The Journal of Open Source Software

NEEP: null empirically estimated p-values for high-throughput molecular survival analysis

Sean West[1], **Hesham Ali**[1], **and Dario Ghersi**[1]

1 School of Interdisciplinary Informatics, University of Nebraska at Omaha

DOI: 10.21105/joss.02044

Software

- Review
- Repository
- Archive

Editor: Mark A. Jensen

Reviewers:

- @wrathematics
- @SiminaB
- @rhagenson

Submitted: 19 December 2019
Published: 27 January 2020

Summary

When conducting survival analysis for molecular expression, a researcher has two main options: Cox-Proportional Hazards (CPH) or Kaplan-Meier (KM). CPH uses regression to calculate the survival-significance of each expression vector. It has two assumptions that are frequently violated in cancer research. First, CPH assumes that the censoring mechanism is not associated with patient survival. When using patients from cancer research, those who survive to the end of the study are likely to have a higher rate of survival than those who did not (**???**). Second, the core of CPH is the proportional hazards assumption. For molecular data, this assumes that the effect that a molecule has on survival is constant over time. This cannot be the case as the pathology of cancer changes across stages and time.

KM is non-parametric and uses a log-rank test to check if patients with low expression of the molecule of interest have altered survival rates from those with high expression. KM has the same two assumptions mentioned for CPH. In addition, high-throughput KM survival analysis using a single threshold has been shown to be sensitive to patient group re-sampling (**???**). However, we can get around the first assumption by conducting the log-rank test along a range of splits, where the threshold that splits the expression-ordered list of patients into low- and high-expression is tested at multiple points. The second assumption is weaker in KM than in CPH, but it is important that survival-significant molecules do not have KM curves that cross.

Conducting KM across a range of thresholds for each molecules produces a range of p-values, of which we could sample the minimum. However, taking the lowest p-value from a range will produce a non-uniform, right-skewed distribution of p-values. Since p-values should be uniform under the null distribution, the skewed distribution cannot be used for valid statistical analysis. An equation was developed that could predict the correct p-values was developed (**???**); however, it is not precise enough for p-value correction procedures that are sensitive to very small p-value changes. Thus, we developed NEEP, which overcomes these issues by re-sampling permutations of the patients to construct a null distribution in parallel. Using this null distribution, NEEP transforms the p-values so they are statistically valid. Finally, NEEP conducts FDR p-value correction and calculates effect sizes, the hazard ratio and the 1, 2, and 5 year mortality ratios.

Statement of Need

Research purpose: `NEEP` offers non-parametric, high-throughput, and statistically valid survival analysis of molecular expression vectors.

Problem solved: `NEEP` permutes different orders of patients and calculates their p-values across a range in order to produce a 'Null Empirically Estimated P-value' for each molecular

expression vector, thus overcoming the assumptions of Cox-PH survival analyiss and the issues of correct p-value estimation.

Target audience: The target audience is anyone conducting molecular, high-throughput survival analysis that does not have confounding clinical variables and whose expression vectors may violate Cox-PH assumptions. Full documentation is available in the project repository.

Acknowledgements

This project was partly funded by a University of Nebraska Collaboration Initiative/ System Science Seed Grant to Sushil Kumar, Hesham Ali, and Dario Ghersi and by the NIH AA026428 R21 grant to Sushil Kumar. The funder website is https://nebraska.edu/collaborationinitiative/. The funders had no role in project design, data collection and analysis, decision to publish, or preparation of the manuscript.

References