Table S1: A Description of the Variables of the LSM-10+ Year Datasets

|  |  |  |
| --- | --- | --- |
| Variables included | Description | Values |
| *race* | race of patient | white, black, Asian, American Indian or Alaskan native, native Hawaiian or other Pacific islander |
| *ethnicity* | ethnicity of patient | not Hispanic, Hispanic |
| *smoking* | smoking history of patient | ex smoker, non smoker, cigarettes, chewing tobacco, cigar |
| *alcohol usage* | alcohol usage of patient | moderate, no use, use but nos (non otherwise specified), former user, heavy user |
| *family history* | family history of cancer | cancer, no cancer, breast cancer, other cancer, cancer but nos |
| *age\_at\_diagnosis* | age at diagnosis of the disease | 0-49, 50-69, >69 |
| *menopausal\_status* | inferred menopausal status | pre, post |
| *side* | side of tumor | left, right |
| *TNEG* | triple negative status in terms of patient being ER, PR, and HER2 negative | yes, no |
| *ER* | estrogen receptor expression | neg, pos, low pos |
| *ER\_percent* | percent of cell stain pos for ER receptors | 0-20, 20-90, 90-100 |
| *PR* | progesterone receptor expression | neg, pos, low pos |
| *PR\_percent* | percent of cell stain pos for PR receptors | 0-20, 20-90, 90-100 |
| *P53* | whether P53 is mutated | neg, pos, low pos |
| *HER2* | HER2 expression | neg, pos |
| *t\_tnm\_stage* | prime tumor stage in TNM system | 0, 1, 2, 3, 4, IS, 1mic, X |
| *n\_tnm\_stage* | # of nearby cancerous lymph nodes | 0, 1, 2, 3, 4, X |
| *stage* | composite of size and # positive nodes | 0, 1, 2, 3 |
| *lymph\_nodes\_removed* | number of lymph nodes removed | 0-11, 12-22, > 22 |
| *lymph\_nodes\_positive* | number of positive lymph nodes | 0, 1-8, >8 |
| *lymph\_node\_status* | patient had any positive lymph nodes | neg, pos |
| *histology* | tumor histology | lobular, duct |
| *size* | size of tumor in mm | 0-32, 32-70, >70 |
| *grade* | grade of disease | 1, 2, 3 |
| *invasive* | whether tumor is invasive | yes, no |
| *histology2* | tumor histology subtypes | IDC, DCIS, ILC, NC |
| *invasive\_tumor\_location* | where invasive tumor is located | mixed duct and lobular, duct, lobular, none |
| *DCIS\_level* | type of ductal carcinoma in situ | solid, apocrine, cribriform, dcis, comedo, papillary, micropapillary |
| *re\_excision* | removal of an additional margin of tissue | yes, no |
| *surgical\_margins* | whether residual tumor | res. tumor, no res. tumor,  no primary site surgery |
| *MRIs\_60\_surgery* | MRIs within 60 days of surgery | yes, no |
| *distant recurrence (Metastasis)* | This is the target variable | yes, no |