Model Development for Prediction of Diabetic Retinopathy

Completed Research Paper

Mohsen Sangi
University of Wollongong
ms819@uowmail.edu.au

Khin Than Win
University of Wollongong
win@uow.edu.au

Farid Shirvani
University of Wollongong
farids@uow.edu.au

John Fulcher
University of Wollongong
john@uow.edu.au

ABSTRACT

This research focuses on presenting an empirical method to gather necessary data and then developing several models to predict the chance of diabetic retinopathy (proliferative and non-proliferative) by observing HbA1c, duration of disease and albumin excretion rate of diabetic patients. We gathered required knowledge from other studies that have investigated the relation of different risk factors and complications in diabetes. In order to create 1-1 models, curve fitting was performed by using two different software applications: Tiberius (Brierley 2011) and SPSS (IBM 2010), which work based on ANN and least square regression, respectively. To start producing the model, seven different patterns, i.e. linear, logarithmic, quadratic, cubic, power, s and exponential, have been chosen as the best regression options. Using R-squared, it can be clearly seen that the best selected regression models fit the data in all the dataset tables better than ANN, as well as the other six regression patterns.

Keywords

Diabetes, decision model, HbA1c, diabetic retinopathy, ANN

INTRODUCTION

Early and appropriate intervention in diabetes is able to reduce the rate of complications, to prolong life expectancy and to reduce the financial cost (Habacher, Rakovac, Gorzer, Haas, Gfrerer, Wach and Pieber, 2007) . Since there is no simple relation between risk factors and their associated complications, predictive modeling can help to make models for forecasting future disease behavior based on current observations. In this research we developed a set of models by using the results of the previous studies and applying a novel combination of different techniques. The implementation of the published Guidelines for Diabetes Care recommends optimal management and the early detection of existing complications to reduce further disease progression (European Diabetes Policy 1999). Many aspects of diabetes complications can be limited or even prevented in some instances, with appropriate management of the condition (Huang, McCullagh, Black and Harper, 2007). This benefits the patients as well as having the potential to reduce the overall healthcare expenditures. Diabetic retinopathy is a complication of diabetes that damages the blood vessels in the eye. There are two main types of diabetic retinopathy: non-proliferative and proliferative. Early detection and proper treatment of diabetes can prevent up to 90% of blindness, (Boulton and Vileikyte, 2000). These complications can be prevented only if the treatment starts before the development of non-reversible clinical symptoms. This means that being able to predict such complications is essential in order to intervene successfully.

There is a very intricate relationship between diabetic risk factors and their related complications. In order to overcome the complexities of this interrelated network, we need to use modeling techniques to simulate the relationships and to break down the ultimate model into three simpler steps including modeling the relation between one risk factor and one complication (step 1), modeling the relation between all risk factors related to one complication (step 2) and then modeling the relation between all risk factors and all complications (step 3).

In this research we developed a set of models by using observations recorded between HbA1c, duration of disease and albumin exertion rate (AER) as selected risk factors and PDR and NPDR as two complications .

The contribution of this research is:

- The development of a method to gather and standardize the knowledge from other research and make a structured format for it.
- The application of two important data mining methods, namely ANN and regression analysis, in order to create a predictive model to indicate the relationship between individual risk factors and complications.
- Measurement of the prediction power of the created models using R-squared, F-ratio and adjusted R-squared equations.

METHODOLOGY

Diabetes and its complications were gathered from peer reviewed publications on clinical trials, meta-analysis and the Cochrane review from Medline, Cinahl, federated database searches and diabetes management guidelines published from 1981 to 2012. All titles of result articles were examined by the research team independently and rated each paper as 'potentially relevant', 'of doubtful relevance' or 'not relevant'. Then the process was repeated for the first two groups by reviewing the abstracts and then the full-text versions and all 'not relevant' studies were omitted from this research. Afterwards these were coded in Nvivo8. In brief, both chronic diabetes complications and diabetes predisposing factors have been classified hierarchically by 27 and 47 nodes, respectively. The two are connected by 126 relationships, categorized into three different types, namely 'is a cause for', 'prevent from or decrease' and 'are the same'. All these nodes and relationships are supported by more than 590 identified facts from 99 different sources. Having scanned the abstract of nearly 550 primitive results, about 140 of the most probably related subjects have been chosen, thus enabling us to use and follow up the data of more than 300,000 patient/years (Sangi, Win and Fulcher, 2010).

Data Conversion was conducted to identify a relationship between one specific risk factor and one complication. Data expressions were paraphrased into a 'from-to' table record format to indicate that the research has independently studied the relation between factors and complications. In other words, these tables are designed in a three-dimensional format in order to make data as clear as possible, but later on would be used as two-dimensional records. This format shows by how much the probability of a specific complication will change, if, independent of all other factors, the value of a specific risk factor changes.

Some studies express their observations exactly in our desired format. For instance, according to the DCCT research group (DCCT 1996), 'A reduction in HbA1c from 11 to 9.9% yields a reduction in risk of retinopathy progression from 10.78 to 4.21 cases per 100 patient-years. In contrast, a reduction from 8 to 7.2% yields a reduction in risk from 2.43 to 1.48 cases per 100 patient-years'. Each of these rates per 100 patient-years is in fact the percentage of retinopathy risk in one year. Although some research studies have independently examined the relationship between each risk factor and complication, in order to be comparable, the studies must have examined data from patients suffering from the same duration of the disease to be able to compare them. So, in this study we calculate the magnitude of absolute risks for a 5-year time period. According to Bulmer (Bulmer 1979) and Johnson (Johnson and Bhattacharyya, 2010), the probability of risk during 'n' years can be estimated by addition law since the probability of the risk in one year is known:

$$P(risk_n) \cong (P(risk_1) \times n) - \left(\binom{n}{2} \times \left(P(risk_1) \right)^2 \right) + \left(\binom{n}{3} \times \left(P(risk_1) \right)^3 \right) - \cdots \left(\binom{n}{n} \times \left(P(risk_1) \right)^n \right)$$

where $P(risk_1)$ and $P(risk_n)$ are the probability of risk in the period of one and n years, respectively. Thus, the above sentence can be converted into the desired record format as shown in Table 1.

HbA1c			Retinopathy	
From	To		From	To
11%	9.9%		43.5%	19.3%
8%	7.2%		11.6%	7.1%

Table 1. Risk of retinopathy is changing based on changes in HbA1c levels

Most of the studies, however, express their results in other formats .They report only the factor of change in severity occurring in a complication, but the absolute amount of probability for the complication is not known. As the DCCT research group (DCCT 2000) writes: 'as the HbA1c value increased from 8.1 to 10.1 percent, the odds of retinopathy increased by a factor of 2.8, but as the value increased from 6.1 to 8.0 percent, the odds were increased by a factor of 2.6'. In such cases, by having the risk percentage of the first point, the risk percentage of other points would be calculated easily as the factors of change are available. There are other studies which indicate the risk of complications in another format. For example, 'Patients whose intensive insulin therapy resulted in HbA1c levels 2% lower than those receiving conventional insulin therapy had a 76% lower incidence of retinopathy' (Brownlee, Aiello, Cooper, Vinik, Nesto and Boulton 2011). Similarly, as long as the absolute risk of retinopathy in conventional therapy is available, the risk percentage for the intensive therapy group would be calculated. In this case, the mean amounts of HbA1c in conventional and intensive therapy are nearly 7.2% and 9.1%, respectively. Finally, some studies demonstrate the changes in graphical representations, but the data from these needed to be extracted and presented in the format we needed. As mentioned above, the reason we chose the 'from-to' format was to take into consideration the fact that the research has independently studied the relation between factors and complications, but here we use them as two-dimensional records. Table 2, as an example, shows a dataset of the relation between HbA1c and non-proliferative diabetic retinopathy (NPDR).

HbA1c level	6.8	6.95	7.85	7.95	8	8.95	9.2	9.5	9.9	9.95	10.5	10.55	11.7	12
Risk of NPDR	14.5	3	20	3.8	14	7.1	27	20	27	7.9	9.9	28	51	32

Table 2. A dataset showing the relation between HbA1c and non-proliferative diabetic retinopathy

Model Validation

Evaluate of Gathered Data

One of the disadvantages of secondary data is that the user has no control over their accuracy. Although timely and pertinent secondary data may fit the researcher's requirements, the data could be inaccurate. Research conducted by other persons may be biased to support the vested interest of the source. If the possibility of bias exists, the secondary data should not be used. If the accuracy of the data cannot be established, the researcher must determine whether using the data is worth the risk.

We accepted data from reliable studies such as DCCT (The Diabetes Control and Complications Trial and Follow-up Study) and UKPDS (The United Kingdom Prospective Diabetes Study). Nevertheless, we assessed the reputation of the journals or conferences which have published the data and finally we have done cross-checks of data, that is, the comparison of data from one source with data from another source to determine the similarity of independent projects.

Model Comparison and Testing

The accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen data. There are two important questions to ask: 1) how well does the model fit the observed data and 2) can the model generalize to other samples? These questions have been tested by R square, adjusted R square and F test.

Using the mean as a model, we can calculate the difference between the observed values and the values predicted by the mean. This sum of squared differences is known as the 'total sum of squares' (SS_T). This value represents how good the mean is as a model of the observed data. Now, if we fit the more sophisticated model to the data, such as a line of best fit, we can again work out the differences between this new model and the observed data. The result is known as 'residual sum of squares' (SS_R). The improvement in prediction resulting from using the predictive model rather than the mean is calculated by computing the difference between SS_T and SS_R. This difference shows us the reduction in the inaccuracy of the model resulting from fitting the predictive model to the data. This improvement is the 'model sum of squares' (SS_M). A useful measure arising from these sums of squares is the proportion of improvement due to the model. This is easily calculated by dividing the sum of squares for the model by the total sum of squares. The resulting value is called R^2:

SS_M (Model Sum of squares) = SS_T ((Total Sum Square) - SS_R (Residual Sum of Square)

$$R^2 = \frac{SS_M}{SS_T}$$

A second assessing the model is through the *F*-test. *F* is based upon the ratio of the improvement due to the model (SS_M) and the difference between the model and the observed data (SS_R). Actually, because the sums of squares depend on the number of differences that we have added up, we use the average sums of squares ('mean squares' or MS). To work out the mean sums of squares we divide by the degrees of freedom. For SS_M the degrees of freedom are simply the number of variables in the model, and for SS_R they are the number of observations minus the number of parameters being estimated. The result is the mean squares for the model (MS_M) and the residual mean squares (MS_R).

$$F = \frac{MS_M}{MS_R}$$

It is important to know *F*-ratio tells us how much variability the model can explain relative to how much it can't explain. A good model should have a large *F*-ratio (greater than 1 at least).

As a second part of this step, we have to assess how well our model can predict the outcome in a different sample. To test whether the model does generalize we can look at cross-validating it. Assessing the accuracy of a model across different samples is known as cross-validation. In this research, not only are the values of R^2 calculated, but also an adjusted R^2. This adjusted value indicates the loss of predictive power or 'shrinkage'. Whereas R^2 tells us how much of the variance in *Y* is accounted for by the predictive model from our sample, the adjusted value tells us how much variance in *Y* would be accounted for if the model had been derived from the population from which the sample was taken. SPSS derives the adjusted R^2 using Wherry's equation (Field 2009). This equation, however, has been criticized because it tells us nothing about how well the regression model would predict an entirely different set of data. One version of adjusted R^2 that does tell us how well the model cross-validates uses Stein's formula (Stevensand NetLibrary, 2002) which is:

$$adjusted\ R^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)\right](1 - R^2)$$

In Stein's equation, R^2 is the unadjusted value, *n* is the number of participants and *k* is the number of predictors in the model. In addition, there is an independent errors assumption in regression, which means for any two observations the residual terms should be uncorrelated (or independent). This eventuality is sometimes described as a lack of 'autocorrelation'. This assumption can be tested with the 'Durbin–Watson' test, which tests for serial correlations between errors. The test statistic can vary between 0 and 4 with a value of 2 meaning that the residuals are uncorrelated. As a rough rule, the results between 1.5 and 2.5 imply independent errors (Hutcheson and Sofroniou, 1999).

Model Structure for Prediction

ANN and Regression Analysis are two common techniques for making a predictive model based on observed data ((Razi and Athappilly, 2005; Smith and Mason, 1997).

ANNs can be trained to predict numerical values, and allows the systems to learn from past experiences and recognizes patterns in the gathered data. The ANN takes a dataset and tries to combine the 'inputs' (HbA1c value) in such a way to model the 'output' (Risk percentage of retinopathy). These models can then be used on new data to predict what the output is likely to be for a given set of inputs. The ANN is used as it works more efficiently when a large amounts of data is available (Griffith 2000; Lawrence 1994).

Statistical regression analysis is widely used to estimate different types of functions including linear, polynomial, exponential, and general nonlinear. By using this method a pattern of the function that is supposed to fit the data is needed. Depending on the estimation from function shape, the pattern may vary from a linear function like $y = ax + b$ to nonlinear functions such as $y = m \times e^{ax} + b$. This pattern helps the modelling process to draw a curve by estimating the value of the parameters. In the above example, 'm', 'a' and 'b' are parameters of the mentioned functions, while 'y' and 'x' are dependent and independent variables, respectively. For each observation record in the data file, this method executes the pattern and computes the value of the function. The computed value of the function is assigned to the variable namely Predicted. The predicted value is then subtracted from the actual value of the dependent variable and this difference is assigned to the Residual variable. This process is repeated for each observation in the data file and the squared residual values are added together. After all of the observations have been processed, the values of the parameters are adjusted and this process is being repeated to minimize the sum of the squared residuals (Sherrod 2010).

Both of these methods create models by using the observed data; the models then are able to get a specific value of HbA1c as input and predict the risk of retinopathy for that particular HbA1c Level.

Artificial neural networks (ANN) and regression analysis are used to create a prediction model based on observed data.

Artificial Neural Networks (ANNs)

Software application Tiberius (Brierley 2011) was used for making predictive models.

Regression Analysis

The objective of regression analysis is to determine the best model that can relate the output variable to various input variables. More formally, regression analysis is the process of determining how a variable Y is related to one or more other variables $X_1, \dots X_n$. Y is usually called the response output or dependent variable, and X_{i-n} are inputs, regressors, explanatory variables or independent variables. Several software packages exist to solve regression problems. Examples include SAS (www.sas.com), NLREG (www.nlreg.com), S-Plus (csan.insightful.com) and SPSS (IBM 2010). This last was the software used in this study.

RESULTS

Figure 1 shows the neural network made by Tiberius which attempts to reveal the relationship between risk factor values as input, and risk percentage of any diabetic complication as output, for all of the records.

The network tries to make a model that relates the input numbers to the output ones. This model is created in such a way as to produce the least amount of error in predicting an output for its respective input. An example of the model produced by the software between HbA1c and all kinds of retinopathy is illustrated in Figure 2-4.
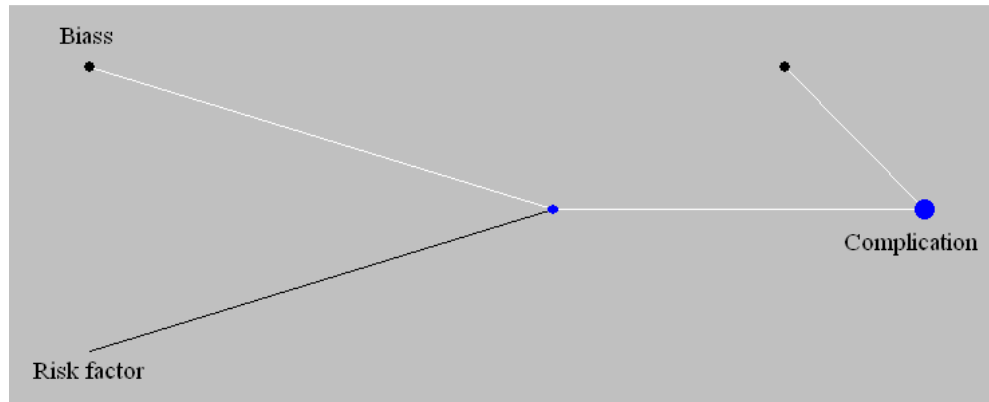

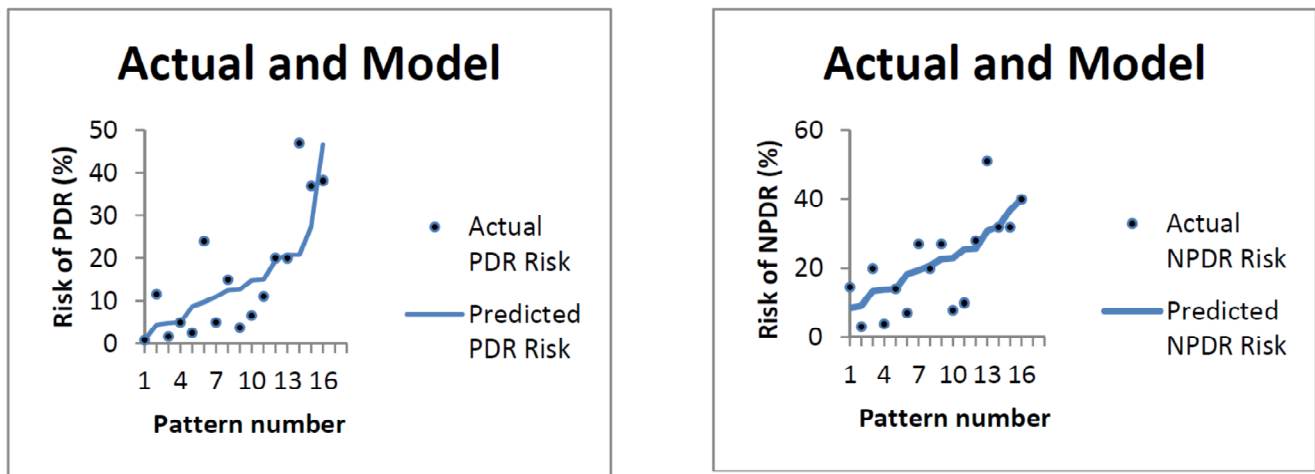**Figure 1. The neural network which created the model based on the observed data**

Figure 2: The model made by the neural network (HbA1c-PDR) and (HbA1c –NPDR)
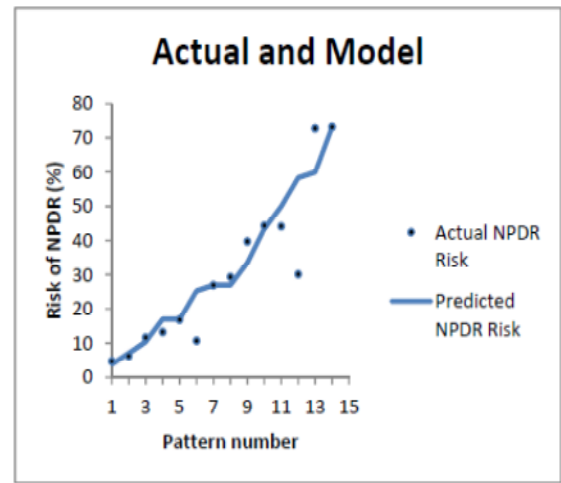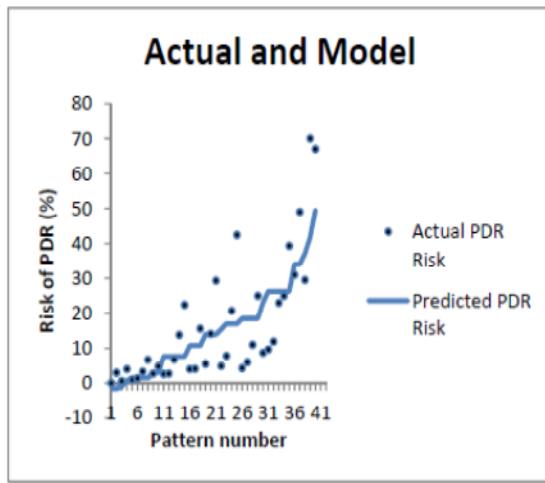
Figure 3: The model made by the neural network (duration-PDR) and (duration –NPDR)
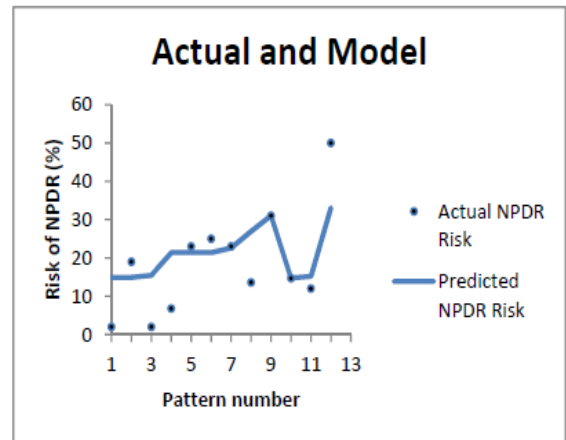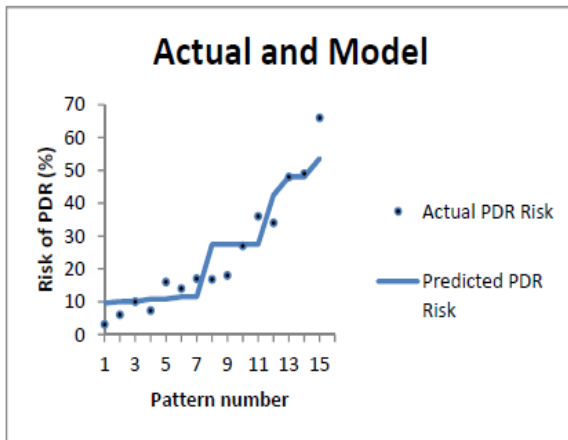
Figure 4: The model made by the neural network (AER-PDR) and (AER –NPDR)

Table 3 shows the statistical indices of ANN patterns for all dataset tables.

Related data table	Std. Error of the Estimate	R	R Square
HbA1c-PDR	9.288	.748	.560
HbA1c-NPDR	9.016	.735	.540
Dur-PDR	10.522	.792	.628
Dur-NPDR	9.509	.898	.807
AER-PDR	6.515	.931	.866
AER-NPDR	8.657	.741	.549

Table 3. The statistical indices of ANN patterns

Regression Analysis

We provide SPSS with the same dataset. Seven different patterns, i.e. linear, logarithmic, quadratic, cubic, power, s and exponential, have been chosen as the best options.

Then, this application uses the process mentioned in the methodology to determine the value of the parameters. Figure 5, as an example, shows the scatter graphs along with plots of the selected best fitted function to the set of data values among all seven patterns.

SPSS calculates the values of parameters for all mentioned patterns to achieve the functions. Table 4-9 show the statistical indices of these seven patterns for all dataset tables. The highlighted rows in each table indicate the selected best fitted function to the set of data values among all seven patterns.

Models	R	R Square	Wherry's Adjusted R Square	Std. Error of the Estimate	F-ratio	
					F	Sig
Linear	.755	.570	.539	9.815	18.537	.001
Logarithmic	.768	.590	.561	9.583	20.136	.001
Quadratic	.776	.602	.541	9.794	9.840	.002
Cubic	.780	.608	.548	9.720	10.089	.002
Power	.764	.584	.555	.799	19.689	.001
S	.793	.628	.602	.756	23.678	< .001
Exponential	.713	.509	.474	.869	14.494	.002

Table 4. Statistical indices of the selected seven patterns for HbA1c-PDR data tables

Models	R	R Square	Wherry's Adjusted R Square	Std. Error of the Estimate	F-ratio	
					F	Sig
Linear	.736	.542	.509	9.616	16.549	.001
Logarithmic	.723	.523	.489	9.808	15.365	.002
Quadratic	.739	.546	.479	9.933	7.817	.006
Cubic	.739	.547	.477	9.925	7.839	.006
Power	.676	.457	.419	.635	11.800	.004
S	.668	.447	.407	.641	11.294	.005
Exponential	.676	.457	.418	.635	11.778	.004

Table 5. Statistical indices of the selected seven patterns for HbA1c-NPDR data tables

Models	R	R Square	Wherry's Adjusted R Square	Std. Error of the Estimate	F-ratio	
					F	Sig
Linear	.799	.639	.630	10.635	67.268	< .001
Logarithmic	.675	.455	.441	13.064	31.761	< .001
Quadratic	.837	.701	.685	9.813	43.312	< .001
Cubic	.840	.706	.682	9.854	28.865	< .001
Power	.779	.606	.596	1.001	58.564	< .001
S	.780	.608	.598	.999	58.899	< .001
Exponential	.706	.499	.486	1.129	37.862	< .001

Table 6. Statistical indices of the selected seven patterns for Dur-PDR data tables

Models	R	R Square	Wherry's Adjusted R Square	Std. Error of the Estimate	F-ratio	
					F	Sig
Linear	.905	.819	.804	9.939	54.420	< .001
Logarithmic	.823	.677	.650	13.287	25.167	< .001
Quadratic	.906	.820	.788	10.351	25.123	< .001
Cubic	.907	.822	.769	10.801	15.416	< .001
Power	.932	.868	.857	.325	79.216	< .001
S	.806	.649	.620	.531	22.204	.001
Exponential	.891	.794	.777	.407	46.193	< .001

**Table 7. Statistical indices of the selected seven patterns for Dur-NPDR data tables**

Models	R	R Square	Wherry's Adjusted R Square	Std. Error of the Estimate	F-ratio	
					F	Sig
Linear	.932	.870	.860	6.904	86.644	< .001
Logarithmic	.842	.709	.686	10.320	31.603	< .001
Quadratic	.946	.894	.877	6.470	50.736	< .001
Cubic	.962	.926	.906	5.640	46.104	< .001
Power	.908	.825	.812	.371	61.271	< .001
S	.868	.754	.735	.441	39.783	< .001
Exponential	.864	.747	.728	.446	38.413	< .001

Table 8. Statistical indices of the selected seven patterns for AER-PDR data tables

Models	R	R Square	Wherry's Adjusted R Square	Std. Error of the Estimate	F-ratio	
					F	Sig
Linear	.765	.585	.544	9.091	14.122	.004
Logarithmic	.595	.354	.289	11.349	5.479	.041
Quadratic	.801	.641	.562	8.914	8.043	.010
Cubic	.859	.738	.640	8.083	7.505	.010
Power	.542	.294	.223	.891	4.156	.069
S	.289	.083	-.008	1.015	.909	.363
Exponential	.632	.399	.339	.822	6.650	.027

Table 9. Statistical indices of the selected seven patterns for AER-NPDR data tables
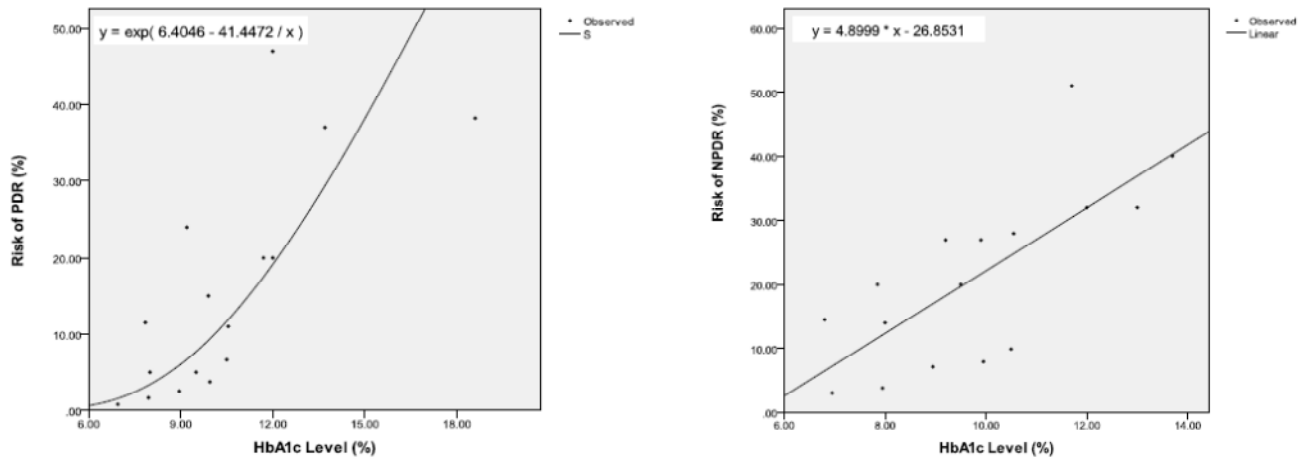
Figure 5. A scatter graph between HbA1c level and risk of PDR and NPDR

We applied the Durbin-Watson test for all dataset tables and all the results were between 1.5 and 2.5, which implies the existence of independent errors (Hutcheson et al. 1999). Then, using R-squared it can be clearly seen that the best selected regression models fit the data in all dataset tables better than ANN, as well as other six regression patterns (Table 10).

For all dataset tables, F is greater than one and also is significant at $p < .01$, which tells us that there is a less than 1% chance that F-ratios this large would happen if the null hypothesis were true. Therefore, we can conclude that our regression models result in significantly better prediction of complications than if we used the mean value of complications.

DISCUSSION AND CONCLUSION

In order to create 1-1 models, curve fitting was performed by using two different software applications: Tiberius (Brierley 2011) and SPSS (IBM 2010), which work based on ANN and least square regression, respectively. To start producing the model, seven different regression patterns, i.e. linear, logarithmic, quadratic, cubic, power, s and exponential, have been chosen as the best options. Using R-squared, it can be clearly seen that the best selected regression models fit the data in all the dataset tables better than ANN, as well as the other six regression patterns do. We also applied the Durbin-Watson test for all the dataset tables and all the results were between 1.5 and 2.5. This implies the existence of independent errors. In

addition, calculation of F-ratio and its associated significant value has been done. For all dataset tables, F is greater than one, and is significant at $p < .01$.

Dataset table	ANN	Best fitted regression patterns						
	R^2	R^2	D-W[1]	Wherry's Adj-R^2	Stein's Adj-R^2	F ratio		Pattern shape
						F	Sig.	
HbA1c-PDR	.560	.628	1.923	.602	.544	23.7	< .001	S
HbA1c-NPDR	.540	.542	2.104	.509	.439	16.6	.001	linear
Dur-PDR	.628	.701	1.822	.685	.668	43.3	< .001	quadratic
Dur-NPDR	.807	.868	2.396	.857	.833	79.2	< .001	power
AER-PDR	.866	.926	1.770	.906	.869	46.1	< .001	cubic
AER-NPDR	.549	.585	1.801	.544	.451	14.1	.004	linear

Table 10: Statistical details of ANN and the best fitted regression models

As mentioned in the methodology section, all of the models are created based on the previously observed relationships which are recorded by a number of surveys. In most cases, each relevant article had only one pair of numeric data to quantify an individual 1-1 relationship. Searching and finding a large number of articles and surveys which included useful information was a difficult part of this research. All of the models are created using the data from diabetic patients. Needless to say, models created based on epidemiologic information can be used only for the same population from which we have created the models. This simply means that as the data are gathered from diabetic patients, the models can be used to predict the complications of patients who already have diabetes.

As we know, however, each complication is affected by a variety of factors. In order to see the affect of all risk factors on one complication, we need to integrate all created models for that specific complication. Therefore, for every one of these factors one 1-1 model is created. By integrating these, N-1 model which shows the affect of all the mentioned factors together on the risk of retinopathy will be created. Ultimately N-K model of risk factors and the complication will be developed. This will indeed assist healthcare providers in focusing complications based on the risk factors. Therefore, this study will have the practical implications in healthcare delivery process by providing decision support tool for chronic disease management.

REFERENCES

1. Boulton, A. J. M., and Vileikyte, L. (2000) The Diabetic Foot, *Journal of Family Practice*, 49,11, S3.

2. Brierley, P.(2011) Tiberius data mining.

3. Brownlee, M., Aiello, L. P., Cooper, M. E., Vinik, A. I., Nesto, R. W., and Boulton, A. J. M. (2011). Complications of Diabetes Mellitus, in Williams textbook of endocrinology, Melmed, S., Polonsky, K., Larsen P.R.,Kronenberg, H.(Eds.), Elsevier/Saunders, Philadelphia.

4. Bulmer, M. G. (1979) Principles of statistics, (2 ed.) Oliver & Boyd, Edinburgh, London.

5. DCCT (The Diabetes Control and Complications Trial Research Group), (1996) The absence of a glycemic threshold for the development of long-term complications: the perspective of the Diabetes Control and Complications Trial. *Diabetes*, 45, 10, 1289.

6. DCCT (The Diabetes Control and Complications Trial Research Group) (2000) Retinopathy and nephropathy in patients with type 1 diabetes four years after a trial of intensive therapy, *New England Journal of Medicine*, 342, 6, 381-389.

7. European Diabetes Policy, G. (1999) A desktop guide to Type 2 diabetes mellitus, *Diabetic Medicine* ,16, 9, 716-730.

---

[1] Durbin-Watson

8. Field, A. P. (2009) Discovering statistics using SPSS, (3rd ed.), SAGE, London.

9. Griffith J (2000) Artificial Neural Networks: Are they ready for use as clinical decision aids? *Medical Decision Making,* 20, 243-4.

10. Habacher, W., Rakovac, I., Gorzer, E., Haas, W., Gfrerer, R.J., Wach, P., and Pieber, T.R. (2007) A model to analyse costs and benefit of intensified diabetic foot care in Austria, *Journal of Evaluation in Clinical Practice* , 13, 6, 906–912.

11. Huang, Y., McCullagh, P., Black, N., and Harper, R. (2007) Feature selection and classification model construction on type 2 diabetic patients' data, *Artificial Intelligence in Medicine* , 41,3, 251-262

12. Hutcheson, G. D., and Sofroniou, N. (1999) The multivariate social scientist : introductory statistics using generalized linear models, Sage Publications: London, Thousand Oaks, Calif.

13. IBM (2010) IBM SPSS Statistics.

14. Johnson, R. A., and Bhattacharyya, G. K. (2010) Statistics : Principles and methods, (6th ed.) John Wiley & Sons, Hoboken, NJ.

15. Lawrence J.(1994) Introduction to Neural Networks: Design. Theory, and Applications. Sixth ed. Nevada City California, California Scientific Software.

16. Razi, M. A., and Athappilly, K. (2005) A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models, *Expert Systems with Applications* 29,1, 65-74.

17. Sangi, M., Win, K. T., and Fulcher, J. (2010) A Knowledge-Based Risk Advisor Model for Chronic Complications of Diabetes, *Proceedings of the Pacific Asia Conference on Information Systems (PACIS 2010)*, 1798-1805.

18. Sherrod, P. H. Nonlinear Regression Analysis Program. (2010) [updated 2010; cited 18 May 2011]; Available from: www.nlreg.com

19. Smith, A. E., and Mason, A. K. (1997) Cost estimation predictive modeling: Regression versus neural network, *The Engineering Economist*, 42 , 2, 137-161.

20. Stevens, J., and NetLibrary, I. (2002) Applied multivariate statistics for the social sciences, (4th ed.) L. Erlbaum, Mahwah, N.J.