Introduction

Oral and pharyngeal cancer, grouped together, rank sixth among most common cancer types in the world [1]. Oral cancer is one of the most common malignancies and contributes the majority of cancer morbidity and mortality worldwide [2]. 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%) [3]. Oral squamous cell carcinoma (OSCC), a common type of oral cancer, can develop from precancerous lesions, such as oral leukoplakia (OLK) [4] [5]. 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 is proved successful method in screening for epithelial dysplasia in situ or invasive carcinomas of the uteri cervix [6]. 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 patent.

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 is much time consuming and sometimes 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 to explain and interpret 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 or it is hard to differentiate the signal from the noises.

Upon fully understanding the challenges inherited in this research, 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 to improve the sensitivity and the specificity of DNA quantitative analysis for OLK patients. To predict the progressing direction of a clinical defined the oral leukoplakia lesion, we tested our SVM model and were not surprised that the prediction results varied across the entire panel of “probability”, from almost normal to almost OSCC. This result was aligned well with the common clinical diagnosis, but it 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.

References

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

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

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

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

5. 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.

6. Bocking A, Sproll C, Stocklein N, Naujoks C, Depprich R, et al. (2011) Role of brush biopsy and DNA cytometry for prevention, diagnosis, therapy, and followup care of oral cancer. J Oncol 2011: 875959.

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

8. Remmerbach TW, Mathes SN, Weidenbach H, Hemprich A, Bocking A (2004) [Noninvasive brush biopsy as an innovative tool for early detection of oral carcinomas]. Mund Kiefer Gesichtschir 8: 229-236.

9. Haws J, Rhodus NL, Williams B, Griffin RJ (2004) Number of apoptotic cells in brush biopsies of patients with oral leukoplakia. ORAL SURG ORAL MED ORAL PATHO 97: 461-461.

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.