SDB Comments 10/3/22

Prognosis prediction in head and neck cancer

Conor Whitley

The University of Liverpool, Department of Physics

Abstract

1 Introduction

eighth

Oral squamous cell carcinoma (OSCC) is the 8th most common form of cancer in the UK [1] with a recent increase in incidence reported [2, 3]. The majority of head and neck cancers are oropharyngeal squamous cell carcinomas (OPSCC), which originate in the upper aerodigestive epithelium. Most OPSCCs are known to be linked to carcinogenic human papillomavirus (HPV), accounting for approximately 51.8% - 71% of cases [4]. In contrast, OSCCs are rarely mediated by HPV [5] and the majority of cases are typically associated with exposure to carcinogens present in tobacco and alcohol [2, 4, 6]. The regions in which head and neck tumours typically develop are anatomically complex and play a vital physiological role in the patient; early diagnosis and selection of appropriate treatment is needed to ensure patient survival and retention of vital organ function.

3 sig figs. "approx"?

A key issue facing clinical decision makers is the determination of the optimal course of treatment for a patient. In cases where lower biological aggression can be demonstrated, a de-escalation of therapy can be possible [7]. Identification of these cases is paramount to minimising the adverse effects of treatment, and improving patient outcomes. Previous work [8–11] has hypothesised that tumours which may be responsive

to novel therapeutic treatment may carry a distinct molecular fingerprint; the identification of which may allow for screening of patients towards appropriate treatment (needs ref)[]. For approximately 50% of HPV negative head and neck squamous cell carcinoma patients, current treatment plans are ineffective **(reference for this?)**[]. Neo-adjuvant therapy has the potential to improve prognoses, and aid clinical decision making if applicable cases can be determined at the time of diagnosis.

Previous studies [12–18] have investigated the viability of a range of prognostic biomarkers for head and neck cancer, with varying degrees of success. Many previously analysed biomarkers are measured on surgically resected tissue - limiting the potential for timely treatment. What is needed are prognostic biomarkers which can be measured in biopsy tissue prior to surgery. The discovery of effective prognostic biomarkers has been difficult, and thus far has largely focused on immunohistochemistry (IHC) techniques. Magnetic Resonance Imaging (MRI) has also been utilised [19–21] to measure physical attributes such as: tumour thickness, depth of invasion, and the presence of sub-volumes in a non-invasive manner. However, MRI based techniques often quantify biomarkers inaccurately when validated against direct measurements of pathological staging sections [22, 23].

Fourier transform infra-red (FTIR) microscopy is a well established technique, which has been utilised to investigate a range of biomedical applications in recent years. Due to its ability to access chemical information present within the sample; FTIR microscopy data and accompanied multivariate analysis has been used to diagnose cancer in biofluids [24–26], surgically resected tissue [27, 28], and cells [29–31]. FTIR microscopy allows for imaging of sample specimens at thousands of infra-red wavelengths simultaneously using a typical spectrometer. Marginal spectral differences in biochemical compounds of interest are typically located in a region known as the fingerprint region (1000cm^{-1}-1800cm^{-1}). It is differences in these absorption bands which contain information which can be utilised to discriminate between samples of interest.

Zawlik et.al [32] investigated FTIR coupled with principal component analysis (PCA) to investigate the efficacy of chemotherapy in triple-negative breast cancer. They

determined that it was possible to monitor changes in the biochemical composition of the tissue in order to monitor the effectiveness of received treatment. Butler et.al [33] have undertaken development of a high-throughput ATR-FTIR based instrument for use in biofluid assays. Their work concluded that it was possible to triage brain cancer using utilising FTIR spectroscopy of biofluid samples. Their analysis comprised a large retrospective cohort of 724 patients with a range of brain cancer subtypes and stages. They utilised a binary support vector machine (SVM) classifier, and were able to achieve a sensitivity and specificity of 93.2% and 92.8% respectively.

This work explores the potential efficacy of FTIR microscopy in combination with a known prognostic biomarker: α-smooth muscle actin (ASMA) expression, as a method of identifying cases which may be appropriate for therapeutic treatment. Previous work [15–17] has explored the efficacy of ASMA and SERPINE1 [17] as predictive variables for extra capsular spread (ECS), and as prognostic biomarkers for OSCC. ASMA expression is closely associated with the presence of activated fibroblasts, also known as myofibroblasts in tumour associated stroma. The degree of ASMA expression can be interrogated through the use of appropriate chemical stains, and evaluated using an optical microscope.

2 Materials and Methods

Tissue preparation

The dataset comprised FTIR spectra taken from primary tumour sites of 29 patients with a diagnosis of OSCC. The specimens are a subset of those arranged in a previously described tissue microarray (TMA; [17]). Inclusion criteria for this study were: a diagnosis of OSCC; the presence of OSCC in the TMA core; the ability to co-register adjacent H&E stained and FTIR imaged sections; a follow-up period after surgery of at least 24 months. Patients gave written, informed consent and the study was undertaken under ethical approval (Northwest - Liverpool Central REC number EC47.01). All

samples were 1mm diameter cores of formalin fixed paraffin embedded (FFPE) tissue arranged into a tissue micro array.

Four adjacent sections of 4μm thickness were taken from the TMA, the first and last sections were stained with Haematoxylin and Eosin (H&E) and used to assess the presence and location of tumour material in the 2nd and 3rd sections. Specimens were removed from the sample set if no clear area containing mostly tumour cells was discernable in the H&E stained sections. Samples were also removed from the sample set if the outline of the regions containing tumour cells were markedly different between the 1st and 4th sections. Images of stained sections were scanned using an Aperio CS2 scanner (Leica Biosystems) and used for IR image annotation. The 2nd and 3rd TMA sections were mounted onto CaF_2 disks for FTIR microspectroscopy.

FTIR Microspectroscopy

FTIR measurements of TMA cores were taken at room temperature using a Varian Cary 670-FTIR spectrometer with an attached Varian Cary 620-FTIR microscope produced by Varian (now Agilent Technologies, Santa Clara CA, USA); with a liquid nitrogen cooled 128×128 pixel mercury-cadmium-telluride (MCT) focal plane array with an effective field of view for each pixel of 5.5µm. The sample stage was enclosed in a perspex box and pumped with dry air until a humidity of 1% was achieved into order to mitigate the effects of water contributions on measured IR spectra. Images were acquired at a resolution of 6cm^{-1} over a spectral range of 990cm^{-1} to 3800cm^{-1} using a co-addition of 128 scans. Attenuator and integration time of the focal plane array (FPA) were chosen to gain the maximum signal-to-noise ratio. Background scans were acquired using a blank CaF_2 disk situated within the perspex box before each session of measurements.

Data Preprocessing & Analysis

Selection of tissue areas to include in the analysis was undertaken by a consultant oral pathologist (AT), who identified regions containing high proportions of tumour cells on the H&E images. These were subsequently co-registered with IR images at 1650cm^{-1} (the amide-I peak), from the same tissue core in order to extract IR data for analysis.

In order to correct for atmospheric scattering, extracted spectra were pre-processed using an open-source extended multiplicative scattering correction (EMSC) code provided by Kohler et al [34]. The following preprocessing steps were carried out on the dataset before a final classification step was performed using a logistic regression (LR) classifier. An unsupervised quality control check of all data was used to eliminate anomalous spectra through the use of the multivariate Hotelling's T^2 statistic [35, 36]. Spectra determined to have a T^2 value lying outside the 95^{th} percentile were deemed to be anomalous and were omitted from further analysis.

Vector normalisation was used in order to account for sample thickness; wavenumber absorbance features were mean-centered; and variance scaled to one; before a final principal component analysis (PCA) step to reduce dimensionality of the dataset. Seven principal components were taken to assist convergence when fitting the LR classifier. A large L2 regularisation term (1×10^5) was applied to the objective function when fitting the LR model in order to mitigate the potential for overfitting.

Scientific Python packages [37–39] were used to implement classification models and survival analysis. The classification power of an FTIR spectrum as a biomarker was estimated using the AUC of the receiver operating characteristic (ROC) curve, and a precision-recall (PR) curve. In order to obtain an estimate of the variability of the classification power of FTIR, bootstrap out-of-bag sampling was utilised as follows. A training data set was constructed by drawing 80% of patients in the total dataset without replacement. The remaining 20% was used as the "out of bag" test set on which fitted models were evaluated, and statistics calculated. This process was repeated 100 times ensuring that no two sample sets were identical. When fitting the

LR model, data points were inversely weighted against differing number of acquired spectra per patient and by risk group to mitigate the imbalanced nature of the dataset. Predictions of risk group from the LR model are outputted as a list of probabilities for each risk group. Final prediction scores for each patient are taken to be the median probability predicted for each patient.

Calculated statistics included: AUC, Matthews correlation coefficient (MCC), specificity, sensitivity, positive predictive value (PPV), and negative predictive value (NPV) all statistics were calculated using appropriate weightings to compensate for class/patient imbalance and to ensure a classification statistics were not skewed. Prognostic efficacy was investigated using Kaplan-Meier survival-analysis, a Cox proportional hazards regression, and a log-rank test.

Prediction of patient outcomes

To facilitate improved clinical decision making, it was decided to stratify the cohort into "high" and "low" risk categories. The choice of risk group for each patient was determined through an optimisation routine which maximised the log-rank statistic with respect to the groupings of patients solely using outcome data. The optimisation procedure was performed using a genetic algorithm (GA) based approach utilising the distributed evolutionary algorithms for python (DEAP) library [40]. The "individuals" involved in the GA routine are vectors comprising the identity of the risk group of each patient in the analysis. The "fitness" of an individual set is the log-rank statistic calculated using the patient risk groups specified by that individual.

Maximum individual fitness of the generation against the generation number ~~which~~ showed a plateauing of the log-rank statistic at around 40 (fig. 1A). The resulting risk groups show a clear distinction in clinical outcomes (fig. 1B,C).

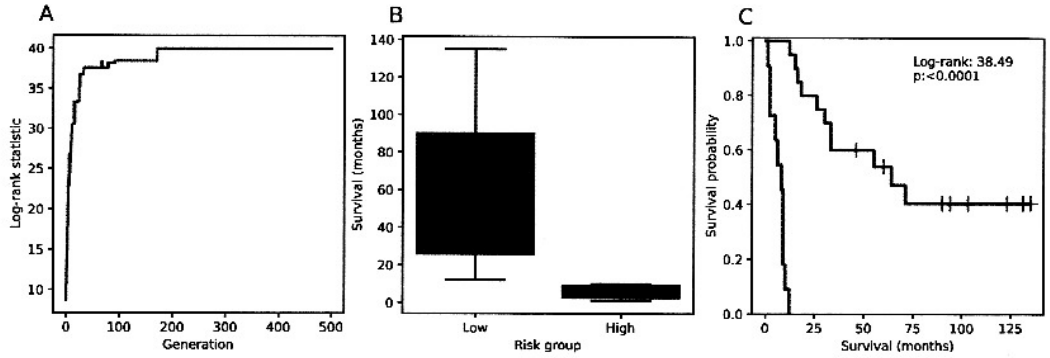

Figure 1: Stratification of patients into high and low risk. A: Maximum log-rank statistic vs GA generation, plateauing around 40. B: whisker box plots of survival duration in each risk group. C: Kaplan-Meier Survival curves showing optimal risk stratification of the patient cohort, also shown are the log-rank statistic and corresponding p-value for the optimal groupings.

The risk groups derived from the GA process were then utilised as prediction targets for classification. This process sought to identify objective groupings to identify using FTIR and ASMA as predictive variables.

3 Results

The inclusion criteria designated in Materials and Methods gave a sample set of 29 patients from the original 102 previously described [17]. The cohort was representative of this original patient set except that the male:female ratio was inverted (Table 1) but now corresponds more closely with larger datasets (Table 1; final column). In agreement with the published cohort, our sample set was enriched for cases with extracapsular spread (i.e. poor prognosis) in comparison to the general HNSCC population [41]. Of these 29 patients, 19 remained alive 12 months after surgery, while 14 survived to two years.

Table 1: Characteristics of the sample cohort

	All (N=29)	Outcome at 12 months		Outcome at 24 months		Original cohort [17]	Larger, local cohort [41]
		Dead	Alive	Dead	Alive	N=102	N=489
Age (years)							
Mean	60	70.4	56.3	64.5	57.2	60	
Range	29-85	59-85	29-72	48-85	29-68	29-89	
Median	61	70.5	56.5	64	60		
α-SMA							
High/Intermediate	26 (94)					60 (64)	ND
Low	3 (6)					33 (36)	ND
Gender							
F	7 (24)†	0	7	2	5	57 (56)[a]	187 (38)[b]
M	22 (76)	10	12	13	9	45 (45)	302 (62)
T Stage							
1	1 (3)	0	1	0	1	8 (8)	123 (25)[b]
2	14 (48)	4	10	7	8	57 (54)	175 (35)
3	2 (7)	0	2	2	0	12 (11)	47 (10)
4	8 (28)	4	4	3	5	20 (19)	144 (30)
4a	4 (14)	2	2	2	2	9 (8)	
N Stage							
0	7 (24)	2	4	1	5	38 (37)[b]	314 (64)[c]
1	7 (24)	2	4	2	4	18 (17)	64 (14)
2a	1 (3)	1	0	1	0		101(20)
2b	16 (55)	4	9	8	5	45 (44)	
2c	3 (10)	1	2	3	0		
Pathological Site							
Floor of mouth	8 (28)	2	6	4	4	35 (34)[b]	162 (33)
Other	12 (41)	5	7	5	7	24 (24)	183 (36)
Tongue	9 (31)	3	6	6	3	37 (36)	144 (30)

a: $p<0.005$; b: p=NS; c: $p<0.00001$
*: T stage 1+2 v 3+4

†: numbers in parenthesis are percentages
#: N stage 0 v 1 v 2

Suggest: a – *, b – +, c – #

otherwise b and c get lost.

Prediction of death within one year

FTIR and ASMA data from the reduced sample set were evaluated as prognostic indicators of death within one year of surgery, both separately and together (Figure 2).

A total of 168,460 FTIR spectra were obtained from the 19 patients who survived beyond one year and 96,402 spectra were obtained from 10 patients who died within 12 months.

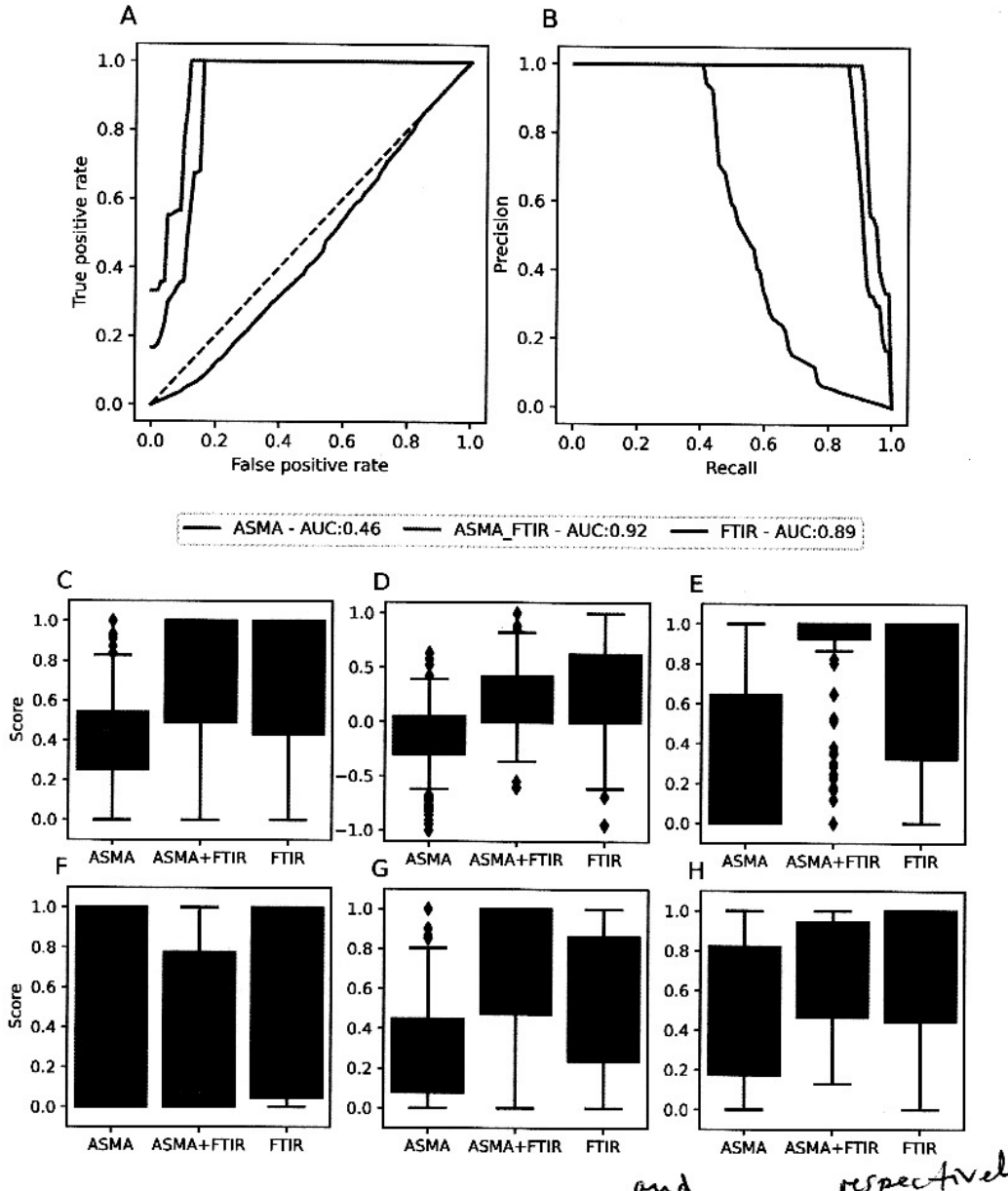

Figure 2: Median ROC and PR curves shown in solid/dashed lines showing classification power of the LR classifier. AUC scores are shown for each set of prognostic indicators. Whisker boxplots of classification statistics calculated across all datasets subsets: AUC (C); Matthew's correlation coefficient (D); specificity (E); sensitivity (F); positive predictive value (G); negative predictive value (H). Boxes show the median, 25^{th}, and 75^{th} percentiles; whiskers extend to points that lie within 1.5 inter quartile ranges of the lower and upper quartiles; points lying outside this range are shown as individual diamonds.

The median AUC (area under the receiver operating characteristic curve) obtained from FTIR alone was 0.89 (Figure 2 A,C); incorporation of the ASMA data into this analysis, increased the AUC slightly to 0.92, while ASMA alone achieved a significantly poorer score of 0.46. Precision and recall scores both remained high for the FTIR

and combined ASMA/FTIR models across a range of decision thresholds, indicating that both models can balance both statistics effectively (Figure 2 B). Furthermore, additional classification statistics produced comparable conclusions, showing that the model was a good predictor of poor outcome (Figure 2 C-H).

Table 2: Median classification statistics

	AUC	F1	MCC	specificity	sensitivity	PPV	NPV	threshold
ASMA	0.46	0.22	0.00	0.18	0.50	0.19	0.48	0.48
ASMA+FTIR	0.92	0.51	0.00	1.00	0.08	0.95	0.66	0.69
FTIR	0.89	0.54	0.17	0.83	1.00	0.62	0.85	0.34

Scores show (table 2) that ASMA alone is a poor predictive variable for this dataset with low scores in all metrics. The joint ASMA+FTIR model shows high specificity and low sensitivity, with a high PPV and moderate NPV; numerous false negatives indicate the model struggles to identify patients with better prognoses. The FTIR model scores are consistently high, showing FTIR is prognostically informative. A slightly lower PPV suggests the FTIR model incurs more false positives - patients with better prognoses will be given inappropriate treatment.

A large degree of variation was observed across some statistics, potentially signifying a large degree of biological heterogeneity in the dataset. Classification thresholds (table 2) used to convert probabilities to binary decisions were determined to be those that gave an maximal log-rank test/most significant result.

Figure 3[A-C] survival curves for high(low) risk groups are shown in red(blue) for each model. Figure 3[A] shows some overlap between risk groups, but risk group predictions are inverted to what might be expected. With respect to this inverted prediction, a significant log-rank test p-value of 0.03 is misleading. Figure 3[B] shows a clear distinction between risk groups ($p < 0.005$; indicating that the model is prognostically useful. Figure 3[C] is again significant but with less prognostic utility than the combined model.

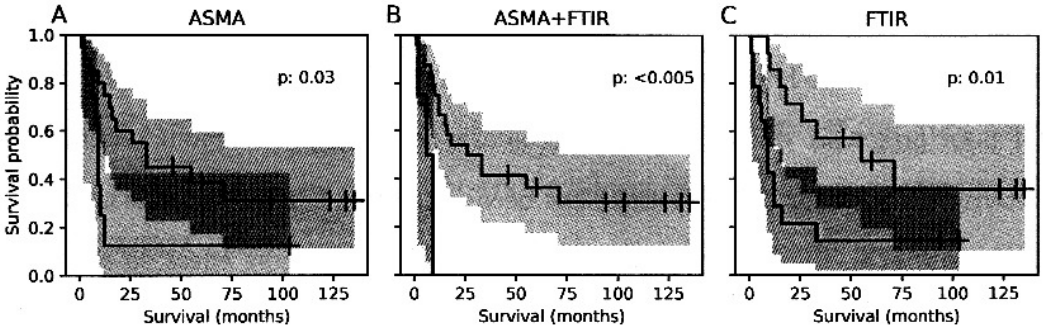

Figure 3: Kaplan-Meier survival curves for each risk group according input variables. Low-risk:blue, high-risk:red. Confidence intervals are computed using the exponential Greenwood method [42].

A univariate Cox's proportional hazard model was fit to prediction scores outputted by the LR model to assess the prognostic utility of the prediction score before conversion to a binary decision. Both models using FTIR data have significantly higher hazard ratios than the pure ASMA model, suggesting the LR classifier is able to stratify risk groups effectively.

Table 3: Cox proportional hazards model fit statistics

	Coefficients	z	p	Hazard ratio
ASMA	-0.03	-0.02	0.98	0.97 (0.06-15.08)
ASMA+FTIR	1.84	2.12	0.03	6.29 (1.14-34.59)
FTIR	2.02	2.07	0.04	7.52 (1.12-50.62)

4 Discussion

Results obtained here show that FTIR was proven to be capable at stratifying a patient cohort into useful clinical risk groups. Survival analysis showed that groups allocated by the classifier had significantly different outcomes. The joint model (fig. 3 [B]) is nearly able to perfectly replicate survival the ideal survival curves in (fig. 1 [C]). Model predictions using FTIR data showed a marked improvement over the pure ASMA model. Hazard ratios of 6.29 and 7.5 for the ASMA+FTIR and FTIR model respectively, show that the prognoses of patients allocated to the high-risk group are significantly poorer.

A significant improvement was observed when taking the median prediction score for each patient. This is potentially a reflection of the fact that the molecular fingerprint for poor prognosis varies in magnitude across the measured tumour section - genetic heterogeneity within OSCC lesions has been noted previously [43]. The aggregation of scores across an entire tumour section may have a regularising effect on the final score - reducing the variance in the final score, thus mitigating the effects of overfitting. This regularisation effect may be dependent upon the size of the measured tumour section present within the core. Data subsets containing patients with relatively few measured spectra may experience lower scores as a result; possibly explaining the large degree of variation within sample scores to some extent.

The use of FTIR in clinical diagnostics is growing quickly, ~~however~~ relatively little work has been aimed towards prognostic biomarkers. The combination of both FTIR microscopy with techniques more familiar to oncologists such as immuno histo chemical (IHC) staining has the potential to improve prognostic predictive capabilities significantly as shown in this work.

Many other prognostic biomarkers exist [44–46] but a large proportion are still in the "discovery phase" - requiring further study to ascertain prognostic benefit [47]. The use of FTIR within clinical diagnostics likely fits into this category, as many potential barriers facing other potential biomarkers are still present.

A relatively small sample set was a key issue facing this study due to the difficulty in acquiring and imaging large numbers of samples. Despite attempts to determine the feasibility of FTIR as a prognostic tool through multiple sampling of the dataset; a larger study would likely need to be conducted in the future to estimate wider clinical utility.

5 Conclusion

The use of FTIR in a clinical setting is still in its infancy, however the work covered here shows that it has the potential to be of significant benefit as a prognostic tool. The addition of ASMA information was shown to be beneficial in one case, and demonstrates that additional information from other modalities could lead to the creation of a novel and informative prognostic tool. In order to progress the work covered here a larger patient sample set would be needed to affirm the conclusions reached here. The addition of information from a broader range IHC stains or imaging modalities could further increase the prognostic benefit outlined here and yield powerful tools for clinicians.

References

[1] *Head and neck cancers statistics.* Sept. 2021. URL: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/head-and-neck-cancers#heading-Zero.

[2] Mauricio E. Gamez et al. "Treatment outcomes of squamous cell carcinoma of the oral cavity in young adults". In: *Oral Oncology* 87.August (2018), pp. 43–48. ISSN: 18790593. DOI: 10.1016/j.oraloncology.2018.10.014. URL: https://doi.org/10.1016/j.oraloncology.2018.10.014.

[3] Torbjörn Ramqvist and Tina Dalianis. "An epidemic of oropharyngeal squamous cell carcinoma (OSCC) due to human papillomavirus (HPV) infection and aspects of treatment and prevention." In: *Anticancer research* 31.5 (2011), pp. 1515–9. ISSN: 1791-7530. URL: http://www.ncbi.nlm.nih.gov/pubmed/21617204.

[4] Matt Lechner et al. "HPV-associated oropharyngeal cancer: epidemiology, molecular biology and clinical management". In: *Nature Reviews Clinical Oncology* 0123456789 (2022). ISSN: 1759-4774. DOI: 10.1038/s41571-022-00603-7.

[5] Victor Lopes et al. "Squamous cell carcinoma of the oral cavity rarely harbours oncogenic human papillomavirus". In: *Oral Oncology* 47.8 (2011), pp. 698–701.