Rank-Based Identification of High-Dimensional Surrogate Markers : Application to Vaccinology

SFdS JdS 2025 - Marseille

Arthur Hughes[1], Layla Parast[3], Rodolphe Thiébaut[12], Boris Hejblum[1]

[1]University of Bordeaux, BPH INSERM U1219, INRIA SISTM
[2]CHU de Bordeaux, Service d'information Médicale,
[3]University of Texas at Austin, USA

Digital Public Health Graduate Program

université de BORDEAUX

VRI VACCINE RESEARCH INSTITUTE

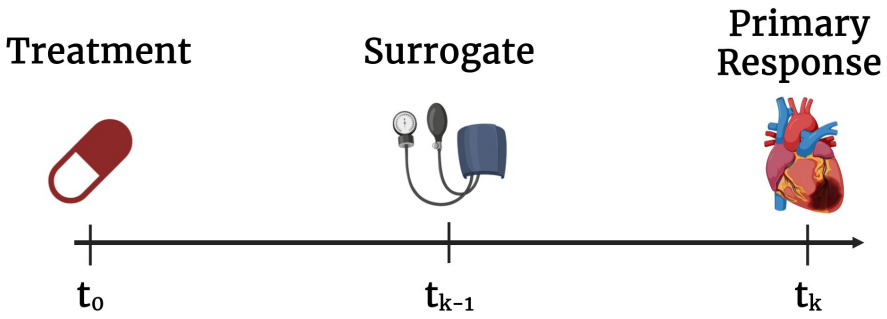
Inserm La science pour la santé From science to health

Inria

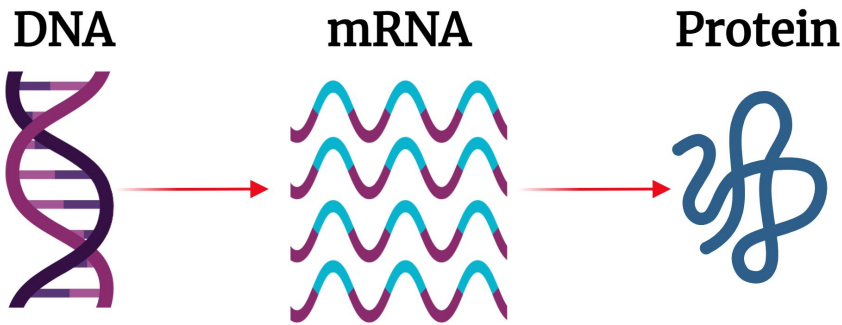
Background

What is a surrogate marker?

- **Intermediate endpoint**
- Treatment effect on surrogate **predicts treatment effect on primary outcome**



Transcriptomics



Could gene expression markers serve as surrogates?

Limitations of Current Methodology

Existing methods require...

- **Restrictive assumptions**
- **Large sample sizes**
- **Low-dimensional setting**

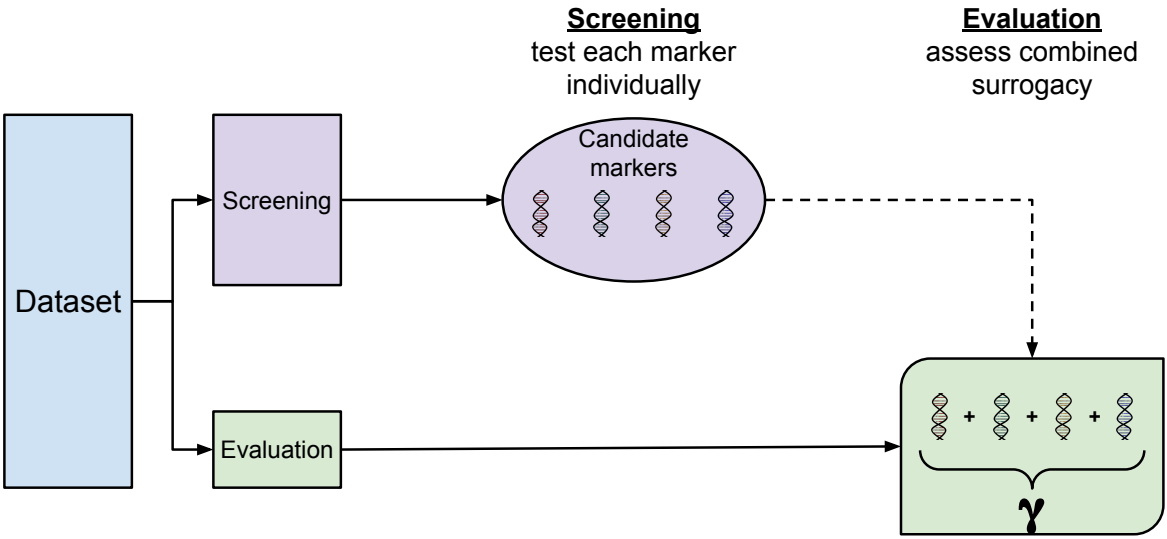
Limitations of Current Methodology

Existing methods require...

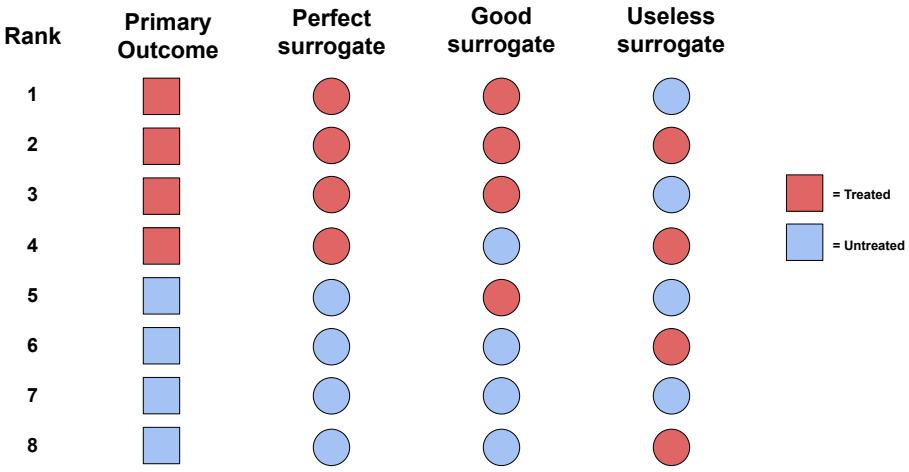
- **Restrictive assumptions**
- **Large sample sizes**
- **Low-dimensional setting**

New approach to **evaluate high-dimensional surrogate markers** in small-sample setting

Materials and Methods


Screening
test each marker
individually
Evaluation
assess combined
surrogacy
Dataset
Screening
Candidate
markers
Evaluation
γ

Intuition of the rank-based test



Notation

- n - sample size
- $A \in \{0, 1\}$ - binary treatment indicator
- Y - continuous response
- $\boldsymbol{S} = (S_1, ..., S_p)$ - candidate surrogates
- Y^a - response had treatment been a
- S^a - surrogate candidates had treatment been a

A non-parametric test for surrogacy of a single marker

- $U_Y = \mathbb{P}(Y^1 > Y^0) + \frac{1}{2}\mathbb{P}(Y^1 = Y^0)$
  - $0.5 < U_Y \leq 1 \implies$ positive treatment effect
- $U_{S_j} = \mathbb{P}(S_j^1 > S_j^0) + \frac{1}{2}\mathbb{P}(S_j^1 = S_j^0)$
- $\delta_j = U_Y - U_{S_j}$
  - i.e. $\delta_j \approx 0 \implies S_j$ approximates treatment effect on Y
- **Non-Inferiority Test** $H_0 : \delta_j \geq \epsilon$ vs $H_1 : \delta_j < \epsilon$

Estimation with rank-sum statistics

- Define $G(A,B) = \begin{cases} 1, & \text{if} \quad A > B \\ \frac{1}{2}, & \text{if} \quad A = B \\ 0, & \text{if} \quad B < A \end{cases}$
- $\widehat{U_Y} = (n_1 n_0)^{-1} \sum_{i=1}^{n_1} \sum_{k=1}^{n_0} G(Y_{i1}, Y_{k0})$
- $\widehat{U_{S_j}} = (n_1 n_0)^{-1} \sum_{i=1}^{n_1} \sum_{k=1}^{n_0} G(S_{ji1}, S_{jk0})$
- $\widehat{\delta_j} = \widehat{U_Y} - \widehat{U_{S_j}}$

Screening stage

- One-sided $(1-\alpha)\%$ C.I. estimated as $[-1, \widehat{\delta}_j + \Phi^{-1}(1-\alpha)\widehat{\sigma}_{\delta_j}]$
- p-value is $p_j = P(Z < \widehat{\delta}_j)$ where $Z \sim N(\epsilon, \widehat{\sigma}_{\delta_j})$
- Test every candidate $S_1, ..., S_p$ and correct p-values for test multiplicity
- Define candidate surrogates $\mathcal{S} = \{j : p_{j,\text{adj}} \leq \alpha\}$

Evaluation Step

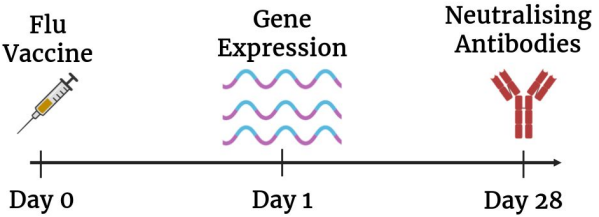
Evaluate combined surrogacy of candidates

- $\widehat{\gamma_{\mathcal{S}}} := \sum_{j \in \mathcal{S}} \widehat{\delta_j}^{-1} \bar{S}_j$
  - $\bar{S}_j$ is S_j standardised
  - Weighted by $\widehat{\delta_j}^{-1} \implies$ stronger surrogates contribute more
- Re-apply rank-test to evaluate $\widehat{\gamma_{\mathcal{S}}}$

Application

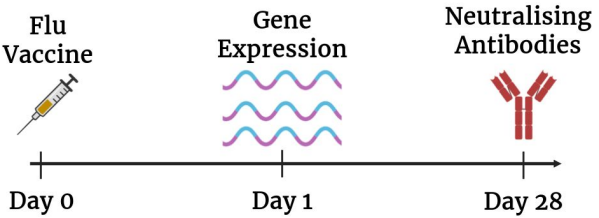
Data description

- **Open data** from Human Immune Project Consortium
- SDY1276 - GE/Antibody response for *Trivalent Inactivated Influenza* ($n = 103$)



Data description

- **Open data** from Human Immune Project Consortium
- SDY1276 - GE/Antibody response for *Trivalent Inactivated Influenza* ($n = 103$)



Can the treatment effect on GE at day 1 predict the treatment effect on day 28 on the antibodies?

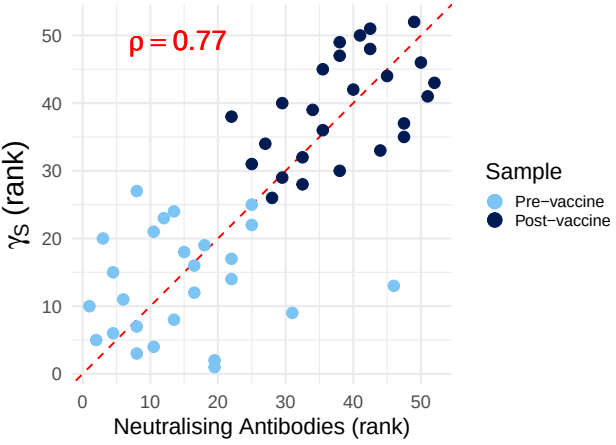
Screening Results

222 significant genes after multiple testing correction

Gene	Gene set	δ (95% C.I.)	p_{adj}
CNDP2		-0.026 (-0.056, 0.004)	1.6e-43
IFI44L	M8.3 (Type 1 Interferon)	-0.026 (-0.056, 0.004)	1.6e-43
IFITM3	M15.127 (Interferon)	-0.026 (-0.056, 0.004)	1.6e-43
NPC2		-0.026 (-0.056, 0.004)	1.6e-43
PSME1		-0.026 (-0.056, 0.004)	1.6e-43
SERPING1	M15.127 (Interferon)	-0.026 (-0.056, 0.004)	1.6e-43
VAMP5		-0.026 (-0.056, 0.004)	1.6e-43
EPB41L3	M12.2 (Monocytes)	-0.013 (-0.05, 0.024)	1.1e-30
IFI6	M8.3 (Type 1 Interferon)	-0.013 (-0.05, 0.024)	1.1e-30
IRF7	M10.1 (Interferon)	-0.013 (-0.05, 0.024)	1.1e-30
MX1	M8.3 (Type 1 Interferon)	-0.013 (-0.05, 0.024)	1.1e-30
MYOF	M16.6 (Monocytes), M16.15 (Cell death)	-0.013 (-0.05, 0.024)	1.1e-30
OAS3	M8.3 (Type 1 Interferon)	-0.013 (-0.05, 0.024)	1.1e-30
PSMB9	M13.17 (Interferon),M15.64 (Interferon)	-0.013 (-0.05, 0.024)	1.1e-30
RHBDF2	M15.37 (Inflammation), M15.64 (Interferon)	-0.013 (-0.05, 0.024)	1.1e-30

Table: Top 15 genes by adjusted p-value from screening stage on 75% of the data.

Evaluation results



- $\delta_{\gamma_S} = -0.0385(-0.102, 0.0248)$
- $p = 0.00311$
- $\implies \gamma_S$ a suitable surrogate for the day 28 treatment effect of TIV on neutralising antibodies

Discussion

Conclusions

- New method to **identify high-dimensional surrogate markers** of continuous responses
- Application to influenza vaccination
  - **222-gene signature** of mainly interferon genes **predicts vaccine effect** on antibodies
- Perspectives
  - Generalisability : other years of TIV, other vaccines?
  - Extension to other data types (survival, binary outcome) and complex designs

arXiv > stat > arXiv:2502.03030

Statistics > Methodology

[Submitted on 5 Feb 2025]

Rank-Based Identification of High-dimensional Surrogate Markers: Application to Vaccinology

Arthur Hughes, Layla Parast, Rodolphe Thiébaut, Boris P. Hejblum

Thank you for listening

Appendix

How to choose the threshold ϵ?

- ϵ depends on n, treatment effect on Y, desired power and significance
- If desired power $100 \times (1-\beta)\%$ to test a treatment effect on Y based on a test with S_j
- ϵ can be chosen adaptively as $\epsilon = \max\{0, \widehat{u}_Y - u^*_{\alpha,\beta}\}$
  - where $u^*_{\alpha,\beta} = \frac{1}{2} - \sqrt{\frac{n_0+n_1+1}{12 n_0 n_1}}[\Phi^{-1}(\beta) - \Phi^{-1}(1-\alpha)]$

Estimation - paired case

- **Data:** $\boldsymbol{Y_i} = (Y_i^1, Y_i^0)^T$ and surrogate candidate $\boldsymbol{S_{ij}} = (S_{ij}^1, S_{ij}^0)^T$.
- $\widehat{U}_Y = n^{-1}\sum_{i=1}^{n} G(Y_i^1, Y_i^0)$
- $\widehat{U}_{S_j} = n^{-1}\sum_{i=1}^{n} G(S_{ij}^1, S_{ij}^0)$

Two one-sided test procedure

We want to test $\delta \in [-\epsilon, \epsilon]$
Perform **Two one-sided tests** :

$$H_0^{(1)} : \delta \geq \epsilon, \quad \text{and} \quad H_0^{(2)} : \delta \leq -\epsilon.$$

resulting in two p-values $p^{(1)} = \Phi\left(\frac{\widehat{\delta}-\epsilon}{\widehat{\sigma}_\delta}\right)$, $p^{(2)} = 1 - \Phi\left(\frac{\widehat{\delta}+\epsilon}{\widehat{\sigma}_\delta}\right)$
Final p-value is $p = \max\{p^{(1)}, p^{(2)}\}$ and $(1-2\alpha) \times 100\%$ C.I. is

$$\left[\widehat{\delta} - \Phi^{-1}(1-\alpha)\,\widehat{\sigma}_\delta,\ \widehat{\delta} + \Phi^{-1}(1-\alpha)\,\widehat{\sigma}_\delta\right]$$

Simulation Setup

- $P = 500$ candidate surrogates
- Response : $Y_a \sim \mathcal{N}(\mu_{y_a}, \sigma_{y_a})$,
  - with $\mu_{y_1} = 3$, $\mu_{y_0} = 0$, and $\sigma_{y_a} = 1$
- Setting 1 : 100% invalid surrogates $S_{j,a} \sim \mathcal{N}(m_j, \sigma_j)$
  - $m_j \sim U(0.5, 2.5)$, $\sigma_j \sim U(0.5, 2)$
- Setting 2 : 10% valid surrogates $S_{j,a} = y_a + \mathcal{N}(0, \sigma_{\text{valid}})$
  - σ_{valid} controls surrogate strength

Setting 1 : no true surrogates

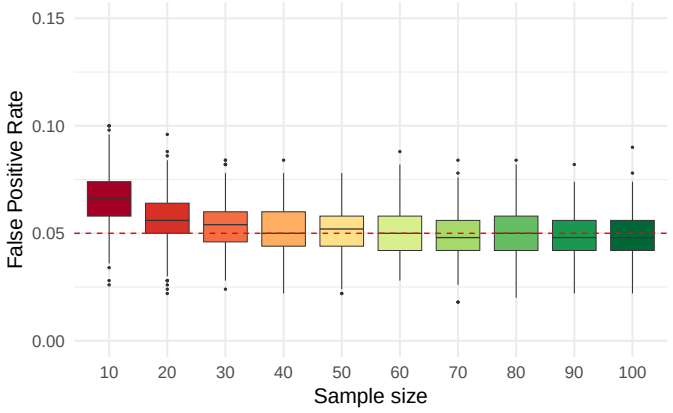


Figure: False positive rate across 500 data generations.

Setting 2 : 10% true surrogates

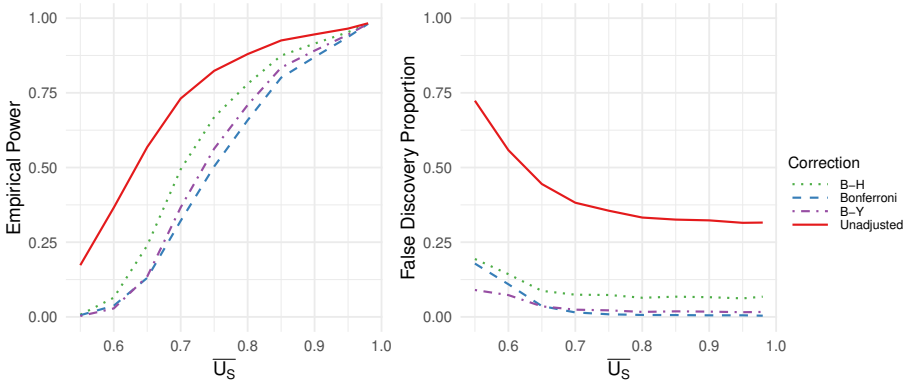


Figure: Power and FDP across 500 data generations with different multiple correction methods.