

# Survival Bump Hunting For Identification and Characterization of Informative Prognostic Subgroups

Michael Choe<sup>1\*</sup>, Jean-Eudes Dazard, PhD<sup>1\*†</sup>, J. Sunil Rao, PhD<sup>3†</sup>

<sup>1</sup>Division of Bioinformatics, Center for Proteomics and Bioinformatics, School Of Medicine, Case Western Reserve University; <sup>2</sup>Department of Biostatistics, School of Public Health, University of Washington, Washington, Public Health Sciences, <sup>2</sup>Fred Hutchinson Cancer Research Center; <sup>3</sup>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. SurvivalPRIM is a modified algorithm of PRIM (patient 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 SurvivalPRIM in a proof of concept endeavor to identify extreme prognostic groups among 20 publically available clinical datasets where the number of covariates remains small in comparison to the number of observations ( $p < n$ ). 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 of predictors, hazard ratios, ranking statistics, event-free probabilities, and median time-to-events against subgroup supports. SurvivalPRIM was able to identify clear survival bumps based on Kaplan Meier plot analysis in 15 of the 20 datasets ( $p < 0.01$ ). The identification of these groups remained robust in 10 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 SurvivalPRIM on large omics data is further explored.

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