Running title: Quantitative risk stratification of oral leukoplakia with oral exfoliative cytology

Yao Liu1#, Jianying Li2#, Zheng Sun1\* and Xiaoxin Chen3\*

2 Frontier Bioinformatics Solution, Cary, NC 27519, USA

#These authors contributed equally

\*Corresponding Authors

**Abstract**

DNA ploidy status directly reflects the cellular neoplasm activity and the abnormal cell division can be detected when the aneusomy or aneuploidy is observed. Several methods have been developed to directly measure the DNA content and further convert to the ratio of G0/G1. This enables us to determine the ploidy status, which can serve as a reliable marker of cell proliferation, even before the clear histopathological sign is observed. Exfoliative cytology, which offers a simple and non-invasive procedure, has emerged as a prominent technology in early oral cancer diagnosis, though many technical hurdles largely limited this method from becoming an automated and robust clinical standard protocol. In this research, we explored a novel data analysis procedure which allows automatically processing the data and re-constructing informative new variables by integrating the expert guided parameters. We then leveraged the modern machine learning technique to efficiently utilize the newly constructed data to build statistical prediction models. Using resampling methods for pruning the model core parameters, we tested a series of models and a successful Support Vector Machine (SVM) model was finally determined. Our method showed high sensitivity (median > 0.98) and specificity (median > 0.99) obtained both during the training process and in predicting on a hold-off test data. Finally, we proposed a risk index metrics for the oral leukoplakis (OLK) diagnosis, clinically defined lesions likely leading to oral squamous cell carcinoma (OSCC). Such an index reflects the probability leading to OSCC predicted from our statistical model, and it will provide a valuable guide for the clinical professionals to develop a meaningful patient’s follow up plan.

**Introduction**

Oral and pharyngeal cancer, grouped together, rank sixth among most common cancer types in the world (ref). Oral cancer is one of the most common malignancies and contributes the majority of cancer morbidity and mortality worldwide (ref). Even with the sufficient surgery, radiotherapy and chemotherapy; unfortunately, the 5-year survival rate of patients with distant metastases at the time of the first diagnosis is only 19%. On the contrary, early diagnosis for operable tumors with localized stage renders a much higher survival rate (approximates 80%) (ref). Oral squamous cell carcinoma (OSCC), a common type of oral cancer, can develop from precancerous lesions, such as oral leukoplakia (OLK) (ref). Therefore, successfully predication of potential OSCC from the commonly observable OLK carries potential value in clinical practice; and it attracts great attention worldwide (ref). Deeper understanding the OLK clinical implication will provide meaningful guide for patient follow up plan, which almost ensures an early diagnosis of possible cancer before the distant metastasis occurs.

Histopathological diagnosis acts as the current golden criteria, but it is a much invasive method and could cause unnecessary trauma for the patients especially when they were diagnosed negative in the end. DNA ploidy status directly reflects the cellular neoplasm activity and the abnormal cell division can be detected when the aneusomy or aneuploidy is observed (ref). Several methods have been developed to directly measure the DNA content and further convert to the ratio of G0/G1. This primary measurement (commonly called DNA index, or D.I. value) can be converted to the equivalent assessment of the ploidy status, which can serve as a reliable marker of cell proliferation, even before the clear histopathological sign is observed. Exfoliative cytology is the currently accepted method for measuring the cellular DNA content (ref) worldwide; it proves a successful method in screening for epithelial dysplasia in situ or invasive carcinomas of the uteri cervix (ref). Due to the progress of Thin Cytologic Test (TCT) and Automatic Imaging Cytometer (AICM), the exfoliative cytology has also been used in diagnosing the oral cancer and premalignant diseases. Since the location of the carcinoma of oral mucosa is superficial, brushing the exfoliate cell can be done during common dental checkup, Therefore, not only does it offer a safe and convenient practice, it also reduces the traumatic injury to the patients.

Although exfoliative cytology offers the greatest potential to be an effective method for early prediction of malignant or pre-malignant, many technical hurdles largely limited this method from becoming an automated and robust clinical standard protocol. Firstly, exfoliative cytology offers a quantitative DNA content measurement, it needs much human intervention to review series of files and look for aneuploidy peaks as well as count the number of cells with excessive increased DNA ploidy. It could be time consuming and sometimes the assessment results could be quite subjective. Secondly, exfoliate cells often consist of mixture of different populations and the results based on thousands of cells presents even harder situations when explaining and interpreting the findings. Lastly, owing to the unbalanced cell populations, statistical models, which have been proved successful in handling mixture of populations (ref), could fail in handling the exfoliative cytology data. The major difficulty has been that the useful signals often buried under the unbalanced amount of the non-informative data. In other words, it is hard to differentiate the signal from the noises.

Upon fully understanding the challenges inherited in this approach, our effort was focused on the data processing and cleaning. In the report, we proposed an Expert-guided data cleaning and reconstruction (**ExGCRn)**, in which we implemented a sequential process to strip out different cell populations while retaining the summary statistics and other useful parameters. Next, we reconstructed a new data set based on *a priori* defined parameters from a mixture of several “populations”. In the end, we defined a set of variables along the axis for initial DNA index values and empirically estimated the density under finite number of the intervals. With the newly constructed the dataset, we then leveraged the modern machine learning technique to build and evaluate a series of statistical prediction models. For each predication model, using resampling methods for pruning the model core parameters, we evaluated the model performance and finalized on best hyper-parameters. Among all the tested models and a successful Support Vector Machine (SVM) model was finally determined. Overall, our method showed high sensitivity (median > 0.98) and specificity (median > 0.99) obtained both during the training process and in predicting on a hold-off test data.

The aim of this study was to establish an analytical protocol and promote a clinical standard for clinical diagnosis of OLK patient, and improve the sensitivity and the specificity of DNA quantitative analysis for potential OSCC early diagnosis. To predict the progressing direction of a clinical defined the oral leukoplakia lesion, we tested our SVM model and found out that the prediction results scattered across the entire panel of “probability”, from almost normal to almost OSCC. It is not surprise that the result was aligned well with the common clinical diagnosis, but it also provides additional risk factor for those OLK patients. Finally, we proposed a risk index metrics for the oral leukoplakia (OLK) diagnosis. Such an index reflects the probability leading to OSCC predicted from our statistical model, and it will provide a valuable guide for the clinical professionals to develop a meaningful patient’s follow up plan.

**Method and materials**

**Patients and clinical specimens collection (yao)**

**Exfoliated cells preparation, DNA staining and imaging analysis (classifier) (yao)**

**Expert-guided parameters** The exfoliative cytology experiment reports the DNA contents of a given cell using the quotient between G0/G1 stages in the normal cell cycle. In this research, we used DNA index (D.I. values) as the raw measurement. The density of the collected D.I. values will peak at different locations along the x-axis, which further represents the central tendency for hidden cell population associated with corresponding DNA ploidy status (table 1). We started off the implementation of the automated data transformation and construction, by testing and finalizing on the several sets of parameters. Most of the parameters match directly or indirectly to the clinical diagnosis standards and/or conventions commonly referred in the laboratory practice (ref). Our goal was to detect and differentiate a mixture of three possible cell populations corresponding to diploid, tetraploid and hypertetraploid/aneuploidy in an exfoliative experiment result. We defined the range thresholds for peaks that represented each cell population: diploid [0.8, 1.2]; tetraploid [1.5, 1.7]; hypertetraploid/aneuploidy [>2.3]. The second set of parameter was for data reconstruction. To balance the model flexibility and robustness, we set ratios among these three populations as R1, R2, and R3 respectively, with the constraint that R1 + R2 + R3 = 1 (equation 1). In practice, we randomly sample R1 from a uniform distribution ~Unif [0.95 , 0.998], R2 ~Unif [(1- R1,-0.05), (1- R1,-0.005),], and R3 will be computed accordingly. If only a single diploid population was detect, the theoretical tetraploid population was sampled from a normal distribution ~ Norm (2.0, 0.3) and hypothetical hypertetraploid/aneuploidy sampled from ~ Norm (2.3, 0.3). The last set of parameters was defined at the reconstruction step. The new variables represent the discrete interval ranging between 0 and 8 (D.I. 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.

**Data density and peak identification**

The D.I. values obtained from the imaging processing software normally represent a mixture of cell populations and often displayed as a histogram (figure 1c); the analytical procedure relied on the estimation of the density to further estimate the parameters for each individual cell population. To do so, we adopted the kernel density estimation procedure often started from the kernel density estimator, (equation 2), where represents a “kernel function” K and a “bandwidth” *h*. Owing to the mathematically requirement, we had assumed that D.I. values were independently selected from a background distribution (although unknown). In smoothing the histogram, we chose Gaussian distribution as the kernel (ref), and finalizing 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. Generally, to identify the candidate peak(s) from the aforementioned density estimation depended on identifying a reflection point, which was achieved from the first derivative sign changes.

**Expert-guided data transformation and reconstruction (EdTAR) pseudo code**

All the parameter estimation, data transformation and reconstruction was done with R ([R\_Core\_Team 2014](#_ENREF_2)). The major procedures displayed in pseudo code format as following:

Data transformation

1. With the parameter set θdt for data transformation
2. Estimate the density distribution from the empirical D.I. values and identify candidate peaks
3. **for** each i = 1..n peaks **do**

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

3.2. Estimate the right side of the distribution

3.3. Filter out the candidate ith family and retain the mean, standard deviation 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 hypertetraploid/aneuploidy, and store the summary statistics of each population
   1. The total number of D.I. values
   2. Sample mean and sample standard deviation
   3. The number of peaks and location of the peaks

Data reconstruction

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 hypertetraploid/aneuploidy

Compute the ratio between diploid and tetraploid population, and sample the three population ratios

Integrate mixture of three theoretical families

Create densities for all 16 discrete intervals

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

**If** the maximum D.I. value ofhypertetraploid/aneuploidy > 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

**Prediction model performance evaluation metrics**

To train a statistic prediction model on a particular classifier, the confusion matrix can be summarized in various standard performance metrics

The Confusion Matrix

|  |  |  |
| --- | --- | --- |
| Total Samples (*TS*) | Actual Positives (*AP*) | Actual Negatives (*AN*) |
| Predicted Positives (*PP*) | True Positives (*TP*) | False Positives (*FP*) |
| Predicted Negatives (*PN*) | False Negatives (*FN*) | True Negatives (*TN*) |

Sensitivity is a measure of actual positives correctly identified as such and the specificity measures the proportion of negatives correctly identified:

*Sensitivity = TP / AP = TP / (TP + FN)*

*Specificity = TN / AN = TN / (TN + FP)*

There is a trade-off between sensitivity and specificity, making models difficult to compare on the basis of these performance metrics alone. In contrast, the receiver operating characteristic (ROC), or ROC curve not only provides a visual evaluation of the model performance by illustrating the balance between the sensitivity and specificity; it also provides the basis to generate summary statistics. In our model assessment and selection, we used the area under the ROC curve as the measuring metric, and we also reported sensitivity and specificity.

**Statistical model selection and model performance evaluation**

All statistical modeling, variable selection and model evaluation were done with R ([R\_Core\_Team 2014](#_ENREF_2)) and caret package (<http://caret.r-forge.r-project.org/>). Dataset containing “normal” and “OSCC” samples (total 195 samples) were used to for building the prediction models. First of all, we randomly separated the dataset into two parts with 70:30 ratio, 70% of the samples were used in model selection and optimization; 30% hold-off samples were used for testing and evaluation. We selected six robust models and compared their performances based on the assessment metrics, these included: 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 assess each model’s performance, we started off with the default parameters and further optimized the hyperparameters to achieve the best performance. The overall assessing procedure involved the resampling processes, which included (1)10 fold cross-validation within each pass (2) and repeated for five times. The model performance was ranked according to the area under the ROC curve, individual sensitivity and specificity for each round.

**Establishing the Oral Cancer Risk Index (OCRI)**

To build the oral cancer risk index, (1) we finalized on one set of EdTAR parameters and processed D.I. values from all three clinically defined classes, (2) and selected the Support Vector Machine with a radial kernel function implemented in an R kenlab ([Karatzoglou 2004](#_ENREF_1)) package. To optimize the hyperparameters, we used two-class samples and the same random sampling procedure to recreate the training data set and hold off dataset. The training data 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 cost C = 32 and hyperparameter sigma = 0.6456. Since the model was built based on pathological classified “normal” and “OSCC” cases, and tested on the hold off dataset using the selected support vectors determined during the training process. Eventually, the model will report the probability an unknown sample being a class of “OSCC” vs. “Normal”. In our research, we extended the prediction probability into a quantifiable index representing the risk that a clinical sample would be classified as the carcinoma. The ORCI ranges between 0 and 1, where 0 indicates the lowest risk of OSCC and 1 indicates the highest risk of OSCC.

**Results**

**DNA staining and imaging analysis with classifier (Yao for Figure 1 a-c)**

The D.I. values obtained from the imaging processing software normally represent a mixture of cell populations and often displayed as a histogram (figure 1c); this presents to us a typical problem involving a mixture of hidden distributions. Strategies had been reported to handle such case by successfully estimated parameters for each of the individual distribution. However, in the exfoliative cytometry experiment, the portion of potential population was largely skewed. To match up the real case, we simulated three populations representing the normal cell diploid population, the tetraploid population and hypertetraploid/aneuploidy population 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 aneuploidy cell population (blue; µ=2.300, σ=0.5). However, the ratio among the three populations was set at: 0.893: 0.092:0.05. This reflects the general exfoliative cytometry results even when the aneuploidy cell population does exit, the actual signal (peak) from such a cell population is often buried (figure 1d).

**Data stripping and signal magnification**

In common exfoliative cytometry practice, hypertetraploid/aneuploidy population usually has less domination in the experimental sample. This challenge is further exemplified with the clinically undefined OLK cases. Therefore, the “real signal” for diagnosis could be overlooked as shown in our simulated data (figure 1d). To overcome such a challenge, we explored and for the first time proposed to sequentially strip out cell population from the left most and reveal the “real signal” which could provide information about abnormal cellular prognosis. In figure 2, it showed the detail work flow of expert-guided data transformation and reconstruction (EdTAR). The detail process started from a collective raw D.I. values produced from the image processing software (ref), then the empirical density estimation was applied, which provides the estimator function for peaks to be identified from the data. For each sample diagnosed with different pathological classification, peaks were successfully revealed (figure 3 A, D, G). In a normal sample (3A), the peak was located at D.I. value = 0.995. It was close to the G0/G1 normal quotient for the diploid cell population. Sometimes, an extra peak(s) could also be revealed (D.I. = 0.594) to the left of the actual peak. This could be caused either by the noise in the data, which can largely due to the feature extraction; or the empirically determined band width (*h*) in the statistical procedure. In a pathologically diagnosed OLK sample (3D), not only did it reveal the major peak (D.I. = 0.798), more peaks were clearly revealed to the right side of the peak for diploid population. Noticeably, the first peak/population that represented the normal diploid cell population took the majority of the distribution density ( > 95%, data not shown); where some peaks were located in the range close to or great than the threshold for a hypertetraploid/aneuploidy population (2.3). As explained in the methodology, the population (where those peaks resided) with extreme low density would almost impossible to be correctly estimated. In the example OSCC sample (3G), similar density plot and peak pattern as OLK sample were observed. In the OSCC sample, not only was the normal diploid cell population and peak, a much clearly tetraploid population was also revealed. The density ratio between these two populations was much prominently revealed, and these two “normal” population again took the major density of the distribution plot ( > 95%, data not shown). Although, we observed that peaks that represented the possible hypertetraploid/aneuploidy population were located much to the right side on the x-axis, for the same sake explained earlier, such population with extreme low density would almost impossible to be correctly estimated without further effort.

As shown the EdTAR work flow (figure 2), three more steps would be necessary if two or more cell populations were determined from the initial density estimation and peak identification procedure. Follow the work flow,

**Creating new variables from the processed data (figure 3 C, F, I)**

**Statistical predication model assessment and selection (figure 4)**

**Oral cancer risk index (OCRI) for clinical samples (figure 5)**

**An OLK diagnosis with high ORCI progressed to OSCC in follow up (figure 6, yao C & D)**

**Discussion and conclusion**

**Tables**

Table 1. **Parameter setting for ExGCRn process**

Table 2. **Prediction results on “olk” sample**

**Table 3.**

**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 0.893:0.092:0.05, 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). This patient was regularly followed up in outpatient clinic. A tumor was observed in August 2011. Histopathology of the surgically resected tumor confirmed the diagnosis of squamous cell carcinoma (D).

Reference

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.

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