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

**Main contribution and objectives:**

1. A novel way to process the image data -- ExGCRn (figure 2/3)
2. Model selection and assessment -- model\_assessment table (table 1/figure 4/5)
3. Render high-risk vs. low risk for OLK patient, and guide follow up schedule accordingly (figure 6)

**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 check up, 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**

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

**Expert-guided data cleaning and reconstruction**

For each DNA imaging sample, all available DNA index (D.I.) values were exported from the “software name” [reference here]. With the unknown distribution of the random variable(s), kernel density smoothing function was applied to the mixtures of three potential cell populations (normal, mitotic, and aneuploidy). Gaussian kernel was chosen for the smoothing purpose. With the nonparametric estimation of the probability density function, a mathematical procedure [reference here] was applied to search for local peak(s), which was assumed to represent the mode for each cell population. Based on our biological theory, a few assumptions were introduced. It was assumed that the mean D.I. value for normal cell population was “1c”; for mitotics cell population was “2c”; and the cell population with abnormal dividing (aneuploidy population) was above “2.2c” [reference here].

From here a sequential parsing the D.I. value was implemented. (1) To successfully extract the normal cell population, we searched along the mixed density probability density functions (pdf) and located the peak fell around the vicinity of “1c”. Then, we only used the data on the left of the peak to estimate the mean and standard deviation for the first population. Using the background normal distribution as our assumption with the estimated mean and standard deviation, the normal population was extracted out of the dataset. (2) A similar procedure was applied to extract the mitotic cell population and in the end only data left (if any) belonged to the abnormally dividing population. In case, the dataset only contained normal cell population, no cleaning was applied.

The next step is to reconstruct the “useful” dataset for the statistical modeling. To do so, we need summary statistics (mean and standard deviation) from the first two populations and number of data points in each population. The newly constructed data would normally represent the mixtures of two or three population at a control ratio, which provides the basis for building the following prediction model. (1) If only the normal population was determined, a ratio at: 98:1.5:0.5 was used to reconstruct the final data (2) If both normal and mitotic population were determined, the actual ratio between the two families was used and together consisted 99.5% toward the total leaving unchanged 0.5% for the abnormal population (3) If all three populations were determined, the actual ratio between the two families was used and together consisted 90% toward the total leaving unchanged 10% for the abnormal population.

The last step is to convert the reconstructed D.I. values to sixteen measurements of hidden variables. Now, with the known mixture ratios of population(s) and both the summary statistics and raw data value, we expand the D.I. measurement value on the scale between 0 and 8 (for samples with D.I. values greater than 8, they were converted 8). A mixed density was further consulted and then discretized into 16 bins, the density for each bin was stored as the “measurement” for each hidden variable.

**Prediction performance metrics**

For 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. In contrast, such measures as accuracy, the proportion of correct predictions, the ROC curve, and the Matthews Correlation Coefficient (MCC)74, 75, enable a single parameter comparison of performance of binary classification models. The MCC is defined as:

**Statistical prediction model performance evaluation and selection**

**Predication on clinical patient samples**

**Results**

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

**Data stripping and signal magnification (figure 1d, 2, 3)**

**Creating new variables from the processed data (figure 4)**

**Selecting SVM as predication model (figure 5)**

**Predication on OLK sample (figure 6)**

**Building clinical risk index metrics for OLK patients**

Discussion and conclusion

**Tables**

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

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

**Table 3.**

**Figures Legends**

**Figure 1.** **Distribution of cellular DNA contents in exfoliative cytology experience “*please ask Yao to proof read A - C*”** **(A)** Selected cells with abnormally high DNA contents above D.I. value > 2.3. **(B)** Scatter plots of cells in the in exfoliative cytology study, y-axis indicates the area of captured nucleus image and x-axis indicates the corresponding DNA Index (D.I.) values. **(C)** Distribution histogram of D.I. values of all captured nucleus image. **(D)** Distribution histogram of D.I. values of simulated three cell populations and mixture of three. Red density was simulated from normal distribution (µ=1.001, σ=0.19); green density was simulated from normal distribution (µ=2.002, σ=0.25); blue density was simulated from normal distribution (µ=2.300, σ=0.5); black density was the mixture of three populations at ratio: 0.893:0.092:0.05.

**Figure 2. Expert-guided data transformation and reconstruction (EdTAR) work flow:** The EdTAR data process starts with DNA Index (D.I.) values. Briefly, the density of all D.I. values was explored and all candidate peaks were identified from the smoothed curve. With the expert-guided major parameters, i.e. theoretical mean of diploid or tetraploid cells, clinical OSCC diagnosis D.I. threshold, etc., diploid cell population was extracted and further filtered (if more than one populations were detected). Then, same procedure was applied onto existing tetraploid cell population leaving all the remaining cells of the aneuploidy population. Using the same expert-guided parameters (i.e. missing ratios of candidate populations) reconstruct new variables (of D.I.) across a wide rage [0 – 8] using the discrete density at each interval. The newly constructed data was used in training the statistical model and establish an Oral Cancer Risk Index (OCRI) panel.

**Figure 3. EdTAR processing results on three candidate samples clinically differentiated classes: Normal, OLK, and OSCC.**  (A-C) was from a normal sample, (D-F) was from an OLK sample, and (G-I) was from an OSCC sample. All density plots have x-axis of DNA Index (D.I.) value, where y-axis as density. Vertically, 3A, 3D and 3G showed density plot of all available D.I. values from each sample respectively. In the normal sample (3A), a major peek located at 0.995 represents the mean of diploid cell population, where another small peaks (D.I. = 0.594) was a minor hump possibly due to measurement error from the image process. In the OLK sample (3D), a major peek located at 0.798, represents the mean of diploid cell population. It deviated toward the left from “1” owing to the smoothing algorithm and it left another minor peak at 1.25. Three more almost invisible peaks were located at 1.75, 2.22, and 2.74 respectively. It was further indicated that the first cell population consist the main density (3590 cells). In the OSCC sample (3G), a major peek located at 1.02, represents the mean of diploid cell population. Another obvious peak was located at 1.79, which was deemed to represent the tetraploid population. The second peak represented the mean of the tetraploid population, which again deviated toward the left from “2” owing to the smoothing algorithm. Three more almost invisible peaks were located at 3.25, 3.57, and 3.99 respectively. These were deems to represent the aneuploidy cell population. The second column three plots showed the net results from EdTAR process, which was to reveal the hidden signals. (3B) was same as (3A) since only a single diploid population was identified and cleaning was unnecessary. (3E) and (3H) clearly showed the benefit from the EdTAR process where the tetraploid cell populations were showed and all possible candidate aneuploidy cells (cell population) were revealed for each of the two samples respectively. The last column of three plots (3C), (3F) and (3I) showed boxplot of newly constructed variables from EdTAR process for three samples. The x-axis indicated the new variables along the D.I range [0 – 8 ] and y-axis was the boxplot of available values for each variable.

**Figure 4. Statistical model assessment** In order to predict the OCRI, we explored a series of prediction models. Seven models (SVM, RRF, PLR, NNET, KNN, and CART) were shown with three major performance assessment metrics (ROC, sensitivity and specificity), ranging between 0 and 1. Each model were trained on the training data and tested on testing data. Each boxplot showed the distribution each of the three metrics from the model resample assessment (R caret package <http://cran.r-project.org/web/packages/caret/index.html>)

**Figure 5. Oral Cancer Risk Index (OCRI) panel** Based on the model selection and assessment, an oral cancer risk index (OCRI) was established. The process was tested on a hold-off dataset with all three (known) clinical classifications: Normal, OLK, and OSCC. Samples from each class were assessed with the newly proposed ORCI panel. The y-axis showed the ORCI index ranging between 0 and 1, where 0 indicates the lowest risk (of cancer) and 1 indicates the highest risk (of cancer).