Datasets

Breast Cancer microarray datasets, ER+ individuals and DMFS response
8 datasets, with 1555 patients

Cleaning

(Dataset section)

- Linear scaling of expression values
- Missing values in gene expressions filled with KNN imputation
- Remove 60% low variance genes
- Remove duplicate samples or samples with missing values
- Remove studies with too few samples
- Curation of covariates: size, age, grade *(Supplementary table 1 for availability)*

7 datasets, with 1021 patients (*Table 1*)

Simulate Independent Datasets

(Simulation Approach section)

Step 1: Resample studies with replacement

Step 2: Resample individuals with replacement

Step 3: Use generative model to simulate response

Non-parametric Bootstrap

Parametric Bootstrap

True Model:

Fit CoxBoost -> true coefficients

True cumulative hazards

Randomly draw samples from U(0,1) and Compute T and C
(function 2-4)

Within- and across study validation

(Simulation Approach section)

Repeat for 100 iterations:

Simulate a list of datasets with the method above

For all i, j

Train on dataset i, Test on dataset j

4-fold CV on set i

A matrix of C Index:

(i, j) element

(i, i) element

1 CV and 1 CSV value: Average of diagonal and off-diagonal elements

(Supplementary figures 1)

Alterations of sources of heterogeneity

Proportions of covariates: size, age, grade	Re-weight probabilities of resampling	Improves CV and CSV, the difference is not equalized *(Supplementary figure 1, panel b)*
Gene Expression Covariance	Filter genes with high Integrative Correlation	Increases the variances of CSV, does not eliminate the difference *(Supplementary figure 1, panel c)*
Difference in true models	Use identical model from all sets combined	Eliminate the difference *(Supplementary figure 1, panel d)*
All sources combined		The gap is greatly reduced and almost eliminated *(Supplementary figure 1, panel f)*
Remove outlier CAL		Only the absolute value of the gaps are changed, the trends stay the same. *(Supplementary figure 7)*