Sains Malaysiana 41(5)(2012): 633–640

Determining the Critical Success Factors of Oral Cancer Susceptibility Prediction in Malaysia Using Fuzzy Models

(Menentukan Faktor Kejayaan Kritikal dalam Peramalan Kecenderungan Terjadinya Kanser Mulut di Malaysia Menggunakan Model Kabur)

ROSMA MOHD DOM, BASIR ABIDIN, SAMEEM ABDUL KAREEM,
SITI MAZLIPAH ISMAIL & NORZAIDI MOHD DAUD*

ABSTRACT

The aim of the study was to determine the success factors of oral cancer susceptibility prediction using fuzzy models. Three fuzzy prediction models including fuzzy logic, fuzzy neural network and fuzzy linear regression models were constructed and applied to a Malaysian oral cancer data set for cancer susceptibility prediction. The three models' prediction performances were evaluated and compared. All the three fuzzy models were found to have 64% prediction accuracies for 1-input and 2-input predictor sets. However, when the number of input predictor set was increased to 3-input and 4-input, both fuzzy neural networks' and fuzzy linear regression's prediction accuracies increased to 80%, while fuzzy logic prediction accuracy remains at 64%. Fuzzy linear regression model was found to have the capability of quantifying the relationships between input predictors and the predicted outcomes and also suitable for small sample size. Fuzzy neural network model on the other hand, handles ambiguous relationship between variables well but lacks the ability to describe input-output association. The third model, fuzzy logic, is easy to construct but highly dependent on human expert-input. The outcome of this study is a computer-based prediction tool which can be used in cancer screening programs.

Keywords: Fuzzy logic; fuzzy neural networks; fuzzy regression; oral cancer; prediction performance

ABSTRAK

Matlamat kajian ini adalah untuk mengenal pasti faktor kejayaan dalam menentukan kecenderungan terjadinya kanser mulut menggunakan model kabur. Tiga model kabur termasuk model mantik kabur, rangkaian neuro-kabur dan regresi kabur telah dibangun dan diaplikasikan ke atas data kanser mulut di Malaysia bagi tujuan menentukan kemungkinan terjadinya kanser. Kejituan ramalan terjadinya kanser mulut untuk ketiga-tiga model diukur dan dibandingkan. Ketiga-tiga model kabur didapati memberikan 64% kejituan ramalan semasa diuji menggunakan satu dan dua faktor penentu. Walau bagaimanapun, apabila bilangan faktor penentu ditambah kepada tiga dan empat, kejituan ramalan model rangkaian neuro-kabur dan regresi kabur meningkat kepada 80% tetapi kejituan ramalan model mantik kabur kekal di paras 64%. Model regresi kabur mampu mengukur kuantiti hubung kait di antara faktor penentu dengan hasilannya dan ia juga sesuai digunakan untuk sampel yang kecil. Model rangkaian neuro-kabur mengambil kira hubungan ketidaktentuan di antara faktor penentu dan hasilannya tetapi tidak mempunyai keupayaan mengukur kuantiti hubung kait di antara mereka. Model ketiga iaitu mantik kabur pula mudah untuk dibangunkan tetapi terlalu bergantung kepada input pakar. Kajian ini menghasilkan satu alat pengukur berasaskan komputer yang boleh digunakan bagi tujuan saringan pesakit kanser.

Kata kunci: Kanser mulut; kejituan ramalan; mantik kabur; rangkaian neuro-kabur; regresi kabur

INTRODUCTION

A computational intelligence concept known as fuzzy theory has been widely used in medical diagnosis and prognosis because it can bridge the gap between the numerical world, in which often symptoms are observed and measured, and the symbolic world, in which knowledge was expressed in order to be easy to read and understand by human users (Castellano et al. 2005). Fuzzy concepts have been proven to be a powerful tool for decision making systems, such as expert systems and pattern classification systems (Muzio et al. 2005). In Malaysia the use of computational intelligence in the medical sector is slowly gaining acceptance with the government's effort in encouraging the use of Tele-Medicine and Electronic-health databases (Mohan & Yaacob 2004). Hybrid fuzzy models are systems that combine fuzzy concepts with other concepts as an effort to enhance the ability of the systems. Examples include fuzzy neural network, fuzzy support vector machine and fuzzy linear regression. The objective of this study was to identify the critical success factors in oral cancer susceptibility prediction using fuzzy models in a Malaysian case study. This paper discussed three fuzzy prediction models in particular fuzzy logic, fuzzy neural network and fuzzy linear regression models. The models were used in an oral cancer case study whereby oral cancer predictions were made based on individual's demographic profiles and risk habits.

FUZZY LOGIC MODEL

Fuzzy logic is a superset of conventional (Boolean) logic. It was first introduced in the 1960s by Lofti A. Zadeh of the University of California, Berkeley as a means to handle the concept of partial truth (Zadeh 1965). Fuzzy logic is considered as one of the most powerful tools for dealing with imprecision and uncertainties. It was designed to mathematically represent uncertainty and vagueness and to provide formalized tools for dealing with the imprecision intrinsic to many problems differentiated by the membership functions (Coppin 2004).

A membership function is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The simplest membership functions are formed using straight lines. Examples include triangular membership function and trapezoidal membership function. On the other hands, Gaussian and sigmoidal membership functions are not made up of straight lines. Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic (Moraga 2000). The mapping formulated provides the basis from which decisions can be made, or patterns discerned. The fuzzy inference process can be divided into five parts which include input fuzzification, applying fuzzy operators, applying implication methods, output aggregation and finally output defuzzification as depicted in Figure 1.

FUZZY NEURAL NETWORK MODEL

In fuzzy neural network modeling, either Radial Basis Function (RBF) or Feed Forward networks were combined together to perform some form of pattern classifications or data mining tasks. A neuro-fuzzy classifying system, in general, has *n* inputs (attributes or features), x_1, x_2, x_3,…, x_n, and an output which has the form of a possibility distribution over the set $Y= \{y_1, y_2,…, y_H\}$ of class labels. Each input x_i represents one input medical attribute which could be either a 'symptom' for diagnostic purposes or a 'risk factor' for prognostic purposes (Gorzalczany & Piasta 1999). The input variable could be in numerical form like body weight, age and blood pressure or non-numerical character like pain level. Numerical-type attributes can be described by numbers or by linguistic terms represented by fuzzy sets (e.g 'age' could be 'very young', 'young', 'old', 'very old'). The output set Y, in medical and dental field, could be a set of potential diseases or possible outcomes of a particular treatment or possibly the state of a patient after some interval time.

In the effort to obtain models that are both accurate and understandable, the learning capability of neural networks can be combined with the expressiveness of fuzzy if-then rules using linguistic variables to produce Neuro Fuzzy models like ANFIS. The ANFIS system was first introduced by Jang in 1992. It posses the three main components of fuzzy inference system which are fuzzification, implication and defuzzification (Jang 1993).

In this particular case study, fuzzy neural network models were constructed and tested as a mean to provide the baseline comparison for fuzzy logic and fuzzy linear regression models in predicting oral cancer susceptibility. In the process, we also hope to be able to identify the limitations and advantages of the different fuzzy prediction models used. Fuzzy neural network model was chosen to be the baseline comparison model since literature have shown that fuzzy neural network models provide good alternative to statistical and other artificial intelligent prediction models (Muzio et al. 2005). In this study, ANFIS methodology with Sugeno's model of first order was constructed and used.

FUZZY LINEAR REGRESSION MODEL

Regression analysis is an estimation method used in finding a crisp relationship between the dependent and independent variables and also used to estimate the variance of measurement error. Fuzzy regression analysis is an extension of the classical regression analysis in which some elements of the models are represented by fuzzy numbers. If a phenomenon under study is governed by possibility variables, it is more appropriate to seek a fuzzy functional relationship for the given data regardless of the data being crisp or fuzzy (Tanaka & Lee 1998). Fuzzy linear regression provides a means for tackling regression problems lacking a significant amount of data with vague relationships between the dependent and independent variables (Savic & Pedrycz 1991). Fuzzy regression methods have also been successfully applied to modeling problems in forecasting and engineering (Nasrabadi & Nasrabadi 2004). There are two categories of fuzzy regression analysis; the first is possibilistic regression analysis which is based on possibility concepts. Possibilistic regression analysis uses fuzzy linear system

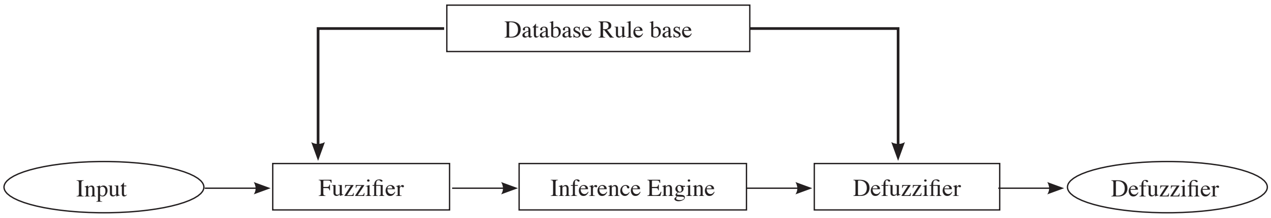

FIGURE 1. Fuzzy inference process

as a regression model whereby the total vagueness of the estimated values for the dependent variables is minimized. It was first proposed by Tanaka in 1982. In this particular case study, fuzzy linear regression model introduced by Tanaka was used to determine the association between different markers and oral cancer susceptibility. In Tanaka's possibilistic regression the response variable *Y* is written as:

$$Y = A_0 + A_1x_1 + A_2x_2 + \dots + A_k\, x_k,$$

where *Y* is the fuzzy output, $x = [x_1, x_2, \dots, x_k]$ is the real-valued input vector of independent variables and each regression coefficient A_j; $j = 0, 1, \dots, k$ was assumed to be a symmetric triangular fuzzy number with center $\alpha_j \in \Re$ and half-width $c_j \geq 0$. The fuzzy linear regression model can be rewritten as follows:

$$Y = (\alpha_0, c_0) + (\alpha_1, c_1)x_1 + \dots + (\alpha_k, c_k)x_k.$$

The following linear programming (LP) formulation was employed to estimate Aj = (α_j, c_j)

$$J = \sum_{j=0}^{k}\left(c_j \sum_{i=1}^{n} x_{ij}\right)$$

subject to $\sum_{j=0}^{k} \alpha_j x_{ij} + (1-h)\sum_{j=0}^{k} c_{jx} x_{ij} \geq y_i$ and

$$\sum_{j=0}^{k} \alpha_j x_{ij} - (1-h)\sum_{j=0}^{k} c_{jx} x_{ij} \leq y_i$$

$$a_j \in \Re, c_j \geq 0,\ j = 0,1,2,K\ ,k$$

$$x_{i0} = 1, i = 1,2,K\ ,n$$

$$0 < h < 1,$$

where *J* is the total fuzziness of the fuzzy regression model. The *h* value is the threshold level that determines the degree of fitness of the fuzzy linear model to its data (Wang & Tsaur 2000).

ORAL CANCER SUSCEPTIBILITY PREDICTION EXPERIMENTATION

The development of a computer-based oral cancer prediction tool is inline with the Oral Cancer Research and Coordinating Center, OCRCC's efforts in coming up with a better understanding of oral cancer incidence and prevalence in Malaysia. OCRCC hopes to provide better guidelines for clinicians and health care providers in handling oral cancer patients as well as in providing alternative tools for oral cancer screening initiatives in this country. Thus an unmatched case-control study was conducted using 84 newly diagnosed oral cancer patients and 87 non-cancer subjects selected from the same locations as cases. Sociodemographic data was obtained from the Malaysian Oral Cancer Database and Tumour Bank System (MOCDTBS) provided by the OCRCC, University of Malaya, Malaysia. Cancer patients' and control group's demographic profiles (age, gender) and oral cancer risk habits (cigarette smoking, alcohol drinking, tobacco chewing) were used as input variables and the outcome refers to health condition of 'cancer' or 'healthy'. Through the MOCDTBS, peripheral blood was obtained from consented individuals, genomic DNA extracted and the GSTM1 and GSTT1 genotypes were determined using Polymerase Chain Reaction (PCR) and restriction enzyme digestion at the Cancer Research Initiatives Foundation (CARIF) laboratory. The data was transformed into binary input variable as to enhance the efficiency of the systems thus contributes to an improved ability in pattern recognition. Table 1 summarizes the descriptions of variables used in the models.

Demographic and disease variables of patients that were reported to be associated risk factors to oral cancer were used as the predictor variables in developing fuzzy logic, fuzzy regression and fuzzy neural network prediction models. The full dataset were split randomly into a modelling dataset (65% of the total) and testing dataset (the remaining 35%).The dichotomous output refers to the health state of either "cancer" (1) or "healthy" (0). In developing the three fuzzy models, we have used 'supervised' machine learning techniques whereby the systems were developed based on algorithms that generate functions which map inputs to desired outputs. Fuzzy logic models used in this study were constructed based on fuzzy rules compiled from a group of 27 oral cancer clinicians' responses. Fuzzy neural network models used were developed in a MATLAB 7.0 environment using the ANFIS system. Fuzzy regression models used were the improvised version of Tanaka's possibilistic regression. Similar oral cancer data were fed into the three fuzzy models and their prediction performances in terms of sensitivity, specificity and percent accuracy were measured and compared.

The number of input features and the sample size of data used in this case study may be considered insufficient by some researchers in the area of artificial classification modeling. However, this is inevitable in this particular case study. Large medical databases are simply non-existence in Malaysia since medical informatics research in this country is very new, thus medical databases are seldom found. To overcome this limitation, we have applied a re-sampling technique known as split sample cross validation in our experiments (Mohd Dom 2009). Cross-validation was used in this study by building models from a training set and its predictive accuracy was then measured by applying a test set to the model. This process was then repeated by a number of times and the average performance was quoted as the performance measure of the algorithm used to build the model from the given data set. The advantage of split sample cross validation technique is that all the examples in the data set are eventually used for both training and testing. The goal of cross-validation is to verify that the result of a particular experiment is replicable and that it is not by random chances alone.

TABLE 1 Input variables' descriptions and categorizations

Variable	Variable descriptions	Binary Classifications

GSTM1 (Gm)	A gene mapped to chromosome 1p13.3. Polymorphism of GSTM1 results in either production of an enzyme known to have a role in Phase II detoxification of polycyclic aromatic hydrocarbons found in tobacco smoke or no production of enzyme (deletion polymorphism).	Negative=0
		Positive=1

GSTT1 (Gt)	A gene mapped to chromosome 22q11.2. Polymorphism of GSTT1 results either in production of an enzyme known to activate ethylene oxide, epoxybutanes, halomethanes, and methyle bromide or no production of enzyme (deletion polymorphism). GSTT1 is also involved in Phase II detoxification of polyaromatic hydrocarbons (PAHs) found in tobacco smoke.	Negative=0
		Positive=1

Smoke (S)	Cigarette smoking risk habit	No=0
		Yes=1

Drink (D)	Alcohol drinking risk habit	No=0
		Yes=1

Chew (C)	Betel-quid and tobacco chewing risk habit	No=0
		Yes=1

Age (A)	Age of patient as in years	> 40 years : 0 < 40 years : 1

Ethnicity (E)	Ethnicity of patient as belongs to the aborigine or the non-aborigine group. Aborigines refer to the people of the soil while the non-aborigine group refers to the Malay, Chinese and Indian races living in Malaysia.	Non-aborigines=0
		Aborigines=1

Gender (G)	Sex of patient as in male or female	Female=0
		Male=1

RESULTS AND DISCUSSION

Before any modelling works were done on the oral cancer dataset, a simple cross-tabulation and risk analysis was carried out using SPSS to detect any possible patterns or associations between the variables. Results of the analysis are summarised in Table 2.

TABLE 2 Descriptive analysis of input variables

Variable	Odds Ratio	Variable description	Healthy	Cancer	Sub Total	Total
GSTM1	1.068	Negative=0	47 (51.6%)	44 (48.4%)	91	171
		Positive=1	40 (50%)	40 (50%)	80	
GSTT1	1.501	Negative=0	35 (57.4%)	26 (42.6%)	61	171
		Positive=1	52 (47.3%)	58 (52.7%)	110	
Smoke	0.581	No=0	54 (46.6%)	62 (53.4%)	116	171
		Yes=1	33 (60%)	22 (40%)	55	
Drink	8.2	No=0	82 (59.4%)	56 (40.6%)	138	171
		Yes=1	5 (15.2%)	28 (84.8%)	33	
Chew	9.3	No=0	77 (67%)	38 (33%)	115	171
		Yes=1	10 (17.9%)	46 (82.1%)	56	
Age	0.12	> 40 years : 0	53 (40.5%)	78 (59.5%)	131	171
		< 40 years : 1	34 (85%)	6 (15%)	40	
Ethnicity	0.171	Non-aborigines=0	24 (29.3%)	58 (70.7%)	82	171
		Aborigines=1	63 (70.8%)	26 (29.2%)	89	
Gender	0.659	Female=0	45 (46.4%)	52 (53.6%)	97	171
		Male=1	42 (56.8%)	32 (43.2%)	74	

The cross tabulation and risk analysis results summarised in Table 2 in percentages form suggest that oral cancer incidence as reflected by the sample collected is higher among alcohol drinker (84.8%) and betel quid chewer (82.1%) but not among smokers (40%). This finding is consistent with what is reported in the literature except for the smoking habit (Zain 2001). The difference could possibly be due to the sample bias towards betel quid chewers and alcohol drinkers. The analysis also shows that higher oral cancer incidence is also found among the non-aborigines ethnic group of the country (70.7%) and in the older age group (59.5%). These findings support the age-related factor for oral cancer susceptibility. However, the report on the association between ethnic group and oral cancer susceptibility in this country is not yet conclusive since there were conflicting reports. Our cross tabulation results show that the aborigine ethnic group is not more susceptible to oral cancer as compared to the non-aborigines of the country. Finally, only slight differences is found between the male and the female gender in oral cancer susceptibility with the females having slightly stronger tendency (53.6%) for developing oral cancer as compared to the male counterpart (43.2%). This finding is consistent with the literature especially among the Indians in this country (Zain 2001). Similarly, not much difference is found in oral cancer incidence between the positive and negative genetic markers of GSTM1. However, GSTT1 positive polymorphism indicates higher association with oral cancer. There is not much being reported regarding this matter in this country thus no further comparison could be made at this point. Generally, the desired features in an intelligent prediction model for dental diagnostic task include good performance in terms of discrimination and calibration, ability to deal with missing data and noisy data (error in data), transparency of diagnostic knowledge and the ability to explain the decision made (Kononenko 2000).

COMPARISON OF MODELS' PREDICTION PERFORMANCES

Comparisons of models' prediction performances were carried out based on two categories of (1) 1-input and 2-input variable sets, and (2) 3-input and 4-input variable sets. The objective of comparing the models prediction performances using these categories was to investigate the differences in models' prediction abilities based on the number of input variables used for the models thus revealing the impact of the 'curse of dimensionality' on the three fuzzy models. Common measures of discrimination are sensitivity, specificity and percent accuracy (Dreiseitl & Ohno-Machado 2002). The area under the Receiver-Operating-Characteristics Curve (ROC) is normally used to depict the graphical representation of discrimination. The ROC was originally used for signal detection during the Second World War before it was used in medical diagnostic and prognostic tests. The ROC is used to determine the accuracy of predicted values and can be used across different classification tools (Abd-Kareem 2002; Speight et al. 1995).

The plot of an ROC curve shows the false positive rate on the *x*-axis and the 1 minus the false negative rate on the *y*-axis. It is normally termed as the sensitivity versus one minus specificity. A good diagnostic test is one that has small false positive and false negative rates across a reasonable range of cut off values. A bad diagnostic test is one where the only cut offs that make the false positive rate low have a high false negative rate and vice versa. The larger the area, the better the diagnostic test is. An ideal test will have an Area Under the receiver operating characteristic Curves (AUC) of 1 because it achieves both 100% sensitivity and 100% specificity (Lasko et al. 2005). Comparisons of the models' prediction performances were made based on the AUC by grouping the single-input with two-input predictor set in one group, and the three-input with four-input predictor set in another as shown in Table 3.

TABLE 3 Models' prediction performances measured by the area under the ROC curves (AUC)

Model	Variable*	AUC Fuzzy Logic (FL)	AUC Fuzzy Regression (FuReA)	AUC Fuzzy Neural Network (FNN)
1- input predictor	D	0.634	0.634	0.634
	A	0.684	0.684	0.684
	C	0.714	0.713	0.713
	S	0.440	0.455	0.456
	G	0.50	0.452	0.456
2-input predictor	CS	0.617	0.672	0.675
	CD	0.766	0.766	0.767
	CA	0.714	0.782	0.782
Average AUC		0.634	0.645	0.646
3-input predictor	CDS	0.636	0.810	0.785
	CDA	0.766	0.826	0.816
4-input predictor	ADCG	0.651	0.737	0.785
	ADCS	0.472	0.824	0.828
Average AUC		0.631	0.799	0.803

*Note: See Table 1 for details

Two statistical measurements were employed in the comparison procedure namely the Analysis of Variance (ANOVA) for the 1-input and 2-input variable sets and the Non-Parametric Mann-Whitney U test for the 3-input and 4-input variable sets. The ANOVA One-Way Between-Groups test was run by taking the AUC values as the "dependent variable" and the five prediction models as the "factor". In the comparison carried out in this study, the ANOVA test was appropriately used for comparing the differences in the areas under the receiver operating characteristic curve (AUC) values associated with the 1-input and 2-input variable sets since the normality assumption is satisfied as reflected by the *p*-values greater than 0.05 for both the Kolmogorov-Smirnov and Shapiro-Wilk tests results shown in Table 4.

However, normality assumption is not satisfied for 3-input and 4-input variable sets as reflected by the skewness statistic values shown in Table 5.

Hence, in such cases where normality assumptions are not met, the Non-Parametric Mann-Whitney U test was employed to check on the significant differences among the models. Mann-Whitney U test is the non-parametric alternative to the *t*-test for independent sample, commonly used for comparing the mean value for some variable of interest between two samples. The ANOVA and Mann-Whitney U test result summaries are given in Table 6.

A *p*-value of greater than 0.05 implies that the Null hypothesis H_0 is true and therefore H_0 is accepted. Accepting H_0 in this case implies that the means of the AUC values for the three models are significantly similar and vice-versa. Thus the prediction performance of fuzzy regression model is significantly the same as the prediction performances of fuzzy neural network and fuzzy logic models for single-input and 2-input variable sets. However, for 3-input and 4-input variable sets, fuzzy regression and fuzzy neural network models are found to have compatible prediction performance to one another which is significantly different from the prediction performance of fuzzy logic prediction models. In conclusion table 6 shows that fuzzy neural network and fuzzy regression models have superior prediction abilities compared to fuzzy logic prediction model.

TABLE 4 Tests of Normality by Kolmogorov-Smirnov(a) and Shapiro-Wilk tests for 1-input and 2-input variable sets

Model	Kolmogorov-Smirnov	Shapiro-Wilk
Fuzzy Logic (FL)	0.191	0.915
Fuzzy regression Adapted (FuReA)	0.216	0.859
Fuzzy neural network (FNN)	0.216	0.856

TABLE 5 Skewness statistics values for normality check for 3-input and 4-input -variable sets

Model	Skewness	Kurtosis
Fuzzy Logic (FL)	-0.596	1.625
Fuzzy regression Adapted (FuReA)	-1.837	3.379
Fuzzy neural network (FNN)	0.253	-4.557

TABLE 6. ANOVA test results for single-input and 2-input variable AND Mann-Whitney U test results for 3-input and 4-input variable of fuzzy regression (FuReA), fuzzy logic (FL) and fuzzy neural network (FNN) prediction models

Pair		*p*-value	Implication on comparison of Prediction performance
1-input and 2-input variable sets	FL & FNN	1.0	No Significant Difference
	FL & FuReA	1.0	No Significant Difference
	FuReA & FNN	1.0	No Significant Difference
3-input and 4-input variable sets	FL & FuReA	0.043	Significant Difference
	FL & FNN	0.020	Significant Difference
	FuReA & FNN	1.0	No Significant Difference

TABLE 7 Advantages and drawbacks of fuzzy prediction models understudied

Model	Advantages	Drawbacks
Fuzzy Regression Adapted model (FuReA)	1. Easy to construct	1. Sensitive to outliers.
	2.High calibration and discrimination power	
	3. Ability to explain relationship between response and input variables	
	4. Does not require big sample size	
	5. Suitable for variables governed by vague ambiguous relationship	
Fuzzy Logic Prediction model (FL)	1. Easy to construct	1. Number of fuzzy rules could be too big to handle "Curse of Dimensionality" problem
	2.Handles ambiguity relation between variables well	2. Unsuitable for large number of input variables
	3. Easy to understand	3. Difficulties in constructing fuzzy rules. Dependent on experts intervention in rules formation
Fuzzy neural network prediction model (FNN)	1. Handles ambiguity relation between variables well.	1. Inadequate ability in explaining relationship between response and input variables
	2. High calibration and discrimination power	

Several other differences were discovered between the three fuzzy prediction models when the oral cancer susceptibility experiments were conducted on them. Table 7 summarizes the findings in terms of the advantages and drawbacks associated with the different fuzzy prediction models (Mohd-Dom 2009).

CONCLUSION

In this paper we have demonstrated the feasibility of using fuzzy models to predict oral cancer susceptibility in a Malaysian sample based on the critical success factor which includes four input predictors age, alcohol drinking habit, tobacco chewing habit and cigarette smoking. Each of the fuzzy models constructed and tested in this case study has its own advantages and limitations as described in the earlier sections. The advantages and drawbacks of the computer models found in this study are consistent with what is reported in related literature (Abd-Kareem 2002; Kononenko 2000; Nasrabadi & Nasrabadi 2004; Savic & Pedrycz 1991). The use of computational intelligence in this study provides alternative initiatives in oral cancer screening (Kononenko 2000; Speight et al. 1995).

Fuzzy linear regression prediction model is found to be the superior prediction model in this study. On top of having high prediction accuracy, fuzzy linear regression prediction model produces transparent prediction equation and works well with small sample size. Thus, the quantification of the relationship between the input predictor variable and the predicted outcome is made possible by the fuzzy linear regression prediction model. This valuable information may assist clinician in providing better guidelines for oral cancer patients. Though fuzzy logic prediction accuracies found to be the lowest, its performance can be enhanced by reviewing the fuzzy rules supplied by the oral cancer clinicians since fuzzy logic modelling is highly dependent on expert input (Castellano & Fanelli 2005).

The development of these computer-based oral cancer prediction models is a step forward towards a more effective screening of oral cancer in the country. The findings on this research work will lead to the identification of the more suitable fuzzy models to be used for screening and educational purposes. The information revealed by the prediction equations may serve as a guide for future fuzzy prediction of oral cancer. It may also provide foundation for setting up of a computer-based fuzzy prediction tool to be incorporated in oral cancer screening program**.** Currently OCRCC is gathering other important parameters such as information on patients' dietary intake and their CYP1A genetic marker. This information could be incorporated into the existing fuzzy prediction models. The outcome is expected to be a more accurate computer prediction tool since more information is being fed into the prediction engine.

ACKNOWLEDGEMENT

The authors would like to thank the staff of Oral Cancer Research and Coordinating Centre (OCRCC), University of Malaya and Ministry of Health Malaysia for the contribution of sociodemographic data; and the Cancer Research Initiative Foundation (CARIF), for the contribution of the GSTM1 and GSTT1 data.

References

Abd-Kareem, S. 2002. Application of artificial neural network for the prognosis of nasopharyngeal carcinoma. Computer Science Department. University of Malaya, Kuala Lumpur, PhD thesis (unpublished).

Castellano, G. & C., Fanelli, A. 2005. Knowledge discovery by a neuro-fuzzy modeling framework. *Fuzzy Sets and Systems* 149: 187–207.

Coppin, B. 2004. *Artificial Intelligence Illuminated*. Sudbury, Massachusetts: Jones & Barllett Publishers.

Dreiseitl, S. & Ohno-Machado, L. 2002. Logistic regression and artificial neural network classification models: a methodology review. *Jonal of Biomedical Informatics* 35: 352-359.

Gorzalczany, M & Piasta Z. 1999. Neuro-fuzzy approach versus rough-set inspired methodology for Intelligent decision support. *Information Sciences* 120(1): 45-48.

Jang, R. 1993. ANFIS: Adaptive- Network-Based Fuzzy Inference System. *IEEE Trans. On Systems, Man and Cybernetics* 23(3): 665-685.

Kononenko, I. 2000. Machine learning for medical diagnosis: History, state of the art and perspective. *Artificial Intelligence in Medicine* 23(1): 89-109.

Lasko, T., Bhagwat, J., Zou, K., Machado, L. 2005. The use of receiver operating characteristic curves in biomedical informatics. *Jounal of Biomedical Informatics* 38: 404-415.

Mohan, J. & Yaacob, R. 2004. The Malaysian Telehealth Flagship Application: A national approach to health data protection and utilisation and consumer rights. *International Journal of Medical Informatics* 73: 217-227.

Mohd Dom, R. 2009. A fuzzy regression model for the prediction of oral cancer susceptibility.Computer Science Department, University of Malaya, Kuala Lumpur, PhD Thesis (unpublished).

Moraga, C. 2000. Neuro-fuzzy modeling with standard feedforward neural networks. *European Symposium on Intelligent Techniques Proceedings*. Aachen, Germany.

Muzio, L., D'Angelo, M., Procaccini, M., Bambini, F., Calvino, F., Florena, AM., Franco, V., Giovanelli, L., Ammatuna, P. And Campisi, G. 2005. Expression of cell cycle markers and human papilloma virus infection in oral squamous cell carcinoma: Use of fuzzy neural networks. *Int. J. Cancer* 115: 717- 723.

Nasrabadi, M. & Nasrabadi, E. 2004. A mathematical-programming approach to fuzzy linear regression analysis. *Applied Mathematics and Computation* 155: 873-881.

Savic, D. & Pedrycz, W. 1991. Evaluation of fuzzy linear regression models. *Fuzzy Sets and Systems* 39(1): 51-63.

Speight, P., Elliott, A., Jullien, J., Downer & M., Zakzrewska, J. 1995. The use of artificial intelligence to identify people at risk of oral cancer and precancer. *British Dental Journal* 179: 382-387.

Tanaka, H. & H. Lee 1998. Interval Regression Analysis by Quadratic Programming Approach. *IEEE Transactions on Fuzzy Systems* 6(4):

Wang, H. & Tsaur, R. 2000. Insight of a fuzzy regression model. *Fuzzy set and systems* 112: 355-369.

Zadeh, L. A, 1965. Fuzzy Sets. *Information and Control* 8: 338-353.

Zain, R. 2001. Cultural and dietary risk factors of oral cancer and pre-cancer – A brief overview. *Oral Oncology* 37:205-210.

Rosma Mohd Dom
Department of Mathematics
Faculty of Computer & Mathematical Sciences
Universiti Teknologi MARA
40450 Shah Alam, Selangor D.E.
Malaysia

Basir Abidin
Director
Foundation in Science
Cyberjaya University College of Medical Sciences
Unit No 2, Street Mall 2
63000 Cyberjaya, Selangor D.E.
Malaysia

Sameem Abdul Kareem
Faculty of Computer Science & Information Teknology
University of Malaya
50603 Kuala Lumpur
Malaysia

Siti Mazlipah Ismail
Oral Cancer Research & Coordinating Center
Faculty of Dentistry
Universiti of Malaya
50603 Kuala Lumpur
Malaysia

Norzaidi Mohd Daud*
Faculty of Business Management/Accounting Research
Institute/Institute of Business Excellence
Universiti Teknologi MARA
40450 Shah Alam, Selangor
Malaysia

*Corresponding author; email: zaidiuitm2000@yahoo.com

Received: 3 February 2010
Accepted: 18 November 2011