PCN428

Scalable Prediction of Cancer Patient Clinical Events in Real-World Clinical Settings

Ransom JF[1], Buderi R[1], Galaznik A[1], McLean C[1], Shilnikova A[1], Lempernesse B[1], Berger M[1]

[1]Medidata Solutions, Boston, MA, USA

Background

- Patient safety concerns, both related and unrelated to treatments, are a critical challenge to managing patient care in clinical oncology, which can limit treatment adherence and impact optimal clinical outcomes.[1,2]
- The concept of machine learning to identify predictors of adverse events is currently under active exploration by many organizations, including regulatory bodies.
- There is an increasing role of machine learning in healthcare, as indicated by the United States (US) Food and Drug Administration's recent proposed guidance on artificial intelligence and machine learning-based software as a medical device.[3]
- Starting with a population of patients diagnosed with female genitourinary cancer (fGU), including both ovarian and endometrial malignancies, we outline an approach leveraging common data standards for predicting patient clinical events across multiple real-world data sources, in this case thrombocytopenia, which is common to oncology treatments in these populations.[4-6]
- We then demonstrate how this approach can be scaled up to simultaneously assess multiple sources of patient morbidity or adverse events and identify common predictors across them.

Methods

Data Source

- Health plan claims were obtained from the HealthVerity Marketplace platform of data suppliers from Feb 2014 – Dec 2018.
    - HealthVerity™ has the most complete coverage of US healthcare, consumer, and purchase data with access to over 330 million patients and 30 billion transactions.[7]
    - HealthVerity™ Private Source (PS) 14 was derived from an institutional medical claims provider of predominantly Medicare Fee-For-Service population, and PS34 from medical claims of a predominantly commercially insured population. Unique source patient IDs from HealthVerity™ were used to ensure non-duplication of patients between datasets.

Data Transformation and Analysis

- Data were converted to the Observational Medical Outcomes Partnership Common Data Model, version 5.[7]
- Analyses were conducted in SHYFT Quantum version 6.7.0 and Python version 3.6.

Study Design

Inclusion/Exclusion Criteria (Figure 1)

- Diagnosis codes ≥1 diagnosis of a fGU cancer (International Classification of Diseases [ICD]-10: C51.xx-c58.xx or ICD-9 equivalent)
- At least one treatment for gemcitabine, platinum-based agents or taxanes
- Index date and continuous enrollment
    - Minimum diagnosis dates were defined for all patients for 3-digit ICD-10 diagnosis codes for target conditions evaluated, and were set as earliest diagnosis date for patients with at least 90 days pre-index continuous enrollment
- Descriptive statistics were assessed for baseline demographics and characteristics

Descriptive Summary Statistics

- Age, gender, and endometrial cancer diagnosis (Table 1).

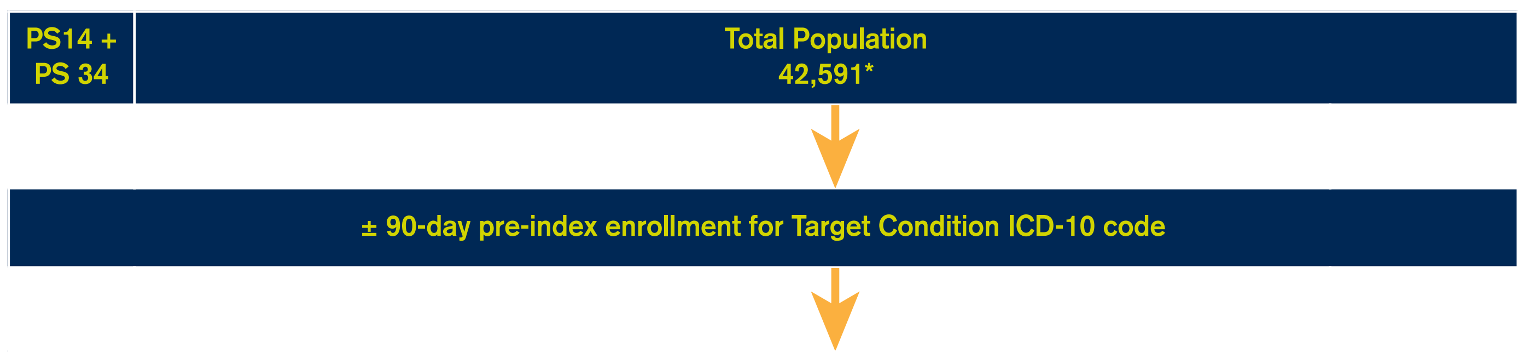
Thrombocytopenia Prediction

- Cohort Selection
    - Index event defined as diagnosis for thrombocytopenia (using ICD-10 D69) post-treatment initiation for target group. Non-target group was set as at least 7,000 individuals, without thrombocytopenia, selected at random, to keep classes balanced.
    - Patients had at least 90 days pre-index continuous enrollment.
- Modeling Methods: A variety of machine learning approaches (Logistic Regression, Gaussian Naïve Bayes, Multi-layer Perceptron, K-Neighbors, Decision Tree, Random Forest, Gradient Boosted Tree, XG Boosted Tree Classifiers)[8,9] were used to develop a predictive algorithm for each underlying comorbidity, with a 3:1 training/testing split, and cross-validation scoring with a K-fold of 6 was used to train the model.
- Models were scored on area under the curve (AUC) scores.
- Training and Validation: To test for overfitting, a 3:1 training/testing split was employed on each data set using cross-validation scoring with a K-fold of 6 for training evaluation. The model was refit and retrained with the parameters resulting in best K-fold training score after a grid search,[10] and prediction probabilities were generated for the test holdout set.
- Feature Importance: Importance values were calculated using XGBoost models for each condition. Feature importance was assessed using the default "gain" measurement from the Python XGBoost package (Table 2).[9]

Expanded Prediction

- Modeling approach expanded to include 19 additional conditions (Figure 1).
    - Conditions included a variety associated with oncology and others to serve as negative controls. Endometrial and ovarian cancer were also included as positive controls.
    - Index event was defined as 3-digit ICD-10 code for target condition for prediction.
    - Patients had at least 90 days pre-index continuous enrollment.
    - At least 7,000 individuals were chosen randomly from other conditions to serve as non-target control group.
    - Features were generated for Diagnosis, Procedure, and Drug Exposures for medical histories for each patient within the 90-day lookback period before index.
    - Top 300 features were selected using Recursive Feature Elimination.[11]
- Modeling, Training, and Validation repeat as above
    - Top model selected for each condition: Models with AUC >0.7 were selected for further analysis (Table 3, Figure 2).
    - Receiver operating characteristic (ROC) curves were plotted for AUC, and Correlation Matrices were plotted to assess potential multicollinearity.
- Feature Importance:
    - For top performing models, importance values were calculated for top features. Features were ranked by selection as a predictor across all 20 conditions assessed.

Figure 1: Cohort Attrition Flow: All Conditions



PS14 + PS 34

Total Population 42,591*

± 90-day pre-index enrollment for Target Condition ICD-10 code

ICD-10 Code	Target Condition	Target Population**
T78	Adverse effects, not elsewhere classified	5,147
D72	Other disorders of white blood cells	4,501
D89	Other disorders involving the immune mechanism, not elsewhere classified	4,673
D69	Purpura and other hemorrhagic conditions	3,405
D64	Other anemias	6,684
K59	Other functional intestinal disorders	3,723
K76	Other diseases of liver	3,507
M99	Biomechanical lesions, not elsewhere classified	7,719
J98	Other respiratory disorders	8,419
N28	Other disorders of kidney and ureter, not elsewhere classified	5,450
K31	Other diseases of stomach and duodenum	6,373
G99	Other disorders of nervous system in diseases classified elsewhere	6,946
R40	Somnolence, stupor, and coma	6,952
N39	Other disorders of urinary system	7,880
B99	Other and unspecified infectious diseases	7,828
M12	Other and unspecified arthropathy	5,500
L99	Other disorders of skin and subcutaneous tissue in diseases classified elsewhere	6,144
E34	Other endocrine disorders	5,351
E74	Other disorders of carbohydrate metabolism	3,497
E07	Other disorders of thyroid	3,317

* not yet de-duplicated
** de-duplicated

Table 1: Baseline Demographics and Characteristics

	HV PS34 n=29,506	HV PS14 n=13,085
Age		
Median (25-75 percentile)	64 (55-72)	73 (69-78)
Gender		
Female	99.3%	98.8%
Other/Unknown	0.7%	1.2%
Endometrial Cancer (C54)	19.6%	38.3%
Ovarian Cancer (C56)	17.9%	59.4%

Table 2: Thrombocytopenia: Top Predictors, XGBoost

Importance	Code	Feature
0.103	D18	Hemangioma and lymphangioma, any site
0.047	Z86	Personal history of certain other diseases
0.018	Z84	Family history of other conditions
0.018	T39	Poisoning by, adverse effect of, and underdosing of nonopioid analgesics, antipyretics, and antirheumatics
0.017	N99	Intraoperative and postprocedural complications and disorders of genitourinary system, not elsewhere classified
0.015	T48	Poisoning by, adverse effect of, and underdosing of agents primarily acting on smooth and skeletal muscles and the respiratory system
0.014	Z85	Personal history of malignant neoplasm
0.012	T78	Adverse effects, not elsewhere classified
0.012	C49	Malignant neoplasm of other connective and soft tissue
0.010		Dexamethasone
0.009	D64	Other anemias

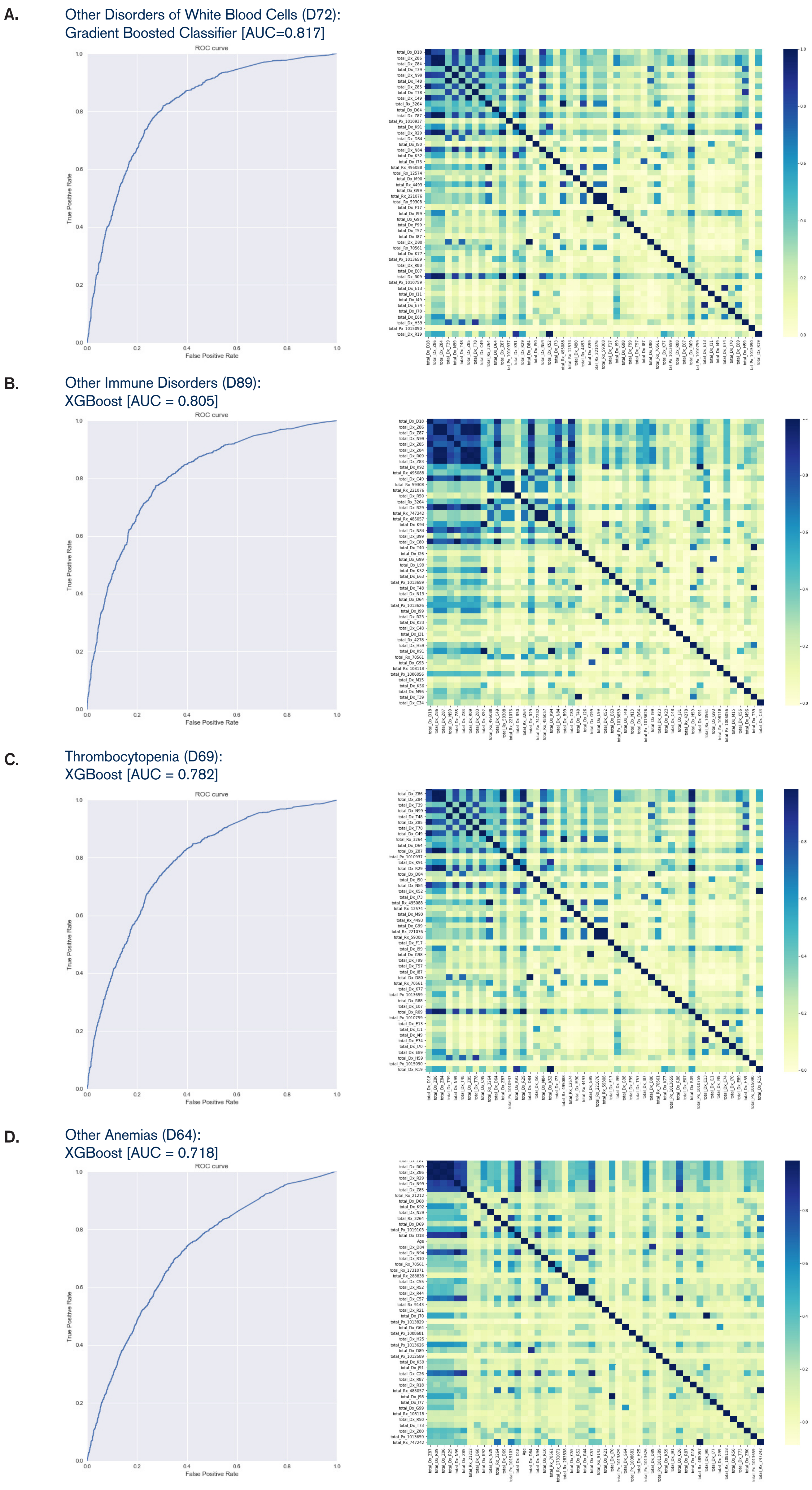
Table 3: Top Performing Target Condition Models by AUC

ICD-10 Code	Target Condition	Best Performing Model	Test AUC
T78	Adverse effects, not elsewhere classified	XGBoost	0.834
D72	Other disorders of white blood cells	Gradient Boosted Classifier	0.817
D89	Other disorders involving the immune mechanism	XGBoost	0.805
D69	Purpura and other hemorrhagic conditions	XGBoost	0.782
D64	Other anemias	XGBoost	0.718
K59	Other functional intestinal disorders	Gradient Boosted Classifier	0.713

Table 4: Top Predictive Features Across All 20 Target Groups

Code	Predictive Feature
	Age at Index Date
Z86	Personal history of certain other diseases
Z85	Personal history of malignant neoplasm
Z84	Family history of other conditions
Z81	Family history of mental and behavioral disorders
Z80	Family history of primary malignant neoplasm
T39	Poisoning by, adverse effect of, and underdosing of nonopioid analgesics, antipyretics, and antirheumatics
S23	Dislocation and sprain of joints and ligaments of thorax
R68	Other general symptoms and signs
R58	Hemorrhage, not elsewhere classified
D75	Other and unspecified diseases of blood and blood-forming organs
R44	Other symptoms and signs involving general sensations and perceptions
R39	Other and unspecified symptoms and signs involving the genitourinary system
R29	Other symptoms and signs involving the nervous and musculoskeletal systems
R18	Ascites
R10	Abdominal and pelvic pain
R09	Other symptoms and signs involving the circulatory and respiratory system
R07	Pain in throat and chest
N99	Intraoperative and postprocedural complications and disorders of genitourinary system, not elsewhere classified
N94	Pain and other conditions associated with female genital organs and menstrual cycle

Figure 2: Correlation Matrices and ROC Curves for Selected Top Performing Condition Models



A. Other Disorders of White Blood Cells (D72): Gradient Boosted Classifier [AUC=0.817]

B. Other Immune Disorders (D89): XGBoost [AUC = 0.805]

C. Thrombocytopenia (D69): XGBoost [AUC = 0.782]

D. Other Anemias (D64): XGBoost [AUC = 0.718]

Summary

- Among treated patients, comorbidity and adverse event rates were consistent with previous reports.[4-6]
- Use of common vocabularies and data modeling allowed for an expansion of modeling across both claims data sources with consistent cohort and covariate definitions. This scaling allowed for simultaneous examination of multiple conditions.
- For thrombocytopenia prediction, the top performing model was XGBoost, with a holdout test AUC of 0.782. Top features included: hemangioma/lymphangioma, other anemias, dexamethasone exposure, and personal history of malignant neoplasm (Table 2).
- Modeling expansion identified top performing models in other adverse effects often associated with chemotherapy treatment, such as white blood cell disorders, autoimmune disorders, and anemias, with AUCs ranging from 0.72-0.83 (Table 3, Figure 2).
    - Negative control conditions not directly associated with oncology had AUC scores below the 0.7 threshold (data not shown).
    - Positive control conditions had AUC scores above the threshold (data not shown).
- The highest AUC scores were seen with Gradient Boosted Classifier and XGBoost.
- Visual inspection of correlation matrices, unsurprisingly, showed some degree of multi-collinearity between hematologic conditions (Figure 2).
- When top features were assessed, 20 predictors were identified that were common across all 20 conditions evaluated.

Limitations

- Data was derived from institutional claims, limiting assessment of pharmacologic treatment to those administered within a health institution, such as injectable or infusible agents. Given the focus of this study on chemotherapeutic agents, this is not anticipated to be a significant limitation.
- Expanding pharmacologic coverage in future analyses, however, will allow for inclusion of emerging oral therapies, such as PARP-inhibitors.
- Given the severity of the conditions in the datasets and use of cytotoxic agents, high AUC scores for conditions pertaining to adverse events are expected. This approach, however, is being developed as a way to find, at scale, associations between medical events and common drivers across them. Future implementations can incorporate enhancements with respect to feature engineering (time-based predictors, grouping to reduce multi-collinearity) or propensity matching techniques to allow for explorations of possible causality.

Conclusions

- Through common data modeling and commonly available machine learning packages, we demonstrate here a predictive modeling approach to identify factors associated with thrombocytopenia, an often treatment-limiting side-effect in oncology populations. This approach is then scalably deployed across multiple comorbidities. This has positive implications for patient care by not only facilitating identification of potential factors preceding adverse events, but also identifying common predictors across multiple adverse events in this high-risk population.

References

1. Lipitz-Snyderman A, et al. Cancer. 2017;123(23):4728-4736.
2. Kuter, DJ. Oncology (Williston Park). 2015; 28 (4):282-94.I
3. US Food and Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SAMD). 2019. (https://www.fda.gov/media/122535/download)
4. Cassidy CA, et al. Anti-Cancer Drugs. 2001, 12(4): 383-385.
5. Ten Berg MJ et al. Drug Safety. 2001, 34(12):1151-60.
6. Mahner S. Eur J Cancer. 2015;51(3):352-8.
7. HealthVerity-Overview. https://healthverity.com/wp-content/uploads/HealthVerity-Overview.pdf Accessed October 18, 2019.
8. https://github.com/scikit-learn/scikit-learn
9. https://github.com/dmlc/xgboost/tree/master/python-package
10. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
11. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

Disclosures

- JR. RB, AG, CM, AS, and BL are employees of Medidata Solutions. MB is a consultant to Medidata Solutions.

QR CODE

Presented at ISPOR Europe 2019, 2-6 November 2019
Copenhagen, Denmark