**Quantitative risk stratification of oral leukoplakia with exfoliative cytology**

Yao Liu1,a, Jianying Li2,a, Xiaoyong Liu1, Xudong Liu3, Waqaar Khawar4, Fan Wang5, ……

Xiaoxin Luke Chen6,b, Zheng Sun1,b

1 Department of Oral Medicine, Beijing Stomatological Hospital, Capital Medical University, Beijing, 100050, China

2 Frontier Bioinformatics Solution, Cary, NC 27519, USA

3 Ongwandada Resource Center, Queen’s University, 191 Portsmouth Avenue, Kingston, ON, K7M 8A6, Canada

4 Ross University School of Medicine, 2300 SW 145th Avenue, Miramar, FL, 33027, USA

5

6 Cancer Research Program, JLC-BBRI, North Carolina Central University, Durham, NC, 27707, USA

a Equal contributions

b Correspondence to: Xiaoxin Luke Chen, MD, PhD, Cancer Research Program, Julius L. Chambers Biomedical Biotechnology Research Institute, North Carolina Central University, 700 George Street, Durham, NC 27707, USA. Tel.: 919-530-6425; Fax: 919-530-7780; Email: [lchen@nccu.edu](mailto:lchen@nccu.edu). or Zheng Sun, DMD, PhD, Department of Oral Medicine, Beijing Stomatological Hospital, Capital Medical University, 4 Tiantanxili, Doncheng District, Beijing, 100050, China. Tel.: 86-(10)-57099016; Email: [zhengsun12@vip.126.com](mailto:zhengsun12@vip.126.com). **Abstract**

Exfoliative cytology has been widely used for early diagnosis of oral squamous cell carcinoma (OSCC). Test outcome is reported as “negative”, “atypical” (defined as abnormal epithelial changes of uncertain diagnostic significance), and “positive” (defined as definitive cellular evidence of epithelial dysplasia or carcinoma). The major challenge is how to properly manage the “atypical” patients in order to diagnose OSCC early and prevent OSCC. In this study, we collected exfoliative cytology data, histopathology data, and clinical data of normal subjects (n=102), oral leukoplakia (OLK) patients (n=82), and OSCC patients (n=93), and developed a data analysis procedure for quantitative risk stratification of OLK patients. This procedure involving a step called expert-guided data transformation and reconstruction (EdTAR) allows automatic data processing and reconstruction and reveals informative signals for subsequent risk stratification assessment. Modern machine learning techniques were utilized to build statistical prediction models on the reconstructed data. Among the several models tested using resampling methods for parameter pruning and performance evaluation, Support Vector Machine (SVM) was found to be optimal with a high sensitivity (median>0.98) and specificity (median>0.99). With the SVM model, we constructed an oral cancer risk index (OCRI) for guiding the clinical follow-up of OLK patients. One OLK patient with an initial OCRI of 0.88 developed OSCC after 3.5 years of follow-up. In conclusion, we have developed a statistical method for qualitative risk stratification of OLK patients. We believe this method will improve cost-effectiveness of clinical follow-up of OLK patients, and help design clinical chemoprevention trial for high-risk populations.

**Key words:** Oral leukoplakia, Oral squamous cell carcinoma, exfoliative cytology, DNA index, oral cancer risk index

**Introduction**

Oral cancer is one of the major public health problems worldwide, as well as a major cause of cancer morbidity and mortality [1,2]. In the United States, approximately 28,030 new cases will be diagnosed and 5,850 cases will die in 2014 [1]. Oral squamous cell carcinoma (OSCC) is the most common type of oral cancer, which usually develops from precancerous lesions such as oral leukoplakia (OLK) and erythroplakia, and histopathologically follows a step-wise pattern of hyperplasia, dysplasia and SCC [3,4]. Overall survival of OSCC patients remained unchanged despite the advances in radiotherapy and chemotherapy [1]. The five-year survival rate for patients with early, localized lesions is ~80%, whereas it is only 19% for patients with distant metastasis [5]. Thus it is important to assess precancerous lesions and diagnose OSCC early.

OLK is defined as “a white plaque of questionable risk having excluded other known diseases or disorders that carry no increased risk for cancer” [6,7]. And the annual age-adjusted incidence rates of OLK varied from 1.1 to 2.4 in male and from 0.2 to 1.3 in female per 1,000 person in India, and the prevalence varied from 0.2 to 4.9% [8]. In Japan, the age-adjusted incidence rate was 40.9 in male and 7.0 in female per 1,000 person-years [9]. Histopathologically, OLK presents as hyperkeratosis of the squamous epithelium in oral cavity. Several months or years are needed for hyperkeratosis progress to cancer. The overall chance of malignant transformation is 3.6% [10] and can be up to 12.9% in some populations [11,12,13]. This situation creates a huge burden on health care and therefore, there is a need of risk stratification for OLK patients to improve the cost-effectiveness of clinical follow-up.

Several measures are available for clinicians to assess OLK lesions: (1) Visual assessment of mucosal appearance: Lesions with a red component, ulceration, or certain topography described (granular, nodular, or verrucous) are more likely to develop malignancy [7,14]. Being subjective in nature, this method of assessment depends on clinical experience of the examiner. Moreover, mucosal appearance of early-stage cancer may appear benign [14]. (2) Visual assessment of physico-chemical properties, such as toluidine blue staining [15], fluorescence spectroscopy [16]: These methods are easy and quick to use, yet less specific [14,15]. It is a challenge to distinguish the high-risk from low-risk OLK lesions [17]. (3) Laboratory assessment of molecular markers: Chromosome *in situ* hybridization, immunohistochemistry, real-time PCR, gene microarray and proteomics have been used for detection of alterations in DNA, mRNA and protein [18]. Although these molecular tools have shown promising results with improved accuracy of cancer diagnosis, they are usually expensive and require high-quality clinical samples. (4) Laboratory assessment of cellular markers: Exfoliative cytology in conjunction with DNA quantitative analysis [19], micronucleus analysis [20] and nucleolar organizer regions [21], has already been used routinely for diagnosis of OSCC in recent years [22]. Its sensitivity and specificity has been reported up to 100% [5,19,23,24,25]. Although qualitative assessment (“negative for OSCC”, “positive for OSCC”, or “atypical lesion”) works well for OSCC diagnosis, this method has limited use in assessing cancer risk of those negative and atypical cases. Additionally, it is unfortunate to lose data other than DNA index (DI) that can be collected by exfoliative cytology.

In this study, we developed a statistical method for quantitative risk stratification of OLK. Our main purpose is to distinguish high-risk OLK from low-risk OLK based on data collected by exfoliative cytology, and therefore to improve cost-effectiveness of clinical follow-up.

**Materials and Methods**

**Clinical subjects, clinical data and follow-up**

Exfoliated cells were collected from oral mucosa of patients with OLK (n=82), OSCC (n=93), and healthy subjects (n=102) in outpatient clinic of the Beijing Stomatological Hospital, Capital Medical University (Table 1). Those who smoked 1 or more cigarettes per day for at least ? year were regarded as smokers, and those who had 1 or more drinks per day for 3 or more times per week as drinkers [11]. The study was approved by the ethical committee of our institution, and all patients signed the informed consent before the study.

Every patients need to follow up to record the changes in signs and symptoms, and find whether there is a malignant transformation. PLEASE DESCRIBE IN DETAIL HOW TO FOLLOW UP PATIENTS

**Exfoliative cytology**

Exfoliative cells were collected in fixed liquid (Motic, China) before Feulgen staining.

The exfoliative cells were transferred to a dry glass slide using the liquid-based preparation. The smears were stained by Feulgen staining kit (Motic, China) according to the manufacturer’s instructions. DNA-image cytometry (Motic, China) was used for the measurements of the DI and others cytologic features in the Feulgen-stained slides. Fifteen percent (20 of 132) features was useful, such as DI, DNA amount, intensity, radius and area. In this study, we only use the DI value, as show in Figure 1.

Yao, YOU NEED TO DESCRIBE THE PROCEDURE IN DETAIL:

1. How to do this? Procedure, instruments, software, parameters collected from image (GIVE A FULL LIST), quality control,
2. You need to describe Figure 1A, B, C
3. What criteria for judging “negative”, “atypical” and “positive”? ADD ONE COLUMN TO SUPPLEMENTARY TABLE 1 AS “DIAGNOSIS” TO PUT THIS INFORMATION FOR EACH CASE OF OLK

**Histopathology**

For OLK and OSCC, a resection biopsy was taken immediately from the same area under local anesthesia after brush biopsy. Tissues were fixed with buffered formalin and processed for clinical histopathology. Paraffin tissue sections were evaluated by our pathologist according to the standard criteria of the WHO Classification System of Head and Neck Tumors (2006) [26]. DESCRIBE THE CRITERIA OF HISTOPATHOLOGY (mild dysplasia,

For healthy subjects, DNA quantitative analysis and Papanicolaou exam were used in diagnosing health oral

**Expert-guided data transformation and reconstruction (EdTAR)(Figure 2)**

In this proof-of-concept study, we only used DI for statistical analysis. EdTAR was made up of four parts, peak identification, extraction of diploid/tetraploid and isolation of aneuploid, signal amplification, and data reconstruction. Parameter estimation, signal amplification and data reconstruction were carried out with R ([R\_Core\_Team 2014](#_ENREF_29)).

**Peak identification**

We first aimed to differentiate three possible cell populations, diploid, tetraploid and aneuploidy. According to the currently accepted clinical diagnosis practice(ref? jyl), we defined the thresholds for peaks that represented each cell population, diploid [0.8, 1.2], tetraploid [1.5, 2.2] and hypertetraploid/aneuploidy [>2.3]. The DI values obtained from Classify normally represent a mixture of cell populations and can be displayed in a histogram (Figure 1C). In order to estimate the parameters for each individual cell population, we adopted the procedure of kernel density estimation with the kernel density estimator, (Equation 2), where represents a “kernel function” K and a “bandwidth” *h*. We assumed that DI values were independently selected from a background distribution. To smoothen the histogram, we chose Gaussian distribution as the kernel ([Wand 2000](#_ENREF_42)), and finalized on the bandwidth to minimize the mean integrated squared error (MISE), as (Equation 3). When two or more populations were observed with fairly large proportion of overlap, bimodal or multi-normal based assumption was made. For such cases, a reflection point was identified if a change in the first derivative sign was observed.

**Extraction of diploid/tetraploid peaks and isolation of aneuploid peak**

One key component in our approach was to extract non-informative cell populations, i.e. diploid/tetraploid cell population. To do so, we proposed the sequential steps in the following pseudo code.

1. With the parameter set θdt for data transformation
2. Candidate peaks obtained on the density distribution from the empirical DI values were stored
3. **for** each i = 1..n peaks **do**

3.1. Estimate the sample statistics from the left part of the peak

3.2. Estimate the right part of the distribution

3.3. Filter out the candidate ith family and retain the mean, SD and count

3.4. Check the next available peak, if any, against threshold

3.5. Go back to 3.1 if the peak is < upper bound

**end**

1. Summarize for candidate cell populations representing the diploid, tetraploid and aneuploid, and store the summary statistics of each population
   1. The total number of DI values
   2. Sample mean and sample standard deviation
   3. The number of peaks and location of the peaks

**Signal amplification**

Our main goal was to quantify the risk via sufficient stratification in an attempt to amplify the “real signal”. First, we defined ratios of these three populations as R1, R2, and R3 respectively, with the constraint that R1 + R2 + R3 = 1 (Equation 4). If all three cell populations were detected and their peaks were retained, we achieved the amplified signal of aneuploidy population by redistributing the ratio among R1, R2, and R3. The original ratio between two populations (R1 and R2 ) were retained and was together weighed as 0.9. If only diploid and tetraploid populations were detected, the original ratio between two populations (R1 and R2) were retained and was together weighed as 0.995, the hypothetical aneuploid population from ~ Norm (2.3, 0.3).If a single diploid population was detected, R1 will be sample from a uniform distribution ~Unif [0.75, 0.8], and kept R1 + R2 = 0.995 and R3 [1-R1-R2]. The hypothetical tetraploid population was sampled from a normal distribution ~ Norm (2.0, 0.3) and the hypothetical aneuploid population from ~ Norm (2.3, 0.3).

**Data reconstruction**

For data reconstruction, new variables were created to represent the discrete interval ranging between 0 and 8 (DI values) with 0.5 increments. For each interval, the density estimated from the actual data was used. If any interval is missing, 0.0001 was used as the filler. The procedure is shown in pseudo code format as follows:

1. With the parameter set θdr for data reconstruction
2. **If** only diploid population exists **do**

Sample the three population ratios

Integrate mixture of three theoretical families

Create densities for all 16 discrete intervals

1. **Else if** both diploid and tetraploid populations exist **do**

**If** tetraploid population mean and standard deviation exist **next**

**Else do**

Sample the tetraploid population mean and standard deviation

Sample the aneuploidy population

Compute the ratio between diploid and tetraploid populations, and sample the ratios of three populations

Integrate mixture of three theoretical families

Create densities for all 16 discrete intervals

1. **Else** all three populations exist **do**

**If** the maximum DI value ofaneuploidy cells > 8, set it as 8

**If** tetraploid population mean and standard deviation exist **next**

**Else do**

Sample the tetraploid population mean and standard deviation

Compute the ratio between diploid and tetraploid population

Finalize ratios for all three populations

Integrate mixture of three theoretical families

Create densities for all 16 discrete intervals

**Statistical models and performance evaluation**

Statistical modeling, variable selection and performance evaluation were done with R ([R\_Core\_Team 2014](#_ENREF_29)) and caret package (<http://caret.r-forge.r-project.org/>). Datasets of “normal subjects” (n=102) and “OSCC patients” (n=93) were used to build the prediction models. First of all, we randomly separated the dataset into two parts with 70% samples for model selection and optimization and 30% for testing and evaluation. We selected six statistical models and evaluated their performance, Support Vector Machine (SVM), Random Forest (RRF), Penalized Logistic Regression (PLR), Neural Network (NNET), K-nearest neighbor (KNN), and Classification and Regression Training (CART). To evaluate each model’s performance, we started with the default parameters and further optimized the hyperparameters to achieve the best performance. Evaluation involved a resampling process, which included 10-fold cross-validation within each pass and repetition for five times. To ensure objective evaluation, we implemented the same random data parsing procedures for internal cross-validation by setting the same seed for any random number generation ([Kuhn 2013](#_ENREF_17)). These models were ranked according to the area under receiver operating characteristic (ROC), sensitivity and specificity. Based on the performance evaluation, the SVM model was chosen for the following calculation.

**Calculation of the Oral Cancer Risk Index (OCRI)**

With the finalized set of EdTAR parameters, the exfoliative cytology data was processed and further used in building the SVM model with a radial kernel function using R kenlab ([Karatzoglou 2004](#_ENREF_14)) package. To optimize the hyperparameters, we used two-class samples (normal and OSCC) and the same random sampling procedure to recreate the training dataset and test dataset. The training dataset was processed with median centering and column scaling. For the best outcome, we used leave-on-out cross validation and evaluated the model performance on the nine grid cost parameter between 2(-2) – 64. The final model had a cost of 32 and a hyperparameter sigma of 0.6456. OCRI was calculated as the probability of OSCC for an unknown sample. It ranges between 0 and 1, where 0 indicates the lowest risk of OSCC, and 1 the highest risk of OSCC.

**Results**

Yao, write something about these clinical cases, Classify results … (Figure 1A, 1B).

**Data transformation and reconstruction by EdTAR**

The DI values obtained from Classify represented a mixture of cell populations, diploid, tetraploid and aneuploid, and were displayed in a histogram (Figure 1C). The ratio of the three populations was roughly 0.893:0.092:0.005. Among these three populations, the most informative was the aneuploidy cell population. We simulated these three populations from three normal distributions, diploid cell population (red; µ=1.001, σ=0.19), tetraploid cell population (green; µ=2.002, σ=0.25), and aneuploid cell population (blue; µ=2.300, σ=0.5) (Figure 1D).

After identification of all peaks, a typical normal sample had one peak located at the DI value of 0.995 which indicated a diploid cell population (Figure 3A). A typical OLK sample showed multiple peaks in addition to the major diploid peak (e.g., DI = 0.798) (Figure 3D). A typical OSCC sample showed a peak pattern similar to that of an OLK sample (Figure 3G) often with more peaks beyond D.I. = 2.3. In case there was only one diploid cell population, no more data processing was conducted (Figure 3B). Otherwise, data were further processed for extraction of the diploid and tetraploid cell populations, isolation of the aneuploidy cell population, and signal amplification. For a typical OLK sample and a typical OSCC sample, after the first cell population was extracted, the second peak and other small peaks became much more prominent (Figure 3E, 3H).

The major statistics of the diploid, tetraploid and aneuploid cell populations were then pooled together for data reconstruction. Along the x-axis of DI value, we defined finite number new variables with a range between 0 – 8 and 16 intervals. As shown in boxplots, normal samples (n=102), OLK samples (n=82), and OSCC samples (n=93) showed different patterns (Figure 3C, F, I).

**Selection of SVM as the statistical model**

Six statistical models were tested using data of normal and OSCC samples, which were pathologically distinct. Sensitivity, specificity and the area under the ROC curve were reported (Figure 4). Median sensitivity ranged between 0.83 and 1, with the SVM having the highest median sensitivity (Supplementary Table 1). Specificity was high for all six models indicating low false positive rates. Taking both the sensitivity and specificity into account, the area under the ROC curved provided a general fair assessment on the performance of a model. The median ranged between 0.91 and 1. As compared with the other five models, SVM performed the best, and thus was chosen as the statistical model for calculation of OCRI.

**Quantitative risk stratification of OLK patients**

We further fine-tuned the SVM model with leave-one-out cross validation strategy and finalized on the key hyperparameters (cost C = 32 and hyperparameter sigma = 0.6456). The model was built with data of 70% cases (72 normal and 66 OSCC) and tested on data of the remaining cases. A sensitivity of 0.939, a specificity of 0.9444, and an area under ROC of 0.968 were reached. To calculate OCRI, we applied the model to data of a new sample, and let the model compute the probability that this sample was sampled from an OSCC population given the variables. OCRI was shown on the scale between 0 and 1 (y-axis). Data of 30 normal samples, 27 OSCC samples, and 82 OLK samples tested with OCRI were shown in the same scale (Figure 5).

The majority of normal samples were predicted with an OCRI<0.5 with two exceptions (0.98, 0.59). The majority of OSCC samples were predicted with an OCRI>0.5 with two exceptions (0.30, 0.43). Consistent with the clinical nature of OLK, OCRI of the OLK samples spread across a wide range. Of the 82 OLK samples, 14 had an OCRI above 0.5.

**Clinical follow-up of OLK patients**

According to our calculation of OCRI, we attempted to set the cutoff lines to assess the risk of malignant transformation of OLK patients. A high risk was defined as OCRI ≥ 0.7, a medium risk as OCRI between 0.5 and 0.7, and a low risk as ORCI ≤ 0.5. The mean follow-up time for the OLK patients was 3.8 years.

In one case (Case 128141), the density plot of DI values obtained from initial exfoliative cytology showed multiple peaks in April 2008 (Figure 6A). The first two peaks represented diploid and tetraploid cell populations as the majority. Following data processing with EdTAR, the second peak become prominent after the first population was successfully extracted (Figure 6B). An OCRI was calculated as 0.88. Although biopsy histopathology reported mild dysplasia (Figure 6C), this patients was regularly followed up in our outpatient clinic. A tumor was observed in August 2011, and the histopathology confirmed the diagnosis of OSCC (Figure 6D).

**Discussion**

In this study, we developed a statistical modeling method for quantitative risk stratification of OLK patients. Using a data transformation method (EdTAR) and a machine learning technique (SVM), we generated a quantitative index, OCRI, for assessment of cancer risk. This index is potentially useful for guiding clinical follow-up of OLK patients and improving cost-effectiveness. Although a statistically valid conclusion can not be reached at this moment, further follow-up of our cases of OLK will allow us set a cutoff threshold.

OLK, as a definite premalignant lesion of OSCC, is known to carry a cancer risk higher than normal subjects [11]. However, OLK may develop from multiple mechanisms some of which may not be associated with cancer risk at all. Visual inspection by clinicians with the aid of various tools tends to have a high rate of false positivity. As a well-established and widely used method for early detection of oral cancer, exfoliative cytology provides qualitative result of diagnosis. The major advantages are its being minimally invasive and inexpensive, and thus better acceptance by patients [23,29]. In clinical setting, physicians have to reply on multiple tests during follow-up before the patient is definitely proved to be “positive”. Therefore there is a need of quantitative risk stratification of OLK. In this study, using DI values of exfoliative cytology we successfully developed EdTAR as the method for data transformation and reconstruction. This strategy overcomes the major problem in statistical analysis of exfoliative cytology data, which usually contain a big population of diploid cells, a smaller population of tetraploid cells, and a very small population of aneupoid cells. After EdTAR, the signal of aneuploid cell population is amplified. Reconstruction of data of three cell populations allows SVM for pattern recognition and calculation of OCRI. One of our OLK cases had a high OCRI and was found to develop OSCC 40 months later during follow-up.

Several approaches have been employed for quantitative stratification of cancer risk. Cancer risk index based on clinical parameters, for example Harvard Cancer Risk Index [34], had only a modest discriminatory accuracy for several cancers. It is mainly used for the general population, but not in a tissue or cancer-specific manner for OLK patients [33]. Recently there has been a tremendous enthusiam of using molecular markers for cancer risk stratification, such as mRNA expression data (using gene array, qRT-PCR) [36] and protein expression data (using immunohistochemical staining) [37]. This approach has been well developed for clinical use in breast cancer [35] and colon cancer [36]. However, performance of molecular markers is not much better than established risk factors. In one study [36], the four tested gene expression-based risk scores provide prognostic information but only contributed marginally to improving models based on established risk factors. It is believed that selection of prognostic gene lists and unclear biological meanings of gene signatures contributed to this limitation. Combination with clinical data and inferring biologically relevant pathway deregulation scores have been proposed as potential solutions [38]. In oral cancer, a 29-gene predictive model showed marked improvements in terms of prediction accuracy over the models using previously known clinicopathological risk factors. The prediction error curves showed that Model 1 (only using microarray data) can markedly improve the prediction accuracy over Model 3 (clinical data and protein data). Model 2 (using microarray data, clinical data and protein data) was slightly better than Model 1, both models have similar performance with 8% prediction error rate beyond 2 years of follow-up time. Although this approach is promising, high cost, special expertise in sample analysis and data analysis, and high-quality of sampling are obvious hurdles to overcome before it is routinely used in clinical setting.

As a laboratory assessment of cellular markers, exfoliative cytology remains a practical and reliable method for quantitative risk stratification of OSCC. It has been well established that DNA aneuploidy can predict histologically obvious malignancy 1 to 15 months prior to histology [31,32]. As a non-invasive and inexpensive method, this approach has advantages over other methods. Cellular morphology tend to be relatively stable than molecular markers. However, our method has its limitations….. Two cases of normal with high OCRI…. Two cases of OSCC with low OCRI….. Exfoliative cytology may be repeated for outliers.

There are several useful parameters collected by exfoliative cytology besides DI value. In our study only used DI value as the parameter for model construction. We believe if other parameters are taken into consideration, the performance of this model may be further improved. On the other hand, in combination with other existing methods, according to visual inspection by clinicians, physic-chemical properties and molecular markers, especially, genomics data obtained by the multiplexed, high-throughput NextGen sequencing [39,40], this model may be further improved.

**Acknowledgement:**

This article was partially supported by Beijing Municipal Administration of Hospitals Key Medical Project (ZYLX201407) and a grant from National Natural Science Foundation of China (81372897).

Table 1. General characteristics of normal subjects, OLK patients and OSCC patients

|  |  |  |  |
| --- | --- | --- | --- |
|  | Normal)  (n=102) | OLK  (n=82) | OSCC  (n=93) |
| Age (yr) |  |  |  |
| Mean ± SD | 44.00 ± 16.00 | 58.16 ± 11.48 | 61.70 ± 11.11 |
| Range | 22 - 80 | 25 - 85 | 21 - 83 |
| Gender |  |  |  |
| Male (%) | 46 (45.1) | 37 (45.1) | 45 (48.4) |
| Female (%) | 56 (54.9) | 45 (54.9) | 48 (51.6) |
| Site |  |  |  |
| Tongue (%) | 28 (27.5) | 22 (26.8) | 41 (44.1) |
| Gingival (%) | 15 (14.7) | 33 (40.2) | 27 (29.0) |
| Other (%) | 59 (57,8) | 27 (32.9) | 25 (26.9) |
| Smoking |  |  |  |
| Yes (%) | 32 (31.4) | 29 (35.4) | 31 (33.3) |
| No (%) | 70 (68.6） | 53 (64.6) | 62 (66.7) |
| Drinking |  |  |  |
| Yes (%) | 29 (28.4) | 19 (23.2) | 29 (31.2) |
| No (%) | 73 (71.6) | 63 (76.8) | 64 (68.8) |

**Figure Legends**

**Figure 1.** **Distribution of DNA contents in exfoliative cytology.** (A) Selected cells with abnormally high DI values (>2.3). (B)A scatter plot with y-axis as the area of nucleus and x-axis as DI value. (C) Distribution histogram of DI values of all nuclei. (D) Distribution histogram of DI values of the three cell populations after simulation from normal distribution, diploid cell population (red; µ=1.001, σ=0.19), tetraploid cell population (green; µ=2.002, σ=0.25) and aneuploidy cell population (blue; µ=2.300, σ=0.5). When these three cell populations are merged at the ratio of 0.893:0.092:0.005, a composite distribution histogram (black) can be generated.

**Figure 2. Work flow of expert-guided data transformation and reconstruction (EdTAR).** Starting with DI values as the raw data, EdTAR first identified candidate peaks of cell populations. Diploid cell population was extracted and further filtered if more than one population is detected. The same procedure was applied to extract the tetraploid cell population and thus the aneuploid cell population was isolated. Data of these three cell populations were reconstructed across a wide rage [0 – 8] using the discrete density at each interval. The newly constructed data was used for training the statistical model and calculation of the Oral Cancer Risk Index (OCRI).

**Figure 3. Application of EdTAR in processing data of three samples with pathological diagnosis of normal (A-C), OLK (D-F), and OSCC (G-I).** All density plots have x-axis as DI value and y-axis as density. Panel A, D and G showed density plots before data processing by EdTAR. In Panel A, a major peek with a DI of 0.995 represents the diploid cell population, where another small peaks (DI = 0.594) was a minor population possibly due to image processing. In Panel D, a major peek with a DI of 0.798 represents the diploid cell population (3,590 cells). Other than this peak, four peaks with DI values of 1.25, 1.75, 2.22, and 2.74, were present. In Panel G, a major peek with a DI of 1.02 represents the diploid cell population, and a second peak with a DI of 1.79 represents the tetraploid cell population. Other than these two peaks, three peaks with DI values of 3.25, 3.57, and 3.99 were present, and were believed to represent the aneuploidy cell population. Panel B, E and H corresponding with Panel A, D and G respectively were three plots showing the net results of data processing by EdTAR. Signals of the aneuploidy cell populations were amplified in Panel E and H. Panel C, F and I showed boxplots of newly constructed variables after data processing with EdTAR. The x-axis indicated the new variables along a range of DI [0 – 8] and y-axis the boxplot of available values for each variable.

**Figure 4. Assessment of statistical models.** Seven models (SVM, RRF, PLR, NNET, KNN, and CART) were tested for their performance using three parameters, ROC, sensitivity and specificity. Each model was trained on the training data and tested on the testing data. Each boxplot showed the distribution of these three parameters (R caret package <http://cran.r-project.org/web/packages/caret/index.html>).

**Figure 5. Calculation of Oral Cancer Risk Index (OCRI).** OCRI was calculated for each case with known pathology. The y-axis showed the ORCI between 0 and 1, where 0 indicates the lowest risk of OSCC and 1 indicates the highest risk of OSCC.

**Figure 6. Application of EdTAR in clinical follow-up of one patient (Case 128141).** Exfoliative cytology was performed in April 2008 and a density plot of DI data was generated (A). With EdTAR, positive signals were relatively amplified and an OCRI was calculated as 0.88 (B). Histopathology of biopsy showed mild dysplasia on H&E stained section (C). A tumor was observed in August 2011 with the histopathological diagnosis of squamous cell carcinoma (D).

References

Ahmed, S. M., Mubeen, et al. (2009). "Molecular biology: an early detector of oral cancers." Ann Diagn Pathol **13**(2): 140-145.

Arpino, G., D. Generali, et al. (2013). "Gene expression profiling in breast cancer: a clinical perspective." Breast **22**(2): 109-120.

Awan, K. H., P. R. Morgan, et al. (2011). "Evaluation of an autofluorescence based imaging system (VELscope) in the detection of oral potentially malignant disorders and benign keratoses." Oral Oncol **47**(4): 274-277.

Brouns, E. R., J. A. Baart, et al. (2013). "The relevance of uniform reporting in oral leukoplakia: definition, certainty factor and staging based on experience with 275 patients." Med Oral Patol Oral Cir Bucal **18**(1): e19-26.

Cassidy, A., S. W. Duffy, et al. (2006). "Lung cancer risk prediction: a tool for early detection." Int J Cancer **120**(1): 1-6.

Cervigne, N. K., J. Machado, et al. (2014). "Recurrent genomic alterations in sequential progressive leukoplakia and oral cancer: drivers of oral tumorigenesis?" Hum Mol Genet **23**(10): 2618-2628.

Chaturvedi, P., S. K. Majumder, et al. (2010). "Fluorescence spectroscopy for noninvasive early diagnosis of oral mucosal malignant and potentially malignant lesions." J Cancer Res Ther **6**(4): 497-502.

Di Narzo, A. F., S. Tejpar, et al. (2014). "Test of four colon cancer risk-scores in formalin fixed paraffin embedded microarray gene expression data." J Natl Cancer Inst **106**(10).

Domany, E. (2014). "Using high-throughput transcriptomic data for prognosis: a critical overview and perspectives." Cancer Res **74**(17): 4612-4621.

Gupta, P. C., F. S. Mehta, et al. (1980). "Incidence rates of oral cancer and natural history of oral precancerous lesions in a 10-year follow-up study of Indian villagers." Community Dent Oral Epidemiol **8**(6): 283-333.

Jadhav, K., N. Gupta, et al. (2011). "Micronuclei: An essential biomarker in oral exfoliated cells for grading of oral squamous cell carcinoma." J Cytol **28**(1): 7-12.

Jeck, W. R., J. Parker, et al. (2014). "Targeted next generation sequencing identifies clinically actionable mutations in patients with melanoma." Pigment Cell Melanoma Res **27**(4): 653-663.

Kammerer, P. W., F. P. Koch, et al. (2013). "Prospective, blinded comparison of cytology and DNA-image cytometry of brush biopsies for early detection of oral malignancy." Oral Oncol **49**(5): 420-426.

Karatzoglou, A. S., Alex; Hornik, Kurt; and Zeileis, Achim (2004). "kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software." Journal of Statistical Software **11**(9): 1-20.

Khiabanian, H., P. Van Vlierberghe, et al. (2010). "ParMap, an algorithm for the identification of small genomic insertions and deletions in nextgen sequencing data." BMC Res Notes **3**: 147.

Kim, D. J., B. Rockhill, et al. (2004). "Validation of the Harvard Cancer Risk Index: a prediction tool for individual cancer risk." J Clin Epidemiol **57**(4): 332-340.

Kuhn, M. J., Kjell (2013). Applied Predictive Modeling, Springer New York Heidelberg Dordrecht.

Lee, J. J., H. C. Hung, et al. (2006). "Carcinoma and dysplasia in oral leukoplakias in Taiwan: prevalence and risk factors." Oral Surg Oral Med Oral Pathol Oral Radiol Endod **101**(4): 472-480.

Lingen, M. W., J. R. Kalmar, et al. (2007). "Critical evaluation of diagnostic aids for the detection of oral cancer." Oral Oncol **44**(1): 10-22.

Ma, J. M., T. J. Zhou, et al. (2014). "Brush biopsy with DNA-image cytometry: a useful and noninvasive method for monitoring malignant transformation of potentially malignant oral disorders." Eur Arch Otorhinolaryngol **271**(12): 3291-3295.

Maraki, D., J. Becker, et al. (2004). "Cytologic and DNA-cytometric very early diagnosis of oral cancer." J Oral Pathol Med **33**(7): 398-404.

Maraki, D., U. R. Hengge, et al. (2006). "Very early cytological and DNA-cytometric diagnosis of in situ carcinoma in an immunosuppressed liver transplant recipient." J Oral Pathol Med **35**(1): 58-60.

Maraki, D., S. Yalcinkaya, et al. (2006). "Cytologic and DNA-cytometric examination of oral lesions in lichen planus." J Oral Pathol Med **35**(4): 227-232.

Mehrotra, R., M. Hullmann, et al. (2009). "Oral cytology revisited." J Oral Pathol Med **38**(2): 161-166.

Messadi, D. V. (2013). "Diagnostic aids for detection of oral precancerous conditions." Int J Oral Sci **5**(2): 59-65.

Nagao, T., N. Ikeda, et al. (2005). "Incidence rates for oral leukoplakia and lichen planus in a Japanese population." J Oral Pathol Med **34**(9): 532-539.

Pektas, Z. O., A. Keskin, et al. (2006). "Evaluation of nuclear morphometry and DNA ploidy status for detection of malignant and premalignant oral lesions: quantitative cytologic assessment and review of methods for cytomorphometric measurements." J Oral Maxillofac Surg **64**(4): 628-635.

Pentenero, M., W. Giaretti, et al. (2009). "DNA aneuploidy and dysplasia in oral potentially malignant disorders: association with cigarette smoking and site." Oral Oncol **45**(10): 887-890.

R\_Core\_Team (2014). " R: A language and environment for statistical computing."

Ramshankar, V., V. T. Soundara, et al. (2014). "Risk Stratification of Early Stage Oral Tongue Cancers Based on HPV Status and p16 Immunoexpression." Asian Pac J Cancer Prev **15**(19): 8351-8359.

Remmerbach, T. W., H. Weidenbach, et al. (2003). "Earliest detection of oral cancer using non-invasive brush biopsy including DNA-image-cytometry: report on four cases." Anal Cell Pathol **25**(4): 159-166.

Remmerbach, T. W., H. Weidenbach, et al. (2003). "Diagnostic value of nucleolar organizer regions (AgNORs) in brush biopsies of suspicious lesions of the oral cavity." Anal Cell Pathol **25**(3): 139-146.

Remmerbach, T. W., H. Weidenbach, et al. (2001). "Cytologic and DNA-cytometric early diagnosis of oral cancer." Anal Cell Pathol **22**(4): 211-221.

Rhodus, N. L., A. R. Kerr, et al. (2014). "Oral cancer: leukoplakia, premalignancy, and squamous cell carcinoma." Dent Clin North Am **58**(2): 315-340.

Scheifele, C., A. M. Schmidt-Westhausen, et al. (2004). "The sensitivity and specificity of the OralCDx technique: evaluation of 103 cases." Oral Oncol **40**(8): 824-828.

Sciubba, J. J. (1999). "Improving detection of precancerous and cancerous oral lesions. Computer-assisted analysis of the oral brush biopsy. U.S. Collaborative OralCDx Study Group." J Am Dent Assoc **130**(10): 1445-1457.

Siegel, R., J. Ma, et al. (2014). "Cancer statistics, 2014." CA Cancer J Clin **64**(1): 9-29.

Silverman, S., Jr. (2001). "Demographics and occurrence of oral and pharyngeal cancers. The outcomes, the trends, the challenge." J Am Dent Assoc **132 Suppl**: 7S-11S.

Stelow, E. B. and S. E. Mills (2005). "Squamous cell carcinoma variants of the upper aerodigestive tract." Am J Clin Pathol **124 Suppl**: S96-109.

van der Waal, I. (2014). "Oral potentially malignant disorders: is malignant transformation predictable and preventable?" Med Oral Patol Oral Cir Bucal **19**(4): e386-390.

Vijayavel, T. and N. Aswath (2013). "Correlation between histological grading and ploidy status in potentially malignant disorders of the oral mucosa: A flow cytometric analysis." J Oral Maxillofac Pathol **17**(2): 169-175.

Wand, M. P. J., M.C. (2000). Kernel Smoothing, Taylor & Francis.

Warnakulasuriya, S. (2009). "Global epidemiology of oral and oropharyngeal cancer." Oral Oncol **45**(4-5): 309-316.

Warnakulasuriya, S., N. W. Johnson, et al. (2007). "Nomenclature and classification of potentially malignant disorders of the oral mucosa." J Oral Pathol Med **36**(10): 575-580.

1. Siegel R, Ma J, Zou Z, Jemal A (2014) Cancer statistics, 2014. CA Cancer J Clin 64: 9-29.

2. Warnakulasuriya S (2009) Global epidemiology of oral and oropharyngeal cancer. Oral Oncol 45: 309-316.

3. Stelow EB, Mills SE (2005) Squamous cell carcinoma variants of the upper aerodigestive tract. Am J Clin Pathol 124 Suppl: S96-109.

4. Silverman S, Jr. (2001) Demographics and occurrence of oral and pharyngeal cancers. The outcomes, the trends, the challenge. J Am Dent Assoc 132 Suppl: 7S-11S.

5. Maraki D, Becker J, Boecking A (2004) Cytologic and DNA-cytometric very early diagnosis of oral cancer. J Oral Pathol Med 33: 398-404.

6. Brouns ER, Baart JA, Bloemena E, Karagozoglu H, van der Waal I (2013) The relevance of uniform reporting in oral leukoplakia: definition, certainty factor and staging based on experience with 275 patients. Med Oral Patol Oral Cir Bucal 18: e19-26.

7. Warnakulasuriya S, Johnson NW, van der Waal I (2007) Nomenclature and classification of potentially malignant disorders of the oral mucosa. J Oral Pathol Med 36: 575-580.

8. Gupta PC, Mehta FS, Daftary DK, Pindborg JJ, Bhonsle RB, et al. (1980) Incidence rates of oral cancer and natural history of oral precancerous lesions in a 10-year follow-up study of Indian villagers. Community Dent Oral Epidemiol 8: 283-333.

9. Nagao T, Ikeda N, Fukano H, Hashimoto S, Shimozato K, et al. (2005) Incidence rates for oral leukoplakia and lichen planus in a Japanese population. J Oral Pathol Med 34: 532-539.

10. Vijayavel T, Aswath N (2013) Correlation between histological grading and ploidy status in potentially malignant disorders of the oral mucosa: A flow cytometric analysis. J Oral Maxillofac Pathol 17: 169-175.

11. Lee JJ, Hung HC, Cheng SJ, Chen YJ, Chiang CP, et al. (2006) Carcinoma and dysplasia in oral leukoplakias in Taiwan: prevalence and risk factors. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 101: 472-480.

12. van der Waal I (2014) Oral potentially malignant disorders: is malignant transformation predictable and preventable? Med Oral Patol Oral Cir Bucal 19: e386-390.

13. Cervigne NK, Machado J, Goswami RS, Sadikovic B, Bradley G, et al. (2014) Recurrent genomic alterations in sequential progressive leukoplakia and oral cancer: drivers of oral tumorigenesis? Hum Mol Genet 23: 2618-2628.

14. Rhodus NL, Kerr AR, Patel K (2014) Oral cancer: leukoplakia, premalignancy, and squamous cell carcinoma. Dent Clin North Am 58: 315-340.

15. Messadi DV (2013) Diagnostic aids for detection of oral precancerous conditions. Int J Oral Sci 5: 59-65.

16. Chaturvedi P, Majumder SK, Krishna H, Muttagi S, Gupta PK (2010) Fluorescence spectroscopy for noninvasive early diagnosis of oral mucosal malignant and potentially malignant lesions. J Cancer Res Ther 6: 497-502.

17. Awan KH, Morgan PR, Warnakulasuriya S (2011) Evaluation of an autofluorescence based imaging system (VELscope) in the detection of oral potentially malignant disorders and benign keratoses. Oral Oncol 47: 274-277.

18. Ahmed SM, Mubeen, Jigna VR (2009) Molecular biology: an early detector of oral cancers. Ann Diagn Pathol 13: 140-145.

19. Pentenero M, Giaretti W, Navone R, Demurtas A, Rostan I, et al. (2009) DNA aneuploidy and dysplasia in oral potentially malignant disorders: association with cigarette smoking and site. Oral Oncol 45: 887-890.

20. Jadhav K, Gupta N, Ahmed MB (2011) Micronuclei: An essential biomarker in oral exfoliated cells for grading of oral squamous cell carcinoma. J Cytol 28: 7-12.

21. Remmerbach TW, Weidenbach H, Muller C, Hemprich A, Pomjanski N, et al. (2003) Diagnostic value of nucleolar organizer regions (AgNORs) in brush biopsies of suspicious lesions of the oral cavity. Anal Cell Pathol 25: 139-146.

22. Mehrotra R, Hullmann M, Smeets R, Reichert TE, Driemel O (2009) Oral cytology revisited. J Oral Pathol Med 38: 161-166.

23. Scheifele C, Schmidt-Westhausen AM, Dietrich T, Reichart PA (2004) The sensitivity and specificity of the OralCDx technique: evaluation of 103 cases. Oral Oncol 40: 824-828.

24. Ma JM, Zhou TJ, Wang R, Shan J, Wu YN, et al. (2014) Brush biopsy with DNA-image cytometry: a useful and noninvasive method for monitoring malignant transformation of potentially malignant oral disorders. Eur Arch Otorhinolaryngol 271: 3291-3295.

25. Kammerer PW, Koch FP, Santoro M, Babaryka G, Biesterfeld S, et al. (2013) Prospective, blinded comparison of cytology and DNA-image cytometry of brush biopsies for early detection of oral malignancy. Oral Oncol 49: 420-426.

26. Maraki D, Yalcinkaya S, Pomjanski N, Megahed M, Boecking A, et al. (2006) Cytologic and DNA-cytometric examination of oral lesions in lichen planus. J Oral Pathol Med 35: 227-232.

27. Pektas ZO, Keskin A, Gunhan O, Karslioglu Y (2006) Evaluation of nuclear morphometry and DNA ploidy status for detection of malignant and premalignant oral lesions: quantitative cytologic assessment and review of methods for cytomorphometric measurements. J Oral Maxillofac Surg 64: 628-635.

28. Lingen MW, Kalmar JR, Karrison T, Speight PM (2007) Critical evaluation of diagnostic aids for the detection of oral cancer. Oral Oncol 44: 10-22.

29. Sciubba JJ (1999) Improving detection of precancerous and cancerous oral lesions. Computer-assisted analysis of the oral brush biopsy. U.S. Collaborative OralCDx Study Group. J Am Dent Assoc 130: 1445-1457.

30. Remmerbach TW, Weidenbach H, Pomjanski N, Knops K, Mathes S, et al. (2001) Cytologic and DNA-cytometric early diagnosis of oral cancer. Anal Cell Pathol 22: 211-221.

31. Remmerbach TW, Weidenbach H, Hemprich A, Bocking A (2003) Earliest detection of oral cancer using non-invasive brush biopsy including DNA-image-cytometry: report on four cases. Anal Cell Pathol 25: 159-166.

32. Maraki D, Hengge UR, Becker J, Boecking A (2006) Very early cytological and DNA-cytometric diagnosis of in situ carcinoma in an immunosuppressed liver transplant recipient. J Oral Pathol Med 35: 58-60.

33. Cassidy A, Duffy SW, Myles JP, Liloglou T, Field JK (2006) Lung cancer risk prediction: a tool for early detection. Int J Cancer 120: 1-6.

34. Kim DJ, Rockhill B, Colditz GA (2004) Validation of the Harvard Cancer Risk Index: a prediction tool for individual cancer risk. J Clin Epidemiol 57: 332-340.

35. Arpino G, Generali D, Sapino A, Del Matro L, Frassoldati A, et al. (2013) Gene expression profiling in breast cancer: a clinical perspective. Breast 22: 109-120.

36. Di Narzo AF, Tejpar S, Rossi S, Yan P, Popovici V, et al. (2014) Test of four colon cancer risk-scores in formalin fixed paraffin embedded microarray gene expression data. J Natl Cancer Inst 106.

37. Ramshankar V, Soundara VT, Shyamsundar V, Ramani P, Krishnamurthy A (2014) Risk Stratification of Early Stage Oral Tongue Cancers Based on HPV Status and p16 Immunoexpression. Asian Pac J Cancer Prev 15: 8351-8359.

38. Domany E (2014) Using high-throughput transcriptomic data for prognosis: a critical overview and perspectives. Cancer Res 74: 4612-4621.

39. Jeck WR, Parker J, Carson CC, Shields JM, Sambade MJ, et al. (2014) Targeted next generation sequencing identifies clinically actionable mutations in patients with melanoma. Pigment Cell Melanoma Res 27: 653-663.

40. Khiabanian H, Van Vlierberghe P, Palomero T, Ferrando AA, Rabadan R (2010) ParMap, an algorithm for the identification of small genomic insertions and deletions in nextgen sequencing data. BMC Res Notes 3: 147.