DE GRUYTER OPEN

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 39 (52) 2014

DOI: 10.2478/slgr-2014-0043

The Use of Principal Component Analysis and Logistic Regression in Prediction of Infertility Treatment Outcome

Anna Justyna Milewska[1], **Dorota Jankowska**[1], **Dorota Citko**[1], **Teresa Więsak**[2], **Brian Acacio**[3], **Robert Milewski**[1]

[1] Department of Statistics and Medical Informatics, Medical University of Bialystok, Poland

[2] Department of Gamete and Embryo Biology, Institute of Animal Reproduction and Food Research of Polish Academy of Sciences in Olsztyn, Poland

[3] Acacio Fertility Center, Laguna Niguel, California, USA

Abstract. Principal Component Analysis is one of the data mining methods that can be used to analyze multidimensional datasets. The main objective of this method is a reduction of the number of studied variables with the maintenance of as much information as possible, uncovering the structure of the data, its visualization as well as classification of the objects within the space defined by the newly created components. PCA is very often used as a preliminary step in data preparation through the creation of independent components for further analysis. We used the PCA method as a first step in analyzing data from IVF (in vitro fertilization). The next step and main purpose of the analysis was to create models that predict pregnancy. Therefore, 805 different types of IVF cycles were analyzed and pregnancy was correctly classified in 61–80% of cases for different analyzed groups in obtained models.

Introduction

Generally, clinical studies produce a large number of measurements that have an effect on the size of a database. To appropriately analyze such a large amount of information, data mining methods are usually employed (Milewski et. al., 2009, 2011, 2013b). The most popular methods of data mining are: neural networks, cluster analysis, correspondence analysis, or basket analysis (Milewska et. al., 2011, 2012, 2013; Milewski et al., 2009). Principal Component Analysis has a different approach to the problem of dimensionality of the database. This technique relies on transformation of the initial set of features into new uncorrelated variables. New

ISBN 978–83–7431–428–2 ISSN 0860-150X

variables are called the principal components. They represent linear combinations of original variables and they can describe relationships between studied characteristics. The main purpose of the analysis is the most optimal reduction of database size with minimum information loss, while at the same time maintaining as much variability in the data as possible. PCA allows for the selection of the best diagnostic characteristics and can be used in situations where subsequent analysis requires analyzing uncorrelated variables.

Principal Component Analysis

Principal Component Analysis was described for the first time by Pearson in 1901 (Pearson, 1901). Over twenty years later, Fisher and McKenzie proposed the first algorithm to PCA that is presently known as a NIPALS (Fisher et al., 1923). However, Hotelling (1933) made the major developmental impact on the method. Since then, PCA is quite popular (Daszykowski et al., 2008) as a chemometric method in the interpretation of complicated environmental and biological samples (Petrisor et al., 2012; Szefer, 2003). There are many examples of application of PCA in different fields of science: analytical chemistry (Suchacz et al., 2010), geology (Nowicki et al., 2013), agriculture (Kolasa-Więcek, 2012; Ukalska et al., 2008), psychology and sociology (Brzyski et al., 2012, Raskin et al., 1988) or in the analysis of food quality (Czernyszewicz, 2008; Koter et al., 2003; Rymuza et al., 2013) as well as in image and signal processing (Hladnik, 2013; Mudrova et al., 2005; Pandey et al., 2011).

The essence of Principal Component Analysis is to convert collections of the p variables $X_1, X_2, \ldots, X_p$ into a system of orthogonal variables:

$$\begin{aligned} Z_1 &= a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ Z_2 &= a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ &\vdots \\ Z_p &= a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p \end{aligned}$$

Newly created variables represent a new coordinate system obtained by the rotation of the initial axis of the system. PCA can be applied when observed variables are correlated. However, Tabachnick (1996) suggests that application of the method makes sense only when the correlation between some features is greater than 0.3. Coefficients a_{ij} are established based on the covariance matrix if analyzed variables are comparable. When

variables are expressed in different ranges of values or units, the matrix is converted into a correlation matrix using a standardization procedure (Webb, 2003). The PCA algorithm assumes that eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ and their corresponding eigenvectors of appropriate matrix should be calculated. Eigenvectors are normalized to become unit vectors. They can be interpreted as coefficients of principal components. Moreover, the i-th eigenvalue is equal to the variance of the i-th component. Newly created variables $Z_1, Z_2, \ldots, Z_p$ are arranged in descending order according to the explained variation. They are ordered by eigenvalues, from highest to lowest, so they are arranged with significance. The variable Z_1 is determined as the component that corresponds to the greatest eigenvalue. For a univocal solution, Z_1 is established in such a way that the following conditions should be executed:

$$\sum_{i=1}^{p} a_{1i}^2 = 1$$

The first component explains the highest part of variability of the collected data. Each of the succeeding components is determined by the maximization of the variability that was not explained by previous features. Because the succeeding principal components are in orthogonal position to each other, the sum of their variances provides information about the total variance of the initial variables. This allows one to establish what percentage of the total variability is represented by the i-th component:

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}$$

It offers a view on the significance of that component in analysis.

The result of the PCA analysis is the same number of variables as the number of observed characteristics. However, during further considerations only the first few and most important ones are taken into account because they explain the majority of the variability of the original set of features. The selection of the components can be based on the following criterion:

a) Criterion of sufficient quality of representation – it allows one to take into consideration such initial components that the sum of the variances corresponding to them determines a majority of the total observed variability of data (so it is greater than a certain, predetermined level e.g. 70%),
b) Keiser criterion – it allows one to select components whose variance is larger than 1,

c) Criterion based on the scree plot – on a linear graph, which presents the eigenvalues, there is a chosen and marked point to the right of which a mild decrease in values occurs. According to this criterion, only components whose variances are on the left of that point are taken into consideration.

Interpretation of the principal components is possible by exploiting component loadings (component coordinates). This parameter informs one about the correlation between the analyzed variable and selected component. Depending on what matrix was used in PCA, the parameter is set up as follows:

a) for the covariance matrix

$$r_{X_i Z_j} = \frac{cov(X_i, Z_j)}{s_i \sqrt{\lambda_i}} = \frac{\lambda_i a_{ij}}{s_i \sqrt{\lambda_i}} = \frac{\sqrt{\lambda_i} a_{ij}}{s_i}$$

b) for the correlation matrix

$$r_{X_i Z_j} = \sqrt{\lambda_i} a_{ij}$$

The coefficient of determination is used to establish what percentage of variability of some feature is explained by the chosen component. It is expressed as a square of component coordinates. A parameter called communality is obtained by summing coefficients of determination for all components included in further analysis. In this way it is possible to define how much of the variability of a feature is represented by the chosen set of components.

The analyzed data is projected into a space defined by the chosen principal components to reveal their structure. It is also possible to include components' coordinates in this space to see the effect of studied characteristics on the defined components. Figure 1 represents such an example.

Variables are represented by components' coordinates. Each time they are included into the unit circle of correlation. The length of the vector that links a point with the beginning of the coordinate system provides information about the communality of that characteristic. If the points are close to each other, then the correlation between the variables is strong. If the vectors are perpendicular, the characteristics are not correlated.

Principal Component Analysis is a tool that allows the size of enormous databases to be reduced, while at the same time maintaining control over loss of information. In addition, it enables visualization of observations. The representation of a sample in the reduced space permits one to establish relationships between variables. PCA is one of the data mining methods that allows one to discover connections hidden in the data and better their understanding. On the other hand, it can be used as a preliminary method when the final statistical tests require analyzing independent variables. For example, it is used as a first step in the analysis of regression.

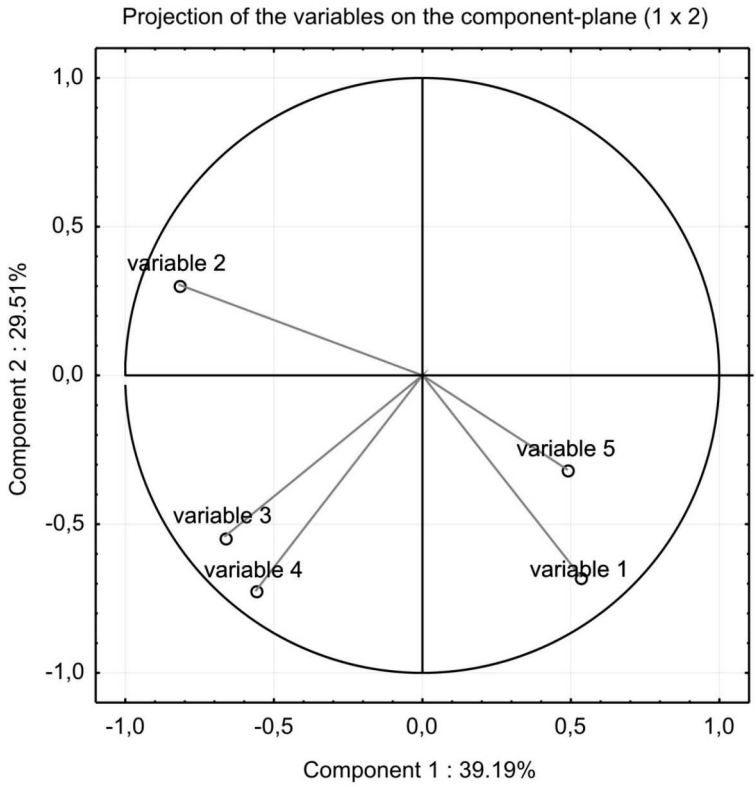

Figure 1. The representation of the components' coordinates in the space defined by the selected components

Application of PCA in Medical Science

Principal Component Analysis is primarily used to analyze high dimensional data sets. In practice, many variables are strongly correlated. It is sufficient to take into account only a small subset of variables to obtain a full picture of the described phenomenon. When analyzing all of the variables, it is usually impossible (or very difficult) to find all the relationships among the data. PCA allows a huge amount of information enclosed in initially correlated data to be transformed into a set of new orthogonal components. Therefore, the main applications of this method are: detection of such data structures, discovering concealed relationships, data visualization and object classification within the newly defined dimensions. PCA is very often used to obtain independent components in the preliminary preparation of the data. To simplify subsequent calculations, dimensionality reduction is required.

Such examples can be found in the modeling data by the neural networks (Duch et al., 2000), grouping data (Daszykowski et al., 2001) and in some methods of regression analysis (Martens et al., 1991; Næs et al., 2002).

The method is also applicable in biomedical studies. PCA usefulness has been proven in cancer detection epidemiology studies (Giuliani et al., 2000). The literature presents examples of PCA application: in genetic epidemiology to construct quantitative phenotypes for alcoholism (Scholz et al., 1999), in nutritional epidemiology to assess dietary patterns (Hoffmann et al., 2004; Varraso et al., 2012), in cardiology to predict clinical cardiovascular events (Agarwal et al., 2012), in radiology to compress magnetic resonance imaging (Furman-Haran et al., 2014), and in pharmacology (Konieczna et al., 2008; Nascimento et al., 2012).

There has been a lot of interest in applying this technique for gene expression studies (Biffi et al., 2010; Patterson et al., 2006; Raychaudhuri et al., 2000). In analyzing microarray experiment results, with thousands of variables in the database, PCA is primarily used for dimensionality reduction. Classification methods of explored data are later used to classify that data. Because the sample space of microarray data is probably nonlinear in nature, a popular generalization of linear PCA is therefore applied very often – namely Kernel Principal Component Analysis (KPCA) (Gastinel, 2012; Reverter et al., 2012). KPCA with radial basis kernel was used among others in two databases: the leukemia data set and the lymphoma data set. Both of these databases contained a few thousand expressed genes in three classes. The application of the KPCA with radial basis kernel method in both cases allowed for detection of the group structure in reduced dimension as well as full separation of studied classes (Reverter et al., 2012).

The described method found application in medical imaging processes. PCA is used in digital image compression such as in the structural image of the brain obtained during magnetic resonance treatment (Santo, 2012). The way the number of main components affects the quality of the picture has been shown (the fewer principal components used in the characteristics vector, the more degraded the quality of the image recovered). Moreover, it has been observed that the compressed medical images maintain the principal characteristics until they reach approximately one-fourth of their original size. This can be exploited towards saving storage space of medical images.

PCA is also used in the analysis of tomographic PET and SPECT images of the brain. That method is primarily used to reduce high dimensionality of the neuroimaging data (Stuhler et al., 2012). The effects of age related changes in the brain were analyzed using the PCA method. Age was significantly correlated with the first two principal components (Zuendorf et al.,

2003). Additionally, the PCA method allows for the selection of characteristics in the magnetic resonance picture study of brain tumors (Pushpa Rathi et al., 2012) or, in conjunction with the 3TP method, in breast cancer diagnostics (Furman-Haran et al., 2014). The technique was used for dimensionality reduction in diagnostics of atherosclerosis from Carotid Artery Doppler Signals (Latifoglu et al., 2008) and in the analysis of electrocardiogram (ECG) signals to diagnose cardiac arrhythmia (Martis et al., 2013). The above method was also applied to demonstrate pathological voice changes by reduction of the principal parameters number obtained from the acoustic analysis of speech (Panek, 2014).

PCA is applicable in the analysis of regression (Agarwal et al., 2012; Aguilera et al., 2006; Akinsola et al., 2014; Kaur et al., 2012; Ma, 2007) when a multidimensional database has to be analyzed and it is impossible to include all the variables in a statistical model because the data suffers from multicollinearity. The way to exclude correlated variables is to replace them with principal components.

As an example of PCA application in the analysis of regression, a study was conducted to determine the effects of different factors on cancer diagnosis (Belasco et al., 2012). The first step of the study was the Health Care Access Index (HCAI index) determination of many socio-economic components using the PCA method. The results of the two first components of the obtained index were then included into an analysis of regression. Performing such an operation eliminated the problem of correlated variables, reduced the degrees of freedom in the regression models and improved goodness-of-fit. Models based on PCA were better fitted to the data than models that contained all variables.

Application of PCA and Logistic Regression Analysis in Pregnancy Prediction

More and more people struggle with infertility and for most of them the IVF procedure is a chance to have a baby. The current IVF success rate is approximately 40% and diminishes with age (Milewski et al., 2008, 2013a). Many factors affect IVF success and some of them are still unknown (for example – idiopathic infertility). Many studies have been carried out to improve the success rate in IVF (Milewski et al., 2009, 2013b). The aim of our analysis was to create a statistical model that will be able to predict pregnancy. The data (805 IVF cycles) for our analysis are from one of the IVF clinics in the USA (Acacio Fertility Center, CA) (Table 1).

Table 1. Different types of cycles in infertility treatment

Group IP	n=610	The embryos of intendant parents were transferred to the female/mother uterus.
Group IP-PGD	n=84	The embryos of intendant parents after PGD testing were transferred to the female/mother uterus.
Group D-IP	n=68	An anonymous egg donor provided oocytes.
Group IP-GC	n=22	The embryos of intendant parents were transferred to the uterus of a gestational carrier.
Group D-IP-GC	n=21	An anonymous egg donor provided oocytes and the embryos of intendant parents were transferred to the uterus of a gestational carrier.

Two groups of predictors were selected. The first group included the quality and number of oocytes retrieved. The second group included the quality and number of embryos obtained after fertilization (e.g. number of oocytes inseminated with ICSI, number of different types of embryos) and information about the embryo transfer (number of embryo transferred, day of embryo transfer). There were strong correlations between the variables included in statistical analysis. Therefore, first the PCA method was employed. The analyses were carried out for the each type of procedure (Table 1). Table 2 presents a median for the ages in the groups.

Table 2. Age and percentage of pregnancy in compared groups

Group	Age (median)			pregnancy
	IP – mother	D – egg donor	GC – gestational carrier	n (%)
IP	38 years	–	–	n=215 (35%)
IP-PGD	37 years	–	–	n=28 (33%)
D-IP	43 years	25 years	–	n=52 (76%)
IP-GC	37.5 years	–	32 years	n=11 (50%)
D-IP-GC	42 years	26 years	31 years	n=15 (71%)
together GC	39 years	26 years	31 years	n=26 (60.5%)

Principal Component Analysis was performed by Statistica Data Miner + QC 10.0 software (StatSoft). Logistic regression was performed by STATA 12.0 software. Statistical significance was determined at the $p < 0.05$ level.

Statistically significant models were obtained in 6 cases: two in the patients of the IP group (the oocytes and embryos of intendant parents were transferred to the female/mother uterus), two in the D-IP group (the oocytes were collected from an egg donor), one model in the IP-PGD group (the embryos were subjected to genetic testing) and one in the group where a gestational carrier was used. Comparing the groups according to age, the oldest were mothers (median 37–43 years), followed by gestational carriers (31–32 years), and the youngest were egg donors (median 25–26 years). The best pregnancy outcome was obtained when egg donors were used (71% and 76 %) and the lowest percentages were observed with mother/father IVF cycles (33–35%) – Table 2.

The first predictive model (I) for the IP group was created for the variables of quality and number of oocytes. The 11 components were created using the PCA method (selected components are shown in Table 3).

Table 3. Selected components obtained with the PCA method for the I and III models

	Model I				Model III
	Comp1	Comp2	Comp4	Comp6	Comp1
# follicles	−0.37	−0.15	0.14	0.90	−0.41
# egg retrieved	−0.41	−0.12	0.13	−0.21	−0.48
M2 after ER	−0.35	−0.22	0.38	−0.25	−0.39
M1 after ER	−0.29	0.10	−0.11	−0.06	−0.18
GV after ER	−0.29	−0.11	−0.56	−0.06	−0.28
OTH after ER	−0.21	0.64	0.08	−0.02	−0.19
M2 at ICSI	−0.35	−0.22	0.38	−0.25	−0.39
M2* at ICSI	−0.22	0.04	−0.06	−0.09	−0.11
M1 at ICSI	−0.24	0.12	−0.15	0.01	−0.17
GV	−0.29	−0.11	−0.56	−0.07	−0.28
OTH	−0.20	0.64	0.09	0.02	−0.19

Next, the components were used as independent variables in the logistic regression analysis. The dependent variable was pregnancy HB, where the heart beat was detected during a scan of the uterus (USG). The predictive model I showed components 1, 2, 4 and 6, which had a significant effect on pregnancy (Table 4).

For that model – the percentage of pregnancies correctly classified was 61%. Sensitivity was only 54% and specificity 65%. The area under the

Table 4. Predictive models for pregnancy

Model I	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Component1	-.1590062	.037776	-4.21	0.000	-.2330459	-.0849665
Component2	-.1888785	.0698509	-2.70	0.007	-.3257837	-.0519732
Component4	.3080229	.0800911	3.85	0.000	.1510472	.4649985
Component6	.4023885	.1938359	2.08	0.038	.0224771	.7822998
_cons	-.646302	.0887914	-7.28	0.000	-.82033	-.472274
Model II						
Component1	.2384806	.0387992	6.15	0.000	.1624355	.3145256
Component3	.1940687	.0871819	2.23	0.026	.0231953	.364942
_cons	-.6371075	.0886235	-7.19	0.000	-.8108063	-.4634087
Model III						
Component1	-.4239164	.1874435	-2.26	0.024	-.7912989	-.0565338
_cons	1.363461	.336031	4.06	0.000	.7048525	2.02207
Model IV						
Component1	.4872743	.1805951	2.70	0.007	.1333145	.8412341
Component3	1.157432	.5380945	2.15	0.031	.1027866	2.212078
Component6	-1.220713	.5177525	-2.36	0.018	-2.235489	-.2059362
_cons	1.89315	.4941718	3.83	0.000	.9245915	2.861709
Model V						
Component9	-1.286289	.5557162	-2.31	0.021	-2.375473	-.1971055
_cons	-.7491325	.2440403	-3.07	0.002	-1.227443	-.2708223
Model VI						
Component11	-80.85404	32.9964	-2.45	0.014	-145.5258	-16.18229
group	-2.489083	1.095572	-2.27	0.023	-4.636366	-.3418014
_cons	4.616272	1.850286	2.49	0.013	.9897788	8.242766

ROC curve (ability of model to differentiate positive and negative results) was AUC=0.67.

Model I can be expressed by the following formula:

$$ln\frac{\pi(x)}{1-\pi(x)} = -0.64 - 0.16 \cdot comp1 - 0.19 \cdot comp2 + 0.31 \cdot comp4 + 0.4 \cdot comp6$$

where $\pi(x)$ means probability of pregnancy, and the components $(1, 2, \ldots)$ are the linear combination of the standardized variables with the coefficients shown in Table 3.

The second predictive model was also obtained for the IP group with variables of quality and number of embryos. PCA created 12 uncorrelated components (Table 5).

The logistic regression model (II) includes only two components (1 and 3) that affected pregnancy (Table 4). The ability of that model to predict pregnancy was 64%. Sensitivity was 55% and specificity 69%. The area under the ROC curve was AUC=0.68, similarly to model I.

Table 5. Selected components obtained with PCA analysis for the models II, IV, V and VI

	Model II		Model IV			Model V	Model VI
	Comp1	Comp3	Comp1	Comp3	Comp6	Comp9	Comp11
# eggs ICSI	0.39	0.01	0.39	0.02	−0.02	0.05	0.53
2PN	0.39	0.12	0.41	0.06	0.19	0.12	0.36
NEF 2PB	0.03	−0.60	0.01	−0.26	−0.07	0.06	0.02
NEF	0.09	−0.37	0.13	−0.43	−0.20	−0.06	−0.05
1,3,>3PN	0.16	−0.03	0.06	−0.20	−0.33	−0.09	−0.11
Dead	0.15	0.18	0.05	0.77	−0.46	−0.17	−0.11
clvd day3	0.40	0.09	0.40	0.04	0.19	0.11	−0.75
≥7 cell day3	0.38	0.07	0.37	0.17	0.12	−0.50	−0.01
Blasts day5	0.36	−0.07	0.36	0.09	0.09	−0.27	−0.01
Blasts day6	0.36	−0.05	0.33	−0.12	0.25	0.65	0.00
#ET	−0.00	0.65	−0.24	0.26	0.60	0.22	0.01
ET Day	0.29	−0.12	0.26	−0.05	−0.33	−0.36	0.01

Model II can be expressed by the following formula (components are described in Table 5):

$$ln\frac{\pi(x)}{1-\pi(x)} = -0.64 + 0.24 \cdot comp1 + 0.19 \cdot comp3$$

Model III was created for the variables of quality and number of oocytes in the group (D-IP) with oocyte donation. The 7 components (Table 3) were created using the PCA method. Only component 1 in model III significantly affected pregnancy (Table 4). Correctness of classification was 63%, sensitivity was 63%, specificity was 62% and the area under the ROC curve was AUC=0.7.

Model III can be expressed by the following formula (component 1 is described in Table 3):

$$ln\frac{\pi(x)}{1-\pi(x)} = 1.36 - 0.42 \cdot comp1$$

Model IV describes the effect of the quality of embryos on pregnancy in the group (D-IP), where the oocytes were retrieved from an anonymous egg donor. The PCA method created 12 uncorrelated components (Table 5). Model IV included three components – 1, 3 and 6 – that significantly affected pregnancy (Table 4). Correctness of classification for model IV was slightly

higher than for previous models and was 69%, sensitivity and specificity were 69%, and the area under the ROC curve was AUC=0.82.

Model IV can be expressed by the following formula (components described in Table 5):

$$ln\frac{\pi(x)}{1-\pi(x)} = 1.89 + 0.49 \cdot comp1 + 1.16 \cdot comp3 - 1.22 \cdot comp6$$

Genetic testing of the embryo was performed in 84 cycles before the embryo transfer in the group IP-PGD. The predictive model V was created with the variables of embryo quality. The PCA method created 12 components (Table 5). In model V, only component 9 significantly affected pregnancy (Table 4). Correctness of classification was 67%, sensitivity was 68%, specificity was 66% and the area under the ROC curve was AUC=0.68.

Model V can be expressed by the following formula (component 9 described in Table 5):

$$ln\frac{\pi(x)}{1-\pi(x)} = -0.75 - 1.29 \cdot comp9$$

The last obtained model (VI), demonstrates presence of pregnancy based on the quality of embryos in the group GC with the use of a gestational carrier (in all cases). PCA created 12 components (Table 5). In this model, component 11 and the Group variable (with the value of 1 for the patients from group D-P-GC and the value of 2 for the patients from group IP-GC – see Table 4) significantly affected pregnancy. It is the best model among the created ones. Correctness of classification was 80%, sensitivity was 83%, specificity was 75% and area under the ROC curve was AUC=0.88.

Model VI can be expressed by the following formula (component 11 described in Table 5):

$$ln\frac{\pi(x)}{1-\pi(x)} = 4.62 - 80.85 \cdot comp11 - 2.49 \cdot group$$

Conclusions

Application of the Principal Component Analysis method allowed models to predict pregnancy to be built. The basis for modeling was the linear combination of the standardized variables describing the quality of the retrieved oocytes and embryos. Models I and III predicted pregnancy in 61% and 63% of cases, respectively, based on the quality of oocytes. However,

correctness of classification for models II, IV, V and VI, which predicted pregnancy based on embryo quality, was higher: 64%, 69%, 67% and 80%, respectively. The best prognostic results for pregnancy were obtained in the gestational carrier group (80%) and in the group with egg donation (69%). Our models demonstrate that good quality oocytes (retrieved from a young, healthy donor) or healthy gestational carriers significantly increase the chances for pregnancy. The PCA and logistic regression methods are the appropriate methods to demonstrate this.

REFERENCES

Agarwal, S., Jacobs Jr., D. R., Vaidya, D. Sibley, Ch. T., Jorgensen, N. W., Rotter, J. I., Chen, Y.-D. I., et al. (2012). Metabolic Syndrome Derived from Principal Component Analysis and Incident Cardiovascular Events: The Multi Ethnic Study of Atherosclerosis (MESA) and Health, Aging, and Body Composition (Health ABC). *Cardiology Research and Practice*, 2012. DOI:10.1155/2012/919425.

Aguilera, A. M., Escabias, M., & Valderrama, M. J. (2006). Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics & Data Analysis*, 50(8), 1905–1924.

Akinsola, O. M., Nwagu, B. I., Orunmuyi, M., Iyeghe-Erakpotobor, G. T., Eze, E. D., Abanikannda, O. T. F., Onaadepo, O., Okuda, E. U., & Louis, U. (2014). Prediction of bodyweight from body measurements in rabbits using principal component analysis. *Annals of Biological Sciences*, 2(1), 1–6.

Belasco, E., Philips, B. U., & Gong, G. (2012). The Health Care Access Index as a Determinant of Delayed Cancer Detection Through Principal Component Analysis. In P. Sanguansat (Ed.), *Principal Component Analysis – Multidisciplinary Applications* (pp. 143–166). InTech. DOI:10.5772/38460.

Biffi, A., Anderson, Ch. D., Nalls, M. A., Rahman, R., Sonni, A., Cortellini, L., Rost, N. S., et al. (2010). Principal-Component Analysis for Assessment of Population Stratification in Mitochondrial Medical Genetics. *The American Journal of Human Genetics*, 86(6), 904–917.

Brzyski, P., Tobiasz-Adamczyk, B., & Knurowski T. (2012). Trafność i rzetelność skali GARS w populacji osób w starszym wieku w Polsce, *Gerontologia Polska*, 20(3), 109–117.

Czernyszewicz, E. (2008). Zastosowanie analizy głównych składowych do opisu konsumenckiej struktury jakości jabłek. *Żywność. Nauka. Technologia. Jakość*, 2(57), 119–127.

Daszykowski, M., & Walczak, B. (2008). Analiza czynników głównych i inne metody eksploracji danych. In D. Zuba & A. Parczewski (Eds.), *Chemometria w analityce*. Kraków: IES.

Daszykowski, M., Walczak, B., & Massart, D. L. (2001). Looking for natural patterns in data: Part 1. Density-based approach. *Chemometrics and Intelligent Laboratory Systems*, 56, 83–92.

Duch, W., Korbicz, J., Rutkowski, L., & Tadeusiewicz, R. (2000). *Biocybernetyka i Inżynieria Biomedyczna 2000. Tom 6: Sieci neuronowe*. Warszawa: Akademicka Oficyna Wydawnicza Exit.

Fisher, R., & MacKenzie, W. (1923). Studies in crop variation II. The manurial response of different potato varieties. *Journal of Agricultural Science*, 13, 311–320.

Furman-Haran, E., Shapiro Feinberg, M., Badikhi, D., Eyal, E., Zehavi, T., & Degani, H. (2014). Standardization of Radiological Evaluation of Dynamic Contrast Enhanced MRI: Application in Breast Cancer Diagnosis. *Technology in Cancer Research & Treatment*, 13(5), 445–454.

Gastinel, L. N. (2012). Principal Component Analysis in the Era of "Omisc" Data. In P. Sanguansat (Ed.), *Principal Component Analysis – Multidisciplinary Applications* (pp. 21–42). InTech. DOI:10.5772/37099.

Giuliani, A., & Benigni, R. (2000). Principal Component Analysis for Descriptive Epidemiology. In R. W. Brause & E. Hanisch (Eds.). *Medical Data Analysis. Lecture Notes in Computer Science*, 1933, 308–313.

Hladnik, A. (2013). Image compression and face recognition: two image processing applications of principal component analysis. *International Circular of Graphic Education and Research*, 6, 56–61.

Hoffmann, K., Schulze, M. B., Schienkiewitz, A., Nothlings, U., & Boeing, H. (2004). Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology. *American Journal of Epidemiology*, 159(10), 935–944.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441.

Kaur, G., Arora, A. S., & Jain, V. K. (2012). Multiple Linear Regression Model based on Principal Component Scores to Study the Relationship between Anthropometric Variables and BP Reactivity to Unsupported Back in Normotensive Post–graduate Females. *International conference: 1st, Energy and environment technologies and equipment. Advances in Environment, Biotechnology and Biomedicine* (pp. 373–377). Greece: WSEAS.

Kolasa-Więcek, A. (2012). Application of PCA in the analysis of parameters related to agricultural greenhouse gases emissions in Europe. *Journal of Research and Applications in Agricultural Engineering*, 57(1), 77–79.

Konieczna, L., & Lamparczyk, H. (2008). Wpływ płci na farmakokinetykę wybranych leków. *Zastosowania metod statystycznych w badaniach naukowych III* (pp. 299–310). Kraków, Polska: StatSoft Polska. Retreived from: http://www.statsoft.pl/portals/0/Downloads/Wplyw_plci.pdf.

Koter, S., & Wesołowska, K. (2003). Zastosowanie metody PCA do opisu wód naturalnych. *II Ogólnopolska Konferencja Naukowo-Techniczna "Aktualne zagadnienia w uzdatnianiu i dystrybucji wody"* (pp. 413–420). Szczyrk, Poland.

Latifoglu, F., Polat, K., Kara, S., & Gunes, S. (2008). Medical diagnosis of atherosclerosis from Carotid Artery Doppler Signals sing principal component analysis (PCA), k-NN based weighting pre-processing and Artificial Immune Recognition System (AIRS). *Journal of Biomedical Informatics*, 41(1), 15–23.

Ma, S. (2007). Principal Component Analysis in Linear Regression Survival Model with Microarray Data. *Journal of Data Science*, 5, 183–198.

Martens, H., & Næs, T. (1991). *Multivariate calibration.* Chichester: Jon Wiley & Sons.

Martis, R. J., Acharya U. R., & Min, L. Ch. (2013). ECG beat classification using PCA, LDA, ICA and Discrete Wavelet Transform. *Biomedical Signal Processing and Control*, 8(5), 437–448.

Milewska, A. J., Górska, U., Jankowska, D., Milewski, R., & Wołczyński, S. (2011). The use of the basket analysis in a research of the process of hospitalization in the gynecological ward. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 25(38), 83–98.

Milewska, A. J., Jankowska, D., Cwalina, U., Więsak, T., Morgan, A., & Milewski, R. (2013). Analyzing outcome of intrauterine insemination treatment by application of Cluster Analysis or Kohonen Neural Networks. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 35(48), 7–25.

Milewska, A. J., Jankowska, D., Górska, U., Milewski, R., & Wołczyński, S. (2012). Graphical representation of the relationships between qualitative variables concerning the process of hospitalization in the gynecological ward using correspondence analysis. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 29(42), 7–25.

Milewski, R., Jamiołkowski, J., Milewska, A. J., Domitrz, J., Szamatowicz, J., & Wołczyński, S. (2009). Prognosis of the IVF ICSI/ET procedure efficiency with the use of artificial neural networks among patients of the Department of Reproduction and Gynecological Endocrinology. *Ginekologia Polska*, 80(12), 900–906.

Milewski, R., Malinowski, P., Milewska, A. J., Czerniecki, J., Ziniewicz, P., & Wołczyński, S. (2011). Nearest neighbor concept in the study of IVF ICSI/ET treatment effectiveness. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 25(38), 49–57.

Milewski, R., Milewska, A. J., Czerniecki, J., Leśniewska, M., & Wołczyński, S. (2013a). Analysis of the demographic profile of patients treated for infertility using assisted reproductive +techniques in 2005–2010. *Ginekologia Polska*, 84(7), 609–614.

Milewski, R., Milewska, A. J., Domitrz, J., & Wołczyński, S. (2008). In vitro fertilization ICSI/ET in women over 40. *Przegląd Menopauzalny*, 7(2), 85–90.

Milewski, R., Milewska, A. J., Więsak, T., Morgan, A., (2013b). Comparison of artificial neural networks and logistic regression analysis in pregnancy prediction using in the in vitro fertilization treatment Networks. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 35(48), 39–48.

Mudrova, A., & Prochazka, A. (2005). *Principal Component Analysis in Image Processing*. Technical Computing Conference, Prague, Czech Republic.

Næs, T., Isaksson, T., Fearn, T., & Davies, T. (2002). *A user-friendly guide to multivariate calibration and classification*. Chichester UK: NIR Publications.

Nascimento, E. C. M., & Martins, J. B. L. (2012). Pharmacophoric Profile: Design of New Potential Drugs with PCA Analysis. In P. Sanguansat (Ed.), *Principal Component Analysis – Multidisciplinary Applications* (pp. 59–74). InTech. DOI:10.5772/37426.

Nowicki, J., Żylińska, A., & Kin, A. (2013). Zastosowanie metod statystycznych i graficznych w analizie zdeformowanych tektonicznie trylobitów z rodziny Ellipsocephalidae Matthew, 1887 z kambru Gór Świętokrzyskich. In M. Kędzierski & B. Kołodziej (Eds.), *XXII Konferencja Naukowa Sekcji Paleontologicznej Polskiego Towarzystwa Geologicznego "Aktualizm i antyaktualizm w paleontologii"* (pp. 38–39). Tyniec, Poland: Polskie Towarzystwo Geologiczne.

Pandey, P. K., Singh, Y., & Tripathi, S. (2011). Image Processing using Principle Component Analysis. *International Journal of Computer Applications*, 15(4), 37–40.

Panek, D. (2014). Ocena parametrów analizy akustycznej w detekcji patologii mowy. *Przegląd Elektrotechniczny*, R. 90(5), 126–129. DOI: 10.12915/pe.2014.05.29.

Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genetics*, 2(12), 2074–2093. DOI:10.1371/journal.pgen.0020190.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2, 559–572.

Petrisor, A. I., Ianos, I., Iurea, D., & Vaidianu, M. N. (2012). Applications of Principal Component Analysis integrated with GIS. *Procedia Environmental Sciences*, 14, 247–256.

Pushpa Rathi, G. V. P., & Palani, S. (2012). Brain Tumor MRI Image Classification with Feature Selection and Extraction using Linear Discriminant Analysis. *International Journal of Information Sciences & Techniques*, 2(4), 131–146.

Raskin, R., & Terry, H. (1988). A Principal-Components Analysis of the Narcissistic Personality Inventory and Further Evidence of Its Construct Validity. *Journal of Personality and Social Psychology*, 54(5), 890–902.

Raychaudhuri, S., Stuart, J. M., & Altman, R. (2000). Principal Component Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. *Pacyfic Symposium on Biocomputing*, 2000, 455–466.

Reverter, F., Vegas, E., & Oller, J. M. (2012). Kernel Methods for Dimensionality Reduction Applied to the "Omics" Data. In P. Sanguansat (Ed.), *Principal Component Analysis – Multidisciplinary Applications* (pp. 1–20). InTech. DOI:10.5772/37431.

Rymuza, K., & Radzka, E. (2013). Zastosowanie analiz wielowymiarowych do oceny jakości wody pitnej. *Nauka. Technologia. Jakość*, 6(91), 165–174.

Santo, R. do E. (2012). Principal Component Analysis applied to digital image compression. *Einstein (Sao Paulo)*, 10(2), 135–139.

Scholz, M., Schmidt, S., Loesgen, S., & Bickeböller, H. (1999). Analysis of principal component based quantitative phenotypes for alcoholism. *Genetic Epidemiology*, 17(l), 313–318.

Stuhler, E., & Merhof, D. (2012). Principal Component Analysis Applied to SPECT and PET Data of Dementia Patients – A Review. In P. Sanguansat (Ed.), *Principal Component Analysis – Multidisciplinary Applications* (pp. 167–186). InTech. DOI:10.5772/38010.

Suchacz, B., & Wesołowski, M. (2010). Relacje pomiędzy zawartością cynku, miedzi, ołowiu i niklu w wodnych ekstraktach z mieszanek ziołowych. *Bromatologia i Chemia Toksykologiczna*, 43(4), 485–492.

Szefer, P. (2003). Zastosowanie technik chemometrycznych w analitycznej ocenie próbek biologicznych i środowiskowych. In J. Namieśnik, W. Chrzanowski & P. Szpinek (Eds.), *Nowe Horyzonty i Wyzwania w Analityce i Monitoringu Środowiskowym* (pp. 599–629). Gdańsk, Poland: CEEAM.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using Multivariate Statistics*. Boston: Pearson.

Ukalska, J., Ukalski, K., Śmiałowski, T., & Mądry, W. (2008). Badanie zmienności i współzależności cech użytkowych w kolekcji roboczej pszenicy ozimej (*Triticum aestivum* L.) za pomocą metod wielowymiarowych. Część II. Analiza składowych głównych na podstawie macierzy korelacji fenotypowych i genotypowych. *Biuletyn Instytutu Hodowli i Aklimatyzacji Roślin*, 249, 45–57.

Webb, A. R. (2003). *Statistical Pattern Recognition*. Wiley.

Varraso, R., Garcia-Aymerich, J., Monier, F., Le Moual, N., De Batlle, J., Miranda, G., Pison, Ch., Romieu, I., Kauffmann, F., & Maccario, J. (2012). Assessment of dietary patterns in nutritional epidemiology: principal component analysis compared with confirmatory factor analysis. *The American Journal of Clinical Nutrition*, 96(5), 1079–1092.

Zuendorf, G., Kerrouche, N., Herholz, K., & Baron, J. C. (2003). Efficient principal component analysis for multivariate 3D voxel-based mapping of brain functional imaging data sets as applied to FDG-PET and normal aging. *Human Brain Mapping* 18(1), 13–21.