# Application of PTE methods to a Hypothetical Vaccine Trial

Florian Stijven* Geert Molenberghs† Ariel Alonso‡

## Contents



# 1 Introduction

In this document, the analysis of a set of hypothetical vaccine trials is described. This document supplements "Proportion of Treatment Effect Explained: An Overview of Interpretations". All concepts discussed in that paper are assumed known in this document. All data in this document are hypothetical, but meant to be realistic. The goal of this document is **not** to show how different versions of the PTE are estimated in practice. Instead, the goal is to illustrate three conceptual points regarding the PTE:

1. The PTE can be interpreted in a variety of ways. However, for the PTE to have a relevant interpretation, unverifiable assumptions are **always** needed.
2. The PTE is often defined independent of baseline covariates. Without taking such baseline covariates into account, the PTE can often not be interpreted in the desired manner.
3. The validity of the non-causal interpretation of the PTE is inherently linked to the potential application trials. The validity of the causal interpretation does not involve assumptions regarding potential application trials.

In our hypothetical scenario, we have data from a single evaluation trial in which the efficacy of a vaccine is evaluated. In this trial, the occurrence of infection in the year after vaccination is the primary endpoint,

*KU Leuven - University of Leuven & Universiteit Hasselt, I-BioStat, B-3000 Leuven, Belgium, florian.stijven@kuleuven.be
†KU Leuven - University of Leuven & Universiteit Hasselt, I-BioStat, B-3000 Leuven, Belgium
‡KU Leuven - University of Leuven & Universiteit Hasselt, I-BioStat, B-3000 Leuven, Belgium

i.e., the true endpoint, $T$. At the same time, the neutralizing antibody level has been measured 2 weeks after vaccination. This serves as a potential surrogate endpoint, $S$. In this trial, the vaccine operates mainly through inducing an antibody response, hence, $S$ is an important mediator.

Independent of the evaluation trial, we consider two potential application trials. These are trials in which the results of the previous surrogacy analysis are used to replace the true endpoint with the surrogate endpoint. The population in these two application trials differs from the evaluation trial's population in terms of age and general health at baseline. In addition, the vaccine in the second application trial has a different mechanism of action. These differences are meant to emulate a real-life setting where, indeed, there may be important differences between trials for the same disease. This also puts the spotlight on the reasoning required for justifying the use of a potential surrogate endpoint in a new trial, and the potential pitfalls.

The remainder of this document is structured as follows. First, the data generating model underlying the evaluation trial is explained. For this trial, two versions of the PTE are then estimated and interpreted in a causal framework. Next, the data generating model underlying the two potential application trials is explained. The results of the surrogacy evaluation in the evaluation trial are then used to predict the treatment effects in these two application trials *using only surrogacy information*. The accuracy of these predictions is then related to (violations of) assumptions and the PTEs that were estimated in the evaluation trial.

```
# The tidyverse packages are loaded for data manipulation and plotting.
library(tidyverse)
# The mediation package is loaded for computing the PTE and related quantities.
library(mediation)
# Set a seed for reproducibility.
set.seed(1)
```

# 2 Evaluation Trial

## 2.1 Data Generating Model

The data generating model is based on the selection diagram in Figure 1. The nodes in that diagram correspond to the following variables:

- **Z**. This is the treatment indicator where $Z = 0$ corresponds to the control vaccine and $Z = 1$ corresponds to the experimental vaccine. Treatment is randomized in this scenario. Therefore, there are no arrows into $Z$.
- **S**. This is the surrogate endpoint, i.e., the neutralizing antibody level. Treatment has a causal effect on $S$, and $S$ has a causal effect on the true endpoint. Therefore, the neutralizing antibody level is a mediator for the effect of vaccination.
- **T**. This is the true endpoint where $T = 0$ corresponds to infection in the year following vaccination and $T = 1$ to no infection in this period.
- **X**. This represents the confounding variables for the causal effect of the neutralizing antibody level on the infection outcome. In this data generating model, there are only 2 confounding variables. These variables are measured at baseline.
  - Age, $X_1$. Older patients tend to have a weaker immune response to vaccination and are more susceptible to infection.
  - General health at baseline, $X_2$. Patients in better general health (higher $X_2$) tend to have a stronger immune response to vaccination and are less susceptible to infection.
- **A**. This is the trial indicator where $A = 1$ corresponds to the evaluation trial and $A = 0$ corresponds to the application trial.

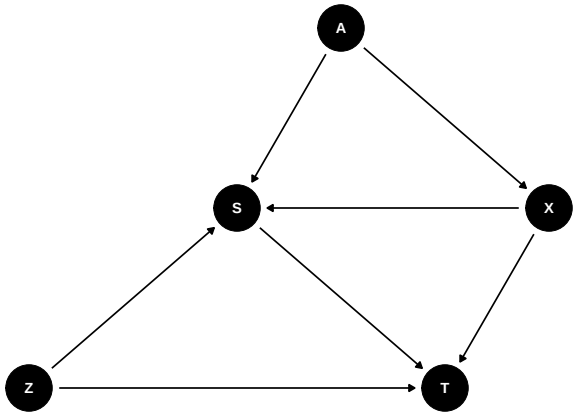


Figure 1: Selection diagram underlying the data generating model. The evaluation and first application trial are consistent with this diagram, but the second application trial is not. The assumptions in Lemma 4.2 of the paper are satisfied for trials that are consistent with this selection diagram.

We now discuss the distributions of these variables in the evaluation trial. Concomitant with these distributions, we provide the code that is used to generate the data in the evaluation trial.

#### 2.1.1 Treatment

The treatment is randomized in a 1:1 fashion such that there are 10.000 patients in each treatment arm.

```
# Sample size per treatment arm is fixed.
n = 1e4
# Construct integer vector where the first n elements are zero, and the next n
# elements 1.
Z = c(rep(0L, n), rep(1L, n))
```

#### 2.1.2 Age and General Health

As mentioned previously, age and general health are baseline covariates that confound the causal effect of $S$ on $T$. These variables are distributed in the evaluation trial as follows.

- Age, $X_1|A = 1 \sim N(35, 7^2)$. This population is thus quite young (in comparison with the application trials, see further).
- General health, $X_2|A = 1 \sim N(5, 5^2)$. This population has a good general health (in comparison with the application trials, see further).

Because of randomization, these two variables are independent of the treatment variable. However, the distribution of these variables is different in the application trials.

```
# Generate observations for the baseline covariates.
age = rnorm(n = 2 * n, mean = 35, sd = 7)
health = rnorm(n = 2 * n, mean = 5, sd = 5)

# Combine the treatment vector and the baseline covariates into a tibble.
eval_trial_tbl = tibble(
```

```
  Z = Z,
  age = age,
  health = health
)
```

### 2.1.3 Neutralizing Antibody Levels, $S$

As mentioned before, the vaccine in the evaluation trial mainly exerts its effect through inducing an antibody response. In other words, the vaccine reduces the risk of infection by increasing the patient's neutralizing antibody levels. In the evaluation trial, $S|Z, X, A = 1$ is a normal distribution with unit variance and mean $E(S|Z, X, A = 1) = 1.5 \cdot Z - 0.05 \cdot (X_1 - 45) + 0.15 \cdot X_2$.

```
# The surrogate endpoint is simulated taking into account the baseline
# covariates and the treatment assignment.
eval_trial_tbl = eval_trial_tbl %>%
  mutate(S = 1.5 * Z - 0.05 * (age - 45) + 0.15 * health + rnorm(n = 2 * n))
```

### 2.1.4 Infection Status, $T$

The true endpoint is a binary endpoint that depends on the treatment, the neutralizing antibody level, and the baseline covariates. Its distribution is completely determined by the mean, which depends on the mentioned variables in the following way,

$$E(T|Z, S, X, A = 1) = expit(0.8 + 0.1 \cdot Z + 0.5 \cdot S - 0.03 \cdot (X_1 - 45) + 0.1 \cdot X_2).$$

```
# The true endpoint is simulated in two steps. First, the logit(mean) as a
# linear function of the covariates is computed. Given this conditional mean,
# the true endpoint is sampled from a Bernoulli distribution.
eval_trial_tbl = eval_trial_tbl %>%
  mutate(
    eta = 0.8 + 0.1 * Z + 0.5 * S - 0.03 * (age - 45) + 0.1 * health,
    infection_free = rbinom(
      n = 2 * n,
      size = 1,
      prob = 1 / (1 + exp(-1 * eta))
    )
  )
```

## 2.2 Surrogacy Analysis

In this subsection, we first do several simple analyses of the evaluation trial that will help us to interpret the results regarding the PTE. Key concepts regarding the PTE, which are discussed in the paper, are repeated here for completeness. Next, two versions of the PTE are estimated with these data: (i) the "usual" PTE that ignores baseline covariates, $PTE_{WT}$, and (ii) the PTE that does take into account baseline covariates, $PTE^X_{WT}$. These estimates are then interpreted in the causal frameworks discussed in the paper.

### 2.2.1 Preliminaries

We start by computing the marginal treatment effect.

```
# Proportion infection-free after 1 year in the control group.
eval_trial_prop0 = mean(eval_trial_tbl$infection_free[eval_trial_tbl$Z == 0])
# Proportion infection-free after 1 year in the experimental group.
eval_trial_prop1 = mean(eval_trial_tbl$infection_free[eval_trial_tbl$Z == 1])
```

The proportions infection-free in the control and experimental groups are, respectively, 0.862 and 0.932. The risk difference is thus 0.07. In this context, the treatment effect is often summarized in the vaccine efficacy (VE): $VE = 1 - RR$ where $RR$ is the relative risk of infection in the experimental vs control group. The VE in this trial is 0.509.

In what follows, $PTE_{WT}$ is defined as $\frac{\Delta - \Delta_S}{\Delta}$ where

$$\Delta = E(T_1 - T_0 | A = 1)$$

and

$$\Delta_S = \int E(T_1 | S_1 = s, A = 1)\, dF_{S_0|A}(s|a = 1) - \int E(T_0 | S_0 = s, A = 1)\, dF_{S_0|A}(s|a = 1). \quad (1)$$

The above two quantities are termed the marginal (or total) treatment effect, and the residual treatment effect, respectively. In light of the general definition used throughout the paper, we thus have that $h(u) = u$ and $g(\cdot)$ is the expectation.

From the definition of $PTE_{WT}$, it follows that $PTE_{WT}$ is close to 1 when the residual treatment effect, $\Delta_S$, is close to 0. From the definition of $\Delta_S$ in (1) follows that $\Delta_S$ is close to 0 when the regression functions of $T$ on $S$ are similar in both treatment groups. Indeed, we can rewrite $\Delta_S$ as the weighted difference of these two regression functions,

$$\Delta_S = \int E(T_1 | S_1 = s, A = 1) - E(T_0 | S_0 = s, A = 1)\, dF_{S_0|A}(s|a = 1),$$

where the weights depend on the distribution of the surrogate in the control group. These two regression functions are plotted next in Figure 2.

```
eval_trial_tbl %>%
  ggplot() +
  geom_smooth(aes(x = S, y = infection_free, color = Treatment), se = FALSE) +
  geom_density(aes(x = S, color = Treatment)) +
  xlim(c(-2.5, 7.5)) +
  theme_bw() +
  ylab("Proportion Infection-Free") +
  xlab("Neutralizing Antibody Level")
```

These regression functions differ between the treatment groups. Indeed, the regression function in the control group is larger than in the experimental group for all values of $S$. Hence, $PTE_{WT}$ will be larger than 1.

We can also include baseline covariates in the definition of the proportion explained. This leads to $PTE^X_{WT} = \frac{\Delta - E[\Delta_S(X)]}{\Delta}$ where

$$\Delta_S(x) = \int E(T_1 | S_1 = s, X = x, A = 1) - E(T_0 | S_0 = s, X = x, A = 1)\, dF_{S_0|X,A}(s|X = x, a = 1).$$

If $X$ contains all confounders, $E[\Delta_S(X)]$ corresponds to the interventional direct effect and $\Delta - E[\Delta_S(X)]$ to the interventional indirect effect. Under the additional cross-world independence assumption, these effects correspond to their natural counterparts. The ratio of the interventional/natural indirect effect and the total effect is classically termed the proportion mediated. As for the PTE, the *proportion* mediated is actually a misnomer because it is not a proportion.

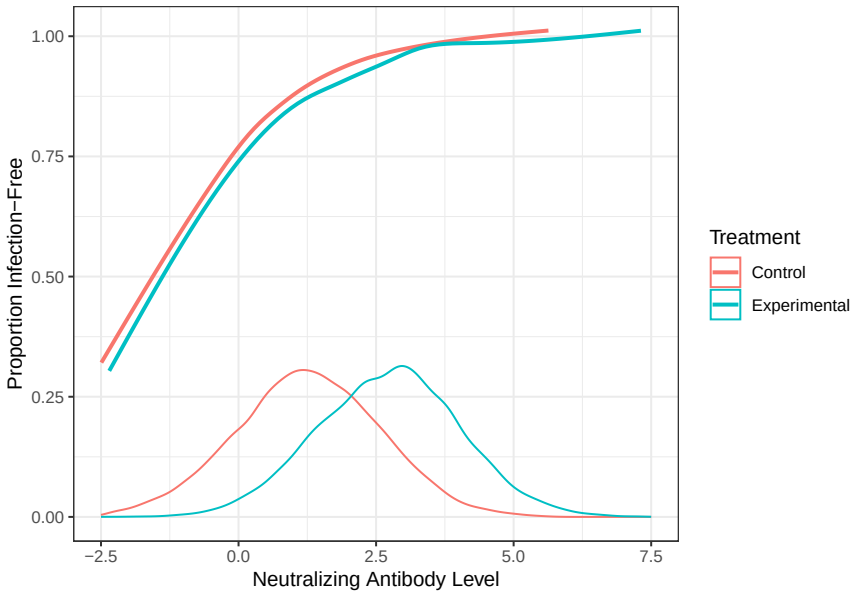


Figure 2: Regression function of the regression of the true endpoint on the surrogate endpoint in each treatment group separately. These regression functions are estimated by local regression. The corresponding density estimates of the distribution of the surrogate endpoint are superimposed on this plot.

### 2.2.2 Estimation of the PTE

As explained in Appendix E of the review paper, the $PTE_{WT}$ can be estimated with existing software for causal mediation analysis. In this document, we use the `mediate()` function from the `mediation` package for this purpose. For a conventional mediation analysis, this function relies on two regression models.

1. A regression model for the mediator given the treatment and the confounders measured at baseline.
2. A regression model for the true endpoint given the treatment, the mediator and the confounders measured at baseline.

For estimating the $PTE_{WT}$, we replace the mediator with the surrogate, and we do not include baseline covariates. For estimating the $PTE^X_{WT}$, we do include baseline covariates.

The corresponding models are fitted in turn next. For the surrogate, we choose a linear regression model. For the true endpoint, we choose a logistic regression model. When we include the baseline covariates in these models, the models will be correctly specified. However, the logistic regression model will be misspecified when we do not include the baseline covariates.

```
# Fit the models that do not include the baseline covariates.
surrogate_marg_model = lm(S ~ Z, data = eval_trial_tbl)
true_endpoint_marg_model = glm(infection_free ~ Z + S,
                               data = eval_trial_tbl,
                               family = binomial())
# Fit the models that include the baseline covariates.
surrogate_model = lm(S ~ Z + age + health, data = eval_trial_tbl)
true_endpoint_model = glm(infection_free ~ Z + S + age + health,
                          data = eval_trial_tbl,
                          family = binomial())
```

We can now estimate $PTE_{WT}$ and $PTE_{WT}^{X}$. This is done in the following 2 code chunks. For completeness, we call the summary method on the object returned by `mediate()`. This method returns numerous estimates, which are discussed in turn. We only focus on the estimates with `(treated)` in their name. These estimates correspond to components of $PTE_{WT}$ as defined previously. If we switch both treatment arms, then the estimates with `(control)` in their name are obtained.

- Average causal mediation effect, `ACME (treated)`. This is often termed the indirect effect. The estimator in `mediate()` is only consistent for the indirect effect if all confounders are included in the models in `mediate()`.
  - This estimator is in any case consistent for $\Delta - \Delta_S$ in the definition of $PTE_{WT}$ when no baseline covariates have been included and the models are correctly specified.
  - This estimator is in any case consistent for $\Delta - E\left[\Delta_S(X)\right]$ in the definition of $PTE_{WT}^{X}$ when baseline covariates have been included and the models are correctly specified.
- Average direct effect, `ADE (treated)`. As for the indirect effect, the estimator in `mediate()` is only consistent for the direct effect if all confounders have been included in the models in `mediate()`.
  - This estimator is in any case consistent for $\Delta_S$ in the definition of $PTE_{WT}$ when no baseline covariates have been included and the models are correctly specified.
  - This estimator is in any case consistent for $E\left[\Delta_S(X)\right]$ in the definition of $PTE_{WT}^{X}$ when baseline covariates have been included and the models are correctly specified.
- `Total effect`. This estimator is in any case consistent for $\Delta$ because treatment is randomized.
- Proportion mediated, `Prop. Mediated (treated)`. This estimator is the ratio of `ACME (treated)` and `Total effect`.
  - This is a consistent estimator for $PTE_{WT}$ when no baseline covariates have been included and the models are correctly specified.
  - This is a consistent estimator for $PTE_{WT}^{X}$ when baseline covariates have been included and the models are correctly specified.

```
PTE_WT = mediate(
  # Model for the mediator, i.e., the surrogate.
  model.m = surrogate_marg_model,
  # Model for the outcome, i.e., the true endpoint.
  model.y = true_endpoint_marg_model,
  # Number of MC samples for the numerical approximation.
  sims = 1e3,
  # Name of the treatment variable.
  treat = "Z",
  # Name of the mediator (surrogate) variable.
  mediator = "S"
)
summary(PTE_WT)
```

```
##
## Causal Mediation Analysis
##
## Quasi-Bayesian Confidence Intervals
##
##                         Estimate 95% CI Lower 95% CI Upper p-value
## ACME (control)            0.0838       0.0788         0.09  <2e-16 ***
## ACME (treated)            0.1011       0.0929         0.11  <2e-16 ***
## ADE (control)            -0.0323      -0.0448        -0.02  <2e-16 ***
## ADE (treated)            -0.0150      -0.0206        -0.01  <2e-16 ***
```

```
## Total Effect               0.0688       0.0605         0.08  <2e-16 ***
## Prop. Mediated (control)   1.2176       1.1210         1.34  <2e-16 ***
## Prop. Mediated (treated)   1.4700       1.2605         1.72  <2e-16 ***
## ACME (average)             0.0924       0.0864         0.10  <2e-16 ***
## ADE (average)             -0.0237      -0.0326        -0.01  <2e-16 ***
## Prop. Mediated (average)   1.3438       1.1910         1.53  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sample Size Used: 20000
##
##
## Simulations: 1000
```

```r
PTE_WT_X = mediate(
  model.m = surrogate_model,
  model.y = true_endpoint_model,
  sims = 1e3,
  treat = "Z",
  mediator = "S"
)
summary(PTE_WT_X)
```

```
##
## Causal Mediation Analysis
##
## Quasi-Bayesian Confidence Intervals
##
##                          Estimate 95% CI Lower 95% CI Upper p-value
## ACME (control)            0.06561      0.06077         0.07  <2e-16 ***
## ACME (treated)            0.06194      0.05429         0.07  <2e-16 ***
## ADE (control)             0.00903     -0.00376         0.02    0.16
## ADE (treated)             0.00536     -0.00218         0.01    0.16
## Total Effect              0.07097      0.06294         0.08  <2e-16 ***
## Prop. Mediated (control)  0.92432      0.83231         1.03  <2e-16 ***
## Prop. Mediated (treated)  0.87046      0.72259         1.06  <2e-16 ***
## ACME (average)            0.06378      0.05769         0.07  <2e-16 ***
## ADE (average)             0.00719     -0.00297         0.02    0.16
## Prop. Mediated (average)  0.89739      0.77676         1.05  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sample Size Used: 20000
##
##
## Simulations: 1000
```

### 2.2.3 Interpretation of Estimates

The relevant results are summarized in the Table 1. We discuss these estimates in turn. Sampling variability is ignored.

Table 1: Estimates for the proportions explained and related quantities. $\Delta_1$ and $\Delta_2$ are the estimates for marginal treatment effect, ignoring and including the baseline covariates, respectively.

| $\Delta_S$ | $\Delta_1$ | $PTE_{WT}$ | $E(\Delta_S(X))$ | $\Delta_2$ | $PTE^X_{WT}$ |
|---|---|---|---|---|---|
| -0.015 | 0.069 | 1.47 | 0.005 | 0.071 | 0.87 |

- The estimated marginal treatment effect, $\Delta$, quantified by the risk difference, is 0.069 or 0.071. The estimate changes slightly when ignoring versus including baseline covariates. This provides the reference to compare the other estimates with.
- The residual treatment effect, $\Delta_S$, is $-0.015$. Under the conditional independence assumption, $S_0 \perp T_1|S_1$ and $S_1 \perp T_0|S_0$, this is equal to the *causal residual treatment effect*, $\Delta_{S,c}$. If $S$ is a principal surrogate, then $\Delta_{S,c} = 0$ by Lemma 3.1 in the review paper.
  - This quantity **cannot** be interpreted as a direct effect because there is confounding for the causal effect of $S$ on $T$ by age and general health.
- The $PTE_{WT}$ is 1.47. In this context, principal surrogacy may not be satisfied exactly. So, $\Delta_{S,c}$ may be close to zero, but not exactly zero. The $PTE_{WT} = \frac{\Delta - \Delta_{S,c}}{\Delta}$ provides a natural scale to assess closeness to zero. Indeed, closeness to zero is quantified relative to the marginal treatment effect.
  - Under the conditional independence assumption, $S$ is thus not very close to being a principal surrogate.
- The $E(\Delta_S(X))$ is 0.005. Because $X$ contains all confounders, this is a valid estimate of the interventional direct effect. If we additionally make the cross-world independence assumption, this is also a valid estimate of the natural direct effect.
  - The direct effect of treatment is small. Therefore, a large portion of the treatment effect is mediated by the surrogate.
- The $PTE^X_{WT}$ is 0.870. Because we have controlled for all confounders in the computation of this quantity, this is a valid estimate of the proportion mediated. Most of the treatment effect is thus mediated by the surrogate endpoint.

# 3 Application Trials

In this section, we introduce two potential application trials. First, the data generating model for these trials is explained. Next, we show how the estimates in Table 1 can be interpreted in reference to these application trials. Indeed, under a set of assumptions, these estimates tell us how good we can predict treatment effects in new trials using only the surrogate endpoint and possibly baseline covariates.

## 3.1 Data Generating Model

In this subsection, we discuss the differences between the evaluation and application trials. As in the evaluation trial, the application trials have 10.000 patients in each treatment arm.

### 3.1.1 Application Trial 1

The first application trial is consistent with the selection diagram in Figure 1. This application trial thus only differs from the evaluation trial in the distribution of the baseline covariates and the distribution of $S|Z, X$:

- The population in application trial 1 is older and less healthy than in the evaluation trial.
  - Age, $X_1|A = 0 \sim N(40, 5^2)$.
  - General health, $X_2|A = 0 \sim N(2, 4^2)$
- The control vaccine is assumed to be the same as in the evaluation trial. Therefore, we have that $S|Z = 0, X = x, A = 0$ and $S|Z = 0, X = x, A = 1$ are equal in distribution for all possible values of $x$. However, the experimental vaccine now has a smaller effect on the neutralizing antibody level. Therefore, $S|Z, X, A = 0$ is a normal distribution with unit variance and mean $E(S|Z, X, A = 0) = 0.75 \cdot Z - 0.05 \cdot (X_1 - 45) + 0.15 \cdot X_2$.

```
# Generate observations for the baseline covariates. These distributions differ
# from those in the evaluation trial.
age = rnorm(n = 2 * n, mean = 40, sd = 5)
health = rnorm(n = 2 * n, mean = 2, sd = 4)

# Combine the treatment vector and the baseline covariates into a tibble. Note
# that the treatment variable, Z, is the same as for the evaluation trial.
appl_trial1_tbl = tibble(
  Z = Z,
  age = age,
  health = health
)
# The surrogate endpoint is simulated. The distribution of the surrogate as a
# function of treatment and baseline covariates is different from the evaluation
# trial.
appl_trial1_tbl = appl_trial1_tbl %>%
  mutate(S = 0.75 * Z - 0.05 * (age - 45) + 0.15 * health + rnorm(n = 2 * n))
# The true endpoint is simulated as in the evaluation trial.
appl_trial1_tbl = appl_trial1_tbl %>%
  mutate(
    eta = 0.8 + 0.1 * Z + 0.5 * S - 0.03 * (age - 45) + 0.1 * health,
    infection_free = rbinom(
      n = 2 * n,
      size = 1,
      prob = 1 / (1 + exp(-1 * eta))
    )
  )
# Add more informative variable names and labels.
appl_trial1_tbl = appl_trial1_tbl %>%
  mutate(Treatment = factor(
    Z,
    levels = 0:1,
    labels = c("Control", "Experimental")
  ))
```

### 3.1.2 Application Trial 2

The second application trial is **not** consistent with the selection diagram in Figure 1. This application trial differs from the evaluation trial in the distribution of the baseline covariates, the distribution of $S|Z, X$, and the distribution of $T|Z, X, S$. The latter is not allowed by the selection diagram in Figure 1.

- The population in application trial 2 is the same as in application trial 1. So, the distribution of baseline covariates is also the same.

- The control vaccine is assumed to be the same as in the evaluation trial. Therefore, we again have that $S|Z=0, X=x, A=0$ and $S|Z=0, X=x, A=1$ are equal in distribution for all possible values of $x$. However, the experimental vaccine now mainly operates through cellular immunity. This type of immunity has not been measured. Therefore, $S|Z, X, A=0$ is a normal distribution with unit variance and mean $E(S|Z, X, A=0) = 0.1 \cdot Z - 0.05 \cdot (X_1 - 45) + 0.15 \cdot X_2$. Indeed, the treatment effect on the surrogate is now much smaller.
- Since the experimental vaccine mainly operates through cellular immunity, which was not measured, the direct effect is now important. This is reflected in the distribution of $T|Z, X, S, A=0$. Indeed, the corresponding mean is

$$E(T|Z, S, X, A=0) = expit\left(1 + 1 \cdot Z + 0.5 \cdot S - 0.03 \cdot (X_1 - 45) + 0.1 \cdot X_2\right)$$

```
# Generate observations for the baseline covariates. These distributions are the
# same as in application trial 1.
age = rnorm(n = 2 * n, mean = 40, sd = 5)
health = rnorm(n = 2 * n, mean = 2, sd = 4)

# Combine the treatment vector and the baseline covariates into a tibble. Note
# that the treatment variable, Z, is the same as for the evaluation trial.
appl_trial2_tbl = tibble(
  Z = Z,
  age = age,
  health = health
)
# The surrogate endpoint is simulated. The distribution of the surrogate as a
# function of treatment and baseline covariates is different from the evaluation
# trial.
appl_trial2_tbl = appl_trial2_tbl %>%
  mutate(S = 0.1 * Z - 0.05 * (age - 45) + 0.15 * health + rnorm(n = 2 * n))
# The true endpoint is simulated. This conditional distribution differs from the
# evaluation trial. This is not consistent with the previously mentioned
# selection diagram.
appl_trial2_tbl = appl_trial2_tbl %>%
  mutate(
    eta = 1 + 1 * Z + 0.5 * S - 0.03 * (age - 45) + 0.1 * health,
    infection_free = rbinom(
      n = 2 * n,
      size = 1,
      prob = 1 / (1 + exp(-1 * eta))
    )
  )
# Add more informative variable names and labels.
appl_trial2_tbl = appl_trial2_tbl %>%
  mutate(Treatment = factor(
    Z,
    levels = 0:1,
    labels = c("Control", "Experimental")
  ))
```

## 3.2 Prediction of Treatment Effects

As explained in the paper, $PTE_{WT}$ and $PTE_{WT}^{X}$ each imply a prediction rule $\mu$ and $\mu_X$ for the true endpoint. In the context of the application trials, these are prediction rules for the true endpoint, given the observed

surrogate endpoint and possibly baseline covariates:

$$\mu(s) = E(T_1|S_1 = s, A = 1) \text{ and } \mu_x(s) = E(T_1|S_1 = s, X = x, A = 1).$$

The above functions are, respectively, transformations of $S$ and $(S, X)'$: $\tilde{S} = \mu(S)$ and $\tilde{S}_X = \mu_X(S)$.

Under a set of assumptions, $PTE_{WT}$ ($PTE^X_{WT}$) quantifies the relative discrepancy between the treatment effect on $\tilde{S}$ ($\tilde{S}_X$) and on $T$ *in the evaluation trial*. This is shown in Lemma 4.1 in the paper. Therefore, the treatment effect on $\tilde{S}$ ($\tilde{S}_X$) is an accurate prediction for the treatment effect on $T$ *in the evaluation trial* when $PTE_{WT} \approx 1$ ($PTE^X_{WT} \approx 1$). In Lemma 4.2, it is shown that $PTE_{WT}$ ($PTE^X_{WT}$) also quantifies the error in these predictions for **new trials** under additional assumptions. In Table 2, the "true" and predicted treatment effects are summarized for both application trials. The predicted treatment effects are defined as follows,

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

and

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

where unknown quantities are replaced with their estimates.

When baseline covariates are not included in the prediction rule, we do not expect the true and the predicted treatment effects to match. Indeed, $PTE_{WT}$ is not close to 1 in the evaluation trial (see Table 1). So, by Lemma 4.1, the predicted treatment effect in the evaluation trial, using a prediction rule estimated in that same trial, is not accurate. Therefore, we should **not** expect better predictions in application trials. This is confirmed by the estimates in Table 2.

In the evaluation trial, we saw that $PTE^X_{WT} \approx 1$. Hence, the predicted treatment effect in the evaluation trial matches the true treatment effect, *when taking baseline covariates into account*. So, good predictions in application trials *might be* possible. Lemma 4.2 tells us under which conditions we can expect accurate predictions in new trials. These conditions are only satisfied in application trial 1 conditional on baseline covariates. Indeed, the corresponding predicted treatment effect shown in Table 2 is 0.051, which is very close to the true treatment effect of 0.053. In contrast, application trial 2 does not satisfy the assumptions of Lemma 4.2. So, we do not expect an accurate prediction. The latter is confirmed in Table 2.

```
# Combine application trials into a single data set.
appl_trials_tbl =
  bind_rows(
    appl_trial1_tbl %>%
      mutate(Trial = "Application Trial 1"),
    appl_trial2_tbl %>%
      mutate(Trial = "Application Trial 2")
  )
# Compute the predicted treatment effects in both application trials.
appl_trials_tbl = appl_trials_tbl %>%
  mutate(
    S_tilde = predict(
      # Logistic regression model without baseline covariates. This model gives us
      # the prediction rule for the true endpoint, given the surrogate.
      true_endpoint_marg_model,
      newdata = pick(everything()) %>%
        mutate(Z = 1L), # The prediction rule from the evaluation trial is used.
      type = "response" # Prediction on the probability scale.
    ),
    S_tilde_X = predict(
      # Logistic regression model with baseline covariates. This models gives us
      # the prediction rule for the true endpoint, given the surrogate and
      # baseline covariates (age and health).
```

Table 2: True and predicted treatment effects in the application trials. The mean difference is used as effect measure in this table.

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

Table 3: True and predicted treatment effects in the application trials. The vaccine efficacy is used as effect measure in this table.

| Trial | $VE$ | $\tilde{VE}^*$ | $\tilde{VE}^*_{EX}$ |
|---|---|---|---|
| Application Trial 1 | 0.241 | 0.326 | 0.240 |
| Application Trial 2 | 0.589 | 0.061 | 0.049 |

```
      true_endpoint_model,
      newdata = pick(everything()) %>%
        mutate(Z = 1L), # The prediction rule from the evaluation trial is used.
      type = "response" # Prediction on the probability scale.
    )
  )
```

Alternatively, we can look at the results in terms of (predicted) VE. While Lemma 4.2 still holds for VE as effect measure for the $PTE_{WT}$, issues arise for $PTE^X_{WT}$. Indeed, VE is a non-collapsible effect measure. So, the conditional residual effect $\Delta_S(x)$ and its expectation cannot be compared with the marginal treatment effect. In principle, we could redefine the total treatment effect as $E(VE(X))$ to allow for proper comparisons. However, it is unclear how such quantities can be properly interpreted. Therefore, the predicted treatment effects are redefined as follows,

$$\tilde{VE}^* = 1 - \frac{1 - E\left(\mu(S_1)|A=0\right)}{1 - E\left(\mu(S_0)|A=0\right)}$$

and

$$\tilde{VE}^*_{EX} = 1 - \frac{1 - E\left(\mu_X(S_1)|A=0\right)}{1 - E\left(\mu_X(S_0)|A=0\right)}.$$

As before, we merely replace $T$ with a prediction of $T$ in the definition of the effect measure.

These results are summarized in Table 3. These results are very much in line with Table 2. This shows that good predictions of the treatment effect on one scale, can translate to good predictions of the treatment effect on another scale.

### 3.3 Other Application Trials

The two application trials considered in this document have been cherry-picked to illustrate potential pitfalls in using $PTE_{WT}$ and $PTE^X_{WT}$ as measures of surrogacy. These two trials indeed represent extreme situations. In the first application trial, the treatment operated through the same mechanism (neutralizing antibodies) as the evaluation trial. The second application trial's treatment operated through a very different mechanism: cellular immunity.

Reality is much more complex than these two application trials might suggest. Many application trials will arguably be situated in between the above two extremes. Indeed, there exist a plethora of different types of neutralizing antibodies. Additionally, the same "type" of antibody can potentially be measured by different methods. All these factors should be taken into account in the evaluation of the assumptions of Lemma 4.2

in the paper. Indeed, a $PTE_{WT}$ or $PTE_{WT}^{X}$ close to 1 is *some* evidence of surrogacy, but the relevance of this result depends heavily on unverifiable extrapolating assumptions. Clearly, this is a very difficult endeavor in which no consensus might be reached among scientists. The only empirical way forward is the extension to multiple trials as touched upon in the Discussion of the paper.

# 4 Conclusion

In this document, we have generated three vaccine trials. The first trial is the evaluation trial. In the evaluation trial, the estimates for $PTE_{WT}$ and $PTE_{WT}^{X}$ differ considerably; only the latter is close to 1. At the same time, both estimates can be given a valid causal interpretation in this trial. Indeed, $PTE_{WT}$ quantifies the degree of principal surrogacy *under the conditional independence assumption*. However, it is not a valid estimate for the proportion mediated because there is confounding that has not been adjusted for. In contrast, the estimate for $PTE_{WT}^{X}$ in the evaluation trial is a valid estimate for the proportion mediated because there is no unmeasured confounding. The proportion mediated is close to 1 in the evaluation trial. This means that the treatment decreases the probability of infection primarily by increasing the neutralizing antibody level. Even though these estimates correspond to well-defined causal quantities, it is unclear how these results translate to new trials.

Besides a causal interpretation, $PTE_{WT}$ and $PTE_{WT}^{X}$ also have a non-causal interpretation that is related to the prediction of treatment effects using only surrogate information. These two proportions explained correspond to a prediction rule for the true endpoint given the surrogate (and possibly baseline covariates). Indeed, Lemma 4.1 shows us that these proportions explained quantify how accurate the treatment effect on $T$ is predicted by the respective prediction rules. Lemma 4.2 states the conditions under which $PTE_{WT}$ and $PTE_{WT}^{X}$ quantify how accurate the same prediction rules predict the treatment effect on $T$ in the application trial. The results in Tables 2 and 3 show how the predictions can fail when these assumptions do not hold.

As shown in this document, $PTE_{WT} \approx 1$ and/or $PTE_{WT}^{X} \approx 1$ does not guarantee accurately predicted treatment effects in new trials. However, we can *empirically* evaluate the extrapolating assumptions by considering multiple trials. Indeed, a single prediction rule can be estimated using one or more trials. Next, this prediction rule leads to multiple predicted treatment effects and corresponding prediction errors. There is now replication of the prediction errors, so formal statistical inference regarding the accuracy of predictions (in new unobserved trials) is possible.