Abstract

Chemotherapy for cancer treatment is costly and accompanied by severe side effects, highlighting the critical need for early prediction of treatment outcomes to improve patient management and informed decision making. Predictive models for chemotherapy outcomes face challenges including sparsity in biomarkers and the absence of explicit phenotypes and treatment outcome labels such as cancer progression and toxicity. This study addresses these challenges by employing Large Language models (LLMs) and ontology- based techniques for phenotypes and outcome labels extraction from patients notes. We focused on one of the most occurring cancers i.e., Breast Cancer patient’s dataset with features including vitals, demographics, staging, biomarkers and performance scales. Drug regimens and their combinations were extracted from the chemotherapy plans in the EMR data and shortlisted based on NCCN guidelines, verified with NIH standards, and analyzed through survival modeling. The proposed approach significantly reduced biomarkers sparsity and improved predictive accuracy. Random Survival Forest was used to predict time-to-event achieving a c-index of 73% and utilized as a classifier at a specific time point to predict treatment outcomes, with accuracy and F1 score above 70%. The outcome probabilities were validated for reliability by calibration curves. This research highlights the potential of early prediction of treatment outcomes using LLM based clinical data extraction enabling personalized treatment plans with better and improved patient outcomes.