Doubly robust treatment effect estimation with missing attributes

Effect of tranexamic acid on mortality of patients with traumatic brain injury

Imke Mayer, Julie Josse, Stefan Wager, Tobias Gauss, Jean-Denis Moyer
EHESS; École Polytechnique; Stanford Business School; Traumabase® Group
Statistique, Mathématique et Applications, Fréjus, 3 sept. 2019
Introduction
Traumabase

- 20,000 patients
- 250 continuous and categorical variables: heterogeneous
- 16 hospitals: multilevel data
- 4,000 new patients/year

<table>
<thead>
<tr>
<th>Center</th>
<th>Accident</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>Lactates</th>
<th>BP</th>
<th>shock</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaujon</td>
<td>fall</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NM</td>
<td>180</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Pitie</td>
<td>gun</td>
<td>26</td>
<td>m</td>
<td>NR</td>
<td>NA</td>
<td>131</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Beaujon</td>
<td>moto</td>
<td>63</td>
<td>m</td>
<td>80</td>
<td>3.9</td>
<td>145</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Pitie</td>
<td>moto</td>
<td>30</td>
<td>w</td>
<td>NR</td>
<td>Imp</td>
<td>107</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>HEGP</td>
<td>knife</td>
<td>16</td>
<td>m</td>
<td>98</td>
<td>2.5</td>
<td>118</td>
<td>no</td>
<td></td>
</tr>
</tbody>
</table>

⇒ Estimate causal effect: Administration of the treatment "tranexamic acid" (within 3 hours after the accident) on the outcome mortality for traumatic brain injury patients
Traumabase

- 20,000 patients
- 250 continuous and categorical variables: heterogeneous
- 16 hospitals: multilevel data
- 4,000 new patients/year

<table>
<thead>
<tr>
<th>Center</th>
<th>Accident</th>
<th>Age</th>
<th>Sex</th>
<th>Weight</th>
<th>Lactactes</th>
<th>BP</th>
<th>shock</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beaujon</td>
<td>fall</td>
<td>54</td>
<td>m</td>
<td>85</td>
<td>NM</td>
<td>180</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Pitie</td>
<td>gun</td>
<td>26</td>
<td>m</td>
<td>NR</td>
<td>NA</td>
<td>131</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Beaujon</td>
<td>moto</td>
<td>63</td>
<td>m</td>
<td>80</td>
<td>3.9</td>
<td>145</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Pitie</td>
<td>moto</td>
<td>30</td>
<td>w</td>
<td>NR</td>
<td>Imp</td>
<td>107</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>HEGP</td>
<td>knife</td>
<td>16</td>
<td>m</td>
<td>98</td>
<td>2.5</td>
<td>118</td>
<td>no</td>
<td></td>
</tr>
</tbody>
</table>

⇒ **Estimate causal effect**: Administration of the treatment "tranexamic acid" (within 3 hours after the accident) on the outcome mortality for traumatic brain injury patients
Missing values

- NA
- Not Informed
- Not made
- Not Applicable
- Impossible

Percentage of missing values

0
25
50
75
100

AIS.external
AIS.face
AIS.head
ISS
TBI
Trauma.center
Pupil.anomaly
OTI.MICU
Pupil.anomaly.ph
HR
GSC.init
Cardiac.arrest.ph
Delta.hemoCue
IGS.II
Vasopressor.therapy
Hemoglobin
SBP
DBP
Death.in.ICU
SpO2
Anticoagulant.therapy
SBP.min
HR.max
DBP.min
SpO2.min
GSC.motor.init
Neurosurgery.day0
Medicare.time.ph
Transaminase
Hemorrhage
Colloid.volume
ICP
Decompressive.craniectomy
SpO2.phase1
SBP.MICU
HR.MICU
DBP.MICU
Glasgow.discharge
OSM.CP
Temperature.min
Improv.anomaly.osmo
Cause.of.death
Temperature.micu
Causal inference: classical framework
Potential outcome framework (Neyman, 1923, Rubin, 1974)

### Causal effect

Binary treatment \( w \in \{0, 1\} \) on \( i-th \) individual with potential outcomes \( Y_i(1) \) and \( Y_i(0) \). Individual causal effect of the treatment:

\[
\Delta_i = Y_i(1) - Y_i(0)
\]

- Problem: \( \Delta_i \) never observed (only observe one outcome/indiv).

Causal inference as a missing value pb?

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Treatment</th>
<th>Outcome(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 )</td>
<td>( X_2 )</td>
<td>( X_3 )</td>
</tr>
<tr>
<td>1.1</td>
<td>20</td>
<td>F</td>
</tr>
<tr>
<td>-6</td>
<td>45</td>
<td>F</td>
</tr>
<tr>
<td>0</td>
<td>15</td>
<td>M</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>-2</td>
<td>52</td>
<td>M</td>
</tr>
</tbody>
</table>
Potential outcome framework (Neyman, 1923, Rubin, 1974)

**Causal effect**

Binary treatment $w \in \{0, 1\}$ on $i$-th individual with potential outcomes $Y_i(1)$ and $Y_i(0)$. Individual causal effect of the treatment:

$$\Delta_i = Y_i(1) - Y_i(0)$$

- Problem: $\Delta_i$ never observed (only observe one outcome/indiv).
  Causal inference as a missing value pb?

- **Average treatment effect (ATE)** $\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$: The ATE is the difference of the average outcome had everyone gotten treated and the average outcome had nobody gotten treatment.

$\Rightarrow$ First solution: estimate $\tau$ with randomized controlled trials (RCT).
Observational data

Non random assignment $\rightarrow$ Confounding

Mortality rate 20% - treated 38% - not treated 16%: treatment kills?

|                        | survived  | deceased | Pr(survived | treatment) | Pr(deceased | treatment) |
|------------------------|-----------|----------|-------------|-------------|-------------|
| TA not administered    | 2,167 (68%) | 399 (13%) | 0.84        | 0.16        |
| TA administered        | 374 (12%)  | 228 (7%)  | 0.62        | 0.38        |

**Table 1:** Occurrence and frequency table for traumatic brain injury patients (total number: 3,168).
Unconfoundedness and the propensity score

Assumptions

- \( n \) iid samples \((X_i, Y_i, W_i),\)
- \( Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0) \) (SUTVA)
- Treatment assignment is random conditionally on \( X_i \):
  \( \{Y_i(0), Y_i(1)\} \perp \perp W_i \mid X_i \equiv \text{unconfoundedness} \) assumption.

Propensity score and overlap assumption

\[
e(x) \triangleq P(W_i = 1 \mid X_i = x) \quad \forall x \in \mathcal{X}.
\]

We will assume overlap, i.e. \( 0 < e(x) < 1 \) \( \forall x \in \mathcal{X} \).

Key property

\( e \) is a balancing score, i.e. under unconfoundedness, it satisfies

\[
\{Y_i(0), Y_i(1)\} \perp \perp W_i \mid e(X_i)
\]
Propensity based estimators

**Inverse Propensity Weighted estimator**

\[
\hat{\tau}_{IPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left( \frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)} \right)
\]

⇒ Balance the differences between the two groups
Propensity based estimators

Inverse Propensity Weighted estimator

\[ \hat{\tau}_{IPW} \triangleq \frac{1}{n} \sum_{i=1}^{n} \left( \frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1 - W_i) Y_i}{1 - \hat{e}(X_i)} \right) \]

⇒ Balance the differences between the two groups

Augmented IPW: a doubly robust estimator

Define \( \mu(w)(x) := \mathbb{E}[Y_i(w) | X_i = x] \).

\[ \hat{\tau}_{AIPW} := \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\mu}(1)(X_i) - \hat{\mu}(0)(X_i) + W_i \frac{Y_i - \hat{\mu}(1)(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}(0)(X_i)}{1 - \hat{e}(X_i)} \right) \]

is consistent if either the \( \hat{\mu}(w)(x) \) are consistent or \( \hat{e}(x) \) is consistent.

⇒ Possibility to use any (machine learning) procedure such as random forests, deep nets, etc. to estimate \( \hat{e}(x) \) and \( \hat{\mu}(w)(x) \) without harming the interpretability of the causal effect estimation. R package grf (Athey et al., 2019)
Doubly robust ATE estimation

Define $\mu(w)(x) := \mathbb{E}[Y_i(w) \mid X_i = x]$ and $e(x) := \mathbb{P}(W_i = 1 \mid X_i = x)$.

**Augmented IPW: a doubly robust estimator**

$$\hat{\tau}_{AIPW} := \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + W_i \frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)} \right)$$

is consistent if either the $\hat{\mu}(w)(x)$ are consistent or $\hat{e}(x)$ is consistent.
Define $\mu(w)(x) := \mathbb{E}[Y_i(w) \mid X_i = x]$ and $e(x) := \mathbb{P}(W_i = 1 \mid X_i = x)$.

**Augmented IPW: a doubly robust estimator**

$$\hat{\tau}_{AIPW} := \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + W_i \frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)} \right)$$

is consistent if either the $\hat{\mu}(w)(x)$ are consistent or $\hat{e}(x)$ is consistent.

**Remark 1:** Possibility to use any (machine learning) procedure such as random forests, deep nets, etc. to estimate $\hat{e}(x)$ and $\hat{\mu}(w)(x)$ without harming the interpretability of the causal effect estimation.

**Remark 2:** Cross-splitting is necessary for learning $\hat{\mu}(w)(x)$ and $\hat{e}(x)$ (Chernozhukov et al., 2018). R package grf (Athey et al., 2019).
Efficient score estimator

Given unconfoundedness \( \{Y_i(1), Y_i(0)\} \perp W_i | X_i \) but no further parametric assumptions on \( \mu(w)(x) \) and \( e(x) \), the previously attained asymptotic variance,

\[ V^* := \text{Var}(\tau(X)) + \mathbb{E} \left[ \frac{\sigma^2(X)}{e(X)(1 - e(X))} \right], \]

is optimal and any estimator \( \tau^* \) that attains it is asymptotically equivalent to \( \hat{\tau}_{AIPW}^* \).

\( V^* \) is the semiparametric efficient variance for ATE estimation.

See Robins et al. (1994); Chernozhukov et al. (2018) for more details.
Causal inference: with missing attributes?
Without any changes to the previous framework, the only straightforward – but generally biased – solution is complete-case analysis.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Treatment</th>
<th>Outcome(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Y(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Y(1)</td>
</tr>
<tr>
<td>X_1</td>
<td>X_2</td>
<td>X_3</td>
</tr>
<tr>
<td>NA</td>
<td>20</td>
<td>F</td>
</tr>
<tr>
<td>-6</td>
<td>45</td>
<td>NA</td>
</tr>
<tr>
<td>0</td>
<td>NA</td>
<td>M</td>
</tr>
<tr>
<td>NA</td>
<td>32</td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td>63</td>
<td>M</td>
</tr>
<tr>
<td>-2</td>
<td>NA</td>
<td>M</td>
</tr>
</tbody>
</table>
Without any changes to the previous framework, the only straightforward – but generally biased – solution is complete-case analysis.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Treatment</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$X_2$</td>
<td>$X_3$</td>
</tr>
<tr>
<td>NA</td>
<td>20</td>
<td>F</td>
</tr>
<tr>
<td>-6</td>
<td>45</td>
<td>NA</td>
</tr>
<tr>
<td>0</td>
<td>NA</td>
<td>M</td>
</tr>
<tr>
<td>NA</td>
<td>32</td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td>63</td>
<td>M</td>
</tr>
<tr>
<td>-2</td>
<td>NA</td>
<td>M</td>
</tr>
</tbody>
</table>
Unconfoundedness with missing attributes?

Without any changes to the previous framework, the only straightforward – but generally biased – solution is complete-case analysis.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Treatment</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$X_2$</td>
<td>$X_3$</td>
</tr>
<tr>
<td>NA</td>
<td>20</td>
<td>F</td>
</tr>
<tr>
<td>6</td>
<td>45</td>
<td>NA</td>
</tr>
<tr>
<td>0</td>
<td>NA</td>
<td>M</td>
</tr>
<tr>
<td>NA</td>
<td>32</td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td>63</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>NA</td>
<td>M</td>
</tr>
<tr>
<td>W</td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>T</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>T</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>F</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>T</td>
</tr>
</tbody>
</table>
Without any changes to the previous framework, the only straightforward – but generally biased – solution is complete-case analysis.

→ Often not a good idea! What are the alternatives?

**Two families of methods**

- Unconfoundedness despite missingness
- Classical missing values mechanisms
Unconfoundedness despite missingness

Adapt the initial assumptions s.t. treatment assignment is unconfounded given only the observed covariates and the response pattern.
Unconfoundedness with missing attributes?

Unconfoundedness despite missingness

Treatment is unconfounded given \( X^* \):

\[
\{ Y_i(1), Y_i(0) \} \perp \perp W_i \mid X^*,
\]

(1)

or alternatively:

\[
\{ Y_i(1), Y_i(0) \} \perp \perp W_i \mid X_i, R_i,
\]

\[
\begin{cases}
\text{CIT:} & W_i \perp \perp X_i \mid X^*_i, R_i \\
\text{or} & \text{CIO:} \quad Y_i(t) \perp \perp X_i \mid X^*_i, R_i \quad \text{for } t \in \{0, 1\}
\end{cases}
\]

(2)

Notations

- response pattern \( R \in \{NA, 1\}^p \), \( R_j \triangleq 1\{X_j \text{ is observed}\} + NA \ 1\{X_j \text{ is missing}\} \)
- \( X^* = R \odot X \in \{\mathbb{R} \cup NA\}^p \)
Unconfoundedness with missing attributes?

**Unconfoundedness despite missingness**

Treatment is unconfounded given \( X^* \):

\[
\{ Y_i(1), Y_i(0) \} \perp \perp W_i \mid X^*, \quad (1)
\]

or alternatively:

\[
\{ Y_i(1), Y_i(0) \} \perp \perp W_i \mid X_i, R_i,
\]

\[
\begin{cases}
\text{CIT:} & W_i \perp \perp X_i \mid X_i^*, R_i \\
\text{or} & \\
\text{CIO:} & Y_i(t) \perp \perp X_i \mid X_i^*, R_i \quad \text{for } t \in \{0, 1\}
\end{cases}
\quad \quad (2)
\]

---

**Notations**

- \( \mathbf{R} \in \{ \text{NA}, 1 \}^p \), \( R_j = 1 \) if \( X_j \) is observed, and \( 1 \) otherwise.
- \( X^* = \mathbf{R} \odot X \in \{ \mathbf{R} \cup \text{NA} \}^p \)
Generalized propensity score (Rosenbaum and Rubin, 1984)

\[ e^*(X^*) = P(W = 1 | X^*). \]

→ Allows to balance treatment and control groups on the observed information \( X^* \) in the case of missing values (1).

→ Estimate \( e^* \) using missingness pattern approach (MPA), i.e. estimate one propensity model per pattern.
   → Often impossible in practice if \( p \) (moderately) large (w.r.t. \( n \)).
Generalized propensity score (Rosenbaum and Rubin, 1984)

\[ e^*(X^*) = \mathbb{P}(W = 1 \mid X^*). \]

→ Allows to balance treatment and control groups on the observed information \( X^* \) in the case of missing values (1).

→ Estimate \( e^* \) using missingness pattern approach (MPA), i.e. estimate one propensity model per pattern.
   → Often impossible in practice if \( p \) (moderately) large (w.r.t. \( n \)).

→ Random forests allow incorporating missing values directly since they allow semi-discrete variables (e.g. \( X^* \in (\mathbb{R} \times \text{NA})^p \)).

→ With specific representation/encoding of missing values (MIA), splits are possible either on observed variables or on response pattern (Josse et al., 2019).

→ This procedure targets the Bayes estimate:

\[
\mathbb{E}[W \mid X^*] = \sum_{r \in \{0,1\}^p} \mathbb{E}[W \mid X^*, R = r] 1_{R=r}.
\]
Assumption on missing values mechanism
Assume MAR mechanism and do multiple imputation using \((X, W, Y)\) (Hill (2004); Seaman and White (2014); Leyrat et al. (2019)).
Problem: can we use \(Y\) for propensity score estimation?
What happens with MNAR missing values?
Simulations: importance of unconfoundedness despite missingness and performance of $\hat{\tau}_{\text{mia}}$

**Setup**

- Different data generating models (linear, nonlinear, latent, etc.)
- Different missingness mechanisms

**Figure 2:** Estimated and true average treatment effect ($\tau = 1$, MCAR)
Simulations: importance of unconfoundedness despite missingness and performance of $\hat{\tau}_{mia}$

**Setup**
- Different data generating models (linear, nonlinear, latent, etc.)
- Different missingness mechanisms

**Figure 2:** Estimated and true average treatment effect ($\tau = 1$, MNAR)
Simulations: importance of unconfoundedness despite missingness and performance of $\hat{\tau}_{mia}$

**Setup**
- Different data generating models (linear, nonlinear, latent, etc.)
- Different missingness mechanisms

**Results**
- AIPW estimators perform better than their IPW counterparts.
- For $\hat{\tau}_{mia}$, the *unconfoundedness despite missingness* is indeed necessary.
- $\hat{\tau}_{mia}$ unbiased for all missingness mechanisms, especially for MNAR.
- Multiple imputation (mice) only requires standard unconfoundedness, but needs MAR
Application: Traumabase
Plausibility of underlying assumptions with Traumabase data

- Overlap: cannot be tested but high level of uncertainty at diagnosing severe (internal bleeding) makes it likely
- Unconfoundedness despite missingness: seems plausible (physicians decide based on what they observe+record)
- Many variables have missing non at random data.
40 covariates, 13 confounders. 3,168 patients.

ATE estimations ($\times 100$) for the effect of tranexamic acid on in-ICU mortality for TBI patients. Imputations on all patients (TBI + no-TBI).

(y-axis: estimation approach, solid: DR, dotted: IPW, turquoise: without adjustment), (x-axis: ATE estimation with bootstrap CI)

We compute the mortality rate in the treated group and the mortality rate in the control group (after covariate balancing). The obtained value corresponds to the difference in percentage points between mortality rates in treatment and control.
Conclusion
Conclusion and perspectives

Conclusion

- **Missing attributes** alter causal analyses.
- Additional assumptions guaranteeing either **unconfoundedness given missing values** or MAR.
- New proposal to handle **missing values in causal inference**.
- Prefer **AIPW** to IPW estimators, in theory and in practice.
- First application on real data.

Ongoing work

- Study the variance of $\hat{\tau}_{AIPW}$, $MIA$.
- Alternative to unconfoundedness despite missingness assumption using latent variable models.
- Different missing values mechanism (Sportisse et al. 2018).
- Heterogeneous treatment effects (Athey and Imbens, 2015) and optimal policy learning (Imai and Ratkovic, 2013).
- Compare results to the ones from CRASH3 study (Dewan et al. 2012).
Conclusion and perspectives

Conclusion

- **Missing attributes** alter causal analyses.
- Additional assumptions guaranteeing either **unconfoundedness given missing values** or MAR.
- New proposal to **handle missing values in causal inference**.
- Prefer **AIPW** to IPW estimators, in theory and in practice.
- First application on real data.

Ongoing work

- Study the variance of $\hat{\tau}_{AIPW,MIA}$.
- Alternative to unconfoundedness despite missingness assumption using latent variable models.
- Different missing values mechanism (Sportisse et al. 2018).
- Heterogeneous treatment effects (Athey and Imbens, 2015) and optimal policy learning (Imai and Ratkovic, 2013).
“One of the ironies of Big Data is that missing data play an ever more significant role” (R. Samworth, 2019)

More information and details on missing values: **R-miss-tastic** platform.

→ Theoretical and practical tutorials, popular datasets, bibliography, workflows (in R), active contributors/researchers in the community, etc.

[rmisstastic.netlify.com](rmisstastic.netlify.com)

MERCI
References


- 20,000 patients
- 250 continuous and categorical variables: heterogeneous
- 16 hospitals: multilevel data
- 4,000 new patients/ year

Graph produced using DAGitty (**Textor et al. (2011)**)
Define $\mu(w)(x) := \mathbb{E}[Y_i(w) | X_i = x]$ and $e(x) := \mathbb{P}(W_i = 1 | X_i = x)$.

Doubly robust estimator

\[
\hat{\tau}_{DR} := \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\mu}(1)(X_i) - \hat{\mu}(0)(X_i) + W_i \frac{Y_i - \hat{\mu}(1)(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}(0)(X_i)}{1 - \hat{e}(X_i)} \right)
\]

is consistent if either the $\hat{\mu}(w)(x)$ are consistent or $\hat{e}(x)$ is consistent.

Furthermore, the oracle $\hat{\tau}_{DR*}$ has good asymptotic variance.
# Cross-fitting for ATE estimation

## Cross-fitted ATE estimator

Assume we divide the data into $K$ folds.

\[
\hat{\tau}_{CF} = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{\mu}_{(1)}^{(-k(i))}(X_i) - \hat{\mu}_{(0)}^{(-k(i))}(X_i) \right) \\
+ W_i \frac{Y_i - \hat{\mu}_{(1)}^{(-k(i))}(X_i)}{\hat{\epsilon}^{(-k(i))}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_{(0)}^{(-k(i))}(X_i)}{1 - \hat{\epsilon}^{(-k(i))}(X_i)},
\]

where $k(i)$ maps an observation $X_i$ to one of the $K$ folds and $\hat{\mu}_{(-j)}$ indicates that the estimator has been learned on all the folds except the $j$-th fold.

Assuming overlap, sup-norm consistency of all used machine learning adjustments and risk decay, we have

\[
\sqrt{n} \left( \hat{\tau}_{CF} - \hat{\tau}_{DR^*} \right) \xrightarrow{p} 0. \\
\]

And we can prove that we can build level-$\alpha$ confidence intervals for $\tau$. 
Conditional independences

**CIT:** \( W \sim Z \odot R \)  
(where \( R_{ij} = 1 \{Z_{ij} \text{ is observed}\} \) and \( \odot = \text{Hadamard product} \).  
Example: for fixed \( \alpha \in \mathbb{R}^4 \) and \( \tau \in \mathbb{R} \):  
\[
\begin{align*}
  r^i &= (1, 1, 0, 0, 0, 1, 0, 0, 0, 1) \\
  \logit(\mathbb{P}(W^i = 1|Z^i_{\text{obs}} = z^i_{\text{obs}}, R^i = r^i)) &= \alpha_0 + \alpha_1 z^i_1 + \alpha_2 z^i_2 + \alpha_6 z^i_6 + \alpha_{10} z^i_{10} \\
  r^j &= (0, 1, 0, 0, 0, 0, 0, 0, 0, 0) \\
  \logit(\mathbb{P}(W^j = 1|Z^j_{\text{obs}} = z^j_{\text{obs}}, R^i = r^j)) &= \alpha_0 + \alpha_2 z^j_2
\end{align*}
\]

\(-\text{CIT}: \logit(\mathbb{P}(W^i = 1|Z^i = z^i)) = \alpha_0 + \alpha^T z^i .\)

**CIO:** \( Y \sim Z \odot R \).  
Example: for fixed \( \beta \in \mathbb{R}^4 \) and \( \tau \in \mathbb{R} \):  
\[
\begin{align*}
  r^i &= (1, 1, 0, 0, 0, 1, 0, 0, 0, 1) \\
  \mathbb{E}(Y^i|Z^i_{\text{obs}} = z^i_{\text{obs}}, R^i = r^i, W^i = w^i) &= \beta_0 + \beta_1 z^i_1 + \beta_2 z^i_2 + \beta_6 z^i_6 + \beta_{10} z^i_{10} + \tau w^i \\
  r^j &= (0, 1, 0, 0, 0, 0, 0, 0, 0, 0) \\
  \mathbb{E}(Y^j|Z^j_{\text{obs}} = z^j_{\text{obs}}, R^i = r^j, W^j = w^j) &= \beta_0 + \beta_2 z^j_2 + \tau w^j
\end{align*}
\]

\(-\text{CIO}: \mathbb{E}(Y^i|Z^i = z^i, W^i = w^i) = \beta_0 + \beta^T z^i + \tau w^i .\)