Identification and Validation of Predictive Factors for Glycemic Control: Neural Networks vs. Logistic Regression

CHUN-LIANG LAI[1], CHUNG-LIANG LAI[2],
SHOW-WEI CHIEN[3], KWOTING FANG[4]
[134]Department of Information Management,
National Yunlin University of Science & Technology
123, Section 3, University Road, Douliou, Yunlin 64002, TAIWAN, R.O.C.

[2]Department of Physical Medicine and Rehabilitation, Taichung Hospital,
Department of Health, Executive Yuan, TAIWAN, R.O.C.
199, Section 1, San Min Rd., Taichung 40343, TAIWAN, R.O.C.

Abstract: - From last decade, we are confronted with the rapid growth of diabetic patients who have become one of the most important burdens of public health. Accompanied with different complications, diabetes has considerable influences on the quality of individual living and the use of medical resources in the world in the 21st century. The purpose of this study is twofold. First, from the comparison standpoint logistic regression and neural networks were adopted to pursue the underlying characteristics of the glycemic control of the achieving target, or poor control level, so as to provide guidelines for physicians and diabetes educators. Second, for the cross validity purpose, 512 middle-aged patients, enrolled in Diabetes Healthcare Quality Improvement Program, were divided into training data and holdout data in a teaching hospital in Taiwan. Armed with the comparison, the finding revealed that neural networks is more accuracy than logistic regression. The important factors influence glycemic control are Years of diabetes onset, Education status, Body mass index, Months of enrolled in Diabetes Healthcare Quality Improvement Program, and Patient-Physician relationship.

Key-Words: - Neural networks, Logistic Regression, Diabetes.

1 Introduction

From last decade, we are confronted with the rapid growth of diabetic patients who have become one of the most important burdens of public health. Accompanied with different complications, diabetes has considerable influences on the quality of individual living and the use of medical resources in the world in the 21st century.

In either the developing countries or newly industrialized countries, diabetes incidence and prevalence are rapidly increasing. The prevalence of diabetes for all age-groups worldwide was estimated to be 2.8% in 2000 and 4.4% in 2030 [26]. The excess global mortality attributable to diabetes in the year 2000 was estimated to be 2.9 million deaths, equivalent to 5.2% of all deaths [22]. The complications of diabetes can be slow or even prevented by glycemic control in advance. In general, the diabetic patients were evaluated health care quality by using Hemoglobin A1C [3].

This study used logistic regression and back-propagation neural networks [23] in terms of A1C served as a decision attribute, to analyze 512 middle-aged (from 40 to 60 years old) patients in a teaching hospital in Taiwan. The discovered rules may assist the physicians and diabetes educators in precisely determining the behavior characteristics of patients, in order to provide guidelines to improve glycemic control for diabetic patients.

The remainder of this paper is organized as follows. The next section presents the background of diabetes and thoroughly reviews the previous research in diabetes. Section 3 describes the data and the development of prediction models, logistic regression and neural networks. Section 4 presents the classification results and comparison of two models. In section 5, the discussion of the research is given. Finally, Section 6 provides the conclusions.

2 Background

2.1 Diabetes

The vast majority of cases of diabetes fall into two broad categories [3]. In one category, type 1 diabetes, the cause is an absolute deficiency of insulin secretion. In the other, type 2 diabetes, the cause is a combination of resistance to insulin action and an inadequate compensatory insulin secretory response. In this paper, we focus on type 2 diabetic patients in middle-age (from 40 to 60 years old). Criteria for the diagnosis of diabetes in nonpregnant adults were shown as follows [3]: (1). Symptoms of diabetes and a casual plasma glucose $\geqq$ 200 mg/dl. OR (2). FPG (Fasting plasma glucose) $\geqq$ 126 mg/dl. Fasting is defined as no caloric intake for at least 8 h. OR (3). 2-h plasma glucose $\geqq$ 200 mg/dl during an OGTT (Oral Glucose Tolerance Test).

2.2 Previous research

Demographics such as age, sex, and ethnicity, affect the development of diabetes [10]. Obesity, physical inactivity, smoking, family history of diabetes have been described as risk factors for diabetes [1]. Socioeconomic status, such as education level, occupational status, and income, are implicated in the development of diabetes. Furthermore, the visit patterns of patients, given by physicians, also affected the glycemic control of diabetes.

The Diabetes Control and Complications Trial (DCCT) [12] in 1993, and United Kingdom Prospective Diabetes Study (UKPDS) [25] in 1998 indicated that the intensive control of blood sugar could reduce the risk of complications and diabetes-related death. Consequently, in essential, the complications of diabetes could be slow or prevented by better control on blood sugar. These researches proposed to use A1C as the criterion of glycemic control.

Quinlan [21] applied C4.5 on PIDD and it was 71.1% accurate. Michie et al. [19] applied CART and back-propagation algorithm on PIDD, and showed the accuracy of 74.5% for CART and 75.2% for back-propagation. Barriga et al. [4] used classification and regression tree (CART) [7] to screening for impaired glucose tolerance. The accuracy of simultaneous approach and serial approach were 61.9% and 51.5%, respectively. The important factors are age, BMI, Fasting glucose, and glycohemoglobin. Breault et al. [6] applied CART algorithm to classify the glycemic control, and reported the accuracy was 59.5%. The important attributes are age, Number of office visits in the given time period, Number of major complications, and lipid disorder.

3 Method

3.1 Data Source

This study used the clinical database of a teaching hospital in central Taiwan, where there are 110 physicians, six hundred hospital beds, 60 thousand outpatient services, and ten thousand inpatient services annually. The diabetic patients are collected from one-year outpatient and inpatient services and enrolled in Diabetes Healthcare Quality Improvement Program (DHQIP) [8] from Jan. 1, 2005 to Dec. 31, 2005.

3.2 Data preparation

Armed with the above-mentioned points, the data has collected includes: medical record code, name, date of birth, gender, address, postcode, medical department to visit, date of clinic visit, prescription, date of enrolled in DHQIP, year of diabetes onset, possess blood glucose meter, body mass index (BMI), education status, tobacco and alcohol use, exercise, family history of diabetes, ICD-9 code of diagnosis, and A1C test data. We transformed the data into suitable format for data mining. Date of birth is transformed into age, based on the date Jan.1, 2005. Address and post codes were transformed into the area of residence, grouped into residing in the metropolitan area and sub-metropolitan area according to the region of Administration.

We calculated the times of visit from the clinic visit time records, and the medical department is selected from the highest ratio among the clinic visits. Each time when there are diabetic drugs in the prescription, or insulin injection, during the clinic visit, then that visit is determined as clinic visit for diabetes. The patient-physician relationship (PPR) is determined by the proportion of visiting the same physician. We defined PPR as follow, if PPR$\geqq$70%, patient-physician relationship is stable; if PPR＜70%, patient-physician relationship is unstable.

We classified BMI into two classes according to the standard weight status categories of CDC [9]: overweight and obese, 25.0 kg/m^2 and above; normal and underweight, below 25.0 kg/m^2. The complications, have divided into 6 categories are judged from ICD-9 diagnosis codes which are Diabetes with Acute complication, Diabetes with Renal complication, Diabetes with Ophthalmic complication, Diabetes with Neurological complication, Diabetes with Vascular complication, and Diabetes with Foot complication. Decision attribute A1C is classified into two levels: A1C＜ 9.0% was considered to have targeted glycemic control; A1C≧9.0% [20, 24] was considered to have poor glycemic control. If A1C has more than one data, we adopted the average value. The attribute values of the subjects are listed in Table 1.

The attributes for prediction are Gender, Times of diabetic clinic visit, Patient-Physician relationship, Medical department to visit, Area of residence, Months of enrolled in DHQIP , Possess blood glucose meter, Years of diabetes onset, BMI, Education status, Regular smoking everyday, Regular drinking everyday, Regular exercising everyday, Family history of diabetes, with or without complications, and the decision attribute is A1C.

Table 1 Attribute value of the subjects

Attribute	Attribute values
Gender	Male, Female
Times of diabetic clinic visit	Continuous
Patient-Physician relationship	Stable, Unstable, Only visit once
Medical department to visit	Family department, Internal department
Area of residence	Metropolitan area, Submetropolitan area
Months of enrolled in DHQIP	Continuous
Possess Blood Glucose Meter	Yes, No
Years of diabetes onset	Continuous
Body Mass Index	Obese and Overweight, Normal and Underweight
Education status	Primary school and below, High school, College and above
Regular smoking everyday	Yes, No
Regular drinking everyday	Yes, No
Regular exercising everyday	Yes, No
Family history of diabetes	Yes, No
With complication	Yes, No
Hemoglobin A1C	Target, Poor

3.3 Development of prediction models

The dataset was divided randomly into two sets, one set of 409 cases (80% of the whole dataset) for training [13] and 103 cases for testing the model.

3.3.1 Development of logistic regression model

The use of logistic regression modeling has explored during the past decade. Logistic regression is used primarily for predicting dichotomous dependent variables on the basis of continuous or categorical independent variables. The existence of multicollinearity can affect the parameters of a regression model. The common measures for assessing multiple variable collinearity are the *Tolerance* value and its inverse, the *Variance Inflation Factor (VIF)*. A common cutoff threshold is a *Tolerance* value below 0.10, which corresponds to a *VIF* value above 10 [15]. The more severe criteria for *Tolerance* is that below 0.40 are regarded as indicating multicollinearity [2]. The *Tolerance* value of our training set all above 0.40, which indicated there is no multicollinearity between the independent variables.

We used stepwise selection method for developing the model. The stepwise procedure for selection or deletion of variables from a model is based on a statistical algorithm that checks for the importance of variables. The *p*-value for adding variable that in the range from 0.15 to 0.20 is highly recommended [17]. We followed Hosmer and Lemeshow's suggestion that the significance threshold (*p*-value) we used for adding variable was 0.20. We examined the overall model using the chi-square goodness of fit. The *p*-value of the chi-square goodness of fit is not significant, which indicated the overall model is adequate fit. The built logistic regression model was tested using the holdout data. The training and holdout data were saved for further processing by decision tree and neural networks.

3.3.2 Development of neural networks model

The attractiveness of neural networks comes from the remarkable characteristics such as nonlinearity, high parallelism, robustness, fault and failure tolerance. Neural network are very flexible with respect to incomplete, missing and noisy data. Multi-layer perceptron (MLP) are feed-forward neural networks trained with the standard back-propagation algorithm. They are supervised networks so they require a desired response to be trained. They learn how to transform input data into a desired response, so they are widely used for pattern classification [11].

A high *learning rate* will accelerate training, this may cause the search to oscillate on the error surface and never converge. A small *learning rate* drives the search steadily in the direction of global minimum. Fu [14] recommends *learning rate* from 0.0 to 1.0. *Momentum* >1.0 yields excessive contributions of the weight increments of the previous step and may cause instability [16]. Conversely, an extremely small *momentum* leads to slow training. Fu [14] suggest *momentum* from 0.0 to 1.0. In most function approximation problems, one hidden layer is sufficient to approximate continuous functions [5]. We used feed forward back propagation neural networks in this study. The architecture of MLP consisted of three layers, an input layer, a hidden layer and an output layer. The input layer consisted of 20 input neurons, the hidden layer consisted of seven hidden nodes, and the output layer consisted of one output neuron. The initial *learning rate* and *momentum* for network training were set to 0.3 and 0.9, respectively.

3.4 Performance evaluation

In this study, we used accuracy, sensitivity and specificity as performance measures. The classification accuracy measures the proportion of correctly classified cases. Sensitivity measure the fraction of positive cases that are classified as positive. Specificity measure the fraction of negative cases that are classified as negative [18].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

$$Sensitivity = TP / TP + FN. \qquad (2)$$

$$Specificity = TN/TN + FP. \qquad (3)$$

where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives, respectively. The odds ratio is the ratio of the expected number of times that an event will occur to the expected number of times it will not occur, and is given by the equation [2]

$$OddRatio = \frac{p}{1-p} \qquad (4)$$

where p is the probability of an event.

4 Result

In this study, we used logistic regression and back-propagation neural networks for classification and feature selection. We applied Clementine to classify the dataset. The models were evaluated based on classification accuracy, sensitivity, specificity, and odds ratio.

The results obtained from holdout data are: the logistic regression model achieved a classification accuracy of 71.8 % with a sensitivity of 93.2 % and a specificity of 17.2 %. The neural networks model performed a classification accuracy of 75.7% with a sensitivity of 87.8% and a specificity of 44.8%. The odds ratio of logistic regression and neural networks are 2.88, and 5.87, respectively. Table 2 shows the complete set of results.

Table 2 Model performance

Model	Logistic Regression	Neural Networks
Accuracy %	71.8	75.7
Sensitivity %	93.2	87.8
Specificity %	17.2	44.8
Odds ratio	2.88	5.87

The equation of logistic regression contained six attributes: Years of diabetes onset, BMI, Education status, Medical department to visit, Months of enrolled in DHQIP, and Regular drinking everyday. From the equation we obtained the coefficient and odds ratio of each significant attributes which listed in Table 3. Our analysis of neural networks used Clementine to produce the relative importance of attributes. The top six attributes are Years of diabetes onset, Education status, Patient-Physician relationship, Months of enrolled in DHQIP, Regular smoking everyday, and Family history of diabetes. The results of important predictive factors produced from two models are listed in Table 4.

Table 3 Coefficient of Logistic regression Model

Attribute	β coefficient	Odds ratio (Inversed)
Years of diabetes onset (10 years)	-0.080*10	0.448 (2.232)
Education (primary school)	-1.061	0.346 (2.890)
Education (high school)	-0.702	0.496 (2.016)
BMI (Obese and Overweight)	0.429	1.536 (0.651)
Drinking (No)	-0.529	0.589 (1.698)
Months of enrolled in Diabetes Healthcare Quality Improvement Program (12 months)	0.015*12	1.197 (0.853)
Medical department to visit (family department)	0.4700	1.600 (0.625)

Table 4 Important Predictive Factors

Logistic Regression	Neural Networks
Years of diabetes onset	Years of diabetes onset
Body Mass Index	Education status
Education status	Patient-Physician relationship
Medical department to visit	Months of enrolled in DHQIP
Months of enrolled in DHQIP	Regular smoking everyday
Regular drinking everyday	Family history of diabetes

5 Discussion

As shown in the Table 4, the first important attribute of two models is Years of diabetes onset. From logistic regression equation, we found that the years of diabetes onset increasing every ten years, the glycemic control risk increasing 223%. The education status also played an important role in glycemic control. The risk of glycemic control with primary school to college status, and high school to college status is 289 % and 202 %, respectively. We revealed that education status also influence on glycemic control.

In order to control blood sugar effectively and slow or prevent the complications, Taiwan Government encourages them to enroll in DHQIP, expecting to achieve the above-mentioned purposes by the integrating healthcare of physicians, dietarians, and case managers. The patients in the program diagnosed and given treatment by the physician who were specialist, and taken care by case managers and dietarians. The time of enrolled in DHQIP is affected the glycemic control, this may verify the clinical condition.

Family department varies from the other department because of it is the gatekeeper of the medical treatment. The patients who visited family department, are most in the initial stage of diabetes, thus the ratio of target control is higher. This major point also conforms to the clinical phenomenon.

6 Conclusion

In this paper we used logistic regression and neural networks algorithm with Hemoglobin A1C as a decision attribute to classify the glycemic control status and to discover the behavior characteristics of patients. After going though a long process of data cleansing and transformation, we used it to develop the prediction models. The results indicated that the neural networks algorithm performed better with a classification accuracy of 75.7% which is better than any reported in the published literature using real-life data in diabetic domain, the logistic regression model came out with a classification accuracy of 71.8%. Medical databases may consist of a large volume of heterogeneous data. The heterogeneity of the data may be decreased the classification accuracy rate.

In addition to the prediction model, we also identified important factors to classification in order to gain insight into the independent factors to predict the glyermic control. From these two model we concluded that Years of diabetes onset, Education status, BMI, Time of enrolled in DHQIP, and Patient-Physician relationship are most important factors affected the glyermic control. Ideally, the value of this finding may assist the physicians and diabetes educators in precisely determining the behavior characteristics of patients, in order to provide guidelines to improve glycemic control in diabetic patients.

References:

[1] Agardh, E.E. and colleagues, Explanations of socioeconomic differences in excess risk of type 2 diabetes in Swedish men and women, *Diabetes Care*. Vol. 27, No.3, 2004, pp. 716-721.

[2] Allison, P. D., *Logistic regression using the SAS system: Theory and Application*, SAS Institute, Cary NC, 1999.

[3] American Diabetes Association, Clinical Practice Recommendations 2006, *Diabetes care*, Vol.29, Supplement 1, 2006.

[4] Barriga, K.J., Hamman, R.F., Hoag, S., Marshall, J.A., and Shetterly, S.M., Population screening for glucose intolerant subjects using decision tree analyses, *Diabetes Research and Clinical Practice*, Vol.34, Sup.17, 1996, pp. S17-S29.

[5] Basheer, I.A., Hajmeer, M., Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, Vol.43, No.1, 2000, pp. 3-31.

[6] Breault, J.L., Goodall, C.R., Fos, P.J., Data mining a diabetic data warehouse, *Artif Intell in Med*, Vol.26, No.1, 2002, pp. 37-54.

[7] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., *Classification and regression trees*, Monterey, CA: Wadsworth & Brooks/ Cole Advanced Books & Software, 1984.

[8] Bureau of National Health Insurance, Department of Health, Executive Yuan, R.O.C., *Diabetes Healthcare Quality Improvement Program*, 2005.

Retrieved March 8, 2006, from http://www.nhi.gov.tw/webdata/AttachFiles/Attach_3078_2_w0950059032-a1.pdf

[9] Centers for Disease Control and Prevention (CDC), U.S., *BMI — Body Mass Index: About BMI for Adults*. Retrieved Oct 5, 2006, from http://www.cdc.gov/nccdphp/dnpa/bmi/ adult_BMI/about_adult_BMI.htm

[10] Congdon, P., Estimating diabetes prevalence by small area in England, *J. Public Health Med*, Vol. 28, No.1, 2006, pp. 71-81.

[11] Delen, D., Walker, G., Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods, *Artificial Intelligence in Medicine*, Vol.34, No.2, 2005, pp. 113-127.

[12] Diabetes Control and Complications Trial research group, The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus, *New England Journal of Medicine*, Vol.329, No.14, 1993, pp. 977-986.

[13] Eklund, P. W. and Hoang, A., Classifier Selection and Training Set Features: LMDT, 1998, Retrieved March 8, 2006, from citeseer.nj.nec.com/309003.html.

[14] Fu, L., *Neural Networks in Computer Intelligence*, McGraw-Hill, New York, 1995.

[15] Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C., *Multivariate Data Analysis, 5th edn*., Prentice Hall, Upper Saddle River, NJ, 1998.

[16] Henseler, J., Backpropagation. In: Braspenning, P.J. et al. (eds.), *Artificial Neural Networks: An Introduction to ANN Theory and Practice*, Lecture Notes in Computer Science, Springer, NY, 1995, pp. 37–66.

[17] Hosmer, D. and Lemeshow, S., *Applied Logistic Regression. 2nd edn*. John Wiley and Sons, New York, 2000.

[18] Lavrac, N., Selected techniques for data mining in medicine, *Artificial Intelligence in Medicine*, Vol.16, No.1 , 1999, pp. 3-23.

[19] Michie, D., Spiegelhalter, D.J., et al., *Machine learning, neural and statistical classification*, Ellis Horwood, New York, 1994.

[20] Pollock-BarZiv, S. M. and Caroline Davis, C., Personality Factors and Disordered Eating in Young Women with Type 1 Diabetes Mellitus, *Psychosomatics*, Vol.46, No.1 , 2005, pp. 11-18.

[21] Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Diego, 1993.

[22] Roglic, G., Unwin, N., Bennett, P.H., Mathers, C., Tuomilehto, J., Nag, S., Connolly, V., and King, H., The Burden of Mortality Attributable to Diabetes: Realistic estimates for the year 2000, *Diabetes Care*, Vol.28, No.9 , 2005, pp. 2130-2135.

[23] Rumelhart, D., Hinton, G. and Williams, R., Learning internal representations by error prop agation. In : Anderson, J. and Rosenfeld, E. (eds.): *Neurocomputing*, MIT Press, Cambridge, MA, 1988, pp. 675-695.

[24] Rydall, A.C., Rodin, G.M., Olmsted, M.P., Devenyi, R.G., Daneman, D., Disordered eating behaviour and microvascular complications in young women with insulin-dependent diabetes mellitus, *New England Journal of Medicine*, Vol.336, No.26 , 1997, pp. 1849–1853.

[25] UK Prospective Diabetes Study (UKPDS) Group, Intensive blood-sugar control with sulphonylureas or insulin compared with conventions1 treatment in patients with type 2 diabetes, *Lancet*, Vol.352, 1998, pp. 837-853.

[26] Wild, S., Roglic, G., Green, A., Sicree, R. and King, H., Global Prevalence of Diabetes: Estimates for the year 2000 and projections for 2030, *Diabetes Care*, Vol.27, No.5, 2004, pp. 1047-1053.