, w) = \beta_0 + \beta_1 a + \beta_2 w$.

1. Regress $Y$ on $A$ and $W$, yielding estimate $\bar{Q}_n(a, w) = \beta_{0n} + \beta_{1n} a + \beta_{2n} w$.
2. Calculate G-computed quantities $\psi_{n,G,0}$, $\psi_{n,G,1}$ and $\gamma_{n,G}$:

$$\begin{aligned} \psi_{n,G,a} &= \frac{1}{n} \sum_{i=1}^{n} \bar{Q}_n(a, W_i) = \beta_{0n} + \beta_{1n} a + \beta_{2n} \overline{W}_n \\ \gamma_{n,G} &= \psi_{n,G,1} - \psi_{n,G,0} = \beta_{1n} \end{aligned}$$

How convenient! Is this observation useful, though???

# Estimation based on the G-computation and IPTW formulas

```
#-------------------------------------------------------------
# all code snippets assume that you have:
# n = a numeric indicating sample size
# W = an n-row data.frame of covariates
# A = an n-length vector of binary treatment assignments
# Y = an n-length vector of binary or continuous outcomes
#-------------------------------------------------------------

# fit a glm regressing Y onto functions of A and W
# here, we use a main terms linear regression
fit_or <- glm(Y ~ ., data = data.frame(A,W))

# predict on data setting A=1
Qbar1W <- predict(fit_or, newdata = data.frame(W,A=1,Y))

# predict on data setting A=0
Qbar0W <- predict(fit_or, newdata = data.frame(W,A=0,Y))

# take means
psi_nG1 <- mean(Qbar1W)
psi_nG0 <- mean(Qbar0W)

# average treatment effect
gamma_nG <- psi_nG1 - psi_nG0
```

# Estimation based on the G-computation and IPTW formulas

As another example, suppose we are willing to assume that $g(w) = \text{expit}(\alpha_0 + \alpha_1 w)$.

1. Perform a logistic regression of $A$ on $W$, yielding estimate
$$g_n(w) = \text{expit}(\alpha_{0n} + \alpha_{1n} w) \ .$$
2. Calculate IPTW estimates $\psi_{n,IPTW,0}$, $\psi_{n,IPTW,1}$ and $\gamma_{n,IPTW}$:

$$\begin{aligned}
\psi_{n,IPTW,0} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I(A_i=0)}{1-g_n(W_i)}\right]Y_i = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I(A_i=0)}{1-\text{expit}(\alpha_{0n}+\alpha_{1n}W_i)}\right]Y_i \\
\psi_{n,IPTW,1} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I(A_i=1)}{g_n(W_i)}\right]Y_i = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{I(A_i=1)}{\text{expit}(\alpha_{0n}+\alpha_{1n}W_i)}\right]Y_i \\
\gamma_{n,IPTW} &= \psi_{n,IPTW,1} - \psi_{n,IPTW,0}
\end{aligned}$$

# Estimation based on the G-computation and IPTW formulas

```
# fit a glm for the propensity score regression A on W
# here we fit a main terms logistic regression
fit_ps <- glm(A ~ ., data=data.frame(W), family = binomial())

# probability of receiving treatment
g1W <- predict(fit_ps, newdata = data.frame(W), type = "response")

# take means
psi_nIPTW1 <- mean(as.numeric(A==1)/g1W * Y)
psi_nIPTW0 <- mean(as.numeric(A==0)/(1-g1W) * Y)

# average treatment effect
gamma_nIPTW <- psi_nIPTW1 - psi_nIPTW0
```

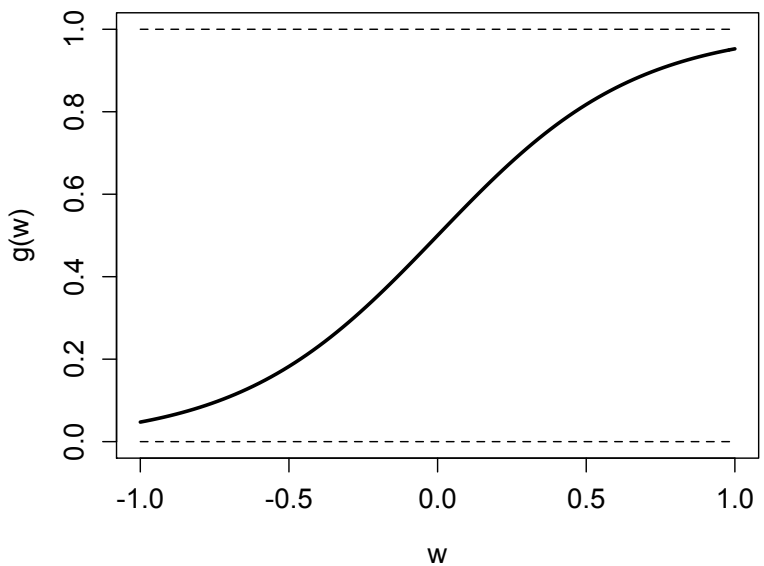
# Estimation based on the G-computation and IPTW formulas

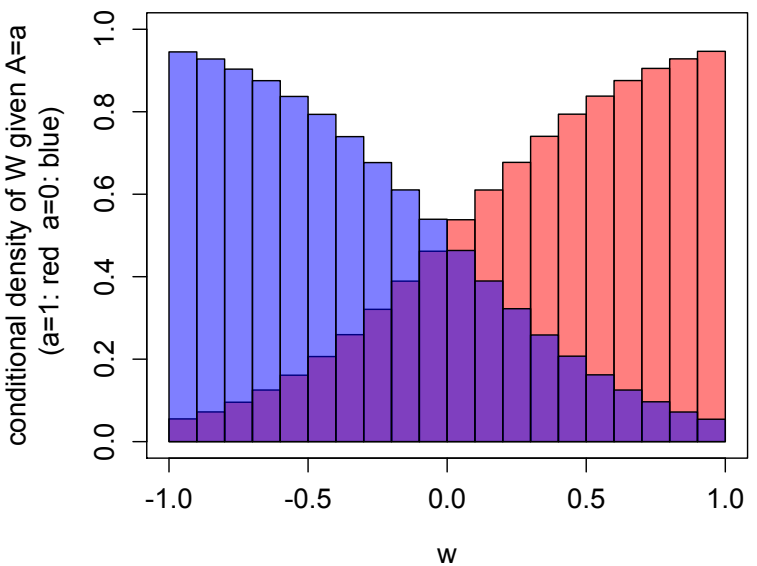
**An illustration with a single, continuous confounder $W$**

Every single data unit $O = (Y, A, W)$ was generated independently as follows:

1. generate $W \sim U(-1, +1)$, a uniform random variable over $(-1, +1)$;
2. given $W = w$, generate $A \sim \text{Bernoulli}(g(w))$ with $g(w) = \text{expit}(3w)$;
3. given $W = w$ and $A = a$, generate $Y \sim N(\bar{Q}(a, w), 1)$ with $\bar{Q}(a, w) = 1 + a - w$.

Here, the true ATE is $\gamma = 1$ while $E(Y \mid A = 1) - E(Y \mid A = 0) = 1.64 - 1.36 = 0.28$.



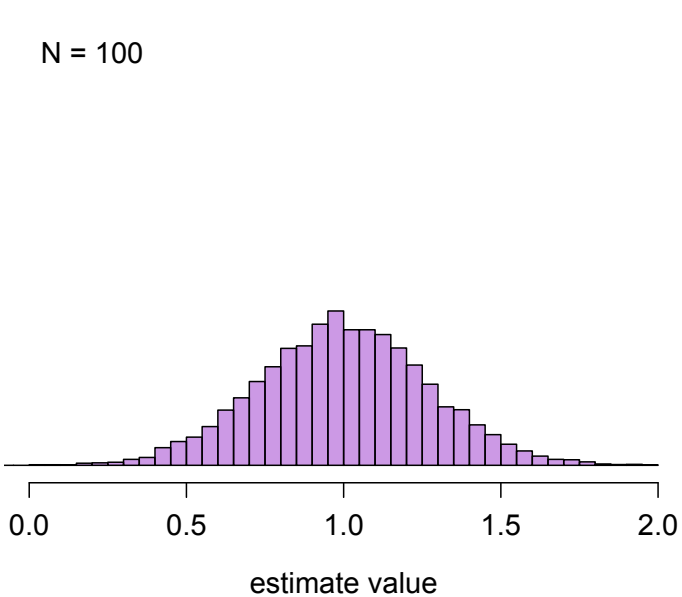



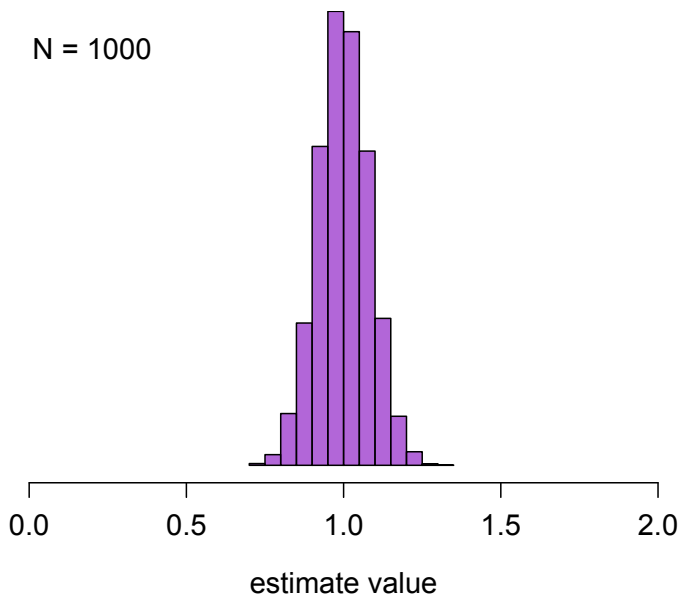
# Estimation based on the G-computation and IPTW formulas

We consider ATE estimation via G-computation using the linear regression model

$$\bar{Q}(a, w) + \beta_0 + \beta_1 a + \beta_2 w + \beta_3 aw + \beta_4 w^2 \ .$$

Below are histograms of the sampling distribution of $\gamma_{n,G}$ for $n \in \{100, 1000\}$.



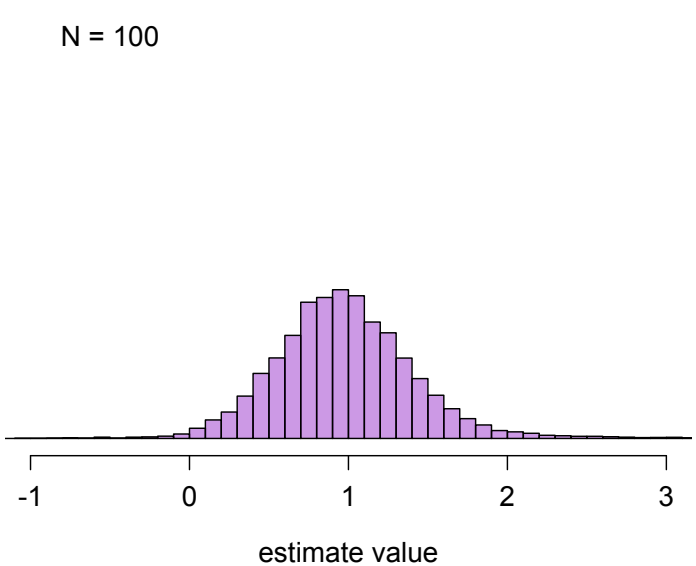



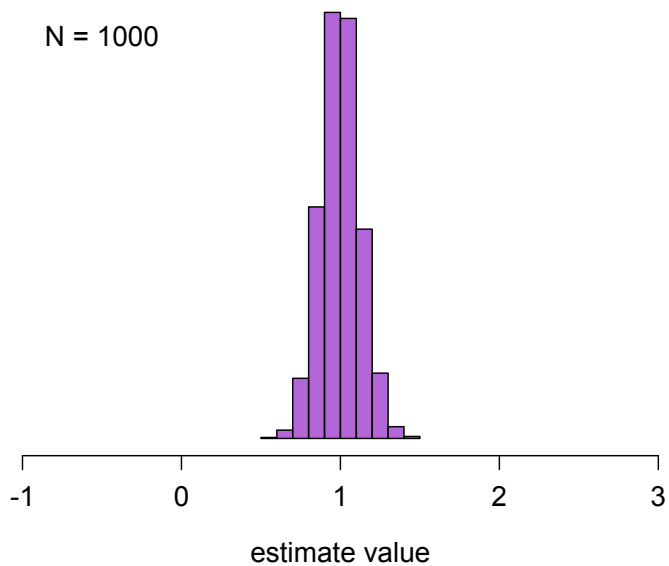
# Estimation based on the G-computation and IPTW formulas

Here, we instead estimate the ATE via IPTW using the logistic regression model

$$g(w) = \alpha_0 + \alpha_1 w \ .$$

Below are histograms of the sampling distribution of $\gamma_{n,IPTW}$ for $n \in \{100, 1000\}$.



# Estimation based on the G-computation and IPTW formulas

So far, we have discussed how to obtain estimates of the ATE via the G-computation or IPTW formulas. What about inference (i.e., confidence intervals, $p$-values)?

Formulas for the variance of the (empirical or parametric) G-computation and IPTW estimators exist but are complex. We may use the bootstrap though.

Say we wish to construct a 95% CI for the ATE using the estimator $\gamma_n$.

1. Draw $O_1^*, O_2^*, \ldots, O_n^*$ from the original data with replacement, and compute $\gamma_n^\#$.
2. Repeat a total of $M$ times to obtain bootstrapped estimates $\gamma_n^{\#,1}, \gamma_n^{\#,2}, \ldots, \gamma_n^{\#,M}$.
3. Compute the empirical standard error $\sigma_n^\#$ of these bootstrapped estimates.
4. Compute the empirical $\alpha$-quantile $\gamma_n^\#(\alpha)$ of the bootstrapped estimates.
5. An approximate 95% CI for the ATE is given by $(\gamma_n^\#(0.025),\ \gamma_n^\#(0.975))$ and a $p$-value of the null hypothesis $H_0 : \gamma = 0$ versus $H_1 : \gamma \neq 0$ can be obtained as

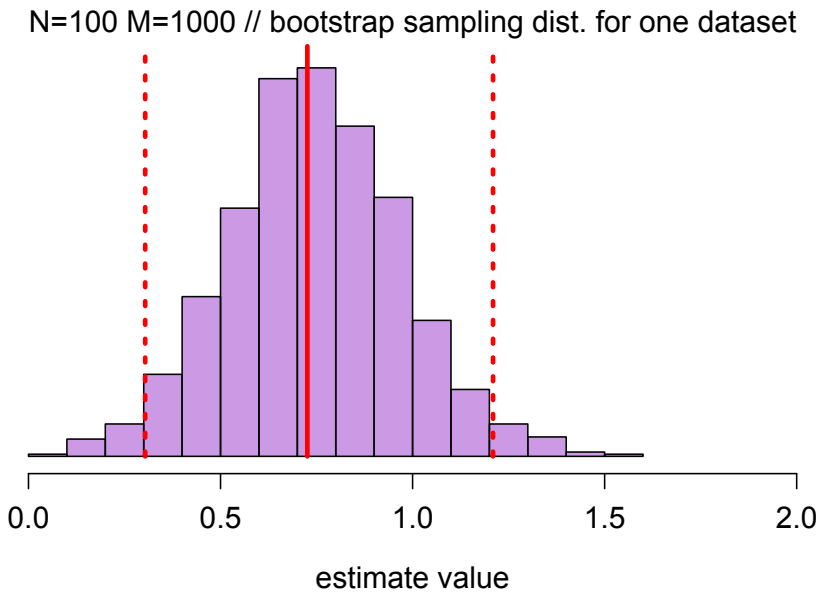
$$p = 2\left[1 - \Phi\left(\frac{|\gamma_n|}{\sigma_n^\#}\right)\right],$$

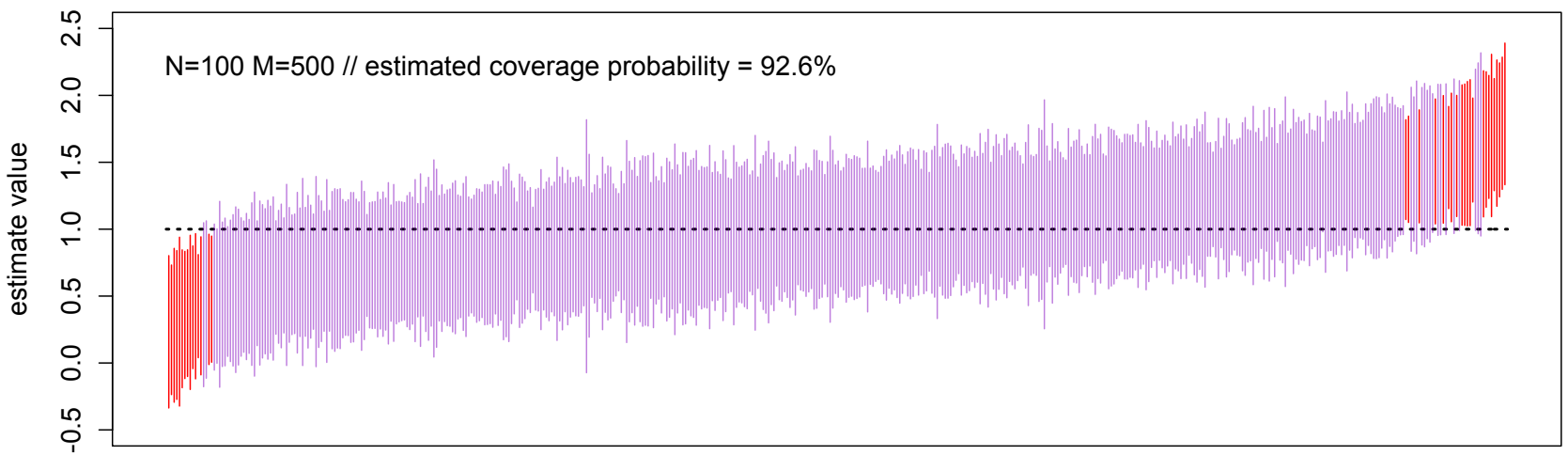
where $\Phi$ is the distribution function of the standard normal distribution.

# Estimation based on the G-computation and IPTW formulas

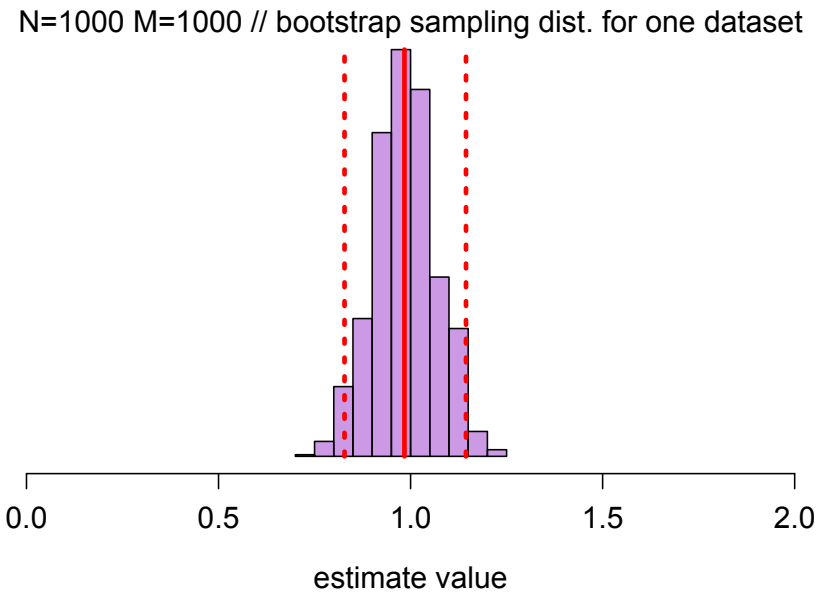
```
# bootstrap 500 samples
M <- 500

# use replicate to generate estimates
gammaVec <- replicate(M, {
  # randomly sample obs. with replacement
  ind <- sample(1:n, replace = TRUE)
  # compute gcomp estimator in resampled data
  fit_or <- glm(Y[ind] ~ ., data=data.frame(A,W)[ind,])
  Qbar1W <- predict(fit_or, newdata = data.frame(A=1,W[ind,]))
  Qbar0W <- predict(fit_or, newdata = data.frame(A=0,W[ind,]))
  psi_nG1 <- mean(Qbar1W)
  psi_nG0 <- mean(Qbar0W)
  gamma_nG <- psi_nG1 - psi_nG0
  return(gamma_nG)
})

# percentile confidence interval
quantile(gammaVec, c(0.025, 0.975))
```

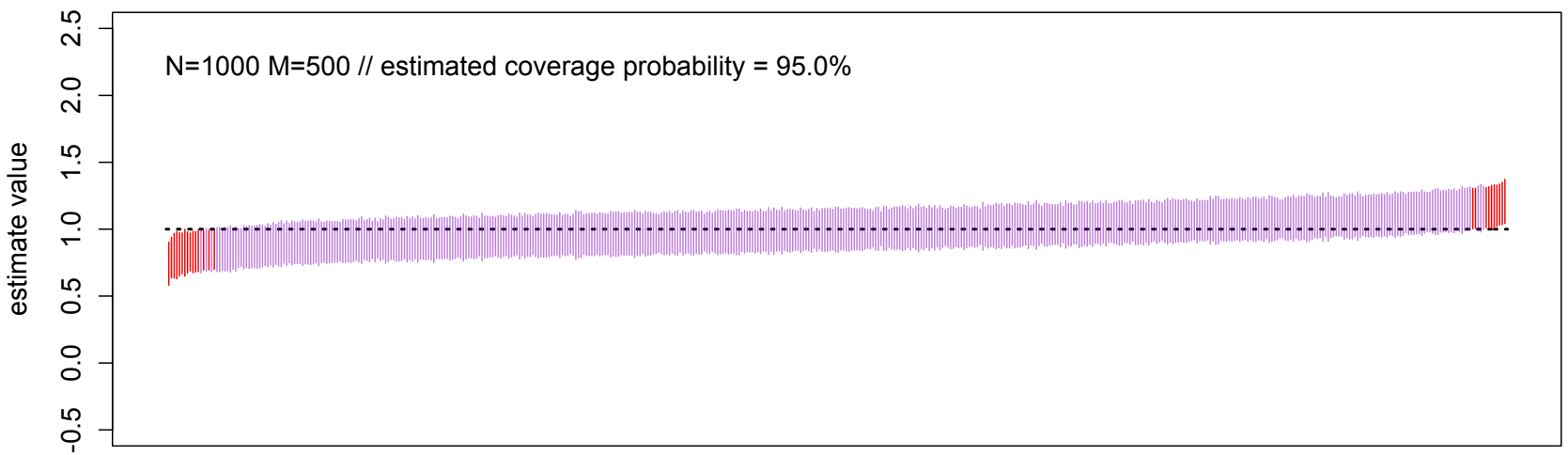
# Estimation based on the G-computation and IPTW formulas



# Estimation based on the G-computation and IPTW formulas



# Estimation based on the G-computation and IPTW formulas

We analyzed data from the BOLD study using the IPTW formula and estimation of the propensity using logistic regression (with main terms only).

Average counterfactual score corresponding to early imaging intervention:

estimate = 8.20, 95% CI: (7.82, 8.51)

Average counterfactual score corresponding to control (no early imaging):

estimate = 8.59, 95% CI: (8.31, 8.81)

Average treatment effect comparing early imaging to control:

estimate = -0.39, 95% CI: (-0.81, -0.09), $p = 0.03$

Based on these results, we would conclude that obtaining early imaging appears to lower disability scores on average at the 12-month mark.

# Estimation based on the G-computation and IPTW formulas

We analyzed data from the BOLD study using the G-computation formula and estimation of the outcome regression using linear regression (with main terms only).

Average counterfactual score corresponding to early imaging intervention:

estimate = 8.07, 95% CI: (7.77, 8.34)

Average counterfactual score corresponding to control (no early imaging):

estimate = 8.56, 95% CI: (8.34, 8.80)

Average treatment effect comparing early imaging to control:

estimate = -0.50, 95% CI: (-0.85, -0.19), $p = 0.005$

Again, based on these results, we would conclude that obtaining early imaging appears to lower disability scores on average at the 12-month mark.

# Estimation based on the G-computation and IPTW formulas

We analyzed data from the BOLD study using the G-computation formula and estimation of the outcome regression using a GLM with logarithmic link (with main terms only) since the outcomes are non-negative.

Average counterfactual score corresponding to early imaging intervention:

estimate = 8.28, 95% CI: (7.96, 8.56)

Average counterfactual score corresponding to control (no early imaging):

estimate = 8.46, 95% CI: (8.26, 8.69)

Average treatment effect comparing early imaging to control:

estimate = -0.18, 95% CI: (-0.55, 0.14), $p = 0.31$

Based on these results, we would conclude that obtaining early imaging does not appear to lower disability scores on average at the 12-month mark.

# Matching to achieve balance

The identification formulas described so far – especially the G-computation formula – form the basis of the more complex approaches we will advocate for in Chapter 4.

Nevertheless, in practice, the most common (but not necessarily best) approach to causal inference is via **matching**.

Randomization is used to ensure that the treated and controls are comparable groups.

The idea of matching is simple:

we may use the available pool of patients
to reconstitute comparable groups of treated and controls.

Seems sensible! Of course, the devil is in the details...

# Matching to achieve balance

There are many different ways of performing matching. Below are some of the key questions that define a particular implementation. (Stuart, 2010)

- Who is the reference group to match to?
  - Find controls for each treated patient? This leads to the ATT!!!
  - Also find similar treated patients for each control?
- How is the similarity between patients adjudicated?
  - Exact matching? Nearest neighbors?
  - How is closeness measured?
  - Based on entire set of covariates? On the propensity score alone?
- How many matches should be selected, and can matched patients be reused?
  - Number of matches used determines bias/variance trade-off.
  - If matches are reused, how is this accounted for?
- How is estimate obtained and how is inference performed?
  - Unadjusted mean difference? Model-based adjustment?
  - Bootstrap inference? How exactly? What theory justifies this?

# Matching to achieve balance

**Some pros and cons of resorting to matching:**

- ⊕ Crux of the approach can easily be explained to non-statisticians.
- ⊕ The quality of the reconstructed groups can be scrutinized and depicted easily.

- ⊖ Easy to use, easy to misuse!
- ⊖ Can result in quite an inefficient use of the data.
- ⊖ There is often confusion about the estimand being targeted (i.e., ATT vs ATE).
- ⊖ Valid inference is very difficult to perform because of the complex dependence induced by matching process.
- ⊖ Many choices to make without clear guidelines: implementation is often an art, which is not amenable to rigorous inference and replicability of findings.

# Key points of Chapter 2

- We must link the counterfactual and observable worlds since we only have data on the observable world yet want to study the counterfactual world.
- G-computation and IPTW identification formulas are distinct but equivalent ways to write mean counterfactual outcomes in terms of the observed data distribution.
- G-computation is based on the outcome regression, while IPTW heavily relies on the propensity score.
- These formulas suggest ways of constructing simple (empirical or model-based) estimators, and inference can be carried out via bootstrapping.
- IPTW estimators behave poorly when the positivity condition is nearly violated.
- Matching is very popular in practice, but it has important shortcomings.

# References and additional reading

**References:**

Horvitz, D, Thompson, D (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260)663–685. doi: 10.1080/01621459.1952.10483446.

Robins, JM (1986). A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9–12)1393-1512. doi: 10.1016/0270-0255(86)90088-6.

Robins, JM, Hernan, MA, Brumback, B (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5)550-560. 10.1097/00001648-200009000-00011.

Stuart, EA (2010). Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25(1), 1-21. doi: 10.1214/09-STS313.

**Additional reading:**

Snowden, JM, Rose, S, Mortimer, KM (2011). Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7)731-738. doi: 10.1093/aje/kwq472.