Statistical Modeling via Bootstrapping and Weighted Techniques Based on Variances

An Oral Health Case Study

W. M. A. W. Ahmad
School of Dental Sciences
Universiti Sains Malaysia
Malaysia
wmamir@usm.my

N. A. Aleng
School of Informatics and
Applied Mathematics, Universiti
Malaysia Terengganu, Malaysia
azlida_aleng@umt.edu.my

Z. Ali
School of Mathematical Sciences
Universiti Sains Malaysia
Malaysia
zalila_ali@usm.my

M. S. M. Ibrahim
School of Dental Sciences
Universiti Sains Malaysia
Malaysia
shafiqmat786@gmail.com

Abstract—**Multiple logistic regression is a methodology of handling dependent variables with a binary outcome. This method is becoming increasingly widespread as a statistical technique that represents a discrete probability model. Many studies have focused on the application but less on the methodology building. This study aims to provide an applied method for multiple logistic regression which is called modified Bayesian logistic regression modeling as an alternative technique for logistic regression analysis that focuses on a combination of the bootstrap method using SAS macro and weighted techniques based on variances using SAS algorithm. Data on oral cancer were applied to illustrate a real scenario of oral health data. This data will be applied to the multiple logistic regression algorithm and modified Bayesian logistic regression. Results from both cases are strongly supported by clinical studies. Through the proposed algorithm, the researcher will have an option whether to analyze the data with the usual or an alternative method. Final results indicate that the modified procedure can provide more efficient results especially for the case which involves statistical inferences.**

Keywords-multiple logistic regression; bootstrap; Bayesian and weighted techniques

I. INTRODUCTION

The logistic regression, analyzes the relationship between multiple independent variables and categorical dependent variables [1]. The multiple logistic response functions is $E\{Y\}=\dfrac{\exp(\boldsymbol{\beta}'X)}{1+\exp(\boldsymbol{\beta}'X)}$ where

$$\boldsymbol{\beta}'X=\beta_0+\beta_1X_1+\cdots+\beta_{p-1}X_{p-1}$$

$$\boldsymbol{\beta}'X=\beta_0+\beta_1X_{i1}+\cdots+\beta_{p-1}X_{i,\,p-1}$$

$$\underset{p\times 1}{\boldsymbol{\beta}}=\begin{pmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_{p-1}\end{pmatrix}\quad \underset{p\times 1}{\mathbf{X}}=\begin{pmatrix}1\\ X_1\\ \vdots\\ X_{p-1}\end{pmatrix}\quad \underset{p\times 1}{\mathbf{X_i}}=\begin{pmatrix}1\\ X_{i1}\\ \vdots\\ X_{i,p-1}\end{pmatrix}$$

The multiple logistic regression models can, therefore, be stated as follows: $\boldsymbol{Y_i}$ are independent Bernoulli random variables with expected values $E\{Y_i\}=\pi_i$ where: $E\{Y_i\}=\pi_i=\dfrac{\exp(\boldsymbol{\beta}'X_i)}{1+\exp(\boldsymbol{\beta}'X_i)}$. The X observations are considered to be constants. Alternatively, if the X variables are random $\mathrm{E}\{Y_i\}$ is viewed as a conditional mean, given the values of $X_{i1}+X_{i2}+\ldots+X_{i,p-1}$.

A. Bootstrapping, Weighted Techniques and Bayesian Approach with SAS

Authors in [2] introduced the bootstrap method which emphasizes on an empirical density function (EDF). The basic concept of bootstrap is that it is initiated with an original sample which is taken from the studied population. The second step is to copy the original sample a number of times in order to create a pseudo-population. Then, it draws several samples considering random sampling approach thus providing a new comprehensive sample from the original sample. It stores the new set of data from the original dataset and creates a new distribution for further analysis [2, 3]. The Bayesian analysis involves the posterior distribution. In the stage of Bayesian estimation procedures, the posterior distribution will play an important role especially in statistical inferential. While running the analysis the summary statistics for the posterior distribution samples are produced by default. The SAS statements of OUTPOST provide an option that saves the samples in the SAS data set for further processing. PROC GENMOD procedure fits generalized linear models with Bayesian methods (considering Bayesian estimation procedures) with a normal error term [4]. In SAS programming procedure, the SEED option is to maintain reproducibility. By default, the uniform prior is a flat prior with a distribution that reflects ignorance of the location of the parameter. Its placing an equal likelihood for all possible values which regression coefficients can take. PROC GENMOD also produces convergence diagnostics where ODS Graphics is enabled in SAS statements which provides a section of assessing Markov chain convergence diagnostics and their interpretation [5].

Weighting is a very important technique which involves adjusting data to reflect dissimilarities in the number of population units that represent each respondent [6]. There are several techniques that can be used as a weighted method like weighted by mean, standard deviation or variances to apply to the model according to the sample population of interest. Moreover, a weighted technique allows assigning different weights to the different cases in data analysis. The aim of the weighted method is to correct the skewness and to make the sample more representative of a true population.

II. Methodology: Algorithm Building and Results

We used secondary oral medical data which involved 23 oral cancer patients from Universiti Sains Malaysia (USM). The selected variables are nerve invasion (nerv_inv), gender (gen), betel quid (bet), tumour site (tum_site) and tumour size (tum_size). To explore the underlying association between nerve invasion and the selected explanatory variables, a set of the regression model is fitted in this section. Let us define the following dichotomous variables for the model: $Y_{ii}=0$ has not nerve invasion and $Y_{ii}=1$ has nerve invasion. The proposed model is given in equation form as follows:

$$\mathbf{b'X}_{ij} = \beta_0 + \beta_1(Gender) + \beta_2(Betel\ Quid) + \beta_3(Tumour\,site) + \beta_4(Tumour\,size) + \varepsilon_{ij}..... \quad (1)$$

Then the logistic regression model for (1) is given as:

$$\mathrm{P}(Y_i = 1 \mid X_i) = \hat{\pi}$$

$$= \frac{\begin{pmatrix}\exp(\beta_0 + \beta_1(Gender) + \beta_2(Betel\ Quid) \\ +\beta_3(Tumour\,site) + \beta_4(Tumour\,size))\end{pmatrix}}{\begin{pmatrix}1+\exp(\beta_0 + \beta_1(Gender) + \beta_2(Betel\ Quid) \\ +\beta_3(Tumour\,site) + \beta_4(Tumour\,size))\end{pmatrix}}$$

$$= \left[1+\exp\left(-\begin{pmatrix}\beta_0 + \beta_1(Gender) + \beta_2(Betel\ Quid) \\ +\beta_3(Tumour\,site) + \\ \beta_4(Tumour\,size)\end{pmatrix}\right)\right]^{-1}$$

Then we obtain (2) as follows:

$$\mathrm{P}(Y_i = 1 \mid X_i) = \hat{\pi} = = \left[1+\exp\left(-\begin{pmatrix}\beta_0 + \beta_1(Gender) \\ +\beta_2(Betel\ Quid) \\ +\beta_3(Tumour\,site) \\ +\beta_4(Tumour\,size)\end{pmatrix}\right)\right]^{-1} \quad (2)$$

The estimated model for our case is given in (2). Before we apply the equation there are two main steps needed to be done which are bootstrapping and weight data.

A. Multiple Logistic Regression Modeling on Nerve Invasion

- Cancer cell data should be entered as follows in SAS algorithm to calculate the multiple logistic regressions..

```
Data cancer;
input Nerv_inv Gen Bet Tum_site  Tum_size ;
datalines;
 0   0   0   1   1
 0   1   0   2   0
 0   1   0   2   0
 0   1   0   2   0
 0   1   0   3   0
 0   1   0   2   0
 0   1   0   3   0
 0   0   0   2   0
 1   0   1   1   1
 0   0   0   4   0
 0   0   1   1   1
 1   0   1   3   1
 0   0   1   4   0
 0   0   1   1   0
 0   0   1   1   1
 0   0   0   2   0
 0   1   0   4   0
 0   0   0   2   0
 0   1   0   2   0
 0   1   0   2   0
 1   0   1   2   0
 1   1   0   2   0
 1   0   0   4   0
 ;
  run;
  ods rtf file='abc.rtf' style=journal;
```

- Run the analysis using multiple logistic regression. Below is the syntax of multiple logistic regression.

```
/* Run The Logistic Regression Through Proc Logistic*/
ods graphics on;
proc logistic descending data=cancer;
    model Nerv_inv(event='1') = Gen Bet Tum_site Tum_size / rsquare expb lackfit;
    roc 'Gen Bet Tum_site Tum_size ' Gen Bet Tum_site Tum_size ;
run;
ods graphics off;
ods rtf close;
```

B. Results: Using Multiple Logistic Regression Modeling on Nerve Invasion

In Figure 1 and Table I the results of using the above syntax are shown. None of the variables in the list is significant. The area under the ROC curve is 0.70556. The model can accurately discriminate 71.0% of the cases (it significantly discriminates more than half of the cases).

C. Modified Multiple Bayesian Logistic Regression on Nerve Invasion

- Adding bootstrapping to the calculation.

The following syntax calculates the data using a bootstrap method and prints them out:

```
%macro bootstrap (data=_last_, booted=booted, boots=4, seed=1234);
data &booted;
pickobs = int(ranuni(&seed)*n)+1; /* This procedure  is randomly select an integer from 1 to n*/
```

```
    set &data point = pickobs nobs = n; /*This procedure point
tells SAS to read value pickobs, nobs sets n to number of
bservation in & data, when the point option is used SAS will
loop through the data step forever*/
     replicate=int(i/n)+1;
    /* and saves number of current bootstrap*/
      i+1;
      if i > n*& boots then stop;
    /* This procedure stop will leave data set when n*&boots
observation have been created*/
      run;
    %mend bootstrap;
```

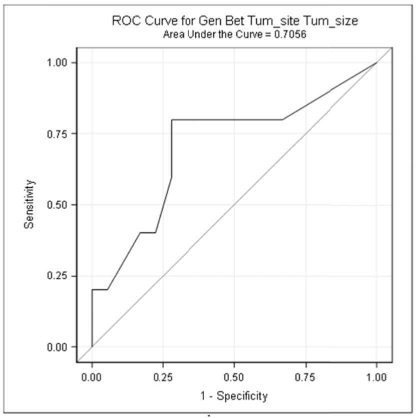

Fig. 1. ROC Curve

TABLE I. MAXIMUM LIKEHOOD ESTIMATES

Parameter	Estimate	Standard Error	Pr>ChiSq
Intercept	-3.2222	2.1179	0.1282
Gen	-0.1454	1.5915	0.9272
Bet	0.9728	1.4332	0.4973
Tum_site	0.5295	0.5908	0.3701
Tum_size	1.2608	1.6022	0.4314

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	4.6580
Pr>ChiSq	0.4590

- Logistic regression using bootstrap and weighted sample.

```
Data cancer;
input Nerv_inv Gen Bet Tum_site  Tum_size ;
datalines;
  0    0    0    1    1
  0    1    0    2    0
  0    1    0    2    0
  0    1    0    2    0
  0    1    0    3    0
  0    1    0    2    0
  0    0    1    3    0
  0    0    0    2    0
  1    0    1    1    1
  0    0    0    4    0
  0    0    1    1    1
  1    0    1    3    1
  0    0    1    4    0
  0    0    1    1    0
  0    0    1    1    1
  0    0    0    2    0
  0    1    0    4    0
  0    0    0    2    0
  0    1    0    2    0
  0    1    0    2    0
  1    0    1    2    0
  1    1    0    2    0
  1    0    0    4    0
;
run;
ods rtf file='abc.rtf' style=journal;
```

- The syntax for generating a bootstrap sample

```
% bootstrap(data= cancer, boots=4);
run;
```

- The syntax for bootstrapping data printing

```
proc print data=booted;
run;
```

- The syntax for logistic regression and bootstrap to get the residuals

```
    proc genmod data= booted descending;
    class Nerv_inv;
    model  Nerv_inv  =  Gen  Bet  Tum_site  Tum_size  /
dist=binomial link=logit;
    output out=pseudo_data  reschi=residual;
    run;
```

- Compute the absolute and squared residuals

```
data data_nerve_invasion1;
set pseudo_data;
absresid=abs(residual);
sqresid=residual**2;
```

- Run a regression with the absolute residuals vs. independent variables to get the estimated standard deviation

```
proc genmod data=data_nerve_invasion1;
model absresid = Gen Bet Tum_site Tum_size;
output out=data_nerve_invasion p=varians_hat;
run;
```

- Compute the weights using the estimated variances

```
data data_nerve_invasion;
set data_nerve_invasion;
varians_weight=1/varians_hat;
label varians_weight = "weights using squared residuals";
```

- Do the weighted least squares using the weights from the estimated variances

```
proc genmod  data=data_nerve_invasion;
weight varians_weight;
model Nerv_inv = Gen Bet Tum_site Tum_size /
dist=binomial
link=logit;
run;
ods graphics on;
```

- Run the logistic regression with weighted data through proc logistic.

```
proc logistic descending data=data_nerve_invasion
plots= effect plots= roc(id=prob);
weight varians_weight;
model Nerv_inv(event='1') = Gen Bet Tum_site
```

```
Tum_size / rsquare expb lackfit;
roc 'Gen Bet Tum_site Tum_size ' Gen Bet Tum_site
Tum_size ;
run;
```

- Run the logistic regression with weighted data through proc genmod.

```
proc genmod data=data_nerve_invasion descending;
weight varians_weight;
model Nerv_inv = Gen Bet Tum_site Tum_size/
dist=binomial
link=logit;
bayes nbi=1000 nmc=10000 thin=2 seed=1
out=posterior;
run;
ods graphics off;
ods rtf close;
```

D. Results: Modified Multiple Bayesian Logistic Regression on Nerve Invasion

The area under the curve of ROC (Figure 2) is 0.78556. The model can accurately discriminate 78.5% of the cases (it significantly discriminates more than half of the cases).

TABLE II. MAXIMUM LIKEHOOD PARAMETER ESTIMATES

Parameter	Estimate	Standard Error	Pr > ChiSq
Intercept	1.9942	1.5181	0.1890
Gen	1.3686	1.1029	0.2146
Bet	-1.6603	0.9156	0.0698
Tum_site	-0.1978	0.4045	0.6248
Tum_size	-0.3001	1.0968	0.7844

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	4.9294
Pr>ChiSq	0.2946

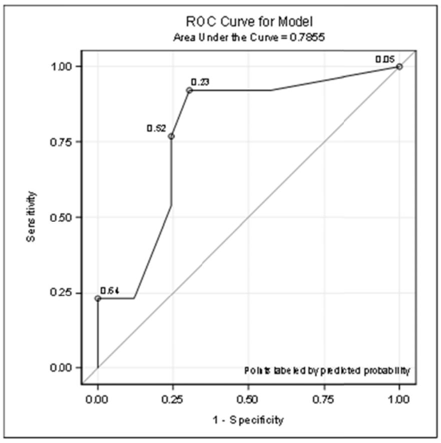

Fig. 2. ROC curve

A converged Markov chain explored the parameter space sufficiently according to the distribution. Besides that, convergence by visualization can also be examined through the trace plot for each variable (assess convergence with visual examination), not just the ones of interest. Inference should be based on convergence Markov chains. Figures 3 to 6 show that Markov chain has stabilized with a good mixing by visual examination of the trace plot. The samples stay close to the high-density region of the target distribution (convergence looks good and consistent in parameters and they are stabilized).

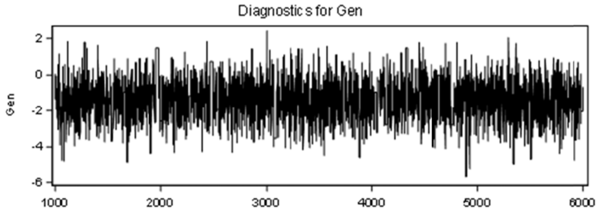

Fig. 3. Trace plot for gender

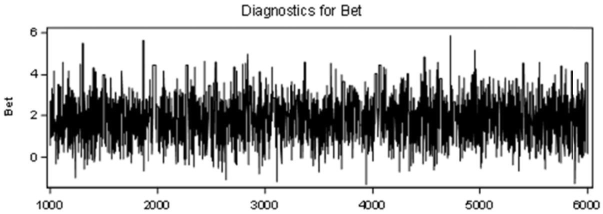

Fig. 4. Trace plot for betel quid

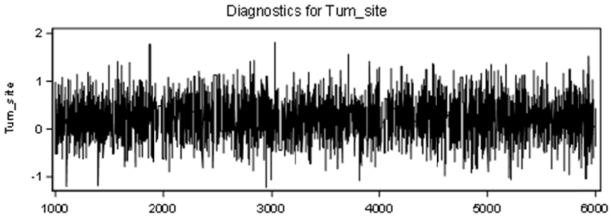

Fig. 5. Trace plot tumour site

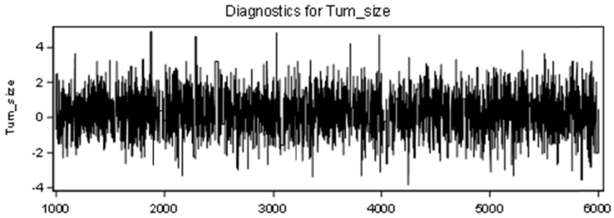

Fig. 6. Trace plot tumour size

E. Comparison of the Multiple Logistic Regression with Modified Multiple Bayesian Logistic Regression Model

Table III shows the summary of the results for both methods, multiple logistic regression and modified multiple Bayesian logistic regression. As expected, the modified method's result is improved. Gender and betel quid factors appear to be more significant at $p<0.25$, which is clinically significant. This study significance factor on gender is also supported in [7] in which the skin tumors with perineural invasion were found in male sexual orientation, in the high-risk mask zone of the face. Authors in [8-10] found that chewing betel quid independently contributes to the risk of cancer which was the same with our findings. Moreover, the factor which is not related to the dependent variable also appears to have a not significant association with nerve invasion with non-significant value. This shows that the alternative method can improve the results by fixing the p-value and the ROC curve. Figure 7 shows that the increased area under curve equals to 0.7056 (obtained from multiple logistic regression) and to 0.7855 (obtained from modified multiple Bayesian logistic regression). The position of the ROC on the graph reflects the accuracy measuring of the model and overall test performance.

TABLE III. RESULTS COMPARISON OF THE TWO METHODS

Multiple Logistic Regression				Modified Multiple Bayesian Logistic Regression		
Parameter	**Estimate**	**Standard Error**	**Sig. Value**	**Estimate**	**Standard Error**	**Sig. Value**
Intercept	-3.2222	2.1179	0.1282	1.9942	1.5181	0.1890
Gender	-0.1454	1.5915	0.9272	1.3686	1.1029	0.2146*
Betel Quid	0.9728	1.4332	0.4973	-1.6603	0.9156	0.0698*
Tumour site	0.5295	0.5908	0.3701	-0.1978	0.4045	0.6248
Tumour size	1.2608	1.6022	0.4314	-0.3001	1.0968	0.7844

Chi-Square 4.6580 Chi-Square 4.9294

Pr > ChiSq 0.4590 Pr > ChiSq 0.2946

Hosmer and Lemeshow Goodness-of-Fit Test

*Significant at p < 0.25

III. CONCLUSION

This paper examined and focused on two main sections, the first of which is the methodology based on algorithm building with bootstrap weighted data by using variances. Second is the output gained as a new finding for the applied research. In our study, the potential variables were influencing the nerve invasion among oral health problem patients. According to Table II, nerve invasion may depend on factors that are related to gender β_1=1.13686, p<0.25 and taking betel quid β2=-1.6603, p<0.25. This paper demonstrates two different techniques that can be used to such relationships. At first, by using the multiple logistic regression we obtained the result as follows:

$$\mathrm{P}\left(Y_i = 1 \mid X_i\right) = \hat{\pi}$$

$$= \left[1 + \exp\left(-\begin{pmatrix} \beta_0 - 0.1454\left(Gender\right) + 0.9728\left(Betel\ Quid\right) \\ +0.5295\left(Tumour\ site\right) + 1.2608\left(Tumour\ size\right) \end{pmatrix}\right)\right]^{-1}$$

And second, the modified Bayesian multiple logistic regression:

$$\mathrm{P}\left(Y_i = 1 \mid X_i\right) = \hat{\pi}$$

$$= \left[1 + \exp\left(-\begin{pmatrix} \beta_0 + 1.3686\left(Gender*\right) - 1.6603\left(Betel\ Quid*\right) \\ -0.1978\left(Tumour\ site\right) - 0.3001\left(Tumour\ size\right) \end{pmatrix}\right)\right]^{-1}$$

This proposed method can be alternatively applied to an ordinary multiple logistic regression in the case of small sample size. This method produces better results compared to the previous one.

IV. DISCUSSION AND RECOMMENDATIONS

The objective of this research is to develop an alternative method for multiple logistics regression with some modification on the programming calculation. Through this procedure, we can determine the associated factors for the nerve invasion (for the case of oral study). On top of that, the effectiveness of calculation by combining several statistical methods is the primary focus, which leads to significantly better results. The combined calculation method between weighted techniques and bootstrap approach showed improvement of the results which can be seen in Tables I and II. Table I shows the standard calculation and Table II shows the modified calculation. Both Tables shown some improvement in the results.

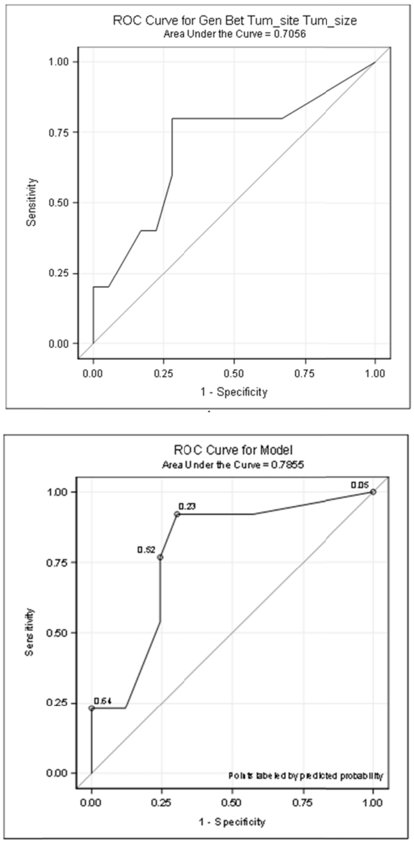

Fig. 7. Comparison of ROC curve of the two methods

In Table II, factors which are clinically associated with the dependent variable show their true association. We can also estimate the precise odd ratio based on the Wald Chi-square. This promising method lead to a successful research and gave better results for the decision making especially for the decision maker. Therefore, in order to keep the effectiveness of the results, it is necessary to have a good way of calculation with some improvement of the proposed strategy. The approached method can have a better predicting result in future for the decision making. In this paper, the algorithms have a good potential to determine the potential factors that lead to oral cancer. Some recommendations are raised in the following study findings:

- There is a need to explore more on the methodology improvement in order to optimize the gained output. This could include a higher level of a combination of theoretical, methodology building and computation which may lead to higher precision and accuracy of the results.
- Performance measurements can be taken into consideration when measuring quality of the recommended algorithms.

This knowledge will empower researchers and serve as a roadmap to improve future studies.

Acknowledgment

Authors would like to express their gratitude to Universiti Sains Malaysia for providing the research funding (Grant no.1001/PPSG/8012278, School of Dental Sciences, Universiti Sains Malaysia).

References

[1] D. W. Hosmer, S. Lemeshow, R. X. Sturdivant, Applied Logistic Regression, 3rd ed, John Wiley & Sons, 2013

[2] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall/CRC, 1993

[3] G. E. Higgins, “Statistical Significance Testing: The Bootstrapping Method and an Application to Self-Control Theory”, The Southwest Journal of Criminal Justice, Vol. 2, No. 1, pp. 54-76, 2005

[4] A. Gelman, J. B Carlin, H. S. Stern, D. B. Rubin, Bayesian Data Analysis, Chapman and Hall/CRC, 2004

[5] M. Stokes, F. Chen, F. Gunes, “An Introduction to Bayesian Analysis with SAS/STAT Software”, SAS Global Forum 2014 Conference, Washington DC, USA, Paper SAS400-2014, March 23-26, 2014

[6] J. Mickey, S. Greenland, “The Impact of Confounder-Selection Criteria on Effect Estimation”, American Journal of Epidemiology, Vol. 129, No. 1, pp 125-137, 1989

[7] I. Leibovitch, S. C. Huilgol, D. Selva, S. Richards, R. Paver, “Basal cell carcinoma treated with Mohs surgery in Australia III. Perineural invasion”, Journal of the American Academy of Dermatology, Vol. 53, No. 3, pp. 458-463, 2005

[8] C. T. Liao, C. J. Tung-Chieh, H. M. Wang, I. H. Chen, C. Y. Lin, T. M. Chen, L. L. Hsieh, A. J. Cheng, “Telomerase as an independent prognostic factor in head and neck squamous cell carcinoma”, Head & Neck, Vol. 26, No. 6, pp. 504–512, 2004

[9] C. T. Liao, J. T. Chang, H. M. Wang, S. H. Ng, C. Hsueh, L. Y. Lee, C. H. Lin, I. H. Chen, S. F. Huang, A. J. Cheng, T. C. Yen, “Analysis of risk factors predictive of local tumor control in oral cavity cancer”, Annals of Surgical Oncology, Vol. 15, No. 3, pp. 915–922, 2008

[10] C. T. Liao, C. J. Kang, J. T. Chang, H. M. Wang, S. H. Ng, C. Hsueh, L. Y. Lee, C. H. Lin, A. J. Cheng, I. H. Chen, S. F. Huang, T.C. Yen, “Survival of second and multiple primary tumors in patients with oral cavity squamous cell carcinoma in the betel quid chewing area”, Oral Oncology, Vol. 43, No. 8, pp. 811–819, 2007