Evaluating High-dimensional Surrogate Markers of Vaccine Response through Causal Mediation Analysis

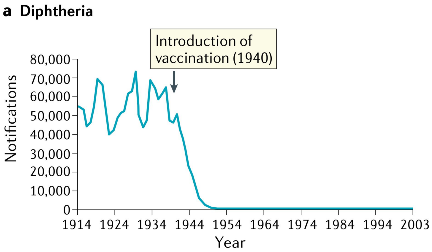
Arthur Hughes

University of Bordeaux

10/06/2024

Vaccination

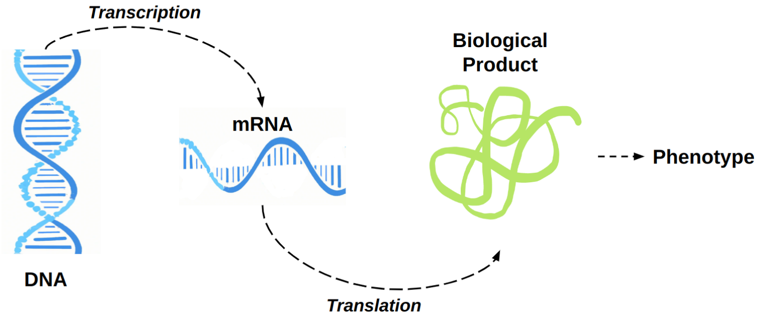
- Exploit immune system
- Most **cost-effective** measure in public health
  - **154 million** lives saved in last 50 years
  - **4-5 million** lives saved per year
- Historically developed empirically



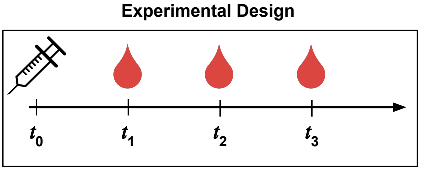
Pollard and Bijker, 2020

Transcriptomics

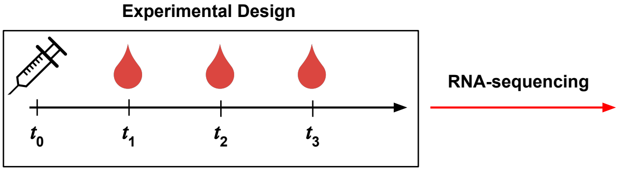
- **Gene expression**: Genes → Product



Transcriptomics in vaccinology



Transcriptomics in vaccinology



Count Matrix

Genes →

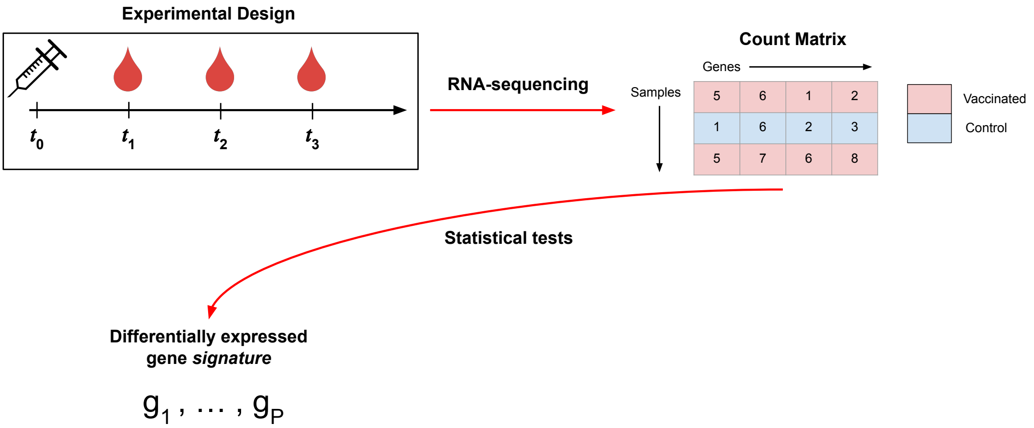
Samples ↓

5	6	1	2
1	6	2	3
5	7	6	8

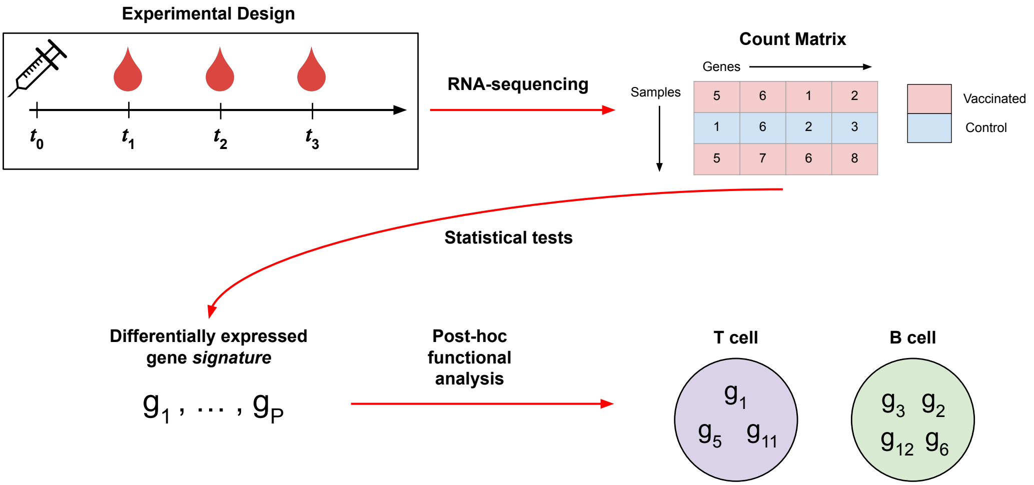
Vaccinated

Control

Transcriptomics in vaccinology



Transcriptomics in vaccinology



Why study gene expression?

Huge potential :

Why study gene expression?

Huge potential :

✓ Holistic view of system

Why study gene expression?

Huge potential :

- ✓ Holistic view of system
- ✓ Reveal vaccine mechanisms

Why study gene expression?

Huge potential :

- ✓ Holistic view of system
- ✓ Reveal vaccine mechanisms
- ✓ Observed early

Why study gene expression?

Huge potential :

- ✓ Holistic view of system
- ✓ Reveal vaccine mechanisms
- ✓ Observed early

But challenges with high-dimensionality...

Why study gene expression?

Huge potential :

✓ Holistic view of system

✓ Reveal vaccine mechanisms

✓ Observed early

But challenges with high-dimensionality...

⚠ Interpretability

Why study gene expression?

Huge potential :

- ✓ Holistic view of system
- ✓ Reveal vaccine mechanisms
- ✓ Observed early

But challenges with high-dimensionality...

- ⚠ Interpretability
- ⚠ Sensitivity to investigator choices

Why study gene expression?

Huge potential :

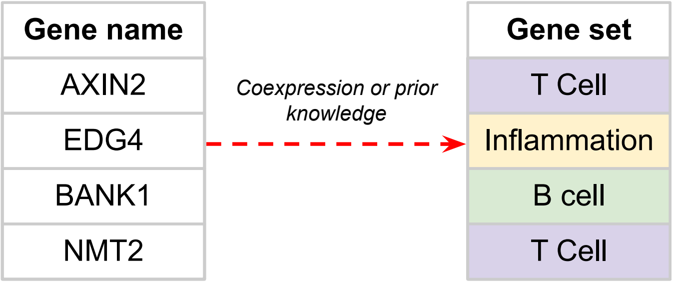
- ✓ Holistic view of system
- ✓ Reveal vaccine mechanisms
- ✓ Observed early

But challenges with high-dimensionality...

- ⚠ Interpretability
- ⚠ Sensitivity to investigator choices
- ⚠ Low signal-noise ratio

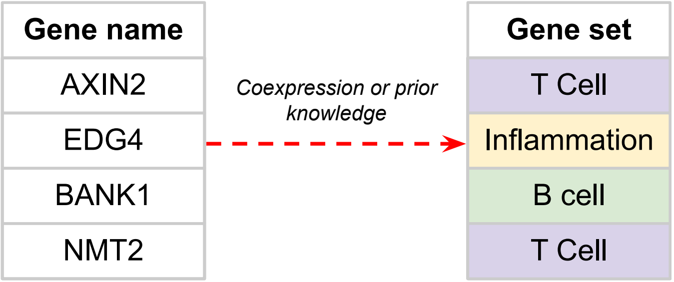
Gene Set Approaches

Investigate groups of biologically related genes



Gene Set Approaches

Investigate groups of biologically related genes



✓ Reduced dimensionality

Gene Set Approaches

Investigate groups of biologically related genes

Gene name	Gene set
AXIN2	T Cell
EDG4	Inflammation
BANK1	B cell
NMT2	T Cell

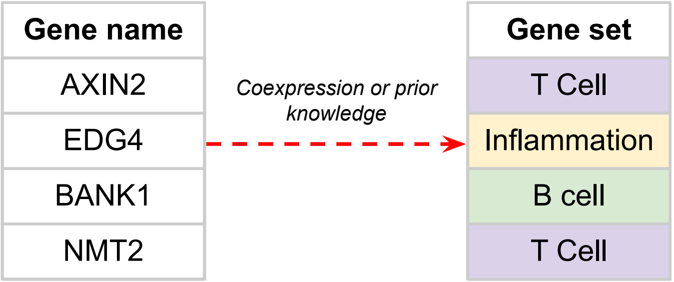
Coexpression or prior knowledge

✓ Reduced dimensionality

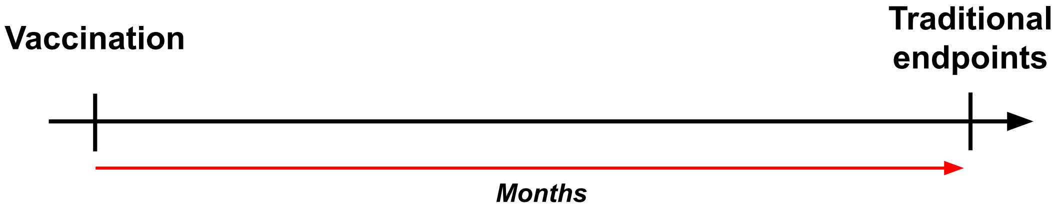
✓ Biological interpretability

Gene Set Approaches

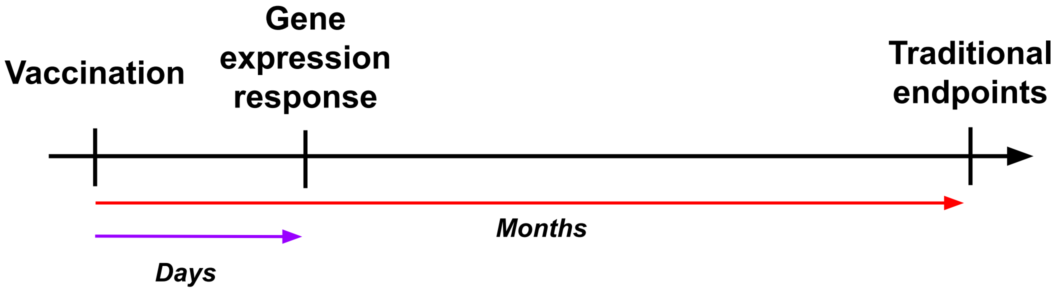
Investigate groups of biologically related genes



- ✓ Reduced dimensionality
- ✓ Biological interpretability
- ✓ Boost signal



Vaccination
Traditional endpoints
Months


Gene
expression
response
Vaccination
Traditional
endpoints
Months
Days

Potential of an early gene expression surrogate

✓ Accelerating validation of candidate vaccines

Potential of an early gene expression surrogate

- ✓ Accelerating validation of candidate vaccines
- ✓ Protection of at-risk groups

Potential of an early gene expression surrogate

- ✓ Accelerating validation of candidate vaccines
- ✓ Protection of at-risk groups
- ✓ Mechanistic inference

Potential of an early gene expression surrogate

- ✓ Accelerating validation of candidate vaccines
- ✓ Protection of at-risk groups
- ✓ Mechanistic inference

Difficulties in vaccine RCT context :

⚠ High-dimensional

Potential of an early gene expression surrogate

- ✓ Accelerating validation of candidate vaccines
- ✓ Protection of at-risk groups
- ✓ Mechanistic inference

Difficulties in vaccine RCT context :

- ⚠ High-dimensional
- ⚠ Small sample size

Potential of an early gene expression surrogate

✓ Accelerating validation of candidate vaccines

✓ Protection of at-risk groups

✓ Mechanistic inference

Difficulties in vaccine RCT context :

⚠ High-dimensional

⚠ Small sample size

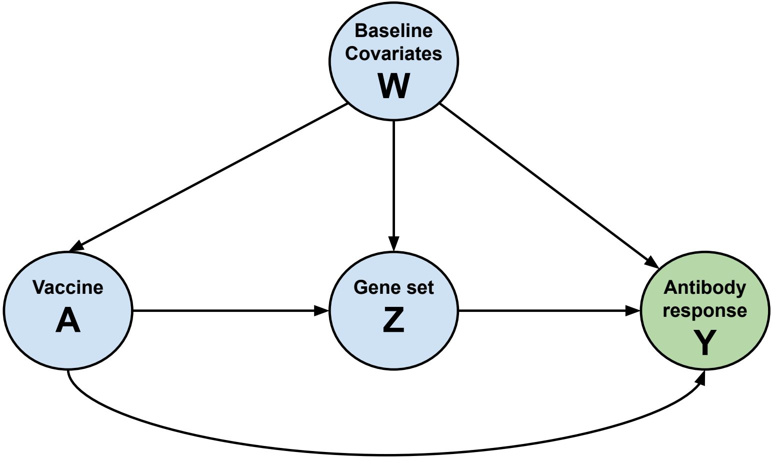
⚠ Complex data structures

Evaluating Gene Set Signals as Surrogates

Goal : Ranked list of gene sets by **proportion of treatment effect explained**

Gene Set	PTE
Z_1	R_1
Z_2	R_2
Z_3	R_3
....	....

where $R_1 > R_2 > R_3 > \ldots$


Baseline
Covariates
W
Vaccine
A
Gene set
Z
Antibody
response
Y

Potential outcomes framework

- Observed data consists of n i.i.d copies of $O = (W, A, Z, Y)$ where
  - W - baseline covariates (e.g. age, sex...)
  - $A \in \{0, 1\}$ - vaccine indicator
  - Z - m genes in the same biological pathway $(m > n)$
  - Y - antibody levels

Potential outcomes framework

- Observed data consists of n i.i.d copies of $O = (W, A, Z, Y)$ where
  - W - baseline covariates (e.g. age, sex...)
  - $A \in \{0, 1\}$ - vaccine indicator
  - Z - m genes in the same biological pathway ($m > n$)
  - Y - antibody levels

Define **potential outcomes**

- $Y(a)$: Response had treatment been $A = a$

Potential outcomes framework

- Observed data consists of n i.i.d copies of $O = (W, A, Z, Y)$ where
  - W - baseline covariates (e.g. age, sex...)
  - $A \in \{0, 1\}$ - vaccine indicator
  - Z - m genes in the same biological pathway $(m > n)$
  - Y - antibody levels

Define **potential outcomes**

- $Y(a)$: Response had treatment been $A = a$
- $Z(a')$ - mediators had treatment been $A = a'$

Potential outcomes framework

- Observed data consists of n i.i.d copies of $O = (W, A, Z, Y)$ where
  - W - baseline covariates (e.g. age, sex...)
  - $A \in \{0, 1\}$ - vaccine indicator
  - Z - m genes in the same biological pathway $(m > n)$
  - Y - antibody levels

Define **potential outcomes**

- $Y(a)$: Response had treatment been $A = a$
- $Z(a')$ - mediators had treatment been $A = a'$
- $Y(a, Z(a'))$ - response had treatment been $A = a$ and mediators under $A = a'$

Definition of effects

- **Total Effect** $= \Delta := \mathbb{E}\{Y(1) - Y(0)\}$

Definition of effects

- **Total Effect** $= \Delta := \mathbb{E}\{Y(1) - Y(0)\}$

Decompose total effect :

$$\underbrace{\mathbb{E}(Y(1) - Y(0))}_{\text{Total Effect}} = \underbrace{\mathbb{E}(Y(1, Z(0)) - Y(0, Z(0))}_{\text{Natural direct effect}} + \underbrace{\mathbb{E}(Y(1, Z(1)) - Y(1, Z(0))}_{\text{Natural indirect effect}}$$

Definition of effects

- **Total Effect** $= \Delta := \mathbb{E}\{Y(1) - Y(0)\}$

Decompose total effect :

$$\underbrace{\mathbb{E}(Y(1) - Y(0))}_{\text{Total Effect}} = \underbrace{\mathbb{E}(Y(1, Z(0)) - Y(0, Z(0))}_{\text{Natural direct effect}} + \underbrace{\mathbb{E}(Y(1, Z(1)) - Y(1, Z(0))}_{\text{Natural indirect effect}}$$

Define **proportion of treatment effect explained** as

$$R_S := \frac{NIE}{\text{Total Effect}} = 1 - \frac{NDE}{\text{Total Effect}}$$

Method 1 : G-computation

Goal : estimate $\Delta := \mathbb{E}\{Y(1) - Y(0)\}$

1. Estimate $\mathbb{E}\{Y|W, A\}$

Method 1 : G-computation

Goal : estimate $\Delta := \mathbb{E}\{Y(1) - Y(0)\}$

1. Estimate $\mathbb{E}\{Y|W, A\}$
2. Predict *potential outcomes* $\forall i \in \{1, ..., n\}$

$$\widehat{Y(1)} = \widehat{\mathbb{E}}\{Y|W, A = 1\},$$
$$\widehat{Y(0)} = \widehat{\mathbb{E}}\{Y|W, A = 0\}$$

Method 1 : G-computation

Goal : estimate $\Delta := \mathbb{E}\{Y(1) - Y(0)\}$

1. Estimate $\mathbb{E}\{Y|W, A\}$
2. Predict *potential outcomes* $\forall i \in \{1, ..., n\}$

$$\widehat{Y(1)} = \widehat{\mathbb{E}}\{Y|W, A = 1\},$$
$$\widehat{Y(0)} = \widehat{\mathbb{E}}\{Y|W, A = 0\}$$

3. Estimate $\widehat{\Delta} = \mathbb{E}\{\widehat{Y(1)} - \widehat{Y(0)}\}$

Method 1 : G-computation

Goal : estimate $\Delta := \mathbb{E}\{Y(1) - Y(0)\}$

1. Estimate $\mathbb{E}\{Y|W, A\}$
2. Predict *potential outcomes* $\forall i \in \{1, ..., n\}$
$$\widehat{Y(1)} = \widehat{\mathbb{E}}\{Y|W, A = 1\},$$
$$\widehat{Y(0)} = \widehat{\mathbb{E}}\{Y|W, A = 0\}$$
3. Estimate $\widehat{\Delta} = \mathbb{E}\{\widehat{Y(1)} - \widehat{Y(0)}\}$

⚠ **Optimises bias-variance tradeoff for** $\mathbb{E}\{Y|W, A\}$ - not the causal parameter of interest

Method 1 : G-computation

Goal : estimate $\Delta := \mathbb{E}\{Y(1) - Y(0)\}$

1. Estimate $\mathbb{E}\{Y|W, A\}$
2. Predict *potential outcomes* $\forall i \in \{1, ..., n\}$
$$\widehat{Y(1)} = \widehat{\mathbb{E}}\{Y|W, A = 1\},$$
$$\widehat{Y(0)} = \widehat{\mathbb{E}}\{Y|W, A = 0\}$$
3. Estimate $\widehat{\Delta} = \mathbb{E}\{\widehat{Y(1)} - \widehat{Y(0)}\}$

⚠ **Optimises bias-variance tradeoff for** $\mathbb{E}\{Y|W, A\}$ - not the causal parameter of interest

⚠ **Needs consistent estimation** of $\mathbb{E}\{Y|W, A\}$

Method 2 : Inverse probability weighting

Create comparable treatment groups w.r.t W

1. Estimate exposure mechanism $g_1 = \mathbb{P}(A = 1|W)$

Method 2 : Inverse probability weighting

Create comparable treatment groups w.r.t W

1. Estimate exposure mechanism $g_1 = \mathbb{P}(A = 1|W)$
2. $\widehat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}(\frac{A_i Y_i}{\widehat{g_{1i}}} - \frac{(1-A_i)Y_i}{1-\widehat{g_{1i}}})$

Method 2 : Inverse probability weighting

Create comparable treatment groups w.r.t W

1. Estimate exposure mechanism $g_1 = \mathbb{P}(A = 1|W)$
2. $\widehat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}(\frac{A_i Y_i}{\widehat{g_{1i}}} - \frac{(1-A_i)Y_i}{1-\widehat{g_{1i}}})$

Needs consistent estimation of $\mathbb{P}(A = 1|W)$

Method 2 : Inverse probability weighting

Create comparable treatment groups w.r.t W

1. Estimate exposure mechanism $g_1 = \mathbb{P}(A = 1|W)$
2. $\widehat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}(\frac{A_i Y_i}{\widehat{g_{1i}}} - \frac{(1-A_i)Y_i}{1-\widehat{g_{1i}}})$

Needs consistent estimation of $\mathbb{P}(A = 1|W)$

Not robust to sparsity

Method 3 : Targeted Maximum Likelihood Estimation

Target causal parameter of interest to reduce bias

Method 3 : Targeted Maximum Likelihood Estimation

Target causal parameter of interest to reduce bias

1. Initial estimation : $\widehat{\mathbb{E}}\{Y|W, A\}$ and $\widehat{\mathbb{P}}(A = 1|W)$

Method 3 : Targeted Maximum Likelihood Estimation

Target causal parameter of interest to reduce bias

1. Initial estimation : $\widehat{\mathbb{E}}\{Y|W, A\}$ and $\widehat{\mathbb{P}}(A = 1|W)$
2. Use $\widehat{\mathbb{P}}(A = 1|W)$ to *update* initial estimator $\implies \widehat{\mathbb{E}^*}\{Y|W, A\}$

Method 3 : Targeted Maximum Likelihood Estimation

Target causal parameter of interest to reduce bias

1. Initial estimation : $\widehat{\mathbb{E}}\{Y|W,A\}$ and $\widehat{\mathbb{P}}(A=1|W)$
2. Use $\widehat{\mathbb{P}}(A=1|W)$ to *update* initial estimator $\implies \widehat{\mathbb{E}^*}\{Y|W,A\}$
3. Predict *targeted* potential outcomes $\widehat{Y^*(a)} = \widehat{\mathbb{E}^*}\{Y|W,A=a\}$

Method 3 : Targeted Maximum Likelihood Estimation

Target causal parameter of interest to reduce bias

1. Initial estimation : $\widehat{\mathbb{E}}\{Y|W,A\}$ and $\widehat{\mathbb{P}}(A=1|W)$
2. Use $\widehat{\mathbb{P}}(A=1|W)$ to *update* initial estimator $\implies \widehat{\mathbb{E}^*}\{Y|W,A\}$
3. Predict *targeted* potential outcomes $\widehat{Y^*(a)} = \widehat{\mathbb{E}^*}\{Y|W,A=a\}$
4. $\widehat{\Delta} = \mathbb{E}[\widehat{Y^*(1)} - \widehat{Y^*(0)}]$

Method 3 : Targeted Maximum Likelihood Estimation

Target causal parameter of interest to reduce bias

1. Initial estimation : $\widehat{\mathbb{E}}\{Y|W, A\}$ and $\widehat{\mathbb{P}}(A = 1|W)$
2. Use $\widehat{\mathbb{P}}(A = 1|W)$ to *update* initial estimator $\implies \widehat{\mathbb{E}^*}\{Y|W, A\}$
3. Predict *targeted* potential outcomes $\widehat{Y^*(a)} = \widehat{\mathbb{E}^*}\{Y|W, A = a\}$
4. $\widehat{\Delta} = \mathbb{E}[\widehat{Y^*(1)} - \widehat{Y^*(0)}]$

✓ **Optimises bias-variance tradeoff for causal parameter of interest**

Method 3 : Targeted Maximum Likelihood Estimation

Target causal parameter of interest to reduce bias

1. Initial estimation : $\widehat{\mathbb{E}}\{Y|W,A\}$ and $\widehat{\mathbb{P}}(A=1|W)$
2. Use $\widehat{\mathbb{P}}(A=1|W)$ to *update* initial estimator $\implies \widehat{\mathbb{E}^*}\{Y|W,A\}$
3. Predict *targeted* potential outcomes $\widehat{Y^*(a)} = \widehat{\mathbb{E}^*}\{Y|W,A=a\}$
4. $\widehat{\Delta} = \mathbb{E}[\widehat{Y^*(1)} - \widehat{Y^*(0)}]$

✓ **Optimises bias-variance tradeoff for causal parameter of interest**

✓ **Doubly robust** - consistent if $\widehat{\mathbb{E}}\{Y|W,A\}$ or $\widehat{\mathbb{P}}(A=1|W)$ consistently estimated

G-Computation

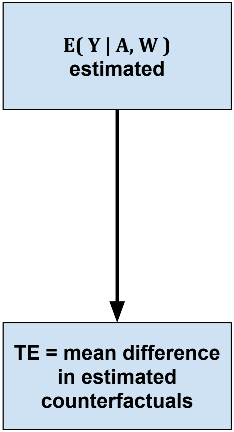


Figure adapted from Schuler and Rose, 2016

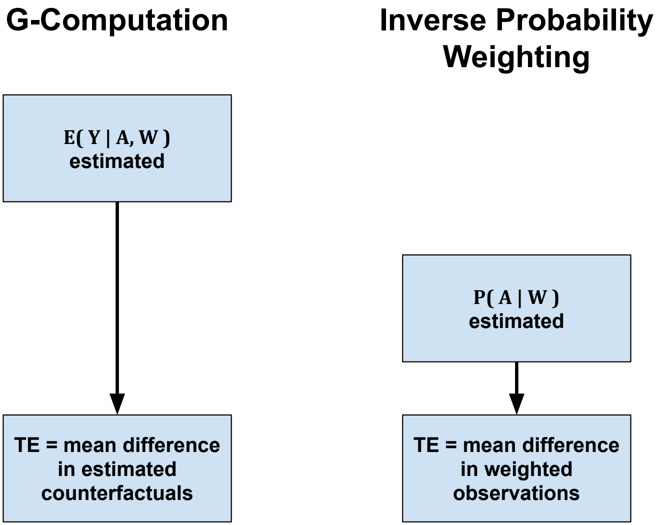


Figure adapted from Schuler and Rose, 2016

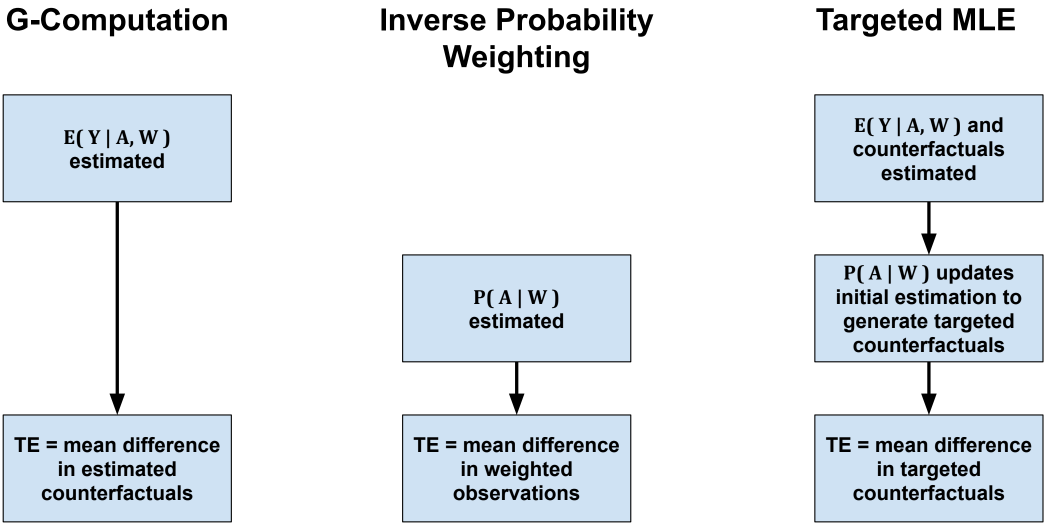


Figure adapted from Schuler and Rose, 2016

TMLE Step 1 : Initial estimation

⚠ $\mathbb{E}\{Y|W,A\}$ and $\mathbb{P}(A=1|W)$ are **complex**

TMLE Step 1 : Initial estimation

⚠ $\mathbb{E}\{Y|W,A\}$ and $\mathbb{P}(A=1|W)$ are **complex**

💡 Estimate with ensemble machine learning

TMLE Step 1 : Initial estimation

⚠ $\mathbb{E}\{Y|W, A\}$ and $\mathbb{P}(A = 1|W)$ are **complex**

Estimate with ensemble machine learning

Super-Learning

1. Choose m algorithms *a priori*

TMLE Step 1 : Initial estimation

$\mathbb{E}\{Y|W, A\}$ and $\mathbb{P}(A = 1|W)$ are **complex**

Estimate with ensemble machine learning

Super-Learning

1. Choose m algorithms *a priori*
2. Use cross-validation to estimate each algorithm performance

TMLE Step 1 : Initial estimation

$\mathbb{E}\{Y|W, A\}$ and $\mathbb{P}(A = 1|W)$ are **complex**

Estimate with ensemble machine learning

Super-Learning

1. Choose m algorithms *a priori*
2. Use cross-validation to estimate each algorithm performance
3. Find **combination of algorithms minimising CV loss** by estimating vector of weights $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)^T$

TMLE Step 1 : Initial estimation

$\mathbb{E}\{Y|W, A\}$ and $\mathbb{P}(A = 1|W)$ are **complex**

Estimate with ensemble machine learning

Super-Learning

1. Choose m algorithms *a priori*
2. Use cross-validation to estimate each algorithm performance
3. Find **combination of algorithms minimising CV loss** by estimating vector of weights $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)^T$
4. Fit each algorithm to full data and use $\widehat{\boldsymbol{\alpha}}$ to generate super-learner estimator

TMLE Step 1 : Initial estimation

$\mathbb{E}\{Y|W,A\}$ and $\mathbb{P}(A=1|W)$ are **complex**

Estimate with ensemble machine learning

Super-Learning

1. Choose m algorithms *a priori*
2. Use cross-validation to estimate each algorithm performance
3. Find **combination of algorithms minimising CV loss** by estimating vector of weights $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)^T$
4. Fit each algorithm to full data and use $\widehat{\boldsymbol{\alpha}}$ to generate super-learner estimator

$\Longrightarrow$ **Initial estimates** $\mathbb{E}\{\widehat{Y|W,A}\}$ and $\widehat{g_a} = \mathbb{P}(\widehat{A=a}|W)$

TMLE Step 2 : Targeting step

1. Calculate *Auxiliary covariate* $H_a(A,W) = \frac{\mathbb{1}(A=1)}{\widehat{g}_1} - \frac{\mathbb{1}(A=0)}{\widehat{g}_0}$

TMLE Step 2 : Targeting step

1. Calculate *Auxiliary covariate* $H_a(A, W) = \frac{\mathbb{1}(A=1)}{\widehat{g}_1} - \frac{\mathbb{1}(A=0)}{\widehat{g}_0}$
2. Regress observed Y on H_a with intercept $\widehat{Y}$

$$Y = \widehat{Y} + \delta H_a$$

TMLE Step 2 : Targeting step

1. Calculate *Auxiliary covariate* $H_a(A, W) = \frac{\mathbb{1}(A=1)}{\widehat{g}_1} - \frac{\mathbb{1}(A=0)}{\widehat{g}_0}$
2. Regress observed Y on H_a with intercept $\widehat{Y}$

$$Y = \widehat{Y} + \delta H_a$$

3. Targeted potential outcomes $\Longrightarrow$

$\widehat{Y^*(1)} = \widehat{Y(1)} + \widehat{\delta} H_1$ and $\widehat{Y^*(0)} = \widehat{Y(0)} + \widehat{\delta} H_0$

TMLE Step 2 : Targeting step

1. Calculate *Auxiliary covariate* $H_a(A, W) = \frac{\mathbb{1}(A=1)}{\widehat{g}_1} - \frac{\mathbb{1}(A=0)}{\widehat{g}_0}$
2. Regress observed Y on H_a with intercept $\widehat{Y}$

$$Y = \widehat{Y} + \delta H_a$$

3. Targeted potential outcomes $\implies$

$\widehat{Y^*(1)} = \widehat{Y(1)} + \widehat{\delta} H_1$ and $\widehat{Y^*(0)} = \widehat{Y(0)} + \widehat{\delta} H_0$

4. Estimate $\widehat{\Delta} = \mathbb{E}[\widehat{Y^*(1)} - \widehat{Y^*(0)}]$

Estimating the Natural Direct Effect

Goal : **estimate the natural direct effect** from n copies of $O = (W, A, Z, Y)$. Let

- $Q_Y(W, A, Z) = \mathbb{E}(Y|W, A, Z)$

Estimating the Natural Direct Effect

Goal : **estimate the natural direct effect** from n copies of $O = (W, A, Z, Y)$. Let

- $Q_Y(W, A, Z) = \mathbb{E}(Y|W, A, Z)$

Define NDE $= \underbrace{\mathbb{E}_W\{\mathbb{E}_Z[\underbrace{Q_Y(W, 1, Z) - Q_Y(W, 0, Z)}_{Q_{\text{diff}}}|A = 0, W]\}}_{\textit{Residual treatment effect} = \Delta_S}$

$\implies$ **two quantities to target** : Q_{diff} and Δ_S

$$\text{NDE} = \mathbb{E}_W\{\mathbb{E}_Z[Q_Y(W,1,Z) - Q_Y(W,0,Z)|A=0,W]\}$$

1. TMLE of Q_{Diff}

Initial estimates with super-learning :

E(Y | W, A, Z), P(Z | W, A), P(A | W)

↓

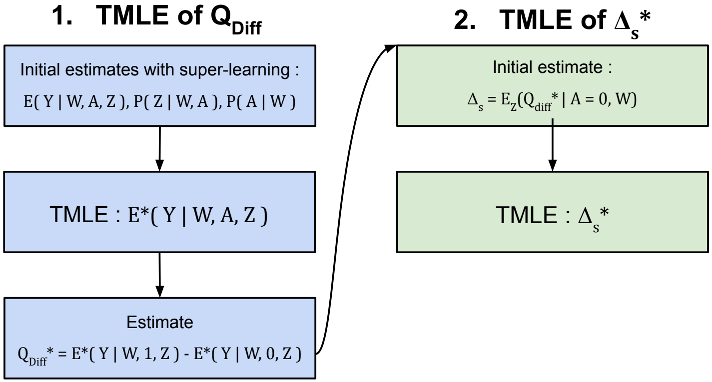
TMLE : E*(Y | W, A, Z)

↓

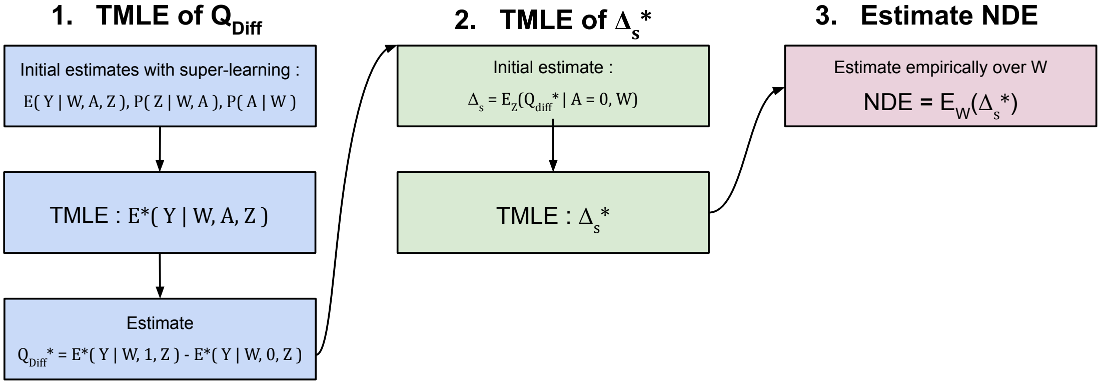
Estimate

Q_{Diff}* = E*(Y | W, 1, Z) - E*(Y | W, 0, Z)

$$\text{NDE} = \mathbb{E}_W\{\mathbb{E}_Z[Q_Y(W,1,Z) - Q_Y(W,0,Z)|A=0,W]\}$$



$$\text{NDE} = \mathbb{E}_W\{\mathbb{E}_Z[Q_Y(W,1,Z) - Q_Y(W,0,Z)|A=0,W]\}$$



Identifiability conditions

Consistency

- $Y(a,m) = Y$ if $A = a$, $Z = m$
- $Z(a') = Z$ if $A = a'$
- $Y(a, Z(a')) = Y(a,m)$ if $Z = m$

Identifiability conditions

Consistency

- $Y(a,m) = Y$ if $A = a,\ Z = m$
- $Z(a') = Z$ if $A = a'$
- $Y(a, Z(a')) = Y(a,m)$ if $Z = m$

Conditional independence

- $A \perp\!\!\!\perp Z(a')|W$
- $A \perp\!\!\!\perp Y(a,m)|W$
- $Z \perp\!\!\!\perp Y(a,m)|W, A = a$

⚠ $Z(a') \perp\!\!\!\perp Y(a,m)|W$

Identifiability conditions

Consistency

- $Y(a,m) = Y$ if $A = a,\ Z = m$
- $Z(a') = Z$ if $A = a'$
- $Y(a, Z(a')) = Y(a,m)$ if $Z = m$

Conditional independence

- $A \perp\!\!\!\perp Z(a') | W$
- $A \perp\!\!\!\perp Y(a,m) | W$
- $Z \perp\!\!\!\perp Y(a,m) | W, A = a$

⚠ $Z(a') \perp\!\!\!\perp Y(a,m) | W$

Positivity

- $\mathbb{P}(A = 1 | W) \in (0,1)$
- $\mathbb{P}(Z = m | W, A = a) \in (0,1)$

Limitations of TMLE workflow

- Potential violation of *cross-world independence* assumption

Limitations of TMLE workflow

- Potential violation of *cross-world independence* assumption
- No associated statistical test

Limitations of TMLE workflow

- Potential violation of *cross-world independence* assumption
- No associated statistical test
- Not simple to extend to **longitudinal mediator setting**

Limitations of TMLE workflow

- Potential violation of *cross-world independence* assumption
- No associated statistical test
- Not simple to extend to **longitudinal mediator setting**
- No tractable form for statistical inference

Limitations of TMLE workflow

- Potential violation of *cross-world independence* assumption
- No associated statistical test
- Not simple to extend to **longitudinal mediator setting**
- No tractable form for statistical inference

Take-home messages

- Potential and challenges evaluating an early gene expression surrogate of vaccine response
- **Targeted learning to reduce bias** in causal estimands
- Methodological workflow for evaluating **PTE by gene sets** on vaccine response

Thank you for listening

References

Pollard, Andrew J. and Else M. Bijker (Dec. 2020). "A guide to vaccinology: from basic principles to new developments". In: *Nature Reviews Immunology* 21.2, pp. 83–100. ISSN: 1474-1741.

Schuler, Megan S. and Sherri Rose (Dec. 2016). "Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies". In: *American Journal of Epidemiology* 185.1, pp. 65–73. ISSN: 1476-6256.

Step 1 : TMLE of Q_{diff}

Define *auxiliary covariate*

$$H_Y(Q_Z, g) = \frac{\mathbb{1}(A=1)}{g(1|W)} \cdot \frac{Q_Z(W,0)}{Q_Z(W,1)} - \frac{\mathbb{1}(A=0)}{g(0|W)}$$

and *working parametric submodel*

$$Q_Y(\epsilon_1) = Q_Y + \epsilon_1 H_Y(Q_Z, g) \tag{1}$$

Step 1 : TMLE of Q_{diff}

Define *auxiliary covariate*

$$H_Y(Q_Z, g) = \frac{\mathbb{1}(A=1)}{g(1|W)} \cdot \frac{Q_Z(W,0)}{Q_Z(W,1)} - \frac{\mathbb{1}(A=0)}{g(0|W)}$$

and *working parametric submodel*

$$Q_Y(\epsilon_1) = Q_Y + \epsilon_1 H_Y(Q_Z, g) \tag{1}$$

TMLE of Q_{diff}

1. Obtain initial estimates $\widehat{Q_Y}$, $\widehat{Q_Z}$, $\widehat{g}$ with *super-learner*

Step 1 : TMLE of Q_{diff}

Define *auxiliary covariate*

$$H_Y(Q_Z, g) = \frac{\mathbb{1}(A=1)}{g(1|W)} \cdot \frac{Q_Z(W,0)}{Q_Z(W,1)} - \frac{\mathbb{1}(A=0)}{g(0|W)}$$

and *working parametric submodel*

$$Q_Y(\epsilon_1) = Q_Y + \epsilon_1 H_Y(Q_Z, g) \tag{1}$$

TMLE of Q_{diff}

1. Obtain initial estimates $\widehat{Q_Y}$, $\widehat{Q_Z}$, $\widehat{g}$ with *super-learner*
2. Find $\epsilon_1^* = \arg\min_\epsilon L_Y(\widehat{Q_Y(\epsilon)})$ w.r.t some loss function L_Y

Step 1 : TMLE of Q_{diff}

Define *auxiliary covariate*

$$H_Y(Q_Z, g) = \frac{\mathbb{1}(A=1)}{g(1|W)} \cdot \frac{Q_Z(W,0)}{Q_Z(W,1)} - \frac{\mathbb{1}(A=0)}{g(0|W)}$$

and *working parametric submodel*

$$Q_Y(\epsilon_1) = Q_Y + \epsilon_1 H_Y(Q_Z, g) \tag{1}$$

TMLE of Q_{diff}

1. Obtain initial estimates $\widehat{Q_Y}$, $\widehat{Q_Z}$, $\widehat{g}$ with *super-learner*
2. Find $\epsilon_1^* = \arg\min_\epsilon L_Y(\widehat{Q_Y(\epsilon)})$ w.r.t some loss function L_Y
3. Plug in $\widehat{\epsilon_1^*}$ for targeted estimate $\widehat{Q_Y^*} = \widehat{Q_Y} + \widehat{\epsilon_1^*} H_Y(\widehat{Q_Z}, \widehat{g})$

Step 1 : TMLE of Q_{diff}

Define *auxiliary covariate*

$$H_Y(Q_Z, g) = \frac{\mathbb{1}(A=1)}{g(1|W)} \cdot \frac{Q_Z(W,0)}{Q_Z(W,1)} - \frac{\mathbb{1}(A=0)}{g(0|W)}$$

and *working parametric submodel*

$$Q_Y(\epsilon_1) = Q_Y + \epsilon_1 H_Y(Q_Z, g) \tag{1}$$

TMLE of Q_{diff}

1. Obtain initial estimates $\widehat{Q_Y}$, $\widehat{Q_Z}$, $\widehat{g}$ with *super-learner*
2. Find $\epsilon_1^* = \arg\min_\epsilon L_Y(\widehat{Q_Y(\epsilon)})$ w.r.t some loss function L_Y
3. Plug in $\widehat{\epsilon_1^*}$ for targeted estimate $\widehat{Q_Y^*} = \widehat{Q_Y} + \widehat{\epsilon_1^*} H_Y(\widehat{Q_Z}, \widehat{g})$
4. Estimate $\widehat{Q_{\text{diff}}^*} = \widehat{Q_Y^*}(W, A=1, Z) - \widehat{Q_Y^*}(W, A=0, Z)$

Step 2 : TMLE of Δ_S

Define *auxiliary covariate*

$$H_Z(g) = \frac{1}{g(0|W)}$$

and *working parametric submodel*

$$\Delta_S(\epsilon_2) = \psi_Z + \epsilon_2 H_Z(g) \quad (2)$$

Step 2 : TMLE of Δ_S

Define *auxiliary covariate*

$$H_Z(g) = \frac{1}{g(0|W)}$$

and *working parametric submodel*

$$\Delta_S(\epsilon_2) = \psi_Z + \epsilon_2 H_Z(g) \tag{2}$$

TMLE of Δ_S

1. Obtain initial estimate of $\Delta_S = \mathbb{E}_{Q_Z}[Q_Y(W,1,Z) - Q_Y(W,0,Z)|A=0,W]$ by regressing $\widehat{Q^*_{diff}}$ on W among controls

Step 2 : TMLE of Δ_S

Define *auxiliary covariate*

$$H_Z(g) = \frac{1}{g(0|W)}$$

and *working parametric submodel*

$$\Delta_S(\epsilon_2) = \psi_Z + \epsilon_2 H_Z(g) \quad (2)$$

TMLE of Δ_S

1. Obtain initial estimate of $\Delta_S = \mathbb{E}_{Q_Z}[Q_Y(W,1,Z) - Q_Y(W,0,Z)|A=0,W]$ by regressing $\widehat{Q^*_{diff}}$ on W among controls
2. Find $\widehat{\epsilon^*_2} = \arg\min_\epsilon L_Z(\psi_Z(\widehat{Q^*_Y})(\epsilon))$ w.r.t some loss function L_Z

Step 2 : TMLE of Δ_S

Define *auxiliary covariate*

$$H_Z(g) = \frac{1}{g(0|W)}$$

and *working parametric submodel*

$$\Delta_S(\epsilon_2) = \psi_Z + \epsilon_2 H_Z(g) \tag{2}$$

TMLE of Δ_S

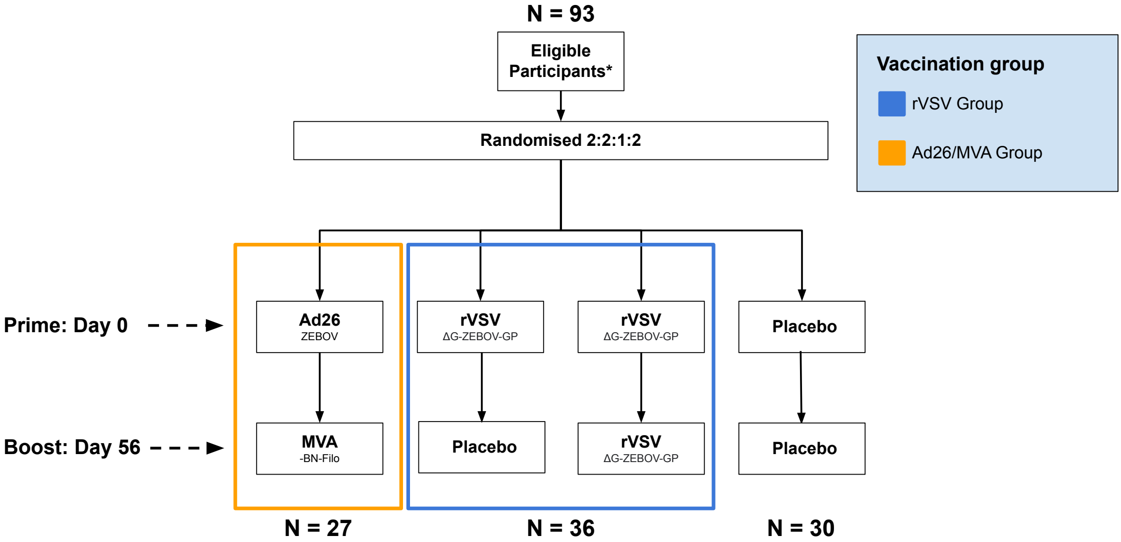
1. Obtain initial estimate of $\Delta_S = \mathbb{E}_{Q_Z}[Q_Y(W,1,Z) - Q_Y(W,0,Z)|A=0,W]$ by regressing $\widehat{Q^*_{diff}}$ on W among controls
2. Find $\widehat{\epsilon_2^*} = \arg\min_\epsilon L_Z(\psi_Z(\widehat{Q^*_Y})(\epsilon))$ w.r.t some loss function L_Z
3. Plug in $\widehat{\epsilon_2^*}$ for targeted estimate $\widehat{\psi^*_Z(Q^*_Y)} = \psi_Z(\widehat{Q^*_Y}) + \widehat{\epsilon_2}^* H_Z(\widehat{g})$

Step 3 : Empirical estimate of Direct effect

$$\psi_{NDE} = \mathbb{E}_{Q_W}(\psi_Z(Q_Y))$$

$\Longrightarrow$ **Empirically average over covariates for NDE**

$$\widehat{\psi_{NDE}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\psi_Z^*(Q_Y^*)}(W_i)$$



N = 93
Eligible Participants*
Randomised 2:2:1:2
Vaccination group
rVSV Group
Ad26/MVA Group
Prime: Day 0
Boost: Day 56
Ad26
ZEBOV
MVA
-BN-Filo
rVSV
ΔG-ZEBOV-GP
Placebo
rVSV
ΔG-ZEBOV-GP
rVSV
ΔG-ZEBOV-GP
Placebo
Placebo
N = 27
N = 36
N = 30