|  |  |  |
| --- | --- | --- |
| Supplementary Table 1. Summary of variables used for imputation in SEER-18 excluding Alaska registries. |  |  |
| Female Breast Cancer Cases Diagnosed in 2010-2013 and followed until December 31, 2014.  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|   |   |   |   | **Overall N=196094** |  |  |  |  |
| **No** | **Variable Type** | **Variable Levels** | **Variable Names** | **N Missing** | **% Missing**  |  |  |  |  |
| 1 | Nominal | 12 | Age at diagnosisa | 0 | 0.00% |  |  |  |  |
| 2 | Nominal | 2 | Ethnicity | 0 | 0.00% |  |  |  |  |
| 3 | Nominal | 17 | Registry | 0 | 0.00% |  |  |  |  |
| 4 | Nominal | 4 | Year of diagnosis | 0 | 0.00% |  |  |  |  |
| 5 | Binary | 2 | Reporting Source | 0 | 0.00% |  |  |  |  |
| 6 | Continuous  |  | Baseline Cumulative Hazardb | 0 | 0.00% |  |  |  |  |
| 7 | Binary | 2 | Vital Status | 0 | 0.00% |  |  |  |  |
| 8 | Nominal | 3 | Poverty 2010 (county level)c, e | 47 | 0.02% |  |  |  |  |
| 9 | Nominal | 3 | Urban 2010 (county level)d, e | 47 | 0.02% |  |  |  |  |
| 10 | Binary | 2 | Surgerye | 651 | 0.30% |  |  |  |  |
| 11 | Nominal | 4 | Racee | 1545 | 0.79% |  |  |  |  |
| 12 | Nominal | 4 | Histology | 2358 | 1.20% |  |  |  |  |
| 13 | Nominal | 4 | Insurance | 4677 | 2.48% |  |  |  |  |
| 14 | Binary | 2 | Nodal status | 5222 | 2.66% |  |  |  |  |
| 15 | Nominal | 4 | AJCC 7th Stage | 6608 | 3.40% |  |  |  |  |
| 16 | Binary | 2 | ER Status | 7965 | 4.06% |  |  |  |  |
| 17 | Binary | 2 | PR Status | 8763 | 4.50% |  |  |  |  |
| 18 | Nominal | 6 | Tumor Size | 8830 | 4.50% |  |  |  |  |
| 19 | Binary | 2 | Marital Status | 11448 | 5.48% |  |  |  |  |
| 20 | Binary | 2 | HER2 Status | 16563 | 8.44% |  |  |  |  |
| 21 | Nominal | 3 | Bloom Richardson Grade | 27148 | 13.80% |  |  |  |  |
| Abbreviations: SEER= Surveillance, Epidemiology, and End Results (SEER); ER = Estrogen receptor; PR = progesterone receptor; HER2= human epidermal growth factor 2-neu.a: Age 0-34 were combined for imputation model. |  |  |  |  |  |  |
| b: Used vital status and the Nelson-Aalen estimator of the cumulative baseline hazard in the imputation model for imputing missing HER2 status for the survival analyses. |
| c: Poverty was used as low (<10% poverty), medium (10.00 -19.99% poverty), and high level (>20% poverty) in the imputation model. |  |
| d: Urban variable was used as less urban (<50%) vs. more urban (>=50%) in the imputation model. |  |  |  |  |
| e: Poverty, urban, race, and surgery variables had very little missing information. Therefore, we used a single imputation model controlling for age, Hispanic ethnicity, registry, year of diagnosis to fill in missing information for these four variables.  |