Validation of Semantic Discretization based Indian Weighted Diabetes Risk Score (IWDRS)

Omprakash Chandrakar
Department of Computer Science and Technology
Uka Tarsadia University, Bardoli, Gujarat, India

Jatinderkumar R. Saini*
Narmada College of Computer Application
Bharuch, Gujarat, India

Lal Bihari Barik
Department of Information Systems
Faculty of Computing & Information Technology in Rabigh, King Abdulaziz University,
Kingdom of Saudi Arabia

Abstract*—The objective of this research study is to validate Indian Weighted Diabetes Risk Score (IWDRS). The IWDRS is derived by applying the novel concept of semantic discretization based on Data Mining techniques. 311 adult participants (age > 18 years), who have been tested for diabetes using the biochemical test in pathology laboratory according to World Health Organization (WHO) guidelines, were selected for this study. These subjects were not included for deriving IWDRS tool. IWDRS is calculated for all 311 subjects. Prediction parameters, such as sensitivity and specificity are evaluated along with other performance parameters for an optimal cut-off score for IWDRS. The IWDRS tool is validated and found to be highly sensitive in diagnosing diabetes positive cases at the same time it is almost equally specific for identifying diabetes negative cases as well. The result of IWDRS is compared with the results of another two similar studies conducted for the Indian population and found it better. At optimal cut-off score IWDRS>=294, the prediction accuracy is 82.32%, while sensitivity and specificity is 82.22% and 82.44%, respectively.

Keywords—Data mining; indian weighted diabetes risk score; semantic discretization; type-2 diabetes risk score

I. Introduction

Undetected diabetes and prediabetes are the major concerns for East Asian countries, including India [1]. In such scenario, Diabetes Risk Score (DRS) tools can be proved effective in detecting undiagnosed diabetes and pre-diabetes cases. DRS tools are simple and easy to use computational tools that calculate the risk of diabetes of an individual's based on some risk factors.

Rest of the paper is organized as follows. Section 2 presents the literature review, which is followed by the discussion on Indian Weighted Diabetes Risk Score (IWDRS) in Section 3. Section 4 presents an outline of the research design. Details of experiments and results are discussed in Sections 5 and 6, respectively. The conclusion of the research study is given in Section 6.

II. Literature Survey

Various DRS tools have been reported in literature [2]-[14]. Basically, DRS tool uses a questionnaire to collect data from the target population. These data are used to build a mathematical model for predicting risk score of an individual. A mass diabetic screening test can be organized to detect undiagnosed and pre-diabetic persons, in which only those person who scored high on DRS, will be pathologically tested for high blood sugar. Developing countries like India, where lack of awareness, lack of pathological testing facilities, shortage of medical fund and late diagnosis is a major problem, DRS tools can be used as a cost-effective solution.

Several DRS tools have been developed and validated for different ethnic groups. A DRS tool, developed for a particular ethnic group, may not be generalized and may not produce similar results if applied on another ethnic group [15]. And that is why, separate DRS tools need to be developed and validated for each ethnic group, society, and country.

Logistic regression and Cox logistic regress models are used for deriving such risk scores, in which β coefficients of the risk factors are computed [10], [11], [14]. But building such logistic regression models are not a fixed, and it cannot be reproduced. Gary et al. [16] have observed that different investigators with the same data set produced different risk models. Anderson et al. [17] have argued that the diagnostic algorithm tools developed using logistic regression model is not perfect and prone to misuse.

To overcome the limitations of logistic regression models, Chandrakar and Saini have proposed a new methodology for deriving risk score and applied for deriving IWDRS [18]. IWDRS is derived by collecting data from a comprehensive questionnaire consisting of more than 60 risk factors [19], [20]. These risk factors are discretized using a novel concept of semantic discretization [21]. Then each risk factor is assigned to appropriate weight using machine learning techniques, and the corresponding risk score is calculated. One study Pima Indian Diabetes Dataset shows that classification accuracy is significantly increased when the dataset is semantically discretized before giving them to classifier [21]. In the present study, researchers validate the proposed IWDRS.

III. Indian Weighted Diabetes Risk Score

IWDRS is developed for Indian population considering demographic, socioeconomic, family and personal indicators. It includes parameters like age, family history of diabetes, blood pressure and high cholesterol, personal history of blood pressure and high cholesterol, BMI, waist circumference, diet quality, stress, physical activity and life quality. Various types of stress faced like work stress, financial stress, family or social

*Corresponding Author.

stress and health-related stress with its perceived intensity are considered. Life quality majors how the subject perceives the quality of his/her life, which includes qualitative indicators like happiness, love, and hope in their life. Responses of these parameters recorded at three different points of time. The responses of these parameters are categorized into three categories, low, moderate and high based on the rules derived using machine learning techniques. Table 1 shows the Indian Weighted Risk Score assigned to each parameter in each category.

TABLE I. INDIAN WEIGHTED RISK SCORE

		IWDRS		
No	**Risk Factor**	**Low**	**Moderate**	**High**
1	Age	10	27	63
2	Family History	16	41	44
3	Personal History	25	36	39
4	BMI	14	39	47
5	Waist Circumference	15	41	44
6	Diet	7	37	56
7	Stress	26	35	38
8	Physical Activity	15	16	69
9	Life Quality	14	22	64

IV. RESEARCH DESIGN

In this study, we validate the IWDRS, with the data which was not used in derivation. Data is collected from Advanced Diabetes Center, Surat, Gujarat (India). 311 adult subjects (age > 18 years), who have been tested for diabetes using the biochemical test in pathology laboratory according to World Health Organization (WHO) guidelines, were selected for this study. Out of total 311 subjects, 180 have tested positive for diabetes. IWDRS is calculated for all 311 subjects. Prediction parameters such as sensitivity and specificity are evaluated along with other performance parameters for an optimal cut-off score for IWDRS. The flow of this research study is as follows:

1) 311 adult subjects' records are used for validation.

2) IWDRS is calculated for each record.

3) Minimum and Maximum value for IWDRS is 142 and 464.

4) Considering 142 as base score, interval 142 – 464 is divided into 10 equidistance cutoffs.

5) Calculate Proportion of population and confusion matrix for each cutoff scores.

6) Calculate Prediction parameters for each cut-off scores.

7) Sensitivity, Specificity, and Accuracy are noted at the Optimal cut-off score.

8) Results are compared with two other Indian DRS.

V. EXPERIMENTS

Data are collected using the same questionnaire which was used to collect data for deriving IWDRS [18]. Data is collected from 311 adult subjects, of both genders, with age more than 18 years. Their diabetes status is confirmed with a biochemical test. 180 out of 311 subjects were diabetic. IWDRS is calculated for each of them.

Minimum and maximum possible score is 142 and 464 respectively. Considering 142 as base score, the IWDRS 142-464, is divided into 10 cutoff scores, which are 142, 175, 207, 239, 271, 303, 336, 368, 400, 432 and 464. Prediction parameters are calculated for the above cut-off score. Results are shown in Table 2.

TABLE II. INDIAN WEIGHTED DIABETIC RISK SCORE: PREDICTION PARAMETERS (MINIMUM AND MAXIMUM POSSIBLE SCORE BEING 142 AND 464, RESPECTIVELY)

IWDRS ≥	**Proportion of Population at High Risk (in %)**	**Prediction**				
		Sensitivity (in %)	**Specificity (in %)**	**PPV (in %)**	**NPV (in %)**	**Accuracy (in %)**
142	100	100	0	57.88	0	57.88
175	99.68	100	0.76	58.06	100	58.2
207	96.46	99.44	7.63	59.67	90.91	60.77
239	90.35	97.78	19.85	62.63	86.67	64.95
271	71.7	90	53.44	72.65	79.55	74.6
287	62.38	86.67	70.99	80.41	79.49	80.06
303	50.16	77.22	87.02	89.1	73.55	81.35
336	33.12	56.11	98.47	98.06	62.02	73.95
368	16.4	27.78	99.24	98.04	50	57.88
400	3.22	5.56	100	100	43.52	45.34
432	0.64	1.11	100	100	42.39	42.77
464	0	0	100	0	42.12	42.12

TABLE III. INDIAN WEIGHTED DIABETIC RISK SCORE: PREDICTION PARAMETERS (MINIMUM AND MAXIMUM POSSIBLE SCORE BEING 271 AND 303, RESPECTIVELY)

IWDRS ≥	**Proportion of Population at High Risk (in %)**	**Prediction**				
		Sensitivity (in %)	**Specificity (in %)**	**PPV (in %)**	**NPV (in %)**	**Accuracy (in %)**
271	71.7	90	53.44	72.65	79.55	74.6
275	70.1	88.33	54.96	72.94	77.42	74.28
278	68.49	87.22	57.25	73.71	76.53	74.6
281	65.92	86.67	62.6	76.1	77.36	76.53
284	63.67	86.67	67.94	78.79	78.76	78.78
287	62.38	86.67	70.99	80.41	79.49	80.06
291	58.2	83.33	76.34	82.87	76.92	80.39
294	54.98	82.22	82.44	86.55	77.14	82.32
297	53.05	82.39	80.39	87.88	72.57	72.99
303	50.16	77.22	87.02	89.1	73.55	81.35

Tables 2 and 3 present the sensitivity and specificity and accuracy of predicting diabetes for different cut-off values for IWDRS. From Tables 2 and 3, the highest prediction accuracy is 82.32% for IWDRS >= 294 and IWDRS >= 300. Sensitivity is 82.22% and 80.56% and specificity is 82.44% and 84.73%, respectively. Though prediction accuracy is same for both cut-off scores, at IWDRS>= 300, sensitivity is less than specificity, meaning that it predicts diabetes negative persons more accurately than diabetes positive persons, while our interest is in identifying diabetes person more accurately. So we choose IWDRS>=294 as the optimal cut-off score.

VI. RESULT ANALYSIS

Our study results are comparable and consistent with other studies reported in scientific literature. Experimental result of validation of IWDRS is shown in Table 3.

Two similar studies are found for Indian population. Mohan et al. [10] have developed simplified Indian Diabetes Risk Score using logistic regression model. Four parameters are used for developing the risk model, namely, 1) Age; 2) Obesity; 3) Physical activity; and 4) History of diabetes in the family. Ramachandran et al. [14] have also developed a DRS for Asian Indian population living in India using a logistic regression model with five parameters. They used 1) BMI; 2) Waist Circumference as a risk factor apart from; 3) Age; 4) Physical activity; and 5) History of diabetes in the family. Initially, Gender and Monthly income were considered as a diabetes risk factor, but not taken into account while developing the model. Table 4 compares the prediction statistics of these two risk score tools with their results with IWDRS.

TABLE IV. COMPARATIVE PREDICTION STATISTICS FOR IDRS, IADRS, AND IWDRS

No.	**DRS Tool**	**Proportion of Population at High Risk (%)**	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**
1	IDRS>=10 **IWDRS>=175**	99.4 **99.68**	100 **100**	0.7 **0.76**	10.7 **58.2**
2	IDRS>=20 **IWDRS>=207**	99.0 **96.46**	99.5 **99.44**	1.1 **7.63**	11.1 **60.77**
3	IDRS>=30 **IWDRS>=239**	93.3 **90.35**	97.7 **97.78**	7.2 **19.85**	16.7 **64.95**
4	IDRS>=40 IADRS>=13 **IWDRS>=271**	75.9 - **71.7**	93.1 91.8 **90**	25.5 33.6 **53.44**	32.4 - **74.6**
5	IDRS>=50 IADRS>17 **IWDRS>=303**	62.8 - **50.16**	84.9 86.4 **77.22**	39.4 47 **87.02**	43.0 - **81.35**
6	IDRS>=60* IADRS>21* **IWDRS>=294***	42.9 - **54.98**	72.5 76.6 **82.22**	60.1 59.9 **82.44**	61.3 - **82.32**
7	IDRS>=70 IADRS>25 **IWDRS>=368**	20.9 - **16.4**	42.7 54.1 **27.78**	81.1 77.4 **99.24**	77.2 - **57.88**
8	IDRS>=80 IADRS>29 **IWDRS>=400**	6.0 - **3.22**	15.1 33.5 **5.56**	95.0 88.5 **100**	86.9 - **45.34**
9	IDRS>=90 **IWDRS>=432**	0.9 **0.64**	2.3 **1.11**	99.3 **100**	89.5 **42.77**
	IWDRS>=464	**0**	**0**	**100**	**42.12**

** Optimal cut-off scores. - Data is not disclosed in the reference.*

VII. Conclusion

With prediction accuracy, 82.32%, IWDRS can be proved useful and inexpensive yet effective tool for a two-phase mass screening test for diabetes, especially in developing and underdeveloped countries like India where the undiagnosed or late diagnosis of diabetes is a major problem. In the first phase of the mass screening test, IWDRS can be calculated using an easy to response questionnaire for all subjects. In the second phase, only those subjects, who scored more than optimal cut-off value for IWDRS, are tested for the induced plasma glucose tolerance test using biochemical methods in the pathology laboratory, as per WHO guidelines. This two-phase mass screening approach will reduce the mass screening cost drastically in comparison single phased mass screening using pathology test only. By conducting a pathological test for only 55% of the population, we can detect 82% of the total diabetic person present in the population. In other words, for any given budget for the diabetes mass detection program, we can identify 20% more diabetic person if we use IWDRS tool in the first phase of two-phase diabetes screening.

References

[1] D. R. Whiting, L. Guariguata, C. Weil, and J. Shaw, “IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030,” Diabetes research and clinical practice, vol. 94(3), pp. 311-321, 2011.

[2] American Diabetes Association, “Risk Test,” 2017. [Online]. Available: http://main.diabetes.org/dorg/PDFs/risk-test-paper-version.pdf/.

[3] Finnish Diabetes Association, “Diabetes Risk Assessment,” 2017. [Online]. Available: https://www.diabetes.fi/files/502/eRiskitestilomake.pdf/.

[4] Diabetes UK, “Type 2 Diabetes: Know Your Risk,” 2017. [Online]. Available: https://riskscore.diabetes.org.uk/start?_ga=2.226350035.859741145.1506596366-1542314797.1506596366/.

[5] American Diabetes Association, “Diabetes Care,” 2017. [Online]. Available: http://care.diabetesjournals.org/

[6] Australian Government Department of Health, “Diabetes risk assessment tool,” 2017. [Online]. Available: http://www.health.gov.au/internet/main/publishing.nsf/content/diabetesriskassessmenttool/.

[7] Leicester Diabetes Centre, “The leicester diabetes risk score,” 2017. [Online]. Available: http://leicesterdiabetescentre.org.uk/The-Leicester-Diabetes-Risk-Score/.

[8] Canada Government, Public Health Agency of Canada, “The Canadian diabetes risk questionnaire,” 2017. [Online]. Available: http://healthycanadians.gc.ca/en/canrisk/.

[9] Diabetes Queensland “Assess your risk of developing type 2 diabetes,” 2017. [Online]. Available: http://www.diabetesqld.org.au/healthy-living/who-is-at-risk/assess-your-risk.aspx/.

[10] V. Mohan, R. Deepa, M. Deepa, S. Somannavar, and M. Datta, “A simplified Indian diabetes risk score for screening for undiagnosed diabetic subjects,” The Journal of the Association of Physicians of India, vol. 53, pp. 759-63, 2005.

[11] S.R. Joshi, “Indian diabetes risk score,” JAPI, vol. 53, pp.755-757, 2005.

[12] K. E. Heikes, D. M. Eddy, B. Arondekar, and L. Schlessinger, “Diabetes risk calculator,” Diabetes Care, vol. 31(5), pp.1040-1045, 2008.

[13] C. Glümer, B. Carstensen, A. Sandbæk, T. Lauritzen, T. Jørgensen, and K. Borch-Johnsen, “A Danish diabetes risk score for targeted screening,” Diabetes Care, vol. 27(3), pp.727-733, 2004.

[14] A. Ramachandran, C. Snehalatha, V. Vijay, N.J. Wareham, S. Colagiuri, “Derivation and validation of diabetes risk score for urban Asian Indians,” Diabetes Research and Clinical Practice, October 2005

[15] C. Glümer, D. Vistisen, K. Borch-Johnsen, and S. Colagiuri, “Risk scores for type 2 diabetes can be applied in some populations but not all”, Diabetes Care, 29(2), pp.410-414, 2006.

[16] G. L Grunkemeier, J. Z. Kathryn, and J. Ruyun. “Cardiac surgery report cards: making the grade.” The Annals of thoracic surgery 72.6, 1845-1848, 2001.

[17] R. P. Anderson, R. Jin, G. L. Grunkemeier, “Understanding logistic regression analysis in clinical reports: an introduction,” The Annals of Thoracic Surgery, Volume 75, Issue 3, Pages 753-757, Elsevier Science Ink, 2003

[18] O. Chandrakar, and J. R. Saini, “Development of Indian weighted diabetic risk score (IWDRS) using machine learning techniques for type-2 diabetes”, In Proceedings of the 9th Annual ACM India Conference, pp. 125-128. ACM. 2016.

[19] O. Chandrakar, and J. R. Saini, “Identification of parameters impacting diabetes risk score,” International Journal of Control Theory and Application, ISSN 0974– 5572, Volume 10 – No 18, 2017

[20] O. Chandrakar, and J. R. Saini, “Designing a comprehensive questionnaire for calculating diabetic risk score for Indian population,” In Proceedings of the International Conference on Artificial Intelligence In Health Care, pp 64, 2016

[21] O. Chandrakar, and J. R. Saini, “Knowledge-based semantic discretization using data mining techniques,” International Journal of Advanced Intelligence Paradigms, Inderscience Publication, in the press.