**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 called expert-guided data transformation and reconstruction (EdTAR) allows automatic data processing and reconstruction by integrating expert-guided parameters. Modern machine learning technique was utilized to build statistical prediction models with the reconstructed data. Among the several models tested using resampling methods to prune the core parameters, Support Vector Machine (SVM) was found to be the optimal model with a high sensitivity (median>0.98) and specificity (median>0.99). Using the SVM model, we generated an oral cancer risk index (OCRI) for 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 major public health problem worldwide, as well as a major cause of cancer morbidity and mortality, especially in Asia [1,2]. 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, especially oral leukoplakia (OLK), and histopathologically follows a step-wise pattern of hyperplasia, dysplasia and SCC [3,4]. Meanwhile, the five-year survival rate remained unchanged despite the advances in radiotherapy and chemotherapy [1]. The five-year survival rate for patients with early, localized lesions is nearly 80%, whereas it is only 19% for patients with distant metastasis [5]. So it is important to assess OLK and early diagnose OSCC. 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]. Its histopathology is hyperkeratosis. From hyperkeratosis to OSCC, that needs several months or couple of years. Overall chance of malignant transformation in the world is 3.6% [10], which can up to 12.9% in Taiwan [11]. It may develop from a variety of reasons, such as smoking habits, elder age, female gender, oral subsite (such as borders of the tongue and floor of the mouth) and clinical subtype (homogeneous versus heterogeneous) [11,12,13]. Millions of OLK patients cannot follow up them all due to the high incidence. So, there is a clinical need of risk stratification in OLK patients to improve the cost-effectiveness.

There are several tools available for early diagnosis OSCC and assessment OLK lesions. (1) Clinical examination. Lesions with a red component, ulceration, or with variable topography described as granular, nodular, or verrucous, are more likely to represent malignancy [7,14]. However, in the early stage, OSCC may appear as small, minor mucosal changes, to which the unsuspecting clinician may not aggressively respond [14]. Meanwhile, the clinical examination is subjective and needs more clinical experience. (2) The examination according to physic-chemical properties of oral mucosa. Those contains toluidine blue [15], fluorescence spectroscopy [16] and so on. They are easy and quick to use. However, due to the high false positive value, the toluidine blue test and fluorescence test (such as VELScope system) appear to be highly sensitive but less specific [14,15]. While the other study showed that the VELScope system cannot distinguish the high-risk from low-risk OLK lesions [17]. (3) Diagnostic tools according to molecular markers. Those include chromosome in situ hybridization, immunohistochemistry, polymerase chain reaction, DNA microarrays and proteomics [18]. Although the molecular tools can improve the accuracy of the OSCC diagnosis, those sophisticated tools is difficult of using in clinical exam due to expensive and critical for sample quality.

The exfoliative cytology has already been used in diagnosing oral cancer and premalignant diseases in recent years. Exfoliative cytology is always assisted with DNA quantitative analysis [19], mivronucleus analysis [20] and nucleolar organizer regions [21]. Currently, exfoliative cytology and DNA quantitative analysis is increasingly used for early detection of oral cancer [22]. It has high sensitivity and specificity, which can up to 100% [5,19,23,24,25]. However, as for risk stratification for OLK patients, DNA quantitative analysis cannot assess it well. Meanwhile, the diagnosis criterion of DNA quantitative analysis only used fewer data of the DNA index (DI), that lost lots of information. Above all, there is a need quantitatively assess the cancer risk of OLK lesions. But the challenge is data analysis. There are 3 cell populations (diploid cells, tetraploid cells and aneupoid cells) with a large diploid cell population. Unfortunately, the very small population of aneupoid cells are important, which need to amplified this signal. Therefore we developed a statistical method to meet this need.

**Materials and Methods**

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

Exfoliative cells including OLK (n=82), OSCC (n=93), and healthy oral mucosa (n=102) were obtained from Beijing Stomatological Hospital, Capital Medical University, using cervibrush (Motic, China). The general information as shown in Table 1. Smokers were defined as smoked 1 cigarette or more per day for at least year. And the patients who had 1 or more drinks 3 times or more per week were categorized as drinkers [11]. The study was approved by the ethical committee and all patients signed the informed consent.

Every patients need to follow up to record the changes in signs and symptoms, and find whether there is a malignant transformation.

Yao, YOU NEED TO 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**

The histopathology biopsy from OLK and OSCC patients was taken after brush biopsy immediately. Histopathological diagnosis criterion was scored according to the standard criteria of the World Health Organization (WHO) classification system (2006) of Head and Neck Tumors. DNA quantitative analysis and Papanicolaou exam were used in diagnosing health oral mucosa [26].

YAO, DESCRIBE THE CRITERIA OF HISTOPATHOLOGY

**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_3)).

**Peak identification**

We first aimed to differentiate three possible cell populations, diploid, tetraploid and aneuploidy. 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 (ref), 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.

Jianying, add an Reference above

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

**Jianying, write something here.**

**Jianying, can you separate “extraction of diploid/tetraploid peaks and isolation of aneuploid peak” and “signal amplification” into two parts each with its own pseudo code? The purpose is to be consistent with the revised Figure 2.**

**Signal amplification**

For signal amplification, we set ratios of these three populations as R1, R2, and R3 respectively, with the constraint that R1 + R2 + R3 = 1 (Equation 1). We randomly sampled R1 from a uniform distribution ~Unif [0.95, 0.998], R2 ~Unif [(1- R1-0.05), (1-R1-0.005)], and R3 [1-R1-R2]. If only a single diploid population was detected, 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). The procedure is shown in pseudo code format as follows:

1. With the parameter set θdt for data transformation
2. Estimate the density distribution from the empirical DI values and identify candidate peaks
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

**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_3)) 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_2)). These models were ranked according to the area under receiver operating characteristic (ROC), sensitivity and specificity. Based its performance, the SVM model was chosen for the following calculation.

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

We finalized one set of EdTAR parameters and processed exfoliative cytology data using SVM model with a radial kernel function implemented in an R kenlab ([Karatzoglou 2004](#_ENREF_1)) 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 with 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 (DI = 0.798) (Figure 3D). A typical OSCC sample showed a peak pattern similar to that of an OLK sample (Figure 3G). 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 with 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 to compute the probability 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 were used for calculation of OCRI (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 arbitrarily 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 (YAO, do not remove the structure before we finish)**

1. Summary

The OLK, as the premalignant lesion, which has 12.9% malignant transformation rates in Taiwan [11]. In contrast to high sensitivity and specificity rates of visual screening of skin cancer, it has not been successfully used in detection for early-stage of oral cancer. One explanation for this is that advanced stage’s classic clinical symptoms are often subtle in early stage, included ulceration, elevation, bleeding [27,28]. As a well-established and widely used method for early detection of oral cancer, exfoliative cytology provides quanlitative result of diagnosis. It was a minimally invasive and inexpensive technique. For example, OralCDx reports “negative or benign”, “positive” (defined as definitive cellular evidence of epithelial dysplasia or carcinoma), or “atypical” (defined as abnormal epithelial changes of uncertain diagnostic significance) [23]. Due to the high sensitivity and specificity, OralCDx has been proved to be a very good method for early detection of oral cancer [23,29]. However, as for risk stratification for OLK patients, OralCDx does not do a good job because qualitative results “negative, positive or atypical” are vague. Clinicians have to reply on multiple tests during follow-up before the patients is definitely proved to be “negative” or “positive”. That is why a quantitative risk stratification of OLK is important.

1. Our method and its outcome

There is a need quantitatively assess the cancer risk of OLK lesions. In this study, we build EdTAR statistical model to assess OLK patients. We extracted DI value of all cells. However, those are big population of diploid cells, relatively small population of tetraploid cells, and a very small population of aneupoid cells. Therefore, amplification of aneuploid cell population was a big challenge of analysis of exfoliative cytology for quantitative risk stratification. In our study, if more than one population is detected, the diploid cell population was extracted and further filtered. And the same procedure was applied to extract the tetraploid cell population and isolated the aneuploid cell population. And then, we used seven models to tested our statistical model, using three parameters (ROC, sensitivity and specificity). Most statistical models showed high sensitivity, specificity and ROC. Some studies have shown that the sensitivity and specificity of cytological diagnosis combined with DNA-image cytometry reached up to 100% [5] [24,30].

Based on the model selection and assessment, the OCRI was calculated. We set up the cutoff line and defined high risk, medium risk and low risk. In our study, there was one case, whose OCRI was belonged to high risk, and the OLK lesion was malignant transformation after 40 months. Meanwhile, other studies have proved that DNA-aneuploidy can predict histological obvious malignancy 1 to 15 months prior to histology [31,32].

1. Our method is complementary to the existing tools
2. According to visual inspection by clinicians
3. According to physic-chemical properties
4. According to molecular markers

There are other methodologies for quantitative risk stratification of OLK. Such as clinical data-based cancer risk index [33,34] and molecular data-based cancer risk index [35,36]. (1) Clinical data-based cancer risk index. For example Harvard Cancer Risk Index [34]. It has been proved that discriminatory accuracy was modest for ovarian cancer (age-adjusted concordance statistic of 0.59), and relatively good for pancreatic cancer (concordance statistic of 0.72), and colon cancer in men and women (concordance statistics of 0.71, 0.67 respectively). However, this kind of risk prediction was only as accurate as the data on which they are based[33]. And it was not tissue or cancer specific. (2) Molecular data-based cancer risk index, such as mRNA expression data (using gene expression array, qRT-PCR) [36] and protein expression data (using immunohistochemical staining) [37]. This method has been developed for clinical use in other cancers, for example, breast cancer [35], colon cancer [36]. Performance of this approach is not great. According to 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. The reason was that different prognostic gene lists have very few shared genes, the biological meaning of most signatures in unclear, and the published success rates are considered to be overoptimistic [38]. Meanwhile, there are two approaches to dissolve the limitation: (1) combining molecular data with clinical data; (2) infers biologically relevant pathway deregulation scores [38].

Molecular data-based cancer risk index method has been used to generate an oral cancer index using biopsy samples from OLK patients. This study developed 29-transcript predictive model showed marked improvements in terms of prediction accuracy over the models using previously known clinicopathologic risk factor. In details, there were three models, Model 1 (only using microarray data), Model 2 (using microarray data, clinical data and protein data), Model 3 (clinical data and protein data), were tested. The prediction error curves in Model 1 showed that microarray data can markedly improve the prediction accuracy over Model 3 that used only clinical and protein data. And Model 2 is slightly better than Model 1, both models have similar performance with 8% prediction error rate beyond 2 years of follow-up time. Overall, limitations of this methodology are two-folds: (1) it is expensive and special expertise is required for sample analysis and data analysis; (2) since sample quality is critical for this method, clinical sampling, sample storage and processing will be a challenge to clinicians.

1. Limitations of our method
2. Future perspectives:
3. Parameters collected by exfoliative cytology other than DI
4. In combination with other existing methods, according to visual inspection by clinicians, according to physic-chemical properties, according to molecular markers, in particular, genomics data obtained by NextGen sequencing etc.

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. Be 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 improved better.

In conclusion, exfoliative cytology in combination with EdTAR and OCRI is a potentially good method for quantitative risk stratification of OLK patients. It may be used for the follow-up large clinically suspicious OLK lesions to find early stage oral cancer.

**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

Jianying’s references

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.

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

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

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.