PT13

Claims-Based Machine Learning Algorithms to Predict ECOG Performance Status (ECOG) for Pan-Cancer (Prostate, Breast, Colorectal, Gastric, Non-small cell lung (NSCLC) and Pancreatic Cancer) Patients.

Authors: Jayati Saha[1], Nicole Zhang[1], Amar Das[1], Jiemin Liao[1]

Affiliations: [1]Guardant Health, Redwood City, CA

GUARDANTINFORM™

Introduction

- ECOG is used extensively in oncology to assess progression and determine treatment and prognosis of patients.
- It is primarily a clinical measurement and therefore claims databases do not capture it.
- Some studies have utilized electronic health records and cancer registry information and applied machine learning techniques to predict ECOG.[1,2]
- Studies that focused on claims databases relied mostly on logistic regression models to predict ECOG.[3,4]
- In this study, we used a claims database of cancer patients to assess the utility of automated machine learning (AutoML) models in enhancing prediction accuracy.

Methods

- **Data Source:** Patients were identified from the claims-based clinical-genomic database GuardantINFORM, which links cell-free circulating tumor DNA (cfDNA) results to de-identified claims data, with study time period from June 2014 to December 2022.
- **Inclusion and exclusion criteria:**
  - Adult patients in the US with prostate, breast, colorectal, gastric, NSCLC or pancreatic cancer diagnosis indicated on their Guardant360 test requisition form, and at least two medical claims with corresponding cancer diagnosis
  - Has a valid ECOG or Karnofsky performance scores extracted from pathology reports
  - Has at least two medical claims within 6 months prior to ECOG test result (baseline period)
  - No multiple primary cancers
- **ECOG status:** Dichotomized to 0-1 (good) vs 2+ (poor). Karnofsky performance scores were converted to ECOG equivalent scores using a crosswalk.[3]
- The process of utilizing AutoML for prediction of ECOG status is summarized in **Figure 1**
- Performance was measured by area under the precision-recall curve (AUPRC) due to imbalanced data
- 85 variables were selected based on previous literature that used claims data to predict ECOG status. [3]

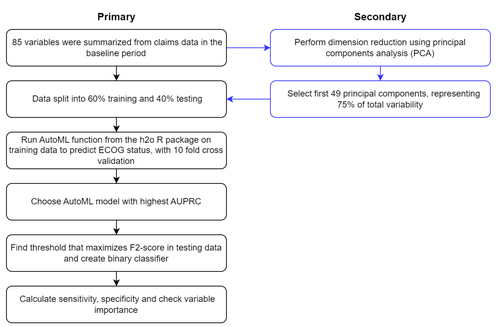


Figure 1. Flowchart summary of methodology

Inclusion Criteria	Prostate Cancer	Breast Cancer	Colorectal Cancer	Gastric	NSCLC	Pancreatic Cancer	Total
Patients with cancer on TRF	20,137	31,466	24,149	3,899	105,146	15,309	200,106
At least 2 claims recorded for the primary cancer diagnosis	18,367	28,987	22,005	3,192	88,124	13,207	173,882
Has a valid ECOG value from pathology reports	846	1825	1,295	144	4,919	751	9,780
Has at least 2 claims within 6 months prior to ECOG test result	699	1,585	1,157	136	4,462	678	8,717
No multiple primary cancers	697	1,579	1,154	136	4,451	674	8,691
Valid ECOG mapping to binary value	696	1,575	1,152	136	4,444	671	8,674

Table 1. Patient attrition table. 8,674 pan-cancer patients were identified from the real-world database.

Results

Demo/Clinical Characteristics	N (%)	Frailty Indicators	N (%)	Healthcare Utilization	N (%)
Age (mean years, SD)	66.5 (11.3)	Arthritis	1,951 (22%)	No. of drug dispensings (mean, SD)	6 (7.2)
Male	3,855 (44%)	Dementia	349 (4%)	No. of OP visits (mean, SD)	12 (11.6)
Region		Difficulty walking	567 (7%)	No. of ER visits (mean, SD)	0.4 (1.0)
Northeast	881 (10%)	Fall	359 (4%)	No. of IP stays (mean, SD)	0.5 (1.0)
Midwest	2,293 (26%)	Lipid abnormality	2,902 (33%)	No. of IP days (mean, SD)	2.6 (7.2)
South	3,265 (38%)	Nursing or personal care services	46 (1%)	No. of E&M claims (mean, SD)	3.9 (4.5)
West	1,792 (21%)	Parkinson's disease	34 (0%)	Home visit	2,024 (23%)
Unknown	443 (5%)	Podiatric care	248 (3%)	Nursing home visit	183 (2%)
Payer type		Psychiatric diagnoses	1,604 (18%)	Hospice visit	452 (5%)
Government	4,614 (53%)	Rehabilitation services	146 (2%)	Restructured BETOS (RBCS)	
Commercial	6,312 (73%)	Sepsis	849 (10%)	Screening	458 (5%)
Charlson Comorbidity Score	1.4 (1.8)	Skin ulcer, pressure ulcer	158 (2%)	Oxygen supply	514 (6%)
COPD	2,444 (28%)	Stroke	753 (9%)	Wheelchair use	72 (1%)
Paralysis	72 (1%)	Vertigo	507 (6%)	Nuclear Imaging	3,019 (35%)
Diabetes Complicated	643 (7%)	Weakness, muscular wasting	1,935 (22%)	Ambulance services	891 (10%)

Table 2. Patient demographic and clinical characteristics. Among 8,674 pan-cancer patients, 1,375 (16%) of patients had poor ECOG score. Individual Charlson score comorbidities were also included. Restructured BETOS was used to group healthcare service codes into the categories of interest. Additional RBCS procedures were also included (not shown).

Rank	Primary Analysis (85 Variables) Model	AUPRC	AUC	PCA Analysis (49 variables) Model	AUPRC	AUC
1	StackedEnsemble_BestOfFamily_1	0.370	0.725	StackedEnsemble_AllModels_1	0.353	0.709
2	GBM_1	0.367	0.717	StackedEnsemble_BestOfFamily_1	0.352	0.711
3	StackedEnsemble_AllModels_1	0.365	0.723	GLM_1	0.350	0.705
4	GLM_1	0.364	0.716	GBM_1	0.340	0.694
5	XGBoost_grid_1_model_1	0.351	0.715	XGBoost_grid_1_model_1	0.337	0.700
6	GBM_2	0.349	0.707	DeepLearning_grid_3_model_1	0.325	0.697
7	GBM_grid_1_model_1	0.344	0.704	XRT_1	0.322	0.689
8	GBM_grid_1_model_2	0.342	0.699	DRF_1	0.320	0.689
9	XGBoost_grid_1_model_2	0.340	0.705	GBM_2	0.319	0.689
10	GBM_5	0.340	0.714	GBM_4	0.317	0.680

Table 3. Output of AML models using 85 variables and 49 principal components. Models were ranked according to AUPRC. The best model generated by AutoML were both stacked ensemble models for primary and PCA analysis.

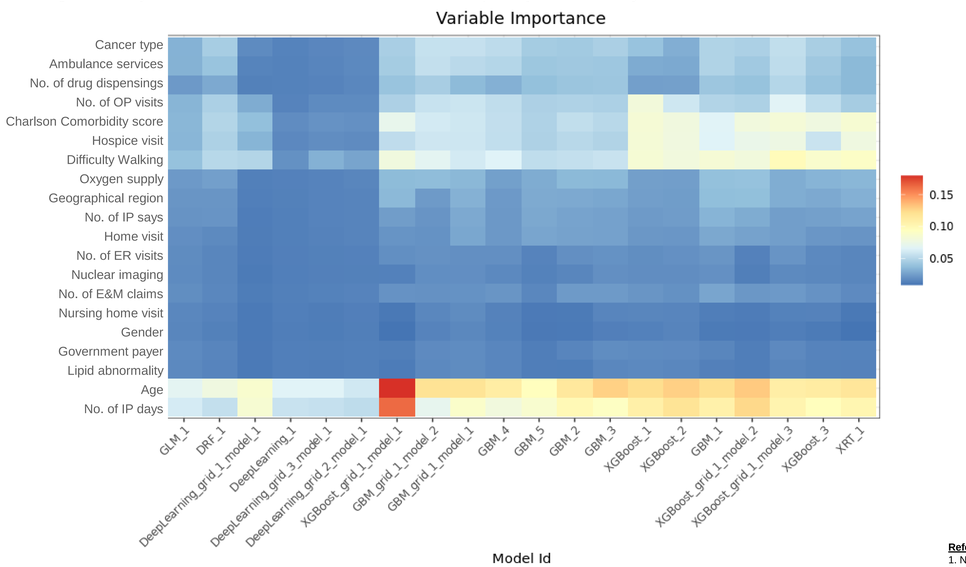
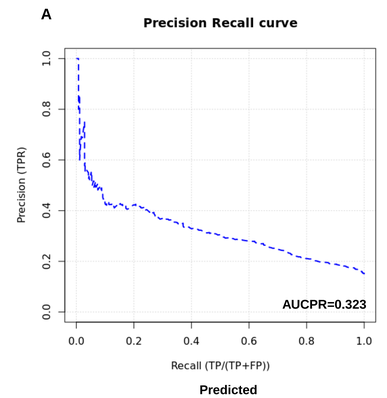
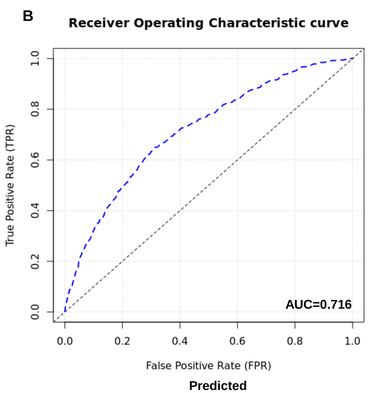


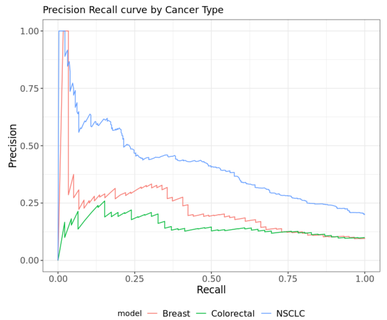
Figure 2. Variable importance plot for predicting ECOG status across top 20 models in primary analysis (excluding deep learning models). Variables that were highly influential in model prediction included age and no. of hospitalization days. Other variables such as no. of OP visits, Charlson Comorbidity Score, hospice visit, and difficulty walking had some influence in the about half of the top 20 models.



Actual \ Predicted	Low	High
Low	1,747	1,194
High	143	380



Actual \ Predicted	Low	High
Low	2,104	837
High	199	324



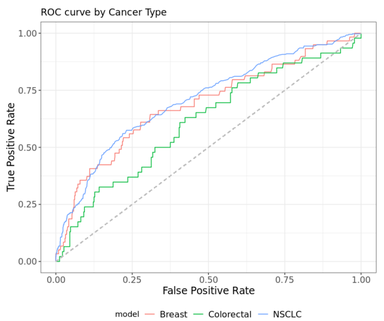


Figure 3A. Precision-recall curve with optimal threshold at 0.128 for maximum F2 score, sensitivity=0.73, specificity=0.59; **Figure 3B.** ROC curve with optimal threshold at 0.157 for maximum F1 score, sensitivity=0.62, specificity=0.72; **Figure 3C, D.** Precision-recall curves and ROC curves by breast cancer, colorectal cancer and NSCLC. NSCLC has the best performance, likely due to its large sample size compared to other cancer types. Optimal cut-off points based on F2 score ranged from 0.06 to 0.14

Conclusions

- We demonstrated that using the AutoML framework, one can easily search through multiple machine learning models and select the best one to predict ECOG score. This framework can be extended to other claims databases with some linked ECOG status from medical records.
- Additional analyses are needed to identify whether the models can be further optimized outside of the AutoML framework.
- Performance of the models is likely driven by sample size, which should be taken into account when developing the framework for other databases.

References

1. Nguyen, H. V. & Byeon, H. Prediction of ECOG Performance Status of Lung Cancer Patients Using LIME-Based Machine Learning. Mathematics 11, 2354 (2023).
2. Graham, S. et al. Machine Learning Approach to Estimating ECOG PS for a Multiple-Myeloma Cohort from Real World Data. Blood 142, 4700–4700 (2023).
3. Sheffield, K. M. et al. Development and validation of a claims-based approach to proxy ECOG performance status across ten tumor groups. J. Comp. Eff. Res. 7, 193–208 (2018).
4. Salloum, R. G., Smith, T. J., Jensen, G. A. & Lafata, J. E. Using claims-based measures to predict performance status score in patients with lung cancer. Cancer 117, 1038–1048 (2011).

Questions can be directed to Jayati Saha. jsaha@guardanthealth.com