Machine Learning Models in Prediction of Head and Neck Squamous Cell Carcinoma Survivability

Rohatgi O[1], Agrawal N[1], Goswami S, [1] Vivek V[1], Chaudhuri M[1], Aparasu RR[2]

[1]Complete HEOR Solutions (CHEORS), PA, 18914, USA;
[2]University of Houston College of Pharmacy, Houston, TX

MSR75

KEY POINTS

Machine learning models were found to have improved efficiency over the logistic regression model, in prediction of Head and Neck Squamous Cell Carcinoma Survivability

BACKGROUND

- Head and neck squamous cell carcinoma (HNSCC) is the sixth most commonly malignancy develop in oral cavity, pharynx, and larynx[1]
- Predicting the survival of a patient is one of the most challenging tasks for a physician
- Machine Learning (ML) is a subfield of artificial intelligence, that learns and discovers complex data patterns.
- Decision trees (DT), Random Forest (RF), and Support vector machines (SVM) are commonly used ML algorithms.
- Although ML algorithms are often used in healthcare, there is limited competitive data to model the survivability among patients with HNSCC.

The 5-year overall survival rate of patients with HNSCC is 40–50%[2]

OBJECTIVE

To evaluate the performance of common ML models in predicting the 5-year overall survival of patients with HNSCC

METHOD

Data Source:
Surveillance, Epidemiology, and End Results (SEER) 17 registries database

Study Population:
Patients of age ≥20 years with HNSCC in Larynx, Nasopharynx, Oral Cavity, Oropharyngeal, and Sinonasal Cavity (identified using ICD-O3 and histology type codes) were selected from SEER database during 2008-19

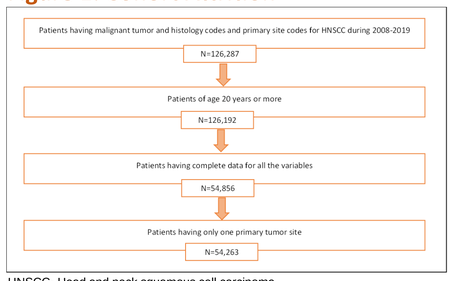
Outcome:
5-year overall survivability among patients with HNSCC

Statistical Analysis:
- Logistic regression (LR), DT, RF, and SVM models were developed
- Dataset was split into training and validation set in 70%-30% ratio. The models were validated on the hold out sample.
- Accuracy, AUROC curve, and F1 score were used to evaluate the model performance

RESULTS

- The study included 54,263 patients with HNSCC. Most of the selected patients were White (84.3%), male (74.0%), and of age ≥65 years (44.9%) (Table 1)

Figure 1: Cohort Attrition



HNSCC- Head and neck squamous cell carcinoma

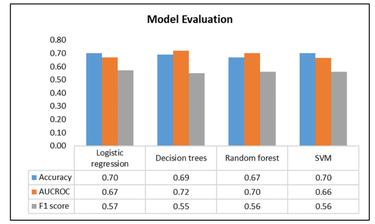
- For majority of selected patients, primary site of tumor was oropharyngeal (39.4%), followed by oral cavity (33.1%), larynx (17.6%), sinonasal cavity (7.7%), and nasopharynx (2.2%) (Table 1)
- Based on AU-ROC score, DT (0.72) showcased the best distinguishing capability, followed by RF (0.70), LR (0.67), and SVM (0.66) models (Figure 3)
- However, LR (0.70) and SVM (0.70) had higher accuracy followed by DT (0.69) and RF (0.67) models (Figure 2)
- The Average F1 score was 0.57 and was similar for all the models (Figure 2)

Table 1: Patient Characteristics

Demographic Characteristics (N=54,263)	
Age groups N (%)	
20-44 years	2,364 (4.4)
45-54 years	9,550 (17.6)
55-64 years	17,958 (33.1)
65+ years	24,391 (44.9)
Sex N (%)	
Female	14,090 (26.0)
Male	40,173 (74.0)
Year of diagnosis N (%)	
2008-11	18,061 (33.3%)
2012-15	20,821 (38.4%)
2016-19	15,381 (28.3%)
Race N (%)	
Black	4,924 (9.1)
White	45,748 (84.3)
Other	3,591 (6.6)
Marital status N (%)	
Married/ domestic partner	30,442 (56.1)
Unmarried	23,821 (43.9)
Income N (%)	
<$55,000	15,063 (27.8)
$55,000-$75,000	23,867 (44.0)
$75,000+	15,333 (28.3)
Urbanicity N (%)	
Metro	46,207 (85.2)
Non-metro	8,056 (14.8)

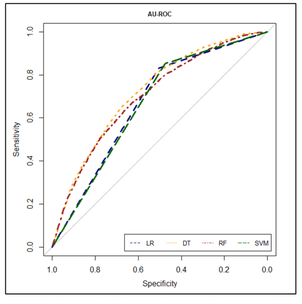
Clinical Characteristics (N=54,263)	
Stage N (%)	
Distant	10,545 (19.4)
Localized	15,752 (29.0)
Regional	27,703 (51.1)
Unknown/unstaged/In situ	263 (0.5)
Radiation N (%)	
Yes	36,958 (68.1)
None/unknown/refused	17,305 (31.9)
Chemotherapy N (%)	
Yes	24,639 (45.4)
No/Unknown	29,624 (54.6)
Grade N (%)	
Moderately differentiated	27,076 (49.9)
Well differentiated	7,314 (13.5)
Poorly differentiated/ Undifferentiated	19,873 (36.6)
Surgery at primary site N (%)	
Yes	32,882 (60.6)
No	21,381 (39.4)
Primary tumor site N (%)	
Larynx	9,525 (17.6)
Nasopharynx	1,168 (2.2)
Oral Cavity	17,981 (33.1)
Oropharyngeal	21,401 (39.4)
Sinonasal Cavity	4,188 (7.7)
Tumor size (in cm) N (%)	
<1	5,064 (9.3)
1-1.9	10,874 (20.0)
2-3.9	23,970 (44.2)
>4	14,355 (26.5)
5-year survival N (%)	
1 (Alive)	32,717 (60.3)
0 (death)	21,546 (39.7)

Figure 2: Model Evaluation Metrics



AUROC- Area under the receiver operating curve; SVM- Support Vector Machine

Figure 3: AUROC Score



AUROC- Area under the receiver operating curve; SVM- Support Vector Machine; LR- Logistic Regression; DT- Decision Trees; RF- Random Forest

CONCLUSIONS

- The study found all ML models were in the acceptable range with DT exhibiting best performance based on the AUROC score.
- This study provides a step forward to utilize ML-based methods to predict patients' chances of survival and can provide decision aid for cancer management
- However, these methods need to be validated externally to develop reliable and generalizable ML models.

LIMITATIONS

- Cases with multiple primary sites can impact survivability and were not considered in this analysis which might influence the model performance
- The SEER database has inherent limitations as it does not report environmental and behavioral risk factors.
- SEER database also lacks human papillomavirus (HPV) status-related information which can impact HNSCC survivability and affect the model performance

For any questions please email: swarnali.goswami@cheors.com

Contact us:

CONFLICT OF INTEREST

• No funding was received from any external organization
• The authors report no other conflicts of interest in this work

References

1. Johnson DE, Burtness B, Leemans CR, Lui VW, Bauman JE, Grandis JR. , "Head and neck squamous cell carcinoma.," Nature reviews Disease primers, vol. 6, no. 1, pp. 1-22, 2020.
2. Lu, L., Wu, Y., Feng, M., Xue, X., Fan, "A novel seven miRNA prognostic model to predict overall survival in head and neck squamous cell carcinoma patients," Molecular Medicine Reports, vol. 20, no. 5, pp. 4340-48, 2019.
3. Du E, Mazul AL, Farquhar D, Brennan P, Anantharaman D, Abedi-Ardekani B, Weissler MC, Hayes DN, Olshan AF, Zevallos JP., "Long-term survival in head and neck cancer: impact of site, stage, smoking, and human papillomavirus status.," The Laryngoscope, vol. 129, no. 11, pp. 2506-13, 2019.
4. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Data, 17 Registries, Nov 2021 Sub (1975-2019) - Linked To County Attributes - Time Dependent (1990-2019) Income/Rurality, 1969-2020 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2022, based on the November 2021 submission.