Annotation of data file used in Ali HR et al. PLoS Med 2016:

Dimensions: 10,988 observations; 52 columns

General rule for dichotomous variables: 0 – negative; 1 - positive

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| inputsample | Unique sample identifier |
| dataset | Study or dataset name |
| er | ER status |
| pgr | PR status |
| her2 | HER2 status |
| size | Tumour size (cm) |
| node | Node status |
| age | Patient age |
| grade | Histological grade (1-3) |
| tdmfs | Distant metastasis free survival, time (days) |
| edmfs | Distant metastasis free survival, events |
| trfs | Relapse free survival, time (days) |
| erfs | Relapse free survival, events |
| tos | Overall survival, time |
| eos | Overall survival, events |
| tbcss | Breast cancer specific survival, time (years) |
| ebcss | Breast cancer specific survival, events |
| surv | Consolidated survival variable used in analysis, see Methods for details. (years) |
| events | Consolidated events variable used in analysis, see Methods for details. |
| tumour\_size\_d | Tumour size, categorical. |
| path\_cr | Pathological complete response |
| t\_stage | T stage |
| n\_stage | N stage |
| strata\_n | Strata variable used to produce stratified estimates |
| ic\_4\_split | IntClust classification where IntClust 4- = IntClust 4, ER-; IntClust 4+ = IntClust 4, ER+ |
| cytolyticactivity | Immune cytolytic activity as calculated by Rooney et al. Cell 2015 for TCGA samples |
| cyt\_mb | Immune cytolytic activity as calculated for METABRIC samples (see Methods). |
| **CIBERSORT OUTPUT (https://cibersort.stanford.edu/)** | |
| pvalue | Empirical CIBERSORT p-value |
| pearsoncorrelation | Pearson’s correlation coefficient limited to genes in the signature matrix |
| rmse | Root mean squared error between original mixture and signature matrix |
| All other variables (bcellsnaive through to neutrophils) represent the proportions of the 22 immune cell subsets estimated by CIBERSORT. | |