Abstract

Focuses on Breast Cancer Chemotherapy Outcome prediction utilizing survival model, results, conclusions, challenges like insufficient biomarkers/indicators, treatment outcomes and their intuition data addressed through LLMs and ontology systems, significantly enhances the model’s reliability and accuracy

Introduction

* Financial Burder, insufficient patient recorded data (biomarkers and staging etc), this study aims to improve patients care with Real world patients profiles
* Background Study

Data overview

* Cohort sizes
* Details may be for patient’s demographics
* Failure/discontinuations (any pattern and prevalent reason for failure)
* Stage and age distribution
* A Table showing stats

Reduce sparsity for biomarkers, performance scores and Treatment outcomes

* Method and its result
* Quantitative effect of the approach

Phenotypes and Treatment Outcome Extraction

* LLM extraction
* Ontology based extraction

Postprocessing

* Postprocessing for hallucinated responses, validation steps

Drug and Drug combinations

* Regimen shortlisting from NCCN guidelines
* NIH verified
* Fetch combination and regimens from patient plans
* Criteria to select the combinations and drugs i.e Patient Support and Encoding

Feature Engineering

* **Vitals**: bsa.
* **Diagnosis**: Diagnostic codes and categories.
* **Lab Results**: Srcr.
* **Demographics**: Age.
* **Staging (Overall, T, N, M)**: Encoded staging data.
* **Grade**: Cancer grades.
* **Performance Scale**: Scales like ECOG or Karnofsky.

Survival Modelling

* Time till failure comparison between and failures and no failures **with graph**
* Classification from Survival model (method) **with Graph** showing the evaluation metrics over time.

Results

* Survival’s C-index + Classification results
* Table for classification results
* Calibration Curves **with Graph**

Conclusion

* Overall strategy
* LLM application in features extraction
* Survival modelling and its usage as classifier and its result

References

* https://www.cancer.gov/about-cancer/treatment/drugs/breast
* https://www.nccn.org/patientresources/patient-resources/guidelines-for-patients/breast-cancer-resources
* https://www.mdpi.com/1648-9144/60/1/168

The study explores how molecular markers and genetic determinants, such as DNA repair genes (ERCC1, BRCA1) and EGFR mutations, influence lung cancer treatment outcomes. It highlights the role of biomarkers like thioredoxin, YB-1, and 14-3-3s methylation in predicting chemotherapy resistance and survival, while emphasizing the potential of EGFR-targeted therapies in non-smokers with adenocarcinoma. These findings underscore the importance of personalized treatment strategies to improve patient outcomes. (<https://www.sciencedirect.com/science/article/abs/pii/S1471489206000907> )

The study highlights advancements in identifying predictive biomarkers like thymidylate synthase (TS), ERCC1, and VEGF to improve chemotherapy outcomes for colorectal cancer (CRC). Emerging high-throughput approaches such as gene expression profiling and proteomics aim to enable personalized treatments by predicting patient responses. The authors emphasize the need for standardized studies and validation to integrate these biomarkers into clinical practice  
(<https://www.sciencedirect.com/science/article/abs/pii/S1471489206000853> )

The study compares machine learning models like Random Forest (RF), Recurrent Neural Networks (RNN), and Hidden Markov Models (HMM) to predict chemotherapy toxicity in breast cancer patients. RNNs outperformed others due to their ability to handle sequential data, emphasizing the importance of comprehensive patient information for accurate predictions. This approach aims to support personalized treatments and reduce toxicity during chemotherapy.

The Random Forest (RF) approach is an ensemble learning method that builds multiple decision trees during training and combines their outputs (via majority voting or averaging) to improve prediction accuracy and reduce overfitting. It works well for classification and regression tasks.

**Random Forest in the Study**

In the context of predicting chemotherapy toxicity for breast cancer patients:

* **Purpose**: RF is used to classify the toxicity level based on patient and treatment features.
* **Advantages**:
  + Handles diverse feature types (categorical and continuous).
  + Robust to outliers and noise.
  + Reduces overfitting compared to single decision trees by aggregating multiple tree predictions.

**Features Used**

The RF model was trained using the following features to predict toxicity outcomes:

1. **Age**: Age of the patient.
2. **BMI**: Body Mass Index of the patient.
3. **Cycle**: The treatment cycle number (sequential context of the treatment).
4. **Regime**: Type of chemotherapy regimen (e.g., adjuvant, neo-adjuvant, palliative).
5. **Previous Performance Status (PS)**: The patient’s performance status prior to the current cycle.
6. **Previous Toxicity Level**: The toxicity level observed in the previous cycle.

**Feature Importance**

The RF model automatically calculates the importance of each feature after training. For example:

* Features like **BMI** and **age** had higher predictive importance compared to others because they were strongly correlated with toxicity outcomes.
* Lower-ranked features like **regime type** were less predictive but still contributed to the model’s overall accuracy.

RF excels in this study by providing interpretable feature importance and handling class imbalances using data augmentation techniques​.

(Key: Missing staging, biomarkers, **detailed info regime and dosage information**, grade, comorbid conditions, Other treatment outcomes(I.e Progression, Death/Hospice), Oversample by duplications)

Abstract

Predicting treatment outcomes for cancer patients is restricted by EMR data sparsity in biomarkers and introduces challenges in phenotype extraction. This study introduces a methodology for phenotype and treatment outcome extraction from patients notes using Large Language models (LLMs) and ontology- based techniques, with a focus on enriching patient’s data for biomarkers, performance scores and treatment outcomes. Breast Cancer patient’s dataset was analyzed, with features including vitals, demographics, staging, biomarkers and performance scales. Phenotypes and treatment outcomes including cancer progression, adverse effect or deterioration of quality of life and death or hospice are extracted using a hybrid approach combining LLMs and domain specific ontologies. Post processing techniques were employed to reduce hallucinated outputs. Drug regimens were extracted and shortlisted based on NCCN guidelines, verified with NIH standards, and analyzed for efficacy through patient treatment plans and survival modeling. The proposed approach significantly reduced biomarkers sparsity improving predictive accuracy. The survival model achieved a C-index of 73% while classification results demonstrated accuracy and F1 score above 70% and reliability as evident by calibration curves. The research highlights the potential of advanced feature engineering an LLMs clinical data extraction, offering a personalized treatment plans with better outcomes.