Rank-Based Identification of High-dimensional Surrogate Markers: Application to Vaccinology
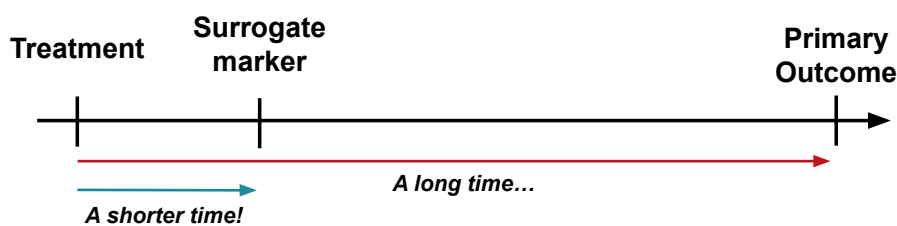
Arthur Hughes[1], Layla Parast[2], Rodolphe Thiébaut[1], Boris Hejblum[1] and PREVAC Study Team

[1]University of Bordeaux, BPH INSERM 1219, SISTM Team
[2]Department of Statistics and Data Science, University of Texas at Austin, USA

Treatment — Surrogate marker — Primary Outcome

A long time…

A shorter time!

Background

- **Surrogates** are reliable markers which can be **observed sooner or easier** than a primary outcome

- **Gene expression** is a **dynamic process** where information in DNA is transformed into proteins

- Changes in gene expression can be **observed soon** after vaccination

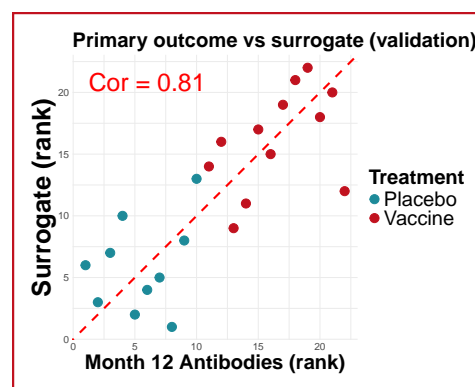- Existing methods for identifying surrogates do not work with $p > n$

Idea of Method

Non-parametric, univariate test to…

1. **Screen 1000s of genes** one-by-one to find surrogate candidates

2. **Combine** candidates into single strong predictor and **validate** on independent data

Application

Goal : identify genes who explain the antibody response to Ebola vaccination

- **12 candidate genes** from 25,165 found in screening stage!

- **Combined expression** of these genes **validated** on independent data!

Primary outcome vs surrogate (validation)
Cor = 0.81
Surrogate (rank) vs Month 12 Antibodies (rank)
Treatment: Placebo, Vaccine

The Maths

Screening

For each gene S_j, $\forall j = 1, ..., P$

- $A \in \{0, 1\}$ - treatment

- Y^a, S_j^a - primary endpoint and surrogate candidate when $A = a$

- $U_Y = \mathbb{P}(Y^1 > Y^0) + \frac{1}{2}\mathbb{P}(Y^1 = Y^0)$
  - **treatment effect on** Y

- $U_{S_j} = \mathbb{P}(S_j^1 > S_j^0) + \frac{1}{2}\mathbb{P}(S_j^1 = S_j^0)$
  - **treatment effect on** S_j

- $\delta_j = U_Y - U_{S_j}$

- $H_0 : \delta_j \geq \epsilon$ vs $H_1 : \delta_j < \epsilon$
  - i.e. is gene S_j **sufficiently good** at detecting a treatment effect?

- p-value is $p_j = P(Z < \widehat{\delta}_j)$ where $Z \sim N(\epsilon, \widehat{\sigma}_{\delta_j})$, $\widehat{\sigma}_{\delta_j}$ is estimated s.d.

- p-values adjusted for number of tests P

Validation

For all genes S_j with adjusted $p_j < 0.05$

- form new surrogate as a weighted sum
  - $\gamma = \sum_j \frac{\bar{S}_j}{\widehat{\delta}_j}$

- validate γ with screening test

- i.e. test $H_0 : U_Y - U_\gamma \geq \epsilon$

1000s of mRNA counts from vaccinated individuals

Screening algorithm to find small number of candidate genes

Validation Can the new surrogate replace the primary outcome?

VS