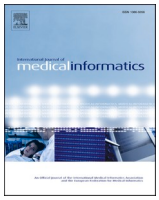
International Journal of Medical Informatics 175 (2023) 105064

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Application of artificial intelligence for overall survival risk stratification in oropharyngeal carcinoma: A validation of ProgTOOL

Rasheed Omobolaji Alabi [a,b,*], Anni Sjöblom [c], Timo Carpén [a,c,d], Mohammed Elmusrati [b], Ilmo Leivo [e], Alhadi Almangush [a,c,e,f,1], Antti A. Mäkitie [a,d,g,1]

[a] Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland
[b] Department of Industrial Digitalization, School of Technology and Innovations, University of Vaasa, Vaasa, Finland
[c] Department of Pathology, University of Helsinki, Helsinki, Finland
[d] Department of Otorhinolaryngology – Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
[e] University of Turku, Institute of Biomedicine, Pathology, Turku, Finland
[f] Faculty of Dentistry, Misurata University, Misurata, Libya
[g] Division of Ear, Nose and Throat Diseases, Department of Clinical Sciences, Intervention and Technology, Karolinska Institute and Karolinska University Hospital, Stockholm, Sweden

ARTICLE INFO

Keywords:
Machine learning
External validation
Internal validation
Overall survival
Prognostication
Web-based tool
Oropharyngeal

ABSTRACT

Background: In recent years, there has been a surge in machine learning-based models for diagnosis and prognostication of outcomes in oncology. However, there are concerns relating to the model's reproducibility and generalizability to a separate patient cohort (i.e., external validation).

Objectives: This study primarily provides a validation study for a recently introduced and publicly available machine learning (ML) web-based prognostic tool (ProgTOOL) for overall survival risk stratification of oropharyngeal squamous cell carcinoma (OPSCC). Additionally, we reviewed the published studies that have utilized ML for outcome prognostication in OPSCC to examine how many of these models were externally validated, type of external validation, characteristics of the external dataset, and diagnostic performance characteristics on the internal validation (IV) and external validation (EV) datasets were extracted and compared.

Methods: We used a total of 163 OPSCC patients obtained from the Helsinki University Hospital to externally validate the ProgTOOL for generalizability. In addition, PubMed, OvidMedline, Scopus, and Web of Science databases were systematically searched according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

Results: The ProgTOOL produced a predictive performance of 86.5% balanced accuracy, Mathew's correlation coefficient of 0.78, Net Benefit (0.7) and Brier score (0.06) for overall survival stratification of OPSCC patients as either low-chance or high-chance. In addition, out of a total of 31 studies found to have used ML for the prognostication of outcomes in OPSCC, only seven (22.6%) reported a form of EV. Three studies (42.9%) each used either temporal EV or geographical EV while only one study (14.2%) used expert as a form of EV. Most of the studies reported a reduction in performance when externally validated.

Conclusion: The performance of the model in this validation study indicates that it may be generalized, therefore, bringing recommendations of the model for clinical evaluation closer to reality. However, the number of externally validated ML-based models for OPSCC is still relatively small. This significantly limits the transfer of these models for clinical evaluation and subsequently reduces the likelihood of the use of these models in daily clinical practice. As a gold standard, we recommend the use of geographical EV and validation studies to reveal biases and overfitting of these models. These recommendations are poised to facilitate the implementation of these models in clinical practice.

* Corresponding author at: Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland.
E-mail address: rasheed.alabi@helsinki.fi (R. Omobolaji Alabi).
[1] The last two authors supervised this work.

https://doi.org/10.1016/j.ijmedinf.2023.105064
Received 1 December 2022; Received in revised form 31 March 2023; Accepted 3 April 2023
Available online 6 April 2023

1. Introduction

The incidence of oropharyngeal squamous cell carcinoma (OPSCC) which is one of the most common carcinomas of the head and neck has increased in recent years with a significant geographical variation in the incidence for males and females, and mortality [1]. OPSCC has an annual incidence of about 100,000 new cases and a mortality rate of 48,143 globally [1]. The risk factors that can be considered from two causal mechanisms [1]. Firstly, typical etiological agents for head and neck squamous cell carcinoma (HNSCC) (i.e. heavy alcohol consumption and smoking) and secondly, human papilloma virus infection [HPV] [1–3].

In the past decade, oncogenic HPV has emerged as the main causative agent for OPSCC in many countries [4]. Recently, HPV-related OPSCC surpassed cervical cancer as the most common HPV-caused cancer [5]. In the absence of available screening methods for OPSCC, most cases are detected at an advanced disease stage, which warrants the use of combined treatment approaches [5]. Delayed diagnosis will still influence the management outcome. As all the therapeutical modalities have the potential to cause side effects affecting quality of life, the required decision making is crucial. Thus, having an assistive tool for risk stratification of the survival OPSCC patients can help in targeted treatment planning and optimal care to meet their psychosocial need and improve their quality of life [6].

In recent years, artificial intelligence (AI), or its subfield, machine learning (ML) has been touted to contribute to improved clinical decision-making and proper management of cancer [7]. Many cancer centers and medical institutions are posited to embracing AI-based technologies targeted at improving clinical efficiency and providing safe and valuable oncological care, which are necessary for achieving improved quality of patient health [8]. Despite these suggested benefits, only few of these developed models and AI-based tools have made it to actual medical practice or patient care [9].

There are several barriers to the implementation of these models or AI-based tools in clinical practices [9]. These barriers have been broadly divided into two categories – barriers that are inherent to the science of ML itself and the barriers that relate to the clinical concerns of these models in healthcare [9]. The examples of barriers inherent to the science of ML include black box concerns, data, results and model interpretability, and generalizability of the models [9–14]. Barriers related to the clinical concerns include explainability, usability, clinical validations, ethical and moral, and regulatory framework [9,10,12] (Fig. 1).

Of note, the generalizability concern is somewhat directly related to the clinical validation concerns of the clinicians. Generalizability signifies the performance ability of the ML models outside the data from which the model was originally trained on [15]. This is generally known as the external validation approach for the developed models. It provides an impression of how the model would perform in actual clinical use. Therefore, once the model is considered generalizable, it implies that the model may be ready for clinical validation. The clinical validations seek to further reassure the clinicians on the performance of the model.

The performance of most of these models has mainly been evaluated using unseen data from the same initial longitudinal dataset set aside for the performance evaluation [16]. These set-aside data are usually known as a test set and the evaluation approach, in this case, is known as internal validation. Remarkably, most of these models, although internally validated, have shown considerably lower performance accuracy when externally validated with datasets from different institutions [16–20]. Therefore, the lack of external validation raises a concern about the true performance of these models. The inability to understand the true performance may impede the model's progression to clinical validations. According to the published secondary research regarding the application of AI/ML models in clinical decision-making, several factors may be responsible for the lack of external validation. For instance, the model may not be publicly available for others to conduct an independent external validation study [21].

This study aims to provide a validation study for a recently published ML model for overall survival risk stratification in OPSCC [3]. Recently, this model has been integrated as a web-based prognostic tool (ProgTOOL) for overall survival risk stratification of OPSCC. Following addressing limitations of AI in oncology [8,22], the clinical significance

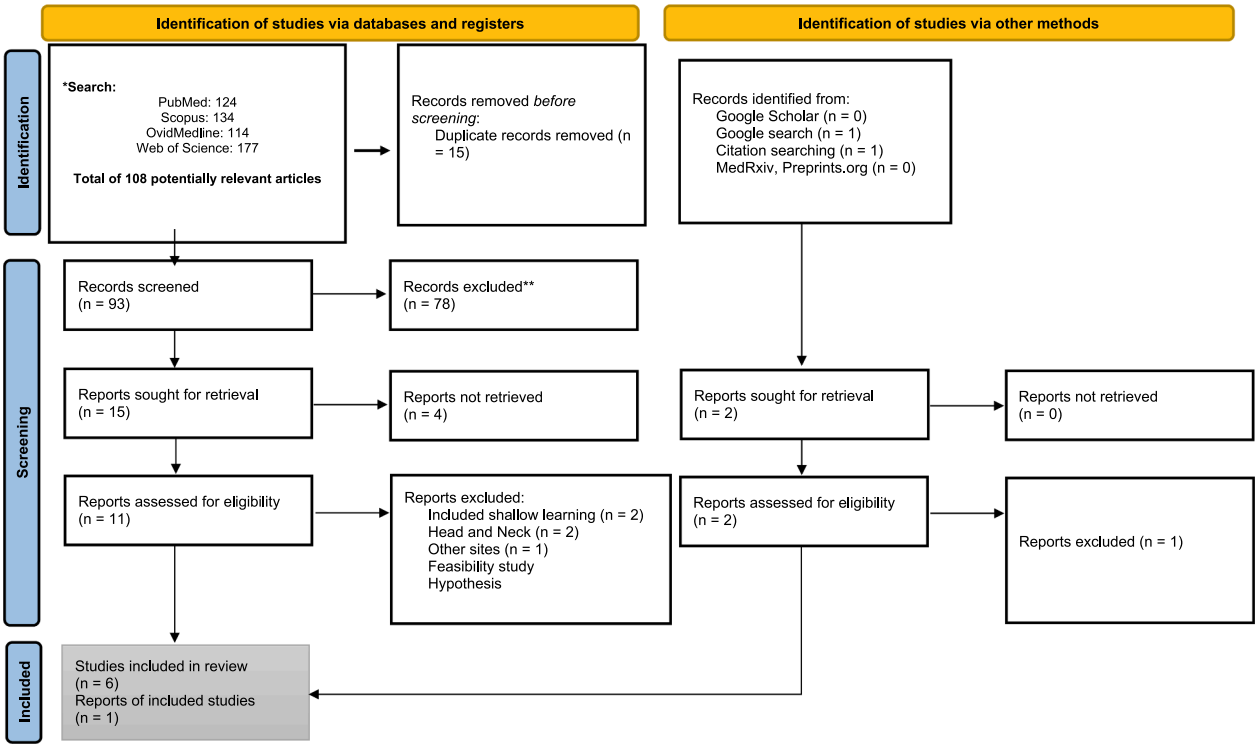


Fig. 1. A schematic of data extraction using the PRISMA flowchart.

of a validly tested and evaluated web-based AI prognostic tool is to serve as an ancillary tool for estimating the chance of survival and consequently providing informed decision-making regarding the proper management for OPSCC patients. This can enhance effective treatment planning of OPSCC and enable lower-costs, time-saving benefits, and better quality of life for the patients. In addition, this study also aims to systematically reviews the published studies that have applied ML to aid the prognostication of OPSCC. The essence of the review is to examine how many of these models in published studies were externally validated. Additionally, we examine the current methods for external validation of these models and compare the reported performances of internal validation to external validation. These afore-mentioned considerations are necessary to assess the readiness of the models for clinical validations. Consequently, paving way for the path to implementation of these models in daily practice. To the best of our knowledge, this is the first study to evaluate current applications of external validations of ML models for OPSCC prognostication. OPSCC was considered in this study as it constitutes a frequent tumor type of the upper aero-digestive and has been characterized with increasing incidence in the last decades [23].

2. Material and methods

2.1. Research questions

The research question was: "What is the predictive performance of a web-based prognostic tool for oropharyngeal squamous cell carcinoma (OPSCC) patients when externally validated with new cases"? To answer this main research question, we externally validated a publicly available web-based tool (https://oncotelligence.com/default.aspx) for risk stratification of OPSCC patients (**sub-section 2.2**). In addition, we reviewed the published studies to evaluate how many of these studies were externally validated (**sub-section 2.3**). Furthermore, we evaluated the modalities of external validation of these models. We explored the importance of external validation as an important approach that can influence the integration of these models into daily clinical practice. The performance of the model following external validation procedure was reported using the checklist for assessment of medical AI (ChAMAI formerly IJMEDI) [Supplementary I] [24].

2.2. External validation of a web-based prognostic tool

2.2.1. Data description, inclusion, and exclusion.

A total of 224 retrospective cases of OPSCC from 2012 to 2016 were extracted from the electronic patient records at the Helsinki University Hospital (Finland) to externally validate our recently introduced machine learning-based web tool for overall risk stratification in OPSCC patients. The inclusion criteria for this study were a prospectively collected cases that included clinical and pathologic characteristics such as gender, age at diagnosis, stage, TNM class, grade of differentiation, marital status, human papillomavirus status, treatment modalities (surgery, and radiotherapy), disease-free survival months, and overall survival status recorded for the patients. Of note, marital status has long been recognized as an important prognostic factor for many cancers [25,26]. More specifically, there are reports which demonstrates a survival benefit for married cancer patients, especially for human papilloma virus related cancers such as oropharyngeal or cervical cancer [27–30]. All these extracted parameters followed the corresponding parameters contained in the web-based tool (ProgTOOL). The included patients were predominantly of Caucasian origin. The exclusion criteria included patients with missing value in any of the above mentioned clinical and pathologic characteristics since being a validation study.

2.2.2. Data preparation

Following the exclusion criteria, that is, cleaning and pre-processing of the data for missing values (deletion of missing rows), a total of 163 cases were finally used to validate the tool (Table 1). There were no outliers in the dataset. Hence, there was no need for imputation analysis for the missing values in this validation study. As this study is aimed at validating an already deployed machine learning model, neither feature preprocessing nor target variable balancing was performed. However, standardization of the dataset was performed (by converting all the variables, except the age of the patients into categorical variable as shown in Table 1). In addition, the imbalance nature of the dataset used for this external validation was mentioned as one of the limitations of this study.

Information regarding marital status was missing from the validation dataset. Thus, we assumed two scenarios (married or single) for the

Table 1
External validated cases obtained from Helsinki University Hospital for oropharyngeal squamous cell carcinoma (N = 163).

Parameters	Total Number of cases for external validation study (n = 163)	Categorization for testing in according to the web-based prognostic tool (ProgTOOL)
Age at diagnosis		
Age ≥ 40	162 (99.4%)	No categorization (Not a parameter for the web-based prognostic tool)
Age < 40	1 (0.6%)	
Gender:		
Male	123 (75.5%)	0 = Male
Female	40 (24.5%)	1 = Female.
Grade:		
Grade I: Well differentiated	4 (2.4%)	Grade I = 1
Grade II: Moderately differentiated	28 (17.2%)	Grade II = 2
Grade III: Poorly differentiated	131 (80.4%)	Grade III = 3
HPV Status:		
Negative	42 (25.8%)	0 = HPV-negative
Positive	121 (74.2%)	1 = HPV-positive
Site:		
Base of tongue	46 (28.2%)	1 = Base of tongue
Oropharynx	21 (12.9%)	2 = Oropharynx
Tonsil	96 (58.9%)	3 = Tonsil
Tumor (T-stage)		
T1	53 (32.5%)	1 = T1
T2	53 (32.5%)	2 = T2
T3	17 (10.4%)	3 = T3
T4	40 (24.5%)	4 = T4
Nodal (N-stage)		
N0; No regional lymph node metastasis	25 (15.3%)	0 = N0
N1; Single node regional lymph node metastasis	16 (9.8%)	1 = N1
N2; One or two lymph nodes metastasize	120 (73.6%)	2 = N2
N3; Cancer has spread to one or more lymph node	2 (1.2%)	3 = N3
Metastases (M−stage)		
AJCC M0; No distant metastasis	161 (98.8%)	0 = M0
AJCC M1; Presence of distant metastasis	2 (1.2%)	1 = M1
Treatment parameters		
Surgery with postoperative radiotherapy (Sx + RT)	23 (14.1%)	1 = Sx + RT
Surgery with chemoradiotherapy (Sx + CRT)	35 (21.5%)	1 = Sx + CRT
Definitive chemotherapy	84 (51.5%)	1 = CRT
Surgery alone	8 (4.9%)	1 = Surgery
No treatment given	13 (8.0%)	0 = No treatment given
Overall Status		
Alive	118 (72.4%)	0 = Alive
Dead	45 (27.6%)	1 = Dead

HPV: Human papillomavirus.

patients to externally validate the web-based tool. The included clinical and pathologic parameters in this validation study were categorized in similar manner as the original study that was used to develop the web-based prognostic tool [3]. All patients included in this study were treated with curative intent. The study was approved by the Research Ethics Board at the Helsinki University Hospital and an institutional permission for the study was granted (Dnr: 51/13/03/02/2013). A detailed description of the development of the ML model has been reported and published by our group [3].

2.2.3. Geographic external validation approach

We used geographical external validation approach recommended by Ramspek et al. for a publicly available machine learning model [31]. This ensures that the external validation cohort used in this study is structurally different from the development cohort. These differences lie in the different region or country and the source of the data which may inform a different type of care setting. In our previous study, the development cohort was obtained from the largest publicly available databases managed by the National Cancer Institute (NCI) through the Surveillance, Epidemiology, and End Results (SEER) Program of the National Institutes of Health (NIH) in the United States. While the external validation was obtained from Helsinki University Hospital (Finland). The former data was a registry data while the latter was a hospital data. Therefore, these two data appeared to be structurally different, and we have presented the area under receiving operating characteristics curve (AUC), Net Benefit, and Brier score to evaluate the model discrimination power, utility, and calibration abilities, respectively.

2.2.4. Evaluation metrics from external validation approach

The performance of the web-based tool from external validation process was evaluated using accuracy (weighted and balanced accuracy), Mathew's correlation coefficient, weighted AUC, Net Benefit, and Brier score. The Brier score was calculated by considering the prediction of the model at every 10th instance within the external validation data (i.e., a total of 16 instances since the data contained 163 cases). Similarly, we used the formular proposed by Vickers et al. to calculate the Net Benefit of the model at 10%−50% probability [32,33].

2.3. Summary of the web-based prognostic tool model development and performance

The web-based tool was developed to classifying OPSCC patients into chance of survival groups (low-chance or high-chance) for overall survival (binary classification task). A trained voting ensemble machine learning algorithm was integrated as the web-based prognostic tool. The model was trained to classify the OPSCC into the chance of overall survival (model output). It was trained using a 5-fold cross validation with hyperparameter tunning (*extreme gradient boosting* [XGB], standardization: *standard scaler wrapper*, *colsample_bytree*: 0.9, *eta*: 0.3, *gamma*:1, *max_depth*: 10, *n_estimators*: 25). Each of the input variables were categorized with the exception of age of the patients (Table 1) before cross validation [3]. The performance of the model from the training phase showed 89.8% accuracy and 86.3% balanced accuracy, respectively [3]. Additionally, the Matthews' correlation coefficient, and weighted area under curve were 0.77 and 0.929, respectively.

Following the training phase, the model was temporally validated prior to integration as a web-based prognostic tool. The model showed an accuracy of 88.3% and Mathew's correlation coefficient of 0.72 from this temporal validation [3]. The feature importance analysis showed that human papillomavirus (HPV) status, age of the patients, T stage, marital status, N stage, and the treatment modality (surgery with post-operative radiotherapy) had significant effects on the predictive performance of the model. In addition, Local Interpretable Model Agnostic Explanations (LIME) and SHapley Additive Explanation (SHAP) frameworks were used to examine the effects of each variable on the predicted outputs by the model. As this is a validation study, the details of the model development and performance metrics have been previously published [3].

Due to the absence of totally independent new dataset, the model was validated using a temporal validation method [31]. A temporal validation method lies between an internal and external validation methods. It is considered as the simplest form of external validation where a small cohort of the same data source were recruited for temporal validation [31]. This cohort were neither used for training nor testing [3,31]. Hence, it is the simplest form of external validation and more robust and stronger than internal validation for prediction model reproducibility and generalizability [31,34]. To further demonstrate the viability, predictive and promising performance of the model, this study aims at validating the model with a totally new dataset (external validation) from a new geographic location. The model was trained using data obtained through Surveillance, Epidemiology, and End Results (SEER) program, United States. The external validation dataset was obtained from Helsinki University Hospital, Finland.

2.4. Search protocol and strategy

The search protocol was developed by combining search keywords: [('deep learning' OR 'machine learning') AND ('oropharyngeal')]. This search word was used to query PubMed, OvidMedline, Scopus, and Web of Science to retrieve relevant articles from inception until September 15, 2022. The retrieved articles were systematically reviewed for relevancy in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Fig. 1). The retrieved hits from this search operation were exported to EndNote software for further analysis. To avoid selection bias and reduce research waste, the reference lists of relevant papers were manually searched. In addition, Google scholar and preprints were searched for possible gray literature.

2.5. Study selection, inclusion, and exclusion criteria

Two independent reviewers (R.A. and A.S.) screened the titles and abstracts of potentially relevant articles for initial inclusion. Subsequently, all potential abstracts were subjected to full-text review by these two independent reviewers. Possible disagreements were resolved through discussion and meeting for consensus. We involved a third reviewer (A.A.) to overcome inclusion discrepancies and to foster agreement on the potentially relevant articles.

Studies were included if they met the following inclusion criteria: (1) original studies conducted in the English language and (2) studies that focused on the application of artificial intelligence or its subfield, machine learning for diagnosis and prognostication of outcomes in oropharyngeal cancer. The following were the exclusion criteria: (i) editorials, commentary articles, conference extracts, and narrative reviews were excluded, and (ii) studies on AI/ML in prognostication of head and neck cancer other than oropharyngeal regions, and of non-malignant oral pathologies were excluded.

2.6. Study extraction

A data extraction sheet was used to compile the list of eligible studies. Relevant information relating to the study author, year, country of publication, number of training data, data types, subfield of AI used, algorithm used, performance metrics, and conclusion was extracted from these studies (Table 2). As this review primarily aims at examining how many of the included studies were externally validated and the methodology of external validation of ML models, we specifically extracted those studies where external validation was performed (Table 3).

Table 2
Extracts of the main findings from the included studies.

Authors, Year and Country	No. of training cases (Data types)	AI Subfield (Applications)	Algorithm used	Performance metrics	Conclusions
*Folkert et al., 2017, United States	174(Features extracted from FDG-PET)	Machine learning (Prediction)	LR	For ACM; AUC: 0.65 For LF; AUC: 0.73 For DM; AUC: 0.66	Robust predictive models from radiomics combined with machine learning can predict all-causal mortality (ACM), local failure (LF), and distant metastasis (DM)
Mascharak et al., 2018, United States	30 (mNBI & WLE)	Machine learning (Diagnosis)	NB	For mNBI:(Accuracy: 65.9% & AUC: 0.72). For WLE: (Accuracy: 52.3% & AUC: 0.54)	Automated clinical detection of OPC might be used to enhance surgical vision, improve diagnosis, and allow for high-throughput screening
*Giraud et al., 2020, France	79(Tabular data extracted from dosimetry CT scan)	Machine learning (Prediction)	XGBoost	AUC: 0.68 Accuracy: 0.64	An interpretable and generalizable machine learning model can yield a good precision regarding locoregional recurrence prediction in OPSCC
Suh et al., 2020, South Korea	60(Tabular data extracted from MRI)	Machine learning (Prediction)	LR, RF, XGBoost	AUC: 0.77	This offers a non-invasive approach for the prediction of HPV status in OPS patients
Hatten et al., 2020, United States	3753(Tabular data)	Machine learning (Prediction)	ML methods^{+}, RF, SVM, LR	AUC: 0.58 – 0.68 Accuracy: 0.67 – 0.68	Common oncologic variables contained in the dataset obtained from the NCDB do not reliably predict extracapsular extension
Ren et al., 2020, China	47(Tabular data extracted from CT)	Machine learning (Segmentation)	k-NN, RF	AUC: 0.95	Three-dimensional segmentation had a better CT extraction for predicting HPV status of OPSCC patients than two-dimensional
*Wang et al., 2020, United States	61(^{18}FDG-PET/CT images)	Deep learning (Prediction)	CNN	SUV of $\frac{GTV}{CTV}$: $\frac{3.50}{1.41}$	AI-based deep learning approach can successfully predict image outcomes to treatment with high quantitative accuracy
Ji et al., 2020, United States	64(^{18}FDG-PET/CT images)	Deep learning (Prediction)	CNN	SUV of 2.45 +/- 0.25	The deep learning model combined both pre-radiation images and radiotherapy dose information for outcome prediction
Fujima et al., 2021, United States	154(FDG-PET)	Deep learning (Prediction)	CNN	AUC: 0.85	The model can predict local treatment outcomes in OPSCCs
*Klein et al., 2021, Germany	273CT)	Deep learning (Diagnosis)	CNN	AUC: 0.80	Detection of HPV association in OPSCC patients using deep learning of H&E-stained images helps to identify that can have favourable prognosis
Lang et al., 2021, Germany	850(CT)	Deep learning (Diagnosis)	CNN	AUC: 0.81	Deep learning models are capable of CT image-based HPV status determination
Onoue et al., 2021, United States	173(CT)	Deep learning (Diagnosis)	CNN	Accuracy: 0.76	Deep learning algorithms performed better than two neuroradiologists. Hence, it provides a diagnostic support tool in cervical lymphadenopathy
Paderno et al., 2021, Italy	45(Endoscopic videos images)	Deep learning (Segmentation)	CNN	DSC: 0.76	The CNN have promising potential in analysis and segmentation of OPSCC video-endoscopic images
Haider et al., 2021, Germany	190(Tabular data extracted from PET/CT scans)	Machine learning (Prediction)	RF	c-index: 0.76	The extracted PET/CT radiomic biomarkers can predict post-radiotherapy locoregional progression in HPV-driven OPSCC
Min Park et al., 2021, South Korea	157(Tabular data + MRI radiomic feature)	Machine learning (Prediction)	LR	AUC: 0.79 Accuracy: 85%	A predictive model that combined clinical variables and MRI radiomics feature showed promising predictive performance for disease recurrence and death in OPC patients
Bos et al., 2021, Netherlands	177(Tabular data + MRI radiomic feature)	Machine learning (Prediction)	LR	AUC: 0.75 (for locoregional prediction)AUC: 0.74 (for overall survival prediction)	The model that was used both clinical and radiomic features outperformed the model that either used clinical or radiomic feature for overall survival prediction
Marsden et al., 2021, United States	53(Fiber-based fluorescence lifetime imaging)	Machine learning & deep learning (Diagnosis)	CNN, Support Vector Machine, Random Forest	AUC: 0.88	The application of machine learning approach demonstrates the potential of fiber-based fluorescence lifetime imaging for fast, reliable interoperative margin without the need for contrast agents
Fouad et al., 2021, United Kingdom	2009(Tissue Micro-Arrays)	Deep learning (Segmentation)	CNN	Accuracy: 91%	This approach outperforms the histopathologists in detecting HPV in OPSCC patients
*Cheng et al., 2021, Taiwan	268(PET)	Deep learning (Prediction)	CNN	c-index: 0.78	The deep learning model can provide a fast, objective, and unbiased assessment of OS prognostication of OPSCC
Rodriguez Outeiral et al., 2021, Netherland	171(MRI)	Deep learning (Segmentation)	CNN	Sorensen-Dice coefficient (Dice): 0.74	The semi-automatic approach that involves human approach to define clipboxes around the tumor yielded a better performance from the deep learning model
*Naser et al., 2021, United States	224PET/CT images	Deep learning (Segmentation)	CNN	DSC: 0.771	A deep learning model coupled to label fusion ensembling approaches are promising for auto-segmentation of OP primary tumor
Bos et al., 2022, Netherlands	153	Machine learning (Prediction)	LR	AUC: 0.871	The model based on clinical variables and radiomic feature outperformed others that used either clinical or radiomic. Such model can predict HPV status in OPC patients
Tewari et al., 2022, Ireland	8106(Tabular data)	Machine learning (Prediction)	Bayesian	AUC: 0.70	The model which is based on mathematical modelling approach can estimate the patient's risk of developing OPSCC
Tardini et al., 2022, United States	536 (CT)	Reinforcement learning (Prediction)	Deep Q-learning	Average Accuracy: 87.35	The deep Q-learning can aid the clinician in determining the course of treatment and assessment of outcome

(*continued on next page*)

Table 2 (*continued*)

Authors, Year and Country	No. of training cases (Data types)	AI Subfield (Applications)	Algorithm used	Performance metrics	Conclusions
*Le Greca Saint-Esteven et al., 2022, Switzerland	602(CT)	Deep learning (Diagnosis)	CNN	AUC: 0.84 Accuracy: 0.76	HPV status in CT images could easily be predicted using a 2.5D CNN
Wahid et al., 2022, United States	124(mpMRI)	Deep learning (Segmentation)	CNN	Average DSC: 0.71	The CNN showed comparable performance with the ground truth in terms of the segmentation of OPC primary gross tumor volume
Kim et al., 2022, South Korea	4039(Tabular data)	Machine learning (Prediction)	DeepSurv, EST, CSF	C-index: 0.77	Machine learning models based on personalized survival predictions can be used to stratify various complex risk factors
Dinia et al., 2022, France	450 (Tabular data)	Machine learning (Prediction)	RF	AUC: 0.89	Patients with HPV-driven OPC which are at a high risk of recurrence could be identified for targeted treatment
Park et al., 2022, South Korea	155(Tabular data)	Machine learning (Prediction)	LR LightGBM	AUC: 0.83 (for HPV status)AUC: 0.85 (for disease recurrence)	The machine learning model can predict pathological factors and treatment outcomes of OPSCC patients
Karadaghy et al., 2022, United States	19,111(Tabular data)	Machine learning (Prediction)	RF	AUC: 0.78 Accuracy: 71%	Machine learning model can predict treatment modality. In addition, tumor and facility-related variables impact the decision-making process
Taku et al, 2022, United States	90(CT)	Deep learning (Segmentation)	CNN	DSC:0.92 AUC: 0.98	The deep learning approach provides the automatic segmentation of lymph node for HPV-OPC patients

Abbreviations: - **AUC**: Area Under the Receiver Operating Characteristic Curve; **c-Index**: Concordance Index; **CNN**: Convolutional Neural Network; **CT**: Computed Tomography; **CTV**: Clinical Target Volume; **DSC**: Dice Similarity Coefficient; **GBM**: Gradient Boosting Machine; **GTV**, Gross Tumor Volume; **H&E**: Hematoxylin & Eosin; **HPV**: Human Papillomavirus; **k-NN**: k-Nearest Neighbour; **LR**: Logistic Regression; **mNBI**: multispectral Narrow Band Imaging; **MRI**: Magnetic Resonance Image; **ML**: Machine Learning; **NB**: Naïve Bayes; **OPC**: Oropharyngeal Cancer; **OPSCC**: Oropharyngeal Squamous Cell Carcinoma; **PET**: Positron Emission Tomography; **RF**: Random Forest; **WLE**: White Light Endoscopy; **SUV**: Standardized Uptake Values; **SVM**: Support Vector Machine; ^{18}FDG-PET/CT: [18F] Fluorodeoxyglucose Positron Emission Tomography/Computed Tomography. * Studies that performed external validation (See Table 3 for details).

Table 3
Comparison between the performance of the model when internally validated vs external validation.

Internal validation/training (IV)		Study	External validation (EV)		Remark
Number of cases	Performance metrics	Study details	Number of cases	Performance metrics	Type of EV
174	For ACM; AUC: 0.65 For LF; AUC: 0.73 For DM; AUC: 0.66	**Folkert et al., 2017, United States**	65	For ACM; AUC: 0.60 For LF; AUC: 0.68 For DM; AUC: 0.65	Geographic
79	AUC: 0.68	**Giraud *et al.*, 2020 (France)**	45	Accuracy: 64.0% AUC: 0.79	Temporal
61	SUV of $\frac{GTV}{CTV}$: $\frac{3.57}{1.51}$	**Wang et al., 2020, United States**	5	SUV of $\frac{GTV}{CTV}$: $\frac{3.50}{1.41}$	Temporal
594	AUC: 0.80	**Klein et al., 2021, Germany**	152	AUC: 0.74	Expert
268	c-index: 0.707	**Cheng et al., 2021, Taiwan**	351 31	c-index: 0.689 c-index: 0.787	Geographic
224	DSC: 0.771	**Naser et al., 2021, United States**	101	DSC: 0770	Temporal
602	AUC: 0.84 Accuracy: 0.76	**Le Greca Saint-Esteven et al., 2022, Switzerland**	80 110	AUC:0.83 Accuracy:0.75 AUC:0.88 Accuracy:0.79	Geographic

2.7. Quality assessment of included studies

We examined the risk of bias and applicability of studies where external validation was performed (Table 3) using the Prediction model Risk Of Bias Assessment Tool (PROBAST) [Table 4]. Further quality assessment of these studies was completed using the guidelines for developing and reporting ML predictive models in biomedical research [35]. The guideline was summarized as previously reported in other studies where a mark was given for each of the guideline topics (Supplementary II) [9,36]. The threshold was set at half of the maximum marks, and the score was presented in Table 5.

3. Results

3.1. Characteristics of the study population for validation study

The detailed characteristics of the external validation data ($n = 163$) are presented in Table 1. The mean age at diagnosis was 61.4 years (Majority [99,4%] were ≥40 while [0.6%] were <than 40). Additionally, the median age was 62 years; SD ± 9.1; range 37–85 with 123 (75.5%) male and 40 (24.5%) female. In terms of the clinical and pathologic characteristics such as grade, 4 (2.4%) out of the 163 OPSCC patients had a well-differentiated grade, 28 (17.2%) moderately differentiated, and 131 (80.4%) poorly differentiated. Regarding the tumor HPV status of these patients, a total of 42 (25.8%) had a HPV-negative while 121 (74.2%) had a HPV-positive tumor.

Table 4
Tabular presentation of PROBAST results.

Study	ROB				Applicability				Overall
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	ROB	Applicability
Folkert et al., 2017	+	+	+	+	+	+	+	+	+
Giraud *et al.*, 2020	+	+	+	+	+	+	+	+	+
Wang et al., 2020	+	+	+	+	+	+	+	+	+
Klein et al., 2021	+	+	+	+	+	+	+	+	+
Cheng et al., 2021	+	+	+	+	+	+	+	+	+
Naser et al., 2021	+	+	+	+	+	+	+	+	+
Le Greca Saint-Esteven et al., 2022	+	+	+	+	+	+	+	+	+

PROBAST = Prediction model Risk Of Bias Assessment Tool; ROB = Risk of Bias.
+ Indicates Low ROB/Low concern regarding applicability.
− Indicates High ROB/high concern regarding applicability.
? Indicates unclear ROB/unclear concern regarding applicability.

Table 5
Quality scores of studies that included external validation.

Studies	Title and abstract		Introduction		Method				Result	Discussion		Quality
	Title	Abstract	Rationale	Objectives	Setting description	Problem definition	Data preparation	Build model	Report performance	Clinical implications	Limitation	Score (%)
Folkert et al., 2017	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100
Giraud *et al.*, 2020	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✘	90.9
Wang et al., 2020	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100
Klein et al., 2021	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✘	90.9
Cheng et al., 2021	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100
Naser et al., 2021	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100
Le Greca Saint-Esteven et al., 2022	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100

A total of 46 (28.2%) tumors originated from the base of the tongue while the remaining 108 (71.8%) cases were from the tonsil. Regarding the staging scheme according to the AJCC TNM, 53 (32.5%) patients had stage T1, 53 (32.5%) stage T2, 17 (10.4%) stage T3, and 40 (24.5%) stage T4. Similarly, 25 (15.3%) had N0, 16 (9.8%) had N1, 120 (73.6%) N2, and 2 (1.2%) N3. Furthermore, 161 (98.8%) M0, and 2 (1.2%) M1. The details of the histopathologic characteristics and the corresponding distributions are given in Table 1.

Considering the treatment modalities in the external validation cohort, 23 (14.1%) had surgery with postoperative radiotherapy, 35 (21.5%) surgery with chemoradiotherapy, 84 (51.5%) definitive chemotherapy, 8 (4.9%) surgery alone and 13 (8.0%) received none of the available treatments. The follow-up time ranged from 0 to 88 months (Mean 45.8; Median 50.0; SD ± 19.9). The number of patients who were

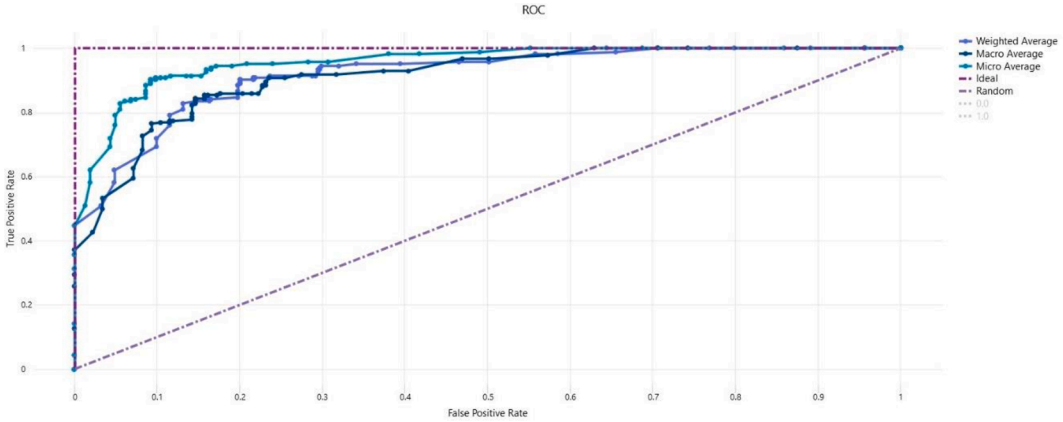


Fig. 2. The area under receiving operating characteristics curve for voting ensemble method during validation study.

alive at last follow-up was 118 (72.4%).

3.2. Performance metrics from the validation study

The web-based tool produced an accuracy of 90.2%, Mathew's correlation coefficient of 0.78 and F1 score of 0.84. The model showed comparable performance from the temporal and external validations. The weighted area under receiving operating characteristic curve was 0.94 (Fig. 2). In terms of predictive values, positive predictive value (precision) was 0.93 while negative predictive value was 0.89. The specificity (recall) and sensitivity were 0.97 and 0.76, respectively. A similar result was obtained when the patients were considered single (in terms of marital status). Other performance metrics relating to the confusion matrix and the predictive rates are given in Table 6. Furthermore, the sensitivity and specificity for 1-, 3-, and 5-year specific time points were 1.00 and 0.50; 0.97 and 0.40; 0.74 and 0.96, respectively (Table 7). The calibration curve for this validation study is presented as Supplementary III.

The calculated Brier score was 0.06 (Brier score is usually between 0.0 and 1.0 where 0.0 indicates perfect model and 1.0 means worst model). Similarly, the Net Benefit value of the model was approximately 0.7 (i.e., acceptable) at 10% − 50% probability threshold.

3.3. Results of the database search

A total of 549 hits were retrieved. After deleting duplicates (N = 245), irrelevant papers (N = 256), and exclusions (N = 20), we found 31 studies eligible to be included in this study as shown in Fig. 1 [23,37–66]. The inter-observer reliability between the independent reviewers for data extraction showed a Cohen's Kappa coefficient (κ = 0.90) (Table 2).

3.4. Characteristics of studies that examined application of machine in oropharyngeal cancer

Of the 31 included studies found to have used ML or its subfield in oropharyngeal cancer, 13 (41.9%) studies had been carried out in the United States [37,38,41,43–45,48,52,56,59,61,65,66], 12 (38.7%) were conducted in Europe, [23,39,46,47,49,50,53,55,57,58,60,63], and 6 (19.4%) in Asia [40,42,51,54,62,64]. The application of artificial intelligence or its subfield for proper management of OPSCC broadly included segmentation, diagnosis, and detection of HPV status [38,40,42,46,47,49,50,52,53,55–58,60,61,66], prediction of treatment outcome [41,43–45,48,59,64,65], local failure, recurrence or distance metastasis [39,50,51,63], and prediction of overall survival [23,37,54,62].

Table 6
Geographical external validation for generalizability (n = 163 cases).

	Performance metrics	ProgTOOL
Confusion matrix parameters	True positive	42
	False positive	3
	False negative	13
	True negative	105
Predictive value	PPV (Precision)	0.93
	NPV	0.89
Rate	False positive rate	0.03
	False negative rate	0.24
Other metrics	Sensitivity (recall)	0.76
	Specificity	0.97
	F1 score	0.84
Accuracy	Accuracy	90.2%
	Balanced accuracy	86.5%
Correlation	Mathew's correlation coefficient	0.78

PPV: Positive predictive value; NPV: Negative predictive value; AUC: Area under curve; ProgTOOL: A web-based prognostic tool for risk stratification of oropharyngeal cancer patients.

3.5. Summary of studies that performed external validation dataset

Out of the 31 studies found eligible (sub-section 3.3) [Table 2], only 7 (22.5%) studies performed external validation (Table 3) [37,39,43,46,54,56,60]. These were the studies that were further evaluated and discussed in our analysis. From the mentioned studies, in terms of ML application, 3 (42.9%) studies emphasized the potential of these models in diagnosis [46,56,60] while 4 (57.1%) studies reported the prognostication abilities of these ML-based models [37,39,43,54]. Regarding the types of external validation, 3 (42.9%) studies each used temporal external validation [39,43,56] or geographic external validation [37,54,60] while a single study (14.3%) employed expert opinion as a form of external validation to examine the potential of the model in terms of generalizability [46].

3.6. Importance of external validation

The need for external validation of the ML models was emphasized in few of the included studies [40–42,50,51,55,66]. Despite the promising results from internal validation of these models, it was suggested that external validation of the trained model can enhance transportability (ability of the ML model to produce comparable performance when tested with an independent dataset that is different from the dataset that it was trained upon) [56,61]. In addition, it was equally observed that reproducibility and generalizability of the ML model can be guaranteed through external validation [40–42,50,51,55,66].

3.7. Quality assessment

According to the PROBAST assessment, all of the studies that examined external validation showed an overall low risk of bias and applicability concerns. Similarly, the quality assessment showed that 5 (71.4%) studies had extremely high quality [37,43,54,56,60] while 2 (28.6%) showed high quality [39,46]. This high-quality assessment indicates that these studies are well posited to answer the research questions in this study.

4. Discussion

This study highlights the importance of a thorough and independent external validation of a machine learning model that has been integrated as a web-based prognostic tool to ensure the model's generalizability and to facilitate its clinical evaluation. The important step of external validation is necessary to demonstrate the readiness of such models from developmental stages to real-world evaluation using a completely different dataset. Several studies have emphasized the significance of validation studies through external validation prior to clinical evaluation to guarantee reproducibility and generalizability [9,15,31]. Traditionally, a model's performance is usually reported based on internal validation. However, it has been reported that making judgments on the potential of the performance of ML-models in the real-world setting based solely on internal validation of underlying data may potentially report an over-optimistic, misleading, and unrealistic expectation of the model's accuracy [31]. Consequently, having implications for clinical applications.

The external validation of a web-based prognostic tool examined in this study demonstrated that the model showed extremely promising performance in its ability to stratify OPSCC patients into risk groups in terms of their chance of overall survival. The model showed a higher performance accuracy, AUC, Mathew's correlation coefficient, and Brier score from the external validation compared to the temporal validation. Presently, there is arguably no standardized approach for external validation procedure. For example, the study by Ramspek et al. emphasized the use of geographic external validation dataset to ensure a structurally different cohort from the development cohort [31]. Cabitza et al. proposed a two-step (*meta*-validation) approach for evaluating an

Table 7
Overall survival at specific time points.

Time points	Sensitivity	Specificity	Precision	Mathew Correlation Coefficient	F1 score	Accuracy
1 year (0–12 months)	1.00	0.50	0.95	0.69	0.97	0.95
3 years (0–36 months)	0.97	0.40	0.91	0.46	0.94	0.89
5 years (0–60 months)	0.74	0.96	0.93	0.75	0.84	0.88
All cases combined	0.76	0.97	0.93	0.78	0.84	0.90

external validation process [67]. In the first step involves building a linear regression model over the training dataset. However, in our study, we externally validated an already developed and web-based integrated model using a structurally different geographic dataset recommended by Ramspek et al. [31]. Additionally, the model may not be suitable for a linear regression since it was aimed at a classification task. In the second step, data similarity and cardinality were evaluated for discrimination, calibration, and utility using AUC, Net Benefit, and Brier score.

Therefore, we combined the suggestions of Ramspek et al. to ensure that we used a structurally different (i.e., geographic external validation) to cater for dataset similarity while we examined the performance of the external validation procedure over arrays of evaluation metrics, specifically AUC, Net Benefit, and Brier score for the model's discriminating power, utility, and calibration abilities for the overall survival risk stratification of OPSCC patients. Thus, the external validation procedure showed that the model is excellent based on AUC and Brier score and may be acceptable based on the Net Benefit score.

With such a potential model, information regarding the estimate of overall survival of OPSCC patients can help guide the clinician for clinical decision-making. For example, clinicians can carefully examine the patients stratified as high chance for poor survival for individualized treatment planning. Remarkably, the initial model integrated as web-based tool was trained using a dataset from one of the largest publicly available databases managed by the National Cancer Institute (NCI) through the Surveillance, Epidemiology, and End Results (SEER) Program of the National Institutes of Health (NIH). The validation study was done using data obtained through the Helsinki University Hospital, Finland. Therefore, the dataset used for this validation study was from a different geographical location. Such an approach is posited to address any concern relating to the generalizability of this web-based tool. Consequently, positioning the web-based tool for further validation studies from other geographical locations prior is important prior to recommending it for clinical evaluation.

Remarkably, for every potential ML model to be transferred for clinical evaluation, two important characteristics should be present. These are model's reproducibility (validity) and generalizability (transportability) [68]. These two important characteristics serves as important cornerstone for the viability of a ML model in clinical practice. While reproducibility ensures that the model shows considerably similar performance when tested with patients similar to the development population, generalizability explores whether the model performs as expected when exposed to a separate population with a relatively similar patient characteristics to the development population. External validation may be used to achieve both characteristics.

However, as shown in this study, only a few studies have performed external validation of their reported models for outcomes prognostication in OPSCC. This raises the concern of generalizability as it is unclear how these models would perform in a newly collected, unseen, and balanced external validation data. The lack of external validation practices in these published studies may have limited the translation of these models from development environments to clinical consideration and evaluation. Without adequate and standard clinical evaluation of these models, the path to implementation in daily clinical practice remains vague. To the best of our knowledge, this may justify why none of these models are presently used in daily clinical practices for OPSCC management.

It is a common practice to validate ML models at the technical level using any of the widely accepted internal validation methods such as cross-validation (k-fold cross-validation or Monte Carlo cross-validation), bootstrapping, or hold-out method (training, validation ± testing). This makes internal validation the most widely used approach because it is tuneable and highly heterogeneous [69,70]. However, the approach of using internal validation has been criticized because it involves significant parameter tuning. In addition, if the original input data are biased, the internal validation approach becomes a biased evaluation [69]. It is not uncommon for a ML model to show promising results during training and perform poorly when externally validated [18,34,71,72]. Several reasons such as underfitting, overfitting, or imbalanced dataset may be attributed to this. To overcome this concern, external validation is warranted.

Conducting an evaluation process such as an external validation study will allow stakeholders to assess and understand the true performance of these models. Validation of a standalone model or of a model integrated as a web-based tool as demonstrated in this study, is one of the crucial steps to secure generalizability prior to clinical evaluation and consequently, to follow other recommended paths to implementation into clinical practice. External validation, also known as a standalone test, is thought to be the approach of testing the predictive performance of a trained ML model with an entirely different set of new patients in order to ascertain that the model works as expected [31].

4.1. Types of external validation

As shown in this study, three forms of external validation approach have been used apart from the traditional internal validation methods. These were temporal, geographical, and expert external validations (Table 3, Fig. 3). In temporal external validation, a certain portion of the same data is reserved [73]. This reserved data is neither used in the training, validation, nor testing. Such reserved data are usually used to evaluate the performance of the trained ML model [73]. The reservation can be done randomly by reserving a certain set of the same data from the first, middle, or last portions. In some cases, cohorts with a certain year within the same data may be reserved for temporal external validation.

Considering the daunting process involved in setting up standard external validation procedures (new set of data and model tested by different researchers) [74], temporal validation approach may be posited as a viable validation process for predicting model reproducibility and generalizability [74]. It is considered the simplest form of external validation, which is more robust and stronger than internal validation [34] even though the subsequent cohorts used for temporal validation were recruited from the same data source [73]. This type of external validation has been criticized for several reasons such as the fact that the patients are not structurally different [34,75]. In addition, the process seems to fall between internal and external validation approaches [75,76] (Fig. 3).

Therefore, geographical external validation is posited to address the concerns of using temporal validation. Geographical external validation ensures that the developed model is validated with a dataset that is structurally different from the development cohorts as demonstrated in this study [71]. These differences may be in terms of region, country, type of care, or treatment protocols [75,76]. The use of a geographical or totally independent dataset from another geographic location has been considered reliable [74]. Despite this, issues relating to the ethical

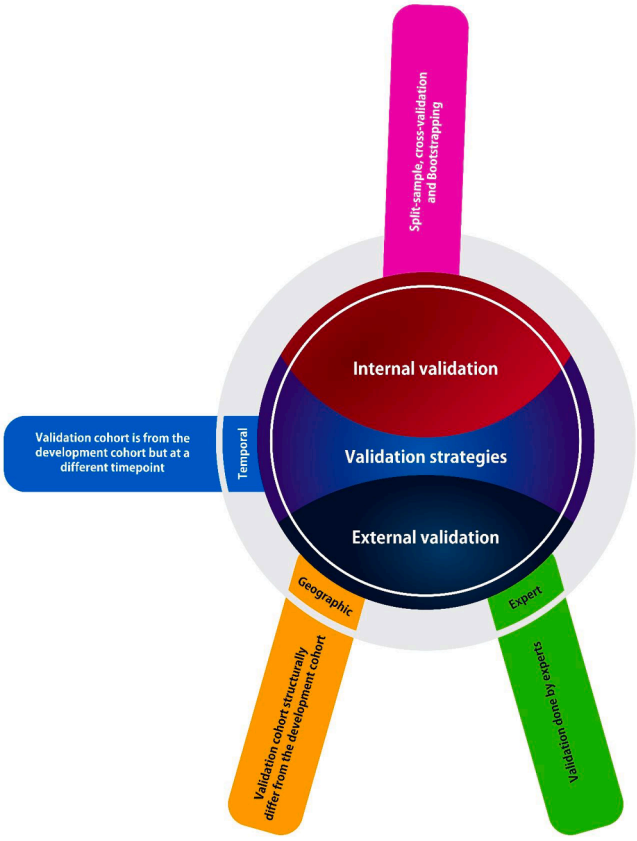


Fig. 3. The various forms of external validation.

concerns regarding the transfer or sharing of data, lack of external validation dataset, and unavailability of the model for independent external validation are some of the reasons for the dearth of external validation. Due to these factors, experts have been saddled with the responsibility of monitoring the performance of these models as a form of external validation. The concern with this approach is that these models are supposed to provide a second opinion to the experts, not otherwise. Hence, geographical external validation remains the widely favored approach [74].

To summarize, either internal validation or temporal validation should not be misconstrued as an adequate form of independent external validation [31]. If a model exhibits comparable performance in rigorous external validation, it may be a strong indication that the model works as expected and may be ready for further clinical evaluation by experts (Fig. 4). Similarly, a clearly low performance when externally validated may indicate that the model may not be generalized (Fig. 4).

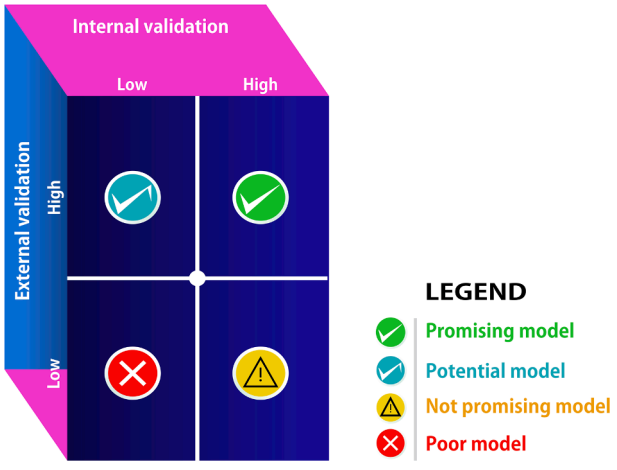


Fig. 4. The model performance box for internal and external validations.

4.2. Concerns relating to external validation

The modalities (how, who, number of cases, effectiveness, and evaluation) to perform external validation are sources of significant concern. For example, as shown in the previously published studies (Table 3), external validation was done by the developer of the model. There is a concern that the developer may fine-tune the model during external validation (i.e. the potential to introduce biasedness in the process of external validation) [18,71,77]. There is growing debate that external validation should be done by separate researchers [77]. The main concern is that how many of these developed models are publicly available in other to be externally validated by independent researchers? This may justify why very few of these models have been externally validated by different authors [18].

Considering these concerns, we would like to emphasize that an adequate framework of external validation may be warranted. This framework should define important parameters such as sample size for external validation, methodological caveats, minimum acceptable differences between the performance of the model during internal and external validations, and reporting standards for the validation process. Concerted efforts are required from different stakeholders to modify the framework to bridge the gap between the development and implementation of prediction models in daily clinical practices.

Of note, it has been discussed whether or not to include the process of external validation in routine ML model development pipelines [71,77]. External validation is thought to be sequel to internal validation as it addresses transportability (use in a different environment from which the model was initially developed) and generalizability (real-world performance), rather than reproducibility [78]. Therefore, considering the ethical and legal concerns relating to data sharing, we opined that the process of external validation should not be integrated as part of the ML development pipeline. This is because data can't be freely shared for geographical external validation due to ethical permission relating to the use of data. However, it may be feasible to include temporal validation as part of the routine ML development process (Fig. 4). Integrating the model as a web-based tool [21] or sending it as a standalone package seeks to address relating to sharing of data and to facilitate external validation.

Remarkably, the fact that a model has been externally validated does not imply straight acceptance in daily clinical practice. Other important aspects should be considered prior to the recommendation of the model for clinical evaluation. For example, the quest for standardized reporting for ML-models in medical oncology. Recently published studies have standardized guideline for reporting AI or ML-based models [34,79,80]. Furthermore, the performance metrics should be standardized such that an array of different metrics should be used to demonstrate the performance of these models.

4.3. Limitations, conclusions, and future research

There are limitations to be considered in this study. First, there was an imbalance in the target variable of overall survival in the external validation data. This may present certain level of model bias. Furthermore, the web-based tool was developed by our group although externally validated with a dataset from a different geographic location as shown in this study. Therefore, this validation study was not completely independent even though it fulfilled the requirements for external validation. In addition, the web-based tool was validated with a relatively small number of cases.

In conclusion, it is a good practice to repeatedly evaluate the trained model with an external geographic validation approach in tandem with internal validation that must have already been performed. If the model

persistently shows good performance after being repeatedly validated externally, this may imply that the model is robust, generalizable, and ready for clinical evaluation. Therefore, we recommend that the developed models should undergo external validation, followed by clinical evaluation, and randomized comparative impact assessment. This carries the potential to increase the likelihood of the utilization of these models in daily clinical practices in the future.

For future research, it is necessary to increase external geographic validation by other researchers to further guarantee an independent external validation and to further assess the generalizability of the model. Similarly, a defined path to the implementation of the ML model should be stated by stakeholders to enable the transition of these models from development to bedside. To enhance model maintenance, performance improvement and updating, we propose that the integration of ML model as a web-based tool should follow modern approaches such as the federated learning and particle swarm optimization paradigms that can preserve data sharing and privacy. This will ensure that the model's quality and performance is continuously improved without any data privacy concerns.

Summary points.

- A publicly available ML web-based prognostic tool (ProgTOOL) was externally validated using a geographic dataset for survival risk stratification of oropharyngeal squamous cell carcinoma (OPSCC).
- A considerable number of published studies on the application of ML models for OPSCC outcome prognostication were not externally validated.
- External validation of ML models has been traditionally done using temporal, expert, or geographic external validation paradigms.
- Geographical external validation was considered the most reliable form of external validation.
- External validation approach through a validation study is important for model generalization and their realization for clinical evaluation and consideration.

Code availability and daily clinical use of the model (web-based tool).

This study is a validation study of a model that has been integrated as a web-based tool that is freely, openly, and publicly available as a web-based tool. Thus, code availability is not needed. Additionally, the web-based tool is not presently used in daily clinical practice. However, detailed description of the development of the mode is available in our previous study [3]. This study also highlights some of the requirements that can facilitate the implementation of ML models in clinical practice. The ProgTOOL will be hosted freely for one year (until the end of 2023). It is intended for research prurpose only.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

K. Albin Johanssons Stiftelse. Sigrid Jusélius Foundation. Helsinki University Hospital Research Fund. Turku University Hospital Research Fund.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105064.

References

[1] V. Lorenzoni, A.K. Chaturvedi, J. Vignat, M. Laversanne, F. Bray, S. Vaccarella, The current burden of oropharyngeal cancer: a global assessment based on GLOBOCAN 2020, Cancer Epidemiol. Biomark. Prevent. 31 (2022) 2054–2062. <https://doi.org/10.1158/1055-9965.EPI-22-0642>.

[2] Anni Sjöblom, Ulf-Håkan Stenman, Jaana Hagström, Lauri Jouhi, Caj Haglund, Stina Syrjänen, Petri Mattila, Antti Mäkitie, Timo Carpén, Tumor-Associated Trypsin Inhibitor (TATI) as a biomarker of poor prognosis in oropharyngeal squamous cell carcinoma irrespective of HPV status, Cancers 13 (11) (2021) 2811.

[3] R. Alabi, A. Almangush, M. Elmusrati, I. Leivo, A.A. Mäkitie, An interpretable machine learning prognostic system for risk stratification in oropharyngeal cancer, Int. J. Med. Inf. 168 (2022), 104896, https://doi.org/10.1016/j.ijmedinf.2022.104896.

[4] A.A.K. Abdel Razek, M. Mansour, E. Kamal, S.K. Mukherji, MR imaging of oral cavity and oropharyngeal cancer, Magn. Reson. Imaging Clin. N. Am. 30 (2022) 35–51, https://doi.org/10.1016/j.mric.2021.07.002.

[5] Haluk Damgacioglu, Kalyani Sonawane, Yenan Zhu, Ruosha Li, Bijal A. Balasubramanian, David R. Lairson, Anna R. Giuliano, Ashish A. Deshmukh, Oropharyngeal cancer incidence and mortality trends in all 50 States in the US, 2001–2017, JAMA Otolaryngol. Head Neck Surg. 148 (2) (2022) 155.

[6] E.L. You, M. Henry, A.G. Zeitouni, Human papillomavirus-associated oropharyngeal cancer: review of current evidence and management, Curr. Oncol. 26 (2019) 119–123, https://doi.org/10.3747/co.26.4819.

[7] P. Rajpurkar, E. Chen, O. Banerjee, E.J. Topol, AI in health and medicine, Nat. Med. 28 (2022) 31–38, https://doi.org/10.1038/s41591-021-01614-0.

[8] Isaac S. Chua, Michal Gaziel-Yablowitz, Zfania T. Korach, Kenneth L. Kehl, Nathan A. Levitan, Yull E. Arriaga, Gretchen P. Jackson, David W. Bates, Michael Hassett, Artificial intelligence in oncology: path to implementation, Cancer Med. 10 (12) (2021) 4138–4149.

[9] Rasheed Omobolaji Alabi, Omar Youssef, Matti Pirinen, Mohammed Elmusrati, Antti A. Mäkitie, Ilmo Leivo, Alhadi Almangush, Machine learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for future—a systematic review, Artif. Intell. Med. 115 (2021) 102060.

[10] Andrés M. Bur, Andrew Holcomb, Sara Goodwin, Janet Woodroof, Omar Karadaghy, Yelizaveta Shnayder, Kiran Kakarala, Jason Brant, Matthew Shew, Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma, Oral Oncol. 92 (2019) 20–25.

[11] M.K. Yu, J. Ma, J. Fisher, J.F. Kreisberg, B.J. Raphael, T. Ideker, Visible machine learning for biomedicine, Cell 173 (2018) 1562–1565, https://doi.org/10.1016/j.cell.2018.05.056.

[12] Bert Heinrichs, Simon B. Eickhoff, Your evidence? Machine learning algorithms for medical diagnosis and prediction, Hum. Brain Mapp. 41 (6) (2020) 1435–1444.

[13] Rasheed Omobolaji Alabi, Mohammed Elmusrati, Iris Sawazaki-Calone, Luiz Paulo Kowalski, Caj Haglund, Ricardo D. Coletta, Antti A. Mäkitie, Tuula Salo, Ilmo Leivo, Alhadi Almangush, Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool, Virchows Arch. 475 (4) (2019) 489–497.

[14] Rasheed Omobolaji Alabi, Mohammed Elmusrati, Iris Sawazaki-Calone, Luiz Paulo Kowalski, Caj Haglund, Ricardo D. Coletta, Antti A. Mäkitie, Tuula Salo, Alhadi Almangush, Ilmo Leivo, Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer, Int. J. Med. Inf. 136 (2020) 104068.

[15] R.O. Alabi, A. Almangush, M. Elmusrati, I. Leivo, A. Mäkitie, Measuring the usability and quality of explanations of a machine learning web-based tool for oral tongue cancer prognostication, IJERPH 19 (2022) 8366, https://doi.org/10.3390/ijerph19148366.

[16] Luisa Oliveira e Carmo, Anke van den Merkhof, Jakub Olczak, Max Gordon, Paul C. Jutte, Ruurd L. Jaarsma, Frank F.A. IJpma, Job N. Doornberg, Jasper Prijs, An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics: are these externally validated and ready for clinical application? Bone Joint Open 2 (10) (2021) 879–885.

[17] Michiel E.R. Bongers, Quirina C.B.S. Thio, Aditya V. Karhade, Merel L. Stor, Kevin A. Raskin, Santiago A. Lozano Calderon, Thomas F. DeLaney, Marco L. Ferrone, Joseph H. Schwab, Does the SORG algorithm predict 5-year survival in patients with chondrosarcoma? an external validation, Clin. Orthop. Relat. Res. 477 (10) (2019) 2296–2303.

[18] G.C.M. Siontis, I. Tzoulaki, P.J. Castaldi, J.P.A. Ioannidis, External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination, J. Clin. Epidemiol. 68 (2015) 25–34, https://doi.org/10.1016/j.jclinepi.2014.09.007.

[19] Kao-Lang Liu, Tinghui Wu, Po-Ting Chen, Yuhsiang M Tsai, Holger Roth, Ming-Shiang Wu, Wei-Chih Liao, Weichung Wang, Deep learning to distinguish pancreatic cancer tissue from non-cancerous pancreatic tissue: a retrospective study with cross-racial external validation, Lancet Dig. Health 2 (6) (2020) e303–e313.

[20] Arkadiusz Gertych, Zaneta Swiderska-Chadaj, Zhaoxuan Ma, Nathan Ing, Tomasz Markiewicz, Szczepan Cierniak, Hootan Salemi, Samuel Guzman, Ann E. Walts, Beatrice S. Knudsen, Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides, Sci. Rep. 9 (1) (2019), https://doi.org/10.1038/s41598-018-37638-9.

[21] Alhadi Almangush, Rasheed Omobolaji Alabi, Antti A. Mäkitie, Ilmo Leivo, Machine learning in head and neck cancer: Importance of a web-based prognostic tool for improved decision making, Oral Oncol. 124 (2022) 105452.

[22] R.O. Alabi, V. Tero, E. Mohammed, Machine learning for prognosis of oral cancer: what are the ethical challenges? CEUR-Workshop Proc. (2020).

[23] Paula Bos, Michiel W.M. van den Brekel, Zeno A.R. Gouw, Abrahim Al-Mamgani, Marjaneh Taghavi, Selam Waktola, Hugo J.W.L. Aerts, Jonas A. Castelijns, Regina G.H. Beets-Tan, Bas Jasperse, Improved outcome prediction of oropharyngeal cancer by combining clinical and MRI features in machine learning models, Eur. J. Radiol. 139 (2021) 109701.
[24] F. Cabitza, A. Campagner, The need to separate the wheat from the chaff in medical informatics, Int. J. Med. Inf. 153 (2021), 104510, https://doi.org/10.1016/j.ijmedinf.2021.104510.
[25] Z. Ding, D. Yu, H. Li, Y. Ding, Effects of marital status on overall and cancer-specific survival in laryngeal cancer patients: a population-based study, Sci. Rep. 11 (2021) 723, https://doi.org/10.1038/s41598-020-80698-z.
[26] Cheng Xu, Xu Liu, Yu-Pei Chen, Yan-Ping Mao, Rui Guo, Guan-Qun Zhou, Ling-Long Tang, Ai-Hua Lin, Ying Sun, Jun Ma, Impact of marital status at diagnosis on survival and its change over time between 1973 and 2012 in patients with nasopharyngeal carcinoma: a propensity score-matched analysis, Cancer Med. 6 (12) (2017) 3040–3051.
[27] G. Inverso, B.A. Mahal, A.A. Aizer, R.B. Donoff, N.G. Chau, R.I. Haddad, Marital status and head and neck cancer outcomes: Marital Status and Head and Neck Cancer, Cancer 121 (2015) 1273–1278, https://doi.org/10.1002/cncr.29171.
[28] Ayal A. Aizer, Ming-Hui Chen, Ellen P. McCarthy, Mallika L. Mendu, Sophia Koo, Tyler J. Wilhite, Powell L. Graham, Toni K. Choueiri, Karen E. Hoffman, Neil E. Martin, Jim C. Hu, Paul L. Nguyen, Marital status and survival in patients with cancer, JCO 31 (31) (2013) 3869–3876.
[29] E.W. Schaefer, M.Z. Wilson, D. Goldenberg, H. Mackley, W. Koch, C.S. Hollenbeak, Effect of marriage on outcomes for elderly patients with head and neck cancer: marriage effect in head and neck cancer, Head Neck 37 (2015) 735–742, https://doi.org/10.1002/hed.23657.
[30] S. El Ibrahimi, P.S. Pinheiro, The effect of marriage on stage at diagnosis and survival in women with cervical cancer: marriage and cervical cancer stage and survival, Psychooncology 26 (2017) 704–710, https://doi.org/10.1002/pon.4070.
[31] C.L. Ramspek, K.J. Jager, F.W. Dekker, C. Zoccali, M. van Diepen, External validation of prognostic models: what, why, how, when and where? Clin. Kidney J. 14 (2021) 49–58, https://doi.org/10.1093/ckj/sfaa188.
[32] A.J. Vickers, B. Van Calster, E.W. Steyerberg, Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests, BMJ (2016), https://doi.org/10.1136/bmj.i6.
[33] A.J. Vickers, E.B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, Med. Decis. Making 26 (2006) 565–574, https://doi.org/10.1177/0272989X06295361.
[34] Karel G.M. Moons, Douglas G. Altman, Johannes B. Reitsma, John P.A. Ioannidis, Petra Macaskill, Ewout W. Steyerberg, Andrew J. Vickers, David F. Ransohoff, Gary S. Collins, Transparent reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration, Ann. Intern. Med. 162 (1) (2015) W1–W73.
[35] Wei Luo, Dinh Phung, Truyen Tran, Sunil Gupta, Santu Rana, Chandan Karmakar, Alistair Shilton, John Yearwood, Nevenka Dimitrova, Tu Bao Ho, Svetha Venkatesh, Michael Berk, Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view, J. Med. Internet Res. 18 (12) (2016) e323.
[36] Wai Tong Ng, Barton But, Horace CW Choi, Remco de Bree, Anne WM Lee, Victor HF Lee, Fernando López, Antti A Mäkitie, Juan P Rodrigo, Nabil F Saba, Raymond KY Tsang, Alfio Ferlito, Application of artificial intelligence for nasopharyngeal carcinoma management – a systematic review, CMAR 14 (2022) 339–366.
[37] Michael R Folkert, Jeremy Setton, Aditya P Apte, Milan Grkovski, Robert J Young, Heiko Schöder, Wade L Thorstad, Nancy Y Lee, Joseph O Deasy, Jung Hun Oh, Predictive modeling of outcomes following definitive chemoradiotherapy for oropharyngeal cancer based on FDG-PET image characteristics, Phys. Med. Biol. 62 (13) (2017) 5327–5343.
[38] S. Mascharak, B.J. Baird, F.C. Holsinger, Detecting oropharyngeal carcinoma using multispectral, narrow-band imaging and machine learning: Multispectral Imaging of Oropharynx Cancer, Laryngoscope 128 (2018) 2514–2520, https://doi.org/10.1002/lary.27159.
[39] Paul Giraud, Philippe Giraud, Eliot Nicolas, Pierre Boisselier, Marc Alfonsi, Michel Rives, Etienne Bardet, Valentin Calugaru, Georges Noel, Enrique Chajon, Pascal Pommier, Magali Morelle, Lionel Perrier, Xavier Liem, Anita Burgun, Jean Emmanuel Bibault, Interpretable machine learning model for locoregional relapse prediction in oropharyngeal cancers, Cancers 13 (1) (2020) 57.
[40] Chong Hyun Suh, Kyung Hwa Lee, Young Jun Choi, Sae Rom Chung, Jung Hwan Baek, Jeong Hyun Lee, Jihye Yun, Sungwon Ham, Namkug Kim, Oropharyngeal squamous cell carcinoma: radiomic machine-learning classifiers from multiparametric MR images for determination of HPV infection status, Sci. Rep. 10 (1) (2020), https://doi.org/10.1038/s41598-020-74479-x.
[41] K.M. Hatten, J. Amin, A. Isaiah, Machine learning prediction of extracapsular extension in human papillomavirus-associated oropharyngeal squamous cell carcinoma, Otolaryngol. Head Neck Surg. 163 (2020) 992–999, https://doi.org/10.1177/0194599820935446.
[42] J. Ren, Y. Yuan, M. Qi, X. Tao, Machine learning-based CT texture analysis to predict HPV status in oropharyngeal squamous cell carcinoma: comparison of 2D and 3D segmentation, Eur. Radiol. 30 (2020) 6858–6866, https://doi.org/10.1007/s00330-020-07011-4.
[43] Chunhao Wang, Chenyang Liu, Yushi Chang, Kyle Lafata, Yunfeng Cui, Jiahan Zhang, Yang Sheng, Yvonne Mowery, David Brizel, Fang-Fang Yin, Dose-Distribution-Driven PET Image-Based Outcome Prediction (DDD-PIOP): a deep learning study for oropharyngeal cancer IMRT application, Front. Oncol. 10 (2020), https://doi.org/10.3389/fonc.2020.01592.
[44] H. Ji, K. Lafata, Y. Mowery, D. Brizel, A.L. Bertozzi, F.-F. Yin, et al., Post-radiotherapy PET image outcome prediction by deep learning under biological model guidance: a feasibility study of oropharyngeal cancer application, Front. Oncol. 12 (2022), 895544, https://doi.org/10.3389/fonc.2022.895544.
[45] Noriyuki Fujima, V. Carlota Andreu-Arasa, Sara K. Meibom, Gustavo A. Mercier, Minh Tam Truong, Kenji Hirata, Koichi Yasuda, Satoshi Kano, Akihiro Homma, Kohsuke Kudo, Osamu Sakai, Prediction of the local treatment outcome in patients with oropharyngeal squamous cell carcinoma using deep learning analysis of pretreatment FDG-PET images, BMC Cancer 21 (1) (2021), https://doi.org/10.1186/s12885-021-08599-6.
[46] S. Klein, A. Quaas, J. Quantius, H. Löser, J. Meinel, M. Peifer, et al., Deep learning predicts HPV association in oropharyngeal squamous cell carcinomas and identifies patients with a favorable prognosis using regular H&E stains, Clin. Cancer Res. 27 (2021) 1131–1138, https://doi.org/10.1158/1078-0432.CCR-20-3596.
[47] D.M. Lang, J.C. Peeken, S.E. Combs, J.J. Wilkens, S. Bartzsch, Deep learning based HPV status prediction for oropharyngeal cancer patients, Cancers 13 (2021) 786, https://doi.org/10.3390/cancers13040786.
[48] Keita Onoue, Noriyuki Fujima, V. Carlota Andreu-Arasa, Bindu N. Setty, Osamu Sakai, Cystic cervical lymph nodes of papillary thyroid carcinoma, tuberculosis and human papillomavirus positive oropharyngeal squamous cell carcinoma: utility of deep learning in their differentiation on CT, Am. J. Otolaryngol. 42 (5) (2021) 103026.
[49] Alberto Paderno, Cesare Piazza, Francesca Del Bon, Davide Lancini, Stefano Tanagli, Alberto Deganello, Giorgio Peretti, Elena De Momi, Ilaria Patrini, Michela Ruperti, Leonardo S. Mattos, Sara Moccia, Deep learning for automatic segmentation of oral and oropharyngeal cancer using narrow band imaging: preliminary experience in a clinical perspective, Front. Oncol. 11 (2021).
[50] Stefan P. Haider, Kariem Sharaf, Tal Zeevi, Philipp Baumeister, Christoph Reichel, Reza Forghani, Benjamin H. Kann, Alexandra Petukhova, Benjamin L. Judson, Manju L. Prasad, Chi Liu, Barbara Burtness, Amit Mahajan, Seyedmehdi Payabvash, Prediction of post-radiotherapy locoregional progression in HPV-associated oropharyngeal squamous cell carcinoma using machine-learning analysis of baseline PET/CT radiomics, Transl. Oncol. 14 (1) (2021) 100906.
[51] Y. Min Park, J. Yol Lim, Y. Woo Koh, S.-H. Kim, Choi E. Chang, Prediction of treatment outcome using MRI radiomics and machine learning in oropharyngeal cancer patients after surgical treatment, Oral Oncol. 122 (2021), 105559, https://doi.org/10.1016/j.oraloncology.2021.105559.
[52] Mark Marsden, Brent W. Weyers, Julien Bec, Tianchen Sun, Regina F. Gandour-Edwards, Andrew C. Birkeland, Marianne Abouyared, Arnaud F. Bewley, D. Gregory Farwell, Laura Marcu, Intraoperative margin assessment in oral and oropharyngeal cancer using label-free fluorescence lifetime imaging and machine learning, I.E.E.E. Trans. Biomed. Eng. 68 (3) (2021) 857–868.
[53] S. Fouad, G. Landini, M. Robinson, T.-H. Song, H. Mehanna, Human papilloma virus detection in oropharyngeal carcinomas with in situ hybridisation using hand crafted morphological features and deep central attention residual networks, Comput. Med. Imaging Graph. 88 (2021), 101853, https://doi.org/10.1016/j.compmedimag.2021.101853.
[54] N.-M. Cheng, J. Yao, J. Cai, X. Ye, S. Zhao, K. Zhao, et al., Deep learning for fully automated prediction of overall survival in patients with oropharyngeal cancer using FDG-PET imaging, Clin Cancer Res 27 (2021) 3948–3959.
[55] R. Rodríguez Outeiral, P. Bos, A. Al-Mamgani, B. Jasperse, R. Simões, U.A. van der Heide, Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning, Phys. Imag. Radiat. Oncol. 19 (2021) 39–44, https://doi.org/10.1016/j.phro.2021.06.005.
[56] M.A. Naser, K.A. Wahid, L.V. van Dijk, R. He, M.A. Abdelaal, C. Dede, et al., Head and neck cancer primary tumor auto segmentation using model ensembling of deep learning in PET/CT images, in: V. Andrearczyk, V. Oreiller, M. Hatt, A. Depeursinge (Eds.), Head and Neck Tumor Segmentation and Outcome Prediction, vol. 13209, Springer International Publishing, Cham, 2022, pp. 121–33. <https://doi.org/10.1007/978-3-030-98253-9_11>.
[57] Paula Bos, Michiel W.M. van den Brekel, Marjaneh Taghavi, Zeno A.R. Gouw, Abrahim Al-Mamgani, Selam Waktola, Hugo J.W.L. Aerts, Regina G.H. Beets-Tan, Jonas A. Castelijns, Bas Jasperse, Largest diameter delineations can substitute 3D tumor volume delineations for radiomics prediction of human papillomavirus status on MRI's of oropharyngeal cancer, Phys. Med. 101 (2022) 36–43.
[58] Prerna Tewari, Eugene Kashdan, Cathal Walsh, Cara M. Martin, Andrew C. Parnell, John J. O'Leary, Donna K. Slonim, Estimating the conditional probability of developing human papilloma virus related oropharyngeal cancer by combining machine learning and inverse Bayesian modelling, PLoS Comput. Biol. 17 (8) (2021) e1009289.
[59] Elisa Tardini, Xinhua Zhang, Guadalupe Canahuate, Andrew Wentzel, Abdallah S R Mohamed, Lisanne Van Dijk, Clifton D Fuller, G Elisabeta Marai, Optimal treatment selection in sequential systemic and locoregional therapy of oropharyngeal squamous carcinomas: deep Q-learning with a patient-physician digital twin dyad, J. Med. Internet Res. 24 (4) (2022) e29455.
[60] Agustina La Greca Saint-Esteven, Marta Bogowicz, Ender Konukoglu, Oliver Riesterer, Panagiotis Balermpas, Matthias Guckenberger, Stephanie Tanadini-Lang, Janita E. van Timmeren, A 2.5D convolutional neural network for HPV prediction in advanced oropharyngeal cancer, Comput. Biol. Med. 142 (2022) 105215.
[61] Kareem A. Wahid, Sara Ahmed, Renjie He, Lisanne V. van Dijk, Jonas Teuwen, Brigid A. McDonald, Vivian Salama, Abdallah S.R. Mohamed, Travis Salzillo, Cem Dede, Nicolette Taku, Stephen Y. Lai, Clifton D. Fuller, Mohamed A. Naser, Evaluation of deep learning-based multiparametric MRI oropharyngeal primary tumor auto-segmentation and investigation of input channel effects: results from a prospective imaging registry, Clin. Transl. Radiat. Oncol. 32 (2022) 6–14.

[62] S.I. Kim, J.W. Kang, Y.-G. Eun, Y.C. Lee, Prediction of survival in oropharyngeal squamous cell carcinoma using machine learning algorithms: a study based on the surveillance, epidemiology, and end results database, Front. Oncol. 12 (2022), 974678, https://doi.org/10.3389/fonc.2022.974678.
[63] Adil Dinia, Samy Ammari, John Filtes, Marion Classe, Antoine Moya-Plana, François Bidault, Stéphane Temam, Pierre Blanchard, Nathalie Lassau, Philippe Gorphe, Events prediction after treatment in HPV-driven oropharyngeal carcinoma using machine learning, Eur. J. Cancer 171 (2022) 106–113.
[64] Y.M. Park, J.-Y. Lim, Y.W. Koh, S.-H. Kim, E.C. Choi, Machine learning and magnetic resonance imaging radiomics for predicting human papilloma virus status and prognostic factors in oropharyngeal squamous cell carcinoma, Head Neck 44 (2022) 897–903, https://doi.org/10.1002/hed.26979.
[65] Omar A. Karadaghy, Matthew Shew, Jacob New, Andrés M. Bur, Machine learning to predict treatment in oropharyngeal squamous cell carcinoma, ORL 84 (1) (2022) 39–46.
[66] Nicolette Taku, Kareem A. Wahid, Lisanne V. van Dijk, Jaakko Sahlsten, Joel Jaskari, Kimmo Kaski, Clifton D. Fuller, Mohamed A. Naser, Auto-detection and segmentation of involved lymph nodes in HPV-associated oropharyngeal cancer using a convolutional deep learning neural network, Clin. Transl. Radiat. Oncol. 36 (2022) 47–55.
[67] Federico Cabitza, Andrea Campagner, Felipe Soares, Luis García de Guadiana-Romualdo, Feyissa Challa, Adela Sulejmani, Michela Seghezzi, Anna Carobene, The importance of being external. methodological insights for the external validation of machine learning models in medicine, Comput. Methods Programs Biomed. 208 (2021) 106288.
[68] T.P.A. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E.W. Steyerberg, K.G. M. Moons, A new framework to enhance the interpretation of external validation studies of clinical prediction models, J. Clin. Epidemiol. 68 (2015) 279–289, https://doi.org/10.1016/j.jclinepi.2014.06.018.
[69] Sung Yang Ho, Kimberly Phua, Limsoon Wong, Wilson Wen Bin Goh, Extensions of the external validation for checking learned model interpretability and generalizability, Patterns 1 (8) (2020) 100129.
[70] Sung Yang Ho, Limsoon Wong, Wilson Wen Bin Goh, Avoid Oversimplifications in machine learning: going beyond the class-prediction accuracy, Patterns 1 (2) (2020) 100025.
[71] E.W. Steyerberg, F.E. Harrell, Prediction models need appropriate internal, internal–external, and external validation, J. Clin. Epidemiol. 69 (2016) 245–247, https://doi.org/10.1016/j.jclinepi.2015.04.005.
[72] R. Argent, A. Bevilacqua, A. Keogh, A. Daly, B. Caulfield, The importance of real-world validation of machine learning systems in wearable exercise biofeedback platforms: a case study, Sensors 21 (2021) 2346, https://doi.org/10.3390/s21072346.
[73] Dino Gibertoni, Paola Rucci, Marcora Mandreoli, Mattia Corradini, Davide Martelli, Giorgia Russo, Elena Mancini, Antonio Santoro, Temporal validation of the CT-PIRP prognostic model for mortality and renal replacement therapy initiation in chronic kidney disease patients, BMC Nephrol. 20 (1) (2019), https://doi.org/10.1186/s12882-019-1345-7.
[74] C.L. Ramspek, K.J. Jager, F.W. Dekker, C. Zoccali, M. van Diepen, External validation of prognostic models: what, why, how, when and where? Clin. Kidney J. 14 (2021) 49–58, https://doi.org/10.1093/ckj/sfaa188.
[75] D.G. Altman, Y. Vergouwe, P. Royston, K.G.M. Moons, Prognosis and prognostic research: validating a prognostic model, BMJ 338 (2009) b605–b, https://doi.org/10.1136/bmj.b605.
[76] E.W. Steyerberg, Clinical Prediction Models, Springer New York, New York, NY, 2009. <https://doi.org/10.1007/978-0-387-77244-8>.
[77] G.C.M. Siontis, J.P.A. Ioannidis, Response to letter by Forike et al.: more rigorous, not less, external validation is needed, J. Clin. Epidemiol. 69 (2016) 250–251, https://doi.org/10.1016/j.jclinepi.2015.01.021.
[78] A.C. Justice, Assessing the generalizability of prognostic information, Ann. Intern. Med. 130 (1999) 515, https://doi.org/10.7326/0003-4819-130-6-199903160-00016.
[79] Gary S Collins, Paula Dhiman, Constanza L Andaur Navarro, Jie Ma, Lotty Hooft, Johannes B Reitsma, Patricia Logullo, Andrew L Beam, Lily Peng, Ben Van Calster, Maarten van Smeden, Richard D Riley, Karel GM Moons, Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence, BMJ Open 11 (7) (2021) e048008.
[80] G.S. Collins, K.G.M. Moons, Reporting of artificial intelligence prediction models, Lancet 393 (2019) 1577–1579, https://doi.org/10.1016/S0140-6736(19)30037-6.