Value of Human Papilloma Virus (HPV) as a marker of prognosis for oropharyngeal cancer: validation and decision curve analysis of an updated prognostic model for patients in Western Europe and the USA

Emilie Dronkers, Daan Nieboer, Stephanie Mes, Marc van der Schroeff, Senada Koljenovic, Elisabeth Bloemena, Ruud Brakenhoff, Peter Snijders†, Daniëlle Heideman, James Lewis, Ewout Steyerberg, Robert Baatenburg de Jong.

submitted

ABSTRACT

Human Papilloma Virus (HPV) infection is a prognostic factor in oropharyngeal squamous cell carcinoma (OPSCC). We developed an updated prognostic model for OPSCC, including HPV status, based on a large consecutive series of patients diagnosed and treated in three international multi-institutional cohorts. An internal- external cross validation procedure was followed and decision curve analysis (DCA) was performed to evaluate the reliability of decisions based on predictions derived from the prognostic model. The updated prognostic model, including 8th TNM classification and a separate variable for HPV, performs reasonably good and very similar to the original model in terms of calibration and discrimination. DCA however shows an improved clinical utility in comparison with the original model. The updated model could therefore be used for counselling patients about their individual prognosis.

INTRODUCTION

Over the last decades, rising incidence rates of oropharyngeal squamous cell carcinoma (OPSCC) in several geographical areas have been reported. Infection with human papillomavirus (HPV) is the major cause of these rising incidence rates.[1,2] The prevalence of HPV-positive OPSCC varies between studies. Ranges from 20 -70% in Europe and up to 90% in the United States have been reported.[1,3,4] HPV-related OPSCC is a distinct entity in contrast to tobacco- and alcohol related head and neck cancer, with regard to cellular, biologic and clinical characteristics.[5] Patients with HPV related OPSCC have an advanced N-status, better loco regional control and improved 5-year survival rates after treatment.[6] Therefore, HPV status has emerged as the main prognostic factor in OPSCC.

Recently the 8th edition UICC/AJCC TNM staging of OPSCC has been divided into two different staging systems for both HPV related OPSCC and non-HPV related OPSCC.[7] The new staging rules permit a more appropriate depiction of the prognosis of HPV-positive disease than is supplied by the 7th edition TNM classification. The discrimination of stages is especially better in HPV related OPSCC patients with smaller tumors and advanced N-status.[8]

The 8th edition of the UICC/AJCC TNM staging of OPSCC uses the immunohistochemical detection of p16 as a surrogate marker for HPV-induced carcinogenic transformation. p16 is a relatively easy, low-cost measurement, which can be easily performed on formaldehyde fixed and paraffin embedded pretreatment (FFPE) tumor samples and is readily available in routine histopathology laboratories.[9,10] However, while the test has a high sensitivity (94%), there is a moderate specificity (82%), especially in comparison with the 'gold standard' technique of HPV-DNA detection.[11] Therefore, positive immunostaining of p16 can occur in the actual absence of HPV. An estimated 10-20% of all OPSCC are p16 positive, but HPV-DNA negative.[12] Studies have shown that the prognosis of patients with p16 positive but HPV-DNA negative OPSCC is almost identical to the prognosis of 'true' HPV (both p16 and HPV-DNA) negative OPSCC patients.[13,14] Several research groups therefore strongly advise to determine HPV status by a bimodal approach with both p16 immunostaining and HPV-DNA or mRNA detection.[13-16]

The favorable prognosis of HPV positive OPSCC has lead towards a need for more specific, or rather individualized, information for both patients and doctors. In general, TNM stage alone is ineffective for predicting outcomes in individual patients, because other tumor factors and patient characteristics such as age, gender, tobacco use or comorbidity are not taken into account in the classification system.[17] Prognostic models are statistical models that calculate the cumulative effect of several prognostic variables on survival. Earlier studies presented models for HPV related OPSCC designed to stratify patients in to risk categories, mostly for the purpose of clinical trials.[3,18-21] However, most of these models are based on clinical trial populations instead of consecutive patient series, do not include combined HPV-DNA and p16 status or have not been externally validated in cohorts from different geographical areas.

In 2010 we presented a prognostic model and internal model validation for patients with newly diagnosed head and neck cancer.[22] This model is based on a Dutch cohort of n=1371 consecutive patients, treated with curative intent between 1981-1999, and externally validated with data from a large referral center in the USA. The model includes the predictors age, gender, 7th TNM classification, prior tumors and comorbidity. Discrimination of this model is good with a Harrel's C-index of 0.73.

With this study we aim to extend, update, improve and validate this prognostic model for OPSCC patients by incorporating the newly published UICC/AJCC 8th TNM staging system (cN status), and both p16 and HPV-DNA status. Our goal is to use this model in clinical practice for counselling of patients.

MATERIALS AND METHODS

This retrospective cohort study was conducted after approval was given by the Medical Ethical Committee. Tissue samples were used and analyzed according to the FEDERA guidelines. Data were analyzed with IBM SPSS Statistics 21.0 and R software V 3.1.1 (packages foreign, mice, rms, survival, Hmisc, stdca). All tests were 2-sided with a significance level of 0.05.

Study design

Five different centers provided data for this retrospective cohort study. Three were located in the Netherlands (Leiden University Medical Center, Erasmus University Medical Center Rotterdam and Amsterdam Medical Center – location VUMC). Two were located in the USA (The Siteman Cancer Center at Barnes-Jewish Hospital St. Louis, Missouri and Washington University School of Medicine). All patients with primary OPSCC curatively treated in the period 1984 – 2011 were deemed eligible for inclusion. The data provided by the five different centers were aggregated in three independent multi institutional cohorts. The characteristics of the different cohorts are described below.

The reference cohort

The prognostic model we presented in 2010 was based on 1371 consecutive patients. Within the original cohort we used back in 2010 for our prognostic model, 15% (n=204) patients were diagnosed with OPSCC. The data were retrieved from the hospital-based cancer registry (ONCDOC) of the Leiden University Medical Center (LUMC). For this study we expanded our search in ONCDOC for all primary OPSCC, diagnosed between 1981 and 2011. Patients of whom FFPE tumor samples were available in the archives of the Pathology Department of the LUMC were included. N=341 pretreatment tumor samples of patients diagnosed with primary OPSCC between 1984 and 2011 were available in the archives. A senior pathologist with elaborate experience in histologic analysis of OPSCC analyzed the tumor samples again on the presence of malignant cells. 311 of 341 (91.2%) tumor samples contained OPSCC.

The external validation cohorts

The Dutch external validation cohort (NL external cohort: Erasmus University Medical Center Rotterdam and VUMC Amsterdam) comprised n=723 patients, diagnosed with primary OPSCC between 2000-2006. The patients were identified through the Dutch Cancer Registries and the data within this cohort were earlier described and used for the development of a prognostic model.18 The USA external validation cohort (USA External Cohort: Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine) consisted of N=305 patients with primary OPSCC, diagnosed and treated between 1996-2009.

Variables

Data of a total of n=1339 consecutive OPSCC patients could be collected from the development and validation cohorts. Variables extracted for each patient were gender, age at diagnosis, date of diagnosis, anatomic subsite of the tumor, cTNM classification (with both 7th and 8th cN classification), comorbidity, prior tumor, smoking behavior, recurrent disease, last date of follow-up and survival. Comorbidity was classified according to the Adult Comorbidity Evaluation 27 (ACE27) index calculator. This classification divides comorbidity into four categories: none, mild, moderate and severe.[23] Main study endpoint was overall survival (OS), calculated from the day of diagnosis. Data were considered right-censored if patients were still alive at the time of last follow-up.

HPV analysis

In all three cohorts the immunohistochemical (IHC) analysis was performed for p16 INK4A (Roche MTM Laboratories AG, Heidelberg, Germany or MTM Laboratories CINTEC, Westborough, MA) on 4 um thick FFPE tumor sections, using a fully automated Ventana BenchMark ULTRA Stainer (Ventana, Tucson Arizona, USA) according to manufacturers' instructions at the pathology department. In all cohorts, stains were reviewed by two independent observers. Strong and diffuse nuclear and cytoplasmic immunostaining in more than 70% of the carcinoma tissue was considered as p16 positive. Partial staining of <70% or no reactivity was considered to be p16 negative. This definition of p16-positivity is consistent with previously published articles.[24] In the Dutch cohorts, DNA was extracted from all p16 positive cases using an automated silica-based extraction system. PCR was performed using the HR HPV GP 5+/6+ PCR with enzyme-immuno-assay.[25] Information on high risk HPV DNA was not available for the USA external cohort.

Statistical methods

Differences in patient characteristics between the three cohorts were assessed using Pearson's χ^2 test and ANOVA with Bonferroni correction. Missing data were handled following the missing at random assumption (MAR). Multiple imputation (MI) was performed with n=30 imputations based on the percentage of incomplete cases. The pooled imputed data were used in all analyses.

The significance of 8th TNM classification (cN status), p16 and HPV-DNA as a marker of prognosis in OPSCC was estimated by the Kaplan-Meier approach in the combined development and validation cohorts and all three cohorts separately. With these data a new Cox proportional hazards regression model was fitted. This model included the same predictors as the previously defined model. The following variables were used in the previously published model: age at diagnosis, gender, cTNM, comorbidity and subsite of the tumor.[22]

The incremental value of adding 8th TNM and HPV status (tested by p16 or PCR) to the previously defined model variables was tested in all cohorts using Wald-test in a nested models analysis. Based on literature, smoking was assumed to be an effect modifier of the incremental prognostic value of HPV.[26] Therefore, adding smoking as an interaction term to the model was also tested using Wald-test.

Cox proportional hazards regression analysis was used to evaluate the effects of all covariates on OS in an internal – external cross validation design. In prognostic modelling, the use of the maximal sample size is preferred. When multiple small datasets are available, an internal – external cross validation design is advised to combine the strength of external validation with the strength of prediction model development on all available data.[27] In an internal-external cross validation design one cohort is non-randomly left out at the time to cross-validate the model developed in the other cohorts. Because the split is not at random, this qualifies as external validation.

In this study different full models were developed by adding one new variable (8th TNM classification, p16, HPV-DNA or smoking) at a time. For each full model, three performance assessments of cross-validation were done based on the three (development and validation) cohorts. Calibration plots and ROC curves for 5 year survival probability were made for all cross-validated models. The area under the ROC curve (AUC) was calculated to evaluate the concordance between predicted and observed responses of individual subjects separately in all nine different cross-validated models. We tested for heterogeneity in baseline risk and performed interaction tests across the different cohorts.

Decision curve analysis (DCA) was performed in order to evaluate the clinical usefulness of the model for decision making. Although the AUC has been the standard for evaluating the discriminating ability of a prognostic model, it has been increasingly recognized that changes in AUC are not sensitive when a new prognostic factor (such as a biomarker) is added to a model that already comprises standard prognostic factors.[28] The AUC typically shows only a small improvement, but the clinical utility of a model including this new prognostic factor may be large. To overcome this limitation, DCA as described by Vickers et al. can be used to summarize the performance of the model in supporting decision making.[29,30]

In this study we used DCA to examine the theoretical relationship between the threshold survival probability at 5 years after diagnosis (for example 5-year survival probability of 65%) and the relative value of benefits (predicting a true positive case) and harms (predicting a false positive case) associated with the different full prognostic models.

The final model was fitted on all data (n=1339). The extent of any overfitting was estimated in an internal validation procedure using 500 bootstrap samples. A shrinkage factor was calculated and used to shrink the regression coefficients to obtain well-calibrated predictions of prognosis for new patients. The bootstrap procedure also yielded an optimism-corrected Harrell's Concordance Index.

RESULTS

Patient characteristics

Patient characteristics of all three cohorts are shown in Table 1. The baseline patient and tumor characteristics differed significantly between the USA cohort and both Dutch cohorts. Especially regarding cT status, cM status, tobacco use and p16 analysis, large differences are shown between the Dutch cohorts and the USA cohort. Over 70% of patients in the USA cohort were p16 positive, in comparison with approximately 30% of patients in

Table 1. Demographic characteristics

		Reference cohort (Dutch, NLREF) 1984-2011		External cohort (Dutch, NLEXT) 2000-2006		External cohort (USA, USEXT) 1996-2009		p- value
		N=311	Table Total %	N=723	Table Total %	N=305	Table Total %	
Age	Mean (SD)	59.7 (49.5 – 69.9)		60.2 (50.8 – 69.6)		55.6 (46.4 – 64.8)		p<0.001*
Gender	male	206	66.2%	482	66.7%	266	87.2%	p<0.001**
	female	105	33.8%	241	33.3%	39	12.8%	
cT	1	58	18.6%	114	15.8%	78	25.6%	
	2	96	30.9%	230	31.8%	110	36.1%	
	3	81	26.0%	243	33.6%	51	16.7%	p<0.001**
	4	76	24.4%	134	18.5%	57	18.7%	
	NA	0	0%	2	0.3%	9	3.0%	
cN (7th TNM)	0	122	39.2%	271	37.5%	39	12.8%	
	1	49	15.8%	101	14.0%	48	15.7%	
	2	119	38.3%	317	43.8%	194	63.6%	p<0.001**
	3	21	6.8%	33	4.6%	21	6.9%	
	NA	0	0%	1	0.1%	3	1.0%	
cN (8th TNM)	0	122	39.2%	271	37.5%	39	12.8%	
	1	60	19.3%	120	16.6%	81	26.6%	
	2	108	34.7%	298	41.2%	157	51.5%	p<0.001**
	3	21	6.8%	33	4.6%	20	6.6%	
	NA	0	0%	1	0.1%	8	2.6%	
cM	0	302	97,1%	714	98.8%	300	100.0%	
	1	9	2,9%	4	0.6%	0	0%	p=0.054***
	NA	0	0%	5	0.7%	0	0%	
Comorbidity (ACE27)	None	133	42.8%	292	40.4%	123	40.3%	
	Mild	63	20.3%	219	30.3%	113	37.0%	
	Moderate	72	23.2%	176	24.3%	42	13.8%	p<0.001**
	Severe	39	12.5%	34	4.7%	22	7,2%	
	NA	4	1.3%	2	0.3%	5	1.7%	
Smoking	never	36	11.6%	64	8.9%	81	26.6%	
	ever	126	40.5%	653	90.3%	212	69.5%	p<0.001**
	NA	149	47.9%	6	0.8%	12	3.9%	
Death	No	123	39.5%	274	37.9%	208	68.2%	p<0.001**
	Yes	188	60.5%	449	62.1%	97	31.8%	
Recurrent disease	No	199	63.9%	385	53.2%	243	79.6%	p<0.001**
	Yes	112	36.1%	338	46.7%	62	21.3%	
P16 analysis	<70% immunostaining	213	68.5%	544	75.2%	70	23.0%	
	≥70% immunostaining	98	31.5%	179	24.8%	230	75.4%	p<0.001**
	NA	0	0%	0	0%	5	1.6%	

Table 1. Demographic characteristics (continued)

		Reference cohort (Dutch, NLREF) 1984-2011		External cohort (Dutch, NLEXT) 2000-2006		External cohort (USA, USEXT) 1996-2009		p- value
		N=311	Table Total %	N=723	Table Total %	N=305	Table Total %	
HPV DNA analysis	negative	241	77.5%	571	79.0%	NA	NA	
	positive	70	22.5%	152	21.0%	NA	NA	p=0.652***
	NA	0	0%	0	0%	305	100%	

*as defined by one way ANOVA using bonferroni p-value adjustment, ** as defined by the χ2 test, *** no data for USA available therefore χ2 test was performed between NL reference and NL external cohort.

the Dutch cohorts. The characteristics of both Dutch cohorts were nearly comparable. The 5-year OS estimates were 70.7% in the p16 positive group and 38.7% in the p16 negative group.

In the Dutch reference cohort, 28.6% (n=28) of patients were p16 positive but HPV-DNA negative. In the Dutch external validation cohort, 15.1% (n=27) of patients were p16 positive but HPV-DNA negative. Since HPV-DNA analysis was not available in the USA external validation cohort, the percentage of true positive HPV cases could not be defined. Figure 1 shows OS as estimated by the Kaplan Meier approach for the significance of p16 and HPV-DNA as a marker for prognosis in all 3 cohorts separately. Log-rank test showed for both factors a significant result regarding the non-equality of survival distributions (p<0.001).

In Table 2 the univariate hazard ratio's (HR's) of all new possible prognostic factors are outlined. The differences in HR in cN status between the Dutch and USA cohorts can be explained by the higher percentage of p16 positive patients in the USA cohort, and thus a different distribution of N-stages. The differences in HR between p16 and HPV-DNA can be explained by the percentage of 'false-positive' p16 cases (p16 positive, but HPV-DNA negative).

The incremental value of adding 8th TNM and HPV status (tested by p16 or PCR) to the previously defined model variables was tested in all cohorts using Wald test in a nested models analysis. Both in the Dutch reference cohort as in the combined internal – external cohort, the Wald test revealed significant results for adding HPV to the previously defined model (p<0.001). This was the case for p16 and HPV-DNA-analysis. Testing for interaction between smoking and HPV positivity did not show significant interaction (p=0.09).

Figure 1. Overall survival in all 3 cohorts

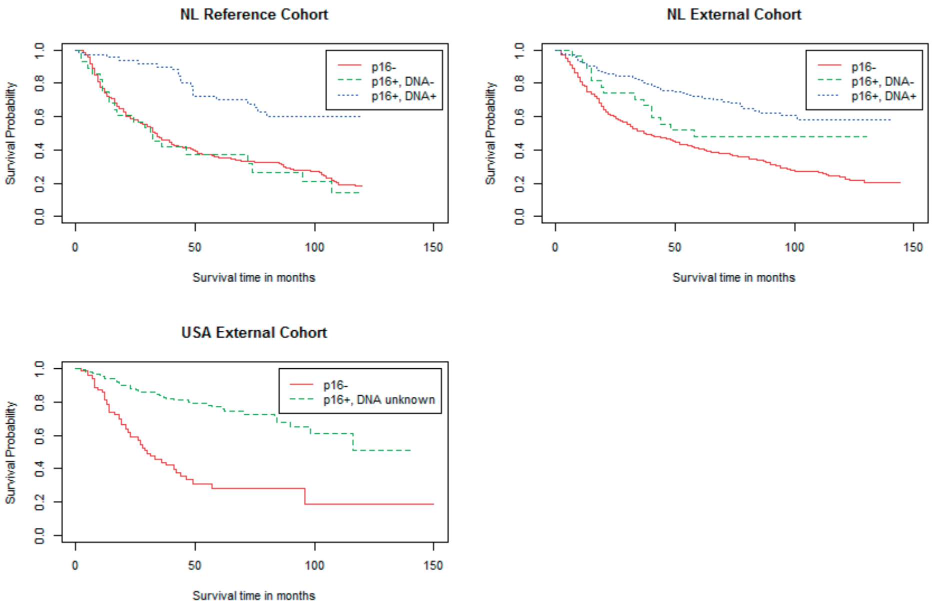


Table 2. Univariate analysis showing unadjusted Hazard Ratio's (HR) for prognostic value of new model variables.

		Dutch reference cohort (LUMC)		Dutch external cohort (EMC/VUMC)		USA external cohort (USA)	
		HR (95% CI)	*p-value*	*HR (95% CI)*	*p-value*	*HR (95% CI)*	*p-value*
cN (8th TNM)	N0	-	-	-	-	-	-
	N1	0.93 (0.61 – 1.43)	0.74	0.91 (0.68 – 1.22)	0.519	0.87 (0.44 – 1.72)	0.685
	N2	1.696 (1.22 – 2.37)	0.002	1.48 (1.20 – 1.83)	0.000	1.02 (0.55 – 1.89)	0.954
	N3	2.33 (1.41 – 3.87)	0.001	2.83 (1.91 – 4.25)	0.000	2.72 (1.21 – 6.14)	0.016
P16	Negative	-	-	-	-	-	-
	Positive	0.48 (0.34 – 0.69)	0.000	0.417 (0.32 – 0.54)	0.000	0.247 (0.164 – 0.37)	0.000
HPV-DNA	Negative	-	-	-	-	NA	
	Positive	0.303 (0.19 – 0.48)	0.000	0.390 (0.29 – 0.52)	0.000	NA	

Following these results, four different prognostic models were fitted:

1. *age + gender + comorbidity + 7th cTNM (original model)*
2. *age + gender + comorbidity + 8th cTNM classification*
3. *age + gender + comorbidity + 8th cTNM classification + P16*
4. *age + gender + comorbidity + 8th cTNM classification + HPV DNA*

The four different prognostic models were cross-validated over the three cohorts using Cox proportional hazards regression analysis, leading to a total of 11 performance assessments. Since HPV-DNA analysis was not available in the USA external validation cohort, a model with HPV DNA could not be validated for this cohort. The calibration of the 11 different models was assessed graphically with a calibration plot. (Figure 2a). The results of 11 model performance assessments (HR's and 95% CI) are presented in Table 3a-c. The Harrell's Concordance Indices of all models differed between 0.64 and 0.74. Models containing 8th TNM and a separate variable for HPV as a prognostic factor performed better than models without HPV.

To summarize the performance of the model in supporting decision making, DCA was performed. Decision curves of 11 model performance assessments for the relationship between threshold survival probability at 5 years after diagnosis and the relative value of benefits and harms are displayed in Figure 2b. Interpretation of the decision curve depends on comparing the net benefit of the different models with that of a strategy of "treat all" (the thin grey line) and "treat none" (parallel to the x axis at net benefit of zero). "Treating" in this setting means any treatment decision that could be made for OPSCC patients dependent on expected survival rate, such as adjuvant chemotherapy, or dose-escalating radiotherapy. The strategy with the highest net benefit at a particular point is optimal, irrespective of the size of the difference. Net benefits of all models were superior at wide range of "decision to treat" thresholds. Across all threshold probabilities, models with a variable for HPV (either p16 or HPV DNA) performed better than models with only the 8th TNM classification. The HPV DNA prognostic model had a slightly greater net benefit compared with the HPV p16 prognostic model.

Regarding the results of internal – external cross validation and the decision curve analysis, we chose to fit the final model in the combined data (n=1339), based on the following variables: gender, age at diagnosis, 8th cTNM classification, comorbidity (ACE27) and p16 analysis. The choice for p16 analysis as a measurement of HPV positivity was also emphasized because of the easy accessibility of this surrogate marker in routine histopathological laboratories. We tested for heterogeneity in baseline risk and performed interaction tests across the different cohorts for the final model. Wald test showed a significant result for comparing the final model with and without an interaction term for the three different cohorts (p=0.021). Therefore, the interaction term for cohorts was also fitted in the updated final model. Harrell's Concordance Index for this model was 0.72. The extent of overfitting was estimated in an internal validation procedure using 500 bootstrap samples. The bootstrap procedure yielded an optimism-corrected Harrell's Concordance Index of 0.70.

3a) Reference cohort (LUMC)

Factors	Model 1 (7th TNM)			Model 2 (8th TNM)			Model 3 (8th TNM + p16)			Model 4 (8th TNM + p16 + DNA)		
	C-stat= 0.672			C-stat= 0.674			C-stat= 0.693			C-stat = 0.703		
	Hazard Ratio	*95% CI*	*p-value*	*Hazard Ratio*	*95% CI*	*p-value*	*Hazard Ratio*	*95% CI*	*p-value*	*Hazard Ratio*	*95% CI*	*p-value*
Male	1.0	-	-	1.0	-	-	1.0	-	-	1.0	-	-
Female	1.19	0.98 -1.44	0.073	1.18	0.98 -1.43	0.083	1.01	0.83 - 1.22	0.90	1.02	0.84 - 1.23	0.85
Age	1.02	1.01 -1.03	<0.001	1.02	1.01 - 1.03	<0.001	1.02	1.01 - 1.03	<0.001	1.02	1.01 - 1.03	<0.001
cT1	1.0	-	-	1.0	-	-	1.0	-	-	1.0	-	-
cT2	1.15	0.87 - 1.51	0.334	1.13	0.86 - 1.49	0.382	1.08	0.82 - 1.42	0.592	1.06	0.81 - 1.40	0.685
cT3	1.69	1.29 - 2.22	<0.001	1.67	1.27 - 2.19	<0.001	1.39	1.06 - 1.84	0.017	1.39	1.05 - 1.83	0.021
cT4	2.69	2.02 - 3.59	<0.001	2.62	1.97 - 3.49	<0.001	2.12	1.58 - 2.84	<0.001	2.08	1.55 - 2.79	<0.001
cN0	1.0	-	-	1.0	-	-	1.0	-	-	1.0	-	-
cN1	0.99	0.75 - 1.31	0.96	0.93	0.71 - 1.21	0.59	1.22	0.93 - 1.59	0.149	1.25	0.95 - 1.63	0.11
cN2	1.23	1.01 - 1.51	0.042	1.29	1.05 - 1.59	0.014	1.77	1.43 - 2.19	<0.001	1.80	1.46 - 2.23	<0.001
cN3	2.76	1.93 - 3.95	<0.001	2.76	1.93 - 3.94	<0.001	3.56	2.49 - 5.10	<0.001	3.70	2.58 - 5.32	<0.001
cM0	1.0	-		1.0	-	-	1.0	-	-	1.0	-	-
cM1	4.75	1.50 - 14.98	0.007	4.54	1.43 - 14.32	0.009	5.74	1.83 - 18.0	0.003	6.04	1.91 - 18.95	0.002
ACE27 Score 0	1.0	-	-	1.0	-	-	1.0	-	-	1.0	-	-
ACE27 Score 1	1.33	1.07 - 1.65	0.008	1.33	1.07 - 1.65	0.009	1.24	0.99 - 1.54	0.052	1.23	0.99 - 1.52	0.065
ACE27 Score 2	2.11	1.68 - 2.66	<0.001	2.09	1.66 - 2.63	<0.001	1.69	1.34 - 2.13	<0.001	1.68	1.33 - 2.12	<0.001
ACE27 Score 3	2.62	1.84 - 3.72	<0.001	2.60	1.84 - 3.69	<0.001	2.65	1.87 - 3.75	<0.001	2.58	1.81 - 3.67	<0.001
p16 negative	-	-	-	-	-	-	1.0	-	-	1.0	-	-
p16 positive	-	-	-	-	-	-	0.34	0.28 - 0.43	<0.001	0.56	0.35 - 0.89	0.017
PCR DNA HPV negative	-	-	-	-	-	-	-	-	-	1.0	-	-
PCR DNA HPV positive	-	-	-	-	-	-	-	-	-	0.55	0.32 - 0.94	0.032

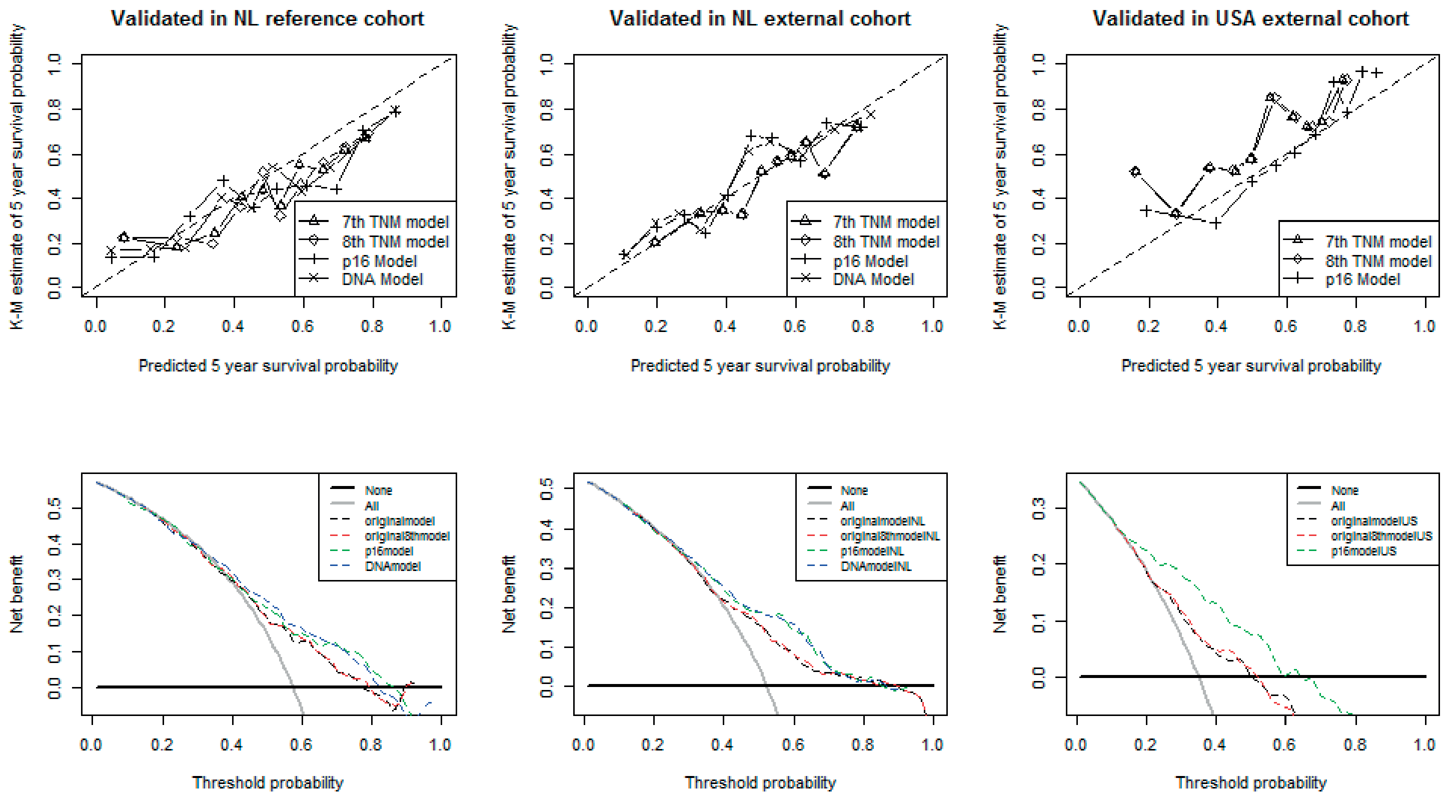
3b) Dutch external validation cohort (EMC/VUMC)

Factors	Model 5 (7th TNM)			Model 6 (8th TNM)			Model 7 (8th TNM + p16)			Model 8 (8th TNM + p16 + DNA)		
	C-stat= 0.643			C-stat= 0.647			C-stat= 0.679			C-stat= 0.678		
	Hazard Ratio	*95% CI*	*p-value*	*Hazard Ratio*	*95% CI*	*p-value*	*Hazard Ratio*	*95% CI*	*p-value*	*Hazard Ratio*	*95% CI*	*p-value*
Male	1.0	-	-	1.0	-	-	1.0	-	-	1.0	-	-
Female	1.12	0.85 - 1.48	0.430	1.14	0.86 - 1.51	0.354	1.03	0.77 - 1.36	0.852	1.01	0.76 - 1.34	0.934
Age	1.02	1.01 - 1.03	0.0021	1.02	1.01 - 1.03	0.0025	1.02	1.01 - 1.03	0.0025	1.02	1.01 - 1.03	0.0041
cT1	1.0	-	-	1.0	-	-	1.0	v		1.0	-	-
cT2	1.81	1.20 - 2.73	0.0046	1.78	1.18 - 2.69	0.006	1.74	1.15 - 2.63	0.0081	1.71	1.12 - 2.58	0.012
cT3	2.38	1.55 - 3.65	<0.001	2.35	1.53 - 3.61	<0.001	2.06	1.34 - 3.17	<0.001	1.97	1.27 - 3.03	0.0023
cT4	4.03	2.68 - 6.06	<0.001	3.96	2.62 - 5.96	<0.001	3.17	2.09 - 4.78	<0.001	2.93	1.93 - 4.43	<0.001
cN0	1.0	-	-	1.0	-	-	1.0	-	-	1.0	-	-
cN1	1.11	0.76 - 1.62	0.597	0.97	0.68 - 1.39	0.887	1.36	0.95 - 1.96	0.095	1.45	1.01 - 2.09	0.0439
cN2	1.09	0.82 - 1.44	0.565	1.15	0.86 - 1.54	0.331	1.69	1.25 - 2.28	<0.001	1.78	1.31 - 2.40	<0.001
cN3	1.84	1.21 - 2.81	0.0047	1.85	1.21 - 2.83	0.0042	2.53	1.65 - 3.88	<0.001	2.53	1.64 - 3.89	<0.001
cM0	1.0	-	-	1.0	-	-	1.0	-	-	1.0	-	-
cM1	8.12	3.76 - 17.5	<0.001	7.89	3.65 - 17.06	<0.001	6.71	3.09 - 14.54	<0.001	6.81	3.13 - 14.82	<0.001
ACE27 Score 0	1.0	-	-	1.0	-	-	1.0	-	-	1.0	-	-
ACE27 Score 1	0.79	0.58 - 1.08	0.147	0.80	0.58 - 1.09	0.152	0.80	0.58 - 1.09	0.164	0.81	0.58 - 1.11	0.195
ACE27 Score 2	1.44	1.06 - 1.95	0.018	1.42	1.05 - 1.93	0.023	1.11	0.81 - 1.52	0.504	1.07	0.78 - 1.47	0.655
ACE27 Score 3	1.39	0.95 - 2.03	0.083	1.39	0.95 - 2.02	0.089	1.15	0.78 - 1.69	0.467	1.11	0.75 - 1.64	0.589
p16 negative	-	-	-	-	-	-	1.0	-	-	1.0	-	-
p16 positive	-	-	-	-	-	-	0.35	0.27 - 0.47	<0.001	0.65	0.43 - 0.99	0.045
PCR DNA HPV negative	-	-	-	-	-	-	-	-	-	1.0	-	-
PCR DNA HPV positive	-	-	-	-	-	-	-	-	-	0.41	0.25 - 0.67	<0.001

3c) USA external validation cohort (St. Louis)

Factors	Model 9 (7th TNM)			Model 10 (8th TNM)			Model 11 (8th TNM + p16)		
	C-stat= 0.677			C-stat= 0.674			C-stat= 0.735		
	Hazard Ratio	*95% CI*	*p-value*	*Hazard Ratio*	*95% CI*	*p-value*	*Hazard Ratio*	*95% CI*	*p-value*
Male	1.0	-	-	1.0	-	-	1.0	-	
Female	1.05	0.87 - 1.25	0.556	1.06	0.89 - 1.25	0.530	1.02	0.86 - 1.21	0.796
Age	1.03	1.02 - 1.03	<0.001	1.03	1.02 - 1.03	<0.001	1.03	1.02 - 1.04	<0.001
cT1	1.0	-	-	1.0	-	-	1.0	-	-
cT2	1.22	0.94 - 1.59	0.142	1.20	0.92 - 1.56	0.178	1.19	0.92 - 1.56	0.189
cT3	1.59	1.22 - 2.06	<0.001	1.55	1.19 - 2.01	<0.001	1.45	1.12 - 1.88	0.0053
cT4	2.39	1.81 - 3.13	<0.001	2.29	1.74 - 3.01	<0.001	1.97	1.49 - 2.59	<0.001
cN0	1.0	-	-	1.0	-	-	1.0	-	-
cN1	1.01	0.78 - 1.29	0.949	0.94	0.74 - 1.20	0.625	1.11	0.86 - 1.41	0.430
cN2	1.38	1.15 - 1.67	<0.001	1.47	1.22 - 1.76	<0.001	1.79	1.47 - 2.16	<0.001
cN3	2.61	1.89 - 3.60	<0.001	2.63	1.90 - 3.63	<0.001	3.23	2.33 - 4.48	<0.001
cM0	1.0	-	-	1.0	-	-	1.0	-	-
cM1	4.25	2.29 - 7.88	<0.001	4.13	2.23 - 7.66	<0.001	4.35	2.33 - 8.11	<0.001
ACE27 Score 0	1.0	-	-	1.0	-	-	1.0	-	-
ACE27 Score 1	1.36	1.11 - 1.66	0.0033	1.36	1.11 - 1.66	0.0028	1.23	1.01 - 1.51	0.0442
ACE27 Score 2	1.89	1.54 - 2.31	<0.001	1.87	1.53 - 2.29	<0.001	1.60	1.31 - 1.96	<0.001
ACE27 Score 3	2.59	1.93 - 3.48	<0.001	2.61	1.95 - 3.51	<0.001	2.25	1.67 - 3.02	<0.001
p16 negative	-	-	-	-	-	-	1.0	-	-
p16 positive	-	-	-	-	-	-	0.43	0.35 - 0.53	<0.001

Figure 2. a) calibration plots of all models, b) decision curves of all models



DISCUSSION

This study describes the update, improvement and validation of an existing prognostic model for OPSCC patients by incorporating the newly published UICC/AJCC 8th TNM staging system (cN sta-tus), and both p16 and HPV-DNA status. Three independent multi-institutional cohorts with OPSCC patients from Western Europe and the USA (period 1984 – 2011) were used in an internal-external cross validation design. In all cohorts HPV, either detected by p16 or PCR DNA, was an independent prognostic factor for overall survival in OPSCC patients. Models with 8th TNM and a separate variable for HPV (PCR DNA or p16) as a prognostic factor performed better than models without HPV. Harrell's Concordance Indices were reasonably good. The final updated prognostic model, including 8th TNM classification and a separate variable for HPV, performs very similar to the original model in terms of calibration and discrimination. Decision curve analysis (DCA) however showed an improved clinical utility in comparison with the original model. To our knowledge, this study is first to report on DCA in prognostic models for head and neck cancer patients. Models with a separate variable for HPV (either p16 or HPV DNA) performed better in terms of supporting decision making, than models with only the 8th TNM classification, despite the incorporation of p16 in this classification system. This statistical method for summarization of model performance in supporting decision making is very interesting given the ongoing studies on de-escalation therapies and other treatment modifications for patients with HPV positive OPSCC.[31] The updated model could be used for counselling patients about their individual prognosis and treatment options. Besides de-escalation therapies for HPV positive patients, tailor-made treatment proposals based on the predicted overall survival for HPV negative patients (with a likely unfavorable prognosis) are of interest.

This study shows that with the introduction of the 8th TNM classification, predictions based only on TNM are improved, but not precise enough for individual patients. Also comorbidity, age, gender and HPV-DNA status should be taken into account. One of the strengths of this study is the joint use of PCR DNA and p16 for scoring HPV positivity in OPSCC patients. The large sizes of the development and validation cohorts, the heterogeneity due to different geographical areas, and the consecutive population based aspect diminishes the risk of selection bias. The same heterogeneity however is a limitation of this study, and is shown best in de USA cohort. The USA cohort might perform better due to several factors: a higher HPV prevalence, lower smoking and a lower share of patients in advanced stage (no M1 disease). Furthermore, there was no information on PCR DNA in this cohort and therefore false-positive (e.g. p16 positive but HPV DNA negative) tumors could have affected the results. This kind of heterogeneity in patient populations can lead to poor calibration in comparison to the derivation cohort. However, internal validation of the full model corrected for optimism using bootstrapping showed reasonably well performance of the model with Harrell's Concordance Index of 0.70.

There was a slight difference in favor of the performance of models with HPV positivity measured by PCR DNA in comparison with the surrogate biomarker p16. Positive immunostaining for p16 can occur in the actual absence of HPV.[12] And since the prognosis of p16 positive but HPV-DNA negative OPSCC is almost identical to the prognosis of double negative patients, a bimodal approach of p16 and HPV-DNA detection is advised for most accurate determination of HPV status.[13-16] The results of our study, and especially the calibration and AUCs of the different models, are in alignment with this recommendation. Alternatively, in terms of DCA and clinical applicability of prognostic models in a decision making process, p16 is a very representative prognostic marker. P16 is readily available in routine histopathology laboratories and not expensive to measure. The easy access to p16 measurements, the reasonably good performance of a model with p16 and the more imprecise prediction of individual prognosis using only 8th TNM classification are arguments to use our proposed updated model with p16 as a marker of HPV positivity. We recommend the use of our model in a clinical setting, especially when counselling patients about their individual prognosis. This could facilitate a shared decision making process. Evidently, when considering de-escalation therapies, one should be sure about HPV positivity and also perform an HPV-DNA analysis.

Acknowledgements

We thank Lisette de Vogel for her technical assistance for immunostaining. We thank Jan Molenaar for all his support to the data management in ONCDOC. We also want to thank René Leemans and Jay Piccirillo for the external data that were used for cross-validation.

REFERENCES

1. Chaturvedi AK, Engels EA, Pfeiffer RM, Hernandez BY, Xiao W, Kim E, et al. Human papillomavirus and rising oropharyngeal cancer incidence in the United States. J Clin Oncol. 2011;29(32):4294-301.
2. Gillison ML, Chaturvedi AK, Anderson WF, Fakhry C. Epidemiology of Human Papillomavirus-Positive Head and Neck Squamous Cell Carcinoma. J Clin Oncol. 2015;33(29):3235-42.
3. Ang KK, Harris J, Wheeler R, Weber R, Rosenthal DI, Nguyen-Tan PF, et al. Human papillomavirus and survival of patients with oropharyngeal cancer. N Engl J Med. 2010;363(1):24-35.
4. Mehanna H, Beech T, Nicholson T, El-Hariry I, McConkey C, Paleri V, et al. Prevalence of human papillomavirus in oropharyngeal and nonoropharyngeal head and neck cancer--systematic review and meta-analysis of trends by time and region. Head Neck. 2013;35(5):747-55.
5. Maxwell JH, Grandis JR, Ferris RL. HPV-Associated Head and Neck Cancer: Unique Features of Epidemiology and Clinical Management. Annu Rev Med. 2016;67:91-101.
6. O'Rorke MA, Ellison MV, Murray LJ, Moran M, James J, Anderson LA. Human papillomavirus related head and neck cancer survival: a systematic review and meta-analysis. Oral Oncol. 2012;48(12):1191-201.
7. Lydiatt WM, Patel SG, O'Sullivan B, Brandwein MS, Ridge JA, Migliacci JC, et al. Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. CA Cancer J Clin. 2017;67(2):122-37.
8. O'Sullivan B, Huang SH, Su J, Garden AS, Sturgis EM, Dahlstrom K, et al. Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study. Lancet Oncol. 2016;17(4):440-51.
9. Bishop JA, Lewis JS, Jr., Rocco JW, Faquin WC. HPV-related squamous cell carcinoma of the head and neck: An update on testing in routine pathology practice. Semin Diagn Pathol. 2015;32(5):344-51.
10. El-Naggar AK, Westra WH. p16 expression as a surrogate marker for HPV-related oropharyngeal carcinoma: a guide for interpretative relevance and consistency. Head Neck. 2012;34(4):459-61.
11. Prigge ES, Arbyn M, von Knebel Doeberitz M, Reuschenbach M. Diagnostic accuracy of p16(INK4a) immunohistochemistry in oropharyngeal squamous cell carcinomas: A systematic review and meta-analysis. Int J Cancer. 2017;140(5):1186-98.
12. Wang H, Sun R, Lin H, Hu WH. P16INK4A as a surrogate biomarker for human papillomavirus-associated oropharyngeal carcinoma: consideration of some aspects. Cancer Sci. 2013;104(12):1553-9.
13. Rietbergen MM, Leemans CR, Bloemena E, Heideman DA, Braakhuis BJ, Hesselink AT, et al. Increasing prevalence rates of HPV attributable oropharyngeal squamous cell carcinomas in the Netherlands as assessed by a validated test algorithm. Int J Cancer. 2013;132(7):1565-71.
14. Sharma SJ, Wagner S, Reder HSF, Kroll T, Wuerdemann N, Klussmann JP, et al. The 8th edition AJCC/UICC TNM staging for p16-positive oropharyngeal carcinoma: is there space for improvement? Eur Arch Otorhinolaryngol. 2018.

15. Smeets SJ, Hesselink AT, Speel EJ, Haesevoets A, Snijders PJ, Pawlita M, et al. A novel algorithm for reliable detection of human papillomavirus in paraffin embedded head and neck cancer specimen. Int J Cancer. 2007;121(11):2465-72.
16. Garnaes E, Frederiksen K, Kiss K, Andersen L, Therkildsen MH, Franzmann MB, et al. Double positivity for HPV DNA/p16 in tonsillar and base of tongue cancer improves prognostication: Insights from a large population-based study. Int J Cancer. 2016;139(11):2598-605.
17. Patel SG, Lydiatt WM. Staging of head and neck cancers: is it time to change the balance between the ideal and the practical? J Surg Oncol. 2008;97(8):653-7.
18. Rietbergen MM, Witte BI, Velazquez ER, Snijders PJ, Bloemena E, Speel EJ, et al. Different prognostic models for different patient populations: validation of a new prognostic model for patients with oropharyngeal cancer in Western Europe. Br J Cancer. 2015;112(11):1733-6.
19. Fakhry C, Zhang Q, Nguyen-Tan PF, Rosenthal DI, Weber RS, Lambert L, et al. Development and Validation of Nomograms Predictive of Overall and Progression-Free Survival in Patients With Oropharyngeal Cancer. J Clin Oncol. 2017;35(36):4057-65.
20. Rios Velazquez E, Hoebers F, Aerts HJ, Rietbergen MM, Brakenhoff RH, Leemans RC, et al. Externally validated HPV-based prognostic nomogram for oropharyngeal carcinoma patients yields more accurate predictions than TNM staging. Radiother Oncol. 2014;113(3):324-30.
21. Granata R, Miceli R, Orlandi E, Perrone F, Cortelazzi B, Franceschini M, et al. Tumor stage, human papillomavirus and smoking status affect the survival of patients with oropharyngeal cancer: an Italian validation study. Ann Oncol. 2012;23(7):1832-7.
22. Datema FR, Ferrier MB, Vergouwe Y, Moya A, Molenaar J, Piccirillo JF, et al. Update and external validation of a head and neck cancer prognostic model. Head Neck. 2013;35(9):1232-7.
23. Piccirillo JF. Impact of comorbidity and symptoms on the prognosis of patients with oral carcinoma. Arch Otolaryngol Head Neck Surg. 2000;126(9):1086-8.
24. Gronhoj Larsen C, Gyldenlove M, Jensen DH, Therkildsen MH, Kiss K, Norrild B, et al. Correlation between human papillomavirus and p16 overexpression in oropharyngeal tumours: a systematic review. Br J Cancer. 2014;110(6):1587-94.
25. Hesselink AT, Berkhof J, van der Salm ML, van Splunter AP, Geelen TH, van Kemenade FJ, et al. Clinical validation of the HPV-risk assay, a novel real-time PCR assay for detection of high-risk human papillomavirus DNA by targeting the E7 region. J Clin Microbiol. 2014;52(3):890-6.
26. Peck BW, Dahlstrom KR, Gan SJ, Caywood W, Li G, Wei Q, et al. Low risk of second primary malignancies among never smokers with human papillomavirus-associated index oropharyngeal cancers. Head Neck. 2012.
27. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245-7.
28. Baker SG, Schuit E, Steyerberg EW, Pencina MJ, Vickers A, Moons KG, et al. How to interpret a small increase in AUC with an additional risk prediction marker: decision analysis comes through. Stat Med. 2014;33(22):3946-59.
29. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak. 2008;8:53.
30. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-74.

31. Masterson L, Moualed D, Liu ZW, Howard JE, Dwivedi RC, Tysome JR, et al. De-escalation treatment protocols for human papillomavirus-associated oropharyngeal squamous cell carcinoma: a systematic review and meta-analysis of current clinical trials. Eur J Cancer. 2014;50(15):2636-48.