Method_Statistical_methods

Zhezheng Jin

2023-12-06

Statistical methods

The primary objective of this research is to evaluate and compare the effectiveness of hormonal treatment and chemotherapy in improving the survival rates of breast cancer patients. Utilizing the comprehensive Rotterdam dataset, which includes data from 2982 primary breast cancer patients, this study aims to discern the impact of these treatments on patient survival. Variables such as patient demographics, cancer characteristics, and treatment details are analyzed to determine how different therapeutic approaches influence survival outcomes.

In our survival analysis of the Rotterdam dataset, we define key variables to accurately capture patient outcomes. The event indicator variable rfs is created using `pmax(rotterdam$recur, rotterdam$death)`, which combines breast cancer recurrence and death into a single event endpoint. This means rfs is marked as 1 if either event occurs. For the survival time, we use rfstime, determined by `ifelse(recur == 1, rtime, dtime)`. This approach selects the time to recurrence or last follow-up (rtime) if it occurred; otherwise, it uses the time to death or last follow-up (dtime). These combined variables allow for a comprehensive analysis of survival outcomes, considering both recurrence and mortality as critical events.

For estimating and comparing survival probabilities among breast cancer patients under different treatments, the Kaplan-Meier estimation method is utilized. This non-parametric approach, crucial in survival analysis, is adept at estimating the survival probability over time for distinct patient groups categorized by their respective treatment regimens—hormonal treatment, chemotherapy. The Kaplan-Meier estimator is particularly valuable for its ability to effectively visualize the survival function. This visualization is instrumental in comprehending the time-to-event distributions for the various treatment groups in our study. By employing this method, we can make preliminary comparisons of survival experiences, providing an initial insight into how different treatments impact patient survival outcomes in the context of breast cancer.

To compare the difference in survival probabilities of different treatment group and test if this difference is significant, we employ the log-rank test, a non-parametric method. This test is applied in our research to assess the survival differences across patient groups categorized by their treatment types: hormonal treatment, and chemotherapy. Additionally, we utilize a stratified log-rank test for a more nuanced analysis. This approach specifically examines the role of hormonal treatment when stratified by the administration of chemotherapy. This stratification gives a more detailed examination of how chemotherapy influences the effectiveness of hormonal treatments in different patient subsets.

We use the Cox proportional hazards regression model to evaluate the relative impact of various covariates on the survival of breast cancer patients. Specifically, our model includes predictors such as age (modeled using penalized splines), tumor size, differentiation grade, the number of positive lymph nodes, progesterone receptor levels, estrogen receptor levels, hormonal treatment, and chemotherapy. This semi-parametric model is particularly beneficial for its capability to manage multiple risk factors and adapt to different types of survival data. To verify the proportional hazards assumption, a fundamental aspect of the Cox model, we employ diagnostic methods like Schoenfeld residuals and graphical checks. These methods are essential to ensure that the hazard ratios remain constant over time, thus validating the reliability of our model.

Our analysis also carefully considers the nature of censoring present in the Rotterdam dataset. The dataset predominantly exhibits right-censoring, where certain patients' follow-up information is incomplete due to their event (death or recurrence) not being observed within the study period, such as loss to follow-up. Therefore, the assumptions applicable for here are Independence of Censoring and Non-Informative Censoring. To check for Independence of Censoring, we plot Kaplan-Meier survival curves for different subgroups of patients based on factors potentially related to censoring (like age, treatment type). If the survival experiences of these subgroups diverge significantly, it might indicate a relationship between these factors and censoring. To check for Non-Informative Censoring, we use Kaplan-Meier curves to visually inspect the survival patterns of censored and uncensored groups. Disproportionate divergence between these curves can be indicative of informative censoring.

Formulae for Selected Methods

Kaplan-Meier Estimation:

$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$

Where $S(t)$ is the survival probability at time t_i, d_i is the number of events at time t_i, and n_i is the number of subjects at risk at time t_i.

Cox Proportional Hazards Model:

$h(t) = h_0(t)\exp(\beta_1 \cdot \text{pspline(age)} + \beta_2 \cdot \text{size} + \beta_3 \cdot \text{grade} + \beta_4 \cdot \text{nodes} + \beta_5 \cdot \text{pgr} + \beta_6 \cdot \text{er} + \beta_7 \cdot \text{hormon} + \beta_8 \cdot \text{chemo})$

Where $h(t)$ is the hazard at time t_i, $h_0(t)$ is the baseline hazard, $\beta_1, \beta_2, ..., \beta_8$ are the coefficients for each covariate, which include age modeled with a penalized spline, tumor size, grade, number of positive lymph nodes, progesterone receptor levels, estrogen receptor levels, hormonal treatment, and chemotherapy, respectively.