Interpreting the Proportion Explained: There is no Escaping Unverifiable Assumptions

Florian Stijven [1] Ariel Alonso [1] Geert Molenberghs [1, 2]

[1]KU Leuven, I-BioStat, B-3000 Leuven, Belgium

[2]UHasselt, I-Biostat, B-3500 Hasselt, Belgium

March 11, 2024

Contact: florian.stijven@kuleuven.be

Introduction

- Freedman et al. (1992) defined the proportion of treatment effect explained (PTE)
  - Model-based definition
- Later, Wang and Taylor (2002) proposed a non-parametric definition $\rightarrow PTE_{WT}$
- Recent literature: focus on estimating PTE_{WT}
- 2 Central questions in this talk
  1. **How can we interpret PTE_{WT}?**
  2. **Do we need additional assumptions for certain interpretations?**

Notation

- Consider a randomized trial with 2 parallel arms
  - Treatment indicator: Z
    - Control treatment: $Z = 0$
    - Active treatment: $Z = 1$
  - Surrogate endpoint: S
  - True endpoint: T
- Under SUTVA[1], we can define corresponding potential outcomes
  - S_0 and S_1
  - T_0 and T_1

[1] Stable unit treatment value assumption (Imbens and Rubin 2015)

PTE_{WT} simplified

Definition (PTE_{WT} simplified)

$$PTE_{WT} = \frac{\Delta - \Delta_S}{\Delta}$$

where

$$\Delta = E(T_1) - E(T_0)$$

is the *total treatment effect* and

$$\Delta_S = \int E(T_1 | S_1 = s)\, dF_{S_0}(s) - E(T_0)$$

is the *residual treatment effect* (Parast et al. 2016)

- Special case of general definition
- Simpler $\rightarrow$ focus on interpretation

Causal Interpretation 1: Principal Surrogacy

Definition (Principal Surrogate (Frangakis and Rubin 2002))

S is a principal surrogate for T iff

$$E(T_1 - T_0 | S_0 = S_1 = s) = 0 \;\forall\; s$$

- $S_1 - S_0 = 0 \Rightarrow$ No expected treatment effect on T
- Accepting this definition, at least two practical issues remain:
  1. Surrogacy is not quantified in a single value
  2. Unlikely that principal surrogacy holds exactly in practice
     $\rightarrow$ How can we determine whether S is approximately a principal surrogate?

Causal Interpretation 1: Principal Surrogacy (ctd.)

1. The *causal residual treatment effect*, $\Delta_{S,c}$, summarizes principal surrogacy

$$\Delta_{S,c} = \int E(T_1 - T_0 | S_0 = S_1 = s)\, dF_{S_0}(s)$$

   - Principal surrogacy $\Rightarrow \Delta_{S,c} = 0$

2. Principal surrogacy approximately satisfied if $\frac{\Delta_{S,c}}{\Delta} \approx 0$

- $\Delta_{S,c}$ is unidentifiable without **additional assumptions**
   - Assume *conditional independence* (Parast et al. 2016):

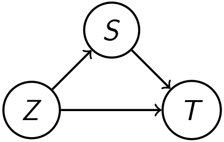
$$S_0 \perp T_1 \,|\, S_1 \text{ and } S_1 \perp T_0 \,|\, S_0$$

   - Then $\Delta_{S,c} = \Delta_S$
- Now PTE_{WT} quantifies principal surrogacy

$$\frac{\Delta_{S,c}}{\Delta} \approx 0 \iff \frac{\Delta_S}{\Delta} \approx 0 \iff PTE_{WT} \approx 1$$

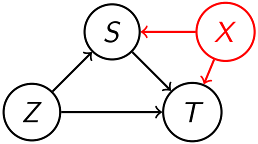
Causal Interpretation 2: Proportion Mediated

- PTE_{WT} often interpreted like a proportion mediated
- Only valid under the following causal diagram
- **Additional assumption:** No confounding of the $S - T$ relation
  - Not satisfied by randomization of Z
  - Extremely unlikely to hold



Causal Interpretation 2: Proportion Mediated

- PTE_{WT} often interpreted like a proportion mediated
- Only valid under the following causal diagram
- **Additional assumption:** No confounding of the $S - T$ relation
  - Not satisfied by randomization of Z
  - Extremely unlikely to hold
- Take baseline covariates, X, into account



More realistic ...

Non-Causal Interpretation: Notation

- *Evaluation trial*, $A = 1$
  - S and T observed
  - Total treatment effect: $\Delta = E(T_1|A = 1) - E(T_0|A = 1)$
- *Application trial*, $A = 0$
  - S observed, T unobserved
  - Total treatment effect: $\Delta^* = E(T_1|A = 0) - E(T_0|A = 0)$

Can we predict Δ and Δ^* using only S?

Non-Causal Interpretation: Evaluation Trial

- $\mu(s) = E(T_1|S_1 = s, A = 1)$ is a prediction rule for T_z
- In the definition of Δ, replace T_z with its prediction:

$$\Delta = E(T_1 - T_0|A = 1) \longrightarrow \tilde{\Delta} = E\left\{\mu(S_1) - \mu(S_0)|A = 1\right\}$$

- Δ_S is the prediction bias:

$$\tilde{\Delta} = \Delta - \Delta_S$$

- PTE_{WT} is the relative prediction bias:

$$PTE_{WT} = \frac{\Delta - \Delta_S}{\Delta} = \frac{\tilde{\Delta}}{\Delta}$$

- **However**, $\tilde{\Delta}$ not relevant because we can estimate Δ directly
  $\rightarrow$ Apply this thinking to the application trial

Non-Causal Interpretation: Application Trial

- In the definition of Δ^*, replace T_z with its prediction:

$$\tilde{\Delta}^* = E\left\{\mu(S_1) - \mu(S_0) | A = 0\right\}$$

Lemma (Treatment effect prediction, Application trial)

Under the following assumptions,

1. $E(T_z|S_z = s, A = 1) = E(T_z|S_z = s, A = 0)$ *for* $z = 0, 1$
2. $F_{S_0|A=1} = F_{S_0|A=0}$,

Δ_S *is the prediction bias in the application trial, i.e.,*

$$\tilde{\Delta}^* = \Delta^* - \Delta_S.$$

- Under these **additional assumptions**

$$PTE_{WT} \approx 1 \iff \Delta_S \approx 0 \Rightarrow \tilde{\Delta}^* \text{ is a good prediction}$$

Application: Evaluation and Applications Trials

- Three simulated vaccine trials ($N = 20,000$)
  - T: infection status
  - S: neutralizing antibody level
  - X: baseline health and age
- Evaluation trial
  - S is an important mediator
- Application trial 1
  - S is an important mediator
  - Assumptions of previous lemma satisfied *conditional on* X
- Application trial 2
  - Vaccine operates mainly through cellular immunity
  - Assumptions of previous lemma not satisfied

Application: Results in Evaluation Trial

- Estimate PTE_{WT} and related quantities
  1. Ignoring $X \rightarrow PTE_{WT}$
  2. Including $X \rightarrow PTE_{WT}^{X}$
     - Estimate components conditional on X, then marginalize over X
- $PTE_{WT} >> 1 \Rightarrow \tilde{\Delta}$ predicts Δ poorly
  - $\tilde{\Delta}^*$ is likely also a poor predictor of Δ^*
- $PTE_{WT}^{X} \approx 1 \Rightarrow$ can predict Δ well
  - Including X, we *might* predict Δ^* well

Ignoring Baseline Covariates			Including Baseline Covariates		
Δ_S	Δ	PTE_{WT}	$E(\Delta_S(X))$	Δ	PTE_{WT}^{X}
-0.015	0.069	1.470	0.005	0.071	0.870

Application: Results in Evaluation Trial

- Estimate PTE_{WT} and related quantities
  1. Ignoring $X \rightarrow PTE_{WT}$
  2. Including $X \rightarrow PTE_{WT}^{X}$
     - Estimate components conditional on X, then marginalize over X
- $PTE_{WT} >> 1 \Rightarrow \tilde{\Delta}$ predicts Δ poorly
  - $\tilde{\Delta}^*$ is likely also a poor predictor of Δ^*
- $PTE_{WT}^{X} \approx 1 \Rightarrow$ can predict Δ well
  - Including X, we *might* predict Δ^* well

Ignoring Baseline Covariates			Including Baseline Covariates		
Δ_S	Δ	PTE_{WT}	$E(\Delta_S(X))$	Δ	PTE_{WT}^{X}
-0.015	0.069	1.470	0.005	0.071	0.870

Results: Treatment Effect Predictions

- $\tilde{\Delta}^*$: prediction ignoring X
  - Related to PTE_{WT}
- $E\left(\tilde{\Delta}^*(X)\right)$: prediction taking X into account
  - Related to PTE_{WT}^X
- Prediction only accurate in **application trial 1** when we **take X into account**
  - Only case where we expected an accurate prediction

	Δ^*	$\tilde{\Delta}^*$	$E\left(\tilde{\Delta}^*(X)\right)$
Application Trial 1	0.053	0.077	0.051
Application Trial 2	0.118	0.015	0.011

Results: Treatment Effect Predictions

- $\tilde{\Delta}^*$: prediction ignoring X
  - Related to PTE_{WT}
- $E\left(\tilde{\Delta}^*(X)\right)$: prediction taking X into account
  - Related to PTE_{WT}^X
- Prediction only accurate in **application trial 1** when we **take** X **into account**
  - Only case where we expected an accurate prediction

	Δ^*	$\tilde{\Delta}^*$	$E\left(\tilde{\Delta}^*(X)\right)$
Application Trial 1	0.053	0.077	0.051
Application Trial 2	0.118	0.015	0.011

Results: Treatment Effect Predictions

- $\tilde{\Delta}^*$: prediction ignoring X
  - Related to PTE_{WT}
- $E\left(\tilde{\Delta}^*(X)\right)$: prediction taking X into account
  - Related to PTE_{WT}^X
- Prediction only accurate in **application trial 1** when we **take X into account**
  - Only case where we expected an accurate prediction

	Δ^*	$\tilde{\Delta}^*$	$E\left(\tilde{\Delta}^*(X)\right)$
Application Trial 1	0.053	0.077	0.051
Application Trial 2	0.118	0.015	0.011

Conclusions

- We identified 3 possible interpretations of the PTE_{WT}
  1. Measure of principal surrogacy
  2. Proportion mediated
  3. Relative trial-level prediction bias
- These interpretations are only possible under **additional unverifiable assumptions**
- These interpretations "make sense" for $PTE_{WT} > 1$
  - Contrary to what is commonly assumed
- Non-causal interpretation aligns most with the final goal of surrogates:
  $\rightarrow$ Prediction of treatment effects using only S

References

Frangakis, Constantine E. and Donald B. Rubin (2002). "Principal stratification in causal inference". In: *Biometrics* 58.1, pp. 21–29.

Freedman, Laurence S. et al. (1992). "Statistical validation of intermediate endpoints for chronic diseases". In: *Statistics in medicine* 11.2, pp. 167–178.

Imbens, Guido W. and Donald B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Parast, Layla et al. (2016). "Robust estimation of the proportion of treatment effect explained by surrogate marker information". In: *Statistics in medicine* 35.10, pp. 1637–1653.

Wang, Yue and Jeremy M.G. Taylor (2002). "A Measure of the Proportion of Treatment Effect Explained by a Surrogate Marker". In: *Biometrics* 58.4, pp. 803–812.

Application: Data Generating Mechanism

- Distribution of $T|S, Z, X, A$ is modeled through its conditional mean, $\pi = E(T|S, Z, X, A)$:

$$\log\left(\frac{\pi}{1-\pi}\right) = 0.8 + \alpha_0 \cdot (1-A) + (0.1 + \alpha_1 \cdot (1-A)) \cdot Z \\ + 0.5 \cdot S - 0.03 \cdot (X_1 - 45) + 0.1 \cdot X_2.$$

  - Application trial 1: $\alpha_0 = \alpha_1 = 0$
  - Application trial 2: $\alpha_0 = 0.2$ and $\alpha_1 = 0.90$

- $S|Z, X, A$ is Gaussian with unit variance and following mean function

$$E(S|Z, X, A) = (1.5 \cdot Z + \beta \cdot (1-A)) \cdot Z - 0.05 \cdot (X_1 - 45) + 0.15 \cdot X_2.$$

  - Application trial 1: $\beta = -0.75$
  - Application trail 2: $\beta = -1.4$

Application: Data Generating Mechanism (ctd.)

- Distribution of age

$$X_1|A = 1 \sim N\left(35, 7^2\right) \text{ and } X_1|A = 0 \sim N\left(40, 5^2\right)$$

- Distribution of baseline health (larger values indicate of better health)

$$X_2|A = 1 \sim N\left(5, 5^2\right) \text{ and } X_2|A = 0 \sim \left(2, 4^2\right)$$

Baseline covariates and PTE_{WT}^{X}

$$PTE_{WT}^{X} = \frac{E\left\{\Delta(X) - \Delta_S(X)|A=1\right\}}{E(\Delta(X)|A=1)} = \frac{E\left\{\tilde{\Delta}(X)|A=1\right\}}{E(\Delta(X)|A=1)}.$$

where

$$E\left\{\Delta(X)|A=1\right\} = E\left\{E(T_1 - T_0|X)\right\} = \Delta$$

$$E\left\{\tilde{\Delta}(X)|A=1\right\} = E\left\{\mu_X(S_1) - \mu_X(S_0)|A=1\right\}$$

$$\mu_x(s) = E(T_1|S_1 = s, X = x, A = 1)$$