Development of non-parametric methods for the evaluation of surrogate markers in the non-continuous data setting

Master Internship

The main objective of a clinical trial is to evaluate the effect of a treatment on a clinical endpoint. However, direct measurement of the clinical endpoint may not be feasible due to cost, time, or ethical reasons. In such cases, one may instead rely on a surrogate marker, a variable which can be used to predict treatment effects on the clinical endpoint.

Evaluating candidate surrogate markers is a long-standing and active area of statistical research. Conventional approaches rely on parametric models whose distributional assumptions are often hard to verify, and inference can be misleading when they fail. This is a real concern for modern biomedical data, which are frequently complex, noisy, and high-dimensional.

To address this, some authors have recently proposed non-parametric methods using rank-based tests for surrogate evaluation (Parast et al., 2024). These methods do not make assumptions about the underlying form of the data, provide interpretable and scale-free surrogacy metrics, and are valid in small samples. This makes them particularly well-suited to high-dimensional applications, such as transcriptomic datasets, which are becoming more commonplace in modern biomedical studies (Hughes et al., 2025, 2026).

A limitation of these approaches is that they have only been developed for the evaluation of continuous variables, both for the clinical endpoint and the surrogate candidates. However, endpoints are often of different types, such as binary/ordinal data (e.g. protection status after vaccination), or time-to-event data (e.g. time to cancer progression). Extending the framework to these settings is not straightforward, since ties and censoring complicate rank-based inference, but it would substantially broaden the range of applications.

The objective of this internship will therefore be to extend the non-parametric, rank-based test procedure proposed in Parast et al. (2024) to binary, ordinal, and/or time-to-event data. The performance of these methods will be evaluated on both simulated and real datasets. The choice of dataset will be adapted based on the direction of the project: for example, Moncunill et al. (2022) describes an open dataset containing gene expression profiling of cases and controls infected with malaria after vaccination, which could be envisaged as an application in the case of a binary clinical endpoint.

Objectives

1. Perform a literature review of surrogate evaluation methods for non-continuous data types (binary, ordinal, and/or time-to-event)
2. Propose an extension of the rank-based test procedure described in Parast et al. (2024) to the setting where one or both of the clinical endpoint and surrogate candidate are of a non-continuous data type
3. Evaluate the proposed methodology on simulated data
4. Apply the proposed methodology to a real dataset (to be determined)

Requirements

- Enrolled in Master's (or equivalent) programme (in France: M2 or école d'ingénieurs)
- Background in biostatistics or similar
- Previous coding experience (preferably in R)
- Written and spoken proficiency in French and/or English
- Interest in biomedical research and scientific curiosity

Hosting laboratory and address

Statistics in Systems Biology and Translational Medicine (SISTM), INSERM U1219, Inria Bordeaux Population Health Research Center
Université de Bordeaux – ISPED
146, Rue Léo Saignat
33076 Bordeaux

Duration:

Internship of 4 to 6 months available starting in January 2027.

Compensation:

Set according to French government recommendations for internships (15% of the social security ceiling, i.e. around 600€/month).

Contact:

Send a detailed CV and a motivation letter to Boris Hejblum (boris.hejblum@u-bordeaux.fr) and Arthur Hughes (arthur.hughes@u-bordeaux.fr).

References

Arthur Hughes, Layla Parast, Rodolphe Thiébaut, and Boris P. Hejblum. Rise: Two-stage rank-based identification of high-dimensional surrogate markers applied to vaccinology. *Statistics in Medicine*, 44(20-22), September 2025. ISSN 1097-0258. doi: 10.1002/sim.70241.

Arthur Hughes, Rodolphe Thiébaut, Layla Parast, and Boris P. Hejblum. Meta-analysis of high-dimensional surrogate markers. *arXiv preprint*, 2026. doi: https://doi.org/10.48550/arXiv.2605.03819.

Gemma Moncunill, Jason Carnes, William Chad Young, Lindsay Carpp, Stephen De Rosa, Joseph J Campo, Augusto Nhabomba, Maxmillian Mpina, Chenjerai Jairoce, Greg Finak, Paige Haas, Carl Muriel, Phu Van, Héctor Sanz, Sheetij Dutta, Benjamin Mordmüller, Selidji T Agnandji, Núria Díez-Padrisa, Nana Aba Williams, John J Aponte, Clarissa Valim, Daniel E Neafsey, Claudia Daubenberger, M Juliana McElrath, Carlota Dobaño, Ken Stuart, and Raphael Gottardo. Transcriptional correlates of malaria in rts,s/as01-vaccinated african children: a matched case–control study. *eLife*, 11, January 2022. ISSN 2050-084X. doi: 10.7554/elife.70393.

Layla Parast, Tianxi Cai, and Lu Tian. A rank-based approach to evaluate a surrogate marker in a small sample setting. *Biometrics*, 80(1), January 2024. ISSN 1541-0420. doi: 10.1093/biomtc/ujad035.