



# Statistical and Multivariate Analysis Applied to a Database of Patients with Type-2 Diabetes

Diana Canales<sup>(✉)</sup>, Neil Hernandez-Gress<sup>(✉)</sup>, Ram Akella<sup>(✉)</sup>, and Ivan Perez<sup>(✉)</sup>

Tecnologico de Monterrey, Mexico City, Mexico  
{canalesd,ngress,ivan.perez}@itesm.mx, akella@soe.ucsc.edu  
<http://tec.mx/en>

**Abstract.** The prevalence of type 2 Diabetes Mellitus (T2DM) has reached critical proportions globally over the past few years. Diabetes can cause devastating personal suffering and its treatment represents a major economic burden for every country around the world. To properly guide effective actions and measures, the present study aims to examine the profile of the diabetic population in Mexico. We used the Karhunen-Loève transform which is a form of principal component analysis, to identify the factors that contribute to T2DM. The results revealed a unique profile of patients who cannot control this disease. Results also demonstrated that compared to young patients, old patients tend to have better glycemic control. Statistical analysis reveals patient profiles and their health results and identify the variables that measure overlapping health issues as reported in the database (i.e. collinearity).

**Keywords:** Type 2 diabetes mellitus · Statistical analysis  
Multivariate analysis · Principal component analysis  
Dimensionality reduction · Data science · Data mining

## 1 Introduction

The number of people suffering from diabetes mellitus globally has more than doubled over the past three decades. In 2015, an estimated 415 million people worldwide (representing 8.8% of the population) developed diabetes mellitus; 91% of these people had type 2 diabetes mellitus (T2DM) [1]. Remarkably the International Diabetes Federation estimates that another 193 million individuals with diabetes remain undiagnosed. These individuals are at a great risk of developing health complications. The evidence documenting the large economic burden of treating T2DM has also risen dramatically in the past decade [2]. The causes of the epidemic are embedded in an extremely complex combination of genetic and epigenetic predisposition interacting within an equally complex combination of societal factors that determine behavior and environmental risks [3]. Great efforts have been taken to build a reliable T2DM patients

database and to determine the methodologies and statistical analysis that will allow researchers to identify the variables that best predict outcomes, and inform public health policies to reduce the epidemic and its associated social and economic costs [4]. The rest of the present work is organized as follows. In Sect. 2 we present the methodology used to develop statistical and multivariate analysis to be performed on the database of patients diagnosed with T2DM provided by the Mexican National Nutrition Institute. Results and conclusions are presented in Sects. 3 and 4, respectively.

## 2 Methods

### 2.1 Patient Database

The present study reports on an analysis of patient data provided by a third level hospital (i.e. highly specialized) from the National Nutrition Institute in Mexico. The database comprises  $p = 40$  health features in  $n = 204$  patients diagnosed with T2DM. The age of the patients ranges from 29 to 90 years ( $\mu = 61$ ,  $\sigma = 11.7$ ), with 80% of these patients between the age of 50 and 70 years, and 60% the patients are females. The health features include in this database comprise four socio-demographic features<sup>1</sup>, three non-modifiable risk factors<sup>2</sup> and 33 modifiable risk factors<sup>3</sup> that are commonly studied in the context of T2DM [4].

### 2.2 Multivariate Analysis

For multivariate analysis, we applied the Karhunen-Loève transform which is a form of principal component analysis (PCA) [5]. PCA allows for the identification of variable subsets that are highly correlated and could be measuring the same health indicator, implying dimensionality reduction. The method works through an orthonormal linear transformation constructed with the idea of representing the data as best as possible representation in terms of the least squares technique [6], which converts the set of health features, possibly correlated, into a set of variables without linear correlation called principal components. The components are numbered in such a manner that the first explains the greatest amount of information through their variability, while the last explains the least. The solution of the computation of the principal components is reduced to an eigenvalue-eigenvector problem reflected in a single positive-semidefinite symmetric matrix called the correlation or covariance matrix.

The eigenvectors of the correlation matrix show the direction in a feature space of  $p = 40$  dimensions in which the variance is maximized. Each principal component contains the cosine of the projection of the patients to the eigenvector correspond. This is relevant because if a variable can be associated with a

<sup>1</sup> Socio-economic strata, residence area, educational levels and occupation.

<sup>2</sup> Age, gender and size.

<sup>3</sup> Weight,  $HbA_{1c}$  (measure glycated hemoglobin), triglycerides, etc.

particular principal component, it must point approximately in the same direction of the eigenvector, and their cosine should approach the value 1. If the value of the cosine approaches 0, then the variable points in an orthogonal direction to the principal component and they are not likely associated. The product of each eigenvector by its corresponding eigenvalue will give each vector a magnitude relative to its importance. These scaled vectors are called *Factor Loadings*. The projection of each sample of  $n = 204$  patients to each eigenvector is called the *Factor Score*. This will help cluster samples to determine patient profiles.

The principal component analysis method requires a standardized database  $X$  for its development, i.e. each of its features have zero mean and variance equal to one.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} - & x_1 & - \\ - & x_2 & - \\ \vdots & \vdots & \\ - & x_n & - \end{pmatrix} = \begin{pmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_p \\ | & | & & | \end{pmatrix},$$

where  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$  represents the  $i$ th patient, and  $X_j = (x_{1j}, x_{2j}, \dots, x_{pj})^T$  represents the  $j$ th health feature,  $i = 1, 2, \dots, n$  and  $j = 1, 2, \dots, p$ , and  $n \geq p$ . Geometrically, the  $n$  patients represent points in the  $p$ -dimensional feature space.

The linear transformation that will take the database to a new uncorrelated coordinate system of features which keeps as much important information as possible and identify if more than one health feature might be measuring the same principle governing the behavior of the patients, will be constructed vector by vector.

Let  $v_1 = (v_{11}, v_{21}, \dots, v_{p1})^T \neq 0$  be this first vector such that, as the technique least squares [6], the subspace generated by it has the minimum possible distance to all instances. This problem can be represented mathematically as the following optimization problem:

$$\min \sum_{i=1}^n \|x_i - y_{i1}\|^2,$$

where  $y_{i1}$  denote the projection of the  $i$ th instance  $x_i$  onto the subspace spanned by  $v_1$ , and  $k_{i1} = \frac{\langle x_i, v_1 \rangle}{\|v_1\|^2}$ , with the usual Euclidean inner product and norm.

Then, by the Pythagoras Theorem  $\|x_i - y_{i1}\|^2 = \|x_i\|^2 - \|y_{i1}\|^2$ , and noting that  $\|x_i\|$  is a constant, the problem turn into  $\min \sum_{i=1}^n \|x_i - y_{i1}\|^2 = \max \sum_{i=1}^n \|y_{i1}\|^2$ . Thus, if  $Y_1 = (k_{11}, k_{21}, \dots, k_{n1})^T$  then  $\|Y_1\|^2 = \sum_{i=1}^n k_{i1}^2 = \sum_{i=1}^n \|y_{i1}\|^2$  and  $\max \sum_{i=1}^n \|y_{i1}\|^2 = \max \sum_{i=1}^n \|Y_1\|^2$ .

With some simple calculations involving the biased variance estimator, the correlation definition, the Euclidean norm, and the properties of  $X$  standardized, we can be concluded that  $\frac{1}{n} \|Y_1\|^2 = Var(Y_1) = v_1^T Corr(X) v_1$ , where  $Corr(X)$  is the correlation matrix of  $X$ . Therefore  $\max \sum_{i=1}^n \|Y_1\|^2 = \max v_1^T Corr(X) v_1$ .

Note that, since  $Corr(X)$  is a constant,  $v_1^T Corr(X)v_1$  increases arbitrarily if  $\|v_1\|$  increases. Thus, the problem turns into the next optimization problem

$$\begin{aligned} \max_{v_1} \quad & v_1^T Corr(X)v_1 \\ \text{subject to} \quad & \|v_1\| = 1. \end{aligned}$$

The Lagrange multiplier technique let conclude that  $v_1$  is the eigenvector of  $Corr(X)$  corresponding to the larger eigenvalue  $\lambda_1$ . Then  $\max v_1^T Corr(X)v_1 = \max \lambda_1$  solve the problem  $\min \sum_{i=1}^n \|x_i - y_{i1}\|^2 = \max \lambda_1$ .

This means that the first vector of the subspace that maintains the minimum distance to all the instances is given by the eigenvector  $v_1$  corresponding to the eigenvalue  $\lambda_1$  of the correlation matrix of the standardized database,  $Corr(X)$ . Then, as the correlation matrix is symmetric and positive-defined, by the Principal Axes Theorem,  $Corr(X)$  has an orthogonal set of  $p$  eigenvectors  $\{v_j\}_{j=1}^p$  corresponding to  $p$  positive eigenvalues  $\{\lambda_j\}_{j=1}^p$ . Which implies that by ordering the eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$  and following an analysis similar to the previous one, we have that the next searched vector is the eigenvector  $v_2$  corresponding to  $\lambda_2$ , the largest eigenvalue after  $\lambda_1$ , and so on for the following vectors. The Main Axes Theorem and the condition of the Lagrange Multipliers that each  $v_j$  must be normal imply that the set of eigenvectors is orthonormal, and  $\{Y_j\}_{j=1}^p$  are called the set of principal axes. So,  $Y_1 = Xv_1$  the first principal component,  $Y_2 = Xv_2$  the second principal component, and so on.

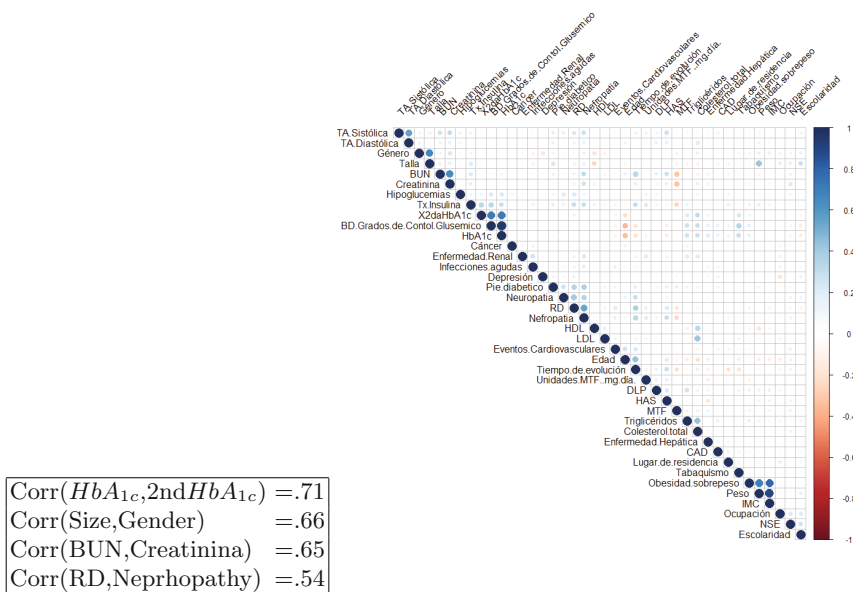
In this way, the new coordinate system that is given by the change to the base  $\{v_j\}_{j=1}^p$ , provides the orthonormal linear transformation that takes the standardized database to a new space of uncorrelated features that maintains the greatest amount of information from the original database. This new database is represented by the matrix of principal components  $Y = (Y_1, Y_2, \dots, Y_p) = X(v_1, v_2, \dots, v_p)$ . In general terms, what this means is that the projection of the database in the new coordinate system results in a representation of the original database with the property that its characteristics are uncorrelated and where the contribution of the information of the original database that each of them keeps, is reflected in the variances of the main components. This property allows extracting the characteristics that do not provide much information, fulfilling the task of reducing the dimensionality of the base. In addition, this technique allows the original database to identify and relate the characteristics that could be measuring the same principle that governs the behavior of the base contributing a plus to this analysis technique.

### 3 Results

Our statistical analysis revealed the following interesting trend in the database of T2DM: old patients tend to have good glyceimic control. Our analysis also shows that patients with poor glyceimic control are commonly young, are overweight or obese (70%), and belong to a low socio-economic strata (85%). Further, patients with poor glyceimic control frequently have an educational level lower than the

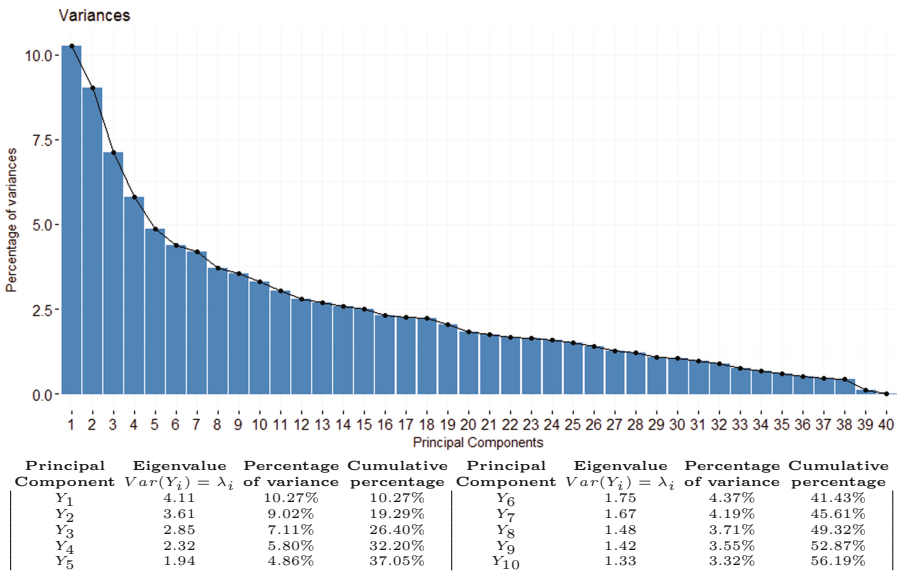
high school level (80%), are unemployed (66%), smoke (80%), and have higher levels of triglycerides and cholesterol. These patients also demonstrate great disease chronicity with a range of complications, such as liver disease (68%), diabetic foot (56%), hypoglycemia (71%), and diabetic Ketoacidosis (82%). They have to undergo insulin (62%) and metformin (52%) treatments. With regard to disease progression, the two glycemic measures ( $HbA_{1c}$  and  $2ndHbA_{1c}$ ) were associated with a 47% reduction in glycemic control, while 53% of patients retained the same level of glycemic control or improved, and 69% retained the same level control or worse. Thus, our results demonstrate that patients in the database who were remained in control in most cases. However, if patients had poor control they tended to retain poor control or even get worse.

Correlation analysis demonstrated that the first and second measures of  $HbA_{1c}$ , size and gender, and blood urea nitrogen (BUN) and creatinina, were significantly associated with diabetic retinopathy (DR) and nephropathy. Regarding the association with height and gender, it should be noted that on average, men are taller than women. Renal failure is usually measured through BUN and Creatinina. Additionally, DR and nephropathy are both known chronic complications of diabetes mellitus (see Fig. 1).



**Fig. 1.** The chart shows the upper triangular correlation matrix of the database. Positive correlations are displayed in blue and negative correlations are displayed in red color. Color intensity and the size of the circle are proportional to the correlation coefficients (see legend provided on right). The most highly correlated values are provided ant the top (bottom). (Color figure online)

Figure 2 shows variance, percentage of variation, and the cumulative percentage of variation for each of the ten principal components obtained via the Karhunen-Loève transform. The percentage of variation analysis indicates the amount of information explained by all of the health features in the database. Here, 32.2% of the health features in the database can be explained through the first four principal components. A list of the health features most highly correlated with each of the principal components is provided in Table 1 such that the values represent the correlation between each health feature and the corresponding principal component. For example 0.63 corresponds to the correlation between the first principal component and the health feature Evolution Time. Broadly, this list means that the Evolution Time, Nephropathy, BUN, and DR are interrelated and can together be represented by the first principal component. This principal component explains 10.3% of variance in health features and may reflect the chronicity of diabetes. This principal component suggests that a long evolution time is associated with a great risk of kidney damage and micro-vascular complications such as elevation of BUN and eye damage (DR). The second principal component, which explains 9.0% of the variance in health features, is associated with the degree of glycemic control. The third principal component, which explains 7.1% of the variance, is predominantly associated with weight. Finally, the fourth principal component which explains 5.8% of the variance in health features, measures patient height.



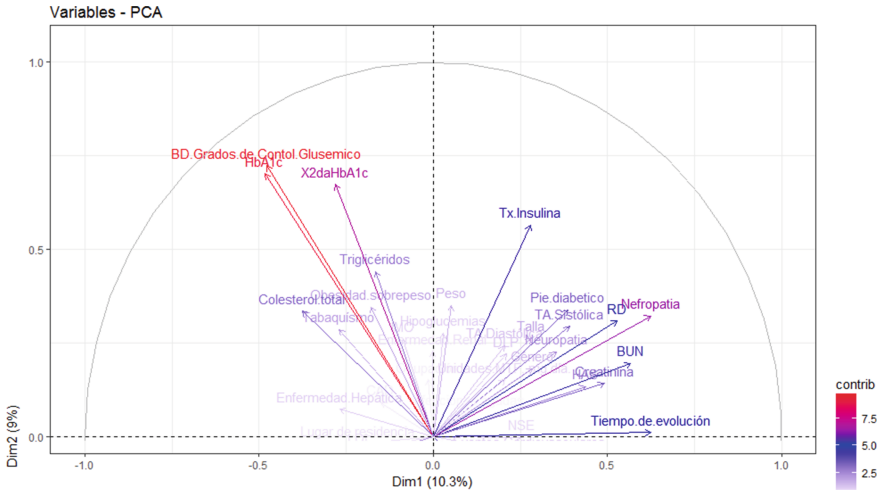
**Fig. 2.** Top: Percent of variance explained by each of the principal components  $Y_i$ ,  $i = 1, 2, \dots, 40$ . Bottom: For each component, the percentage variance and cumulative percentage variance is provided.

Figure 3 shows the graphic of the first two lists, the correlation between the principal components  $Y_1$  and  $Y_2$  with each health feature.

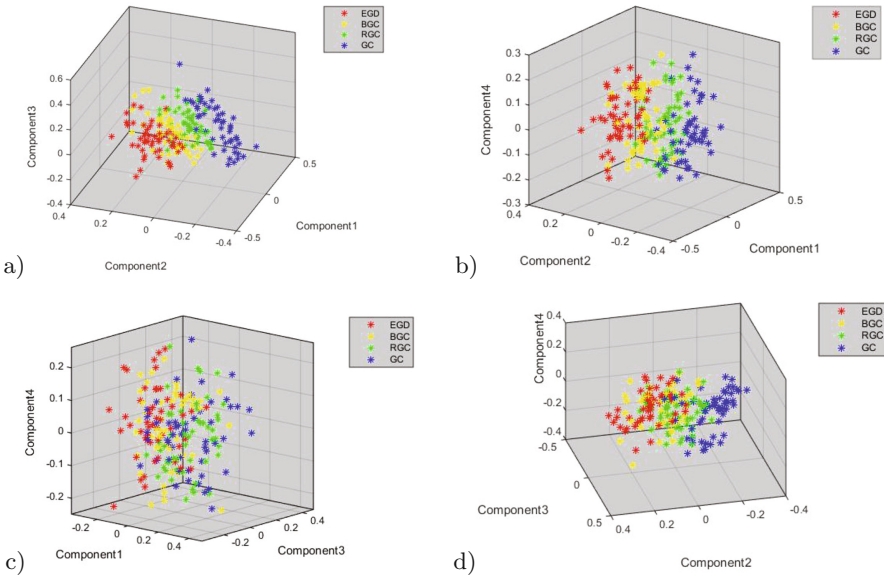
Figure 4 depicts plots for  $Y_1$ ,  $Y_2$ ,  $Y_3$  and  $Y_4$ , arranged in triads. The first graph depicts the plot of the first three principal components, wherein  $Y_1$  is related to chronicity of diabetes through the strong associations with Nephropathy, Evolution Time, BUN and DR,  $Y_2$  is related to glycemic degree control, and  $Y_3$  is related to weight. In each graph, blue points represent patients with high glycemic control (GC), green points represent patients with levels of regular GC (RGC), and yellow points represent patients with bad GC (BGC), and in red points patients in extremely poor GC (EGC). These patient groupings are not retained in the third scatter plot as this plot does not include the second component which determines the degree of GC.

**Table 1.** Correlation between the principal components and health features.

$Y_1$ (10.27%)	$Y_2$ (9.02%)	$Y_3$ (7.11%)	$Y_4$ (5.80%)
(.63)Evolution Time	(.72)Glycemic Control Degree	(.87)Weight	(.64)Size
(.62)Nephropathy	(.70)HbA <sub>1c</sub>	(.77)BMI (Body Mass Index)	(.61)Gender
(.57)BUN (Blood Urea Nitrogen)	(.67)2ndHbA <sub>1c</sub>	(.66)Overweight / Obesity	
(.53)DR (Diabetic Retinopathy)	(.57)TX Insulin		



**Fig. 3.** Correlation between each health feature and components  $Y_1$  and  $Y_2$ , i.e.  $Corr(Y_i, X_j)$ , where  $i = 1, 2$  and  $X_j$  are the health features,  $j = 1, 2, \dots, 40$ .



**Fig. 4.** Comparison of principal components. (a) Displays the first three principal components. (b) Displays the first (Nephropathy, Evolution Time, BUN and DR), second (Control of GC,  $HbA_{1c}$ ,  $2ndHbA_{1c}$  and insulin treatment) and fourth (Height and Gender) principal components. (c) Displays the first (Nephropathy, Evolution Time, BUN and DR), third (Weight, BMI and Overweight/Obesity) and fourth (Size and Gender) principal components. (d) Displays the second (Degree GC,  $HbA_{1c}$ ,  $2ndHbA_{1c}$  and insulin treatment), third (Weight, BMI, and Overweight/Obesity) and fourth (Height and Gender) principal components.

Figure 5, shows the contributions of the health features to the first, second, third and fourth principal components, respectively. Such percentage of contribution is given as follows  $C_i = \frac{Corr(Y_i, X_j)^2}{\sum_{j=1}^{40} Corr(Y_i, X_j)^2} \%$ , where  $X_j$  represents the  $j$ th health feature and  $Y_i$  represents the  $i$ th principal component. In each figure, the red dashed line on the graph above indicates the expected average contribution. If the contribution of the variables were uniform, the expected value would be  $\frac{1}{\#variables} = \frac{1}{40} = 2.5\%$ . Regarding joint contributions, Fig. 6 shows the contributions of health features to the four principal components. This joint contribution is given by  $Cc_4 = \frac{\sum_{i=1}^4 C_i Var(Y_i)}{\sum_{l=1}^4 Var(Y_l)}$ . The red dashed line in each of these figures indicates the linear combination between the expected average contribution and the percentage of the variance of the principal components, i.e.  $\frac{\sum_{i=1}^4 \frac{1}{40} Var(Y_i)}{\sum_{l=1}^4 Var(Y_l)} \%$ .





These profiles revealed that patients with poor GC tended to be young, overweight or obese, belonged to low socio-economic strata, had low education, were unemployed, and had high levels of triglycerides and cholesterol. In addition, patients with poor GC tended to have liver disease, diabetic foot, hypoglycemia, diabetic Ketoacidosis, smoke and take undergo insulin and metformin treatments.

Overall, we found that the poorer GC, the harder it is for them to stay in and the more they tend to get worse: 79% of those who have bad and extremely uncontrol GC remain bad or get worse. In contrast, the better they are, the more they stay in and their rate of decline is not so high: 66% of those in GC and Regular GC remain good or improve it.

In order to reduce dimensionality and extract more information from the relationship between the features of the dataset, in this work we applied the Karhunen-Loève transformation, a form of principal component analysis. Through this method the original dataset was taken to a new coordinate system of 20 dimensions under the least squares principle, with the property that its features (principal components) are not correlated and keep 80.43 % of the information of the original dataset, facilitating the handle and study of the dataset information. In addition, this technique allowed the original dataset to identify and relate the features that could be measuring the same principle that governs the behavior of the dataset through their principal components. Thus, we found that the first principal component, which has the highest amount of variance in the data, explained 10.3% of the health features and was related to diabetes chronicity. This component suggests that along disease evolution time is associated with a great risk of kidney damage and microvascular complications. The second principal component explained 9.0% of health features and was associated with the level of GC. The third principal component explained 7.1% of the variance and was predominantly associated with patient weight. Finally, the fourth principal component, which explained 5.8% of the variance, was associated with patients height. The remaining principal components did not reveal relevant information.

Future research should examine dataset that include a larger number of patients. In addition, we expect the advanced statistical analysis and machine learning tools and techniques will promote a further great discovery.

## References

1. Seuring, T., Archangelidi, O., Suhrcke, M.: International Diabetes Federation, 7th edn. Diabetes Atlas. International Diabetes Federation (2015)
2. Seuring, T., Archangelidi, O., Suhrcke, M.: The economic costs of type 2 diabetes: a global systematic review. *Pharmaco Econ.* **33**(8), 811–831 (2015)
3. Chen, L., Magliano, D.J., Zimmet, P.Z.: The worldwide epidemiology of type 2 diabetes mellitus-present and future perspectives. *Nat. Rev. Endocrinol* **8**, 228–236 (2012)

4. Hernandez-Gress, N., Canales, D.: Socio-demographic factors and data science methodologies in type 2 diabetes mellitus analysis. In: 2016 IEEE International Conference on Computational Science and Computational Intelligence, pp. 1380–1381. IEEE (2016)
5. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (2002). <https://doi.org/10.1007/b98835>
6. Wolberg, J.: Data Analysis Using the Method of Least Squares: Extracting the Most Information From Experiments. Springer Science & Business Media, Heidelberg (2006). <https://doi.org/10.1007/3-540-31720-1>