scientific reports

Check for updates

OPEN

Machine learning for survival outcome in head and neck squamous cell carcinoma: a multicenter validation study

Rasheed Omobolaji Alabi[1,2]✉, Orlando Guntinas-Lichius[3], Mohammed Elmusrati[2,6], Alhadi Almangush[1,4,5,6], Ylva Tiblom Ehrsson[7], Göran Laurell[7] & Antti A. Mäkitie[1,8,9]

Most head and neck squamous cell carcinoma (HNSCC) cases are diagnosed late, with an increased risk of recurrence and distant metastasis. In recent years, there has been a surge in the development of prognostic and predictive machine learning (ML) models for personalized treatment planning. However, only a small number of these have been externally validated. This study aimed to build a prognostic system by combining clinicopathological parameters and treatment-related factors as integrative inputs to build a machine learning (ML) model using data from the Surveillance, Epidemiology, and End Results (SEER, United States) program. We further validated the developed model using multicenter data obtained from the Thuringian Cancer Registry (Germany) and a multicenter prospective observational study obtained from the Uppsala University Hospital (Sweden) to estimate the overall survival (OS) of patients with HNSCC. Additionally, we explored the complementary prognostic potentials of these input parameters using permutation feature importance (PFI). A total of 40,164 patients with HNSCC were recruited from the SEER database and validated with 3950 cases obtained from the Thuringian Cancer Registry and 323 cases recruited from three University Hospitals in Sweden. We evaluated the prognostic significance of the input variables to predict OS in patients with HNSCC using permutation feature importance. The voting ensemble ML algorithm gave an area under receiving operating characteristics curve (AUC) of 0.76 and an accuracy of 70.0%. Independent external validation of the validation model with data from the Thuringian Cancer Registry and the Uppsala University Hospital gave AUCs of 0.68 and 0.76, with decreased performance accuracy in both cohorts. The PFI analysis of the base model showed that age at diagnosis, T stage, tumor site, marital status, and surgical treatment were the most important parameters for the predictive ability of the model for OS. External independent geographic validation is important for performance reproducibility and model generalization before recommending the model for further clinical evaluation. External independent geographic validation may not necessarily increase the performance accuracy. However, it can reveal and demonstrate the performance of the model outside the development data. A generalized ML can lead to individualized risk-based therapeutic decision-making. While independently validating the model may be possible during model development, data privacy and security-related issues may prevent including it as a prerequisite in the ML model development pipeline.

Keywords Machine learning, Head and neck squamous cell carcinoma (HNSCC), Overall survival, External validation, Validation study

[1]Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [2]Department of Industrial Digitalization, School of Technology and Innovations, University of Vaasa, Vaasa, Finland. [3]Department of Otorhinolaryngology, Jena University Hospital, 07747 Jena, Germany. [4]Institute of Biomedicine, Pathology, University of Turku, Turku, Finland. [5]Department of Pathology, University of Helsinki, Helsinki, Finland. [6]Libyan Authority for Scientific Research, Ministry of Higher Education, Tripoli, Libya. [7]Department of Surgical Sciences, Section of Otorhinolaryngology and Head and Neck Surgery, Uppsala University, Uppsala, Sweden. [8]Division of Ear, Nose and Throat Diseases, Department of Clinical Sciences, Intervention and Technology, Karolinska Institute, Karolinska University Hospital, Stockholm, Sweden. [9]Department of Otorhinolaryngology—Head and Neck Surgery, University of Helsinki, Helsinki University Hospital, Helsinki, Finland. ✉email: rasheed.alabi@helsinki.fi

Head and neck cancers (HNCs) consist of a group of malignancies that differ in terms of their etiology, risk factors, histology, and therapeutic management[1]. Their sites include the oral cavity, pharynx, larynx, salivary glands, paranasal sinuses, and nasal cavity[2]. About 90% originate primarily in the squamous cells lining of the mucosal surfaces of the upper aerodigestive tract[3], and these are collectively referred to as head and neck squamous cell carcinomas (HNSCCs)[2]. Their etiology is multifactorial, with alcohol and tobacco use, human papillomavirus (HPV) infection, betel quid chewing, radiation exposure, and genetic mutations identified as major risk factors[3]. The annual HNC incidence has been estimated to be 54,000 new cases and about 11,230 deaths in 2022 in the United States[4]. It represents a major global health concern and is one of the top causes of cancer-related mortality with an estimated 350,000 annual deaths globally[5].

Remarkably, a significant number of HNSCCs are still diagnosed at an advanced stage[6,7]. Treatment of HNSCC consists of single-modality therapy, such as surgery or radiotherapy, or combined modality treatment, which may also include chemotherapy. Therefore, selecting an appropriate treatment for HNC patients becomes pertinent. One approach to selecting appropriate treatment is to stratify the patients into risk groups for personalized treatment planning and thus avoid unnecessary mutilating therapies[8]. Also, it provides a useful insight into effective management decision-making and may guide the selection of a protocol treatment approach[8]. Predicting HNSCC survival is challenging due to different patient-related factors, tumor characteristics, and available treatment modalities. Traditionally, the American Joint Committee on Cancer (AJCC) Tumor-Nodal-Metastasis (TNM) staging system has been shown to be an objective and accurate tool for risk stratification in oncology. However, the TNM staging scheme has been criticized in recent years due to its inability to consider other tumor- and patient-related risk factors. In addition, the TNM staging scheme does not consider cancer patients on an individual basis. Therefore, in the quest for personalized oncology, other risk stratification approaches are warranted.

In recent years, artificial intelligence (AI) and its subfields, such as machine learning and deep learning, have contributed to various applications in cancer management with promising results[9–12]. Several factors have still limited the implementation of ML models in daily clinical practice[13]. One such factor is external validation of the models for generalizability[14]. A key question is how well a ML model can perform when tested with independent geographic validation cohorts[11,14–16]. To this end, the objectives of this study were to: (i) build an ML predictive model to estimate the overall survival of HNSCC patients using the data obtained from a population-based Surveillance, Epidemiology, and End Results (SEER, United States) Program of the National Institutes of Health (NIH) by combining clinicopathologic parameters without treatment-related factors as integrative input sources into the ML algorithm for estimating the OS of patients with HNSCC. Having prior information about survival with only clinicopathologic parameters may provide an adequate individualized treatment approach for this patient population; (ii) analyse the complementary prognostic potentials of these clinicopathological parameters for their order of significance in predicting OS in HNSCC patients; and (iii) externally and independently validate the model using multicenter data obtained from [a] the Thuringian Cancer Registry (Germany) and [b] multicenter prospective observational study obtained from the Uppsala University Hospital (Sweden). The resulting externally validated model shows the extent of the generalizability of the model. Such a generalizable model may aid in prognostication by assisting in the personalized survival stratification of HNSCC patients. In this study, we validated the model using basic parameters (age and sex of patients at diagnosis, ethnicity, TNM staging, and tumor site) to ensure that it could easily be tested using data from other cancer centers. Having prior knowledge of the survival outcomes of a patient, especially using the ML model developed without treatment parameters, can assist clinicians in personalized treatment planning.

Materials and methods

Dataset for model training

The Surveillance, Epidemiology, and End Results (SEER) program database was used for model training. The database was queried for various subsites of HNSCC as shown in Table 1. Following exclusion criteria (all cases with missing and incomplete data were excluded), the results of the database query produced a total of 40,164 histopathologically confirmed HNSCCs. The clinicopathological parameters included for model training were age at diagnosis, ethnicity, sex, marital status, histopathological grade, stage classification according to the American Joint Committee on Cancer (AJCC) tumor-nodal-metastasis (TNM) 7th edition (Table 1), and treatment parameters (surgery, radiotherapy [RT], and chemoradiotherapy [CRT]). Overall survival was the primary endpoint and target variable. All the methods were performed in accordance with the 1964 Declaration of Helsinki and its subsequent amendments. The ethical permission to use the SEER data was granted to the first author.

Machine learning models

The data and the corresponding variables selected from the SEER database (Sect. 2.1) were used for training two different models (*Models A and B*) (Sect. 2.3). *Model-A* was trained using only clinicopathologic parameters (without the treatment parameters) and was further externally and independently validated using [a] multicenter data obtained from the Thuringian Cancer Registry (Germany) and [b] a multicenter prospective observational study obtained from Uppsala University Hospital (Sweden). Therefore, these two cohorts from Germany and Sweden form the basis for geographic independent external validation of *Model-A*. *Model-B*, on the other hand, was considered a control model and it was trained by combining clinicopathologic parameters (including marital status and tumor grade) and treatment-related factors as integrative input sources into the ML algorithm for estimating the OS of patients with HNSCC.

Variables	Total (N = 40,164) (%)	Categorization for machine learning analysis	Data type after categorization
Age at diagnosis			
< 40 years old (Young)	748 (1.9)	No categorization	Integer
>= 40 years old (Old)	39,416 (98.1)		
Sex			
Male	30,090 (74.9)	1 = Male	Integer
Female	10,074 (25.1)	2 = Female	
Marital status			
Married/living together	22,206 (55.3)	1 = Married	Integer
Unmarried/single	17,958 (44.7)	2 = Unmarried	
Ethnicity			
White	34,245 (85.3)	0 = White	Integer
Black	3714 (9.3)	1 = Black	
Other (American Indian/AK Native, Asian/Pacific Islander)	2205 (5.4)	2 = Other	
Tumor site			
Oropharynx	14,082 (35.1)	1 = Oropharynx	Integer
Lip & oral cavity	13,324 (33.2)	2 = Oral cavity	
Larynx	10,884 (27.1)	3 = Larynx	
Hypopharynx	1874 (4.7)	4 = Hypopharynx	
T stage			
Lower T stage (T0-T2)	26,216 (65.3)	T1 = 1	
Upper T stage (T3-T4)	13,948 (34.7)	T2 = 2	
N stage			
N0	20,911 (52.1)	N0 = 0	Integer
N1	5639 (14.0)	N1 = 1	
N2	12,706 (31.6)	N2 = 2	
N3	908 (2.3)	N3 = 3	
M stage			
M0	39,056 (97.2)	M0 = 0	Integer
M1	1108 (2.8)	M1 = 1	
Grade			
Low grade: Well differentiated and Moderately differentiated	27,264 (67.9)	Lower grade = 1	
High grade: Poorly differentiated and Undifferentiated	12,900 (32.1)	Higher grade = 2	
Treatment modalities			
Surgery (Sx) only	10,342 (25.8)	1	Integer
Radiotherapy (RT) only	4770 (11.9)	2	
Chemotherapy (CT) only	631 (1.6)	3	
Chemoradiotherapy (RCT)	10,180 (25.4)	4	
All treatment modalities	6001 (14.9)	5	
Sx + RT	5825 (14.5)	6	
Sx + CT	296 (0.7)	7	
No treatment given	2119 (5.3)	8	
Survival			
Death	22,992 (11.2)	1	Integer
Alive	17,172 (86.4)	0	

Table 1. Baseline demographic and fracture characteristics of head and neck squamous cell carcinoma patients in the training data (N = 40,164).

Machine learning model training process

The model development process is shown in Fig. 1. In the development of *Model-A*, we wanted to build a prognostic model based on available parameters before treatment. Thus, we specifically used basic patient- and tumor-related characteristics i.e. sex, ethnicity, age at diagnosis, tumor site, T stage, N stage, and M stage at diagnosis to accommodate the independent external validation cohorts – [a] the Thuringian Cancer Registry data, and [b] the multicenter prospective observational study obtained from the Uppsala University Hospital.

The data obtained were loaded into a Microsoft Azure automated learning studio and used for model development. Following the loading of data, the necessary training parameters, such as the desired training algorithms (voting ensemble, LightGBM, Extreme Gradient boosting), learning rate, k-fold cross-validation (5-fold cross validation), and performance metrics were defined. We adjusted various training hyperparameters'

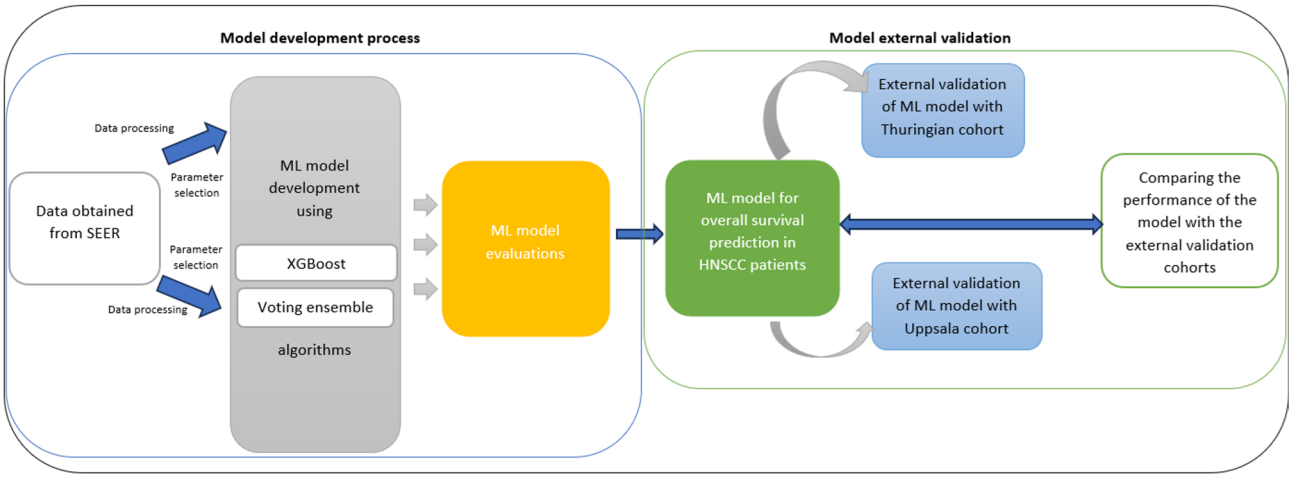

Fig. 1. A schematic representation of ML model development process. *HNSCC* Head and Neck Squamous Cell Carcinoma, *ML* Machine learning, *SEER* Surveillance, Epidemiology, and End Results (SEER).

tuning to maximize the performance of the model (Fig. 1). The hyperparameters with the best accuracy for each model were selected. The models were evaluated primarily based on the area under receiving operating characteristics curve (AUC). We excluded treatment-related parameters to minimize the variation in the treatment protocols between these three cohorts (SEER [United States], Sweden, and Germany). Therefore, we aimed at having an OS prediction insight to guide the treatment approach.

In developing *model-B* (i.e., a control model that includes marital status and tumor grading), TNM 7th edition staging scheme (T stage, N stage, and M stage), age at diagnosis, sex, tumor site, surgery, marital status at diagnosis, tumor grading, and treatment-related parameters (surgery, RT, CRT) were used. The idea of the control model (*Model-B*) was to evaluate the performance accuracy of the model when other factors such as marital status and tumor grading were combined with clinicopathologic parameters as integrative input sources into the ML algorithm for estimating the OS of patients with HNSCC.

Independent external validation of the ML model

Thuringian cancer registry data (Germany)

The clinicopathological variables selected from the Thuringian Cancer Registry included the American Joint Committee on Cancer (AJCC) Tumor-Nodal-Metastasis (TNM) 7th edition staging scheme (T stage, N stage, and M stage), age at diagnosis, sex, tumor site, and treatment-related parameters (surgery, RT, CRT, and combined-modality treatment), and OS status (Table 2). A detailed description of each of the variables and categorizations included is provided in Table 2.

The Ethics Committee of the Jena University Hospital approved the study (IRB No. 3204-07/11) with a waived requirement for informed consent since the study had a non-interventional retrospective design and all data were analysed anonymously[1].

The cohort comprised 8288 HNC cases from a population-based long-term analysis from 1996 to 2016 at the Thuringian Cancer Registry. The detailed characteristics of these cases have previously been published by Dittberner et al.[1]. In the present study, however, we extracted only the squamous cell carcinomas from the cohort presented by Dittberner et al. to ensure that we focused on a histologically homogeneous group of patients[1]. Furthermore, we performed further data pre-processing to remove all cases with missing data. In addition, we only included tumor grade 1–4 as a parameter. All the patients included were Caucasian. Following pre-processing, we had a total of 3950 cases for the training of the ML model. The distribution of patients with HNSCC is shown in Table 2. To validate *Model-A*, we selected the same variables used to train the model (i.e., basic patient- and tumor-related characteristics: T stage, N stage, and M stage, ethnicity, age at diagnosis, sex, and tumor site).

Data from the Uppsala University Hospital (Sweden)

We obtained a total of 419 cases from a multicenter prospective observational study from the Uppsala University Hospital (Sweden) to validate *Model-A*. Following the removal of unknown UICC 7th edition TNM cases, we were left with a total of 323 cases to validate *Model-A*. The cohort included oral, oropharyngeal, laryngeal, and hypopharyngeal cancers. From the variables available, we selected the variables required to validate *Model-A*. The description of the details of the external validation is presented in Table 2. Ethical approval was obtained from the Regional Ethical Review Board of Uppsala (No. 2014/447). Therefore, all the cohorts (SEER, Thuringian, and Uppsala) used in this study are based on 7th edition staging to guarantee homogeneity in terms of staging scheme and to avoid any concerns relating to the study validity.

Variables	Uppsala University Hospital, Sweden Total (*N* = 323) (%)	Thuringian Cancer Registry, Germany Total (*N* = 3950) (%)	Categorization for machine learning analysis
Age at diagnosis			
< 40 years old (Young)	7 (2.2)	84 (2.1)	No categorization
>= 40 years old (Old)	316 (97.8)	3866 (97.9)	
Sex			
Male	236 (73.1)	3343 (84.6)	1 = Male
Female	87 (26.9)	607 (15.4)	2 = Female
Ethnicity			
White	323 (100.0)	4098 (100.0)	
Tumor site			
Oropharynx	213 (65.9)	1144 (30.0)	1
Oral	80 (24.8)	1343 (34.0)	2
Larynx	25 (7.7)	960 (24.3)	3
Hypopharynx	5 (1.6)	503 (12.7)	4
T stage	**(UICC 7)**	**TNM 7th edition**	
Lower T stage (T1-T2)	216 (66.9)	2134 (54.0)	T1 = 1
Upper T stage (T3-T4)	107 (33.1)	1816 (46.0)	T2 = 2
N stage			
N0	98 (30.3)	1916 (48.5)	N0 = 0
N1	35 (10.8)	484 (12.3)	N1 = 1
N2	184 (57.0)	1383 (35.0)	N2 = 2
N3	6 (1.9)	167 (4.2)	N3 = 3
M stage			
M0	323 (100.0)	3803 (96.3)	M0 = 0
M1	–	147 (3.7)	M1 = 1
Treatment modalities			
Surgery (Sx) only	22 (6.8)	1100 (27.8)	1
Radiotherapy (RT) only	119 (36.8)	246 (6.2)	2
Chemoradiotherapy (RCT)	110 (34.1)	451 (11.4)	3
All treatment modalities	72 (22.3)	842 (21.3)	4
Sx + RT	–	1169 (29.6)	
No treatment given	–	142 (3.6)	
Overall survival			
Dead	42 (13.0)	2395 (60.6)	1
Alive	281 (87.0)	1555 (39.4)	0

Table 2. Baseline demographic and tumor characteristics of patients with HNSCC in the validation series from Uppsala university Hospital, Sweden (*N* = 323) and Thuringian cancer Registry, Germany (*N* = 3950).

Permutation feature importance

We performed permutation feature importance (PFI) to examine how each of these variables would contribute to the predictions made by these models (*Model-A* and *Model-B*). PFI works by shuffling the data in such a way that one feature is removed at a time while the corresponding effect of the shuffled feature on the performance metrics of the model is estimated. The larger the change, the more important is the feature to the model's performance in stratifying the patients into risk groups for OS[9].

Results

Patient characteristics

The study cohort used for model training (SEER data) included 40,164 patients with HNSCC: 30,090 males and 10,074 females in a male-to-female ratio of 3:1. The mean age at diagnosis was 63.7 years (SD ± 11.9: range 12–90) and the median age was 63 years. The HNSCC subsites in the series were lip and oral cavity (33.2%), oropharynx (35.1%), hypopharynx (4.7%), and larynx (27.1%) (Table 1). Considering the tumor, the AJCC 7th TNM staging scheme showed that 17 (0.04%) patients had a stage T0 tumor, 14,146 (29.8%) had T1, 12,053 (33.7%) had T2, 6730 (14.8%) had T3, and 7218 (21.7%) had stage T4. Correspondingly, 20,911 (52.1%) had N0, 5639 (14.0%) N1, 12,706 (31.6%) N2, and 908 (2.3%) N3; 39,056 (97.2%) M0, and 1108 (2.8%) M1. Regarding histopathological grading, 27,264 (67.9%) tumors were categorized as low grade (well-differentiated and moderately differentiated) and 12,900 (32.1%) were high grade (poorly differentiated and undifferentiated). The clinicopathologic characteristics are briefly summarized in Table 1.

Ethnicity formed another important parameter, 34,245 (85.3%) being of white origin, 3714 (9.3%) Black, and 2205 (5.34%) from other origins including American Indian/AK Native and Asian/Pacific Islander. Considering marital status, 22,206 (55.3%) were married (including common law), while 17,958 (44.7%) were considered unmarried (single, divorced, widowed, or separated) at the time of diagnosis (Table 1). The follow-up time ranged from 0 to 143 months with a total of 17,172 (42.8%) HNSCC patients being found to be alive at the end of the follow-up. The details of the oncological treatment are given in Table 1.

Performance accuracy and feature importance of the models

The predictive performance of *Model-A* showed a weighted AUC of 0.76 and an accuracy of 70.0% (Fig. 2). The external validation of *Model-A* using the Thuringian Cancer Registry showed a weighted AUC performance of 0.68 and an accuracy of 66.1%. This marked a decrease in the weighted AUC and a predictive performance accuracy when *Model-A* was subjected to external independent geographic validation using the Thuringian cohort. Conversely, the weighted performance AUC of *Model-A* using the Uppsala multicenter cohort was 0.76 and its performance predictive accuracy was 39.3%. For the control model (*Model-B*), the predictive performance of *Model-B* showed a weighted AUC of 0.79 and a predictive accuracy of 71.9% after model training.

Therefore, a marked slight increase in performance accuracy was observed when *Model-A* was subjected to external independent geographic validation using both the Thuringian and Uppsala cohort. In both external validation cases, there was a decrease in the predictive accuracy of *Model-A*.

Evaluating the input variables for aggregate feature importance

In terms of feature importance for Model-A, which was externally validated, the top five features were age at diagnosis, T stage, tumor site, N stage at diagnosis, and ethnicity (Fig. 3).

Conversely, the top eight features for predicting overall survival by Model-B (control model) were age at diagnosis, T stage, tumor site, marital status, surgery, N stage, radiotherapy (RT), and CRT. These were the most important parameters for the predictive ability of the model for OS (Fig. 4). In both Model-A and Model-B, age at diagnosis, T stage, and tumor site are among the top three variables.

Discussion

Independent geographic validation of machine learning (ML) models remains one of the most significant challenges preventing the recommendation of ML models for further clinical validations[13,14]. This is due to concerns of generalizability and model biasness. Therefore, in this study, we developed a ML model using a cohort from the United States and further validated the model using two geographically different cohorts obtained from Germany and Sweden to evaluate how the model would perform with datasets outside the development country and environment. Therefore, the results fulfilled our goal of achieving model generalization to fully understand the real predictive ability of the model, detecting model bias, and, most importantly, facilitating the recommendation of a generalized model for further clinical evaluations. The motivation of this study was not to display a model with a relatively high AUC value or accuracy. Rather, we intended to demonstrate the significance of externally validating ML models for generalizability.

Remarkably, having a generalized model can aid personalized OS risk stratification and consequently enhance targeted treatment plans for patients with HNSCC. In this study, our model was developed and validated using basic parameters (age and sex of patients at diagnosis, ethnicity, TNM staging, and tumor site) to ensure that it could easily be tested using data from other cancer centers. Our model examined the complex relationships between the input variables used for model development using PFI, thereby demonstrating how each of these variables contribute to the predictive performance of the model in terms of the overall survival of HNC patients.

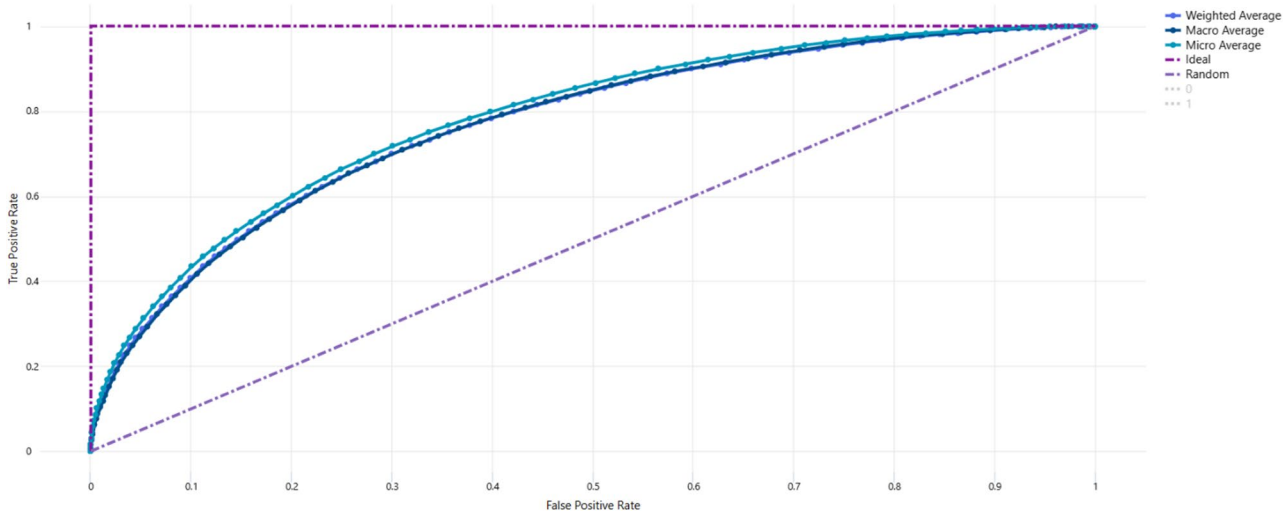

Fig. 2. The area under receiving operating characteristic curve for Model A.

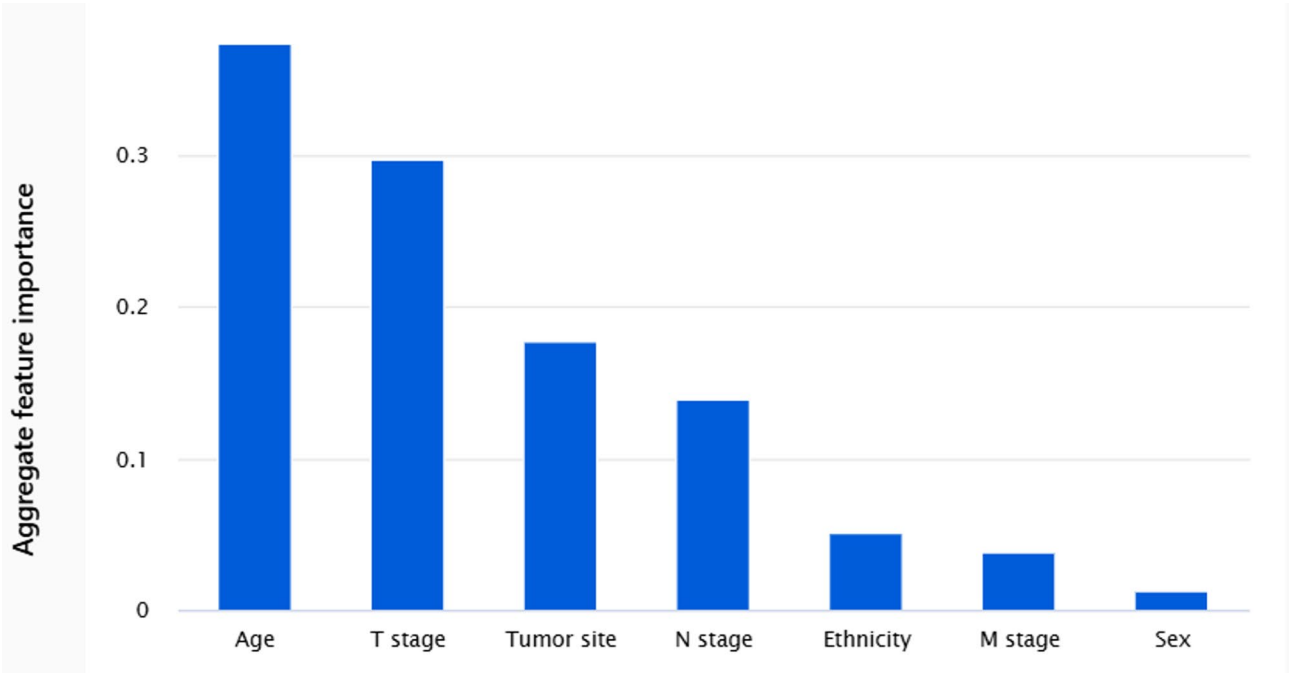

Fig. 3. Aggregate feature importance of input variables for Model-A.

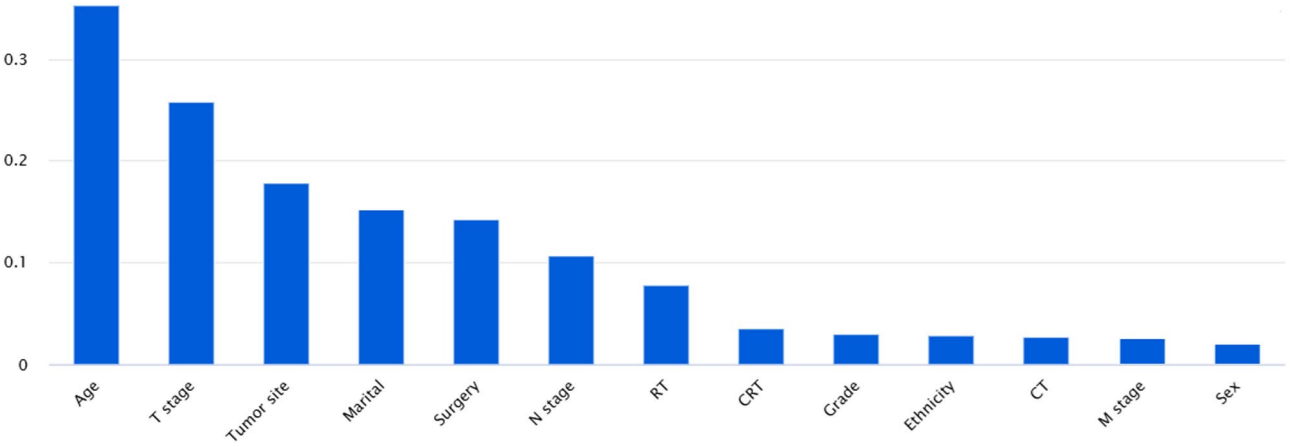

Fig. 4. Aggregate feature importance of input variables for Model-B.

In recent years, there has been a surge in the development of prognostic prediction models for prognostication in various cancers[14,15]. However, only a small number have been externally validated[14]. The model performance cardinal plot presented by Alabi et al. showed that a model may show worse performance in a new patient population compared with the one used in the development phase[14]. Such a model should not be recommended for further clinical validation[14,15]. This is because the performance metric, such as model accuracy, was derived from the development data, which may be imbalanced or biased, and thus the model may not be viable outside the development area. Several validation methods, such as internal and temporal validation methods, have been widely used to ensure that the model is generalizable. However, both internal (split-sample, cross-validation, and bootstrapping) and temporal validation use the same data except that temporal validation may include patients sampled at a later (or earlier) time point. Hence, temporal validation lies between internal and external validation paradigms.

Independent external validation generally means that the validation cohort has been assembled in a completely separate manner from the development cohort[14,15]. The independent geographic external validation presented in this study provided the opportunity to compare head-to-head the performance of the model in OS risk stratification of HNSCC patients before treatment in the development data and data from outside the development country. Considering the sensitive nature of medical oncology in terms of health and safety of patients, clinical decisions based on incorrect prediction models could have dire consequences and adverse

effects on various patient outcomes. Therefore, external validation is necessary to guarantee the model's reproducibility and generalizability. Reproducibility evaluates whether the prediction formula would be valid in new individuals who are similar to the development population (reproducibility). This may be achieved through internal validation techniques. Generalizability (transportability), on the other hand, involves exploring whether the prediction model is transportable (i.e. would show similar predictive performance) to a separate population with different patient characteristics. It is worth noting that, as shown in the model performance cardinal plot presented by Alabi et al., external validation may not guarantee a significant increase in performance accuracy[14]. This corroborates the finding from the study by Ramspek et al., which emphasized that there is usually a difference in performance accuracy when the model is externally validated[15]. In any case, the performance accuracy should be relative to the performance shown in the training cohorts.

We observed a reduction in the performance accuracy of the model when externally validated with validation cohort that was assembled in a completely separate manner from the development cohort[17]. It is typical for prediction models to generally perform more poorly in external validation than in development[18]. Several reasons may be responsible for the reduction in performance accuracy. For example, there seems to be an instance of data distribution shift. For instance, Model-A was trained with data that consisted of ethnicities that included Caucasian, Black, and others (Table 1). However, it was validated with data from predominantly White ethnicity (Table 2). Additionally, there may be issues relating to suspected feature leakage for OS, especially between the base model (Model-A) and the corresponding validation cohorts. For example, the OS was significantly imbalanced for the Uppsala cohort. Furthermore, inadequate processing consistency may be responsible for the overall reduction in accuracy, especially in the Uppsala data where significant number of patients were removed due to missing UICC 7th edition staging. As such, any available external validation should not be perceived as a license for model implementation[19]. Despite these, external validation still represents a reliable way to evaluate ML model generalizability[14,15,17,20]. However, our study argued that multiple external validations are important to properly scrutinize ML model reproducibility and generalizability[19]. Furthermore, we argue that a robust external validation design is imperative as lower methodological quality at model development associates with poorer model performance at external validation[21].

The validated model showed that age at diagnosis, T stage at presentation, and tumor site are among the top variables for OS risk stratification. This finding is supported by several studies emphasizing the clinical tumor stage at presentation as a major predictor of survival in HNSCC[22]. This finding is supported by our study where T-stage was identified in both models (Model-A & Model-B). Understandably, T stage and overall tumor stage of HNSCC are relevant for OS of these patients since a significant number of patients are diagnosed at an advanced stage of cancer, with an increased risk of recurrence and distant metastasis[2,22–24]. This further justifies why Pontes et al. highlighted that patients with recurrent or metastatic HNSCC had poor prognoses[25]. Our model also identified the patients' age at diagnosis as an important prognostic factor for OS. This finding is in tandem with the conclusion of the studies by Cadoni et al. and Talani et al., in which it was suggested that age at diagnosis is a risk factor for early death among patients with HNSCC with curative treatment intent[23,26].

In this study, for the control model (Model-B) we found that marital status has a role to play in the survival of patients with HNSCC. This finding underscores the need to recognize the marital status of patients with HNSCC for future survivorship investigations. This may be due to advantages through social and spousal support mechanisms. This finding corroborates the conclusion from the studies of Shi et al. and Osazuwa-Peters et al., in which it was emphasized that being married confers a survival advantage on HNSCC survivors[27,28]. In addition, this finding is in tandem with the conclusions from some other studies that examined the effect of marital status on patients with HNSCC sites[29–31]. Therefore, social support should be considered an important part of standard care for managing patients with HNSCC[27,32].

For Model-B, both surgery and RT were found to be important factors for the ML model in estimating OS in patients with HNSCC. This is expected because these approaches form the mainstay for treatment of HNSCCs[33]. These treatment options aim to eradicate the tumor, although they are associated with a risk of adverse effects. Therefore, considering the importance of surgery and RT for survival as emphasized in this study, there are overarching considerations that are important for minimizing the associated morbidity and managing its sequelae[33]. RT is frequently used in both primary and adjuvant treatment settings[33,34] and the key role played by RT in OS is, for example, supported by the study of Ronen et al. and Mazul et al.[33,34]. Note that RT should be carefully planned to avoid either delays or extended treatment duration due to interruptions or delays, both of which have been reported to be associated with poor oncologic outcomes[34,35].

Our study has some limitations. First, the SEER data used for ML model development were retrospective in nature, which may introduce some level of bias into the model development. Second, the number of cases used for externally validating the developed ML model development was not uniform among HNSCC sites. Third, there was missing data for ethnicity (in the Uppsala data) and marital status (in the Thuringian data). Therefore, we assumed that all the patients were Caucasian in the Uppsala data, which may introduce some level of bias into the model performance during the external validation step. Therefore, in future, external validation cohorts achieved through a more balanced dataset of tumor sites and HPV status would be preferable. In conclusion, there are growing discussions regarding including an external validation procedure as part of the ML model development pipeline. However, due to data privacy and security concerns, this may not be feasible. In addition, external validation by the same researcher may lead to the temptation to finetune the performance accuracy of the external validation results. Therefore, externally validating these models with multi-institutional datasets and testing them in the context of clinical trials is warranted for safe clinical implementation. Externally validating the model can guarantee that the predictive model can aid in early intervention and treatment planning. Having prior knowledge of the survival outcomes of a patient, especially using the ML model developed without treatment parameters, can assist clinicians in personalized treatment planning and enhance insightful decision-making[9].

Externally validating an ML model may guarantee reproducibility and generalizability. Therefore, the use of multiple and a relatively large number of external validation cohorts is necessary to achieve reproducibility and generalizability. However, to benefit from the variability in these external validation cohorts, a federated ML paradigm is envisioned to ensure that the base model benefits from these external cohorts to enhance generalizability by producing a new model after the external validation process. In the present external validation workflow, a new model is not produced from the external validation process. Rather, it shows the level of generalizability. Therefore, in future studies, the base model will leverage federated ML paradigm to benefit from the variability in the external validation cohorts to form a new ML model that can be recommended for further clinical evaluation. This approach remains crucial for further clinical independent assessment and to guarantee that the model can enhance effective treatment planning and improved disease management. As a sensitivity analysis to Model-A (Model_A_sensitivity_analysis), we included treatment-related parameters to evaluate if there would be any change in the performance accuracy of the model (Supplementary I). No significant difference in the predictive performance of Model-A and Model_A_sensitivity_analysis was observed. This result appears to further support the possibility of having prior information about survival with a ML model trained only on clinicopathologic parameters (without treatment-related parameters). This may facilitate early treatment planning and provide an adequate individualized treatment approach for this patient population.

Data availability

The datasets generated in this study are available from the corresponding author upon reasonable request.

Received: 2 June 2025; Accepted: 17 November 2025
Published online: 29 November 2025

References

1. Dittberner, A. et al. Gender disparities in epidemiology, treatment, and outcome for head and neck cancer in Germany: A population-based long-term analysis from 1996 to 2016 of the Thuringian cancer registry. *Cancers* **12**, 3418 (2020).
2. Johnson, D. E. et al. Head and neck squamous cell carcinoma. *Nat. Rev. Dis. Primers*. **6**, 92 (2020).
3. Lilloni, G. et al. Exploring patient stratification in head and neck squamous cell carcinoma using machine learning techniques: preliminary results. *Curr. Probl. Cancer*. **53**, 101154 (2024).
4. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **72**, 7–33 (2022).
5. Sung, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
6. Chow, L. Q. M. Head and neck cancer. *N. Engl. J. Med.* **382**, 60–72 (2020).
7. Amaral, M. N., Faísca, P., Ferreira, H. A., Gaspar, M. M. & Reis, C. P. Current insights and progress in the clinical management of head and neck cancer. *Cancers* **14**, 6079 (2022).
8. Alabi, R. O. et al. Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer. *Int. J. Med. Informatics*. **145**, 104313 (2021).
9. Alabi, R. O., Elmusrati, M., Leivo, I., Almangush, A. & Mäkitie, A. A. Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. *Sci. Rep.* **13**, 8984 (2023).
10. Alabi, R., Almangush, A., Elmusrati, M., Leivo, I. & Mäkitie, A. A. An interpretable machine learning prognostic system for risk stratification in oropharyngeal cancer. *Int. J. Med. Informatics*. **168**, 104896 (2022).
11. Alabi, R. O., Elmusrati, M., Leivo, I., Almangush, A. & Mäkitie, A. A. Artificial Intelligence-Driven radiomics in head and neck cancer: current status and future prospects. *Int. J. Med. Informatics*. **188**, 105464 (2024).
12. Alabi, R. O., Almangush, A., Elmusrati, M. & Mäkitie, A. A. Deep machine learning for oral cancer: from precise diagnosis to precision medicine. *Front. Oral Health*. **2**, 794248 (2022).
13. Alabi, R. O. et al. Machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for future—A systematic review. *Artif. Intell. Med.* **115**, 102060 (2021).
14. Alabi, R. O. et al. Application of artificial intelligence for overall survival risk stratification in oropharyngeal carcinoma: A validation of progtool. *Int. J. Med. Informatics*. **175**, 105064 (2023).
15. Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C. & van Diepen, M. External validation of prognostic models: what, why, how, when and where? *Clin. Kidney J.* **14**, 49–58 (2021).
16. Mäkitie, A. A. et al. Artificial intelligence in head and neck cancer: A systematic review of systematic reviews. *Adv. Ther.* **40**, 3360–3380 (2023).
17. Steyerberg, E. W. & Harrell, F. E. Prediction models need appropriate internal, internal–external, and external validation. *J. Clin. Epidemiol.* **69**, 245–247 (2016).
18. Siontis, G. C. M., Tzoulaki, I., Castaldi, P. J. & Ioannidis, J. P. A. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J. Clin. Epidemiol.* **68**, 25–34 (2015).
19. La Roi-Teeuw, H. M. et al. Don't be misled: 3 misconceptions about external validation of clinical prediction models. *J. Clin. Epidemiol.* **172**, 111387 (2024).
20. Ho, S. Y., Phua, K. & Wong, L. Bin Goh, W. W. Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns* **1**, 100129 (2020).
21. Helmrich, I. R. A. R. et al. Does poor methodological quality of prediction modeling studies translate to poor model performance? An illustration in traumatic brain injury. *Diagn. Progn Res.* **6**, 8 (2022).
22. Adoga, A. A. et al. The predictive factors of primary head and neck cancer stage at presentation and survival in a developing nation's tertiary hospital. *SAGE Open. Med.* **6**, 205031211879241 (2018).
23. Cadoni, G. et al. Prognostic factors in head and neck cancer: a 10-year retrospective analysis in a single-institution in Italy. *Acta Otorhinolaryngol. Ital.* **37**, 458–466 (2017).
24. Reyes-Gibby, C. C. et al. Survival patterns in squamous cell carcinoma of the head and neck: pain as an independent prognostic factor for survival. *J. Pain*. **15**, 1015–1022 (2014).
25. Pontes, F. et al. Survival predictors and outcomes of patients with recurrent and/or metastatic head and neck cancer treated with chemotherapy plus cetuximab as first-line therapy: A real-world retrospective study. *Cancer Treat. Res. Commun.* **27**, 100375 (2021).
26. Talani, C. et al. Early mortality after diagnosis of cancer of the head and neck – A population-based nationwide study. *PLoS ONE*. **14**, e0223154 (2019).
27. Osazuwa-Peters, N. et al. What's love got to do with it? Marital status and survival of head and neck cancer. *Eur. J. Cancer Care* **28**, (2019).

28. Shi, X., Zhang, T., Hu, W. & Ji, Q. Marital status and survival of patients with oral cavity squamous cell carcinoma: a population-based study. *Oncotarget* **8**, 28526–28543 (2017).
29. Massa, S. T. et al. Demographic predictors of head and neck cancer survival differ in the elderly. *Laryngoscope* **129**, 146–153 (2019).
30. Du, X. et al. Marital status and survival in laryngeal squamous cell carcinoma patients: a multinomial propensity scores matched study. *Eur. Arch. Otorhinolaryngol.* **279**, 3005–3011 (2022).
31. Schaefer, E. W. et al. Effect of marriage on outcomes for elderly patients with head and neck cancer: marriage effect in head and neck cancer. *Head Neck*. **37**, 735–742 (2015).
32. Mäkitie, A. A. et al. Psychological factors related to treatment outcomes in head and neck cancer. *Adv. Ther.* https://doi.org/10.1007/s12325-024-02945-3 (2024).
33. Ronen, O. et al. Emerging concepts impacting head and neck cancer surgery morbidity. *Oncol. Ther.* **11**, 1–13 (2023).
34. Mazul, A. L. et al. Duration of radiation therapy is associated with worse survival in head and neck cancer. *Oral Oncol.* **108**, 104819 (2020).
35. Shaikh, T. et al. The impact of radiation treatment time on survival in patients with head and neck cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **96**, 967–975 (2016).

Acknowledgements

The authors would like to thank the Sigrid Jusélius Foundation and the Finska Läkaresällskapet. The study was supported by the Finnish State Research Funding to the Helsinki University Hospital and the Turku University Hospital, and the Swedish Cancer Society (grant numbers 2015/363, 2018/502, 21 1419 Pj, and 24 3394 Pj). Open access funded by Helsinki University Library.

Author contributions

Study concepts and study design: Alabi, R.O., Laurell G., Ylva, T.E. & Mäkitie A.A.Studies extraction: Alabi R.O., Orlando G.L., Laurell G., & Ylva, T.E.Acquisition and quality control of included studies: Laurell G., Ylva, T.E., Mäkitie AA & Almangush A.Data analysis and interpretation: Alabi RO, Orlando, G.L., and Mäkitie A.A.Manuscript preparation: Alabi R.O. & Mäkite A.A.Manuscript review: Mäkitie A.A., Elmusrati M, Laurell G., & Ylva, T.E.Manuscript editing: Almangush A, Laurell G., Ylva, T.E, Mäkitie AA.Manuscript quality: Laurell G., Orlando G.L., Ylva T., & Mäkitie AA.All authors approved the final manuscript for submission.

Declarations

Competing interests

The authors declare no competing interests.

Informed consent for the data used in this study

The Ethics Committee of the Jena University Hospital approved the study (IRB No. 3204-07/11) with a waived requirement for informed consent since the study had a non-interventional retrospective design and all data were analyzed anonymously. For Uppsala data, informed consent and ethical approval was obtained from the Regional Ethical Review Board of Uppsala (No. 2014/447) and the Swedish Ethical Review Authority (No. 2025-01114-02). The SEER data is a public data with informed consent obtained for all the patients as stated by the National Health Institute (NIH). The ethical permission to use SEER data was granted to the first author.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-29295-6.

Correspondence and requests for materials should be addressed to R.O.A.