Non-parametric survival analysis in breast cancer using clinical and genomic markers

Søren Sønderby and Ole Winther

Background: New survival models based on Gaussian Processes (GP) and Random Forests (RF) have been developed, and have shown good performance in large cancer cohorts.

Purpose: To investigate if these new survival models can improve prediction of 10 year recurrence in a pooled dataset of breast cancer patients and to investigate whether inclusion of gene expression data can improve the baseline clinical predictor.

Data Sources: Breast cancer patients collected by (Haibe-Kains et al. 2012).

Data Extraction: Patient clinical data and gene expression data from several platforms were extracted. Clinical data, including receptor status, was incomplete. Methods for inference of ER, HER2 and PgR receptor status from gene expression data were developed. These methods work independently of the gene expression platform. Recurrence predictors were extracted from expression data.

Results: A pilot study showed that RF survival had worse performance than GP based models. RF survival was not investigated further. Area under curve (AUC) scores for recurrence prediction in breast cancer patients was calculated for the models Cox GP model (CoxGP) and Cox proportional hazard (CoxPH). When appropriate, models were evaluated on dataset with different number of covariates.

Limitations: The included data is a pooled dataset and may be skewed.

Conclusion: CoxGP models show similar performance to CoxPH. It is shown that addition of features extracted from gene expression data improve prediction of 10 year recurrence in both CoxGP and CoxPH models.