USE OF MACHINE LEARNING TO PREDICT THE ONSET OF DIABETES

Vinaytosh Mishra[1], Dr. Cherian Samuel[2], Prof. S.K.Sharma[3]

Department of Mechanical Engineering, IIT (BHU), Varanasi

ABSTRACT

Diabetes is growing like an epidemic in India. Prevalence is not only seen in urban areas but also in the rural parts of the country. The direct and indirect cost of the therapy is a major concern, with the cost rising with the progression of disease. Prediction in early stages can not only reduce the cost of therapy, but also prevent the multiple organ dysfunction and casualties. This paper uses classification techniques, like logistic regression to predict the disease in its early stages. The centres used for the study are diabetes speciality clinics in Varanasi.

KEYWORDS:

Diabetes Prevalence, Risk Score, Logistic Regression, Healthcare Burden

1. INTRODUCTION

Diabetes is becoming a pandemic in world and with 62 million diabetic patients; India is one of the significant contributors [1]. Over the past 30 years, the prevalence of diabetes has increased to 12-18% in urban India and 3-6% in rural India.This rate of increase is 50-80% higher than China (10%) [2].According to International Diabetes Federation (IDF), India is the home of most number of diabetic patients and hence it is rightly termed as the "diabetes capital of the world" [19]. The estimated burden for properly treating diabetes is USD 2.2 billion in India, while government was spending only USD 61 per capita on healthcare in year 2012[20, 21].

The following table shows all major studies done on prevalence of diabetes in India. Most of the studies are regional and none of them are done in Eastern U.P and Bihar. The literature on the studies underlines the fact, that there is a rising trend in the prevalence of Type 2 diabetes in Urban India [18].

[1] Research Scholar ,Industrial Management ,Department of Mechanical Engineering ,IIT-BHU
Corresponding Author: vinaytosh@gmail.com , Mobile: +91-8795832849/51
[2] Assistant Professor, Industrial Management ,Department of Mechanical Engineering ,IIT-BHU
[3] Professor, Industrial Management ,Department of Mechanical Engineering ,IIT-BHU

DOI : 10.14810/ijmech.2015.4202

Year	Authors	Place	Area	Prevalence %(Urban)
1971	Thripathy et al	Cuttak	Central	1.2
1972	Ahuja, et al	New Delhi	North	1.3
1979	Gupta,et al	Multicentre		3
1984	Murthy,et al	Tenali	South	4.7
1986	Patel	Bhadran	West	3.8
1988	Ramchandran,et al	Kudremukh	South	5
1991	Ahuja, et al	New Delhi	North	6.7
1992	Ramchandran,et al	Chennai	South	8.2
1997	Ramchandran,et al	Chennai	South	11.6
2000	Zarger,et al	Kashmir	North	6.1
2001	Ramchandran,et al	National		12.1
2001	Misra,et al	New Delhi	North	10.3
2001	Mohan,et al	Chennai	South	12.1
2001	Kutty,et al	Kerala	South	12.4

Table 1: Rising Prevalence of Diabetes in Urban India

Diabetes is a door way to multiple diseases and the cost of therapy increases with time. According to study on 3010 sample done by Ramchandran et al [18], the prevalence of vascular complication in diabetes is alarming.

Microvascular Complication		Macrovascular Complication	
Retinopathy	23.7	Cardiovascular disease	11.4
Background	20	Peripheral vascular disease	4
Proliferative	3.7	Cerebrovascular accidents	0.9
Nephropathy	5.5	Hypertension	38
Peri-neuropathy	27.5		

Table 2: Prevalence of Vascular Complications in Diabetes

Another study [Bhansali et al] further throws light on the cost of therapy with various complications. The patients with diabetes having foot complications spent 19020 INR, and those who had two complications spent four times more (17633 INR), and patients with renal disease (12690 INR), cardiovascular (13135 INR) and retinal complications (13922 INR) spent three times more than patients without any complications (4493 INR) [3].

Diabetes is associated with the sizeable proportion of the healthcare resource worldwide. Several studies have shown that timely intervention can prevent or postpone the onset of the disease. Therefore appropriate identification of individuals at high risk is important [4, 5]. Several risk scores to predict type 2 diabetes have been developed [6-9].

	ARIC 2005[20]	ARIC 2009[21]	AUS-DRISK[22]	Cam-bridge[23]	DESIR[24]	DPoRT[25]	FINDRISK concise[6]	FINDRISK full[6]	Framingham personal[26]	KORA S4/F4[27]	EPIC-Potsdam[28]	QD Score[29]
Year	2005	2009	2010	2008	2008	2010	2003	2003	2007	2010	2007	2009
Cohort size	7915	12729	6060	24495	3817	19861	4595	4435	3140	873	25167	2540753
Follow-up (years)	9	10	5	4·6	9	9	5	5	7	5	5	10
Definition of incident diabetes	Any incident	Any incident	Treated with drugs, diag-nosed with OGTT	Self-report, registries	Treated with drugs, diagnosed with fasting glucose measure-ments	Any incident	Treated with drugs	Treated with drugs	Treated with drugs, diagnosed with fasting glucose measure-ments	Diag-nosed with OGTT	Diag-nosed with OGTT	Any incident
Number of cases of incident diabetes	1292	2407	362	323	203	1410	194	182	160	91	849	78081
Country	USA	USA	Australia	UK	France	Canada	Finland	Finland	USA	Germany	Germany	UK
Age group included (years)	≥25	45–64	≥25	40–79	30–64	>20	35–64	35–64	54*	55–74	35–65	25–79

Table 3: Various Diabetes Risk Models Worldwide [10]

It is an irony that, a country which is infamous as the diabetes capital of the world, there is a lack of studies on the prediction of diabetes. Moreover, there is significant variation in lifestyle, ritual and eating habits in India, among various states. The models discussed above are based on studies done in specific geography. Existing diabetes prediction models are being used for prediction of diabetes, but the performance of each model varies with country, age, sex, and adiposity [10]. Despite the cultural specificity of Western medicine practices, that it is of a particular cultural tradition-it has been extraordinarily widely diffused throughout the world [11].

2. SCOPE

Thus, there is need of regional studies for diabetes prediction in India. The early intervention can reduce the prevalence of diabetes and hence the economic burden due to it.

OBJECTIVE

To develop an easy to administer diabetes prediction model for Eastern India.

SELECTION OF VARIABLE

1. Twenty five variables were selected using literature review of earlier studies.
2. Twelve diabetes prediction models were revisited, to find out whether those variables are included in the study
3. The variable having more than six inclusion was selected for the study
4. In addition to the above mentioned variables, one additional variable HBA1C was used. The HBA1C is being used as a popular diabetes prediction tool, and has been recommended by the International Expert Committee [17].

3. SAMPLING PLAN

The participants were selected from Eastern U.P and Bihar with Age ≥ 23 years and education more than fifth standard. Only two speciality centre from Varanasi were selected for the study. The definition of the incidence of diabetes is derived from the diagnosis of specialist doctors. Out of 200 participants, 106 were found diabetic, while 94 were adjudged non diabetic.

4. METHODOLOGY

Logistic regression is the model of choice in many medical data classification tasks [12]. The classification tool used for the study is Binary Logistic Regression, and the Software used for the Data Analysis is IBM SPSS 20.0. The power of the study in case of the logistic regression is, between 0.80 to 0.85, for 200 samples [13-16].

The purpose of the study is, not only to find whether the available data is able to classify two data sets namely diabetic and non diabetic, but also to provide a mechanism for removal of superfluous variables, to get more accurate models. This approach will save money, time and effort by dropping unnecessary tests, and considering only relevant questions [12]

The logistic regression model calculates the class membership probability for one of the two categories in the data set:

$$P(1|x,\alpha) = \frac{1}{1+e^{-(\alpha \cdot x)}},$$

and $P(0|x, \alpha)=1-P(1|x,\alpha)$. Here, we write $P(1|x, \alpha)$ to make the dependence of the posterior distribution on the parameters α explicit. The hyperplane of all points x satisfying the equation $\alpha \cdot x=0$ forms the *decision boundary* between the two classes; these are the points for which $P(1|x, \alpha) = P(0|x, \alpha) = 0.5$ [22]

6. RESULTS

The result of SPSS output for Model Summary and Variables used in the model is listed below.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	78.556[a]	.628	.839

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.

The Pseudo R –Square values Cox & Snell R Square and Nagelkerke R Square for the Model are 0.628 and 0.839 respectively. This shows that Model is appropriate using these variables.

Table 4: Model Summary

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	.325	.083	15.454	1	.000	1.384
	SEX	-.931	.926	1.011	1	.315	.394
	SMOKING	-2.865	1.352	4.488	1	.034	.057
	PARENTAL_DM	4.407	1.329	10.989	1	.001	81.998
	HYPERTENSION	4.948	1.346	13.525	1	.000	140.940
	BMI	.127	.142	.808	1	.369	1.136
	WAIST_CRCM	.153	.069	4.834	1	.028	1.165
	HBA1C	.844	.683	1.526	1	.217	2.326
	Constant	-40.591	7.866	26.631	1	.000	.000

a.) Variable(s) entered on step 1: AGE, SEX, SMOKING, PARENTAL_DM, HYPERTENSION, BMI, WAIST_CRCM, and HBA1C.

Table 5: Variables in the Equation

The variables such as Age, Smoking, Parental Diabetes Mellitus, Hypertension & Waist Circumference are significant while other variables like Sex, BMI and HBA1C are found insignificant for 95% confidence interval.

7. CONCLUSIONS

The reading of HBA1C was not found significant in the study. This raises question on over glorification of this tool as predictor of the diabetes. Are the simple anthropometric variables like Waist Circumference better predictor, than the three month average Blood Glucose Level?

LIMITATIONS OF STUDY

Only two centres were used for the study and the selection of participants was done through convenient sampling. More life style & location specific variables can be included as predictor.

FUTURE SCOPE OF STUDY

A multicenter study with more variables can give different results. More variables can be included using Delphi Method. The present methodology used i.e. Logistic Regression can be compared with more advance tools like ANN (Artificial Neural Network) for the results.

REFERENCES

[1] Seema Abhijeet Kaveeshwa, Jon Cornwall, The current state of diabetes mellitus in India, Australia Med J. 2014; 7(1): 45–48.

[2] Mohan V, Sandeep S, Deepa R, Shah B, Varghese C. Epidemiology of type 2 diabetes: Indian scenario. Indian J Med Res. Mar 2007;125(3):217-230

[3] Anil Bhansali, Cost of Diabetes Care : Prevent Diabetes or Face Catastrophe, JAPI • FEBRUARY 2013 • VOL. 61

[4] International Diabetes Federation. IDF Diabetes Atlas, 5th edn. Brussels: International Diabetes Federation, 2011.

[5] 2 Li G, Zhang P, Wang J, et al. The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: a 20-year follow-up study. Lancet 2008; 371: 1783–89.

[6] Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ 2011; 343: d7163

[7] Buijsse B, Simmons RK, Griffin SJ, Schulze MB. Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. Epidemiology Rev 2011; 33: 46–62.

[8] Lindstrom J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes Care 2003; 26: 725–31.

[9] Collins G S, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med 2011; 9: 103.

[10] Andre P K et al. Non-invasive risk scores for prediction of type 2 diabetes (EPIC-InterAct): a validation of existing models

[11] Nicholas A. Ethics are local: engaging cross-cultural Variation in the ethics for clinical research, Sm. SC; Med. Vol. 35, No. 9, pp. 1079-1091, 1992

[12] Stephan, Lucia. Logistic regression and artificial neural network classification models: a methodology review, Volume 35, Issues 5-6, Pages 352–359

[13] Agresti, A. 2002. Categorical Data Analysis, Second Edition Hoboken, NJ: Wiley.

[14] Agresti, A. 1996. An Introduction to Categorical Data Analysis. New York: Wiley.

[15] Cohen, J. 1988. Statistical Power Analysis for the Behavioural Sciences, Second Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

[16] Long, J.S. 1997. Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: SAGE Publications, Inc.

[17] David M Nathan. The International Expert Committee, International Expert Committee Report on the Role of the A1C Assay in the Diagnosis of Diabetes

[18] A. Ramchandran, Socio-Economic Burden of Diabetes in India, JULY 2007 VOL. 55

[19] Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE: Global estimates of diabetes prevalence for 2013 and projections for 2035 for the IDF Diabetes Atlas. Diabetes Res Clin Pract 2013, 49.

[20] Ramchandran A: Socio-economic burden of diabetes in India Assoc Physicians India 2007,55(L):9

[21] World Health Organization. Global Health Expenditure Database. Total expenditure on health/capita at exchange rate. 2012 [16.10.2014]

[22] B. Ripley, Pattern recognition and neural networks, Cambridge University Press, Cambridge (1996)