



Type 2 Diabetes Classification and Prediction Using Risk Score

¹Ramachandro Majji and ²Bhramaramba Ravi

¹Department of CSE,

GMRIT, Rajam.

rama00565@gmail.com

²Department of Information Technology,

GITAM Institute of Technology,

Visakhapatnam.

bhramarambaravi@gmail.com

Abstract

People understand the distribution of diabetes risk in the population associated with kidney problems, heart disease, early identification of patients with undiagnosed type 2 diabetes or those who are at an increased risk of developing type 2 diabetes is an important challenge in today's date. To create a risk prediction website for diabetes using commonly collected survey data. An individual can predict the number of years in which he/she will get diabetes. This website will predict the diabetes of the individual by taking the inputs from the individual and give a risk score depending their answers to the question asked. These questions will basically ask about their habits, age, sex etc., on depending these answers the score will be predicted.

Keywords: Diabetes, Risk Calculation, Classification & Prediction.

1. Introduction

Worldwide the incidence of Type 2 diabetes is increasing rapidly. The World Health Organization predicts that the number of people with Type 2 diabetes will double to at least 350 million by the year 2030 unless appropriate action is taken. In the UK, there are an estimated 2.6 million people with diabetes, with Type 2 diabetes accounting for between 85 and 95% of all cases.

Those from minority ethnic groups, including Black African, Black Caribbean and South Asians (India, Pakistan, Sri Lanka and Bangladesh) accounting for approximately 11% of the UK population have an increased risk for Type 2 diabetes.

The prevalence of Type 2 diabetes in people from the Indian subcontinent is between four- and six-fold higher than their White European peers and onset can be up to 10 years earlier. Early identification of patients at increased risk of developing Type 2 diabetes is thus an important and crucial issue to be resolved so that lifestyle interventions can be implemented to reduce the risk of developing diabetes.

The incidence among those at high risk as defined by impaired glucose tolerance can be significantly reduced by lifestyle changes, but population screening using oral glucose tolerance tests to identify those at increased risk is clearly impractical. Risk scores offer an alternative and simpler approach using key risk factors to identify individuals at increased

2. Literature Review

External validation of QDSCORE for predicting the 10-year risk of developing Type 2 diabetes

A small number of risk scores for the risk of developing diabetes have been produced but not yet been widely used in clinical practice in the UK. The aim of this study is to independently evaluate the performance of QDSCORE for predicting the 10-year risk of developing diagnosed Type 2 diabetes in a large independent UK cohort of patients from general practice. Results showed QDSCORE showed good performance data when evaluated on a large external dataset. There is a slight under estimation of risk in both men and women aged 60 years and above. Conclusion of this paper is that QDSCORE has shown to be a useful tool to predict the 10-year risk of developing Type 2 diabetes in the UK.

Indian Diabetes Risk Score

Variability is an inherent characteristic of the biological world. The globe today faces an epidemic of non-communicable diseases (NCD), which will soon surpass communicable diseases both in the developing and developed world. India is no exception, and both native and migrant.

Thus, for NCD a classic screening or preventive strategy may not work and principles of primordial prevention have to be applied. These lead to identification of “Risk Scores or Tests” like ‘Framingham Risk Score’ or criteria like cluster phenotypes “Metabolic Syndrome”. The classic cardiovascular risk is assessed by the Framingham risk score while diabetes is assessed by the diabetes risk score by American Diabetic Association (ADA).

The new IDRS score is simple user friendly and is currently tested in the CURES cohort. It will need validation in other population based studies from within different Indian states (The IDRS). This score may be incorporated into the proposed Indian National Diabetes Programme and surveillance studies on NCD by WHO and ICMR. The score will no doubt need further validation in future studies.

A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT)

National estimates of the upcoming diabetes epidemic are needed to understand the distribution of diabetes risk in the population and to inform health policy. To create and validate a population-based risk prediction tool for incident diabetes using commonly collected national survey data. Methods With the use of a cohort design that links baseline risk factors to a validated population-based diabetes registry, a model (Diabetes Population Risk Tool (DPoRT)) was developed to predict 9-year risk for diabetes. Results Predictive factors included were body mass index, age, ethnicity, hypertension, immigrant status, smoking, education status and heart disease. This algorithm can be used to estimate diabetes incidence and quantify the effect of interventions using routinely collected survey data.

3. Existing Statement

A small number of risk scores for the risk of developing diabetes have been developed but none has yet been widely used in clinical practice in the UK. There is thus currently no accepted risk prediction score that has been developed and validated for clinical use. QDSCORE, a new multi variable risk score has recently been developed to predict the 10-year risk of acquiring diagnosed. The risk score was developed and validated on a large cohort of patients (3.7million) from the Q Research database, two-thirds of the cohort randomly allocated for model development and one third to model validation. QDSCORE was derived using a Cox proportional hazards model. Fractional polynomials were used to model non-linear risk relationships with continuous risk factors and the presence of interactions between risk factors was tested. The final risk score included 10 risk factor terms (ethnicity, age, sex, body mass index, smoking status, family history of diabetes, Townsend deprivation score, treated hypertension, cardiovascular disease and current use of corticosteroids) and interactions terms between age and each of body mass index, family history and smoking status.

4. Proposed Statement

- We aimed to validate and compare existing prediction models for type 2 diabetes in the world. We assessed variability in predictive performance between countries and by sex, BMI, waist circumference, and age etc.
- We designed a website for Diabetes Risk Calculation based on parameters like BMI (body mass index), Age, Ethnicity, Smoking Status, Steroid usage etc in ECLIPSE IDE as our platform.
- This will calculate risk estimation of the particular user simply by considering few inputs and gives the predicted risk value as output.

5. Implementation Methodology

Classification

Classification is the process of finding model or function which describes and distinguishes data classes or concepts, for the intension of using the model to predict the class of objects whose class label is unknown. For classification we have used WEKA tool. Classification is a mechanism to classify the data set and name the classes. After classification calculate the classification rate using the formula. Using this algorithm, the data set is classified into two class labels namely tested positive and tested negative. WEKA is introduced by Waikato University, itis open source software written in Java and used for different purposes such as research, education. Gini index is used in selecting the splitting attribute. It uses both numeric and categorical attributes for building the decision tree and also uses in-built features to deal with missing attributes.

FT tree is classifier for Building functional trees, which are classification trees that could have logistic regression functions at the inner nodes and/or leaves. Missing values are dealt with by splitting the corresponding instances into pieces. Lad tree is class for generating a multi-class alternating decision tree using the Log it Boost strategy. In this experiment we are using diabetes database having 768 attributes and 7 instances, some of the attributes are preg, plas, pedi ,age ,class etc.

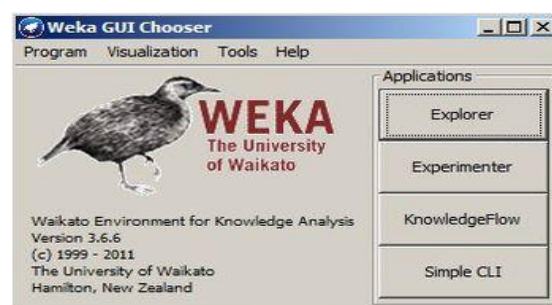


Fig 1: Illustrates WEKA Interface

Classification is a two step process, first, it builds classification model using training data. Every object of the dataset must be pre-classified i.e. its class label must be known, second the model generated in the preceding step is tested by assigning class labels to data objects in a test dataset. Here we are using diabetes dataset because now a day the percentage of diabetes patient is growing very fast.

According to Diabetes Atlas published by the International Diabetes Federation (IDF), there were an estimated 40 million persons with diabetes in India in 2007 and this number is predicted to rise to almost 70 million people by 2025.

India accounts for the largest number of people 50.8 million suffering from diabetes in the world. India continues to be the "diabetes capital" of the world, and by 2030, nearly 9 per cent of the country's population is likely to be affected from the disease. It is estimated that every fifth person with diabetes will be an Indian.

This means that India has highest number of diabetes in any one of the country in the world. Discretization is applied as preprocessing technique, such as preg, plas, pres, insu and it is apply to all attributes. Out of these attributes we preprocess it for preg attribute and it is shown by Fig.2

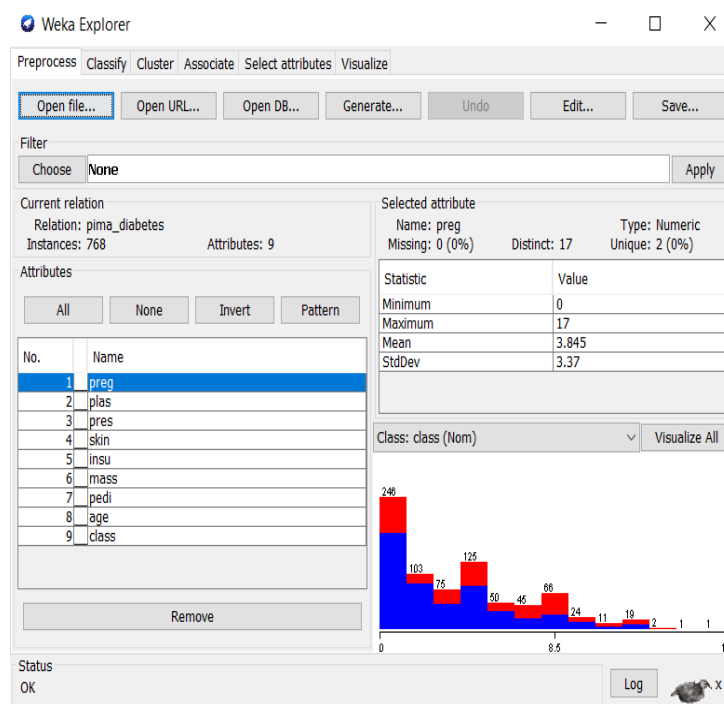


Fig 2: Preprocessing of Diabetes Database

Now we will compare the results of different classification techniques. Fig. 3

can shows Comparison of classifiers using different measures. In this paper out of all other classifiers the Random Tree gives better accuracy and also the time required is less. The below Fig.3 shows this random tree. Out of this Random tree gives good accuracy.

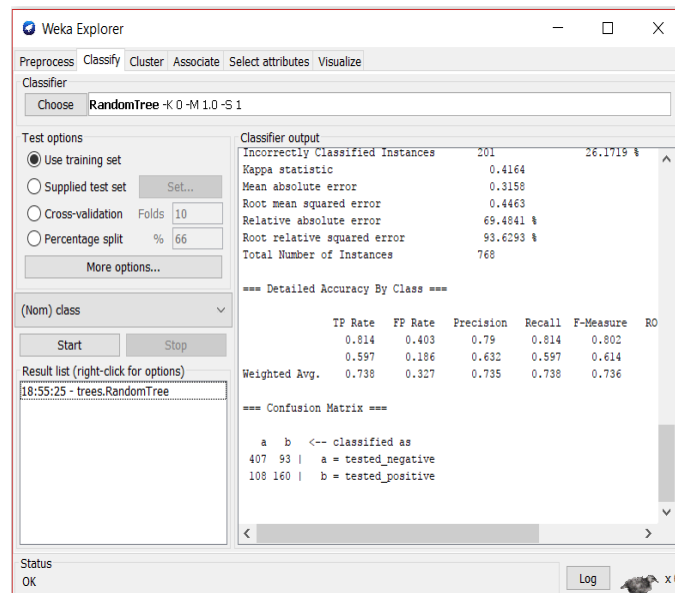


Fig 3: Classification using Random Tree

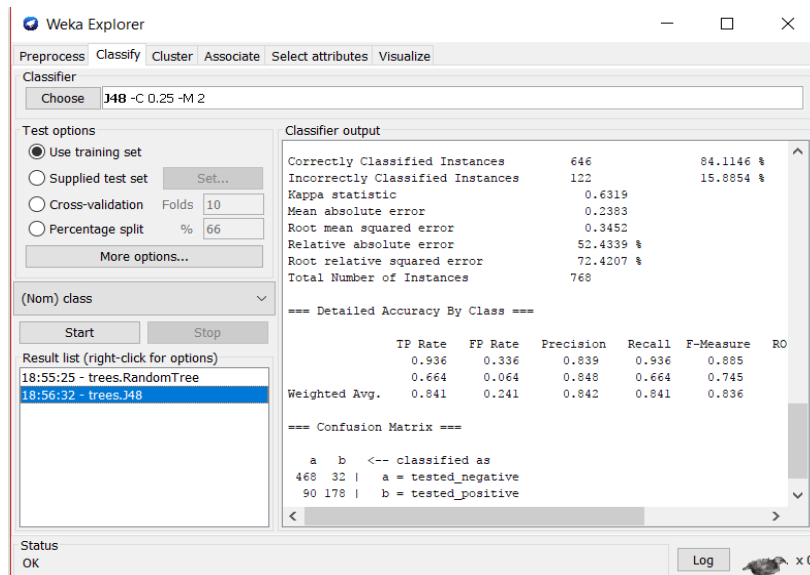


Fig 4: Classification using J48

Diabetes Prediction

In medicine, prediction tools are used to calculate risk, defined as the probability of developing a disease or state in a given time period. Global estimates place the number of people with diabetes at approximately 200 million, and increasing rapidly. There is a growing concern that these trends may slow or even reverse life expectancy gains in the USA and other developed countries. Planning for healthcare and public health resources can be informed by robust prediction tools.

Estimates of future diabetes incidence will alert policy makers, planners and physicians to the extent and urgency of the diabetes epidemic. In addition, a population prediction tool for diabetes can identify the optimal target groups for new intervention strategies, and determine how extensive a strategy must be to achieve the desired reduction in new cases. Clinical risk algorithms have been applied at the population level for other diseases, but with considerable challenges.

Clinical risk tools usually require clinical data that are rarely available at the population level. For diabetes, several clinical risk prediction tools exist, but they require clinical data that are collected infrequently or not at all at the population level, such as fasting blood sugar, or require detailed information, such as diabetes family history. The objective was to study a risk algorithm for diabetes incidence that can be applied at the level of populations using widely available public data. The Diabetes Population risk tool (DPoRT) was created for individuals to check the risk of getting affected by Type-2 diabetes.

Body-Mass Index

If your body-mass index is 25–30, you will benefit from losing weight. If your body-mass index is higher than 30, the adverse health effects of obesity will start to show, and it will be essential to lose weight. The body-mass index is used to assess whether a person is normal weight or not.

The index is calculated by dividing body weight (kg) by the square of body height (m). For example, if your height is 165 cm and your weight 70 kg, your body-mass index will be $70 / (1.65 \times 1.65)$, or 25.7. losing weight; at least you should take care that your weight doesn't increase beyond this.

Diabetes Risk Score Calculation

Interventions to prevent type 2 diabetes should be directed toward individuals at increased risk for the disease. To identify such individuals without laboratory tests, we use Diabetes Risk Score.

The Diabetes Risk Score is a simple, fast, inexpensive, non invasive, and reliable tool to identify individuals at high risk for type 2 diabetes.

Diabetes is associated with increased morbidity and mortality, and accounts for

a substantial proportion of use of health-care resources worldwide.

Two promising areas for further research are interventions that prompt lay people to check their own diabetes risk and use of risk scores on population datasets to identify high risk “hotspots” for targeted public health interventions.

Final Prediction Model (risk score) Includes Following Risk Factor Terms

- Ethnicity
- Age
- Sex
- Body mass index
- Smoking status
- Family history of diabetes
- Hypertension
- Current use of steroids

Table 1: Showing Risk Score Values

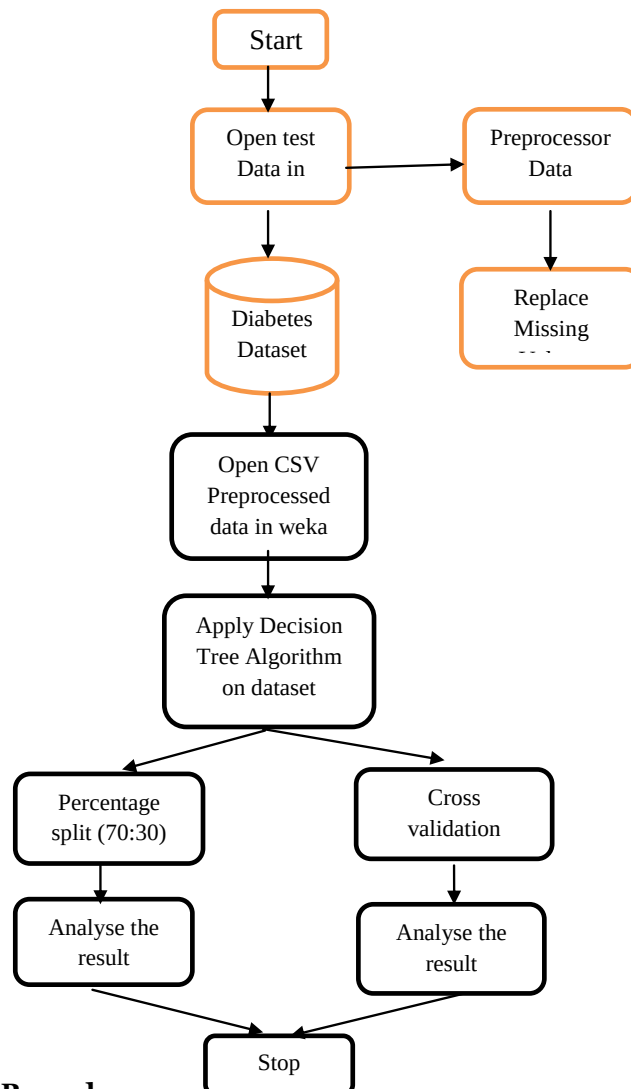
Variables	Score
Age (Years) 45-54	2
≥ 55	3
BMI (Kg/ m ²) >25 to ≤ 30	1
>30	3
Waist circumference (cm) Men, 94 to <102 ; Women, 80 to <88	3
Men, ≥ 102 ; women ≥ 88	4
Have you ever used drugs for high blood pressure? No	0
Yes	2
Has a physician or other health care provider ever told you that you have high blood glucose? No	0
Yes	5
Do you exercise or exert yourself in your spare time or at work at least 30 minutes on most days? No	0
Yes	2
How often do you eat vegetables and fruits or berries? Every day	0
Not every day	1

Algorithmic Procedure

Algorithm 1: J48 Algorithm

INPUT: Diabetes Database of University of California dataset pre-processed in CSV format.

OUTPUT: J48 Decision Tree Predictive Model with leaf node either tested-positive or tested-negative and Naïve Bayes Prediction Results.



Flow Chart Procedure

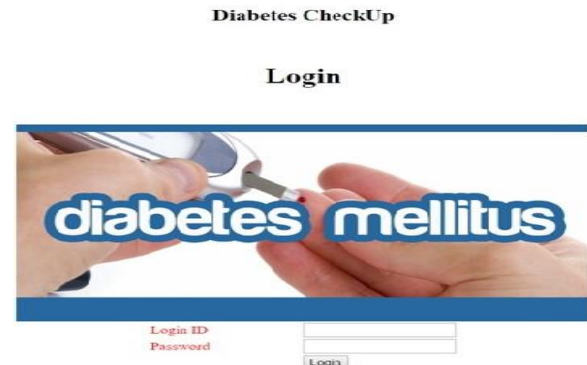
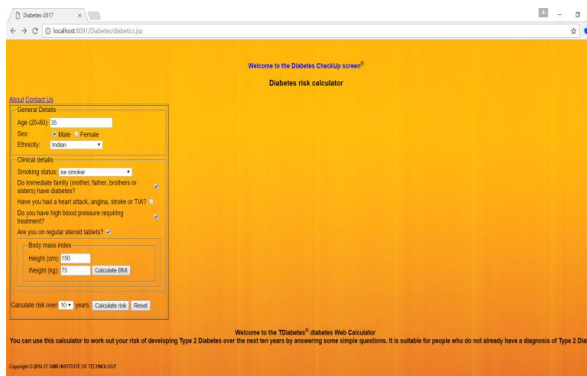
1. The dataset is pre-processed using WEKA tools. Following operations are performed on the dataset
 - a. Replace Missing Values and
 - b. Normalization of values.
2. Processed dataset is passed through feature selection wherein sets of attributes are deleted from the dataset.
3. The final processed dataset is uploaded in WEKA.
4. The J48 Decision Tree and Naïve Bayes algorithm are employed.
5. For purposes of the algorithms, Cross-Validation and Percentage Split techniques are applied for model creation.

6. Results

Output Screens

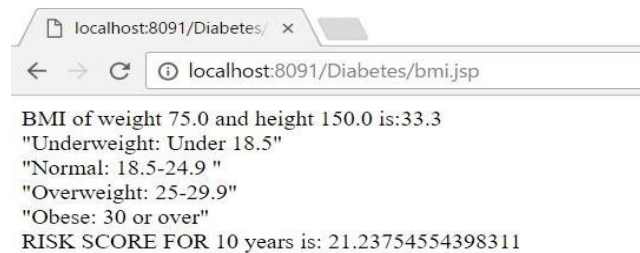
Diabetes CheckUp

Login

Diabetes checkup Entering Details.

Final Output



BMI of weight 75.0 and height 150.0 is:33.3
 "Underweight: Under 18.5"
 "Normal: 18.5-24.9 "
 "Overweight: 25-29.9"
 "Obese: 30 or over"
 RISK SCORE FOR 10 years is: 21.23754554398311

7. Conclusion

Diabetes dataset is classified using FP tree classifier algorithm which is used to estimate diabetes incidence and quantify the effect of interventions using routinely collected survey data and based on the results we developed a website which calculates the diabetes risk score for an individual. The Diabetes Risk Score is a simple, fast, inexpensive, noninvasive and reliable tool to identify

individuals at high risk for type 2 diabetes. This project is aimed to validate and compare existing non in prediction models for type 2 diabetes in the world. Variability is assessed in predictive performance between countries and by sex, BMI, waist circumference and age etc. They can get their risk score by answering simple questions displayed on the website based on their inputs given in the website the risk score will be calculated and displayed to the user. This score displays the person's risk of getting diabetes in the coming 10 years.

References

- [1] G. S. Collins and D. G. Altman, "External validation of QDSCORE for predicting the 10-year risk of developing Type 2 diabetes," Centre for Statistics in Medicine, University of Oxford, Oxford UK, 2011.
- [2] J.Lindstrom, J.Tuomilehto, "The Diabetes Risk Score, Practical Tool to Predict type 2 diabetes risk," Diabetes Care, 2015.
- [3] Aishwarya Iyer and Ronak Sambaly, "Diagnosis of Diabetes Using Classification Mining Techniques," International Journal of Data Mining & Knowledge Management Process (IJDMP) , 2015.
- [4] Murat Koklu and Yavuz Unal, "Analysis of a Population of Diabetic Patients Databases with Classifiers," International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering, 2013.
- [5] P.Dhakate, and Suvarna Patil, "Preprocessing and Classification in WEKA Using Different Classifiers," International Journal of Engineering Research and Applications, 2014.
- [6] Laura C Rosella, Douglas G Manuel, "A population-based risk algorithm for the development of Diabetes:Development and validation of the Diabetes population Risk Tool," BMJ Journals, 2011.
- [7] Gaganjot Kaur and Amit Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes," International Journal of Computer Applications, 2014.
- [8] Denye Ogeh and Segun Fatumo, "Towards Early Diagnosis and Prevention of Type 2 Diabetes: the Role of Smartphones," International Journal of Science and Research, 2012.
- [9] Gary S Collins, Susan Mallett, Omar Omar and Ly-Mee Yu, "Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting," BMC Medicine, 2011.
- [10] Hanley AJG, Williams K, Gonzalez C, et al. Prediction of type 2 diabetes using simple measures of insulin resistance - combined results from the San Antonio Heart Study, the Mexico city

diabetes study, and the insulin resistance atherosclerosis study, 2003.

- [11] Marcano-Cedeno, Alexis and Andina, Diego, "Data mining for the diagnosis of type 2 diabetes," World Automation Congress (WAC), 2012.
- [12] Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes," IEEE, 2012.
- [13] Gray, "Detection of impaired glucose regulation and/or type 2 diabetes mellitus," Diabetologia, 2012.
- [14] Yasodha and Kannan, "Analysis of a population of diabetic patients databases in Weka tool," International Journal of Scientific & Engineering Research, 2011.
- [15] Tuomileht, Lindstrom, Eriksson, Valle, Hamalainen, Ilanne-Parikka, Keinanen Kiukaanniemi, Laakso, Louheranta, Rastas, M., Salminen, V., Uusitupa, "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance.," The New England Journal of Medicine, 2001.

