Research on the Predictive Accuracy of Diabetes based on the BRFSS Database

Zhan Zhao *

College of Cereals, Oil and Food, Henan University of Technology, Henan, China

* Corresponding author: 631401030124@mails.cqjtu.edu.cn

Abstract. To study the factors that influence the disease of people with diabetes based on the BRFSS database, and to find out whether the BRSFF database has the ability to accurately predict whether an individual has the disease. This work adopts the method of data mining based on the 2015 BRFSS dataset, and uses R language to mine the relevant data in the hope of discovering the relationship between the BRFSS dataset and diabetics. The main data mining methods used include regression analysis, boxplot observation and chi-square test. The study found that the BRFSS dataset can accurately predict whether people will have diabetes, and the factors affecting the disease of diabetics in the BRFSS dataset are found. And it can be roughly observed whether these factors have a large or small impact on whether an individual has diabetes. The study provides some ideas for using BRSFF to study the associated effects of diabetes in people with diabetes, broadening the way to study people with diabetes.

Keywords: BRFSS database, diabetes, regression analysis, Chi-square test.

1. Introduction

Diabetes is characterized by persistent disruptions in the body's ability to regulate glucose levels, which can lead to significant medical complications. The global increase in obesity rates has been a major contributing factor to the upward trend of diabetes prevalence in recent decades. Premature mortality is a significant public health concern that requires attention. The World Health Organization and the American Diabetes Association have collaborated on creating diagnostic criteria for diabetes, using measurements of fasting blood glucose or 2-hour post-stress blood glucose, but there has been ongoing discussion surrounding the validity of these criteria [1].

The expenses associated with diabetes are skyrocketing at an alarming rate. It is expected that the overall expenses in the United States will surge by 53%, from $40.8 billion to $62.2 billion, within the timeframe of 2015 and 2030. While the expenses of diagnosing diabetes rose by 41% from 2007 to 2012, the costs incurred by individuals for their diabetes care only saw a 19% increase [2]. The expenses associated with diabetes are skyrocketing at an alarming rate. It is expected that the overall expenses in the United States will surge by 53%, from $40.8 billion to $62.2 billion, within the timeframe of 2015 and 2030. While the expenses of diagnosing diabetes rose by 41% from 2007 to 2012, the costs incurred by individuals for their diabetes care only saw a 19% increase [3].

The premier organization committed to safeguarding the well-being of the public through the application of scientific knowledge and data analysis is the Center for Disease Control and Prevention [4].Due to the multifaceted nature of diabetic complications, which can be influenced by a range of factors such as genetics, age, gender, blood sugar regulation, duration of diabetes, and other cardiovascular risk factors, conducting epidemiological studies on these complications can prove to be a challenging task. A persistent and enduring health condition, diabetes requires ongoing management and care [5].

In just eight years, the prevalence of diabetes rose significantly, with rates jumping from 4.9 per cent to 6.5 per cent in 1998. Across the board, there was a rise observed in every demographic, including males and females, individuals of every age, ethnicity, educational background, and nearly all regions within the country. The occurrence of shifts in prevalence differs according to alterations in the state. There is a strong association between the incidence of diabetes and certain lifestyle choices [6]. Geographic distribution assessments and BRFSS surveys were conducted to evaluate

individuals who were recently diagnosed with diabetes. Over the course of a decade, diabetes rates in 33 states saw a staggering 90 percent rise, from the mid-90s to the mid-2000s, when adjusted for age. The rate of occurrence increased from 4.8 out of every 1,000 people in 1995-1997 to 9.1 out of every 1,000 people in 2005-2007 [6]. The issue with the BRFSS analysis lies in the challenge of reconciling survey data with actual data collection. Despite incorporating mobile phone data and adjusting weight assignment methods, previous BRFSS data collected from landlines has been deemed reliable and valid through reliability and effectiveness research studies [7].

Understanding the magnitude of the issue is equally crucial. As of 2018, the Centers for Disease Control and Prevention (CDC) reported that diabetes affected a total of 34.2 million individuals in the United States. Furthermore, according to the CDC, approximately 20% of individuals with diabetes are unaware of their condition, and 10% are unaware of the potential risks associated with it. Type 2 diabetes is the prevalent form of diabetes, despite the existence of various other types. Individuals with limited financial resources are disproportionately affected by the weight of illness. Statistics show that the estimated cost of managing diagnosed diabetes amounts to roughly $327 billion, while the combined cost of undiagnosed diabetes and projected diabetes is nearly $400 billion [8]. 1984 marked the establishment of the Centers for Disease Control and Prevention (CDC). The BRFSS, which investigates the influence of behavioral risk factors, is determined through the utilization of the CDC's Behavioral Risk Factor Monitoring System. The CDC conducts annual telephone interviews to gather information on the health behaviors, chronic illnesses, and preventive care status of American citizens, with follow-up surveys administered annually. The BRFSS holds significant sway over the CDC's surveillance practices and holds the title for the most extensive system for monitoring risk factors globally [9]. The practice of sifting through vast quantities of data to uncover significant correlations, tendencies, and configurations is known as data mining, and it emerged as a valuable tool for database exploration in the late 1980s, representing an interdisciplinary field that blends various areas of expertise [10].

This paper uses the BRFSS 2015 study to use this survey data to select high correlated predictors There are numerous factors that can impact a person's overall health and well-being, including but not limited to hypertension, hyperlipidemia, tobacco use, hyperglycemia, excessive weight, advanced age, gender, race, dietary habits, physical activity levels, alcohol intake, body composition, income, relationship status, quality of sleep, frequency of medical exams, educational attainment, insurance coverage, mental health status, and various other variables. Investigating if the questions asked in the BRFSS survey are reliable indicators of an individual's likelihood of developing diabetes. What are the most significant risk factors in BRFSS that can accurately predict the likelihood of developing diabetes? Can a person's likelihood of having diabetes be accurately determined by using only a few risk factors simultaneously. Lastly, can feature selection be utilized to generate a concise questionnaire from the BRFSS that can effectively forecast an individual's likelihood of having diabetes or being at a heightened risk of developing diabetes.

2. Method

2.1. Data Source

The data for this experiment comes from a select feature subset of BRFSS 2015.Considering these risk factors, this study attempted to select features (columns or questions) in BRFSS that were associated with these risk factors. To help understand what these columns mean, I consulted the BRFSS 2015 Codebook to review the questions and information about the questions. So decided to try and match the variable names in the codebook to the variable names in the dataset I downloaded from Kaggle. And also refer to some of the same features picked by Xie et al. for their research paper on building a type 2 diabetes risk prediction model using machine learning techniques using the 2014 BRFSS [10].

2.2. Variable Selection

As table 1 shows, the variables selected are significant risk factors, and research in this area has identified the following as significant risk factors (in no strict order of importance) for diabetes and other chronic diseases such as heart disease: There are various factors that can impact one's overall health and wellbeing, including but not limited to blood pressure, cholesterol levels, smoking habits, diabetes, obesity, age, gender, competition, dietary choices, physical activity levels, alcohol consumption, BMI, household income, marital status, quality of sleep, time since last medical examination, level of education, access to healthcare through Medicare, and mental health status. The response variable Diabetes binary is dichotomous with two distinct categories. If your blood sugar level is 0, then you do not have diabetes, but if it is 1 or higher, you may have prediabetes or diabetes. The data collection comprises of 21 distinctive feature variables and lacks equilibrium.

Table 1. Variable description

Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits
0	1	1	1	40	1	0	0	0	0
0	0	0	0	25	1	0	0	1	0
0	1	1	1	28	0	0	0	0	1

2.3. Research Method

First, using boxplots could visualize the presence and absence of diabetes and various possible influencing factors. As shown in Figure 1, in some variables the comparison revealed large differences. Some outliers were observed. The statistical model of logistic regression, typically applied in predictive analytics and classification, has been implemented in this research. Using logistic regression, one can predict the likelihood of an occurrence happening. e. The decision to cast a ballot or abstain from voting is contingent upon the specific independent variables at play. The likelihood of the outcome is determined by the range of values the dependent variable can take, from 0 to 1. There are two distinct categories of samples: those categorized as "Diabetes" and those categorized as "non-Diabetes".

Therefore, the purpose of this classification method is to more sensitively detect the predisposing factors of diabetes. Stratified sampling is used to extract 70% of the observations in the dataset as training data, and the remaining 30% as test data. Furthermore, cross-validation was introduced during the process of building a regression model using training data. Cross-validation aims to assess the model's performance in making accurate predictions on unseen data, identify any problems such as overfitting or selection bias, and provide an understanding of how well the model will perform on diverse datasets. The training data undergoes preprocessing transformations, including centering and scaling.

In order to explain our data set and research topic in more depth, research plot the relationship between the diseased and non-diseased variables using scatterplots and fitting linear relationships. Upon receipt of the manuscript, we operate under the assumption that the authors have granted us the copyright to utilize the material for the specified publication. We operate under the assumption that the authors of the paper have given us permission to use their copyrighted material upon receipt.

3. Results and Discussion

3.1. Descriptive Analysis

First, this research will show the prevalence of diabetes (marked as factor 1) and the absence of diabetes (marked as factor 0) and visualize it with various possible influencing factors, as Figure 1 and 2 showing.

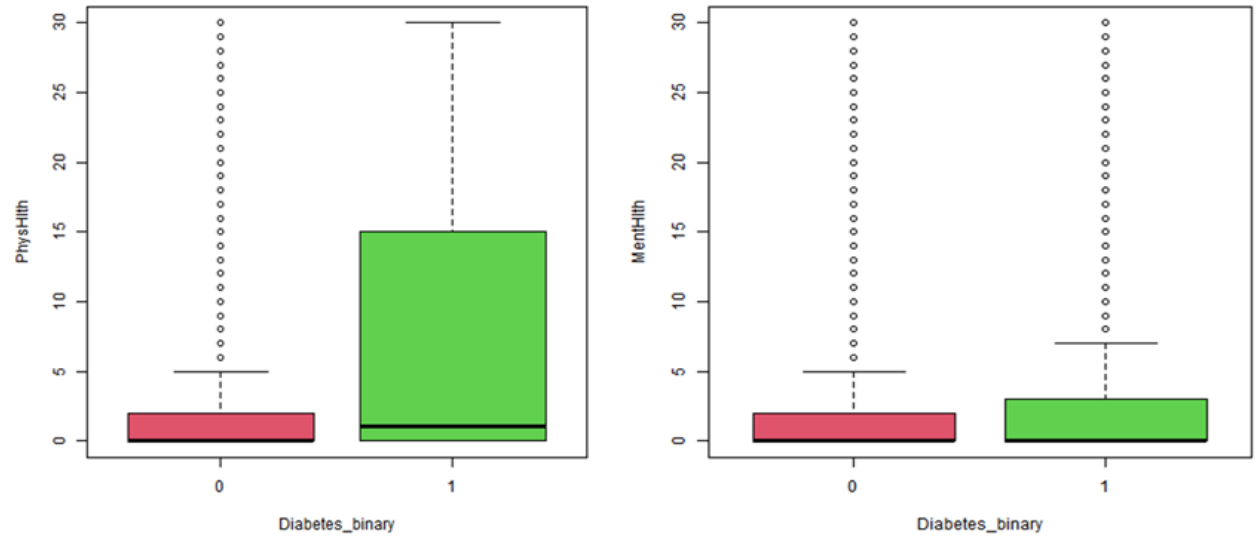

Figure 1. Compare diabetes_binary between physical and mental healthy

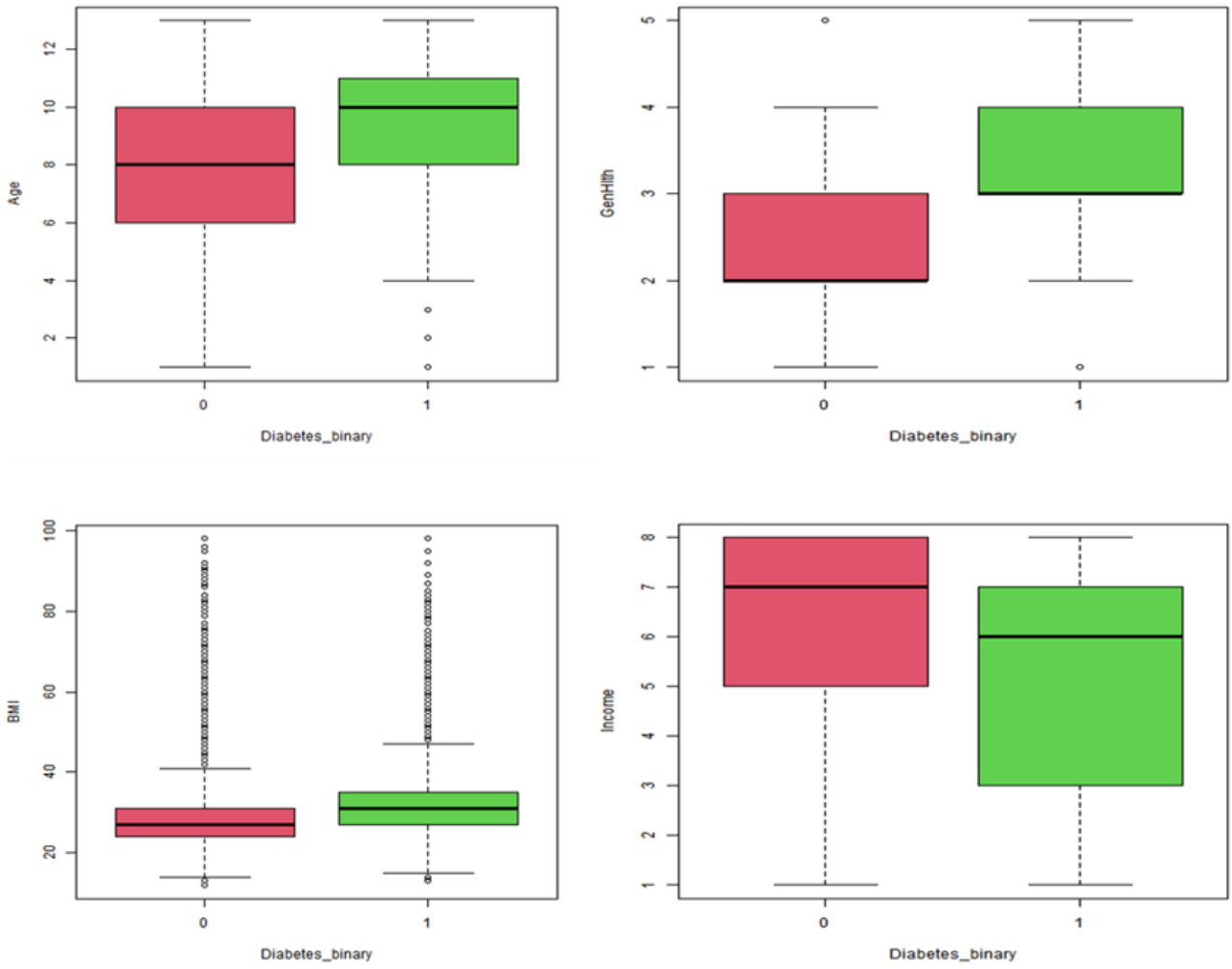

Figure 2. Compare diabetes_binary between the other variables

As shown by the boxplot (Figure 1 and 2) clusters, people with diabetes show larger differences in certain factors when compared to people without the disease. (Such as PhysHlth, PhysActivity, MentHlth, Income, BMI, HighBP, GenHlth, DiffWalk, Age). In addition, the affected and non-affected groups tend to be similar in other factors. No data is lost.

The figure reveals that diabetic individuals tend to have a PhysHlth value ranging from 10-20, while non-diabetic individuals usually maintain a PhysHlth value between 0-5. It shows that the PhysHlth of the sick person is much higher than the PhysHlth of the sick person. Secondly, there was little difference between the MentHlth values of individuals with diabetes and those without diabetes, which suggests that MentHlth may have little effect in predicting whether individuals have diabetes.

In addition, the income of individuals with diabetes is much lower than that of individuals without diabetes, which is about 1 on average. Predicting the likelihood of an individual developing diabetes may involve taking into account this particular factor. Additionally, individuals diagnosed with diabetes have an average age that is almost 2% greater than those who do not have diabetes, which is also an important correlation factor in predicting whether an individual will develop diabetes. Finally, it was also found that there was also a gap in GenHlth between individuals with diabetes and those without diabetes, which was also an important correlation factor.

3.2. Regression Analysis

The application of logistic regression as a statistical tool is widespread in the fields of classification and predictive analytics. The probability of an event occurring can be estimated through the use of logistic regression, for instance, making a decision to attend the concert or not, depending on the available budget and personal preference. The dependent variable can only be expressed as a probability, which necessitates that its range lies within the boundaries of 0 and 1. The samples were categorized into two groups through the process of research, “sick" (1) and "healthy" (0). A linear regression model was performed for all variables and diabetes as shown in Table 2.

Table 2. Linear regression results

Variables	Estimate	Std.Error	t value
(Intercept)	-2.983e-01	6,779e-03	-44.003
HighBp	7.530e-02	1.467e-03	51.321
HighChol	5.599e-02	1.380e-03	40.585
CholCheck	4.359e-02	3.375e-03	12.917
BMI	6.878e-03	1.017e-04	67.638
Smoker	-5.899e-03	1.324e-03	-4.455
Stroke	3.740e-02	3.317e-03	11.275
HeartDiseaseorAttack	6.743e-02	2.333e-03	28.901
PhysActivity	-6.849e-03	1.578e-03	-4.340
Fruits	-1.680e-03	1.379e-03	-1.219
Veggies	-2.781e-03	1.697e-03	-1.639
HvyAlcoholconsump	-5.075e-02	2.766e-03	-18.349
AnyHealthcare	1.506e-02	3.078e-03	4.892
NoDocbcCost	-7.334e-03	2.425e-03	-3.024
GenHlth	4.786e-02	7.895e-04	60.625
MentHlth	-6.044e-04	9.453e-05	-6.394
PhysHlth	1.084e-05	9.199e-05	0.118
Diffwalk	4.431e-02	2.084e-03	21.262
Sex	1.675e-02	1.315e-03	12.741
Age	7.310e-03	2.395e-04	30.528
Education	-3.300e-03	7.354e-04	-4.487
Income	-6.345e-03	3.727e-04	-17.023
F-statistic: 2333 on 21 and 253658 DF, p-value:<2.2e-16			

Through linear regression analysis of the data, as Table 2 shows. it is not difficult to see that Fruits, Veggies, PhysHlth, these variables do not have a strong influence. Given the impact of certain variables on the overall model, it is considered here that the choice of the "optimal" regression model is considered first. What is the "optimal" regression equation? The "optimal" here refers to selecting the variables that have a significant impact on Y from all the variables available to create an equation,

and does not include variables that do not have a significant impact on Y. There are multiple ways to obtain the "optimal" regression equation, such as "all subset regression method", "backward method", "forward method", "stepwise regression method". Among them, the stepwise regression method "is more commonly used because of the simplicity of computer programs. R software provides a more convenient "stepwise regression" calculation function step (), The purpose of eliminating or adding variables can be accomplished by choosing the AIC information statistic with the lowest value as the criterion.

Table 3. AIC test results

	Df	Sum of sq	RSS	AIC
<none>			25497	-582790
-Veggies	1	0.38	25498	-582788
-NoDocbcCost	1	0.91	25498	-582783
-Smoker	1	1.95	25499	-582773
-physActivity	1	2.02	25499	-582772
-Education	1	2.07	25499	-582772
-AnyHealthcare	1	2.42	25500	-582768
-MentHIth	1	4.23	25502	-582750
-Stroke	1	12.78	25510	-582665
-Sex	1	16.64	25514	-582627
-CholCheck	1	16.73	25514	-582626
-Income	1	29.18	25527	-582502
-HvyAlcoholConsump	1	33.73	25531	-582457
-Diffwalk	1	49.26	25547	-582303
-HeartDiseaseorAttack	1	83.98	25581	-581958
-Age	1	93.69	25591	-581862
-HighChol	1	165.99	25663	-581146
-HighBP	1	264.97	25762	-580170
-GenHlth	1	412.41	25910	-578722
-BMI	1	461.78	25959	-578239

The results of stepwise regression are shown in Table 3. The information obtained from the table, after several gradual regressions, the optimal "regression model" was found, although it was improved compared with the previous model, but the obtained AIC value was still too unreasonable, so the model's orientation to the result could not be fully accepted.

3.3. Chi-Square Test

Because the data set contains a high number of binary factor variables, it is suspected that the accuracy of the regression model test may be compromised. Therefore, it was determined that the best approach would be to utilize the chi-square test to analyze these factorial variables. Many researchers rely heavily on this hypothesis testing technique due to its widespread application. The chi-square test measures the extent to which the observed values in a statistical sample differ from the values that would be expected under a given hypothesis. The magnitude of the chi-square value is directly proportional to the extent of difference between the observed and predicted values, with larger chi-square values indicating greater deviation between the two. In contrast, when the two values are nearly identical, the chi-square value approaches 0, indicating a high level of agreement between the observed and expected values. The chi-square test can be used to analyze classified data, including comparisons of two rates or two composition ratios for statistical inference purposes. The statistical methods employed may include Chi-square testing to compare rates or composition ratios across multiple groups, as well as correlation analysis to examine relationships between categorical variables.

Through Chi-square test, the results shows that the effect between these factors is strong, and the p-value is almost less than 0.05. By conducting the chi-square test, it becomes evident whether these

variables hold significant importance in forecasting an individual's likelihood of having diabetes, and based on the test's outcomes, it is clear that variables with a p-value lower than 0.05 have a strong impact. In predicting whether an individual has diabetes, the variables of smoking, experiencing a stroke, consuming vegetables and fruits, engaging in physical activity, and having a history of heart disease or heart attack carry significant weight. An individual's likelihood of developing diabetes can be predicted by considering it as a factor that can influence the outcome.

4. Conclusion

In this study, some factors that influence the prediction of diabetes in an individual are simply identified through the box plot, and the BRFSS dataset has the ability to predict whether an individual will have diabetes. Based on this, the BRFSS dataset was analyzed by a linear regression model, and it can be seen from the fitted linear regression model that most of the variables have a significant influence, but the accuracy of the model is not high. Therefore, the dataset was revisited and found that most of the variables were dichotomous factorial variables, so a chi-square test was performed on some key variables. The comprehensive research discovered that a person's overall health and well-being can be affected by various factors such as blood pressure, cholesterol levels, body mass index, smoking habits, history of stroke or heart disease, level of physical activity, consumption of fruits and vegetables, heavy alcohol intake, access to healthcare services, general health status, mental health, and physical health. Walking differently can affect your earnings. The BRFSS database presents significant evidence that suggests a strong connection between various factors and the likelihood of an individual being afflicted with the disease. Through the data mining method of the chi-square test, it is easy to find that some of the above factors have a good correlation with whether an individual has diabetes. Therefore, the database is more accurate in predicting whether an individual will develop diabetes.

References

[1] Forouhi N G, Wareham N J. Epidemiology of diabetes. Medicine, 2014, 42 (12): 698 - 702.

[2] Rowley W R, Bezold C, Arikan Y, Byrne E, Krohe S. Diabetes 2030: Insights from Yesterday, Today, and Future Trends. Popul Health Manag, 2017, 20 (1): 6 - 12.

[3] Yang Wenying. Epidemiological characteristics and change trend of diabetes mellitus in China. Science in China: Life Science, 2018, 48 (08): 812 - 819.

[4] Duckworth W, Abraira C, Moritz T. Glucose control and vascular complications in veterans with type 2 diabetes. Journal of Vascular Surgery, 2009, 49 (4): 129 - 39.

[5] Mokdad AH, et al. Prevalence of obesity, diabetes, and obesity-related health risk factors. JAMA, 2003.

[6] Klonoff D C. The increasing incidence of diabetes in the 21st century. J Diabetes Sci Technol, 2009.

[7] Pierannunzi C, Hu S S, Balluz L. A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS). BMC Med Res Methodol, 2013, 13: 49.

[8] Anushka, et al. Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. New England Journal of Medicine, 2008.

[9] Ma Xiaoyu, et al. Construction of prediction model of childhood asthma based on artificial neural network applied to BRFSS database. Data Analysis and Knowledge Discovery, 2018, 2(8): 10 - 15.

[10] Group A S, et al. Effects of intensive blood-pressure control in type 2 diabetes mellitus. N Engl J Med, 2010, 362 (17): 1575 - 1585.