Machine Learning vs Traditional Statistics: Developing a Novel Proxy for HPV-associated LA SCCHN

Olivia Shane, Robert Schuldt, Achal Patel, David Fox, Dustin Schrader, Rashed Harun

Genentech Inc, South San Francisco, CA, USA

Please scan using your QR reader application to access this poster on your mobile device. NB: there may be associated costs for downloading data. These costs may be high if you are using your smartphone abroad. Please check your mobile data tariff or contact your service provider for more details. Alternatively this can be accessed at: https://ter.li/nxxkhs

Background

- HPV is a strong prognostic factor for survival in LA SCCHN and valuable for risk stratification
- Existing proxies for predicting HPV status in LA SCCHN are inaccurate and unreliable[1,2]

Study Objectives

- To identify a better performing proxy for HPV status
- Compare the performance of traditional statistical vs machine learning methods

Methods

- Database: SEER Incidence Database - Head and Neck with HPV Status Database (2010-2017) with Census Tract-level SES/Rurality
- Timeframe: 2010-2017
- Inclusion Criteria: 1) LA (locally advanced) (AJCC stage III-IVB), 2) Oropharynx squamous cell carcinoma, 3) Age 18 or older
- Exclusion Criteria: 1) Missing HPV status, 2) Missing data on covariates
- Predictor variables: All available pre-treatment patient characteristics:
  - Sex, Age, Race/ethnicity, national Yost socio-economic status index quintiles, marriage status, urbanicity, year of diagnosis, tumor involvement, node involvement,
- Data Split: Training (80%), Test (20%)
- Models: logistic regression, stepwise logistic regression, LASSO, elastic net, stepwise elastic net, random forest, GBM, XGBoost
- Hyperparameter tuning was performed for ML models
- Base case thresholds for ML models were chosen by maximizing sensitivity & specificity, scenario analysis thresholds were chosen by minimizing cost, where a false positive was 4x more costly than false negative

Table 1. Patient attrition

Description of step	n	% of previous step retained
Patients in SEER Research Plus, Head and Neck with HPV Status and Census Tract-level SES/Rurality Combined Database (2006-2018)	5,531,627	100%
Patients diagnosed from 2010-2017	3,464,655	62.6%
Patients in SEER Research Plus, 21 registries, Nov 2020 Sub (2000-2018) with malignant tumors and known age	107,936	3.1%
With TNM staging available from the derived AJCC, 7th edition (2010-2015) or derived SEER combined stage (2016-2017)	105,019	97.3%
Patients with tumors in oral cavity, oropharynx, larynx, or hypopharynx sites	94,923	90.4%
Patients with tumors only in Oropharynx site	36,128	38.1%
Patient has HPV Positive or HPV Negative status	19,489	53.9%
Patients with AJCC7 staging III or IVA/B (HPV- or HPV+) SCCHN involving oropharynx	15,115	77.6%
Patients 18 years or older	15,113	99.9%
Patients without missing data for all explanatory variables (race/origin, SES, marital status, urban/rural status)	13,645	90.3%

Results

Table 2. Patient Characteristics

Characteristics	N = 13,645
Female n (%)	2,032 (14.9%)
Age in years: mean (SD)	60.96 (9.90)
Race n (%)	
•Non-Hisp American Indian/ Alaska Native	70 (0.5%)
•Non-Hisp Asian/Pacific Islander	377 (2.8%)
•Non-Hisp Black	1,028 (7.5%)
•Non-Hisp White	11,280 (82.7%)
•Hispanic (all races)	890 (6.5%)
SES n (%)	
•Group 1 (lowest SES)	1,796 (13.2%)
•Group 2	2,037 (14.9%)
•Group 3	2,373 (17.4%)
•Group 4	3,136 (23.0%)
•Group 5 (highest SES)	4,303 (31.5%)
Urban vs Rural n (%)	
•All Urban	8,689 (63.7%)
•All Rural	893 (6.5%)
•Mostly Urban	3,052 (22.4%)
•Mostly Rural	1,011 (7.4%)
Year of Diagnosis n (%)	
•2010	621 (4.6%)
•2011	972 (7.1%)
•2012	1,309 (9.6%)
•2013	1,668 (12.2%)
•2014	1,926 (14.1%)
•2015	2,104 (15.4%)
•2016	2,460 (18.0%)
•2017	2,585 (18.9%)
Tumor Size n (%)	
•T0	69 (0.5%)
•T1	3,529 (25.9%)
•T2	4,900 (35.9%)
•T3	2,734 (20.0%)
•T4	2,348 (17.2%)
•TX	65 (0.5%)

Characteristics	N = 13,645
Marriage Status n (%)	
•Married (including common law)	8,408 (61.6%)
•Divorced	1,834 (13.4%)
•Separated	170 (1.2%)
•Single (never married)	2,485 (18.2%)
•Unmarried or domestic partner	98 (0.7%)
•Widowed	650 (4.8%)
Urban vs Rural n (%)	
•All Urban	8,689 (63.7%)
•All Rural	893 (6.5%)
•Mostly Urban	3,052 (22.4%)
•Mostly Rural	1,011 (7.4%)
Node Score n (%)	
•N0	900 (6.6%)
•N1	2,552 (18.7%)
•N2	9,506 (69.7%)
•N3	680 (5.0%)
•NX	7 (0.1%)
Metastasis Score n (%)	
•M0	13,645 (100%)
SES n (%)	
•Group 1 (lowest SES)	1,796 (13.2%)
•Group 2	2,037 (14.9%)
•Group 3	2,373 (17.4%)
•Group 4	3,136 (23.0%)
•Group 5 (highest SES)	4,303 (31.5%)
Marriage Status n (%)	
•Married (including common law)	8,408 (61.6%)
•Divorced	1,834 (13.4%)
•Separated	170 (1.2%)
•Single (never married)	2,485 (18.2%)
•Unmarried or domestic partner	98 (0.7%)
•Widowed	650 (4.8%)

Table 3. Performance of predictive models for HPV-associated LA SCCHN

Threshold chosen by maximizing sensitivity and specificity (F1 score)

	Threshold	Sensitivity	Specificity	NPV	PPV	Test AUC †
Proxy 1 (All HPV+)	NA	1.000	Undefined	Undefined	1.000	0.000
Proxy 2 (Oropharynx site, young (<65), white, male)	NA	0.507	0.643	0.300	0.812	0.556
Logistic Regression	0.76	0.690	0.637	0.403	0.853	0.719
Stepwise Logistic Regression	0.78	0.628	0.679	0.375	0.856	0.720
Lasso*	0.76	0.683	0.636	0.397	0.851	0.720
Elasticnet*	0.76	0.682	0.637	0.396	0.851	0.721
Stepwise Elasticnet*	0.78	0.620	0.688	0.373	0.858	0.721
Random Forest*	0.78	0.597	0.709	0.366	0.862	0.718
GBM*	0.78	0.628	0.690	0.378	0.860	0.722
XGBoost*	0.77	0.643	0.675	0.383	0.858	0.723

*: Hyperparameter tuning for these models were performed using 5-fold cross-validation

†: Test AUC was calculated using test dataset, separate from the training dataset

AUC = Area under ROC curve (measures model's ability to discriminate between two classes). For example, AUC of .72 means 72% of HPV+ patients have correctly higher P(HPV+)

Figure 1. Impact of variables on predicted HPV status (XGBoost Model)

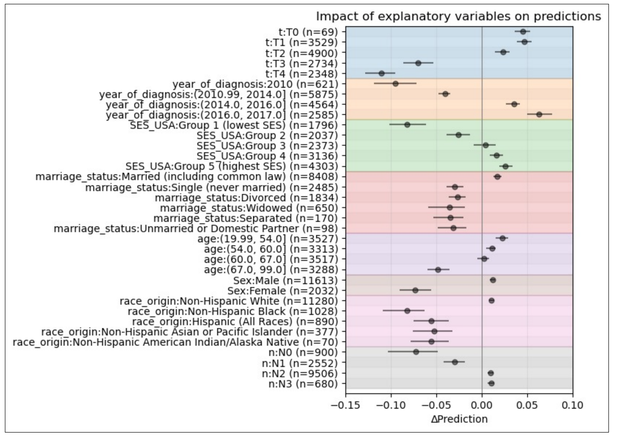
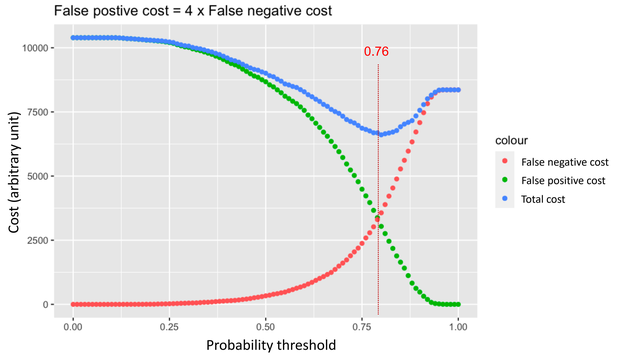


Figure 2. Threshold selection for scenario analysis of "cost" minimization



Note: Y-axis cost units are arbitrary, as the costs of FN/FP would yield identical results as long as FP = 4x FN costs. Threshold was chosen using the probability at which "cost" was lowest.

Table 4. Performance of predictive models for HPV-associated LA SCCHN

(Scenario analysis using threshold chosen by minimizing "cost")

	Threshold	Sensitivity	Specificity	NPV	PPV	Test AUC †	Relative Cost
Proxy 1 (All HPV+)	NA	1.000	Undefined	Undefined	1.000	0.000	130.7%
Proxy 2 (Oropharynx site, young (<65), white, male)	NA	0.507	0.643	0.300	0.812	0.556	100%
Logistic Regression	0.82	0.690	0.637	0.403	0.853	0.719	81.8%
Stepwise Logistic Regression	0.80	0.628	0.679	0.375	0.856	0.720	82.5%
Lasso*	0.79	0.683	0.636	0.397	0.851	0.720	82.7%
Elasticnet*	0.78	0.682	0.637	0.396	0.851	0.721	82.6%
Stepwise Elasticnet*	0.80	0.620	0.688	0.373	0.858	0.721	82.0%
Random Forest*	0.78	0.597	0.709	0.366	0.862	0.718	81.6%
GBM*	0.80	0.628	0.690	0.378	0.860	0.722	81.0%
XGBoost*	0.79	0.591	0.715	0.365	0.863	0.723	81.4%

Discussion and limitations

- We were unable to take full advantage of our ML methods due to low number of covariates
- Missing important variables such as smoking & alcohol history, geographic info, HPV risk-factors
- No data if patient receives care in non-SEER region (reducing generalizability)
- Exclusion of patients with missing data could have biased results

Conclusions and recommendations

- Both ML models and logistic regression-based methods outperformed existing proxy methods for identifying HPV associated LA SCCHN
- ML performed similarly to logistic regression in this limited dataset
- ML may further outperform traditional statistics in datasets with larger number of covariates & patients as ML uniquely characterizes non-linear relationships between variables
- Learnings from the methods & interpretation of this analysis can be applied in future predictive models & HEOR analyses

Abbreviations

AJCC: American Joint Committee on Cancer; AUC: Area under the curve; GBM: Gradient-boosting machine; HPV: Human papillomavirus; LA: locally advanced; LASSO: Least Absolute Shrinkage and Selection Operator; ML: Machine learning; NPV: Negative predictive value; PPV: Positive predictive value; SCCHN: Squamous cell carcinoma of the head and neck; SEER: Surveillance, Epidemiology, and End Results; SES: Socio-economic status; SHAP: SHapley Additive exPlanations; XGBoost: Extreme Gradient Boosting

References

1. Oncologist. 2022 Feb 3;27(1):48-56. doi: 10.1093/oncolo/oyab001.
2. Cancer. 2019;125(2):249-260. doi:10.1002/cncr.31800

Acknowledgments/ Disclosures

This study was funded by Genentech, Inc.

This analysis used data from the SEER Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Plus Data (Specialized Head and Neck Fields with Census Tract SES/Rurality), 18 Registries (excl AK), National Cancer Institute, DCCPS, Surveillance Research Program, based on the November 2020 submission.

All authors were employees of Genentech Inc at the time of this analysis. RS, AP, DF, DS, and RH are shareholders of F. Hoffmann-La Roche.