PCN219
PREDICTING CHEMOTHERAPY-ASSOCIATED THROMBOCYTOPENIA IN REAL WORLD CLINICAL SETTINGS

Ransom, Joshua F PhD[1]; Galaznik, Aaron MD MBA[1]; Lempernesse, Bruno MS[1]; Shilnikova, Alexandra[1]; Berger, Marc MD[1]

[1]SHYFT 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[1]. In this study, we assess real-world rates of chemotherapy-associated thrombocytopenia, measure the impact on patient dosing, and explore the potential of machine learning methods, using commonly available clinical variables, to predict development of this common, potentially treatment-limiting side effect.

Methods

Data Sources

- This retrospective cohort study was conducted in US professional and institutional medical claims sources and a US ambulatory practice electronic medical records source from the HealthVerity® Marketplace platform of data suppliers from 1/1/2015 – 12/31/2018

Data Transformation and Analysis

- Data was transformed into the OMOP Common Data Model, version 5
- Analyses were conducted using the SHYFT Quantum V6.7.0 solution and Python v3.6

Inclusion Criteria

- Cohort: Patients with ≥1 diagnosis of a female genitourinary (fGU) cancer (ICD10: C51.xx-c58.xx or ICD9 equivalent)
- Index date:
  - Incidence Rate: first diagnosis of a fGU cancer
  - Machine Learning: first diagnosis of thrombocytopenia (TCP, ICD10: D69.3-6 or ICD9 equivalent) or first diagnosis of a fGU cancer in patients lacking a TCP diagnosis
- Age ≥18 at index
- ≥6-month pre- and post-index continuous 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 descendants (e.g., A00.xx)
  - Medication covariates were defined as the existence of a patient record prior to the index date for groups based on RxNorm ingredients and their descendants, including HCPCS J-codes
  - Lab test covariates were defined as the existence of a patient record prior to the index date for groups based on LOINC codes
- **Modeling Methods:** Classification models used included Logistic Regression, Gaussian Naïve Bayes, Multi-level Perceptron, K-Nearest Neighbor, Linear Discriminant Analysis, Support Vector Machine, Decision Tree, Random Forest, Gradient Boosted Classifier, and XG Boosted Classifier. Models were scored on accuracy and AUC-scores
- **Measures of Association:** Features were tested for collinearity and excluded from the data frame if Pearson Correlation and Variance Inflation Factor scoring, with respective exclusionary thresholds of >0.8 and >5.0 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. Features importance was also tested using Recursive Feature Elimination

Results

Attrition flow for each dataset's cohort is shown in Figure 1. The incidence rate for all fGU patients was in line with prior work (0.03-0.06 per person-year @ risk in medical claims and 0.01 per person-year @ risk in EMR)[3,4,5,6]. Among treated patients the proportion diagnosed with TCP was in line with other prior publications (29-34% in medical claims, 48% in EMR). The TCP incidence rate among treated fGU patients was also in line with expectations (0.06-0.07 per person-year @ risk in medical claims and 0.03 per person-year @ risk 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 injectable treatments and high-grade TCP diagnoses. The best performing models were from the class of ensemble methods, especially Gradient Boosted Classifier and XG Boosted Classifier. Out of the approximately 60,000 variables possible, only 1,700 – 2,500 were populated in any given data set. The top 1% of variables across each of three models were blended in to a single model and showed strong performance across all three datasets (Figure 2, Table 3). These models showed a 46% improvement in accuracy and 274% improvement in AUC score when compared against previously presented algorithms[2].

Conclusion

Overall the study identified real-world rates of thrombocytopenia consistent with previous publications[3,4,5]. 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

1. Kuter, DJ. Managing Thrombocytopenia Associated with Cancer Chemotherapy. Cancernetwork.com 29(4). Accessed 12/20/2018 (http://www.cancernetwork.com/oncology-journal/managing-thrombocytopenia-associated-cancer-chemotherapy)
2. Lord R, Mirza MR, Woelber L, et al. Safety and dose modification for patients with low body weight receiving niraparib in the ENGOT-OV16/NOVA phase III trial. Presented at: SGO Annual Meeting on Women's Cancer; March 24-27, 2018. New Orleans. Abstract 20.
3. Cassidy CA, et al. Incidence of thrombocytopenia with gemcitabine-based therapy and influence of dosing and schedule. Anti-Cancer Drugs April 2001, 12(4): 383-385.
4. Ten Berg MJ et al. Thrombocytopenia in adult cancer patients receiving cytotoxic chemotherapy: results from a retrospective hospital-based cohort study. Drug Safety, Dec 2001, 34(12):1151-60.
5. Mahner S. Carboplatin and pegylated liposomal doxorubicin versus carboplatin and paclitaxel in very platinum-sensitive ovarian cancer patients: results from a subset analysis of the CALYPSO phase III trial. Eur J Cancer. 2015 Feb;51(3):352-8.
6. Internal SHYFT analysis. Data on file. May 2019

Figure 1. Attrition flow and number of covariates populated for dataset a) institutional and professional medical claims – private source 14, b) professional medical claims – private source 34, c) Ambulatory EMR – private source 42

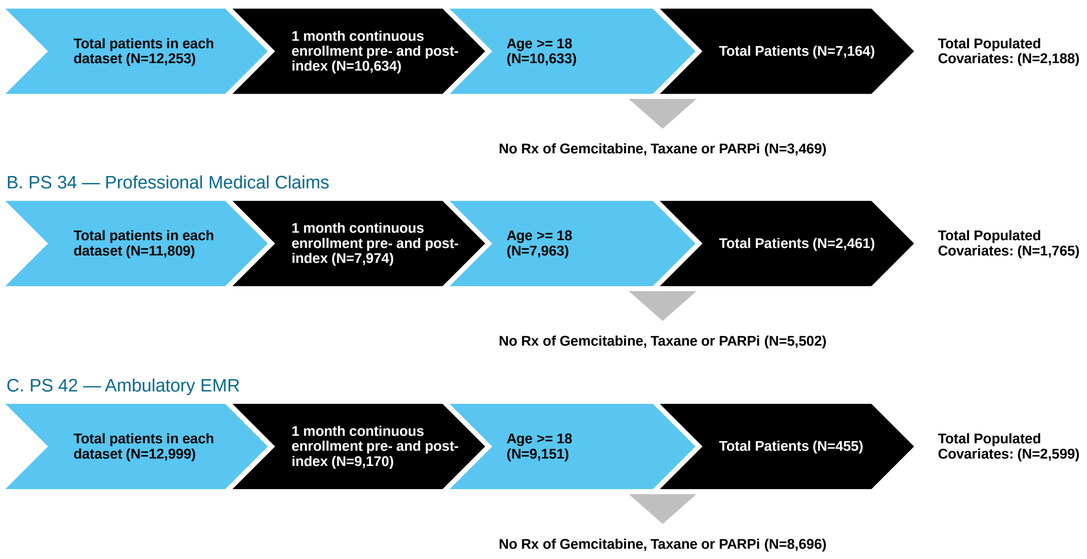

Table 1. Average model accuracy on training data set for each analytic method

Table : Model Training Performance	Institutional and Professional Medical Claims		Ambulatory EMR	
Classification Model	**Avg Accuracy**	**Avg St Dev**	**Avg Accuracy**	**Avg St Dev**
Logistic Regression	93.45%	0.70%	97.43%	0.38%
Naïve Bayes	87.23%	0.77%	91.25%	0.56%
Multi-level Perceptron	84.54%	7.04%	97.25%	0.23%
K-Nearest Neighbor	89.16%	0.65%	95.81%	0.49%
Linear Discriminant Analysis	94.58%	0.67%	97.24%	0.51%
Support Vector Machine	94.47%	0.64%	95.75%	0.41%
Decision Tree	93.23%	0.79%	96.03%	0.29%
Random Forest	94.79%	0.80%	97.38%	0.33%
Gradient Boosted	95.52%	0.76%	97.39%	0.40%
XG Boost	95.61%	0.78%	97.43%	0.35%

Table 2: Model accuracy and AUC-score on validation data set for each analytic method

Model Performance — Validation	Institutional and Professional Medical Claims		Ambulatory EMR	
Classification Model	**Accuracy**	**ROC AUC Score**	**Accuracy**	**ROC AUC Score**
Logistic Regression	92.25%	88.79%	97.87%	74.93%
Naïve Bayes	86.41%	86.83%	91.49%	78.33%
Multi-level Perceptron	57.53%	68.74%	97.49%	75.15%
K-Nearest Neighbor	88.33%	81.01%	96.34%	54.37%
Linear Discriminant Analysis	94.17%	89.36%	97.45%	74.71%
Support Vector Machine	93.96%	89.36%	96.37%	54.39%
Decision Tree	92.37%	89.76%	96.34%	75.39%
Random Forest	93.84%	90.49%	97.91%	74.10%
Gradient Boosted	94.94%	91.01%	97.70%	75.68%
XG Boost	95.10%	91.08%	98.05%	75.86%

Figure 2. ROC-AUC plots for logistic regression using the top features for dataset a) institutional and professional medical claims – private source 14, b) professional medical claims – private source 34, c) Ambulatory EMR – private source 42

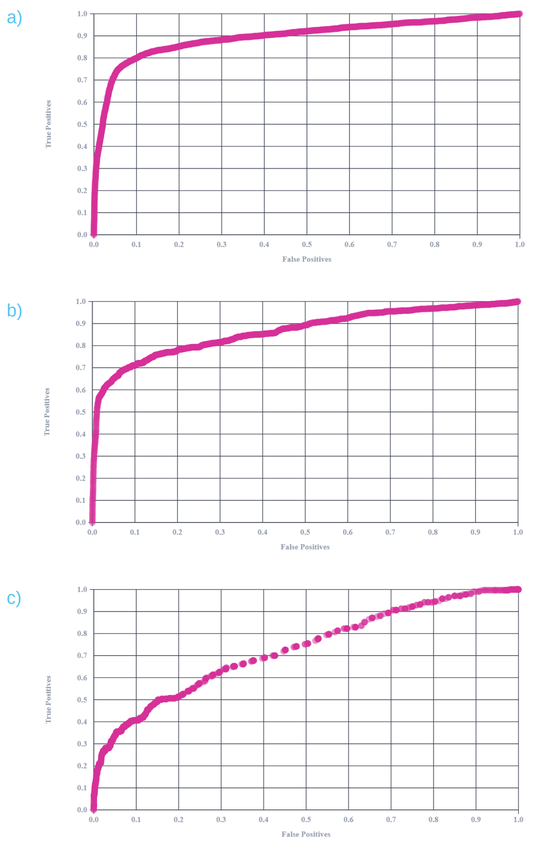

Table 3. Top features across both medical claims and EMR datasets. Statistical significance, odds ratio and confidence intervals are based on logistic regression from Quantum, run on the professional and institutional medical claims dataset (PS 14)

Statistically Significant Features				
Feature	**P-value**	**Odds Ratio**	**2.5% CI**	**97.5% CI**
Age	0.0062	0.9892	0.9816	0.997
Ovarian Cancer	<0.0001	2.056	1.7723	2.3864
Digestive Organ Cancer	0.0125	1.3572	1.0682	1.7252
Polyp	0.0081	0.6244	0.4395	0.8824
Other Female Genitourinary Cancers	0.0276	1.1719	1.0181	1.3502
Secondary Cancer	<0.0001	2.7577	2.3258	3.2685
Artery Disease	<0.0001	1.5342	1.265	1.8595
Esophagus Disorder	0.0043	1.3036	1.0859	1.5635
Count of Visits	<0.0001	1.0014	1.001	1.0019
Fluoxetine	<0.0001	11.5481	8.7192	15.3739
Norfloxacin	0.0083	1.3541	1.0785	1.6923
Thiosalicylate	<0.0001	2.4201	2.0696	2.8286