PCN219
PREDICTING CHEMOTHERAPY-ASSOCIATED THROMBOCYTOPENIA IN REAL WORLD CLINICAL SETTINGS

Ransom, Joshua F PhD1; Galaznik, Aaron MD MBA1; Lempemeseke, Bruno MS1; Shilnikova, Alexandra1; Berger, Marc MD1
1SHYFT Analytics, a Medidata Company, Boston, MA, USA

Introduction

Chemotherapy-induced thrombocytopenia is a frequent challenge in the management of cancer patients and can limit the ability to maintain effective dosing and treatment duration. In this study, we assessed real-world rates of chemotherapy-associated thrombocytopenia, measured the impact on patient dosing, and explored the potential of machine learning methods, using commonly available clinical variables, to predict development of this common, potentially limiting side effect.

Methods

Data Sources

The retrospective cohort study was conducted in US professional and institutional medical claims sources and in US ambulatory practice electronic medical records sourced from the HealthmarkIQ® Marketplace platform of data suppliers from 1/2013 – 12/2018.

Data Transformation and Analysis

Data was transformed into the CMOP® Common Data Model version 5. Analyses were conducted using the SHYFT Quantum V5.7.0 solution and Python v3.6.

Inclusion Criteria

- Cohort: Patients with ≥1 diagnosis of a female gynaecological (FGU) cancer (ICD10: O31.0-1 ¦ O38.0-2) or ICD9 criterion
  - Index date: 
    - Incidence Rate: first diagnosis of a FGU cancer
      - Machine learning: first diagnosis of thrombocytopenia (TCP; ICD10: D98.3-4 ¦ E852.0; ICD9 criterion) or first diagnosis of a FGU cancer in patients lacking a TCP diagnosis
    - Age ≥18 at index
    - All patients pre- and post-index continued enrollment

Incidence Rate at Risk:

- Numerator: Number of patients with the first TCP diagnosis in the dataset occurring after the index FGU cancer diagnosis
- Denominator: Person-years at risk from index date to the incident TCP diagnosis or death or end of patient observation period or end of the dataset

Machine Learning Prediction of Thrombocytopenia Events

Feature Engineering:

- Covariate pipelines were set up for supervised learning:
  - Patient demographics of Age at Index, gender, race, and ethnicity, using OneHotEncoder on all nominal variables
  - Diagnosis covariates were defined as the existence of a patient record prior to the index date for groups based on 3-digit ICD10 and their descendents (e.g. 490.xx)
  - Medication covariates were defined as the existence of a patient record prior to the index date for groups based on ICD10 and their descendents, including NCPDS J.xx
  - Lab test covariates were defined as the existence of a patient record prior to the index date for groups based on ICD9 codes

- Modeling Methods: Classification models used included Logistic Regression, Gaussian Naive Bayes, Multi-level Perceptron, k-Nearest Neighbor, Linear Discriminate Analysis, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosted Classifier, and K-Boosted Classifier. Models were scored on accuracy and AUC-scores.

- Measures of Association: Features were tested for collinearity and excluded from the data if Pearson Correlation and Variance Inflation Factor scoring, with respective exclusivity thresholds of <0.8 and <0.5 respectively.

- Training and Validation:
  - To test for overfitting, a 4-1 training/testing split was employed on each data set using cross-validation scoring with a K-fold of 10.

- Feature Importance: Relative feature importance was extracted from the ensemble models and the top 1% and top 10% were used to validate the models between datasets. Feature importance was also tested using Recursive Feature Elimination.

Results

Attrition flow for each dataset’s cohort is shown in Figure 1. This incidence rate for all FGU patients was in line with prior work (0.03-0.06 per person-year) (OR in medical claims and 0.01 per person-year (OR in EMR) (1-15). Among treated patients the proportion of diagnosed TCP was in line with prior publications (29-34); in medical claims, 40% (EMR) and 75% in EMR. The TCP incidence rate among treated FGU patients was in line with expectations (0.03-0.07 per person-year (OR in medical claims and 0.03 per person-year (OR in EMR).

Overall model performance was highest in the institutional and professional medical claims and lowest in the ambulatory EMR dataset (Table 1 & 2). This is likely in part because the EMR had fewer patients with recorded treatable infections and high-grade TCP diagnoses. The top performing models were from the class of ensemble methods, especially Gradient Boosted Classifier and K-Boosted Classifier. Out of the approximately 60,000 variables possible, only 1,750 to 2,500 were populated in any given data set. The top 1% of variables across each of three models were blended into a single model and showed strong performance across all three datasets (Figure 2 & Table 2). These models showed a 16% improvement in accuracy and 27% improvement in AUC scores when compared against previously presented algorithms.

Conclusion

The overall study identified real-world rates of thrombocytopenia consistent with previous publications (1-15). Results also indicate opportunities to improve the management of chemotherapy-associated thrombocytopenia through both active management of dose adjustment as well as potential use of point of care predictive algorithms. Future work will focus on hyperparameter tuning, expanding the clinical domains for covariates to procedures, and exploring deep learning approaches to further test and validate the predictive modeling.

References

6. Internal SHYFT analysis. Data as of May 2019