Research Article

Andrea Callegaro*, Fabian Tibaldi and Dean Follmann

Principal surrogates in context of high vaccine efficacy

<https://doi.org/10.1515/scid-2020-0003>

Received March 20, 2020; accepted August 21, 2021; published online November 4, 2021

Abstract

Objectives: The use of correlates of protection (CoPs) in vaccination trials offers significant advantages as useful clinical endpoint substitutes. Vaccines with very high vaccine efficacy (VE) are documented in the literature (95% or above). Callegaro, A., and F. Tibaldi. 2019. “Assessing Correlates of Protection in Vaccine Trials: Statistical Solutions in the Context of High Vaccine Efficacy.” *BMC Medical Research Methodology* 19: 47 showed that the rare infections observed in the vaccinated groups of these trials poses challenges when applying conventionally-used statistical methods for CoP assessment such as the Prentice criteria and meta-analysis. The objective of this work is to investigate the impact of this problem on another statistical method for the assessment of CoPs called Principal stratification.

Methods: We perform simulation experiments to investigate the effect of high vaccine efficacy on the performance of the Principal Stratification approach.

Results: Similarly to the Prentice framework, simulation results show that the power of the Principal Stratification approach decreases when the VE grows.

Conclusions: It can be challenging to validate principal surrogates (and statistical surrogates) for vaccines with very high vaccine efficacy.

Keywords: causal inference; high vaccine efficacy; principal surrogate; surrogate endpoint; vaccine clinical trial.

Introduction

An important factor influencing the duration and complexity of clinical trials is the choice of the endpoint used to assess vaccine efficacy. It would be extremely convenient to replace a late, costly or rare true endpoint by an immunological surrogate, which is measured earlier, cheaper, or more frequently. However, from a regulatory perspective, a surrogate endpoint (called sometimes Surrogate of protection, or Correlate of Protection) is not considered acceptable for the determination of efficacy, unless it has been validated, i.e. shown to predict clinical benefit. Prentice (1989) introduced a formal definition of surrogacy based on the concept of mediation in a single-trial setting. Although appealing, Prentice's definition and criteria received criticism, such as

*Corresponding author: Andrea Callegaro, Statistics, GSK Vaccines, Rue de l'institut, 89, 1330, Rixensart, Belgium, E-mail: andrea.x.callegaro@gsk.com. <https://orcid.org/0000-0003-4805-5118>

Fabian Tibaldi, Statistics, GSK Vaccines, Rixensart, Belgium

Dean Follmann, National Institute of Allergy and Infectious, Rockville, USA

(i) the assumption that the surrogate explains 100% of the VE is too restrictive; (ii) the approach can be susceptible to post-randomization selection bias; (iii) the immune response cannot be constant in the control group; etc. (Burzykowski, Molenberghs, and Buyse 2005; Frangakis and Rubin 2002). In subsequent decades, many statistical methods have been proposed for the evaluation of surrogate endpoints, most of them framed within the causal inference (Follmann 2006; Frangakis and Rubin 2002; Gilbert, Qin, and Self 2008) and meta-analytic paradigms (Buyse et al. 2000; Daniels and Hughes 1997; Gail et al. 2000).

Although not common, vaccines with very high efficacy are documented in the literature (Black et al. 2000; Lin et al. 2001; Mitra et al. 2016; Phua et al. 2012; Prymula et al. 2014; Wei et al. 2016). These include the salmonella typhi vi conjugate (Mitra et al. 2016), or the combined measles-mumps-rubella-varicella immunisation (Prymula et al. 2014). Assessing Correlate of Protections (CoPs) in the context of high Vaccine Efficacy (VE) using classical statistical methods is problematic. Indeed, a very small number of cases/infections (corresponding to the vaccinated groups) can trigger considerable issues for such statistical models. There is therefore a need to evaluate the statistical methods for CoP assessment to the context of high efficacy vaccines. Callegaro and Tibaldi (2019) showed that the validation of a surrogate endpoint using the Prentice criteria and meta-analytic frameworks (by randomized subgroups in single trial setting) can be problematic in case of high VE because of the rare events available in the vaccine group. The aim of this paper is to evaluate the performance of the causal framework, specifically the Principal Surrogate approach (Follmann 2006; Gilbert, Qin, and Self 2008) in case of high VE.

Methods

The Prentice criteria

The following set of notation will be used throughout the manuscript: Y_j and S_j are random variables denoting the observed clinical (binary) and the surrogate endpoint for subject $j = 1, \dots, n$ and Z_j is a binary treatment indicator ($Z = 1$ for treatment and $Z = 0$ for control group).

Key concepts, including the hypothesis-testing approach to the validation of substitute endpoints using randomised clinical trial data, were introduced by Prentice in 1989 (Prentice 1989). Prentice's four criteria for the validation of a surrogate endpoint can be evaluated using the following 4 models:

$$\begin{aligned}\text{logit}(P(Y_i = 1)) &= \mu_T + \beta Z_i, \\ S_i &= \mu_S + \alpha Z_i + \varepsilon_{S_i}, \\ \text{logit}(P(Y_i = 1)) &= \mu + \gamma S_i, \\ \text{logit}(P(Y_i = 1)) &= \tilde{\mu}_T + \beta_S Z_i + \gamma_Z S_i.\end{aligned}$$

In this paper we will mainly focus on criterion 4. This criterion is met if the null hypothesis $H_{01}: \gamma_Z = 0$ is rejected (p-value(s) $< \alpha$) and the null hypothesis $H_{02}: \beta_S = 0$ is not rejected (p-value(z) $\geq \alpha$). Note that when this criterion is met ($\beta_S = 0$ and $\gamma_Z \neq 0$), then model 4 degenerates to model 3. The significance level α is not adjusted for multiplicity because the null hypothesis is the intersection of two null hypothesis.

Principal surrogate framework

Many causal inference approaches/methods have been published in the literature. In what follows, we describe the Vaccine Efficacy Framework of Follmann and Gilbert (Follmann 2006; Gilbert, Qin, and Self 2008). Since S_i can be affected by treatment, there are 2 naturally occurring counterfactual values of S_i : $S_i(1)$ under treatment, and $S_i(0)$ under control. The observed clinical endpoint (binary) is denoted by Y_i and the counterfactual values are $Y_i(1)$ under treatment, and $Y_i(0)$ under control. Criteria for S to be a good surrogate are based on risk estimands that condition on the potential surrogate responses (Gilbert, Qin, and Self 2008)

$$\begin{aligned}\text{risk}_{(1)}(s(1), s(0)) &= \Pr(Y(1) = 1 | S(1) = s(1), S(0) = s(0)) \\ \text{risk}_{(0)}(s(1), s(0)) &= \Pr(Y(0) = 1 | S(1) = s(1), S(0) = s(0))\end{aligned}$$

A contrast in $\text{risk}_{(1)}(s(1), s(0))$ and $\text{risk}_{(0)}(s(1), s(0))$ is a causal effect on the clinical endpoint. A classical contrast used in vaccines is the Vaccine Efficacy (VE)

$$\text{VE}(s(1), s(0)) = 1 - \frac{\Pr(Y(1) = 1 | S(1) = s(1), S(0) = s(0))}{\Pr(Y(0) = 1 | S(1) = s(1), S(0) = s(0))}$$

A Principal Surrogate (PS) is a biomarker satisfying two conditions: causal necessity

$$\text{VE}(s(1), s(0)) = 0 \quad \text{for all } s(1) = s(0)$$

and Wide Effect Modification (WEM) which means that

$$\text{VE}(s(1), s(0)) \quad \text{increasing in } s(1) - s(0),$$

WEM is similar in spirit to the Individual Causal Association (ICA) (Alonso et al. 2015), which is the correlation between the individual causal effect on the endpoint and on the surrogate.

In this paper we only focus on WEM. In fact, current works (Gabriel and Follmann 2016; Gabriel and Gilbert 2014; Gilbert, Qin, and Self 2008; Huang and Gilbert 2011; Wolfson and Gilbert 2010) suggest that WEM criterion is of primary importance for a biomarker to be a PS. Furthermore, Alonso et al. (2015) showed that the average causal necessity definition may be extremely restrictive.

Estimating VE: Assumptions A1–A3 (A1: Stable unit treatment value assumption; A2: Ignorable treatment assignments; A3: Equal individual clinical risk up to the time of surrogate measurements) imply that $\text{risk}(Z)(s(1), s(0))$ would be identified if we knew the potential outcomes $S_i(Z)$ of subjects assigned the opposite treatment $1 - Z$ (Wolfson and Gilbert 2010)

$$\text{risk}_{(1)}(s(1), s(0)) = \Pr(Y = 1 | Z = 1, S(1) = s(1), S(0) = s(0))$$

$$\text{risk}_{(0)}(s(1), s(0)) = \Pr(Y = 1 | Z = 0, S(1) = s(1), S(0) = s(0)).$$

It follows that it is necessary to impute (or integrate out) the missing potential biomarkers. The risk can be modeled using the following logistic model

$$\text{logit}(P(Y_i = 1 | z_i, s_{1i}, s_{0i})) = \beta_0 + \beta_z z_i + \beta_{s(1)} s_{1i} + \beta_{s(1)z} s_{1i} z_i + \beta_{s(0)} s_{0i} + \beta_{s(0)z} s_{0i} z_i.$$

The model can be simplified in case of a Constant Biomarker ($S_i(0) = c$)

$$\text{logit}(P(Y_i = 1 | z_i, s_{1i})) = \beta_0 + \beta_z z_i + \beta_{s(1)} s_{1i} + \beta_{s(1)z} s_{1i} z_i, \quad (1)$$

where the VE curve is used

$$\text{VE}(s(1)) = 1 - \frac{\Pr(Y = 1 | Z = 1, S(1) = s(1))}{\Pr(Y = 1 | Z = 0, S(1) = s(1))}.$$

The constant biomarker assumption is reasonable when subjects have been selected to have no meaningful exposure to the pathogen, so that $S(0) = 0$. Examples include HIV (Follmann 2006) or varicella vaccine trials (Chan et al. 2002). This assumption is also reasonable for populations exposed to the pathogen when the biomarker S_i is the log10 Fold-Increase from baseline (FI_i), which is the difference between the log10 post (A_i) and the log10 baseline (B_i) values ($FI_i = A_i - B_i$).

Missing values imputation/integration: The key challenge in estimating these risk estimands is solving the problem of conditioning on counterfactual values that are not observable. This involves integrating out (or imputing) missing values based on some models, and under some set of assumptions and/or trial augmentations. Gilbert, Qin, and Self (2008) and Follmann (2006) proposed to use the estimated maximum likelihood followed by bootstrap. Huang, Gilbert, and Wolfson (2013) suggested a pseudoscore estimation procedure that does have a closed form variance estimator. Miao et al. (2013) used a multiple imputation approach. In this paper we fit model 1 using the method implemented in the R package `pseval` (Sachs and Gabriel 2016): Baseline Immunogenicity Predictor (BIP); parameters estimated using estimated maximum likelihood (missing information is integrated out) and the variance is estimated by bootstrap. Rcode is provided in the Appendix. This approach is similar in spirit to the method used in Follmann (2006).

Results

Simulations of Callegaro and Tibaldi (2019)

To evaluate the impact of high vaccine efficacy on the PS validation, we repeated the simulations of Callegaro and Tibaldi (2019). The Dunning regression model (Dunning 2006) was used to simulate the data in an ideal CoP setting, where the treatment effect is fully explained by the post values (A_i) as follows:

$$P(Y_i = 1|\pi, A_i) = \pi \frac{e^{\mu + \gamma A_i}}{1 + e^{\mu + \gamma A_i}}. \quad (2)$$

Here, π can be interpreted as the probability of being exposed to the disease. This model corresponds to the classical logistic model when all subjects are exposed ($\pi = 1$).

Simulations were run using the following parameter assumptions: total sample size $n = 5,000$, 1:1 randomization, $\pi = 0.1$, $\mu = 8.3$, $\gamma = \log(1 - 0.95)$; the immune response post vaccination is normally distributed $A|Z = 0 \sim N(3, 0.2)$ in the placebo group and $A|Z = 1 \sim N(3 + \Delta, 0.2)$ in the vaccine group, where $\Delta = 0.33, 0.75, 1, 1.5$. The value of the immune response at baseline is generated as $B \sim N(3, 0.2)$ with correlation between A and B of 0.90 in the placebo group and 0.50 in the vaccine group (0.2 is the variance of the normal distribution). For each scenario, 1,000 clinical trials were simulated.

We fit Prentice model 4 on the simulated data with Fold-Increase ($S_i = FI_i = A_i - B_i$) as surrogate adjusting for the baseline (B_i) using logit regression

$$\text{logit}(P(Y_i = 1)) = \tilde{\mu}_T + \beta_S Z_i + \gamma_Z FI_i + \gamma_B B_i$$

and the scaled logistic model Dunning (2006)

$$P(Y_i = 1) = \pi \frac{e^{\tilde{\mu}_T + \beta_S Z_i + \gamma_Z FI_i + \gamma_B B_i}}{1 + e^{\tilde{\mu}_T + \beta_S Z_i + \gamma_Z FI_i + \gamma_B B_i}}.$$

Note that this model is consistent with the model used to generate the data (Eq. (2)), with a slightly different parametrization. The power to meet Prentice criterion 4 (PC4) was measured as the proportion of simulated trials with $p\text{-value}(s) = 2\Phi(-|\hat{\gamma}_Z|/\sqrt{\text{Var}(\hat{\gamma}_Z)}) < \alpha$ and $p\text{-value}(z) = 2\Phi(-|\hat{\beta}_S|/\sqrt{\text{Var}(\hat{\beta}_S)}) \geq \alpha$, $\alpha = 0.05$.

Furthermore, we applied the Principal surrogate approach on vaccine induced fold-increase ($S(1)_i = FI(1)_i$) where missing information is integrated out using the baseline surrogate measurement (B_i). The power of the WEM approach was measured as the proportion of simulated trials with significant Wald statistics for the $s(1)z$ coefficient of model (1) ($p\text{value}(s(1)z) < \alpha$, $\alpha = 0.05$). Appendix contains the R code used to apply the PS approach is provided.

Table 1 shows that the power of both PC4 and WEM decreases when the VE increases. This is due to the fact that there is less information (number of events) as the VE increases. Note that the power of the Prentice approach is higher than in Callegaro and Tibaldi (2019) because of the inclusion of the baseline surrogate as covariate. Simulation results suggest similar power for PC4 and WEM approaches.

The performance of the two approaches depends on the correlation between A and B . In fact, larger is the correlation, more informative is the covariate B . To assess the role of the correlation on the results, we replicated Table 1 with smaller correlation between A and B ($\text{Cor}(A, B) = 0.5$ in the placebo and in the vaccine group). Simulation results are shown in Table 2. We can see that when the correlation is smaller (i.e. when the covariate B is less informative) there is a greater loss of power for high VE for both approaches, especially for the PS approach. These results are aligned with the simulation results of Callegaro and Tibaldi (2019), showing a similar loss of power of Prentice method without covariates.

Table 1: Simulation results of data generated using scaled logit Prentice model 3.

Δ	$\hat{VE}$	PC4 logistic	PC4 scaled logistic	WEM
0.33	0.41	0.94	0.96	0.92
0.75	0.75	0.93	0.96	0.92
1.00	0.87	0.89	0.95	0.90
1.50	0.96	0.80	0.88	0.73

Power ($\alpha=0.05$) to assess Prentice criterion 4 (PC4) using logistic and scaled logistic model and power to assess the Wide Effect Modification (WEM) of a Principal Surrogate using the logistic model (p-value of the interaction $s(1)z$).

Table 2: Simulation results of data generated using scaled logit Prentice model 3 with smaller correlation between A and B ($\text{cor}(A,B)=0.5$ in placebo and in the vaccine group).

Δ	$\hat{VE}$	PC4 logistic	PC4 scaled logistic	WEM
0.33	0.41	0.94	0.96	0.98
0.75	0.75	0.89	0.96	0.97
1.00	0.86	0.79	0.95	0.92
1.50	0.96	0.69	0.95	0.62

Power ($\alpha=0.05$) to assess Prentice criterion 4 (PC4) using logistic and scaled logistic model and power to assess the Wide Effect Modification (WEM) of a Principal Surrogate using the logistic model (p-value of the interaction $s(1)z$).

Simulations with constant biomarker under placebo

In the previous simulations the Fold-Increase was not constant in placebo (it was normally distributed). To evaluate the performance of the Prentice and PS approach in case of constant biomarker under placebo, which mimics vaccine trials in a naive population, we simulated data using the model described above. However, in the inferential models, we replaced FI by FI^* which is constant in Placebo. FI^* is defined as

$$FI^* = \begin{cases} FI & \text{if } FI > c \\ c & \text{if } FI \leq c \end{cases}$$

where c is the 99% quantile of the distribution of FI in Placebo.

Table 3 shows some loss of power of the PS approach when the VE increases. Even if the use of the Prentice framework is not justified in this context, Table 3 shows the results of the Prentice criteria 4 (PC4 logistic model). Results from PC4 scaled logistic are not shown because the model is not converging. We observe a dramatic loss of power of Prentice criterion 4 when the VE is high.

Note that Table 3 shows simulation results where the inferential models do not agree with the data generating mechanism, so it represents a situation of model miss-specification.

Table 3: Simulation results with constant biomarker (inferential models do not agree with the data generating mechanism).

Δ	$\hat{VE}$	PC4 logistic	WEM
0.33	0.41	0.80	0.67
0.75	0.75	0.45	0.85
1.00	0.87	0.49	0.84
1.50	0.96	0.37	0.69

Power ($\alpha=0.05$) to assess Prentice criterion 4 (PC4) using logistic model and power to assess the Wide Effect Modification (WEM) of a Principal Surrogate using the logistic model (p-value of the interaction $s(1)z$).

Table 4: Simulation results with constant biomarker (inferential models agree with the data generating mechanism).

Δ	$\hat{VE}$	PC4 logistic	WEM
0.33	0.29	0.96	0.73
0.75	0.62	0.94	0.85
1.00	0.78	0.96	0.84
1.47	0.94	0.79	0.80

Power ($\alpha=0.05$) to assess Prentice criterion 4 (PC4) using logistic model and power to assess the Wide Effect Modification (WEM) of a Principal Surrogate using the logistic model (p-value of the interaction $s(1)z$).

To disentangle the problem of model miss-specification from the constant biomarker problem, we generate additional constant biomarker data using a model consistent with the “inferential” model used to fit the data. We simulated data using the following Dunning regression model:

$$P(Y_i = 1 | \pi, FI_i^*, B_i) = \pi \frac{e^{\mu + \gamma FI_i^* + \gamma_B B_i}}{1 + e^{\mu + \gamma FI_i^* + \gamma_B B_i}}.$$

Here, $\pi = 0.1$ and the other parameters are chosen to mimic Table 1 data: $\Delta = 0.33, 0.75, 1, 1.47$, $\mu = (8.66, 9.45, 9.82, 9.41)$, $\gamma = (-5.39, -5.15, -4.8, -4.45)$ and $\gamma_B = (-2.31, -2.63, -2.79, -2.66)$.

Table 4 shows that the loss of power of Prentice approach shown in Table 3 was mainly due to model miss-specification. In fact, Table 4 shows a relatively higher power of PC4 logistics than Table 3 when VE is large.

Simulations with low/moderate VE

For comparison, we considered simulations with low VE. We simulated data as described above with $\mu_1 = E(A|Z = 1) = 3, 3.075, 3.15, 3.23$, corresponding to estimated VE about 0%, 10%, 20% and 30%, respectively. Note that Prentice criteria 1 will not be met in this situation. For simplicity, we focused only on Prentice criterion 4. Table 5 shows that both approaches (PC4 and WEM) are powerful in the case of low/moderate VE. Prentice criterion 4 seems to be slightly more powerful than PS.

Simulations using random intercept logistic (correlated potential outcomes)

Finally, we generated data in a different way more aligned with the causal inference setting (potential outcomes). We generated correlated post-vaccination values ($A(0), A(1)$) using a bivariate normal distribution

$$\begin{pmatrix} A(0) \\ A(1) \end{pmatrix} \sim N \left[\begin{pmatrix} 3 \\ 3 + \Delta \end{pmatrix}, \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix} \right]$$

Table 5: Simulation results with small/moderate VE.

Δ	$\hat{VE}$	PC4 logistic	PC4 scaled logistic	WEM
0.000	−0.01	0.95	0.96	0.92
0.075	0.09	0.95	0.96	0.91
0.150	0.20	0.95	0.97	0.91
0.250	0.31	0.95	0.97	0.93

Power ($\alpha=0.05$) to assess Prentice criterion 4 (PC4) using logistic and scaled logistic model and power to assess the Wide Effect Modification (WEM) of a Principal Surrogate using the logistic model (p-value of the interaction $s(1)z$).

with $\Delta = (0.33, 0.75, 1.1, 1.6)$. The mean and the variance of the baseline are the same as the post-dose surrogate in Placebo. The correlation between baseline and post is 90% in Placebo and 50% in Vaccinated, respectively. We generated the correlated clinical outcomes using a logistic model with individual random intercept (b_i)

$$\text{logit}(P(Y(z)_i = 1|A(z)_i, b_i)) = \mu + A(z)_i\gamma + b_i.$$

The variables $Y(0), Y(1)$ are conditionally independent given b but unconditionally (averaged over b) correlated. The extent of correlation depends on the variance of the random effect ($\text{var}(b)$).

We generated bridge distributed random intercept (using R package `bridgedist` Swihart (2016)) such that the resultant marginal distribution follows a logistic regression model Wang and Louis (2003). In fact, the marginal logistic regression model is $\text{logit}(P(Y(z) = 1|A(z))) = \mu/c + A(z)\gamma/c$ for $z = 0, 1$ with $c = \sqrt{1 + 3\text{var}(b)/\pi^2}$. We simulated data with the following parameters: $\text{var}(b) = 10$ (scale=0.5), $\mu = 3.6$ and $\gamma = -3.8$. In this way, $p_0 = P(Y = 1|Z = 0) = 0.05$ and the estimated VE is about 0.45, 0.75, 0.85 and 0.95.

Table 6 shows that Prentice criterion 4 is more powerful than WEM.

The Prentice framework is more powerful than PS for different reasons: (i) PS tests for an interaction, which is less powerful than a test for the main effect; (ii) the covariate S (observed surrogate in vaccinated and placebo) has greater range in the Prentice model 4 than the covariate $S(1)$ in the PS model. It is easier to estimate a slope for a covariate with a bigger range. Figure 1 illustrate these differences.

Case study: analysis of a simulated data-set with large VE

In this section we analyze one simulated dataset from the scenario with largest VE of Table 1. The sample size is $n=5,000$, with 1:1 randomization. The number of events observed in the two groups are 3 and 90, with an estimated VE of 96% (95%CI, 89–98%). Figure 2 shows that the vaccine and placebo groups had similar log10 titer distributions at baseline while there is a small overlap in distributions post vaccination. Antibody responses clearly increased from baseline to post-dose in vaccine recipients but not in placebo recipients.

Figure 3 shows the Spearman correlation between baseline and post (left panel) and between baseline and Fold-Increase (right panel).

Prentice framework

First we examine the interaction between surrogate and the treatment. Table 7 shows that there is no interaction (p-value=0.49).

Secondly, we assess the four Prentice criteria. Table 8 shows that all criteria are met. In particular, the last 4 rows shows the results related to criterion 4. We can see that the effect of the surrogate is significant (p-value(s)=0.019), while the treatment effect is not significant, but is close to 5% (p-value(z) = 0.078).

Table 6: Simulation results with random intercept (bridge) logistic regression.

Δ	VE	PC4 logistic	WEM
0.33	0.46	0.95	0.85
0.75	0.75	0.94	0.84
1.10	0.87	0.91	0.77
1.60	0.95	0.89	0.74

Power ($\alpha=0.05$) to assess Prentice criterion 4 (PC4) using logistic model and power to assess the Wide Effect Modification (WEM) of a Principal Surrogate using the logistic model (p-value of the interaction $s(1)z$).

Slightly better results are obtained if Dunning model is used (see Table 9).

In summary, there is suggestive though not strong evidence that the Fold-Increase is a Statistical Surrogate.

Principal surrogate framework

Table 10 shows the results from R package pseval with 50 bootstrap (R codes are provided in the Appendix). We can see that the interaction between the treatment group and FI(1) (test for wide effect modification) is borderline (p-value=0.053).

Figure 4 shows the estimated VE curve for Fold-Increase. The estimated VE curve is an increasing function of FI(1), however we can see large variability for small values of FI(1) and negative VEs for vaccine recipients with no rise.

In summary, there is suggestive though not strong evidence that the Fold-Increase is a Principal Surrogate.

Figure 1: Simulated trial with $n = 5,000$ per arm under the scenario with $\Delta = 1.6$. The true probability of infection is graphed as a function of observed A s and $A(1)$ s respectively. The top panel is the data used for the Prentice criteria while the bottom panel is used to test WEM. Red denotes the placebo group while blue denotes the vaccine group. The events are shown at the top and the non-events at the bottom of the graph.

Figure 2: Distribution of the surrogate endpoint: baseline, post and fold-increase (post-baseline).

Figure 3: Correlation between baseline and post (left panel) and between baseline and fold-increase (right panel).

Table 7: Logistic model with interaction between treatment group and surrogate.

	Estimate	Std. error	z Value	p-Value
(Intercept)	0.833	0.717	1.161	0.245
Z	-0.576	1.696	-0.339	0.734
FI	-1.060	0.555	-1.909	0.056
B	-1.434	0.257	-5.583	0.000
group:FI	-0.966	1.401	-0.690	0.490

Table 8: Prentice criteria: logistic and linear models.

Criterion	Variable	Estimate	Std error	z Value	p-Value
1	(Intercept)	0.433	0.698	0.620	0.535
1	Z	-3.448	0.588	-5.865	0.000
1	B	-1.293	0.249	-5.196	0.000
2	(Intercept)	0.956	0.031	30.734	0.000
2	Z	1.472	0.009	163.653	0.000
2	B	-0.317	0.010	-31.166	0.000
3	(Intercept)	1.092	0.702	1.556	0.120
3	FI	-1.983	0.285	-6.970	0.000
3	B	-1.545	0.250	-6.176	0.000
4	(Intercept)	0.825	0.717	1.150	0.250
4	Z	-1.644	0.933	-1.763	0.078
4	FI	-1.205	0.514	-2.345	0.019
4	B	-1.432	0.257	-5.574	0.000

Table 9: Prentice criterion 4 using Dunning model.

Variable	Estimate	Std error	z Value	p-Value
(Intercept)	8.528	3.442	2.478	0.013
FI	-2.662	1.131	-2.353	0.019
Z	-0.620	1.305	-0.475	0.635
B	-2.978	0.963	-3.092	0.002
logit(pi)	-2.386	0.370	-6.440	0.000

Table 10: Principal surrogate Evaluation.

	Estimate	Boot se	Lower CL 2.5%	Upper CL 97.5%	p-Value
(Intercept)	−7.81	1.146	−10.35	−6.147	9.13^{-12}
FI(1)	2.64	0.573	1.79	3.891	4.18^{-6}
Z	2.90	2.846	−3.66	6.593	3.08^{-1}
FI(1):Z	−3.98	2.053	−8.11	−0.157	5.28^{-2}

Figure 4: Estimated vaccine efficacy curve across levels of vaccine-induced fold-increase from baseline to post-vaccination, with 95% confidence intervals (dashed lines).

Discussion

Although not common, vaccines with very high efficacy (95% or above) are documented in the literature (Black et al. 2000; Lin et al. 2001; Mitra et al. 2016; Phua et al. 2012; Prymula et al. 2014; Wei et al. 2016). These trials raise the problem of assessing CoPs in the context where small number of cases/infections in vaccinated groups are available.

Callegaro and Tibaldi (2019) showed that the validation of a surrogate endpoint using the Prentice criteria and meta-analytic frameworks (by randomized subgroups in single trial setting) can be problematic in case of high VE. In this paper, we evaluate the performance of the causal framework, specifically the Principal Surrogate (PS) approach (Follmann 2006; Gilbert, Qin, and Self 2008) in case of high VE.

First, we replicated the simulation study of Callegaro and Tibaldi (2019) where the clinical outcome was simulated using Prentice model 3 (assuming full mediation) and using the Dunning model (Dunning 2006). These simulation results show that i) adjustments for important covariates (such as baseline surrogate) considerably improves the power of the Prentice approach (even if the model is miss-specified) in case of high VE. Furthermore, these simulation results show similar power of Prentice and PS frameworks. The power of both approaches decreases when VE grows.

Second, we slightly changed the Callegaro and Tibaldi scenario to consider the case of constant biomarker under placebo and the case of small/moderate VE. Simulation results show that i) PS is more powerful than Prentice in case of constant biomarker when the inferential model is miss-specified, otherwise Prentice is more powerful; ii) Prentice criteria 4 and PS frameworks are powerful when the VE is small (see Table 3). However, in this case Prentice criteria 1 is not met, so the two approaches give different conclusions.

Finally, we simulated correlated potential outcome data using a bivariate (random intercept) logistic regression. In this case the Prentice framework is more powerful than the PS approach. This can be due to the following reasons: (i) Prentice model 4 corresponds to the model used to generate the data and so there is

no lack of fit in the Prentice framework; (ii) PS tests for an interaction, which is less powerful than a test for the main effect; (iii) the covariate S (observed surrogate in vaccinated and placebo) has greater range in the Prentice model 4 than the covariate $S(1)$ in the PS model. It is easier to estimate a slope for a covariate with a bigger range (see Figure 1); (iv) Principal stratification has to impute $S(1)$ for placebo participants which increases the variability of estimates relative to knowing $S(1)$. In contrast S is known in all for the Prentice criterion.

For computational reasons, we performed relatively small number of iterations (1,000). Larger number of iterations can be considered in the future using multiple processors. What is computationally intensive is the bootstrap of the PS approach. As an example, 200 re-sampling on the case study required 14 min. To mitigate the computational load, it may be useful in the future to derive asymptotic formulas approximating the bootstrap approach.

It is important to highlight that the power comparison between the two approaches should be interpreted with care. In fact, the two approaches measure two different things: Prentice framework evaluates if the surrogate is a “statistical surrogate” while the PS evaluates if the surrogate is a “principal surrogate” (see Gilbert et al. (2015) for more details).

For illustration, we analyzed one data-set simulated with full mediation (Dunning model 3) and with high VE ($\hat{VE} = 96\%$). Results showed suggestive thought not strong evidence that the FI is a Statistical Surrogate (Prentice criteria) or a PS. These results are due to the lack of power of these approaches in case of high VE. An interesting topic for future research is the implementation of the two approaches in a Bayesian framework with weakly informative priors (WIP). In fact, Callegaro and Tibaldi (2019) showed that WIP can considerably increase the power of the meta-analytical approach in case of high VE.

In conclusion, we evaluated by simulation the impact of high VE on the PS approach. Similarly to the Prentice framework, we showed that the power decreases when the VE grows. It follows that it can be challenging to validate a principal surrogate (and a statistical surrogate) when rare infections are observed in the vaccinated groups.

Research funding: GlaxoSmithKline Biologicals SA was the funding source for all costs associated with the development and the publishing of the present manuscript.

Conflict of Interest: AC and FT are employees of the GlaxoSmithKline group of companies. AC and FT own stock options in the GlaxoSmithKline group of companies. Prof DF declares that he has no conflict of interest.

Appendix: R Code used to apply the PS approach

```
library("pseval")
  binary.
  est <- psdesign
  (data, Z = group, Y = y, S = FI, BIP = B) +
  integrate_parametric
  (S.1 ~ BIP) +
  risk_binary
  (model = Y ~ S.1 * Z, D = 50, risk = risk.
  logit) +
  ps_estimate
  (method = "BFGS")
  binary.
  boot <- binary.
  est + ps_bootstrap
  (n.boots = 200,
  progress.
```

```

bar = FALSE, start = binary.
est$estimates$par,
method = "BFGS")

```

References

- Alonso, A., W. Van der Elst, G. Molenberghs, M. Buyse, and T. Burzykowski. 2015. "On the Relationship between the Causal-Inference and Meta-Analytic Paradigms for the Validation of Surrogate Endpoints." *Biometrics* 71: 15–24.
- Black, S., H. Shinefield, B. Fireman, E. Lewis, P. Ray, J. R. Hansen, L. Elvin, K. M. Ensor, J. Hackell, G. Siber, F. Malinoski, D. Madore, I. Chang, R. Kohberger, W. Watson, R. Austrian, and K. Edwards. 2000. "Efficacy, Safety and Immunogenicity of Heptavalent Pneumococcal Conjugate Vaccine in Children." *The Pediatric Infectious Disease Journal* 19 (3): 187–95.
- Burzykowski, T., G. Molenberghs, and M. Buyse. 2005. *The Evaluation of Surrogate Endpoints*. New York: Springer.
- Buyse, M., G. Molenberghs, T. Burzykowski, D. Renard, and H. Geys. 2000. "The Validation of Surrogate Endpoints in Meta-Analyses of Randomized Experiments." *Biostatistics* 1: 49–67.
- Callegaro, A., and F. Tibaldi. 2019. "Assessing Correlates of Protection in Vaccine Trials: Statistical Solutions in the Context of High Vaccine Efficacy." *BMC Medical Research Methodology* 19: 47.
- Chan, I. S. F., S. Li, H. Matthews, C. Chan, R. Vessey, J. Sadoff, and J. Heyse. 2002. "Use of Statistical Models for Evaluating Antibody Response as a Correlate of Protection against Varicella." *Statistics in Medicine* 21 (22): 3411–30.
- Daniels, M. J., and M. D. Hughes. 1997. "Meta-analysis for the Evaluation of Potential Surrogate Markers." *Statistics in Medicine* 16: 1965–82.
- Dunning, A. J. 2006. "A Model for Immunological Correlates of Protection." *Statistics in Medicine* 25 (9): 1485–97.
- Follmann, D. 2006. "Augmented Designs to Assess Immune Response in Vaccine Trials." *Biometrics* 62 (4): 1161–9.
- Frangakis, C. E., and D. B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58 (1): 21–9.
- Gabriel, E., and D. Follmann. 2016. "Augmented Trial Designs for Evaluation of Principal Surrogates." *Biostatistics* 17 (3): 453–67.
- Gabriel, E., and P. Gilbert. 2014. "Evaluating Principal Surrogate Endpoints with Time-To-Event Data Accounting for Time-Varying Treatment Efficacy." *Biostatistics* 15 (2): 251–65.
- Gail, M. H., R. Pfeiffer, H. C. V. Houwelingen, and R. Carroll. 2000. "On Meta-Analytic Assessment of Surrogate Outcomes." *Biostatistics* 1: 231–46.
- Gilbert, P. B., E. E. Gabriel, Y. Huang, and I. S. F. Chan. 2015. "Surrogate Endpoint Evaluation: Principal Stratification Criteria and the Prentice Definition." *Journal of Causal Inference* 3: 157–75.
- Gilbert, P. B., L. Qin, and S. G. Self. 2008. "Evaluating a Surrogate Endpoint at Three Levels, with Application to Vaccine Development." *Statistics in Medicine* 27 (23): 4758–78.
- Huang, Y., and P. Gilbert. 2011. "Comparing Biomarkers as Principal Surrogate Endpoints." *Biometrics* 67 (4): 1442–51.
- Huang, Y., P. B. Gilbert, and J. Wolfson. 2013. "Design and Estimation for Evaluating Principal Surrogate Markers in Vaccine Trials." *Biometrics* 69: 301–9.
- Lin, F. Y. C., V. A. Ho, H. B. Khiem, D. D. Trach, P. V. Bay, T. C. Thanh, Z. Kossaczka, D. A. Bryla, J. Shiloach, J. B. Robbins, R. Schneerson, S. C. Szu, M. N. Lan, S. Hunt, L. Trinh, and J. B. Kaufman. 2001. "The Efficacy of a Salmonella Typhi Vi Conjugate Vaccine in Two-To-Five-Year-Old Children." *New England Journal of Medicine* 344 (17): 1263–9.
- Miao, C., X. Li, P. Gilbert, and I. Chan. 2013. "A Multiple Imputation Approach for Surrogate Marker Evaluation in the Principal Stratification Causal Inference Framework." In *Risk Assessment and Evaluation of Predictions*. New York: Springer.
- Mitra, M., N. Shah, A. Ghosh, S. Chatterjee, I. Kaur, N. Bhattacharya, and S. Basu. 2016. "Efficacy and Safety of Vi-Tetanus Toxoid Conjugated Typhoid Vaccine (Pedatyph) in Indian Children: School Based Cluster Randomized Study." *Human Vaccines & Immunotherapeutics* 12 (4): 939–45.
- Phua, K. B., F. S. Lim, Y. L. Lau, E. A. S. Nelson, L. M. Huang, S. H. Quak, B. W. Lee, L. J. van Doorn, Y. L. Teoh, H. Tang, P. V. Suryakiran, I. V. Smolenov, H. L. Bock, and H. H. Han. 2012. "Rotavirus Vaccine RIX4414 Efficacy Sustained during the Third Year of Life: A Randomized Clinical Trial in an Asian Population." *Vaccine* 30 (30): 4552–7.
- Prentice, R. L. 1989. "Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria." *Statistics in Medicine* 8 (4): 431–40.
- Prymula, R., M. R. Bergsaker, S. Esposito, L. Gothefors, S. Man, N. Snegova, M. Štefkovičová, V. Usonis, J. Wysocki, M. Douha, V. Vassilev, O. Nicholson, B. L. Innis, and P. Willems. 2014. "Protection against Varicella with Two Doses of Combined Measles-Mumps-Rubella-Varicella Vaccine versus One Dose of Monovalent Varicella Vaccine: A Multicentre, Observer-Blind, Randomised, Controlled Trial." *The Lancet* 383 (9925): 1313–24.
- Sachs, M. C., and E. E. Gabriel. 2016. "An Introduction to Principal Surrogate Evaluation with the Pseval Package." *The R Journal* 8: 277–92.
- Swihart, B. 2016. *Bridgedist: An Implementation of the Bridge Distribution with Logit-Link as in Wang and Louis (2003)*. Also available at <https://CRAN.R-project.org/package=bridgedist>, r package version 0.1.0.

- Wang, Z., and T. A. Louis. 2003. "Matching Conditional and Marginal Shapes in Binary Random Intercept Models Using a Bridge Distribution Function." *Biometrika* 90: 765–75.
- Wei, M., F. Meng, S. Wang, J. Li, Y. Zhang, Q. Mao, Y. Hu, P. Liu, N. Shi, H. Tao, K. Chu, Y. Wang, Z. Liang, X. Li, and F. Zhu. 2016. "Two-Year Efficacy, Immunogenicity, and Safety of Vigoo Enterovirus 71 Vaccine in Healthy Chinese Children: A Randomised Open-Label Study." *The Journal of Infectious Diseases* 215: 56–63.
- Wolfson, J., and P. Gilbert. 2010. "Statistical Identifiability and the Surrogate Endpoint Problem, with Application to Vaccine Trials." *Biometrics* 66 (4): 1153–1161.