Survival Bump Hunting For Identification and Characterization of Informative Prognostic Subgroups

Michael Choe[1]*, Jean-Eudes Dazard, PhD[1]*†, J. Sunil Rao, PhD[3]†

[1]Division of Bioinformatics, Center for Proteomics and Bioinformatics, School Of Medicine, Case Western Reserve University;
[2]Department of Biostatistics, School of Public Health, University of Washington, Washington, Public Health Sciences,
[2]Fred Hutchinson Cancer Research Center;
[3]Division of Biostatistics, Department of Epidemiology and Public Health, The University of Miami, Florida.

*Equal contributions, †Corresponding authors.

Abstract

The emergence of large volume microarray data creates a unique opportunity in applying high-dimensional search algorithms to identify biomarkers relevant for either classification or prognostication. Survival Bump Hunting (SBH) is a rule induction method utilizing clinical covariates for the induction of the most relevant indicator variables of extreme survival groups or bumps. We propose the utilization of SBH in a proof of concept endeavor to identify extreme prognostic groups among 20 publically available clinical and genomic datasets where the number of covariates either remains small in comparison to the number of observations ($p < n$) or dominates it ($p >> n$). Datasets included studies of various pathologies (11 breast cancer, 1 lung cancer, 1 prostate cancer, 1 multiple myeloma, 1 Hodgkin's lymphoma, 1 bladder cancer, 1 follicular cell lymphoma, 1 primary biliary cirrhosis, 2 HIV) where the response variables included either overall survival, cause specific survival, disease free survival, progression free survival, or metastasis free survival. We report peeling trajectories against subgroup supports of covariates, hazard ratios, log-rank statistics and prediction-error statistics as well as event-free times and probabilities and median time-to-events. Trace curves of covariates variable importance and usage as well as Kaplan-Meier survival probability curves with log-rank p-values for these subgroups are also evaluated. SBH was able to identify clear survival bumps based on Kaplan Meier plot analysis in 12 of the 20 datasets ($p < 0.01$). The identification of these groups remained robust in 9 out of 20 datasets after cross-validation-replication ($p << 0.01$). The clinical implications of the selected covariates are discussed in detail and the applications of SBH on large omics data is further explored. An R package `PRIMsrc` is available on GitHub.

Keywords

Bump Hunting, Patient Rule Induction Method, Non-Parametric Methods, Survival/Risk Analysis, Prognostication, Predictive Prognosis, Precision Medicine