MACHINE LEARNING-BASED PREDICTION OF IN-HOSPITAL MORTALITY IN DIFFUSE LARGE B-CELL LYMPHOMA: A NATIONWIDE ANALYSIS FOR ENHANCED RISK STRATIFICATION

Mariana Marrero Castillo MD[1], Benedict Amalraj MD MPH[1], Shiva Jashwanth Gaddam MD[1], Sheikh Abdullah MD[1], Poornima Ramadas MBBS[1]

1. Louisiana State University Shreveport, Shreveport, LA, USA

Introduction

In-hospital mortality for patients with diffuse large B-cell lymphoma (DLBCL) remains a concern. Early identification of high-risk patients is crucial for timely interventions and optimizing care. Machine learning (ML) offers potential for improved risk stratification.

Objectives

To develop and evaluate ML models, incorporating techniques for imbalanced data, using nationwide inpatient data to predict in-hospital mortality among adult patients with DLBCL, focusing on enhancing the identification of fatal outcomes.

Methods

Design: A retrospective cohort study using the National Inpatient Sample from 2018–2022. Features for prediction were selected, including patient demographics (age, sex, race, payer), hospital characteristics (bed size, location/teaching, region), All Patient Refined–Diagnosis Related Groups (APR-DRG) risk of mortality score, pre-existing comorbidities, and acute in-hospital complications. Numerical features were scaled, and categorical features were one-hot encoded. Synthetic Minority Over-Sampling Technique (SMOTE), Balanced Random Forest, and cost-sensitive learning were used. Models, including Logistic Regression, Random Forest, LightGBM, Balanced Random Forest with SMOTE, cost-sensitive XGBoost, and a Voting Classifier, were evaluated on an 80/20 train/test split.

Setting: United States inpatient hospitalizations.

Patients: The study included 71,890 adult hospitalizations with a DLBCL diagnosis after data cleaning for missing values from an initial 73,474 records. The overall in-hospital mortality rate in the analyzed cohort was 5.1%.

Main Outcome Measures: The primary outcome was in-hospital mortality. Model performance was assessed by accuracy, precision, recall, F1-score (for the "Died" class), and area under the receiver operating characteristic curve (ROC AUC).

Results

The cost-sensitive XGBoost model demonstrated excellent discriminative ability (ROC AUC, 0.922) and a strong focus on identifying mortality.

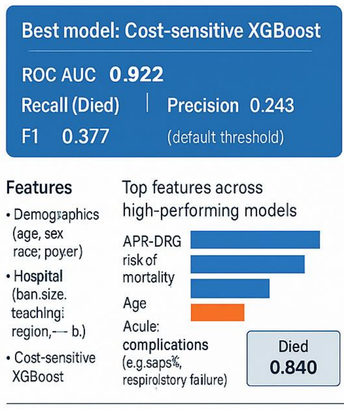
At its default threshold, it achieved a recall (sensitivity) of 0.840 for the "Died" class, a precision of 0.243, and an F1-score of 0.377.

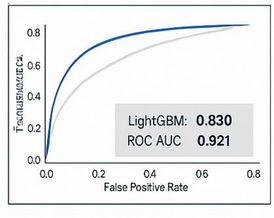
Other models, like LightGBM (recall, 0.839; ROC AUC, 0.921), also showed high recall. Key predictors consistently influencing models included APR-DRG risk score, age, and acute complications.

Conclusions

Advanced ML models incorporating strategies for imbalanced data, particularly cost-sensitive XGBoost, can effectively predict in-hospital mortality with high sensitivity in patients with DLBCL.

These tools can aid in early identification of patients at high risk of dying, facilitate timely palliative care consultations, help discuss goals of care, and potentially optimize resource use by focusing on intensive interventions.





Clinical implications

- High-sensitivity ML can flag patients at risk of in-hospital death.
- Enables eanjer palliative care consults and goals-of-care discussions.
- Supports respurce allocation for intensive interventions

Keywords: DLBCL, in-hospital mortality, machine learning; imbalanced learning; palliative care

Data source: National Inpatient Sample (2018–2022).