Journal of Universal Computer Science, vol. 24, no. 6 (2018), 742-752
submitted: 22/5/17, accepted: 23/4/18, appeared: 28/6/18 © J.UCS

Adapting an Evidence-based Diagnostic Model for Predicting Recurrence Risk Factors of Oral Cancer

Chien-Sheng Cheng
(Department of Pediatrics, Jen-Ai Hospital, Taichung, Taiwan
johnson92144@yahoo.com.tw)

Pei-Wei Shueng
(Division of Radiation Oncology, Far Eastern Memorial Hospital, New Taipei City, Taiwan
Faculty of Medicine, School of Medicine, National Yang-Ming University, Taipei, Taiwan
Department of Radiation Oncology, Tri-Service General Hospital
National Defence Medical Center, Taipei, Taiwan
shuengsir@gmail.com)

Chi-Chang Chang
(School of Medical Informatics, Chung Shan Medical University, Taichung, Taiwan
IT Office Chung Shan Medical University Hospital, Taichung, Taiwan
threec@csmu.edu.tw, corresponding author)

Chi-Wen Kuo
(Department of Pharmacy, Jen-Ai Hospital, Taichung, Taiwan; School of Pharmacy, China Medical University, Taichung, Taiwan; Basic Medical Education Center, Central Taiwan University of Science and Technology, Taichung, Taiwan; Department of Health Policy and Management, Chung Shan Medical University, Taichung, Taiwan
kcwen@ms3.hinet.net)

Abstract: Although the relationship between prognosis and oral cancer has been extensively investigated, its impact on recurrence and surgical margin has not been well studied. Clinical evaluation of a positive surgical margin in recurrent oral cancer is often challenging. The aim of this study was to propose an evidence-based diagnostic model using machine learning techniques for the prediction of risk factors of recurrent oral cancer. In addition, the performance of each technique was evaluated using accuracy, sensitivity, specificity, Fallout, F1 score, and Matthews correlation coefficient (MCC). An oral cancer dataset was provided by cancer registries of three hospitals in Taiwan. Of the 1,428 patients included in the current study, each patient in the dataset had 20 predictor variables. The results indicated that the KSTAR technique showed the best performance compared with other techniques. The GainRaito (RT) method was used in the screening to exclude five insignificant variables. The KSTAR technique also showed larger values for accuracy (77.04%), recall (77.98%), specificity (75.48%), Fallout (36.62%), F1 score (81.17%), and MCC (50.54%). Furthermore, the important risk factors for predicting recurrence in relation to the surgical margin in oral cancer were pathologic stage, behavior code, and lifestyle factors (smoking and betel nut chewing). Application of this proposed diagnostic model may facilitate targeted intervention to reduce the incidence of recurrence; however, our results suggest that adaptive machine learning techniques require incorporation of significant variables for optimal prediction.

Keywords: Oral Cancer, Recurrence, Machine Learning Techniques, Diagnostic Model
Categories: I.2.1, M.4

1 Introduction

Over the past few decades, the number of cases of cancer has been increasing worldwide. Oral cancer is the fifth most common cancer globally and the most common head and neck cancer [Fitzmaurice et al., 2016]. The worldwide incidence of oral cancer of 529,500, accounting for 3.8% of all cancer cases in 2017, has been predicted to increase by 62% to 856,000 cases by 2035 [Shield et al., 2017]. The number of long-term oral cancer survivors has increased recently; recurrence can reflect the late sequelae of treatment in this trend. In Taiwan, the incidence rate of oral cancer is higher than that in America or Europe. In 2017, oral cancer was the sixth most common neoplasm in Taiwan. Data from Taiwan Cancer Registry show that the annual incidence rate of oral cancer has increased from 3.46 cases per 100,000 people in 1979 to 22.69 cases per 100,000 people in 2015 (Figure 1 and Table 1)."

The 5-year survival for oral cancer depends on the stage at diagnosis. In general, the survival rates are as follows: stage 0 (76.6%), stage I (80.3%), stage II (70.5%), stage III (56.0%), and stage IV [The Ministry of Health and Welfare of Taiwan, 2018]. Wang et al. (2012) had reported the recurrence rate of oral cancer to be approximately 35.5% in Taiwan. For early-stage disease, surgery alone or in combination with local therapy is generally curative. Once the primary treatment has failed, the opportunity for a secondary cure is slim. Till date, the mechanisms involved in the occurrence of recurrence have not been elucidated. Therefore, there is a lack of adequate information about a causal relationship between risk factors and recurrence [Chang, 14]

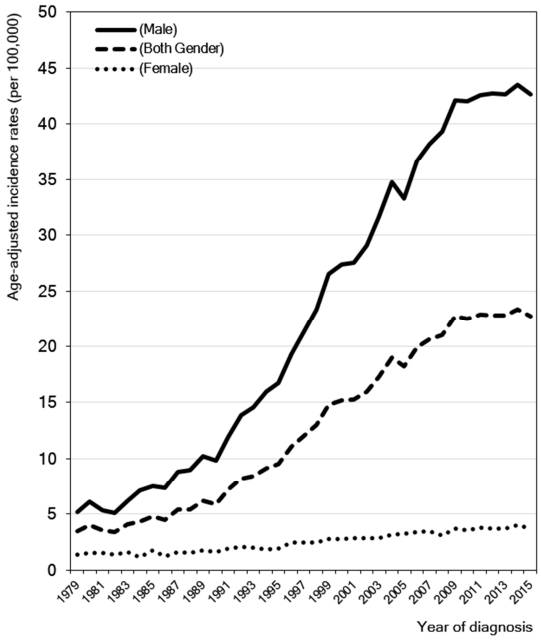

Figure 1: Incidence Rates of Oral Cancer in Taiwan, 1979–2015. (Source: Taiwan Cancer Registry, 2018)

	Both Gender		Male		Female	
Year of diagnosis	NO. of cases	Age-adjusted incidence rates	NO. of cases	Age-adjusted incidence rates	NO. of cases	Age-adjusted incidence rates
1979	439	3.46	364	5.17	75	1.35
1980	509	3.98	424	6.08	85	1.49
1981	469	3.58	380	5.34	89	1.52
1982	445	3.37	361	5.10	84	1.38
1983	552	4.06	453	6.21	99	1.57
1984	620	4.34	541	7.13	79	1.17
1985	687	4.80	574	7.47	113	1.74
1986	662	4.44	581	7.35	81	1.19
1987	837	5.39	718	8.79	119	1.59
1988	878	5.42	760	8.94	118	1.54
1989	1018	6.21	879	10.23	139	1.76
1990	989	5.86	865	9.80	124	1.58
1991	1247	7.16	1087	11.92	160	1.94
1992	1485	8.19	1309	13.87	176	2.05
1993	1570	8.47	1397	14.54	173	1.98
1994	1751	9.12	1586	15.99	165	1.82
1995	1866	9.53	1684	16.74	182	1.94
1996	2242	11.08	2004	19.31	238	2.45
1997	2504	12.02	2261	21.23	243	2.46
1998	2804	13.05	2553	23.35	251	2.44
1999	3257	14.76	2962	26.50	295	2.75
2000	3458	15.19	3160	27.40	298	2.74
2001	3588	15.27	3266	27.54	322	2.86
2002	3851	15.99	3518	29.06	333	2.81
2003	4318	17.28	3964	31.63	354	2.88
2004	4854	18.95	4459	34.76	395	3.15
2005	4799	18.21	4384	33.29	415	3.20
2006	5390	19.94	4938	36.71	452	3.35
2007	5728	20.65	5243	38.12	485	3.46
2008	5986	21.02	5541	39.34	445	3.10
2009	6605	22.68	6050	42.15	555	3.69
2010	6724	22.47	6184	42.02	540	3.51
2011	7003	22.85	6410	42.56	593	3.77
2012	7153	22.82	6557	42.72	596	3.67
2013	7330	22.77	6706	42.65	624	3.72
2014	7660	23.31	6969	43.51	691	4.01
2015	7628	22.69	6965	42.62	663	3.69

Note 1: This is the incidence of invasive cancer data which was calculated by using the mid-year population.
Note 2: Age-adjusted incidence rate was calculated by the direct method using the 2000 WHO world standard population.

Table 1: Statistical Trend of Oral Cancers in Taiwan (1979-2015)

Traditionally, the clinical diagnosis of recurrent oral cancer was dependent on the physician's experience with various risk factors. However, as the risk factors are of broad categories, years of clinical study and experience have attempted to identify key risk factors for recurrence, such as recurrent ovarian cancer [Tseng, 17] and recurrent cervical cancer [Tseng, 13].

While the major of surgical treatment is to remove all local malignant with no residual malignant cells left [Reis, 17]. The status of the surgical resection and several clinical characteristics are important predictors of outcome for recurrence in oral cancer [Wong, 12]. In addition, the number of long□term oral cancer survivors have increased, as recurrence are considered to be an important challenge for clinical management. Nevertheless, the recurrence risk factors of oral cancer have not yet been clarified in Taiwan. The purpose of the present study was to develop an evidence-based diagnostic model for predicting risk factors of recurrent oral cancer. The remainder of this paper is organized as follows. Section 2 gives a brief introduction about KNN-IBK, KSTAR, Randonizable-Filtered-Classifier and RandomTree algorithms. The proposed medical diagnostic scheme is described in Section 3. Section 4 presents the experimental results. Finally, the paper is concluded in Section 5.

2 Research Methods

Taking into account the characteristics of clinical recurrence, we put forward biased machine learning techniques to analyze and predict the associations between variables and recurrence effectively and accurately. Furthermore, four classification methods and different association rule techniques were combined with the three cancer registry databases for constructing classification models and extracting association rules. The models and the rules can be used to identify the risk factors to predict recurrence in patients with oral cancer. The test dataset contained 20 independent variables, with one dependent variable being the presence or absence of recurrence. This study was designed to compare the differences in the screening of important variables, as shown in Figure 2. In the research design, the classifiers were evaluated under the following two scenarios: (1) features with their original 20 independent variables and (2) features with screening to exclude five insignificant variables.

In addition, using 10-fold cross-validation and cross-validation, the dataset was split into 10 subsets. Finally, the risk factors extracted from the classification models and association rules were used to provide valuable information for clinical outcomes. All classification algorithms were implemented in the Weka toolkit [WEKA, 2018]. In this process, the following four most often used machine learning algorithms were selected:

K-Nearest Neighbor (KNN-IBK)–A assessing nearest neighbor approach, where the distance between two feature variables is calculated and decide a belong class is assigned based on the nearest neighbor [Aha et al., 1991; Brighton and Mellish, 2002].

KSTAR (instance-based classifiers)–The KSTAR model was able to attain higher predictive sensitivities and specificities based on random oversampling techniques. In general, it uses an entropy-based distance function to evaluate the similarity between two classes [Wang et al., 2006; Seyed et al., 2016].

RandomizableFilteredClassifier (RFC)–The structure of FilteredClassifier, which is based on attribute weights, has been passed through an arbitrary filter. Before they are passed to the classifier, interface to something that has random behavior that is able to be seeded with an integer. In general, it is useful for constructing an ensemble classifier using the RandomCommittee meta learner [Seyed et al., 2016].

RandomTree (RT)–The RandomTree is based on bagging, which implies no pruning and only selects several properties to construct a tree instead of selecting all the properties (Dugan et al., 2015). The process of selecting and generating split points is as follows [Imbus et al., 2017]: (1). Set a number of attributes K; (2). Sampling attributes without replacement of all attributes; (3). Calculate the information gain of the target attribute; (4). Repeat K times to decide the split node when the information gain is the largest; and (5). Construct the child's tree.

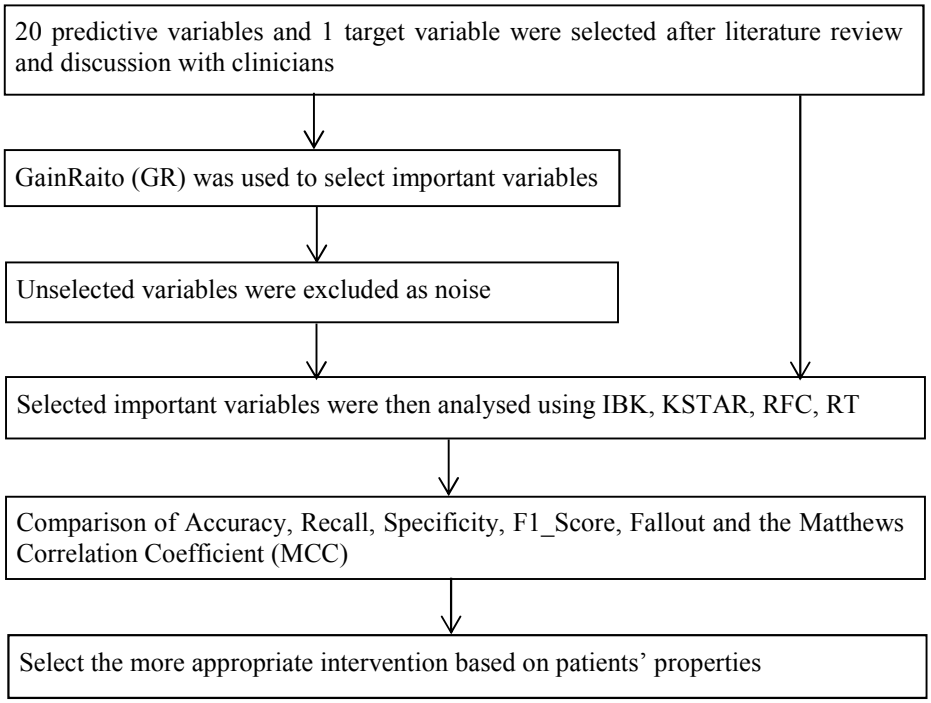

Figure 2: Design of the Research Process

Based on the research design, the GainRaito (GR) method was used to screen and rank features by calculating the information gain of the features, which is based on entropy. Entropy is a commonly used measure in the information theory, which characterizes the purity of an arbitrary collection of examples. The performance of the classification algorithms was evaluated using accuracy, recall, specificity, F1 score, Fallout, and the Matthews correlation coefficient (MCC), which are well-known and

standard measures for evaluating the proposed method. These criteria are calculated as follows

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Recall = \frac{TP}{P} = \frac{TP}{TP+FN} \tag{2}$$

$$Fallout = \frac{FP}{N} = \frac{FP}{FP+TN} \tag{3}$$

$$Specificity = \frac{TN}{N} = \frac{TN}{FP+TN} = 1 - Fallout \tag{4}$$

$$F1_score = \frac{2 \times TP}{2 \times TP+FP+FN} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{((TP+FP)(TP-FN)(TN+FP)(TN+FN))}} \tag{6}$$

Where TP, TN, FP, and FN stand for the case of true positives, true negatives, false positives, and false negatives, respectively. MCC stands for Matthews correlation coefficient.

3 Empirical Study

In this study, the oral cancer dataset provided by the Chung Shan Medical University Hospital, the Jen-Ai Hospital, and the Far Eastern Memorial Hospital Tumour Registry was used to verify the feasibility and effectiveness of K-nearest IBK, KSTAR, RandomizableFilteredClassifier (RFC), and RandomTree (RT). The data of each patient in the dataset contained 20 predictor variables, as follows: (1) age, (2) primary site, (3) histology, (4) behavior code, (5) differentiation, (6) tumor size, (7) pathologic stage, (8) surgical margin, (9) surgical, (10) radiotherapy (RT), (11) radiotherapy (RT) surgery, (12) sequence of local regional therapy and systemic therapy, (13) dose to clinical target volumes (CTV)_high, (14) number to clinical target volumes (CTV)_high, (15) dose to clinical target volumes (CTV)_low, (16) number to clinical target volumes (CTV)_low, (17) body mass index (BMI), (18) smoking, (19) betel nut chewing, and (20) drinking. The data also contained one dependent variable, i.e., recurrence or not. Excluding incomplete records, there were a total of 1,429 patients in the dataset. All datasets were scaled into the range of [-1.0, 1.0] by utilizing min-max normalization method before using machine learning techniques. Further, we repeated each machine learning technique using the training dataset and evaluated it using the validation dataset. The WEKA settings of all machine learning techniques were applied as shown in Table 2.

Machine Learning Techniques	Classification Scheme	Settings
IBK	Lazy	Classifier Model: IB1 instance-based classifier using 1 nearest neighbor for classification
KSTAR	Lazy	The KSTAR algorithm implemented in WEKA
RandonizableFiltered Classifier (RFC)	Meta	Classifier Model: IB1 instance-based classifier using 1 nearest neighbor for classification
RandomTree	Trees	Seed number: 1, Number of generated trees: 98

Table 2: Machine learning techniques used and the settings of training parameters

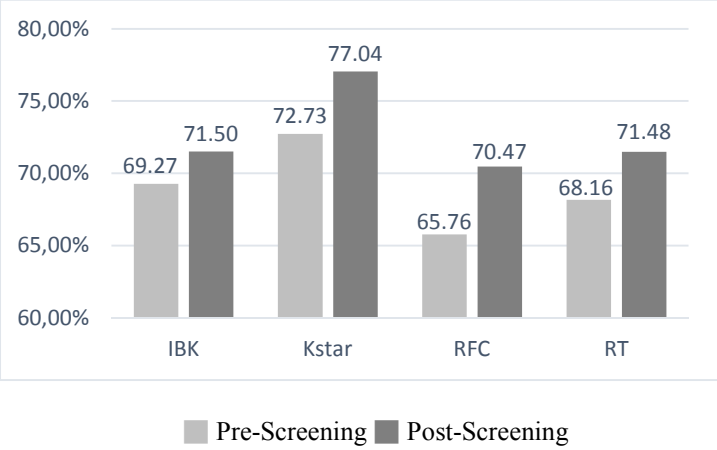

Figure 3: Accuracy comparison of machine learning techniques between pre- and post-screening

Methods	Accuracy	Recall	Specificity	Fallout	F1_Score	MCC
KSTAR	72.73	71.60	72.69	27.04	70.31	43.33
IBK	69.28	66.97	70.00	31.83	67.84	36.94
RFC	65.76	63.24	66.50	35.30	64.05	29.67
RT	68.16	66.14	68.56	31.96	66.42	34.73

Table 3(a): The Performance of the Machine learning techniques of Pre-Screening

Methods	Accuracy	Recall	Specificity	Fallout	F1_Score	MCC
KSTAR	77.04	77.96	75.48	36.62	81.17	50.54
IBK	71.50	73.57	68.46	38.64	76.11	41.16
RFC	70.47	73.46	66.08	37.98	74.97	39.08
RT	71.48	73.79	68.01	38.17	76.04	41.06

Table 3(b): The Performance of the Machine learning techniques of Post-Screening

Screening results after literature review and discussion with clinicians showed that in the part of pre-screening, the accuracy with the surgical margin was the best using KSTAR (67.78%), and the ranking of the risk factors was as follows: behavior code, radiotherapy, number to clinical target volumes (CTV)_high, dose to clinical target volumes (CTV)_high, and pathologic stage. Moreover, the accuracy without the surgery margin was the best using KSTAR (77.40%), and the ranking of the risk factors was as follows: radiotherapy surgery, behavior code, tumor size, histology, and pathologic stage. It appears that behavior code affects the incidence of recurrent oral cancer. Conversely, in the part of post-screening, the accuracy with the positive surgical margin was the best using IBK (68.61%), and the ranking of the risk factors was as follows: betel nut chewing, smoking, behavior code, and difference. Furthermore, the accuracy with the negative surgical margin was the highest using KSTAR (81.35%), and the ranking of the risk factors was as follows: behavior code, tumor size, pathologic stage, and sequence. Considering a similar case in the pre-screening, one of these factors is behavior code, which SEER has defined as the "Fifth digit of the ICD-O Morphology code which designates the malignancy or behavior of this tumor." The behavior code is a common risk factor and has been frequently described as an important predictor of recurrence [April, 18]. In addition to lifestyle factors, the role of betel nut chewing has been investigated by several researchers [Chou, 17]. Furthermore, several investigators have studied the role of smoking in recurrence in oral cancer [Bezerra, 18]. The results of analysis of variance between pre- and post-screening are presented in Figure 4.

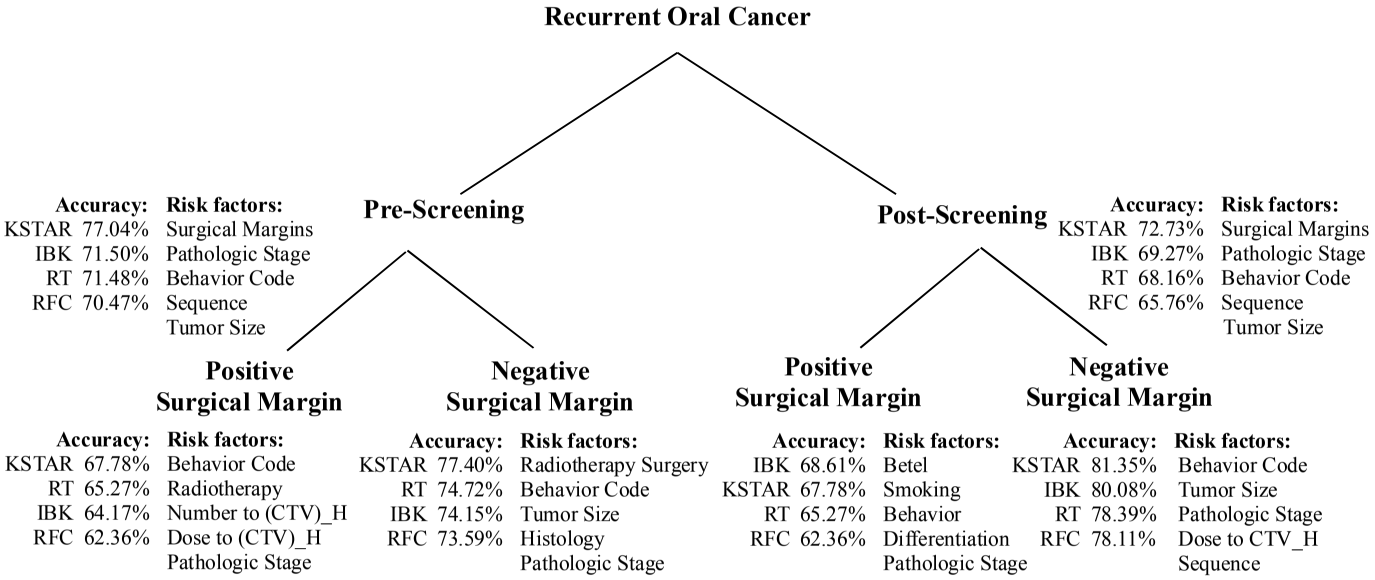

Figure 4: Accuracy-Risk Factors Treemap of Machine learning techniques between Pre-and Post-Screening

4 Conclusion

Recently, there has been an increase in the population of oral cancer survivors. This trend essentially reflects that an early prediction of recurrent risk factors is necessary in oral cancer research. Furthermore, to increase the cure rates and the effective medical resources, it is extremely important to discover the factors predicting recurrence in the actual diagnosis and treatment records for clinicians. The objective of this study was to propose an evidence-based diagnostic model using machine learning techniques for the prediction of risk factors of recurrence in oral cancer survivors. Results showed that KSTAR was better than other techniques in both pre- and post-screening. Our results demonstrated that surgical margin is the most precondition risk factor for recurrence of oral cancer. Furthermore, after screening the five insignificant parameters, various factors associated with positive or negative surgical margin were found to be risk factors for predicting the recurrence outcome. In addition, the lifestyle risk factors associated with recurrent oral cancer included smoking and betel nut chewing. Moreover, the combined effects were often not only additive but also greater than multiplicative. However, the influence of both these risk factors declined rather rapidly following cessation of the habits, with the relative risks decreasing compared with those among nonsmokers. Given the association with lifestyle risk factors, it is important that we continue emphasizing the impact of quitting smoking and stopping betel nut chewing. Clinicians may use the prediction results as a reference and could make better clinical decisions.

Acknowledgements

This work is supported by the Chung Shan Medical University and Jen-Ai Hospital: CSMU-JAH-104-02.

References

[Aha, 91] Aha, D., Kibler, D., Albert, M. (1991) Instance-Based Learning Algorithms. Mach Learn, 6, pp. 37-66.

[April, 18] April F., Bethesda, M., Andrew, J., Kanagaratnam, S., Leslie, S., Max, P., Sharon, W. (2018) lnternational Classification of Diseases for Oncology V3.0, 2018, http://whqlibdoc.who.int/publications/2000/9241545348_eng.pdf (January 11, 2018)

[Bezerra, 18] Bezerra, N.V.F., Leite, K.L.F., de Medeiros, M.M.D., Martins, M.L., Cardoso, A.M.R., Alves, P.M., Padilha, W.W.N., Cavalcanti, Y.W. (2018) Impact of the anatomical location, alcoholism and smoking on the prevalence of advanced oral cancer in Brazil. Med Oral Patol Oral Cir Bucal. 23(3), pp.e295-e301.

[Brighton, 02] Brighton, H., Mellish, C. (2002) Advances in Instance Selection for Instance-Based Learning Algorithms. Data Min Knowl Discov, 6, pp.153-172.

[Chang, 14] Chang, C. K., Chang, C. C. (2014) Bayesian Imperfect Information Analysis for Clinical Recurrent Data. Therapeutics and Clinical Risk Management. 11, pp.17-26.

[Chou, 17] Chou, C.H., Chou, Y.E., Chuang, C.Y., Yang, S.F., & Lin, C.W. (2017) Combined effect of genetic polymorphisms of AURKA and environmental factors on oral cancer development in Taiwan. PLoS ONE, 12(2), pp.e0171583.

[Dugan, 15] Dugan, T., Mukhopadhyay, S., Carroll, A., Downs, S. (2015) Machine Learning Techniques for Prediction of Early Childhood Obesity. Applied Clinical Informatics, 6(3), pp.506-520.

[Fitzmaurice, 17] Fitzmaurice, C., Allen, C., Barber, R., Barregard, L., Bhutta, Z., Brenner, H., Fleming, T. (2017) Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. JAMA Oncol., 3(4), pp.524-548.

[Imbus, 17] Imbus, J., Randle, R., Pitt, S., Sippel, R., Schneider, D. (2017) Machine learning to identify multigland disease in primary hyperparathyroidism. J Surg (219), pp.173-179.

[Nason, 09] Nason, R.W., Binahmed, A., Pathak, K.A., Abdoh, A.A., Sándor, G.K. (2009) What is the adequate margin of surgical resection in oral cancer? Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology. 107(5), pp.625-629.

[Ravi, 14] Ravi, S. B., and Annavajjula, S. (2014). Surgical Margins and Its Evaluation in Oral Cancer: A Review. Journal of Clinical and Diagnostic Research. 8(9), pp.ZE01–ZE05.

[Reis, 11] Reis, P., Waldron, L., Perez-Ordonez, B., Pintilie, M., Galloni, N., Xuan, Y. (2011) A gene signature in histologically normal surgical margins. BMC Cancer. 11, p.437.

[Seyed, 16] Seyed, A., Fatemeh, G., Sara, M. (2016) Applying data mining techniques for increasing implantation rate by selecting best sperms for intra-cytoplasmic sperm injection treatment, Computer Methods and Programs in Biomedicine. 137, pp.215-229.

[Shield, 17] Shield, C., Ferlay, J., Jemal, A., Sankaranarayanan, R., Chaturvedi, A., Bray, F., Soerjomataram, I. (2017) The global incidence of lip, oral cavity and pharyngeal cancers by subsite in 2012. CA: a cancer journal for clinicians, 67(1), pp. 51-64.

[The Ministry of Health and Welfare of Taiwan, 18] The Ministry of Health and Welfare of Taiwan (2018). Taiwan Breast cancer, Oral cancer, and Colorectal Cancer Screening Programs. Retrieved from https://www.hpa.gov.tw/EngPages/Detail. Aspx?nodeid=1051&pid=5957 (February 2, 2018)

[Tseng, 13] Tseng, C. J., Lu, C. J., Chang, C. C., Chen, G. D. (2013) Application of Machine Learning to Predict the Recurrence-Proneness for Cervical Cancer, Neural Computing and Applications, 24(6), pp.1311-1316.

[Tseng, 17] Tseng, C. J., Chang, C. C., Lu, C. J. (2017) Integration of ensemble learning and data mining techniques to predict risk factors for recurrent ovarian cancer. Artificial Intelligence in Medicine. 78, pp.47-54.

[Wang, 06] Wang, H., Zheng, H., Simpson, D., Azuaje, F. (2006) Machine learning approaches to supporting the identification of photoreceptor-enriched genes based on expression data. BMC Bioinformatics, 7, 116.

[Wang, 12] Wang W., Han, S., Yao, Z. (2012) A study of epidemiologic and recurrence factors of oral cancer. J Oral Maxillofac Surg. 70(9), pp. 2205-2210.

[Weka 3, 17] Weka 3. (2017) Data mining with open source machine learning software in java. Retrieved from http://www.cs.waikato.ac.nz/ml/weka/ (December 2, 2017)

[Wong, 12] Wong, L.S., McMohan, J., Devine, J., McLellan, D., Thompson, E., Farrow, A. (2012) Influence of close resection margins on local recurrence and disease-specific survival in oral and oropharyngeal carcinoma. British Journal of Oral and Maxillofacial Surgery. 50, pp.102-108.