STUDIES IN LOGIC, GRAMMAR AND RHETORIC

UNDER THE AUSPICES
OF THE POLISH ASSOCIATION FOR LOGIC AND PHILOSOPHY OF SCIENCE

Logical, Statistical and Computer Methods in Medicine

Edited by ROBERT MILEWSKI

35(48) UNIVERSITY OF BIALYSTOK 2013

LOGICAL, STATISTICAL AND COMPUTER METHODS IN MEDICINE

Series: STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35(48)

Under the Auspices of the Polish Association
for Logic and Philosophy of Science

LOGICAL, STATISTICAL AND COMPUTER METHODS IN MEDICINE

edited by
Robert Milewski

University of Białystok
Białystok 2013

Series: STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35(48)
http://logika.uwb.edu.pl/studies/

Subseries issued by the Chair of Logic, Informatics and Philosophy of Science
Faculty of History and Sociology, University of Bialystok

This issue has been financed by the Medical University of Bialystok

Cover design: Krzysztof Tur

ISBN 978–83–7431–392–6
ISSN 0860–150X

WYDAWNICTWO UNIWERSYTETU W BIAŁYMSTOKU
15-097 Białystok, ul. Marii Skłodowskiej-Curie 14, tel. 857457120
http://wydawnictwo.uwb.edu.pl, e-mail: ac-dw@uwb.edu.pl

Druk i oprawa: „QUICK-DRUK" s.c., Łódź

CONTENTS

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0041

Analyzing Outcomes of Intrauterine Insemination Treatment by Application of Cluster Analysis or Kohonen Neural Networks

Anna Justyna Milewska[1], Dorota Jankowska[1], Urszula Cwalina[1], Teresa Więsak[2], Dorota Citko[1], Allen Morgan[3], Robert Milewski[1]

[1] Deartment 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, Olsztyn, Poland

[3] Shore Institute for Reproductive Medicine, Lakewood, USA

Abstract. Intrauterine insemination (IUI) is one of many treatments provided to infertility patients. Many factors such as, but not limited to, quality of semen, the age of a woman, and reproductive hormone levels contribute to infertility. Therefore, the aim of our study is to establish a statistical probability concerning the prediction of which groups of patients have a very good or poor prognosis for pregnancy after IUI insemination. For that purpose, we compare the results of two analyses: Cluster Analysis and Kohonen Neural Networks. The k-means algorithm from the clustering methods was the best to use for selecting patients with a good prognosis but the Kohonen Neural Networks was better for selecting groups of patients with the lowest chances for pregnancy.

Introduction

Infertility refers to an inability to conceive after having regular unprotected sex for a period of at least one year (Radwan J., 2011). More and more often women are experiencing difficulty becoming pregnant. The female, male or both partners can contribute to the couple's infertility. It has been estimated that in approximately 20–30% of couples, both partners suffer from infertility (Kurzawa et al., 2010). A study conducted by the World Health Organization showed that males might contribute in 50% to these couples' infertility (Radwan et al., 2011).

Usually, at the beginning of any treatment, the male and female are evaluated to establish the reason(s) of infertility. Many tests need to be performed to establish a diagnosis for the couple. During a basic evaluation, the potency of the fallopian tubes and uterus, and concentrations of repro-

ISBN 978-83-7431-392-6 ISSN 0860-150X

ductive and non-reproductive hormones in the blood need to be determined. The quality of male gametes (semen analysis) and ovulatory status of the woman are also checked. Superovulation therapy is prescribed when an ovulatory problem exists. The follicular growth is controlled by the injection of different hormones and measurement of estradiol level in the blood. 75–85% of women will ovulate after such a treatment (Pierzyński, 2011) and half of them will become pregnant within the first 3–4 months if no major infertility problems exist on both sides (female – tubal factor or endometriosis and/or male – abnormal semen parameters) (Pierzyński, 2011). In some circumstances, intrauterine inseminations (IUI) or In Vitro Fertilization (IVF) procedures need to be performed.

The first IUIs were performed at the end of the XVIII century (Horák, 2004; Radwan P., 2011). There are many types of insemination but the most common is intrauterine insemination. It has been established that after 6 IUI inseminations 30–40% of women will become pregnant. However, efficiency of the IUI is in the range of 6–18% per insemination and depends on the type of diagnosis (Radwan P., 2011). IUI is prescribed for the following medical conditions: cervical factors, endometriosis, and male and/or immunological or idiopathic factors (Derwich et al., 2008; Radwan P., 2011). The collected semen samples are purified in the laboratory before IUI and a minimum 5 million motile spermatozoa, usually in the 0.5 ml volume, are deposited in the uterus (Tkaczuk-Włach et al., 2006; Wainer et al., 2004).

The fertility of women drastically decreases after they are forty years old. (Milewski et al., 2008, 2013). Usually a few IUIs are recommended but when the woman is older (around 40 plus years old) IVF is more often suggested as the first choice of treatment (Pierzyński, 2011; Radwan P., 2011). The best outcome with IUI treatment might be obtained: when the woman is younger than 30 years old, superovulation is carried out with gonadotropins, two follicles larger than 16 mm in diameter are present on the ovaries, endometrium thickness is greater than 9 mm and there are more than 5 mln/ml of motile spermatozoa with the forward progression class A and B (Radwan P., 2011).

To be successful in providing treatment to patients, the medical staff has to be knowledgeable, experienced and equipped with adequate tools to make proper diagnoses. The correct diagnosis is essential for designing an efficient treatment plan and analyzing collected medical information appropriately. Classic statistical analyses are lagging behind in the quality of results in such situations. Therefore, other more efficient statistical analysis methods, such as Data Mining, should be employed because they give

an opportunity for the creation of suitable predictable algorithms. Medical databases are becoming larger and larger. This allows for accommodation of more sophisticated statistical analyses, which lead to the creation of precise treatment options. For example, to predict efficacy of the IVF with embryo transfer procedure – the Artificial Neural Networks (ANNs) (Milewski et al., 2009), Correspondence (Milewska et al., 2012) or Basket (Milewska et al., 2011) or types of analysis with the application of feature selection algorithms to reduce original dimensionality of the original data set (Milewski et al., 2011, 2012) are applicable. Therefore, the aim of our study is to provide a statistical probability of successful IUI treatment outcomes by grouping patients appropriately (with good or poor prognosis).

Material and Methods

The medical information without personal identifiers of the 825 IUI cycles performed at the Shore Institute for Reproductive Medicine, Lakewood, USA was used in these statistical analyses. Segmentation methods such as Cluster Analysis or Kohonen Neural Networks were applied because they provide an option for uniformly grouping data to determine/estimate percentages of successful pregnancies. The quantitative variables: semen parameters, hormone levels, age of the female (Table 1) and qualitative variables: reason for infertility (Table 2) were analyzed. The Statistica Data Miner + QC 10.0 (StatSoft, Tulsa, OK, USA) software was used. Statistical significance was determined at the $p < 0.05$ level.

Table 1. List of quantitative variables

Quantitative variables	median	min	max
semen – number of motile sperm	17	0.02	614.8
semen vol. – volume of semen – ml	3	0.2	12
sperm ct. – concentration sperm – M/ml	47	1.4	598
sperm mot. – motility sperm – %	66	2	100
age – age of women	35	23	46
no. of follicles – total number of ovulatory follicles per cycle	7	1	46
endometrium thickness at HCG injection – mm	10	5	24
E_2 at HCG – estradiol level at HCG injection – pg/ml	470	89.3	2624
P_4 at HCG – progesterone level at HCG injection – ng/ml	1	0.2	8.8

Table 2. List of qualitative variables used in analysis

Qualitative variables	n	%
clinical pregnancy (variable result)	98	11.9%
infertility reason – idiopathic factor	439	53.0%
infertility reason – AMA – Advanced Maternal Age	126	15.3%
infertility reason – endomertiosis	62	7.5%
infertility reason – MF – Male Factor	41	5.0%
infertility reason – ovulatory factor	67	8.1%
infertility reason – PCOS – polycystic ovary syndrome	36	4.4%
infertility reason – tubal factor	29	3.5%
infertility reason – secondary infertility	17	2.1%
infertility reason – uterine factor	23	2.8%
stimulation drug – CC – clomiphene citrate	387	46.9%
stimulation drug – I – gonadotropins	352	42.7%
stimulation drug – TX – tamoxifene	86	10.4%

Cluster Analysis

Cluster analysis refers to the methods of organizing data according to certain structures. Basically, it is a process of identifying groups of objects similar to each other in some characteristics but distinctively different from elements in other groups. This indirect technique is ranked among unsupervised learning methods. The variables that decisively determine an observational group are not defined. Cluster Analysis is applicable in exploratory data mining because it allows us to reduce the sizes of enormous databases and to organize information for easy access. It also allows us to discover existing relationships, for example, the relationship between a patient's biochemical parameters and occurrence of illness (McLachlan, 1992).

Algorithm Cluster Analysis groups the objects that are more similar to each other but differ in some way from elements in other clusters. Either cases or variables, which characterize a study group, can be grouped into classes. The distances among them are determined to estimate similarity and/or dissimilarity between the objects $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n)$. The following metrics can be used:

- Euclidean distance – the most natural geometric distance in a multidimensional space.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Squared Euclidean distance – usually used to provide higher weight to distant objects.

$$d(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2$$

- Chebyshev distance – useful in demonstrating differences among the objects in one dimension that differentiate elements the most.

$$d(x,y) = \max_{i=1,\ldots,n}|x_i - y_i|$$

- Manhattan distance (City block) – a sum of differences between the objects in all dimensions (in this metric sphere it is the surface area of the cube).

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

- Exponential (Power) distance – allows the weights that are placed on the differences between objects in individual dimensions to be guided.

$$d(x,y) = \Big(\sum_{i=1}^{n}|x_i - y_i|^a\Big)^{\frac{1}{b}}$$

where a, b are the parameters established by the researcher

- Percent disagreement – used in situations where variables are qualitative.

$$d(x,y) = \frac{\text{number of pair of variables such as } x_i \neq y_i}{\text{total number of pair of variables}}$$

$i = 1, 2, \ldots, n$

Groups of algorithms called the Cluster Analysis can be divided into hierarchical and nonhierarchical subgroups. Among the hierarchical methods are the agglomerative and divisive procedures. The agglomerative algorithm is called a "bottom-up" approach. Basically, it assumes that initially each object is a separate cluster and new clusters are formed by combining the nearest objects and groups, which were joined earlier. Contrary to the previously described bottom up method, the top down approach or divisive method relies on the gradual splitting of clustered content into single elements (Stanisz, 2007). In these methods there is no need to define the number of clusters for the analyzed objects up front but an enormous computer processing power is essential. The applications of the agglomerative analyses are more frequently used and run according to the algorithm presented in diagram form in Figure 1 (Timm, 2002).

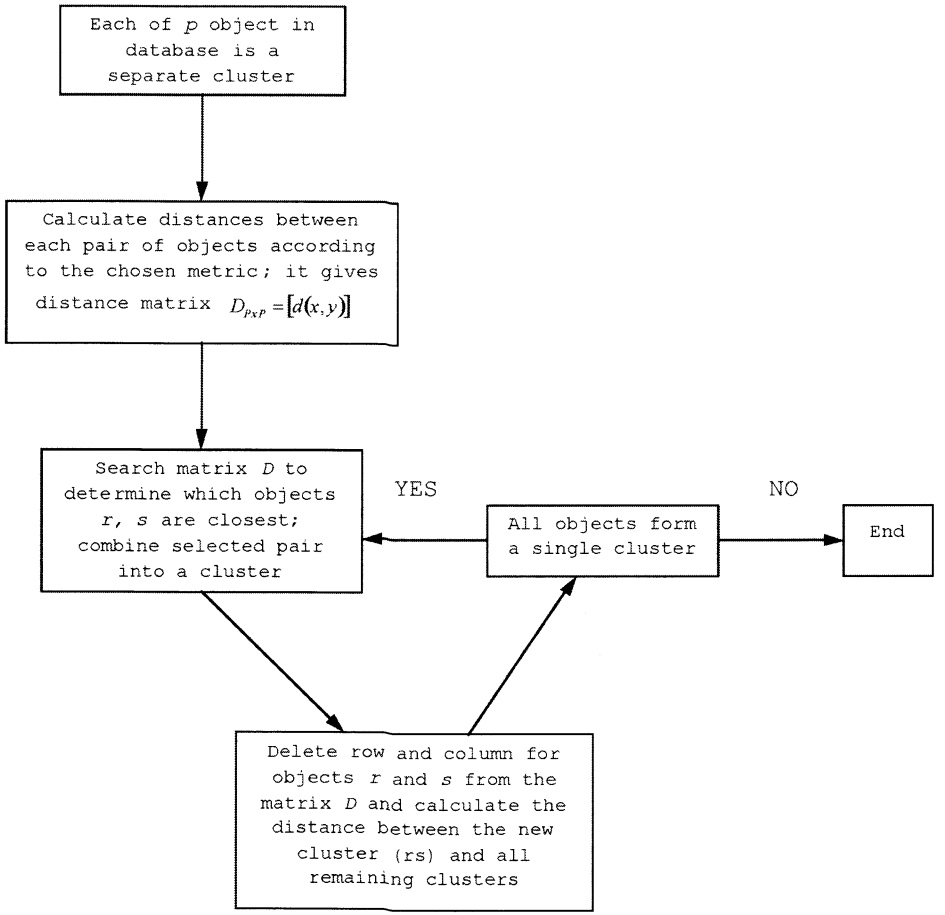


Figure 1. The flowchart of the agglomerative hierarchical clustering procedure

The method of defining distances between the clusters is the major contributing factor to the quality of classification and subsequently the quality of the whole analysis. The following techniques are applied very often (Stanisz, 2007):

- Single Linkage Clustering – also called the Nearest Neighbor Method; it is a distance between the analyzed clusters that is equivalent to the distance between the two nearest objects from two different clusters.
- Complete Linkage Clustering (Farthest Neighbor) – the distance between the clusters is defined as a distance between the two farthest objects in different clusters. This method is in contrast with the single linkage clustering method.

- Average Link Method – UPGMA (unweighted pair-group method using arithmetic averages); distance between clusters is determined as an average of the distances between all pairs of objects from two different clusters.
- Weighted Average Link Method – WPGMA (weighted pair-group method using arithmetic averages); this algorithm is recommended in situations where clusters possibly differ in number from each other. This method is more advanced than the previous one because it uses weights related to the number of elements in clusters.
- Centroid Method – UPGMC (unweighted pair-group method using the centroid average); distance between the clusters is estimated as a distance between their centers of mass.
- Weighted Centroid Method (Median Linkage) – WPGMC (weighted pair-group method using centroid average); determination of the distance is boosted with the weight that accounts for the disproportion in numbers of elements between clusters.
- Ward's Method (Incremental Sum of Squares) – this algorithm minimizes variability within the cluster; among all possible connections of pairs of objects, the one that characterizes the minimal variability is chosen – so this method minimizes the square deviations sum inside the cluster.

The vertical or horizontal tree diagram (dendrogram) is a graphic presentation of hierarchical clustering method results. Figure 2 represents a sample of such a dendrogram.

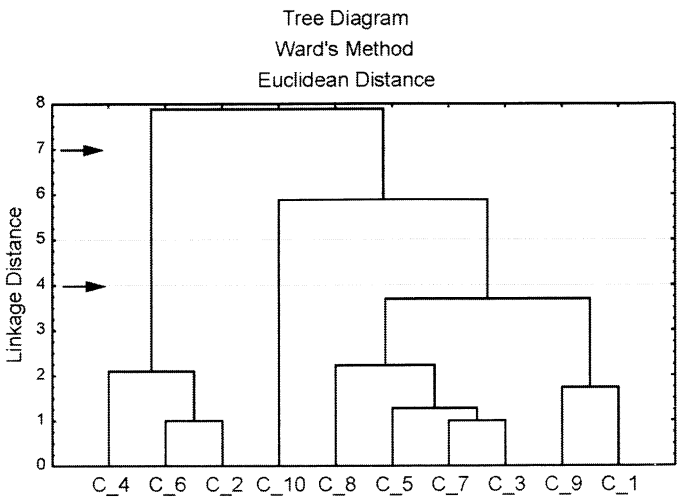


Figure 2. An example of a dendrogram, created as a result of Cluster Analysis using Ward's method

The objects are on OX axes that were initially assigned as single clusters. The most similar objects are linked by gradually diminished criteria of similarity (Stanisz, 2007). Subsequently, the clusters are created by linking more and more different objects. Finally, one cluster is created. The distances at which respective elements were merged into a single new cluster are on the OY axis. The dendrogram allows us to make a decision on the number of groups to be analyzed by cutting the tree diagram at the appropriate height.

The other Cluster Analysis methods are the nonhierarchical analyses, such as the k-means method or Expectation Maximization Method (EM). The principle of these methods is to disperse a group of objects into a known number of separate clusters in such a way that none of those groups are not subgroups of others. In the k-means method, the number of the clusters k has to be determined by the statistician at the beginning of the analysis (for example based on his/her intuition and experience). However, performing v-fold cross-validation is recommended to optimize the analysis. The algorithm divides objects in the study groups into more and more segments while at the same time checking the precision of division for each of them. The precision criterion for the classification is the average of the distances of each object from the center of its own cluster. The number of clusters is chosen such that the increase of precision in divisions is still very distinct. Before conducting analyses using the k-means method, it is necessary to define how the first cluster centers will be chosen. Usually, they are the first k objects. The cluster centers can be chosen using the principle of maximizing initial distances between clusters or by sorting all the distances between objects choosing elements with constant intervals as a center. After setting up initial assumptions, the algorithm runs according the diagram presented in Figure 3. The core of performance k-means is to move objects between clusters to minimize variation inside segments and to maximize variation between the clusters. This algorithm runs towards maximization of the significance of the F-test results in the analysis of variance. The statistical F value in a particular dimension decides on the role of particular variable during creation of clusters (Stanisz, 2007).

The Expectation Maximization Method (EM) is needed at the beginning of the analysis to define the number of segments for the initial data collection. Segregation of objects into separate k clusters is performed based on the principle that distribution of each analyzed variable is a mixture of k distributions. The basis for this method is first, to determine parameters of each component of distribution (such as mean or standard deviation),

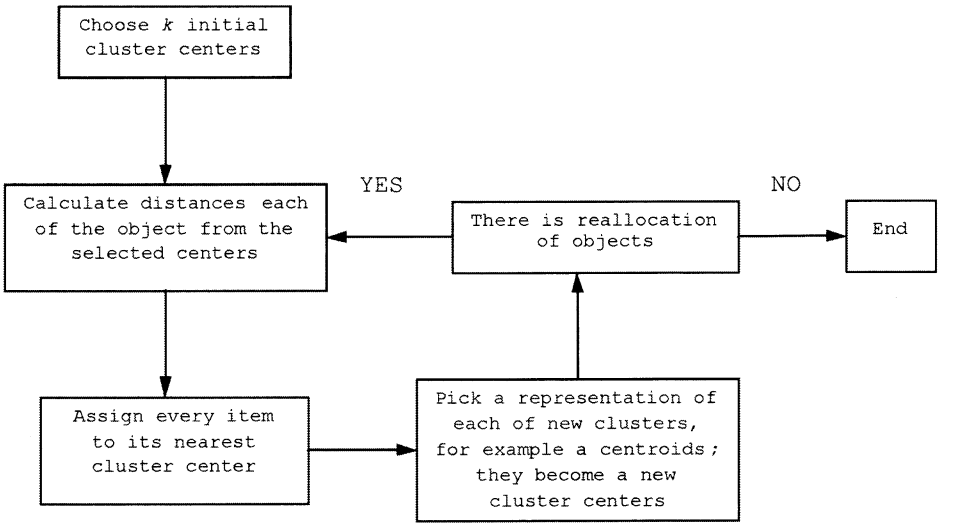


Figure 3. The flowchart of k-means method clustering

then the probability of assigning objects to those particular created clusters is determined. Finally, the objects are assigned to the cluster with their highest probability of belonging to that cluster.

Cluster Analysis is primarily used in analyzing data in the areas of biology, medicine and/or bioinformatics (Xu et al., 2010). The segmentation of data has a wide application and is very useful in: classification of plants or animals, psychological studies to select personalities (Sava et al., 2011), artificial intelligence studies, analyzing results of tomographic studies (distinction of tissue types and blood in the three-dimensional pictures (DeLapaz et al., 1990)) or in computer-aided diagnoses that assist doctors in the interpretation of medical images. Additionally, Cluster Analysis is applied to analyze microarray results (Shannon et al., 2003) because it allows one to determine groups of genes with similar patterns of expression or to establish similarity in the genotypes (Eisen et al., 1998).

Kohonen Neural Networks

Self-Organizing Feature Maps, also called Kohonen Neural Networks (Kohonen, 1982), are one of the basic self-organizing networks. The Kohonen Neural Networks are an example of unsupervised learning. This means that data are entered into a database without an established earlier pattern. It is similar to the way the human brain functions. Patterns are cre-

ated during the learning process. Such an approach is very useful in situations where new relationships that are not detectable using traditional statistical methods need to be found. Kohonen Neural Networks are able to demonstrate new output patterns, which can be identified with earlier unrecognized relationships. Thus, it allows statisticians to better understand data and, subsequently, to be able to apply different tools in further analysis. This method groups cases or creates grouped cases that are quite similar to each other or similar in their characteristics. At the same time, the groups should be as different as possible from each other. Kohonen Networks is a competitive learning method where the winning neuron (the most similar to the input vector) is chosen in competition with other neurons.

The Kohonen Network is built from two layers: an input layer and an output layer that is built out of neurons. The basic algorithm of Kohonen Networks runs in an interactive way. At the beginning, the nodes are chosen randomly and afterward they go through multiple runs according to the plan in Figure 4.

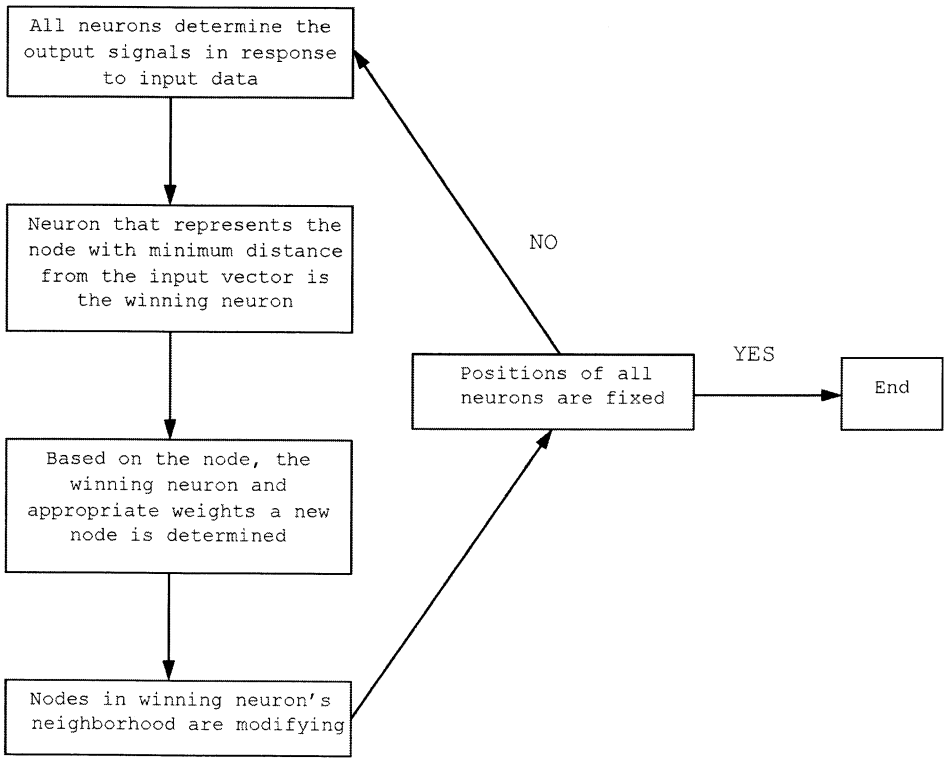


Figure 4. Scheme of Kohonen algorithm

During algorithm runs, the learning coefficient diminishes, affecting vector weights. At the beginning, the changes are usually large but with time they become smaller and smaller. The algorithm tests successive input factors and assigns the most appropriate nodes to them. Next, each node in the neighborhood is modified to resemble the element of the training set. At first an algorithm creates an approximate topological map, which at the end shows neurons that correspond with small clusters. Neurons in Kohonen Network are not connected with each other but they are presented as two-dimensional grids of nodes. This makes interpretation easy and/or allows one to observe similarities between the clusters with ease.

The Kohonen network is usually a one-way network and each neuron is connected to all elements of input vector X. The process starts from the normalization of the N-dimension input vector according to the equation below.

$$x_i' = \frac{x_i}{\sqrt{\sum_{y=1}^{N}(x_y)^2}}$$

After entering the input vector, the neurons compete with each other. The winning neuron w_w is in the smallest distance to X, so it complies with the following equation:

$$d(x, w_w) = \max_{1 \leq i \leq n} d(x, w_i)$$

Where weight vector is:

$$w_i = \begin{bmatrix} w_{i1} \\ w_{i2} \\ \vdots \\ w_{iN} \end{bmatrix}$$

$d(x, w)$ is a distance between the vectors x and w according to the selected metric. The Kohonen algorithm uses the concept of topological neighborhood neurons w_w. In the content of the neighborhood (with diminishing radius in time) neurons surrounding w_w are included. The winning neuron and neighbor neurons are subjected to adaptation according to Kohonen's formula (Kohonen, 1995):

$$w_i(n+1) = w_i(n) + \eta_i(n)[x - w_i(n)]$$

where η is coefficient factor i-of that neuron in neighborhood $Sw(n)$ in k-time. The value of that coefficient factor decreases with the distance to the winner. The basic Kohonen learning algorithm is:

$$w_i(n+1) = w_i(n) + \eta G(i, x)[x - w_i(n)]$$

Kohonen (Kohonen, 1995) proposed two types of neighborhood:

- rectangular (based on Euclid metric)

$$G(i,x) = \begin{cases} 1 & \text{dla } d(i,w) \leq \lambda \\ 0 & \text{dla } d(i,w) > \lambda \end{cases}$$

- Gaussian

$$G(i,x) = \exp\left(-\frac{d^2(i,w)}{2\lambda^2}\right)$$

Kohonen Neural Networks have a wide application where grouping objects into clusters with similar characteristics allows industrial or diagnostic processes to be improved. An interesting example of an application of Kohonen Networks is attempt the prognosis of daily usage of water (Licznar et al., 2006). This method might be used in marketing (Migut, 2009), as grouping clients with similar habits and preferences plays a key role in designing a marketing strategy. Furthermore, Kohonen Networks are useful in analyzing medical pictures (Ahmed et al., 1997).

Results

The 5 clusters (statistically different $p = 0.01$ in terms of pregnancy percentage) were outlined using k-means. The most efficient treatment was in cluster IV with a more than 27% pregnancy rate. However, the lowest pregnancy rate (only 7%) was in cluster II. In the other clusters the pregnancy rates were close to the mean value of allover analyzed cycles and were in the range of 10–13%. Moreover, in cluster IV, 59% of women were diagnosed with PCOS and 23% with endometriosis, and 7% of men were diagnosed with male factor. There were no cycles with idiopathic and tubal factors. In spite of male factor presence, all the semen parameters were above the median and 61% of patients were stimulated with gonadotropins. The median (30.5 years) for the age was the lowest among the studied clusters but the endometrium thickness was the highest – 11 mm (Table 3).

The lowest pregnancy success was in cluster II, probably due to fact that 84% of women were diagnosed with advanced maternal age (median for the age was 42 years old). The male factor diagnosis was not present in this cluster. However, the sperm concentration and motility were similar to those in cluster IV. Sixty-six percent of patients were stimulated with gonadotropins and the median for the estradiol (E_2) was 560 pg/ml. It was the highest value in comparison to other groups/clusters.

Table 3. Clusters characteristics obtained with k-means method

clusters	I	II	III	IV	V
frequency	249	114	72	44	97
clinical pregnancy	12.9%	7.0%	11.1%	27.3%	10.3%
Qualitative variables (percent)					
idiopathic factor	90.4%	0.9%	87.5%	–	–
AMA	–	84.2%	1.4%	–	2.1%
endometriosis	4.4%	4.4%	4.2%	22.7%	16.5%
MF	3.2%	–	–	6.8%	13.4%
ovulatory factor	–	–	–	9.1%	57.7%
PCOS	–	–	–	59.1%	–
tubal factor	0.8%	6.1%	1.4%	–	11.3%
secondary infertility	–	2.6%	2.8%	18.2%	–
uterine factor	–	3.5%	4.2%	2.3%	4.1%
CC	30.9%	23.7%	20.8%	15.9%	69.1%
I	58.2%	65.8%	73.6%	61.4%	23.7%
TX	10.8%	10.5%	5.6%	22.7%	7.2%
Quantitative variables (median)					
sperm no.	15.3	17.0	43.7	28.9	11.4
sperm vol.	2.8	2.9	2.2	3.0	3.0
sperm ct.	42.2	48.5	120.5	48.8	33.2
sperm mot.	60.0	73.5	86.0	75.0	60.0
age	34.0	42.0	34.0	30.5	31.0
no. of follicles	6.0	6.0	33.5	8.5	6.0
endom. thickness	10.0	10.0	9.0	11.0	9.0
E_2 at HCG	463.0	560.0	491.5	467.0	415.0
P_4 at HCG	0.9	0.9	0.8	0.9	0.9

The application of the hierarchical agglomerative clustering method produced the most optimal separation of objects into 5 clusters (Figure 5, Table 4). The pregnancy rate (%) among the clusters was not statistically different and was in the range of 9–15%. In cluster III, the pregnancy rate was the highest (15%) and medians of other characteristics were higher than in other clusters: age of women (36 years of age), endometrium thickness (11 mm), ovulatory follicle numbers (12), and estradiol level in the blood (1316 pg/ml). However, the medians for sperm concentration and motility were lower in comparison to other clusters. Additionally, there were almost 41% of women with idiopathic factors and more than 20% with advanced maternal age diagnosis. Cluster V was characterized by the lowest

Table 4. Clusters characteristics obtained with agglomerative method

clusters	I	II	III	IV	V
frequency	151	133	59	155	78
clinical pregnancy	10.6%	14.3%	15.3%	12.3%	9.0%
Qualitative variables (percent)					
idiopathic factor	50.3%	46.6%	40.7%	53.5%	56.4%
AMA	14.6%	17.3%	20.3%	16.8%	20.5%
endometriosis	9.3%	9.8%	5.1%	9.7%	–
MF	6.0%	3.8%	1.7%	4.5%	2.6%
ovulatory factor	12.6%	15.0%	2.4%	7.7%	9.0%
PCOS	5.3%	3.8%	11.9%	2.6%	2.6%
tubal factor	2.0%	3.8%	6.8%	2.6%	6.4%
secondary infertility	0.7%	2.3%	5.1%	1.3%	5.1%
uterine factor	2.0%	0.8%	5.1%	3.2%	–
CC	44.4%	33.8%	13.6%	34.8%	24.4%
I	27.2%	55.6%	86.4%	63.2%	75.6%
TX	28.5%	10.5%	–	1.9%	–
Quantitative variables (median)					
sperm no.	15.5	26.1	16.0	19.8	14.0
sperm vol.	2.5	2.6	2.8	2.7	2.8
sperm ct.	44.6	49.2	41.0	49.9	59.1
sperm mot.	66.0	67.0	63.0	66.0	70.0
age	33.0	35.0	36.0	35.0	35.0
no. of follicles	6.0	6.0	12.0	7.0	9.0
endom. thickness	10.0	9.2	11.0	9.0	10.0
E_2 at HCG	238.0	379.0	1316.0	590.0	890.0
P_4 at HCG	0.8	0.8	1.2	0.9	1.0

pregnancy rate (only 9%). Cluster V contained women with idiopathic factors (56%) and with advanced maternal age diagnosis (20%). For 76% of patients, ovulation induction was with gonadotropins. The median age was 35 years. The sperm concentration and motility were the highest in comparison to the other clusters.

After running multiple analyses, the most optimal network that contained 9 clusters was chosen (Figure 6, Table 5). The pregnancy rate was in the range of 9–18% and there were no statistical differences between the clusters. In cluster IV, the pregnancy rate was the highest (18.2%) and 59% of women were with secondary infertility (they were all of the women with secondary infertility), the male factor contributed in 41% of

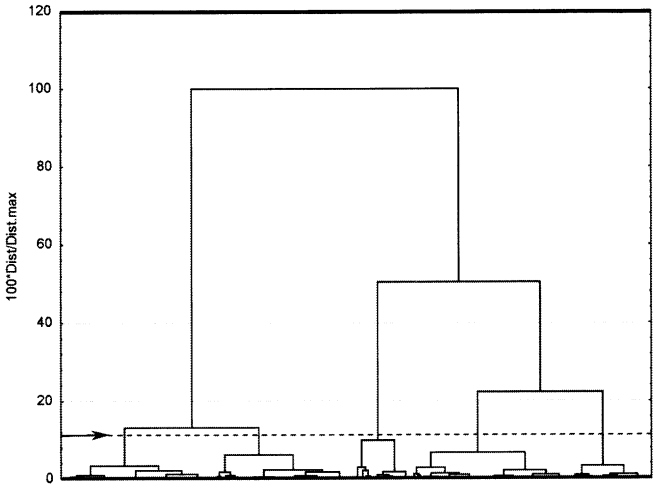


Figure 5. Dendrogram presenting agglomeration of cases

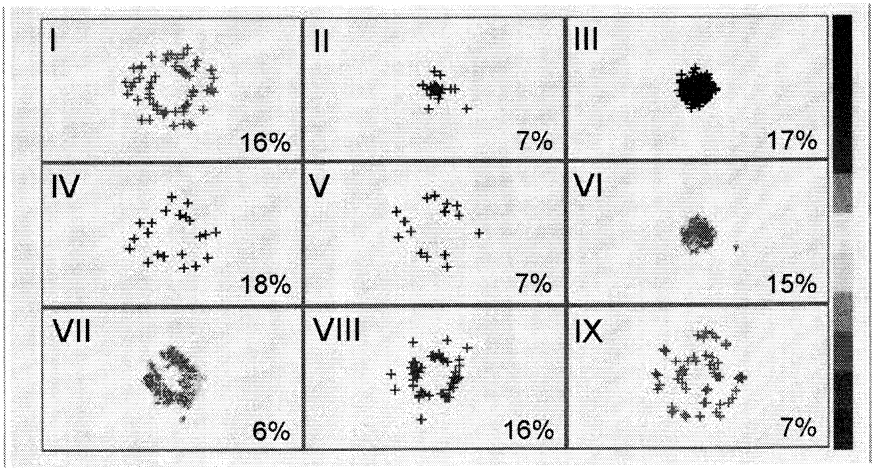


Figure 6. Clusters obtained with Kohonen Networks

the cases (the median for sperm concentration was 45 million/ml and for motility was 72%), the median age of female was 32 years old, 77% of patients were stimulated with gonadotropins and the median for the number of ovulatory follicles/oocytes was 11. The lowest pregnancy rate (6.2%) was in cluster VII with the AMA as a major infertility diagnosis. The median for the female age was 42 years of age and it was the highest among the studied clusters. The median number of ovulatory follicles was 5 after stimulation.

Table 5. Cluster characteristics obtained with Kohonen Networks

clusters	I	II	III	IV	V	VI	VII	VIII	IX
frequency	56	28	169	22	15	92	95	45	54
c. pregnancy	16.1%	7.1%	16.6%	18.2%	6.7%	15.2%	6.2%	15.6%	7.4%
Qualitative variables (percent)									
idiopathic f.	–	100%	100%	–	–	100%	–	–	–
AMA	–	3.6%	–	–	–	1.1%	100%	–	3.7%
endometriosis	–	–	–	–	–	–	–	100%	–
MF	–	–	–	40.9%	–	–	–	8.9%	20.4%
ovulatory f.	48.2%	–	–	–	–	–	–	–	61.1%
PCOS	26.8%	–	–	–	60%	–	–	4.4%	–
tubal factor	17.8%	–	–	–	–	–	–	–	20.4%
secondary i.	–	–	–	59%	–	–	–	–	–
uterine f.	8.9%	–	–	–	40%	–	1.1%	–	–
CC	–	–	–	–	60%	100%	28%	35.6%	90.7%
I	100%	–	100%	77.3%	–	–	61%	51.1%	–
TX	–	100%	–	22.7%	40%	–	10.5%	13.3%	9.3%
Quantitative variables (median)									
sperm no.	17.1	28.4	23.2	25.6	25.3	17.4	15.8	12.8	8.9
sperm vol.	3.0	2.7	2.5	2.0	2.5	2.5	2.8	2.7	3.0
sperm ct.	42.0	60.6	68.0	45.4	51.0	50.3	50.0	29.9	32.7
sperm mot.	72.0	77.5	65.0	72.5	72.0	69.0	72.0	59.0	55.0
age	32.0	33.0	34.0	32.0	29.0	34.0	42.0	33.0	33.5
no. of follicles	12.0	4.0	9.0	10.5	5.0	5.0	5.0	6.0	6.0
endom. thick.	11.0	10.0	10.0	10.5	10.0	9.0	10.0	10.0	8.2
E_2 at HCG	653.0	249.0	605.0	578.0	204.0	390.0	532.0	406.0	361.5
P_4 at HCG	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.9

The idiopathic diagnosis was present in clusters II, III and VI; the highest (16.6% and 15.2%) pregnancy rates were in clusters III and VI, respectively, but in cluster II it was 7.1%. Generally, the medication used for stimulation was different among the clusters; in cluster II all women were stimulated with Tamoxifene (TX), whereas in cluster III they were stimulated with gonadotropins and with clomiphene citrate (CC) in cluster VI. The median ages of females in all of these clusters were quite similar (approx. 33–34 years of age).

According to the the Kohonen algorithm of creation of Neural Networks, the neighboring clusters should be similar to each other. Figure 6 demonstrates the similarities between clusters I and IV in terms of pregnancy rate as well as the significant differences between clusters I and II.

Discussion and Conclusions

The k-means method seems to be the best, in comparison to the other two segmentation methods, in demonstrating the percentage of pregnancies achieved in IUI treatment. The pregnancy rate for the analyzed 825 cycles was 11.9%. However, the cluster with the highest pregnancy rate (27.3%) included the young women with the best chances for successful outcome of IUI treatment and all of the women with PCOS. Only the k-means method was able to demonstrate statistical differences in the pregnancy rates between the studied clusters.

In contrary to the k-means method, the agglomerative method was not able to demonstrate a high pregnancy rate in any of the studied clusters. The clusters were quite similar to each other; therefore pregnancy rates were comparable (9–15%) among them. The application of Kohonen Neural Networks into analysis of our data also did not produce the anticipated results. The internal structure of clusters was more varied here than in other methods. It is evident that Kohonen Neural Networks grouped predominantly qualitative variables. The highest pregnancy rate was 18.2% for that method. It is about 9% lower than using the k-means method. However, the Kohonen Neural Networks was able to detect the lowest chances for pregnancy and it was at the level of 6%.

To conclude – the k-means algorithm from the clustering methods was the best method for selecting patients with good prognosis but the Kohonen Neural Networks was better in selecting groups of patients with the lowest chances for achieving pregnancy.

REFERENCES

Ahmed, M. N., & Farag, A. A. (1997). Two-stage neural network for volume segmentation of medical images. *Pattern Recognition Letters*, 18, 1143–1151.

DeLapaz, R. L., Herskovits, E., Di Gesu, V., Hanson, W. J., & Bernstein, R. (1990). Cluster analysis of medical magnetic-resonance images data: diagnostic application and evaluation. *Proceedings of SPIE*, 1259, Extracting Meaning from Complex Data: Processing, Display, Interaction, 176. DOI: 10.1117/12.19984.

Derwich, K., Jędrzejczak P., & Pawelczyk, L. (2008). Metody wspomaganego rozrodu. In Z. Słomko (Eds.), *Ginekologia* (pp. 516–532). Warszawa: PZWL.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25), 14863–14868.

Horák, S. (2004). Insemination – indications, methods and efficiency. *Ginekologia Praktyczna*, 12(6), 41–49.

Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43, 59–69.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer.

Kurzawa, R., Kaniewska, D., & Bączkowski, T. (2010). Infertility from clinical and social perspective. *Przewodnik Lekarza*, 2, 149–152.

Licznar, P. & Łomotowski, J. (2006). Zastosowanie sieci neuronowych Kohonena do prognozowania dobowego poboru wody, *Ochrona Środowiska*, 28, 45–48.

McLachlan, G. J. (1992). Cluster analysis and related techniques in medical research. *Statistical Methods in Medical Research*, 1(1), 27–48.

Migut, G. (2009). Zastosowanie technik analizy skupień i drzew decyzyjnych do segmentacji rynku. *StatSoft Polska*. Retreived from: http://www.statsoft.pl/czytelnia/artykuly/Zastosowanie_technik.pdf.

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., 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., Malinowski, P., Milewska, A. J., Ziniewicz, P., Czerniecki, J., Pierzyński, P., & Wołczyński, S. (2012). Classification issue in the IVF ICSI/ET data analysis. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 29(42), 75–85.

Milewski, R., Milewska, A. J., Czerniecki, J., Leśniewska, M., & Wołczyński, S. (2013). 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.

Pierzyński, P. (2011). *Zajść w ciążę*. Białystok: CMR.

Radwan, J. (2011). Badanie niepłodnej pary. In J. Radwan & S. Wołczyński (Eds.), *Niepłodność i rozród wspomagany* (pp. 47–66). Poznań: Termedia.

Radwan, J., Krasiński, R., & Gruszczyński, W. (2011). Badanie nasienia. In J. Radwan & S. Wołczyński (Eds.), *Niepłodność i rozród wspomagany* (pp. 67–80). Poznań: Termedia.

Radwan, P. (2011). Inseminacja domaciczna. In J. Radwan & S. Wołczyński (Eds.), *Niepłodność i rozród wspomagany* (pp. 165–178). Poznań: Termedia.

Sava, F. A., & Popa, R. I. (2011). Personality types based on the big five model. A cluster analysis over the Romanian population. *Cognition, Brain, Behavior. An Interdisciplinary Journal*, 15(3), 359–384.

Shannon, W., Culverhouse, R., & Duncan, J. (2003). Analyzing microarray data using cluster analysis. *Future Medicine*, 4(1), 41–52.

Stanisz, A. (2007). *Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny. T. 3. Analizy wielowymiarowe*. Kraków: StatSoft.

Timm, N. T. (2002). *Applied Multivariate Analysis*, Springer.

Tkaczuk-Włach, J., Robak-Chołubek, D., & Jakiel, G. (2006). Male infertility. *Przegląd Menopauzalny*, 5, 333–338.

Wainer, R., Albert, M., Dorion, A., Bailly, M., Berge're, M., Lombroso, R., Gombault, M., & Selva, J. (2004). Influence of the number of motile spermatozoa inseminated and of their morphology on the success of intrauterine insemination. *Human Reproduction*, 19(9), 2060–2065.

Xu, R., & Wunsch, D. C. 2nd. (2010). Clustering algorithms in biomedical research: a review. *IEEE In Biomedical Engineering*, 3, 120–154.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0038

Dipolar Tree Ensemble With and Without Adjustment to Competing Risks: Application to Medical Data

Małgorzata Krętowska[1]

[1] Faculty of Computer Science, Bialystok University of Technology, Poland

Abstract. The analysis of survival data often aims at the prediction of failure time distribution. In cases of competing risk events, the time distributions of more than one failure are under investigation. In this paper, the comparison of two approaches to analyzing survival data with competing risks is presented. The analyses are performed by use of an ensemble of dipolar trees with and without adjustment to competing risks.

Introduction

Collecting survival data, we are very often interested in the prediction of disease-free survival. In such investigations, the observation time for every patient may end because of an accident or the end of follow-up time. The patient may also be lost to follow-up for other reasons. In the first case, the exact time of failure occurrence is known for the patient. In the other two cases, the observation is finished without any event, therefore the exact time of failure occurrence is unknown and the observation is considered as censored. For such patients, we only know that the survival time was no less than the observation time.

In many studies, however, the interest focuses not only on disease-free survival, but also on the time distribution of a specified event occurrence (failure occurrence). If the collected data contain information about one event only, the analysis may be performed by use of 'classical' methods for survival analysis (Kalbfleisch, 1980; Marubini et al., 1995). Competing risks data is a special type of survival data, in which more than one type of failure is under investigation, (Putter et. al., 2007) and, as a result, their analysis requires more sophisticated methods (Pintilie, 2006).

In this paper, I would like to compare the distributions of failure time occurrence obtained by two approaches to analyzing competing risks. According to the first method, all types of failure are considered as separate

ISBN 978-83-7431-392-6 ISSN 0860-150X

failures, treating other cases as censored; according to the other, all investigated failures are taken together. A dipolar tree ensemble, proposed in (Kretowska, 2006), is used as a prediction tool in both cases. The difference is the way the dipolar trees are inducted. The use of data with competing risk events requires additional adjustments (Kretowska, 2012), which make the information of competing risk events possible.

Survival Data with Competing Risk Events

Survival data are represented by a set of observations, which, besides the values of covariates, also contains information about the time of occurrence of a specified event. The event, also called failure, may represent e.g. death or disease relapse. A learning sample L is defined as $L = (\boldsymbol{x}_i, t_i, \delta_i)$, $i = 1, 2, \ldots, n$, where $\boldsymbol{x}_i$ is an N-dimensional covariate vector describing the ith observation (patient), t_i is the follow-up time and $\delta_i \in \{0, 1\}$ indicates whether the failure has occurred. δ_i equal to 0 represents a censored observation – the observation with unknown failure occurrence and δ_i equal to 1 represents an uncensored observation.

In cases of survival data with competing risk events, a patient is at risk of p ($p > 1$) different types of failure (Figure 1). Assuming that the time of occurrence for the ith type of failure is T_j, we are interested only in the failure with the shortest time: $T = \min(T_1, T_2, \ldots, T_p)$. The learning sample L_{CR} for competing risk data is defined as $L_{CR} = (\boldsymbol{x}_i, t_i, \delta_i)$, $i = 1, 2, \ldots, n$, but unlike in cases of a single failure, t_i is the time to the first event observed and $\delta_i \in \{0, 1, \ldots, p\}$ indicates the type of failure. δ_i equal to 0 represents a censored observation, which means that for a given patient no failure has occurred.

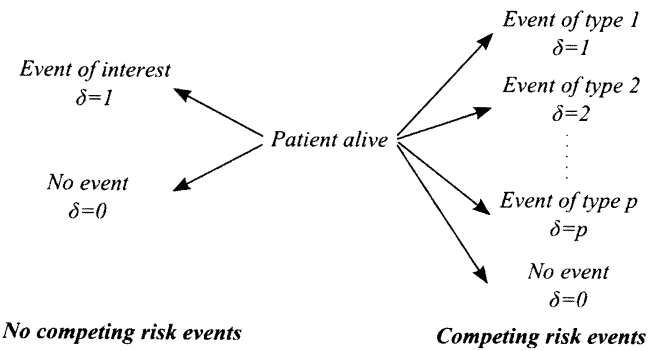


Figure 1. Survival data in cases of absence and presence of competing risk events

The distribution of survival time may be expressed by several functions. The most common approach is the use of the survival function, which presents the probability that the failure does not occur before time t:

$$S(t) = P(T > t).$$

The Kaplan-Meier estimator of the survival function is given as:

$$KM(t) = \prod_{j/t_{(j)} \leq t} \frac{n_j - d_j}{n_j}$$

where $t_{(1)} < t_{(2)} < \ldots < t_{(D)}$ are distinct, ordered survival times from the learning sample L, in which the event of interest has occurred, d_j is the number of events at time $t_{(j)}$ and n_j is the number of patients at risk at $t_{(j)}$ (i.e., the number of patients who are alive at $t_{(j)}$ or experience the event of interest at $t_{(j)}$).

A cumulative incidence function (CIF) is used to describe the failure time distribution in cases of competing risks. The CIF for the ith type of event is defined as:

$$F_i(t) = P(T \leq t, \delta = i)$$

and may be interpreted as the probability that an event of type i occurs before or at time t. The estimate of CIF is given as:

$$\tilde{F}_i(t) = \sum_{j/t_{(j)} \leq t} \frac{d_{ij}}{n_j} \tilde{S}(t_{(j-1)})$$

where d_{ij} is the number of events of type i that occur at time $t_{(j)}$ and $\tilde{S}(t_{(j-1)})$ is the Kaplan-Meier estimator of the probability of being free of any event by time $t_{(j-1)}$.

When there are no competing risk events, the value of $(1 - KM(t))$ is equal to $\tilde{F}_1(t)$, in other cases $\tilde{F}_i(t) < 1 - KM_i(t)$, where $KM_i(t)$ is the Kaplan-Meier estimator calculated for the ith type of event. The value of $1 - KM_i(t)$ can be received from the equation (Pintilie, 2006):

$$1 - KM_i(t) = \sum_{j/t_{(j)} \leq t} \frac{d_{ij}}{n_j} KM_i(t_{(j-1)})$$

As we can see, the second part of the formulae, where instead of disease-free survival $\tilde{S}(t_{(j-1)})$ the $KM_i(t_{(j-1)})$ is present, causes the difference between the functions.

Ensemble of Dipolar Tree

An ensemble of dipolar trees (Kretowska, 2006; Kretowska, 2012) is used as a prediction tool. The appropriate construction of single trees enables the analysis of competing risks data as well as the data without competing risk events.

The ensemble is a set of single dipolar trees. Each tree is inducted on the base of a bootstrap sample, drawing with replacement from the learning data L, or in cases of competing risks, from the learning data L_{CR}. The tree induction starts from the root. Next, other internal nodes are created. Each internal node has two children nodes: internal or terminal ones. Internal nodes contain the test, which causes a vector of covariates to go to the appropriate child node. If certain conditions are fulfilled, the node is set as a terminal node, which does not contain any test, but may be viewed as a set of covariate vectors that have reached the node.

In the dipolar tree, the test in the kth internal node has a form of a hyperplane: $H(w_k) = \{\boldsymbol{x} : w_k^T x = 0\}$ and is constructed by minimizing the dipolar criterion function being a sum over some specified criterion functions connected with dipoles. The dipole (Bobrowski et. al., 1997) is a pair of different covariate vectors $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ from the learning set. Mixed and pure dipoles are distinguished. The pure dipoles are created between pairs that should not be separated, and the mixed ones between pairs that should belong to different groups. Depending on the problem, the dipoles are created in a different manner. In cases of data with no competing risk events, we are interested in dividing the feature space into areas with homogeneous survival times, so the vectors with similar failure times constitute the pure dipoles, and the vectors with distant failure times constitute the mixed dipoles (Kretowska, 2006):

- a pair of feature vectors $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ constitutes the pure dipole, if
  $\delta_i = \delta_j = 1 \wedge |t_i - t_j| < \eta$
- a pair of feature vectors $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ constitutes the mixed dipole, if
  $\delta_i = \delta_j = 1 \wedge |t_i - t_j| > \zeta$
  $\delta_i = 0, \delta_j = 1 \wedge t_i - t_j > \zeta$ or $\delta_i = 1, \delta_j = 0 \wedge t_j - t_i > \zeta$

For data with competing risk events, the analysis aims at dividing the feature space into such areas, which would include the patients with the same cases of failure and similar survival times. Taking into account censored cases, the following rules of dipole construction can be formulated (Kretowska, 2012):

- a pair of feature vectors $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ forms the pure dipole, if
  $\delta_i = \delta_j = z \wedge |t_i - t_j| < \eta_z, z = 1, 2, \ldots, p$

- a pair of feature vectors $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ forms the mixed dipole, if
  $\delta_i = \delta_j = z \wedge |t_i - t_j| > \zeta_z, z = 1, 2, \ldots, p$
  $(\delta_i = 0, \delta_j = z \wedge t_i - t_j > \zeta_z)$ or $(\delta_i = z, \delta_j = 0 \wedge t_j - t_i > \zeta_z), z = 1, 2, \ldots, p$

Parameters η_z and ζ_z are equal to quartiles of absolute values of differences between uncensored survival times for the zth type of failure, $z = 1, 2, \ldots, p$. Based on earlier experiments, the parameter η_z is fixed as 0.3 quantile and ζ_z – 0.6. Parameters η and ζ are calculated in a similar manner, without taking into account the type of failure.

The general algorithms for generating the ensemble of dipolar trees is as follows:

1. Draw k bootstrap samples $(L_1; L_2; \ldots; L_k)$ of size n with replacement from L or in cases of data with competing risks from L_{CR}.
2. Induction of dipolar survival tree $T(L_i)$ based on each bootstrap sample L_i.
3. For each tree $T(L_i)$, distinguish the set of observations $L_i(\boldsymbol{x}_n)$ which belongs to the same terminal node as $\boldsymbol{x}_n$.
4. Build an aggregated sample $L_A(\boldsymbol{x}_n) = [L_1(\boldsymbol{x}_n), L_2(\boldsymbol{x}_n); \ldots; L_k(\boldsymbol{x}_n)]$.
5. Calculate the Kaplan-Meier aggregated survival function for a new observation $\boldsymbol{x}_n$ as $KM_A(t/\boldsymbol{x}_n)$.
6. In cases of no competing risk events, calculate $1 - KM_A(t/\boldsymbol{x}_n)$; for competing risks calculate the aggregated CIF for all types of failure for a new observation $\boldsymbol{x}_n$: $\tilde{F}_i(t/\boldsymbol{x}_n)$, for $i = 1, 2, \ldots, p$.

The output of the ensemble for a new observation is, depending on the problem, the aggregated Kaplan-Meier function (and then $1 - KM(t)$) or the estimator of CIF. In both cases, some other statistics may also be calculated. Median value or lower and upper quartile belong to the most common statistics, but other quantiles may also be calculated for obtained functions.

Experimental Results

The experiments were performed on two datasets: breast cancer data and follicular type lymphoma data, both with competing risk events. For each dataset, two types of tests were done:

1. For whole data sets, with competing risks events (CR).
2. Without competing risks (nCR). For each basic dataset, p separate datasets were created, each containing the information of one distinguished event, treating other events as censored observations.

All the experiments were conducted using the ensemble of 100 single survival trees.

Breast cancer data (Fyles et al., 2004) contained information about 641 women (50 years old or older) who had undergone breast-conserving surgery for an invasive adenocarcinoma 5 cm or less in diameter. They were randomly assigned to receive breast irradiation plus tamoxifen (321 women) or tamoxifen alone (320 women). The data were collected between 1992 and 2000. The last follow-up was conducted in summer 2002. Table 1 contains a description of the variables (Ibrahim et al., 2008), while Table 2 presents characteristics of each variable – for discrete variables – the number of cases having the same value; for continuous ones – minimum, maximum, lower (Q_1) and upper (Q_3) quartile and median values.

Table 1. Description of variables in breast cancer dataset

Variable name	Description
Tx	Randomized treatment: 1 = tamoxifen, 2 = radiation + tamoxifen
Variables assessed at the time of randomization	
Pathsize	Size of tumor (cm)
Hist	1 = ductal, 2 = lobular, 3 = medullary, 4 = mixed, 5 = other
Hrlevel	0 = negative, 1 = positive
Hgb	Hemoglobin (g/l)
Nodediss	Whether axillary node dissection was done: 0 = no, 1 = yes
Age	Age (years)
Outcome variables	
Time	Time from randomization to event or last follow up (years)
d	Status at last follow up: 0 = censored, 1 = death, 2 = relapse, 3 = malignancy

Table 2. Characteristics of variables in breast cancer dataset

Discrete variables	Value (number of cases)
Tx	1 (321); 2 (320)
Hist	1 (397); 2 (31); 3 (5); 4 (174); 5 (34)
Hrlevel	0 (46); 1 (595)
Nodediss	0 (106); 1 (535)
D	0 (503); 1 (14); 2 (69); 3 (55)
Continuous variables	
Path size	Min = 0.2; Q_1 = 1; Med = 1.5; Q_3 = 2; Max = 4.5
Hgb	Min = 96; Q_1 = 128; Med = 135; Q_3 = 142; Max = 169
Age	Min = 50; Q_1 = 59; Med = 67; Q_3 = 73; Max = 88

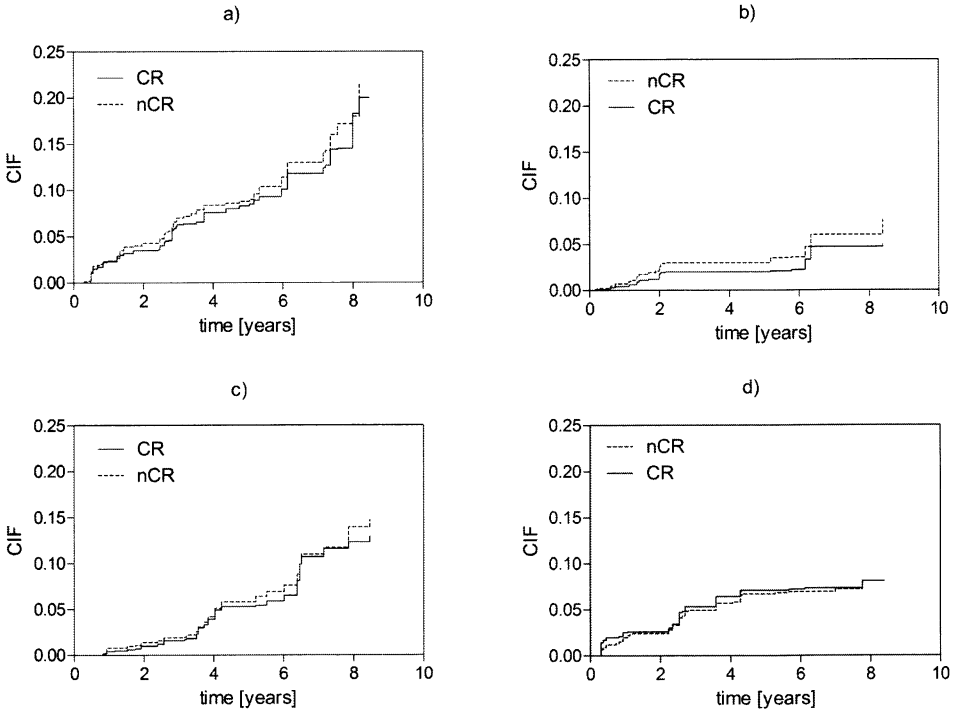


Figure 2. Cumulative incidence functions received for breast cancer data for: a) relapse, Tx = 1; b) relapse, Tx = 2; c) malignancy, Tx = 1; d) malignancy, Tx = 2

Cumulative incidence functions received for breast cancer data are presented in Figure 2. Each graph contains two cumulative incidence functions received for the data with and without competing risks. Figures 2(a) and 2(b) contain CIF's for relapse, while Figures 2(c) and 2(d) – CIF's for malignancy with hist = 1, hrlevel = 1, nodediss = 1, and values of continuous features fixed as their medians.

There are small differences between the two received functions in Figures 2(a) and 2(c) calculated for patients treated with tamoxifen alone, and bigger for the other treatment (Figures 2(c) and 2(d)). The differences may be caused by the previously described relationship between the two functions: $\tilde{F}_i(t) \leq 1 - KM_i(t)$ or by different division of feature space in the two presented approaches, described in "Ensemble of Dipolar Tree" section. The latter is especially visible in Figure 2(d) where $\tilde{F}_i(t)$ is usually greater than $1 - KM_i(t)$.

A Lymphoma patient dataset was created at Princess Margaret Hospital, Toronto (Pintilie, 2006). In the experiments we used the subset

of 541 patients who had follicular type lymphoma, registered for treatment at the hospital between 1967 and 1996, with early stages of disease (I or II) and treated with radiation alone (118 people) or with radiation and chemotherapy (179 people). Each patient was described by four variables, explained in Table 3 and characterized in Table 4. The event of interest was failure from the disease: no response to treatment or relapse. A competing risk type of event was death without failure. There are 272 events of interest and 76 observations of death without relapse.

Table 3. Description of variables in lymphoma patient dataset

Variable name	Description
Variables assessed at the time of diagnosis	
Age	Age (years)
Hgb	Hemoglobin (g/l)
Clinstg	Clinical stage: 1 = stage I; 2 = stage II
Ch	Chemotherapy: 0 = no; 1 = yes
Outcome variables	
Time	Time from diagnosis to event or last follow up (years)
D	Status at last follow up: 0 = censored, 1 = no response to treatment or relapse, 2 = death

Table 4. Characteristics of variables in lymphoma patient dataset

Discrete variables	Value (number of cases)
Clinstg	1 (362); 2 (179)
Ch	0 (118); 1 (179)
D	0 (193); 1 (272); 2 (76)
Continuous variables	
Age	Min = 17; Q_1 = 47; Med = 58; Q_3 = 67; Max = 86
Hgb	Min = 40; Q_1 = 130; Med = 140; Q_3 = 150; Max = 189

In Figure 3, four graphs obtained by an analysis of follicular type lymphoma data are presented. Each of them contains four functions received for death and relapse taking into account CR and nCR approaches and two values of clinical stage (clinstg = 1 and clinstg = 2) and chemotherapy (ch = 0 and ch = 1). The differences between appropriate functions do not always follow the relationship $\tilde{F}_i(t) \leq 1 - KM_i(t)$, which

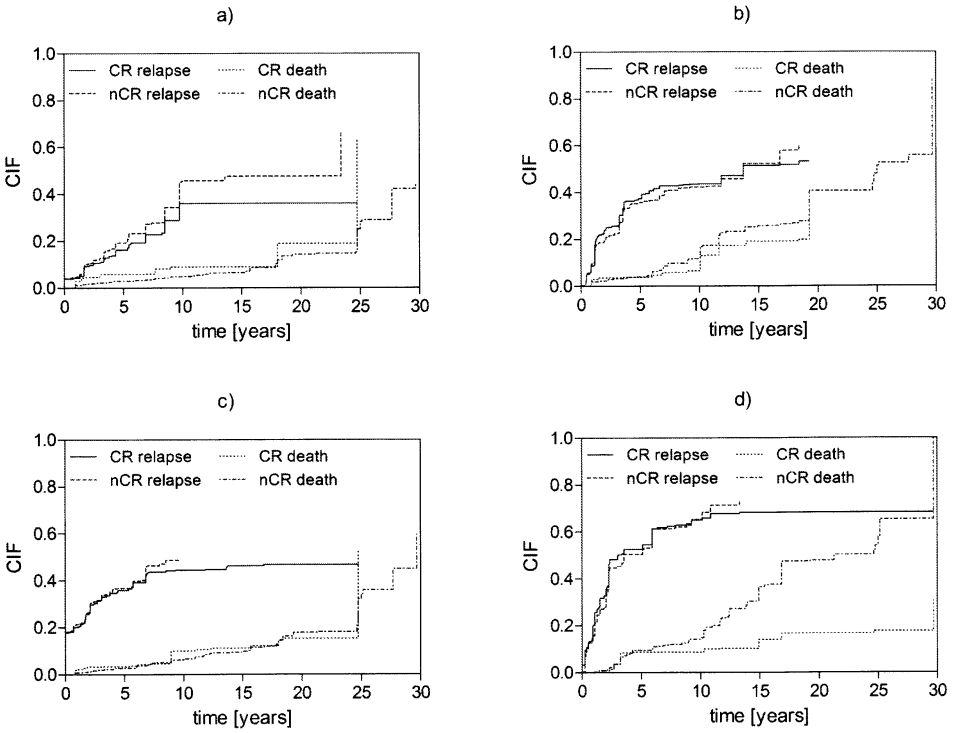


Figure 3. Cumulative incidence functions received for relapse and death in lymphoma data: a) clinstg = 1, ch = 1; b) clinstg = 1, ch = 0; c) clinstg = 2, ch = 1; d) clinstg = 2, ch = 0

may suggest that received groups of patients in the approach adjusted for competing risk events differ from the other approach. This is visible especially in the probabilities received for death, but in Figure 3(a) there are also quite big differences between the two functions calculated for relapse.

In Figures 4 and 5, we can observe the surfaces of lower quartiles obtained from CIF (Figure 4) and $1 - KM$ (Figure 5) functions describing the distribution of no response to treatment or relapse. In cases of the CR approach we can see a better prognosis for patients aged 30–40 with higher values of hemoglobin (150–160 g/l) and for patients aged 50–60 and hemoglobin values of 130–140 g/l. The highest value of the lower quartile received for the CIF function is 3.6 years. In the case of the graph visible in Figure 5, the maximum value of the surface is 7 years. Its shape does not indicate age as a risk factor. The value of the lower quartile depends only on hemoglobin: the higher the value of hemoglobin, the better the prognosis.

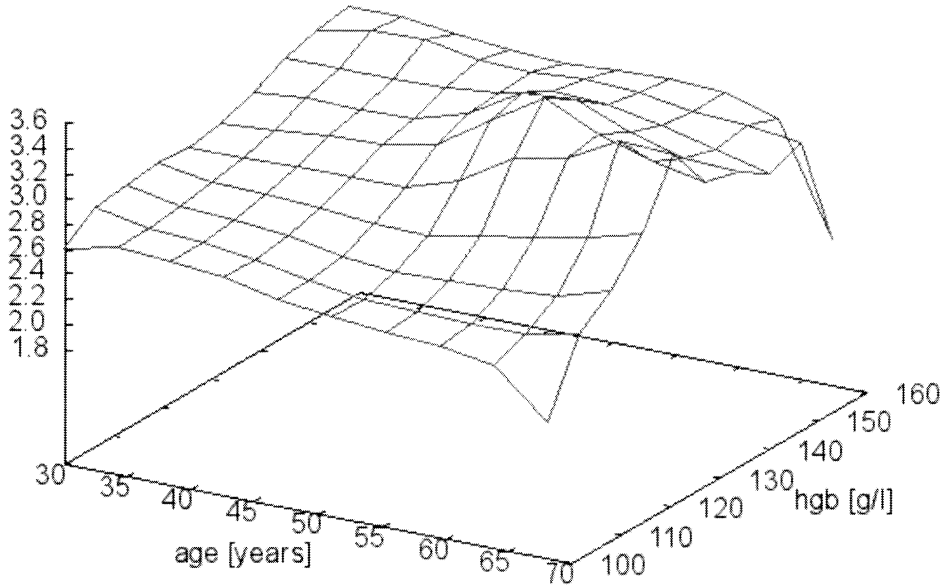


Figure 4. The influence of age and hemoglobin for lower quartiles calculated for CIF functions ($d = 1$) for patients with clinical stage I and chemotherapy

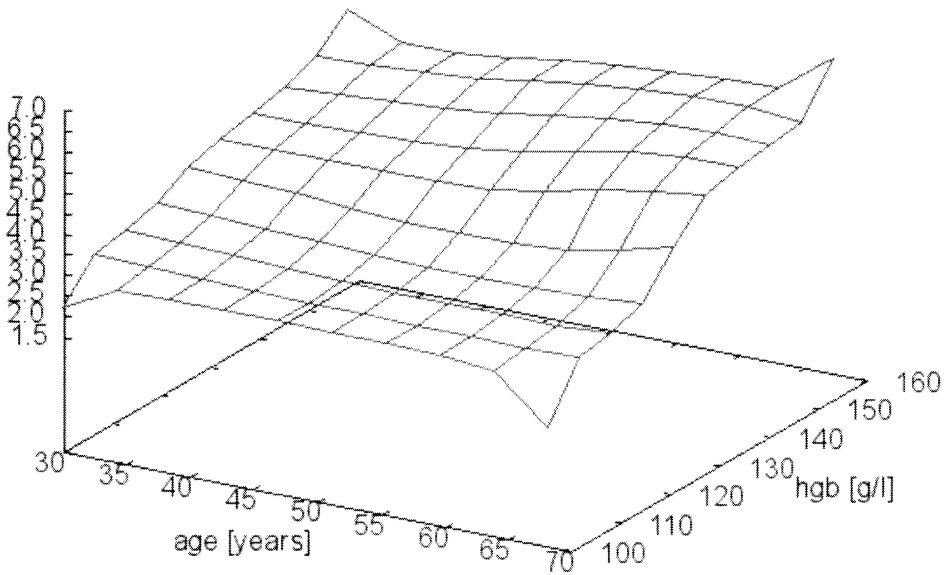


Figure 5. The influence of age and hemoglobin for lower quartiles calculated for $1 - KM$ functions ($d = 1$) for patients with clinical stage I and chemotherapy

Conclusions

In this paper, two approaches to analyzing survival data – with and without adjustment to competing risks – are presented. The experiments were performed on two real datasets: breast cancer data and follicular type lymphoma data, containing a high number of censored observations. To obtain results, cumulative incidence functions or $1 - KM$ estimators were calculated. For follicular type lymphoma data, the surfaces of lower quartiles are also presented. The graph comparisons show how important the use of information about competing risks is. The cumulative incidence functions received for competing risks data differ from the results obtained for single events, treating other events as censored observations. The differences are not only related to the association between the two functions used: $\tilde{F}_i(t) \leq 1 - KM_i(t)$, but also to the way the ensemble of dipolar trees uses the information about competing risks. This leads to dissimilar division of feature space and hence, the established groups of patients with similar survival experience are different for the two approaches. The interpretation of $1 - KM$ estimators obtained for a single event without taking into account competing risks may mislead.

Acknowledgements

This work was supported by the grant S/WI/2/2013 from Bialystok University of Technology.

REFERENCES

Bobrowski, L., Kretowska, M., & Kretowski, M. (1997). Design of neural classifying networks by using dipolar criterions. In proceedings of the Third Conference on Neural Networks and Their Applications, 14–18 October 1997 (pp. 689–694). Kule, Poland.

Fyles, A. W., McCready, D. R., Manchul, L. A., Trudeau, M. E., Merante, P., Pintilie, M., Weir, L. M., & Olivotto, I. A. (2004). Tamoxifen with or without breast irradiation in women 50 years of age or older with early breast cancer. *New England Journal of Medicine*, 351, 963–970.

Ibrahim, N. A., Kudus, A., Daud, I., & Abu Bakar, M. R. (2008). Decision tree for competing risks survival probability in breast cancer study. *World Academy of Science, Engineering and Technology*, 38, 15–19.

Kalbfleisch, J. D., & Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: John Wiley & Sons Ltd.

Kretowska, M. (2006). Random forest of dipolar trees for survival prediction. In L. Rutkowski, R. Tadeusiewicz, L. A. Zadeh, & J. M. Zurada (Eds.), ICAISC 2006, LNCS (LNAI) 4029, 909–918.

Kretowska, M. (2012). Competing Risks and Survival Tree Ensemble, L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, & J. M. Zurada (Eds.), ICAISC 2012, Part I, LNCS 7267, 387–393.

Marubini, E., & Valsecci, M. G. (1995). *Analysing survival data from clinical trials and observational studies.* England: John Wiley & Sons Ltd.

Pintilie, M. (2006). *Competing Risks: A Practical Perspective.* England: John Wiley & Sons Ltd.

Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-stage models. *Statistics in Medicine*, 26, 2389–2430.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0033

Comparison of Artificial Neural Networks and Logistic Regression Analysis in Pregnancy Prediction Using the In Vitro Fertilization Treatment

Robert Milewski[1], **Anna Justyna Milewska**[1], **Teresa Więsak**[2], **Allen Morgan**[3]

[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, Olsztyn, Poland

[3] Shore Institute for Reproductive Medicine, Lakewood, NJ, USA

Abstract. Infertility is recognized as a major problem of modern society. Assisted Reproductive Technology (ART) is the one of many available treatment options to cure infertility. However, the efficiency of the ART treatment is still inadequate. Therefore, the procedure's quality is constantly improving and there is a need to determine statistical predictors as well as contributing factors to the successful treatment. There is a concern over the application of adequate statistical analysis to clinical data: should classic statistical methods be used or would it be more appropriate to apply advanced data mining technologies? By comparing two statistical models, Multivariable Logistic Regression analysis and Artificial Neural Network it has been demonstrated that Multivariable Logistic Regression analysis is more suitable for theoretical interest but the Artificial Neural Network method is more useful in clinical prediction.

Introduction

Based on available statistics, approximately 15% of Polish couples suffer from infertility. Some authors suggest the percentage is even 18–20% (Radwan, 2011). This value generally is in the range of 10–20% and reported differences depend on the data collection methods in different countries: in Denmark it is 11%, in France 16.4% and in the UK 17% (Radwan, 2011). In the United States, using the current duration approach, infertility among women 15–44 years old is 15.5% (Thoma et al., 2013).

Female factors such as: endometriosis, PCO or other ovulatory, uterine, or fallopian tube irregularities (Milewski et al., 2013) and male factors such as: oligoasthenospermia, asthenospermia, teratospermia, azoospermia, rare

ISBN 978-83-7431-392-6 ISSN 0860-150X

oligospermia and immunological factors (Radwan, 2011) contribute to infertility. Idiopathic infertility is a situation in which the clinical evaluation and laboratory tests are normal (within the range) but the couple is not able to conceive naturally.

The infertility treatment depends on the type of diagnosis, ART is one of those options. Many factors influence the efficacy of ART but the age of the female is the most important one because it affects the quality of oocytes. The efficiency of infertility treatment administered to women over 40 years old is only 10% (Milewski et al., 2008). It has been observed recently, that environmental factors contribute to both male and female infertility more and more. Other factors also influence the outcome of the treatment such as the uterine contraction level in women undergoing in vitro fertilization (Milewski, Pierzynski et al., 2012).

In spite of the constant improvement in techniques that enhance the efficacy of ART treatment, the pregnancy rate is still low and remains in the range of 40% (Milewski et al., 2013). To increase the likelihood of achieving pregnancy success, more than one embryo is transferred, subsequently causing multiple pregnancies, which is a major problem of infertility treatment. Therefore, a lot of studies across the world are focused on improving the single embryo transfer method without sacrificing the success of the ART treatment. Such a treatment also reduces the occurrence of multiple pregnancies. The key element of such an approach is to establish prognostic values that would allow the selection of an appropriate treatment protocol and method of treatment with the highest probability of a successful outcome.

It turns out that traditional methods of statistical analysis are inefficient in the precise determination of the reasons behind infertility and in providing effective predictors of treatment. Univariate analysis only determines the relationship between the analyzed factor and treatment outcome. Multivariate analyses (e.g. multivariate logistic regression) provide models that allow the prediction of pregnancy or lack of pregnancy with a higher level of accuracy. However, these methods include some restrictions that influence their effectiveness and limit their wide clinical application. Therefore, there is a necessity to apply new and more advanced statistical analyses such as data mining methods (Witten et al., 2011). The effort in this area is focused on determining which data mining method would be the best suited to analyze data derived from infertility treatment. Siristatidis et al. (2011) advocate for the usefulness of artificial intelligence methods in analyzing data concerning reproductive medicine. A lot of hope lies in the application of Artificial Neural Networks (ANNs), which so far gives excellent results in predict-

ing negative outcomes of infertility treatment (Milewski et al., 2009). There are also some attempts to apply other methods, such as Basket Analysis (Milewska et al., 2011) and Correspondence Analysis (Milewska et al., 2012) into ART data analysis. It is possible that good statistical results could be produced by combining advanced data mining methods with the feature classification procedures (Milewski, Malinowski et al., 2011, 2012).

The aim of this study is to compare logistic regression analysis (a classical statistical method that can predict effectiveness of infertility treatment) with the ANNs method that resembles the concept of the human brain in analytical ability.

Material and Methods

The data of 1995 infertility patients of the Shore Institute for Reproductive Medicine, Lakewood, NJ, USA, in the age range of 21–45 years old were analyzed. Pregnancy (defined according to a positive pregnancy test that is ≥ 5 IU HCG/ml on days 10–12 after embryo transfer) as a binary variable was the dependent feature in our analysis. Twenty-six various variables of patients' treatment were independent variables. Fourteen of these variables were quantitative and 12 were qualitative (Table 1). The quantitative variables were: age of the patients, number of oocytes retrieved and cultured, semen parameters and hormone levels. The qualitative variables were: diagnosis, type of treatment and stimulation protocol.

Table 1. List of independent variables (quantitative and qualitative)

Quantitative variables		Qualitative variables	
Age	Age of woman	Insem_type	1-ICSI, 0-classical IVF
Nr_ET	Number of transferred embryos	Tubal	Tubal factor
Total_nr_eggs	Total number of retrieved eggs	Endometr	Endometriosis
Nr_MII_eggs	Number of mature eggs	Ovulatory	Ovulatory factor
Nr_eggs_ins	Number of inseminated eggs	MF	Male factor
Nr_2pn	Number of fertilized eggs	AMA	Advanced maternal age
Nr_cultured	Number of cultured eggs	PCOS	Polycystic Ovary Syndrome
Nr_clvd	Number of cleavage embryos	DOR	Diminish Ovarian Reserve
Vol_prewash	Volume of the semen	Idiopathic	Idiopathic factor
Ct_prewash	Sperm concentration	Down_regul	1-Lupron, 0-Antagon
Mot_prewash	Sperm motility	Lupron_dur	Lupron duration
Baseline_FSH	FSH level on day 3	HCG_dose	HCG dose to induce
Nr_stim_days	Number of days on stimulation		ovulation
E2_at_HCG	E_2 level at HCG injection		(250, 5000, 7500, 10000 IU)

Univariate and multivariate logistic regression analyses were performed using software Stata/IC 12.1 (Stata Corp LP., College Station, TX, USA) to provide predictions for pregnancy occurrence. The data were also analyzed using Artificial Neural Networks technology with application of the software Statistica Data Miner + QC 10.0 (StatSoft, Tulsa, OK, USA). To determine the quality of obtained predictors, the Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) were analyzed (Hanley at al., 1982). Based on Hanley's algorithm, statistically significant differences between the predictors were obtained (Hanley et al., 1997). Statistical significance was determined at the $p < 0.05$ level.

The Logistic Regression Model

The logistic regression analysis is the most appropriate method to determine the relationship between the described above independent variables and pregnancy occurrence among the classical statistical methods. The univariate analysis provides a statistical significance level, an Odds Ratio with 95% Confidence Intervals and a Standard Error value for the probable effect of each independent variable on the dependent variable (pregnancy). Table 2 shows results of the univariate logistic regression analysis.

Table 2. Results of univariate logistic regression analysis

Pregnancy	Odds Ratio	Std. Error	p-value	95% Confidence Interval	
Age	0.950636	0.0107684	<0.001*	0.929763	0.971978
Nr_ET	1.047353	0.0395560	0.221	0.972625	1.127823
Total_nr_eggs	1.036945	0.0068559	<0.001*	1.023594	1.050470
Nr_MII_eggs	1.054609	0.0084598	<0.001*	1.038158	1.071321
Nr_eggs_ins	1.046144	0.0076691	<0.001*	1.031221	1.061284
Nr_2pn	1.066017	0.0104114	<0.001*	1.045805	1.086620
Nr_cultured	1.090955	0.0131257	<0.001*	1.065530	1.116987
Nr_clvd	1.097478	0.0135132	<0.001*	1.071310	1.124286
Vol_prewash	1.007906	0.0310475	0.798	0.948854	1.070632
Ct_prewash	0.999363	0.0011440	0.578	0.997124	1.001608
Mot_prewash	0.997700	0.0019475	0.238	0.993890	1.001524
Baseline_FSH	0.963693	0.0133326	0.008*	0.937912	0.990181
Nr_stim_days	0.983413	0.0267040	0.538	0.932443	1.037170
E2_at_hCG	1.000151	0.0000498	0.002*	1.000054	1.000249
Insem_type	0.733145	0.0684540	0.001*	0.610538	0.880373
Tubal	0.892669	0.1040469	0.330	0.710357	1.121771
Endometr	0.996447	0.1108244	0.974	0.801280	1.239152
Ovulatory	1.106040	0.1507395	0.460	0.846764	1.444704

Pregnancy	Odds Ratio	Std. Error	p-value	95% Confidence Interval	
MF	1.136840	0.1037123	0.160	0.950704	1.359418
AMA	0.880833	0.1446499	0.440	0.638425	1.215283
PCOS	1.105390	0.3373659	0.743	0.607754	2.010494
DOR	0.810750	0.0949209	0.073	0.644511	1.019868
Idiopathic	0.854709	0.2195203	0.541	0.516653	1.413962
Down_regul	1.292339	0.1439821	0.021*	1.038825	1.607720
Lupron_dur	0.756088	0.0831845	0.011*	0.609429	0.938040
HCG_dose	0.999952	0.0000198	0.016*	0.999913	0.999991

The 13 independent variables statistically affected the pregnancy variable (see p value with the asterisk). Univariate analyses determine only statistical differences between variables but without the power of prediction. The analyzed variables were not able to efficiently demonstrate a relationship with the dependent variable. Therefore, multivariate analyses were applied to create a prediction model. The multivariate model was built by selecting variables with at least $p < 0.2$ from the univariate analysis. The $p < 0.2$ value was chosen to allow for variables with slightly diminished statistical power to enter into our model. Variables that were strongly correlated with each other were excluded. Finally, the 6 selected variables and the intercept were included into the model (Table 3).

Table 3. Multivariate logistic regression model

Pregnancy	Odds Ratio	Std. Error	p-value	95% Confidence Interval	
Age	0.958398	0.0113094	<0.000	0.936486	0.980822
Nr_clvd	1.085128	0.0137621	<0.000	1.058487	1.112439
Insem_type	0.690624	0.0760208	0.001	0.556603	0.856916
MF	1.355477	0.1446616	0.004	1.099635	1.670842
Down_regul	1.226959	0.1455581	0.085	0.972410	1.548143
Lupron_dur	0.630935	0.0739802	<0.000	0.501391	0.793948
Const	2.629308	1.1784360	0.031	1.092294	6.329123

This multivariate logistic regression model allowed us to establish probability of achieving pregnancy by the individual patient (`fit_pr`) and was the searching predictor. Figure 1 presents the ROC curve for the `fit_pr` and Table 4 contains the Area Under the Curve with the 95% Confidence Intervals and Standard Error. The cut-off point for the probability value produced by our model (`fit_pr` = 0.479) was established by applying the minimum sum of squared components method. The sensitivity and specificity were 65.4% and 56.8%, respectively. However, this model is not able

to predict presence or lack of pregnancy with perfect accuracy because it misses around 35% of IVF cycles with true pregnancies and more than 43% of cycles with lack of pregnancy.

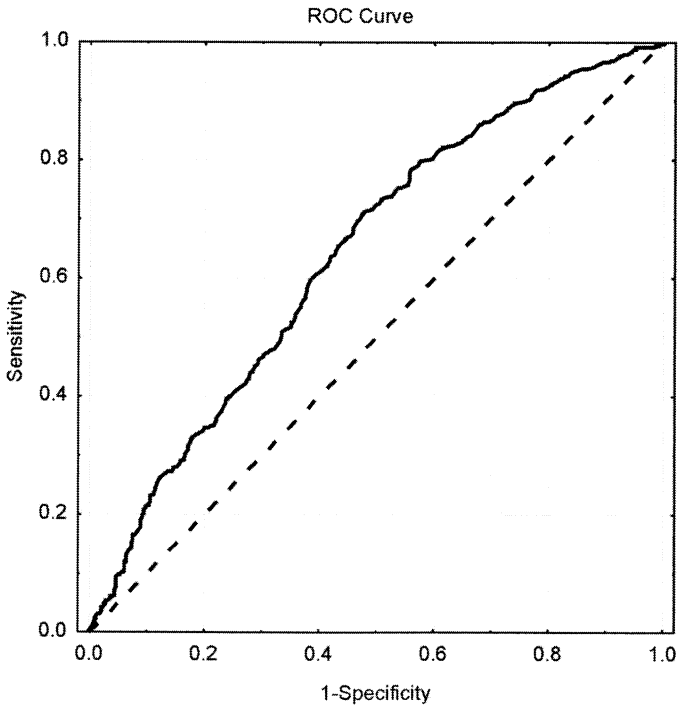


Figure 1. The ROC Curve for the Logistic Regression model

Table 4. Characteristics of the ROC Curve created for the Logistic Regression model

AUC	Std. Error	95% Confidence Interval	
0.642	0.015	0.613	0.671

The Artificial Neural Networks Model

The classical statistical analysis gives us a statistical relationship among the independent and dependent variables with the statistical power for each individual variable. In contrast, the Artificial Neural Networks method

creates such a statistical model which will provide the best prognosis for the studied phenomena based on the entered values, without determining the influence of each variable (Tadeusiewicz, 1993). The ANNs method relies on processing data in such a way as the human brain does. The attractiveness of ANNs comes from their remarkable information processing characteristics pertinent mainly to nonlinearity, high level of parallelism, noise and fault tolerance, and learning and generalization capabilities (Basheer et al., 2000). Therefore, it is a useful method in providing prognostic values for the effectiveness of infertility treatment (Milewski et al., 2009; Siristatidis et al., 2011). There are many types of ANNs that are classified according to used structure, network parameters or selected learning algorithms (Osowski, 2009).

To create the best network, the algorithm was run 30 000 times with random startup parameters (type of network, error function, activation functions, an initial value of the generator) each time. Based on the quality of training, testing and validation, the following classifying network was chosen: three-layer perception with 40 neurons in the input layer, 6 neurons in the hidden layer and two neurons in the output layer (for 26 input variables and one output variable). Table 5 contains the parameters of the selected network.

Table 5. Characteristics of selected ANN

Network name	Quality (training)	Quality (validation)	Quality (testing)	Training algorithm	Error function	Activation (hidden)	Activation (output)
MLP 40-6-2	63.777	65.217	64.251	BFGS 21	Entropy	Logistic	Softmax

Based on the created network MLP 40-6-2 the probability of pregnancy for the each patient (MLP_pr) was set up. This parameter was used as a predictor for the analysis. Figure 2 shows the ROC curve established for the MLP_pr. The Area Under the Curve, along with the 95% Confidence Interval and Standard Error values, is presented in Table 6.

Table 6. Characteristics of the ROC Curve created for the Artificial Neural Networks model

AUC	Std. Error	95% Confidence Interval	
0.703	0.014	0.676	0.730

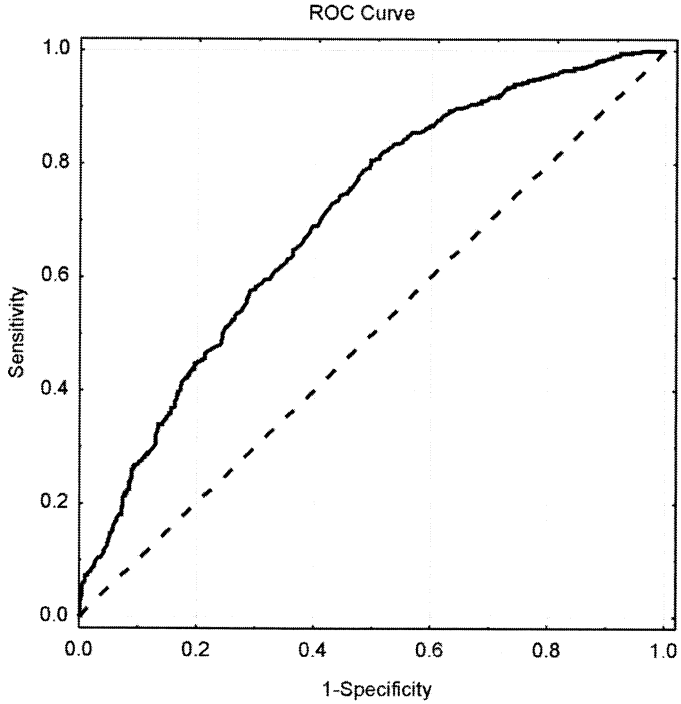


Figure 2. The ROC Curve for the Artificial Neural Networks model

The cut-off point for the probability value produced by the model (`MLP_pr` = 0.477) was established by using the minimum sum of squared components method. The sensitivity and specificity were 69.0% and 60.3%, respectively. It means that our model correctly predicts about 70% of pregnancies and misdiagnoses lack of pregnancy less than 40% of the time.

Comparison of Models and Conclusions

Comparing predictive power of the two studied models, the better results were obtained with the ANNs technologies. The results of the comparison of the AUC for the ROC curves of these two models are shown in Table 7. The statistically significant differences between the predictive powers of both models were at the $p < 0.0001$ level. Comparing the ROC curve in Figure 1 with the curve in Figure 2, it is evident that the shape of the curve for the ANNs is more convex then other one. The Area Under the Curve, sensitivity and specificity for the ANNs were higher (10%, 3.5% and 3.5%, respectively) than for the multivariate logistic regression model.

Table 7. Comparison of two predictors

AUC MLP_pr	0.7026
AUC fit_pr	0.6423
SD MLP_pr	0.0139
SD fit_pr	0.0147
AUC difference	-0.0603
SD AUC difference	0.0116
95% CI AUC difference '1	-0.0829
95% CI AUC difference '2	-0.0376
p-value	<0.0001

To obtain a predictive model for clinical treatment, the ANNs technologies are better suited than classical statistical analysis. However, they are not able to detect which variable and to what degree it influences the final results. In contrast, the logistic regression analyses allow for the selection of variables that affect treatment and their statistical power. Thus, for clinical prediction purposes, the ANNs technologies are better to apply but classical technology (in this case logistic regression analysis) is more appropriate for theoretical (scientific) purposes.

REFERENCES

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods*, 43(1), 3–31.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36.

Hanley, J. A., & Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. *Academic Radiology*, 4(1), 49–58.

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

Milewski, R., Jamiolkowski, J., Milewska, A. J., Domitrz, J., Szamatowicz, J., & Wolczynski, 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., & Wolczynski, 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., Malinowski, P., Milewska, A. J., Ziniewicz, P., Czerniecki, J., Pierzynski, P., & Wolczynski, S. (2012). Classification issue in the IVF ICSI/ET data analysis. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 29(42), 75–85.

Milewski, R., Milewska, A. J., Czerniecki, J., Lesniewska, M., & Wolczynski, S. (2013). 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., & Wolczynski, S. (2008). In vitro fertilization ICSI/ET in women over 40. *Przeglad Menopauzalny*, 7(2), 85–90.

Milewski, R., Pierzynski, P., Milewska, A. J., Zbucka-Kretowska, M., & Wolczynski, S. (2012). Concept of system allowing non invasive detection of uterine contractions in women undergoing in vitro fertilization – embryo transfer treatment. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 29(42), 99–106.

Osowski, S. (2009). Artificial neural networks – basic network structures and learning algorithms. *Przeglad Elektrotechniczny*, 85(8), 1–8.

Radwan, J. (2011). Epidemiologia nieplodnosci. In J. Radwan, & S. Wolczynski (Eds.), *Nieplodnosc i rozrod wspomagany* (pp. 11–14). Poznan: Termedia.

Siristatidis, C., Pouliakis, A., Chrelias, C., & Kassanos, D. (2011). Artificial intelligence in IVF: A need. *Systems Biology in Reproductive Medicine*, 57(4), 179–185.

Tadeusiewicz, R. (1993). *Sieci neuronowe*. Warszawa: Akademicka Oficyna Wydawnicza RM.

Thoma, M. E., McLain, A. C., Louis, J. F., King, R. B., Trumble, A. C., Sundaram, R., & Buck Louis, G. M. (2013). Prevalence of infertility in the united states as estimated by the current duration approach and a traditional constructed approach. *Fertility and Sterility*, 99(5), 1324–1331e1.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.). San Francisco, CA, USA: Morgan Kaufmann.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0039

A Computer-Aided Diagnosis of Liver Tumors Based on Multi-Image Texture Analysis of Contrast-Enhanced CT. Selection of the Most Appropriate Texture Features

Dorota Duda[1], Marek Krętowski[1], Johanne Bézy-Wendling[2,3]

[1] Faculty of Computer Science, Bialystok University of Technology, Poland

[2] Signal and Image Processing Laboratory (LTSI), University of Rennes 1, France

[3] National Institute of Health and Medical Research (INSERM), University of Rennes 1, France

Abstract. In this work, a system for the classification of liver dynamic contest-enhanced CT images is presented. The system simultaneously analyzes the images with the same slice location, corresponding to three typical acquisition moments (without contrast, arterial- and portal phase of contrast propagation). At first, the texture features are extracted separately for each acquisition moment. Afterwards, they are united in one "multiphase" vector, characterizing a triplet of textures. The work focuses on finding the most appropriate features that characterize a multi-image texture. At the beginning, the features which are unstable and dependent on ROI size are eliminated. Then, a small subset of remaining features is selected in order to guarantee the best possible classification accuracy. In total, 9 extraction methods were used, and 61 features were calculated for each of three acquisition moments. 1511 texture triplets, corresponding to 4 hepatic tissue classes were recognized (hepatocellular carcinoma, cholangiocarcinoma, cirrhotic, and normal). As a classifier, an adaptive boosting algorithm with a C4.5 tree was used. Experiments show that a small set of 12 features is able to ensure classification accuracy exceeding 90%, while all of the 183 features provide an accuracy rate of 88.94%.

Introduction

In *Global Cancer Statistics*, Jemal et al. (2011) reported that "Liver cancer in men is the fifth most frequently diagnosed cancer worldwide but the second most frequent cause of cancer death. In women, it is the seventh most commonly diagnosed cancer and the sixth leading cause of cancer death". According to the mentioned research, in 2008, there were estimated to be 748 300 new liver cancer cases in the world, and 695 900 people died from liver cancer. Moreover, Jemal et al. (2011) revealed that the incidence and mortality rate of primary liver cancers were increasing across many

ISBN 978-83-7431-392-6 ISSN 0860-150X

parts of the world. Most patients diagnosed with primary liver cancer die within six months of diagnosis. In this context, the earliest possible detection of such a disease becomes critical to successful treatment.

In clinical practice, a preliminary diagnosis of liver disorders is usually based on several contrast-enhanced Computed Tomography (CT) scans. The first series of images are acquired without contrast media injection. The next two series concern two scenarios (*i*) when injected contrast media reaches the liver through the hepatic artery (*hepatic phase*) and (*ii*) when it reaches the liver through the portal vein (*portal phase*). Thus, an enhancement of two vascular trees (branching from the hepatic artery and from the portal vein) is possible. An evolution of the liver tissue region appearance, over the contrast media propagation, could be a discriminating factor in tumor diagnosis.

Visual analysis of liver CT scans, performed by an experienced radiologist, often is not sufficient to correctly recognize the type of pathology. Due to the fact that those performing the analysis are able to identify only a small part of information stored in images, invasive techniques (such as a needle biopsy) still remain a gold standard for a definitive diagnosis of hepatic disorders. The use of invasive procedures could be avoided if doctors had the appropriate tools to interpret the image content. The solution could be the image-based Computer Aided/Assisted Diagnosis/Detection (CAD) systems, which have recently and rapidly become of growing interest. They include many techniques of image analysis, such as organ segmentation, lesion extraction, and tissue characterization, often based on texture analysis (Bruno et al., 1997), combined with classification algorithms. A large number of publications on the subject proves that the (semi)automatic CAD systems appear to be a powerful tool for supporting medical decisions.

Two main stages of work of a typical image-based CAD system exist, regardless of the imaging technique, the organ analyzed, or the possible diseases to diagnose. The first stage is a system preparation for recognition of a certain number of tissue classes. This stage, called *learning* (or *training*), consists in the classifiers' induction from a set of labeled vectors of features. The learning set is created on the basis of the tissue regions, traced on the images, for which a diagnosis has been verified. The second stage is an application of the classifiers to aid a diagnosis.

Our goal is to develop a (semi)automatic CAD system for aiding a diagnosis of hepatic diseases from dynamic contrast-enhanced CT images. The system we are working on simultaneously analyzes the triplets of images of the same slice of the liver, corresponding to the three different moments of contrast media propagation. Several CAD systems based on liver CT im-

ages have already been investigated. Nevertheless, systems taking texture evolution into account over contrast media propagation in hepatic vessels are still rare. In this study, we focus on the choice of texture features that best describe the "triphase" liver texture.

The rest of the paper is organized as follows: at first, we give a short overview of the proposed CAD systems concerning the hepatic diseases diagnosis, based on CT images. Afterwards, we detail the two stages of work of our system. We also discuss some strategies for choosing the most appropriate texture features for triphase texture characterization: (*i*) an assessment of a feature stability and of its dependency on the size of the analyzed image region, (*ii*) feature selection with two searching directions – *Forward* and *Backward*. Then, the classification results obtained with selected features are discussed. Finally, the conclusions and future works are outlined.

Review of Image-Based CAD Systems Based on Liver CT Images

One of the earliest studies concerning a computer-assisted diagnosis based on liver CT images was undertaken by Chen et al. (1998). Their system was able to automatically find the liver boundaries and to recognize two types of liver tumors: hepatoma and hemangioma. In this system, the image texture was characterized by its fractal dimension and the features obtained from the co-occurrence matrices. A probabilistic neural network was used as a classifier. A similar system, but one able to distinguish a healthy liver and liver disease, was later presented by Husain et al. (2000).

In 2003, Gletsos et al. described a system adapted for recognizing the four types of liver tissue: healthy, hepatic cysts, hemangioma, and hepatocellular carcinoma (HCC). They characterized the textures with the features calculated from the co-occurrence matrices. The classifier was composed of three sequentially placed feed-forward neural networks, trained with a back-propagation algorithm. The same tissue types were recognized in the study presented by Stoitsis et al. (2006). Their system tested several sets of texture features, derived from: the gray-level histogram, the co-occurrence matrices, the run-length matrices, the Laws' texture energy method, and the fractal models. A feature selection, based on genetic algorithms, was performed in order to find the most useful features. Classification was carried out by neural networks and statistical methods. Further continuation of this research has resulted in the creation of a telematics-enabled system for image archiving, management, and diagnosis support (Mougiakakou et al., 2009).

This integrated system performed an image preprocessing, a semi-automatic image segmentation, an extraction of texture features, and a classification.

All of the aforementioned systems processed only one image at a time, acquired without contrast media. Later systems used contrast-enhanced images, but they were still adapted for the analysis of a single image. For example, in 2004, Bilello et al. presented a system working on portal-phase images. It combined the methods for detection, characterization and classification of liver hypodense hepatic tissue (cysts, hemangiomas, and metastases). The texture analysis was performed with frequency methods. As classifiers, the support vector machines were used.

The system described by Smutek et al. (2006) focused on the analysis of focal liver lesions. It used the first- and second order texture features. The analyzed images corresponded to the late portal phase. Another system, developed by Lambrou et al. (2006), differentiated healthy and tumor tissue. To extract texture features, it used a wavelet transform method, in combination with three statistical methods (based on the gray level histogram, the co-occurrence matrices, and the run length matrices). In both systems, an ensemble of Bayesian classifiers was applied.

Still, in 2006, Mala et al. described a system adapted for a recognition of HCC, cholangiocarcinoma, hemangioma, and hepatic adenoma. Their system was able to automatically detect regions affected by a disease, characterize a tissue (using methods based on wavelet transform), select the best texture features, and finally classify tissues, using a probabilistic neural network.

In 2009, Wang et al. tested yet another diagnostic system, which worked with the three types of liver tissue: HCC, hemangioma, and normal tissue. This system used four texture analysis methods (based on the gray level histogram, the co-occurrence matrices, the gray level difference matrices, and the run length matrices), and the support vector machines as classifiers.

Finally, in the work presented by Duda et al. (2004), a simultaneous analysis of the three images, corresponding to typical moments of contrast propagation (without contrast, arterial phase, portal phase) was proposed. At first, the three corresponding textures were characterized separately by features obtained from: the gray level histogram, Laws' texture energy method, the co-occurrence matrices, and the run length matrices. Then, the features corresponding to the three related textures were placed together in a one feature vector, characterizing a triphase texture. Three types of liver tissue were recognized: healthy, HCC, and cholangiocarcinoma. The classification results obtained with the triphase textures were significantly better than the results corresponding to each acquisition moment separately. Fur-

ther work on similar data has confirmed that considering texture evolution over contrast media, propagation could considerably improve classification accuracy (Duda et al., 2006).

The idea of a multiphase texture characterization was also exploited by Quatrehomme et al. (2013). In this case, four acquisition moments were considered: the first one was a pre-injection phase, the next three corresponded to arterial, portal, and late phase of contrast media propagation. Five types of hepatic lesions were classified: cysts, adenomas, hemangiomas, HCC, and metastasis. The multiphase vectors were composed of features calculated separately for four acquisition moments and united in a multiphase vector. The results obtained for the "four-phase" textures were significantly better than the single-phase ones. However, this work, like the two previous ones, did not select the most relevant features for each of the considered acquisition moments.

In the work described by Ye et al. (2009) the quadruples of images were analyzed in order to differentiate four tissue classes: normal, cyst, hemangioma, and HCC. Only some combinations of the mean pixel values were considered as temporal features. Nevertheless, they did not outperform the texture features (based on the gray level histogram and the co-occurrence matrices) considered separately for each of four phases.

Methods

The two previously described stages of work can also be distinguished in the system that we are developing. The first stage – the construction of the classifiers from the preprocessed database of image triplets – is depicted in Figure 1. After a database creation, triplets of images are formed. An (ordered) triplet contains the images acquired at the same slice location. Each of the images in the triplet corresponds to a different moment of contrast media propagation in the hepatic vessels. The first of them is acquired without contrast, the second and the third – after its injection, in the two typical phases of its propagation, arterial and portal, respectively. The next step, the preprocessing of images, could be optional. It aims at improving the contrast, eliminating the noise or the artifacts. Then, a Region of Interest (ROI) is drawn on each of the three images. The three corresponding ROIs are of the same size and of the same anatomical position. Afterwards, a label is attributed to each triplet of ROIs. It refers to a tissue class, determined on the basis of a verified diagnosis, for example, confirmed by a histopathological study. Then follows a tissue characterization. It is based on the texture

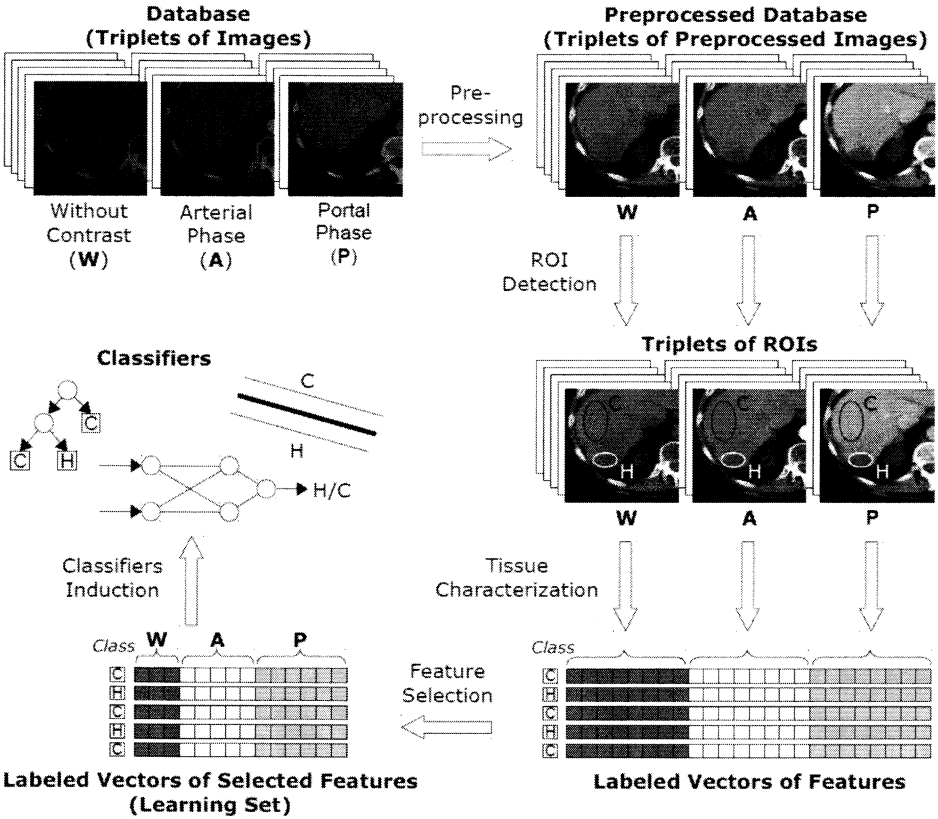


Figure 1. The system for texture-based classification of liver tissues. First stage of work: a construction of the classifiers from the preprocessed database of triplets of images. C and H stand for liver tissue classes, cirrhosis and HCC, respectively

analysis and consists of calculating a set of numerical parameters in order to measure different texture properties, e.g. coarseness, homogeneity, entropy, or local contrast. Such numerical descriptions of the texture are called *texture features*. The tissue characterization is firstly made separately for each of the three corresponding ROIs. Each of the ROIs is characterized with the same features. Then, features corresponding to the three ROIs (thus – to the three acquisition moments) are placed together in the one triphase complex vector, characterizing the triplet of ROIs. This vector contains the information about tissue properties that change over different contrast product concentrations in the vessels. The set of labeled feature vectors is called the training (or the learning) set. Often, not all features are equally useful for a tissue description. For this reason, the selection of the most suitable ones

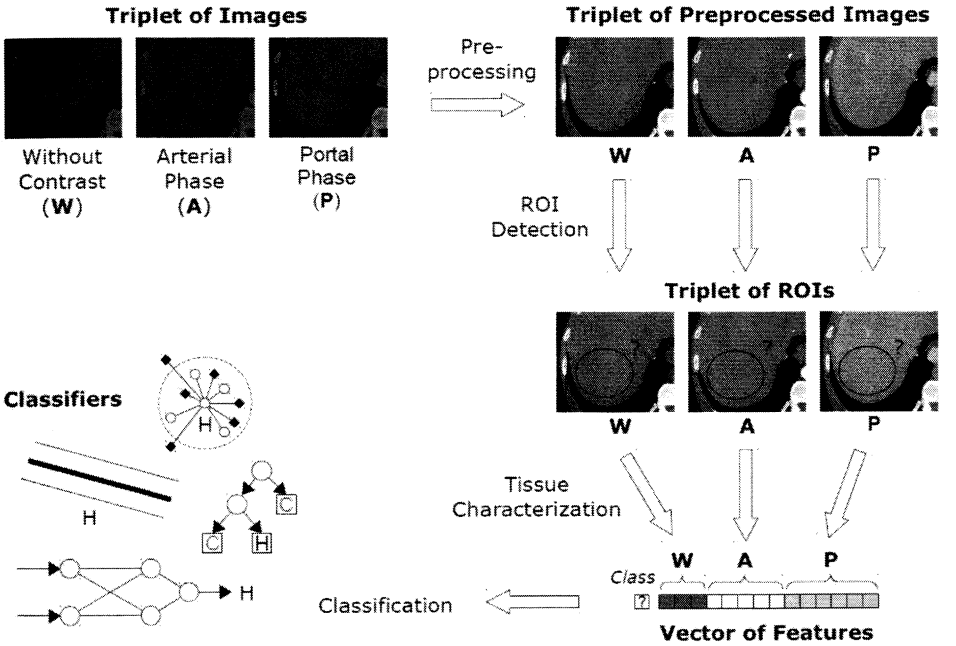


Figure 2. The system for texture-based classification of liver tissues. Second stage of work: application of the classifiers to aid a diagnosis. C and H stand for liver tissue classes, cirrhosis and HCC, respectively

is made. The subsets of features selected for different acquisition moments could differ. Finally, one or several classifiers are constructed on the basis of the training set, composed of vectors of selected features.

The second stage of the system work, i.e., its application to aid a diagnosis, is presented in Figure 2. At this stage, a triplet of images, visualizing the same part of liver, but in the three typical phases of contrast propagation, is necessary. If the image preprocessing was performed at the first stage, now it is also performed, in order to obtain the same properties of the images, as they were in a previous stage. Then, a triplet ROIs is traced, one ROI on each image. Each ROI is characterized with the features that were selected in the previous stage. Features obtained for the three ROIs are placed in one triphase vector. Finally, the classifiers are used in order to propose the most probable tissue class. Such a class is one of the classes considered in the first stage of system work.

One of the major problems encountered during the construction of such systems is a choice of features for the proper tissue characterization. So far, a wide variety of methods for texture feature extraction have been proposed.

They allow the calculation of tens or even hundreds of features describing the various properties of visualized tissue. There is no such set of features that fits each problem, regardless of imaging techniques, acquisition conditions, or the organ that is visualized. However, applying all available features could be impossible for many reasons. Too many features, especially far in excess of the number of objects, could result in data overfitting. The presence of redundant or not reliable features could diminish the classification accuracy. Moreover, handling a relatively big set of observations, described by a large number of features, requires considerable memory resources and can be very time consuming. In the next part of the work, some important aspects for choosing the most relevant features will be discussed.

The ability to properly characterize tissue could be preliminarily assessed taking into account the instability (or the stability) of the feature. If a slight displacement of the ROI results in a significant change in the feature value, the feature can be considered unstable, thus not reliable in the tissue characterization process.

Another way to evaluate the usefulness of the feature could be assessment of its dependency on a ROI size. This is particularly important when the ROI size is different for different patients or for different images (which is true in the case of the database that we explore).

The instability of a feature and its dependency on the ROI size can be determined by a standard coefficient of variation (CV). The coefficient of variation is calculated from a set of feature values, obtained for ROIs located at very close positions, or having almost the same size. It expresses the importance of the variability of several feature values (given by the standard deviation) compared to the absolute value of their average. The more unstable the feature is, the greater its coefficient of variation and the lower its reliability in the process of tissue characterization. We consider that the CV of a feature should not exceed a certain threshold, in order to consider the feature as stable.

In our work, we use the same sets of square ROIs in order to assess both the stability of the feature and its dependency on the ROI size. In each case, the coefficient of variation is calculated on the basis of the same number of feature values obtained, respectively, from ROIs of almost the same locations (approach *Displace*) or from the ROIs of almost the same sizes (approach *Size Changing*). For the first approach, the initial ROI size is first slightly decreased, then the reduced ROI is successively displaced, in order to take all the possible positions within its initial boundaries. For the second approach, the successive ROI vertices are moved by one pixel towards the ROI center, in order to obtain smaller and smaller square ROIs.

The coefficient of variation, obtained with the approaches *Displace* and *Size Changing* is denoted, respectively, CV_D and CV_S. Only the features with relatively small values of both CV_D and CV_S, thus insensitive both to small ROI displacements and to small changes in ROI size, are considered in further analyses.

After identifying the features with a relatively low coefficient of variation, the next step is to assess which subset of them ensures the best possible classification accuracy. When a set of available texture features is large, an evaluation of all of its possible subsets is impossible; the number of subsets of an N-element set is 2^N. Also, there can exist many different feature combinations that ensure comparable classification results. Due to these facts, it is possible and sufficient to test only a small part of the whole subset space.

We perform the feature selection using two searching directions (*Forward*, and *Backward*) in order to test a certain number of feature subsets. Each time, the quality of candidate feature subset is assessed by the classification accuracy ensured with the proposed subset. When the classification accuracy does not grow significantly while adding (in the *Forward* method) or eliminating (*Backward*) the new features, the search stops.

In our work, we propose repeating the following procedure many times: At first, a certain part of observations is randomly chosen from the initial set of observations. Afterwards, a feature selection is performed only on this part of chosen observations. After a multiple repetition of this procedure, we count how often each feature occurs in the final subset of selected features. On the basis of this "feature incidence frequency" a ranking of features is made. Then, we examine how many first features from the ranking are sufficient to ensure the best possible classification accuracy, using the whole set of observations. We suspect that such an approach ensures a better generalization than one single execution of feature selection on the entire observation set. We conduct our experiments separately for *Forward* and *Backward* searching directions.

Experimental Setup

The images were gathered in two hospitals in Rennes (France): Eugene Marquis Anticancer Center (Department of Medical Imaging, Radio Diagnosis Unit) and the University Hospital Pontchaillou (Department of Radiology). The images from the first center were acquired on the *HiSpeed NX/i* device, produced by *GE Medical Systems*. In total, studies on 28 patients

from this center were available. 21 of them were characterized by a slice thickness of 7 mm, the remaining 7 – by a slice thickness of 5 mm. The images from the second center were acquired on the *LightSpeed16* device (*GE Medical Systems*). They corresponded to 48 patients, and were all characterized by a 5 mm slice thickness.

All the images from both hospitals (about 100 images for each patient examination) were acquired using a helical scanner. A standard amount of contrast agent (100 ml) was injected into the patient's arm vein, at a rate of 4 ml/s. The acquisition of images corresponding to the arterial phase of the contrast agent propagation started about 20 seconds after the beginning of the injection. The images corresponding to the portal phase were acquired from 30 to 40 seconds later.

The images were initially stored in DICOM format, of 4096 gray levels. The minimum gray level (zero) corresponded to –2048 Hounsfield Units (HU) and the maximum gray level (4095) corresponded to 2047 HU. All the images were of the size 512×512 pixels. Their reconstruction diameter was between 330 mm and 500 mm.

Four classes of liver tissue were distinguished: HCC (the most frequently observed primary liver tumor), cholangiocarcinoma, the cirrhotic liver, and the healthy liver.

ROIs were manually drawn, as large as possible, on the images, avoiding the big vessels. They entered into the two disjoint ROI sets. The first set (named *Square-ROI-Set*), composed of only square ROIs of the same size (60×60 pixels), was used for assessing the stability of the texture features. It included 62 ROIs for each of: 2 available slice thicknesses, 4 tissue classes, and 3 acquisition moments ($2 \cdot 4 \cdot 3 \cdot 62 = 1488$ ROIs in total). The second one (named *Circular-ROI-Set*) was composed of only circular ROIs with a diameter ranging from 20 to 130 pixels. This ROI set was created on the basis of 2961 images (987 images for each of the three acquisition moments). In total, it included 4533 ROIs (1511 triplets of ROIs), $3 \cdot 319$ ROIs for HCC, $3 \cdot 222$ ROIs for cholangiocarcinoma, $3 \cdot 433$ ROIs for the cirrhotic liver, and $3 \cdot 537$ ROIs for the normal liver.

All the considered images were pre-processed. Since it was found that the range of pixel values characterizing the pixels belonging to all considered ROIs was less than $256 = 2^8$ (more precisely, it was 248), the DICOM images were converted to a 8-bit BMP format. Such a conversion (*windowing*) was done taking the *window width* of 256 levels (HU), and the *window center* of 70 HU. Thus, all the differences in gray levels between all the pixel pairs belonging to considered ROIs were preserved in the resulting images.

On the basis of 8-bit BMP images, 61 texture features were calculated

separately for each of the three acquisition moments. For this purpose, the *MIP (Medical Image Processing) Application* was used. This application was created at the Faculty of Computer Science, Bialystok University of Technology (Poland) and at the Signal and Image Processing Laboratory, Rennes 1 University (France). One of its modules, *Texture Analyzer*, enabled us to extract the texture features, based on the following methods:

- First Order statistics, abbreviated FO,
- Gradient based, GB,
- Co-Occurrence Matrices, COM (Haralick et al., 1973),
- Run Length Matrices, RLM (Chu et al., 1990; Galloway, 1975),
- Gray Level Difference Matrices, GLDM (Weszka et al., 1976),
- Laws Texture Energy, LTE (Laws, 1980),
- Fractal based, FB (Chen et al., 1989),
- Texture Feature Coding Method, TFC (Horng et al., 1996),
- Autocorrelation, AC (Gonzalez et al., 2002).

Table 1 contains the names of features obtained from each method.

Table 1. Texture features chosen for evaluation

Method	Features
AC	*Autocorr*
COM	*AngSecMom, InvDiffMom, Entr, Corr, Contrast, DiffAvg, DiffEntr, DiffVar, SumAvg, SumEntr, SumVar*
FB	*FractalDim, FractalArea*
FO	*Avg, Var, Skew, Kurt*
GB	*GradAvg, GradVar, GradSkew, GradKurt*
GLDM	*DAvg, DEntr, DAngSecMom, DInvDiffMom, DContrast*
LTE	*E3L3, S3L3, S3E3, E3E3, S3S3, E5L5, S5L5, W5L5, R5L5, S5E5, W5E5, R5E5, W5S5, R5S5, R5W5, E5E5, S5S5, W5W5, R5R5*
RLM	*Fraction, RLNonUni, GLNonUni, LongEmp, ShortEmp, LowGLREmp, HighGLREmp, RLEntr*
TFC	*MeanConv, CodeEntr, Coarse, Hom, CodeVar, ResSim, CodeSim*

When applying the COM, GLDM, and RLM methods, the number of gray levels was reduced from 256, used initially, to 64. Other methods used the whole range of 256 levels.

The co-occurrence matrices and the gray level difference matrices were constructed separately for 4 standard directions (0°, 45°, 90°, 135°) and for 5 different distances between the pixel pairs, going from 1 to 5. From each

of 20 thus obtained matrices, the same features were calculated, 11 features by the COM method and 5 features by the GLDM method. Then, 20 values of the same feature, corresponding to different directions and pixel distances, were averaged in order to obtain only one value per feature. The run-length matrices were also constructed for 4 standard directions. Each of them served to calculate the 8 features. The 4 values of the same feature, obtained for different directions of pixel runs, were averaged.

The normalized autocorrelation coefficients (AC method) and the 2 among 7 TFC features (*CodeEntr*, *CodeSim*) were also calculated separately for 4 standard directions, and for 5 different pixel distances, going from 1 to 5. Then, an average of 20 feature values were used to characterize a liver tissue. The remaining 5 TFC features were obtained by averaging 4 values of the same feature, calculated separately for 4 standard directions.

The FB method was based on the fractional Brownian motion model (Chen et al., 1998) and considered 4 pixel distances (1, 2, 3, and 4).

The LTE method provided 19 features, obtained either by the application of 24 filtering masks of 5×5 size or the application of 8 masks of 3×3 size. In the first case, 4 symmetric masks and 10 pairs of asymmetric ones, each pair consisting of a mask and its transposition, were used, in the second – 2 symmetric masks and 3 pairs of asymmetric ones (a mask and its transposition). The sum of elements of each convolution matrix was equal to zero. For each pair of asymmetric masks, the resulting images were added. Images obtained with an application of symmetric masks were multiplied by two. Finally, the entropies of 14 or 5 thus obtained images (with, respectively, 5×5 masks, and 3×3 masks) served as texture features.

In order to assess the feature stability and its dependency on the ROI size, the *Square-ROI-Set* was used. The assessment of the coefficient of variation was performed separately for each of its 24 subsets. Two approaches were separately applied for calculating the coefficient of variation for each feature: *Displace* (giving the CV_D value) and *Size Changing* (for CV_S). The coefficient of variation was always calculated on the basis of 9 feature values. In the *Displace* approach, the ROI was reduced to a 58×58 square in order to take the 9 possible positions inside its initial boundaries. In the *Size Changing* approach, the 9 considered ROIs were of sizes going from 60×60 to 52×52. For each of the 24 subsets, the average of 62 values of coefficient of variation, obtained separately for each of 62 ROIs, was used for further analyses.

Next, the experiment consisted of selecting a subset of features ensuring the best possible classification accuracy. In this step, only the triplets of

ROIs from the *Circular-ROI-Set* were used. Each triplet was described by 120 texture features, 40 features for each acquisition moment. Only the features insensitive both to small ROI displacements and to small changes in ROI size were considered. This experiment was performed with the *Weka* software (Hall et al., 2009). The following selection settings were applied: the wrapper method – as an evaluator of each tested subset of features (*WraperSubserEval*), the *C4.5* tree (Quinlan, 1993) (called *J48* in *Weka*), as a classifier, and the *BestFirst* searching strategy with the two searching directions, *Forward* and *Backward* (tested separately). The selection was repeated 100 times. Each time $1007 = 1511 \cdot 2/3$ multiphase feature vectors were randomly chosen for the selection experiment. On the basis of the 100 obtained subsets of selected features, the feature incidence frequency ranking was made.

Finally, for the classification experiment, the entire *Circular-ROI-Set* was used. Each triplet of ROIs was described only by several or several dozen selected texture features, occupying the first positions of the feature incidence frequency ranking. Different numbers of first ranked features were used. The classification was performed with *Weka*, using an *Ensemble of Classifiers* with an adaptive boosting voting scheme (Freund et al., 1997), called *AdaBoostM1* in *Weka*, and a *C4.5* tree (*J48*) as the underlying algorithm. The number of iterations for the *AdaBoostM1* algorithm was 100. The classification accuracy was estimated by 10-fold cross-validation, repeated 10 times.

Results and Discussion

First, we assessed the feature stability and its dependency on the ROI size. Since the presentation of all the coefficients of variation obtained for each of the 61 features (24 values of CV_D, and 24 values of CV_S) would occupy too much space, we limit ourselves, for the moment, to the presentation of general conclusions, drawn from the whole set of results.

- Regardless of the approach (*Displace* or *Size Changing*) and, as far as the same slice thickness is considered, the three averaged coefficient of variation values (CV_D or CV_S), obtained for the three acquisition moments do not differ significantly. Nor do they differ between 4 tissue classes. However, for the *Size Changing* approach, the highest of 4 CV_S values, obtained for different classes, but corresponding to the same slice thicknesses and the same acquisition moment, is observed slightly more frequently for the tumor tissue.

Table 2. Maximum feature CV value (among the values corresponding to 2 approaches, 3 acquisition moments, and 4 classes) obtained for different slice thicknesses: 7 mm and 5 mm. The features are sorted by the increasing maximal CV values, corresponding to the slice thickness of 7 mm. Only the CV values not exceeding a threshold of 0.01 for both slice thicknesses are taken into account

Rank	Feature	7 mm	5 mm	Rank	Feature	7 mm	5 mm
1	*Autocorr*	0.0009	0.0010	21	*S3S3*	0.0034	0.0063
2	*FractalDim*	0.0013	0.0018	22	*DInvDiffMom*	0.0034	0.0049
3	*E3E3*	0.0014	0.0023	23	*Avg*	0.0036	0.0037
4	*CodeEntr*	0.0014	0.0040	24	*InvDiffMom*	0.0036	0.0047
5	*S5L5*	0.0016	0.0024	25	*W5S5*	0.0036	0.0079
6	*S3L3*	0.0017	0.0026	26	*MeanConv*	0.0037	0.0069
7	*E5E5*	0.0017	0.0025	27	*SumAvg*	0.0039	0.0039
8	*S5S5*	0.0017	0.0032	28	*CodeVar*	0.0044	0.0081
9	*ShortEmp*	0.0018	0.0017	29	*GradAvg*	0.0052	0.0060
10	*W5E5*	0.0019	0.0033	30	*DiffEntr*	0.0052	0.0034
11	*E3L3*	0.0021	0.0025	31	*DEntr*	0.0052	0.0033
12	*Fraction*	0.0022	0.0020	32	*R5R5*	0.0054	0.0083
13	*W5L5*	0.0022	0.0028	33	*DAngSecMom*	0.0054	0.0062
14	*R5S5*	0.0025	0.0050	34	*Entr*	0.0055	0.0042
15	*S3E3*	0.0026	0.0059	35	*RLEntr*	0.0058	0.0041
16	*R5E5*	0.0026	0.0050	36	*LongEmp*	0.0062	0.0053
17	*E5L5*	0.0027	0.0030	37	*R5W5*	0.0064	0.0089
18	*R5L5*	0.0027	0.0043	38	*DAvg*	0.0070	0.0072
19	*S5E5*	0.0030	0.0044	39	*SumEntr*	0.0072	0.0050
20	*W5W5*	0.0030	0.0054	40	*HighGLREmp*	0.0074	0.0076

– The CV_D values are almost always lower than the corresponding CV_S values. This may indicate that studied features are more influenced by the ROI size than by the ROI position. An exception to this rule is observed for the most unstable features (with the largest values of coefficient of variation), such as *Skew* (FO method) or *GradKurt* (GB). We can thus conclude that one should be particularly careful in the choice of features when analyzing the ROIs of different sizes.

As a measure of the feature stability, we finally took the maximum of the 24 CV values, obtained for two different approaches, three acquisition moments and four tissue classes. They are listed in Table 2. For each feature,

Table 3. Ranking of features according to their frequency rate [%], obtained with the *Forward* searching direction. Superscript indices indicate the corresponding acquisition moment. Only the first 33 features are considered

Rank	Feature	Frequency
1	$Avg^{(A)}$	3.94
2	$R5S5^{(P)}$	3.54
3	$SumAvg^{(W)}$	3.27
4	$LongEmp^{(A)}$	3.07
5	$HighGLREmp^{(A)}$	2.67
6	$SumAvg^{(A)}$	2.60
7	$HighGLREmp^{(W)}$	2.54
8	$Avg^{(W)}$	2.47
9	$R5S5^{(W)}$	1.94
10–11	$Avg^{(P)}$, $SumAvg^{(P)}$	1.87
12	$HighGLREmp^{(P)}$	1.67
13	$S3S3^{(A)}$	1.60
14	$R5E5^{(P)}$	1.40
15–17	$W5S5^{(W)}$, $ShortEmp^{(A)}$, $R5S5^{(A)}$	1.34
18–19	$R5E5^{(W)}$, $Fraction^{(A)}$	1.27
20–21	$R5E5^{(A)}$, $R5L5^{(P)}$	1.20
22–25	$W5W5^{(W)}$, $W5W5^{(A)}$, $RLEntr^{(P)}$, $S3E3^{(W)}$	1.13
26	$Entr^{(W)}$	1.00
27–33	$ShortEmp^{(W)}$, $S5S5^{(W)}$, $MeanConv^{(A)}$, $Fraction^{(P)}$, $R5R5^{(P)}$, $SumEntr^{(A)}$, $E3E3^{(P)}$	0.93

the maximum CV value was found separately for two different slice thicknesses: 7 mm and 5 mm. Only the most stable features (with the CV not greater than a fixed threshold, 0.01, for both slice thicknesses) are included in the table.

For most of the features, the greater stability (expressed in the lower maximal value of the coefficient of variation) can be observed for thicker slices – those with a thickness of 7 mm. An exception to this rule can be found for the majority of RLM features (*ShortEmp*, *Fraction*, *RLEntr*, *LongEmp*) and for the different measures of texture entropy: *Entr*, *DiffEntr*, *SumEntr* (COM method), *DEntr* (GLDM). When the slice thickness diminishes, smaller and smaller details in the vascular structure became

Table 4. Ranking of features according to their frequency rate [%], obtained with the *Backward* searching direction. Superscript indices indicate the corresponding acquisition moment. Only the first 32 features are considered

Rank	Feature	Frequency
1	$HighGLREmp^{(P)}$	1.46
2	$R5R5^{(P)}$	1.43
3	$HighGLREmp^{(W)}$	1.33
4–6	$RLEntropy^{(P)}$, $R5L5^{(P)}$, $CodeEntr^{(W)}$	1.30
7	$FractalDim^{(A)}$	1.29
8	$SumAvg^{(W)}$	1.25
9–10	$R5W5^{(P)}$, $DInvDiffMom^{(P)}$	1.19
11–12	$S5E5^{(A)}$, $SumAvg^{(A)}$	1.18
13	$S3S3^{(P)}$	1.14
14–18	$HighGLREmp^{(A)}$, $FractalDim^{(P)}$, $SumAvg^{(P)}$, $CodeEntr^{(P)}$, $DEntr^{(P)}$	1.13
19–20	$Autocorr^{(A)}$, $Avg^{(P)}$	1.11
21	$RLEntr^{(A)}$	1.10
22–24	$S5E5^{(W)}$, $Avg^{(A)}$, $SumEntr^{(P)}$	1.08
25–26	$LongEmp^{(A)}$, $CodeEntr^{(A)}$	1.07
27–30	$Avg^{(W)}$, $FractalDim^{(W)}$, $SumEntr^{(A)}$, $S3S3^{(A)}$	1.03
31–32	$W5S5^{(A)}$, $R5S5^{(P)}$	0.99

perceptible in the image. Its texture becomes more and more varied, and its "disorder" increases.

In the next experiment (selection of a subset of features ensuring the best possible classification accuracy), we used only the most stable 40 features (Table 2).

In the case of the *Forward* searching direction, the minimum number of selected features was 7, the maximum – 30, and the average number was about 15. The features corresponding to the arterial phase were slightly more frequently selected. On average, they represented 35.45% of the selected features. The features corresponding to the portal phase, with the frequency of 31.98%, followed.

An application of the *Backward* direction resulted in more numerous subsets of selected features. They included between 45 and 91 features. The average number of features was about 64. In this case, the percentages of features corresponding to each acquisition moment were almost similar.

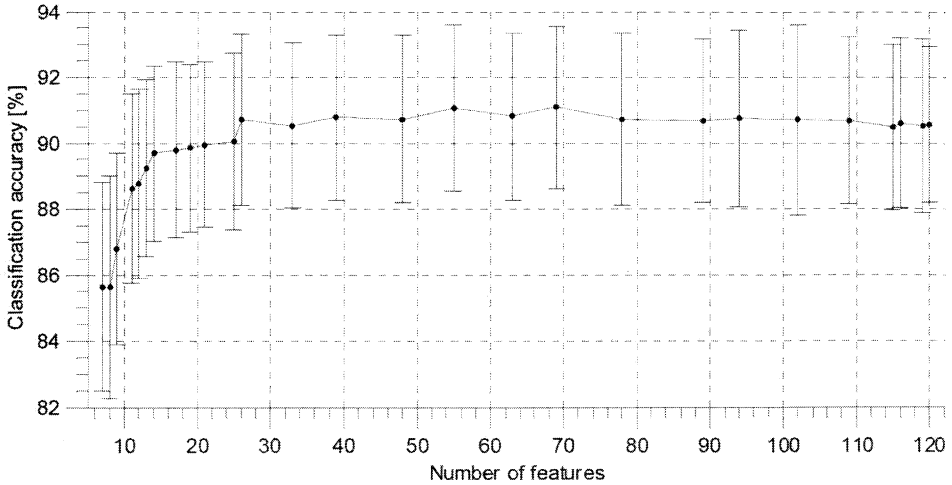


Figure 3. Classification accuracy (with standard deviation) obtained with different numbers of the most frequently selected features with *Forward* direction

The ranking of features, according to their incidence frequency in the entire selection experiment is presented in Tables 3 and 4, for the *Forward* and the *Backward* directions, respectively. For each feature name, superscript indices indicate the corresponding acquisition moment: without contrast (W), arterial phase (A), or portal phase (P). The rankings take into account only a certain number of the best features. In the next experiment, we will see that with such a number of features, it is possible to achieve the highest possible classification accuracy.

Figures 3 and 4 show the classification accuracy obtained for different numbers of first features taken from the rankings presented in Table 3 (concerning the *Forward* direction), and Table 4 (*Backward* direction).

From Figure 3, we can notice that the use of 26 features, most frequently selected with the *Forward* searching direction (Table 3), leads to the correct recognition of 90.71% of the observations. Even if the best possible result (91.10%) is obtained for a set of 69 features, using such a "better" and larger set of features might not be necessary. Taking into account the standard deviations (approximately 2.50%), we can conclude, that the improvement achieved in this case (0.39%) is not significant.

We can also observe that the 26 most selected features derive from only 4 extraction methods: LTE, RLM, COM, and FO (12, 7, 4, and 3 features, respectively). The most represented acquisition moment is, in this case, the arterial phase (10 of 26 features), then follows the no-contrast phase (9 features) and the arterial phase (7 features).

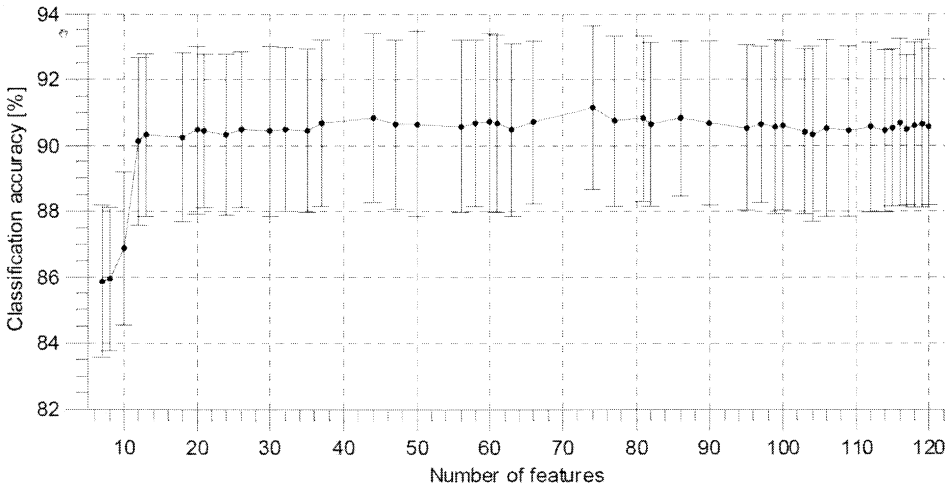


Figure 4. Classification accuracy (with standard deviation) obtained with different numbers of the most frequently selected features with *Backward* direction

In the case of the features most frequently selected with the *Backward* direction (Table 4, Figure 4), the classification accuracy exceeds 90% even with 12 features. The classification results do not change significantly even when more and more features are considered. The best result (91.15% with a standard deviation of 2.49%) is now obtained with the set of 74 features.

This time, the set of the most frequently selected features includes features obtained from the 6 extraction methods: LTE (4 features), RLM (3 features), COM (2 features), GLDM, FB, and TFC (one feature for each of the last three methods). The features corresponding to the portal phase acquisition are more numerously represented in this set. They constitute half of the set. Each of the remaining acquisitions (no-contrast, and the arterial phase one) is represented by 3 features.

For comparison, the classification accuracy obtained with the set of all the $3 \cdot 61 = 183$ features (including the unstable features, and dependent on ROI size) was 88.94%.

Conclusions and Future Work

In this paper, an image-based CAD system was presented. The advantage of the system is its ability to analyze a triplet of images simultaneously. Each image in a triplet visualizes the same liver slice, and corresponds to a different moment of contrast media propagation. The first one is taken

without contrast, the second – in the arterial phase, and the third – in the portal phase. At first, the liver texture in each acquisition moment is characterized. Afterwards – features corresponding to the three acquisition moments are united in one multiphase vector. Thus, the texture evolution over the contrast media propagation is characterized.

The study focused on a choice of the best texture features for description of triplets of textures. 61 features were tested for each of the three acquisition moments (183 features in total). At the first step, $3 \cdot 21$ features that were unstable or dependent on the ROI size were excluded. Then $3 \cdot 40 = 120$ remaining features, united in triphase vectors, were subjected to a selection. The classification experiments showed that a small set of the most frequently selected features (composed of 26 or even 12 features) is able to ensure the classification accuracy that is comparable to, or even better than, the accuracy achieved by using all of the 120 features. The features corresponding to each acquisition moment were different.

In the future, we plan to search for the features that are independent on the image resolution (or on the reconstruction diameter). It seems that it would also be interesting to apply a similar method of multiphase texture characterization for other classification tasks, wherever the organ is visualized repeatedly, each time under different acquisition conditions.

Acknowledgements

We thank Dr. Yann Roland, Dr. Andre Carsin, and Dr. Damien Olivie for their medical contribution to this study. This work was supported by the grant S/WI/2/2013 from Bialystok University of Technology.

REFERENCES

Bilello, M., Gokturk, S. B., Desser, T., Napel, S., Jeffrey, R. B., & Beaulieu, C. F. (2004). Automatic detection and classification of hypodense hepatic lesions on contrast-enhanced venous-phase CT. *Med. Phys.*, 31(9), 2584–2593. DOI: 10.1118/1.1782674.

Bruno, A., Collorec, R., Bezy-Wendling, J., Reuze, P., & Rolland, Y. (1997). Texture analysis in medical imaging. In C. Roux & J.-L. Coatrieux (Eds.), *Contemporary perspectives in three-dimensional biomedical imaging* (pp. 133–164). Amsterdam, Netherlands: IOS Press. DOI: 10.3233/978-1-60750-874-8-133.

Chen, C.-C., DaPonte, J., & Fox, M. (1989). Fractal feature analysis and classification in medical imaging. *IEEE Trans. Med. Imag.*, 8(2), 133–142. DOI: 10.1109/42.24861.

Chen, E.-L., choo Chung, P., Chen, C.-L., Tsai, H.-M., & Chang, C.-I. (1998). An automatic diagnostic system for CT liver image classification. *IEEE Trans. Biomed. Eng.*, 45(6), 783–794. DOI: 10.1109/10.678613.

Chu, A., Sehgal, C., & Greenleaf, J. (1990). Use of gray value distribution of run lengths for texture analysis. *Pattern Recog. Lett.*, 11(6), 415–419. DOI: 10.1016/0167-8655(90)90112-F.

Duda, D., Kretowski, M., & Bezy-Wendling, J. (2004). Texture-based classification of hepatic primary tumors in multiphase CT. In C. Barillot, D. Haynor, & P. Hellier (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI'2004* (Vol. 3217, pp. 1050–1051 Part II). Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-30136-3_133.

Duda, D., Kretowski, M., & Bezy-Wendling, J. (2006). Texture characterization for hepatic tumor recognition in multiphase CT. *Biocybern. Biomed. Eng.*, 26(4), 15–24. Retrieved July 31, 2013, from http://www.ibib.waw.pl/bbe/bbefulltext/BBE_26_4_015_FT.pdf.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1), 119–139. DOI: 10.1006/jcss.1997.1504.

Galloway, M. M. (1975). Texture analysis using gray level run lengths. *Comput. Graph. Image Process.*, 4(2), 172–179. DOI: 10.1016/S0146-664X(75)80008-6.

Gletsos, M., Mougiakakou, S., Matsopoulos, G., Nikita, K., Nikita, A., & Kelekis, D. (2003). A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier. *IEEE Trans. Inf. Technol. Biomed.*, 7(3), 153–162. DOI: 10.1109/TITB.2003.813793.

Gonzalez, R. C., & Woods, R. E. (2002). *Digital image processing* (2nd ed.). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1), 10–18. DOI: 10.1145/1656274.1656278.

Haralick, R., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Trans. Syst. Man Cybern., SMC*-3(6), 610–621. DOI: 10.1109/TSMC.1973.4309314.

Horng, M.-H., Sun, Y.-N., & Lin, X.-Z. (1996). Texture feature coding method for classification of liver sonography. In B. Buxton & R. Cipolla (Eds.), *Computer Vision – ECCV'96* (Vol. 1064, pp. 209–218 Part I). Springer Berlin Heidelberg. DOI: 10.1007/BFb0015537.

Husain, S., & Shigeru, E. (2000). Use of neural networks for feature based recognition of liver region on CT images. In *Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop*, 11–13 December 2000. (Vol. 2, pp. 831–840). New York, USA: The IEEE, Inc. DOI: 10.1109/NNSP.2000.890163.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: A Cancer J. Clin.*, 61(2), 69–90. DOI: 10.3322/caac.20107.

Lambrou, T., Linney, A. D., & Todd-Pokropek, A. (2006). Wavelet transform analysis and classification of the liver from computed tomography datasets. In *Proceedings of the 6th International IEEE EMBS Special Topic Conference*. 26–28 October 2006. Retrieved July 31, 2013, from http://medlab.cs.uoi.gr/itab2006/proceedings/medicalimaging/107.pdf.

Laws, K. I. (1980). *Textured image segmentation.* Unpublished doctoral dissertation, University of Southern California, Los Angeles, California, USA.

Mala, K., Sadasivam, V., & Alagappan, S. (2006). Neural network based texture analysis of liver tumor from computed tomography images. *Int. J. Biol. Life Sci.*, 2(1), 33–40. Retrieved July 31, 2013, from http://www.waset.org/journals/ijbls/v2/v2-1-5.pdf.

Mougiakakou, S., Valavanis, I., Mouravliansky, N., Nikita, K., & Nikita, K. (2009). Diagnosis: a telematics-enabled system for medical image archiving, management, and diagnosis assistance. *IEEE Trans. Instrum. Meas.*, 58(7), 2113–2120. DOI: 10.1109/TIM.2009.2015538.

Quatrehomme, A., Millet, I., Hoa, D., Subsol, G., & Puech, W. (2013). Assessing the classification of liver focal lesions by using multi-phase computer tomography scans. In H. Greenspan, H. Muller, & T. Syeda-Mahmood (Eds.), *Medical Content-Based Retrieval for Clinical Decision Support – MCBR-CDS 2012* (Vol. 7723, pp. 80–91). Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-36678-9_8.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Smutek, D., Shimizu, A., Tesar, L., Kobatake, H., Nawano, S., & Svacina, S. (2006). Automatic internal medicine diagnostics using statistical imaging methods. In *19th IEEE International Symposium on Computer-Based Medical Systems (CBMS'2006)*, 22–23 June 2006 (pp. 405–412). Los Alamitos, California, USA: The IEEE Computer Society Press. DOI: 10.1109/CBMS.2006.56.

Stoitsis, J., Valavanis, I., Mougiakakou, S. G., Golemati, S., Nikita, A., & Nikita, K. S. (2006). Computer aided diagnosis based on medical image processing and artificial intelligence methods. *Nucl. Instrum. Methods Phys. Res., Sect. A*, 569(2), 591–595. DOI: 10.1016/j.nima.2006.08.134.

Wang, L., Zhang, Z., Liu, J., Jiang, B., Duan, X., Xie, Q., Hu, D., Li, Z. (2009). Classification of hepatic tissues from CT images based on texture features and multiclass Support Vector Machines. In W. Yu, H. He, & N. Zhang (Eds.), *Advances in Neural Networks ISNN 2009* (Vol. 5552, pp. 374–381 Part 2). Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-01510-6_43.

Weszka, J. S., Dyer, C. R., & Rosenfeld, A. (1976). A comparative study of texture measures for terrain classification. *IEEE Trans. Syst., Man Cybern., SMC-*6(4), 269–285. DOI: 10.1109/TSMC.1976.5408777.

Ye, J., Sun, Y., Wang, S., Gu, L., Qian, L., & Xu, J. (2009). Multiphase CT image based hepatic lesion diagnosis by SVM. In *2nd International Conference on Biomedical Engineering and Informatics (BMEI'2009)*, 17–19 October 2009 (pp. 1–5). New York, USA: The IEEE, Inc. DOI: 10.1109/BMEI.2009.5304774.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0045

Performance of Resampling Methods Based on Decision Trees, Parametric and Nonparametric Bayesian Classifiers for Three Medical Datasets

Małgorzata M. Ćwiklińska-Jurkowska[1]

[1] Department of Theoretical Foundations of Biomedical Sciences and Medical Computer Science, Collegium Medicum, Nicolaus Copernicus University, Poland

Abstract. The figures visualizing single and combined classifiers coming from decision trees group and Bayesian parametric and nonparametric discriminant functions show the importance of diversity of bagging or boosting combined models and confirm some theoretical outcomes suggested by other authors. For the three medical sets examined, decision trees, as well as linear and quadratic discriminant functions are useful for bagging and boosting. Classifiers, which do not show an increasing tendency for resubstitution errors in subsequent boosting deterministic procedures loops, are not useful for fusion, e.g. kernel discriminant function. For the success of resampling classifiers' fusion, the compromise between accuracy and diversity is needed. Diversity important in the success of boosting and bagging may be assessed by concordance of base classifiers with the learning vector.

Introduction

Combining classifiers with very close discriminant properties is not useful in the models' fusion. Diversity is suspected to be important for the success of merging classifiers (Banfield et al., 2005; Bi, 2011; Brown et al., 2010; Kuncheva et al., 2000; Kuncheva, 2003; Melville et al., 2005), as well for homogenous combining (with the same kind of constituent classifiers) as for heterogonous classifiers (Kuncheva et al., 2002; Shipp et al., 2002). A question arises: how is this diversity realized during the process of resampling datasets according to the most popular fusion procedures: bagging and boosting (Breiman, 1996, 1998; Freund et al., 1997). It is known that for unstable classifiers resampling methods are useful to decrease generalization error in comparison to the single classifier. Popular unstable classifiers are decision trees or neural networks; however, trees are characterized by smaller computational and memory complexity than neural networks. Thus, for resampling ensemble methods such as boosting or bag-

ISBN 978-83-7431-392-6 ISSN 0860-150X

ging (bootstrap aggregating), the most common constituent classifiers are trees. Therefore, most examinations concerning bagging and boosting fusion are connected with decision trees. Few authors consider linear classifiers joined with bagging (Skurichina et al., 2002, Vu et al., 2009) and boosting (Skurichina et al., 2002). Conclusions concerning the usefulness of bagging linear discriminant classifiers are not concordant in the literature. For large data sets, much bigger than the number of variables, Breiman (1996) concluded that bagging LDC is not useful. Also, further works were based on large data sets (Breiman, 1998; Dietterich, 2000). Skurichina et al. (2002) stated that for critical training sample sizes (when the number of training objects is comparable with data dimensionality) a bagging ensemble is useful for LDC, because then LDC is an unstable classifier. Vu et al. (2009) concluded for small sets of microarray data that bagging is useful for unstable trees and neural networks, but not for LDCs. For boosting, however, fusion might be useful for large training sample sizes (Skurichina et al., 2002).

The aim of this work is to examine the usefulness of fusion of trees and other constituent classifiers like Bayesian parametric and nonparametric discriminant functions in the context of diversity, depending on the number of loops, the size and character of the set, the type of constituent classifiers and kind of merging. The visualization of the single and combined Bayesian classifiers are elaborated to examine the trends of learning curves and to aid the exploration of the effectiveness of bagging and boosting fusion based on such base discriminant functions.

Methods

The practical aims of the discriminant analysis pertaining to medical problems are to find variables with the biggest discriminant power, which is useful for differentiation, and next to support medical decisions according to the chosen classification model based on those variables. High performance of classification models confirms the correctness of the selected set of variables. For classifiers based on single or merged trees, selection of variables is incorporated in the modeling step, so is not necessary as the first step, though it may be beneficial. The theoretical aim is to define which of the potential discriminant methods has the lowest misclassification rate. An application of classification methods to three real medical decision problems of which, the most essential information, hopefully, was included into the data sets, was performed (Table 1). Modeling on selected variables sets may support medical decisions for those problems.

Table 1. Characterization of applied medical data sets

Data Set	Medical decision problem, coming from:		Number of of cases	Variables number	Number of groups
WDBC	Malignant or benign tumor of the breast	University of Wisconsin Hospitals	569	30	2
Breast cancer	Relapse of breast cancer	Institute of Oncology University Medical Center Ljubljana	286	9	2
Schizophrenia	Discrimination between schizophrenic and control group based on EEG parameters	Department of Psychiatry Nicolaus Copernicus University	80	36	2

The classification methods applied, single and combined, are presented in Duda et al. (2001), Kotsiantis et al., 2006, Rokach (2009, 2010a, 2010b), Webb (2002). For linear discriminant classifiers (LDC) and quadratic discriminant classifiers (QDC) variables were selected by Wilks statistics (measuring variability between groups in relation to total variability). For the nonparametric kernel classifier, the choice was made according to minimization of the 1-Nearest Neighbor leave-one-out error, because 1-Nearest Neighbor is not complex, so is a quick classifier. Performance of various discriminant methods was assessed using apparent (Resubstitution), cross-validation (CV) and leaving-one-out method errors (LOOUT). Exploring performance of classifiers, single and combined, is based on "learning curves", where apparent, CV or LOOUT errors are plotted versus the number of resampling loops. Figures were obtained by use of independent programs based on the PRTOOLS package for Matlab (Duin et al., 2007).

The classical methodological technique (supplying parametric discrimination functions) assumes a jointly normal distribution of the predictive variables for optimality. However, in many problems, this assumption (or assumption of equal covariance matrices in differentiated groups for quadratic discrimination) can be doubtful. Various procedures have been elaborated as alternatives to classical discriminant analysis. One of them is the Parzen classifier, based on kernel estimation of density in discriminated populations. This discriminant Bayesian procedure, in opposite to Bayesian linear and quadratics discriminant functions is nonparametric, i.e. no assumption on distribution in discriminated populations is made. Another discriminant

procedure, coming from quite different methodology, which does not assume anything of distributions, so is also nonparametric, is the creation of trees (Rokach et al., 2005; Quinlan, 1987).

Currently, classifier researchers tend to combine procedures, based on similar type or different base classifier (Kotsiantis et al., 2006; Rokach, 2009, 2010a, 2010b). Especially big attention is focused on families of classifiers coming from two ideas: bootstrap aggregations and boosting (Breiman, 1996, 1998, Freund et al., 1997). Because these combining procedures are time consuming, the constituent classifiers with great complexity may cause computational problems. Combining simple and not optimal classifiers may to some extent bypass the drawbacks of such classifiers. Additionally, relaxed assumptions connected with resampling of the whole training data set may reduce deficiencies of base classifiers built on training sets.

Datasets

Three data sets, characterized by different difficulties, were examined (Table 1). The material, used in the discriminant analysis, comes from a few medical centers. The patients were divided into two groups. Classification was performed by clinicians. The discriminant problems included in the data sets are described in Table 1. Besides group classification, each patient was described, using clinical variables of a number not bigger than both group sizes. Two first sets from Table 1 come from the UCI Machine Learning Repository (Bache et al., 2013). Those sets are of various difficulties in making decisions. The biggest Wisconsin Diagnostic Breast Cancer data set(WDBC) is relatively simple (212 malignant patients from 569), the Breast Cancer data set is more difficult (85 relapses among 286 patients). Schizophrenic data sets consist of 50 schizophrenic patients among 80.

The data sets presented in Table 1 do not have strict multidimensional normal distribution in discriminated groups. However, the Breast Cancer data set shows the largest deviation from normality. The Schizophrenia data set has 36 variables that are linear combinations of disjoint subsets of 96 primary variables coming from computerized EEG equipment. Before summation, those elementary variables were standardized and transformed by logarithm to approach to one-dimensional gaussianity. The process of summing variables was done to reduce the large number of EEG parameters according to the medical knowledge, i.e. the summation of EEG parameters was done with six brain regions, three on the left and three on the right side of the brain.

Classifiers' Visualization for WDBC Data Set

At the beginning of the analysis let's look at the behavior of tree errors during the boosting procedure, because trees are the most common constituent classifiers applied in this ensemble. On the consecutive figures, there are overlaid learning curves of apparent (resubstitution), cross-validation and leave-one-out errors. Each elaborated learning curve represents the dependence of a chosen kind of errors assessment on the number of resampling loops. The figure with overlaid learning curves contains errors of individual classification models for increasing the number of loops (i.e. constituent classifiers number), represented by horizontal axis ($L = 1, \ldots, 100$: "Number of loops in resampling") with the estimated averages of apparent errors for first x loops ("Err. in conseq.loops" and "Mean err. of const.classif", respectively). Fusion errors after merging results of L loops ($L = 1, \ldots, 100$): apparent, cross validated with ten folds and leave-one-out – are also drawn ("Ensemble err", "Ensemble CV10err","Ensemble LOOUTerr", respectively).

Two following plots (Figures 1 and 6) represent overlaid learning curves connected with the diagnosis of breast cancer based on 30 discriminating variables in the WDBC data set (Table 1).

Because the performance of elaborated learning curves for a number of loops extending $L = 100$ was not meaningfully changed, the number of loops on the horizontal axes with many overlaid lines was cut to 100 in order to make clear overlaying representation of single classifiers loops possible (if it is helpful, some results for bigger numbers of combined constituent classifiers to 200 may be quoted as numerical, not graphical results). In this way, on one plot we can observe the behavior of individual classifiers constructed on subsamples and the combined classifiers built on all numbers of loops from one till the current at the same time. Namely, on those learning curves, the gray lines without marks denote apparent classification errors made by constituent classifiers, a line with diamonds represents the mean of apparent constituent classification errors. A line marked by triangles represents ensemble apparent classification errors (bagging or boosting), a line with stars shows ensemble CV errors for the increasing number of combined classifiers and, similarly, a line with circles displays leave-one-out ensemble errors.

The instability of the base classifier is expressed in the diversity of errors after resampling of the data set (oscillating line). For the decision tree committee (Figure 1) we can observe useful diversity. The smallest apparent error is achieved for the first resampling loop. For boosting trees, a very

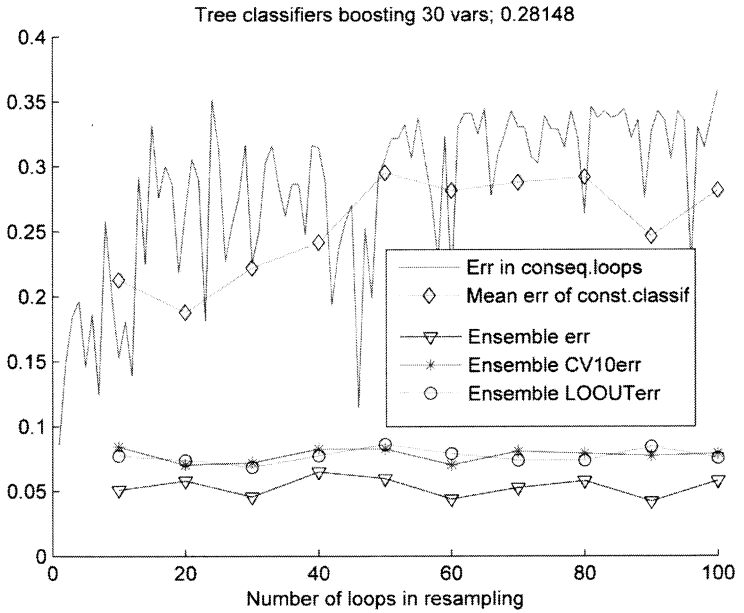


Figure 1. Dependence of classification errors on number of tree boosting loops, for WDBC set recognition based on 30 discriminating variables

high increase for the first 4 loops can be noticed (gray lines severely increasing from 0.09 to 0.2, Figure 1). For decision trees, a boosting aggregation procedure does not have the constant trend of increasing average apparent error over the whole range of the examined number of loops, e.g. significant reduction of mean constituent classifier errors is obtained for 80 resampling loops. The mean of single trees apparent errors have a general increasing tendency till 80 loops. For 50 loops, the maximum level of mean apparent individual errors, equal to 0.29, is obtained and for higher numbers of loops, the levels of individual apparent classifiers are not considerably increasing. The apparent error average of trees constructed on bootstrap subsamples with 100 loops is equal to 0.281 (Figure 1) and for a bigger number of loops, it still generally increases to 0.32 in 200 loops. According to CV and LOOUT assessment of generalization properties, for a relatively easily classified WDBC set, joining thirty loops of boosting trees is sufficient to minimize generalization errors. The smallest CV and LOOUT errors for whole training sets are equal to 0.07 and 0.04 (for 30 loops), respectively. For 100 (Figure 1) and 200 loops CV errors are 0.078 and 0.084, respectively. Similarly, concerning LOOUT errors: for 100 (Figure 1) and 200 loops, LOOUT errors for the whole training set are 0.075 and 0.08, respectively.

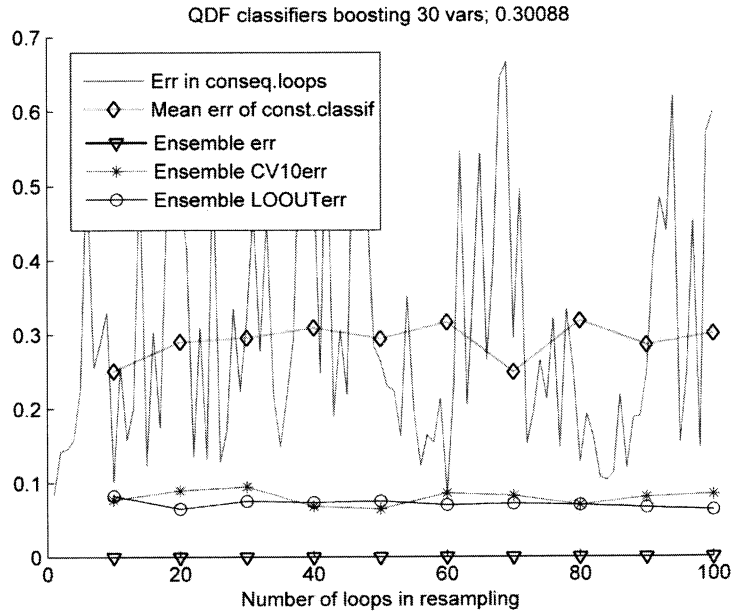


Figure 2. Dependence of classification errors on number of QDC boosting loops for WDBC set

Single classical trees are classifiers characterized by identification regions with boundaries consisting of linear parts parallel (perpendicular) to axes. Constituent decision trees depend strongly on the subset of the training set drawn by resampling, because of trees' instability. A single unstable tree classifier on the whole data set is characterized by a CV error equal to 0.074 and LOOUT error equal to 0.076. The resampling method improves the stability of a decision tree. The diversity of classifiers in consecutive loops is observable in error diversity.

Single quadratic classifier regions have boundaries that are multidimensional quadrics. The kind of quadric depends on the relationship between covariance matrices within discriminated groups. Each loop results in a different boundary. For boosting fusion of the quadratic discriminant function (QDC), the smallest apparent error is achieved for the fist resampling loop. A very high increase for the first 6 loops can be noticed (gray line severely increasing from 0.09 to 0.53, solid black line without additional marks in Figure 2). Looking at the lines with diamonds in Figure 2, we can observe that for QDC, the boosting aggregation procedure does not have the constant trend of increasing average apparent error over the whole range of the examined number of loops. In the QDC boosting combiner, the mean apparent errors of constituent classifiers have a general tendency to increase until

the first 60 loops (Figure 2), but a significant reduction of mean constituent classifier errors is obtained for 70 combined classifiers. The average apparent error follows a strict increasing trend until 40 loops. The mean of apparent errors of trees constructed on subsequent subsamples with 100 loops is equal to 0.281. The smallest LOOUT and CV errors are equal to 0.065, obtained by CV for 50 loops and by LOOUT on the 20^{th} loop. According to CV and LOOUT error assessment, the results of boosting trees are comparable with QDC (Figures 1 and 2).

Classifiers' Visualization for Breast Cancer Data Set

The analysis for the next set will begin by examining the constituent Parzen classifier in detail. The approximated value of optimal radius (r) is chosen with the usage of CV error. In boosting the Parzen classifier, the mean apparent constituent classifier does not have any tendency (Figure 3). Average reclassification errors substantially vary across the whole range of examined loops till 100. In contrast, for linear discriminant function (LDC) on the same Breast Cancer Data Set (Figure 4) there is an observable increasing trend, which confirms the intuition that consecutive loops of boosting procedure work on samples more difficult for classification (boosting is focused on objects problematic for identification). For strong and flexible Parzen classifiers there are fewer difficult patterns to identify than for other Bayesian classifiers (smaller CV and LOOUT errors of the constituent Parzen kernel classifier based on the whole data set, equal to 0.25 and 0.26, respectively). Modeling on a selected set of variables may support medical decisions for those problems. In the context of Parzen boosting results for the Breast Cancer data set, it should be noticed that the Parzen classifier has the beneficial property of tuning the parameter of smoothing (r) on the basis of CV error, so it has better discriminant properties than other single examined Bayesian methods and obtains the smallest generalization error of them – CV 10 equal to 0.21, LOOUT error level of 0.22.

Quite a different situation concerning Breast Cancer data was observed for LDC boosting classifier (Figure 4). For LDC boosting, the mean of apparent single classifier errors, which is interpretable by lines with diamonds, has a tendency to increase till 90 loops and after that to not grow until the number of 200 loops is reached. The single constituent LDC discriminant function obtains classification errors close to errors of the ensemble – about 0.25 – after resampling of the dataset. Few classifiers reach the level

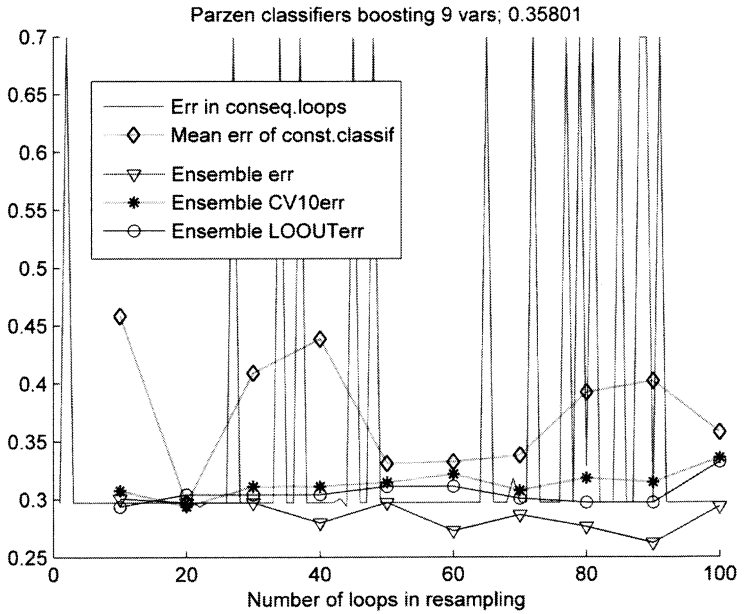


Figure 3. Dependence of classification errors on the number of Parzen boosting loops for breast cancer recognition based on 9 variables

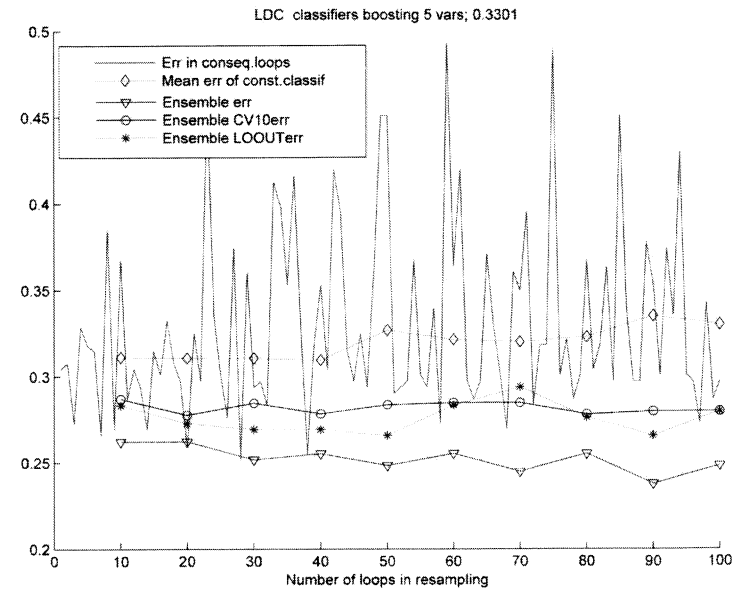


Figure 4. Dependence of classification errors on number of LDC boosting loops for breast cancer recognition based on the 5 best discriminating variables

of error of almost 0.5. The mean of reclassification errors after 100 loops is equal to 0.33, while fusion constructed by 100 loops of bagging obtains a generalization error level of 0.28.

Single linear classifier regions have boundaries that are multidimensional hyperplanes, which depend on the relationship between covariance matrices and centroids of populations. Each loop results in a different boundary. The diversity of discriminant functions in consecutive loops is visible as variability in error levels.

Classifiers' Visualization for Schizophrenia Data Set

To visualize more exactly classification during resampling methods, the schizophrenia data set was reduced to two dimensions, according to the optimization of variables selection criterion. In the data connected with the problem of recognizing schizophrenia, let's analyze the visualization of another resampling method – bootstrap aggregation (bagging).

The QDC classifier base error is 0.24, when estimated by CV, which means, after comparison with CV fusion errors not exceeding 0.14 after fourty loops, that the fusion bagging committee in this classification problem is certainly useful (Figure 5). Assessment of CV errors for bagging QDC shows the greatest decrease till 20 loops, where the smallest level error, 0.10, is reached. Increasing the number of loops above 30 is not useful. In contrast to boosting (e.g. Figure 2), the mean apparent errors of individual classifiers does not have any clear tendency. Because bagging is based on trials that are nondeterministic (as opposed to boosting, in which the draw in subsequent loops is associated with assigning higher weights for the observation poorly classified in the previous step), the average error base classifiers vary and no clear trends are established (line with diamonds on Figure 6). CV errors of the single LDC on the whole data set is equal to 0.216, thus it appears that CV error estimate (not bigger than 0.201) shows substantial reduction in CV error after the LDCs combining. Thus, according to CV error, LDC bagging gives improvement of the generalization properties, though CV errors are very diverse along increasing numbers of loops (Figure 6). For numbers of loops between 100 and 200, the CV is about 0.22, so adding more constituent classifiers is not beneficial (not included in the figure).

Boosting LDC for the schizophrenia dataset with 2 selected variables is not beneficial in comparison to bagging, because CV and LOOUT errors for all numbers of loops till 200 exceed 0.2 (not presented in graphical way).

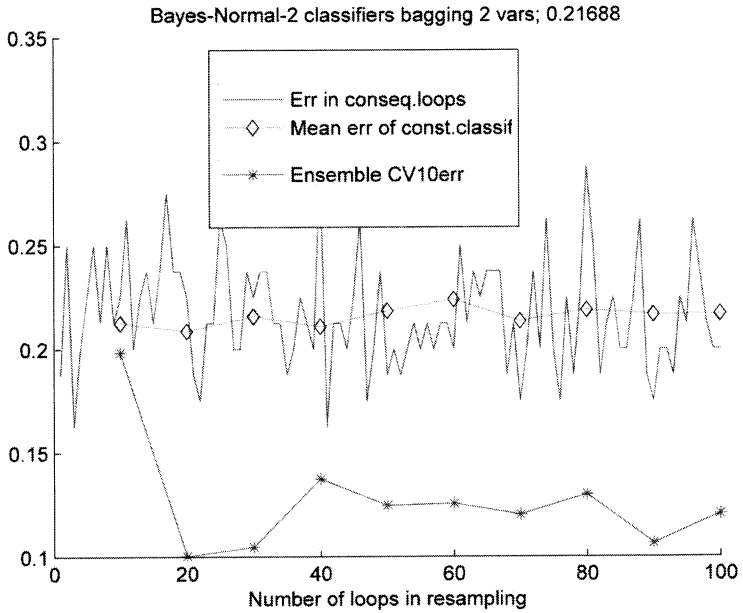


Figure 5. Dependence of classification errors on number of QDC bagging loops for schizophrenia recognition based on 2 best discriminating variables

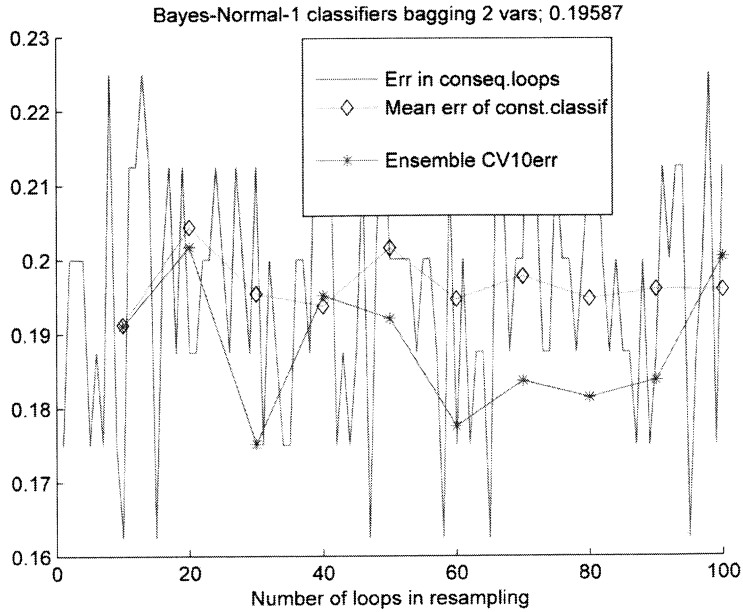


Figure 6. Dependence of classification errors on number of LDC bagging loops for schizophrenia recognition based on 2 best discriminating variables

Bagging is known as appropriate for smaller data sets, while boosting is elaborated for rather bigger data sets (Skurichina, 2001). The Schizophrenia data set is the smallest of the examined sets, though, when this set is considered with the number of chosen two best discriminating variables, it is not very small in comparison to dimensionality.

Discussion

According to all learning curves, apparent constituent error averages are substantially higher and considerably more diverse than ensemble errors. Assessment of misclassification rates for ensemble methods may also be regarded in the context of the diversity of a set of constituent classifiers. From learning curves we can observe that the average apparent error of individual loops is much higher than the error of the weighted voting for boosting loops (Figures 1–4). This corresponds to a linear formula for combining individual models, where the committee error is the sum of individual model error averages and of components related to the measure of the heterogeneity of a set of models. For example, error-ambiguity decomposition was proposed by Krogh et al. (1995) for regression tasks. Krogh et al. (1995) proved that for a single observation (x), the squared error of the combined estimator, obtained by weighing the base linear classifiers results, is expressed as the difference of weighted average base classifiers' squared error and component specifying ambiguity (indicating the diversity of base classifiers). If, instead of considering the arithmetic mean, the geometric mean is chosen as the fusion procedure, then as s a measure of the accuracy of the combined classifier, the mean squared error can no longer be applied, but the Kulback-Leibler (D_{KL}) directed divergence for two distributions can. On the base of Heskes (1998) work, Brown et al. (2010) gave the formula for D_{KL}:

$$D_{KL} = (y \| \bar{f}) = \frac{1}{L} \sum_{l=1}^{L} D_{KL}(y \| f_l) - \frac{1}{L} \sum_{l=1}^{L} (\bar{f} \| f_l)$$

where f_l is the l-th discriminant function ($l = 1, \ldots, L$) and y is the learning vector.

Thus the results coming from figures representing the learning curves that the ensemble errors benefit mean apparent errors can be explained by attached theoretical reasons.

From attached theoretical background we can see that in order to achieve a small error classifier fusion, a compromise is needed between diversity and the average error of the base classifiers. However, those decom-

position models take into account the estimated mean square error of the linear combination function or Kullback-Leibler divergence, while in bagging and boosting methods we are dealing with a merger constructed by a vote, weighted or unweighted. Brown et al. (2010) presented an analogous decomposition of errors in the case of classifiers combined by voting. Decomposition for the majority vote error and *"0–1"* loss function for L base classifiers with labels $\{-1, 1\}$ is the following:

$$e_{MV} = \int e_{AvgInd} - \int_{x+} \frac{1}{L} \sum_{l=1}^{L} d_l(x) + \int_{x-} \frac{1}{L} \sum_{l=1}^{L} d_l(x)$$

where

- e_{MV} is Majority Vote (MV) error
- e_{AvgInd} is individual classifiers error average
- d_l – binary variable denoting the mismatch of l-th base classifier ($l = 1, \ldots, L$) with the vote, subspaces of the data set:
- x_+ where the vote fusion is correct
- x_- where the vote fusion incorrect.

The last decomposition indicates that the mean of individual errors and the diversity both have an impact on the classifier committee error. The second component is beneficial diversity and the third component is unbeneficial diversity. By considering the third component, we can explain the phenomenon that combining diverse but inaccurate classifiers is not beneficial and that the compromise between diversity and accuracy of component classifiers is needed.

The success of boosting and bagging is connected with the diversity of the ensemble. Different sets in resampling cause different classifiers (with different classification regions) and they have different performance. Diversity of classifiers reflects diversity of classification errors. However, they are not the same, because constituent classifiers with the same errors may be different, e.g. may have quite dissimilar boundaries, and therefore different sensitivity and specificity. The component classifiers' diversity, visualized in the presented learning curves, reflects the concordance of the constituent classifier with the learning vector. Another, though to some extent analogous, measure of diversity used for the active enforcement of base classifiers differentiation, to build the accurate ensembles, was applied by Melville et al. (2005); this is the average binary incompatibility of individual classifiers' results with the aggregated classifier.

The tendency of increase of the average resubstitution errors was found for boosting methods. For bagging, such a clear trend cannot be seen. In particular, the fastest increase in apparent errors, compared with fusion meth-

ods of Bayesian classification, was found in the trees. They are known as unstable classifiers. Although in applications of bagging and boosting methods usually the number of loops used is at least 100, for the data sets which do not have the problem of small size relative to the size, exceeding the number of a few tens of loops is not necessary. It may be an essential observation in the context of the complexity of the ensemble procedure. Contrary to the opinions of some authors, linear discriminant functions may also be useful in resampling combining. This is concordant with the results of Skurichina (2001). Additionally, for quadratic discrimination, combining by bagging or boosting committee may also be beneficial. Constituent learners, which do not hold an increasing trend for resubstitution errors in subsequent boosting procedures loops, are not useful for the ensemble. This fact may mean that they little correct errors for patterns close to the classification boundaries. An example of such a classifier is the Parzen classifier. The kernel classifier is known as strong and flexible discriminant procedure. Thus, boosting is beneficial for nonparametric decision tree classifiers, but may not be useful for nonparametric Bayesian classifiers.

Some diversity measures based on oracle outputs, examined by Kuncheva et al. (2001, 2003), use only the information about concordance of pairs for constituent classifier decisions. Additional methods of diversity valuation may be the changes in constituent classifier error assessment, presented in the current work, connected with elaborated learning curves. We can observe the relationships between such differentiation and classification errors. Further development may be the assessment of the correlations between the new measure of diversity suggested by the current examination, variance or standard errors of consecutive constituent classification errors, and ensemble errors. It would also be interesting to study how this diversity correlates with the stability of constituent and combined classifiers.

Conclusions

The usefulness of bagging and boosting methods comes from diversity. Discriminant functions, which do not have an increasing trend for base resubstitution errors in subsequent boosting deterministic procedure loops, are not advantageous for the fusion, like the kernel Parzen discriminant function is. For the resampling success of classifier fusion, a compromise between accuracy and diversity is necessary. Diversity important in the success of boosting and bagging may be evaluated by concordance of component classifier with the learning vector.

Acknowledgments

The author is grateful to Prof. Wiktor Dróżdż from Department of Psychiatry at Nicolaus Copernicus University for the schizophrenic patients data set.

REFERENCES

Bache, K., & Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Retrieved from http://archive.ics.uci.edu/ml.

Banfield, R. E., Hall, L. O., Bowyer, K. W., & Kegelmeyer, W. P. (2005). Ensemble diversity measures and their application to thinning. *Information Fusion*, 6(1), 49–62.

Bi, Y. (2011). Analyzing the Relationship between Diversity and Evidential Fusion Accuracy. In C. Sansone, J. Kittler, F. Roli (Eds.), *Multiple Classifier Systems*, LNCS 6713, 249–258.

Breiman, L. (1996). Bagging predictions. *Machine Learning*, 24(2), 123–140.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, 26(3), 801–849.

Brown, G., & Kuncheva, L. I. (2010). Good and Bad Diversity in majority vote ensembles. In N. El Gayar, J. Kittler, & F. Roli (Eds.), *Multiple Classifiers Systems*, LNCS 5997, 124–133.

Dietterich, T. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, 40(2), 139–157.

Duin, R. P. W., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D., Tax, D. M. J., & Verzakov, S. (2007). *PRTools4.1, A Matlab Toolbox for Pattern Recognition.* Delft University of Technology.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1),119–139.

Heskes, T. (1998). Bias/variance decomposition for likelihood-based estimators. *Neural Computations*, 10(6), 1425–1433.

Kotsiantis, S. B, Zaharakis, I. D., & Pintelas, P. E. (2006). Supervised machine learning. *A review of classification and combining techniques*, 26(3), 159–190.

Krogh, A., & Vedelsby, J. (1995). Neural network ensembles, cross validation and active learning. *Advances in Neural Information Processing Systems*, 7, 231–238.

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, 51(2), 181–207.

Kuncheva, L. I., Skurichina, M., & Duin, R. P. W. (2002). An Experimental Study on Diversity for Bagging and Boosting with Linear Classifiers. *Information Fusion*, 3(4), 245–258.

Kuncheva, L. I., & Whitaker, C. J. (2001). Ten Measures of Diversity in Classifier Ensembles: Limits for Two Classifiers. Proceeding IEEE Workshop on Intelligent Sensor Processing, 14 February 2001.

Kuncheva, L. I., Whitaker, C. J., Ship, C. A., & Duin, R. P. W. (2000). Is independence good for combining classifiers? In International Conference on Pattern Recognition (ICPR'00), 3–8 September 2000 (Volume 2, 168–171). Barcelona, Spain.

Kuncheva, L. I. (2003). That elusive diversity in classifier ensembles. In F. J. Perales, A. J. C. Campilho, N. P. de la Blanca, & A. Sanfeliu (Eds.), *Pattern Recognition and Image Analysis*, LNCS 2652, 1126–1138.

Melville, P., & Mooney, R. J. (2005). Creating diversity in ensembles using artificial data. Diversity in Multiple Classifier Systems. *Information Fusion*, 6(1), 99–111.

Quinlan, J. R. (1987). Simplifying Decision Trees. *Int. J. Man – Machine Studies*, 27(3), 221–234.

Rokach, L. (2010a). *Pattern Classification Using Ensemble Methods*. Series in Machine Perception and Artificial Intelligence, World Scientific Publishing.

Rokach, L. (2010b). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1–2), 1–39.

Rokach, L. (2009). Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics and Data Analysis*, 53(12), 4046–4072.

Rokach, L., & Maimon, O. (2005). Top-down induction of decision trees classifiers – a survey. *IEEE Transactions on Systems, Man and Cybernetics*, Part C, 35(4), 476–487.

Shipp, C. A., & Kuncheva, L. I. (2002). Relationships between combination methods and measures of diversity in combining classifiers. *Information Fusion*, 3(2), 135–148.

Skurichina, M. (2001). *Stabilizing weak classifiers* (Doctoral dissertation). Delft University of Technology.

Skurichina, M., & Duin, R. P. W. (2002). Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis and Applications*, 5(2), 121–135.

Vu, T. T., Braga-Neto, U., & Dougherty, E. R. (2009). Bagging degrades the performance of linear discriminant classifiers. In IEEE International Workshop on Genomic Signal Processing and Statistics, GENSiPS, 17–21 May 2009. Minneapolis, MN, USA.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0040

The Stability of Gene Selection in Microarray Experiments

Magdalena Wietlicka-Piszcz[1]

[1] Department of Theoretical Backgrounds of Biomedical Science and Medical Informatics, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University, Poland

Abstract. This paper addresses the issue of the stability of lists of genes identified as differentially expressed in microarray experiments. The similarities between gene rankings yielded by various gene selection methods performed with resampled datasets were assessed. The mean percentage of overlapping genes for two rankings varied from 10 to 90% depending on the applied gene selection method and the size of the list. The assessment of the stability of obtained gene rankings seems to be relevant in the analysis of microarray data.

Introduction

Microarrays are a new technology applied in the genetics field, enabling simultaneous investigation of expression levels of thousands or tens of thousands of genes. In many cases, the main aim of a microarray experiment is the identification of genes involved in the aetiology of a particular disease or the selection of genes enabling the differentiation between disease subtypes as well as the prediction of future events, such as response to applied therapy, survival times, relapse of a disease or cancer recurrence. Thus, the proper identification of genes differentially expressed, e.g. between the diseased and normal tissue, is crucial because it often determines, to a certain extent, the direction of further research, which is frequently focused on selected genes.

In the vast majority of microarray experiments, the selection procedures of differentially expressed genes restrict the number of investigated genes from the total of tens of thousands to tens and "the results of microarray studies are usually the starting point for further more expensive and time-consuming experiments, which involve only a small number of candidate genes" (Aerts et al., 2006).

Typically, in a microarray experiment the number of arrays is in the range of a few microarrays to over 200 or 300, while the number of exam-

ISBN 978-83-7431-392-6 ISSN 0860-150X

ined genes is in the range between a few thousand and tens of thousands. Therefore, the selection of differentially expressed genes is a very important stage of microarray data analysis and involves the use of methods that can be used when the number of features (genes) is much bigger than the number of samples (microarrays). There have been many methods developed for the selection of active genes in microarray settings. However, the lists of genes identified as differentially expressed produced by those methods may differ substantially.

Another problem associated with the gene selection procedure is the stability of gene lists obtained with a particular method but with slightly modified versions of the dataset (Boulesteix et al., 2009), e.g. subsampled dataset. The gene rankings obtained with a method performed, e.g. with bootstrap samples of the original dataset, may significantly differ from the ranking returned by the same method applied to the whole dataset.

In the vast majority of cases, the standard procedure of microarray data elaboration involves the use of one active gene selection method applied to the whole data set and further analysis and reasoning is partly based on that list.

This work addresses the issue of gene selection stability and its dependence on the applied method of active gene identification. The similarities between gene rankings yielded by various methods and between rankings obtained for the perturbed dataset were considered. The analysis was performed for three datasets. Four methods of gene selection were applied and compared.

Material and Methods

The analysis was performed based on three high density oligonucleotide microarray datasets downloaded from the public repositories. The first dataset contained data from 167 samples of oral squamous cell carcinoma (OSCC) and 17 samples of oral dysplasia tissues (Chen et al., 2008; public repository GEO – GSE30784). The data were used to identify genes enabling differentiation between OSCC and dysplasia. In further considerations this dataset will be called *dataset 1.*

Another dataset concerned the problem of early detection of colorectal cancer (CRC) and consisted of expression data from 100 whole blood samples from patients with CRC and from 100 samples from patients without any symptoms of CRC, inflammatory bowel diseases or polyps (Xu et al., 2013; public repository ArrayExpress – E-MTAB-1532). The

gene expression profiles from this dataset were used to pick up genes useful in the early recognition of CRC. This dataset will be referred to as *dataset 2* from this point forward.

The third dataset comprised expression profiles from a total of 58 cirrhotic tissue samples from liver tissue with HCV infection. Seventeen of them were from patients with hepatocellular carcinoma and 41 were from patients without hepatocellular carcinoma (Mas et al., 2009; public repository GEO – GSE1423). The analysis of this dataset aimed to identify genes enabling the differentiation between cirrhotic tissue and cirrhotic tissue with concomitant hepatocellular carcinoma. This dataset will be referred to as *dataset 3*.

The data were previously pre-processed, so for each probe expression summaries were available. The gene selection was performed by the use of two methods based on the t-test, i.e. Parametric Empirical Bayes Method (Limma) (Smyth, 2004) and Significance Analysis of Microarrays (SAM) (Tusher et al., 2001), and the Nonparametric Empirical Bayes Method based on Wilcoxon rank sums (EbamWilcoxon) (Efron et al., 2002) and the Wilcoxon rank sum test (Wilcox) were also applied.

The above mentioned methods of variable selection return ordered lists of candidate genes, where the genes are ordered according to the criterion used to rank variables, i.e. the absolute value of a particular test statistic. The highest ranked genes are considered for further analysis. The top k-list of candidate genes is the list of k genes with the highest rank values, so there are the genes from the top of the ordered list of candidate genes.

For each of the considered datasets in this work, the identification of differentially expressed genes was performed by the use of the four various methods of feature selection. Additionally, to address the issue of the stability of obtained gene lists, the feature selection methods were applied with the re-sampled datasets. The Jackknife subsampling technique was applied. The dataset was split into 10 disjoint folds of approximately equal size and the samples were created by removing the consecutive folds from the whole dataset. This procedure was repeated 10 times, so 100 samples were created, each comprising approximately 90% of the whole dataset.

To assess the similarity of two gene rankings, the proportion of common genes in the two top k-lists was calculated. This measure is also denoted as a percentage of overlapping genes (POG) (Zhang et al., 2009). To visualize the similarity of two gene rankings (the two top k-lists) versus the size k of the list, a descriptive plot called correspondence at the top (CAT-plot) was applied (Irizarry et al., 2005). CAT curves show the proportion of common genes plotted against the size of the lists.

To assess the stability of a gene ranking obtained with a particular gene selection method, the comparison of the ranking obtained for the original dataset and rankings derived for subsamples was performed. The POG for pairs of rankings (the ranking for the original dataset and for the subsample) was calculated and then the mean value of POG for all these pairs was also determined.

To consider the similarities between rankings derived with various gene selection methods, the POG was calculated for pairs of rankings obtained with different methods. Then, for each pair of methods, the mean value of POG for subsamples was computed.

Additionally, an attempt to aggregate rankings based on subsamples was made and then the comparison of classification, based on the single ranking and the ranking derived from aggregation, was carried out.

To perform the selection of differentially expressed genes and classification based on identified gene sets, the whole dataset was split into 10 disjoint folds of approximately equal size. Each fold was used as a testing set while the remaining arrays were used as a training set. This procedure was repeated 10 times (ten ten-fold cross-validation). For each training set (TS), the selection of genes was performed in two ways. At first, active gene identification was performed for the whole TS, which resulted in a single ranking. Then, the selection procedure was applied to the re-sampled TS. The bootstrap samples were created by drawing samples (arrays) with replacements from the TS. For each bootstrap sample, the selection of genes was performed. The re-sampling was repeated 100 times, so for each TS 100 gene rankings were obtained. Then the rankings were aggregated and for each gene the final score was estimated as a sum of ranks from all the lists. A single rank of a gene g in a list was estimated as $1/r$ (to assign bigger weight to genes on the top of the list) where r is the position of the gene on a single list, so the final score s of a gene g was estimated as $s_g = \sum_{i=1}^{n} 1/r$, where n denotes the number of lists. Therefore, for each TS two gene rankings were derived: the ranking from the single selection of genes from the TS (*ranking 1*) and an aggregated ranking from the resampling of the TS (*ranking 2*). Then, classification based on the two rankings was performed. The classification was carried out for the consecutive subsets of the first $2, 3, \ldots, 100$ highest ranked genes from both rankings.

The procedures of gene selection and classification were repeated for all pairs of training and testing sets. For classification, the classifiers widely used in microarray data analysis (Boulesteix et al., 2008; Van Sanden et al., 2008), such as the Support Vector Machines with linear kernel (SVMl), Support Vector Machines with radial kernel (SVMr), Diagonal

Linear Discriminant Analysis (DLDA) and Diagonal Quadratic Discriminant Analysis (DQDA) were used.

Results and Discussion

The considered gene selection methods were applied to identify the differentially expressed genes in the three analyzed datasets. Each method was applied both with the original dataset and with the re-sampled dataset.

To assess the stability of gene rankings produced by a particular gene selection method, the similarities between the rankings obtained with the original dataset and the rankings obtained with the re-sampled dataset were examined. The values of POG for the top k-list for the whole dataset and the top k-lists corresponding to the sub-samples were calculated and averaged.

The mean values of POG for the consecutive top k-lists, for k = 10, $20, \ldots, 300$ were calculated. Figures 1–3 present the CAT-curves for the top k-lists for the whole and sub-sampled dataset, for various gene selection methods and investigated datasets.

The highest number of common genes was observed for *dataset 2*. The POG for the top 100 genes was over 80%; however, the POG for *dataset 1*

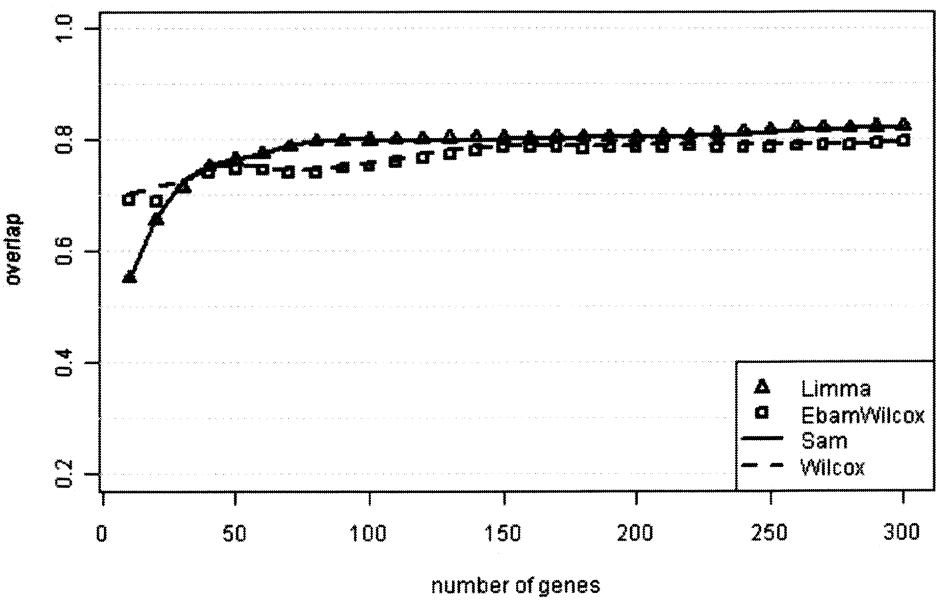


Figure 1. Mean percentage of overlapping genes for *dataset 1*

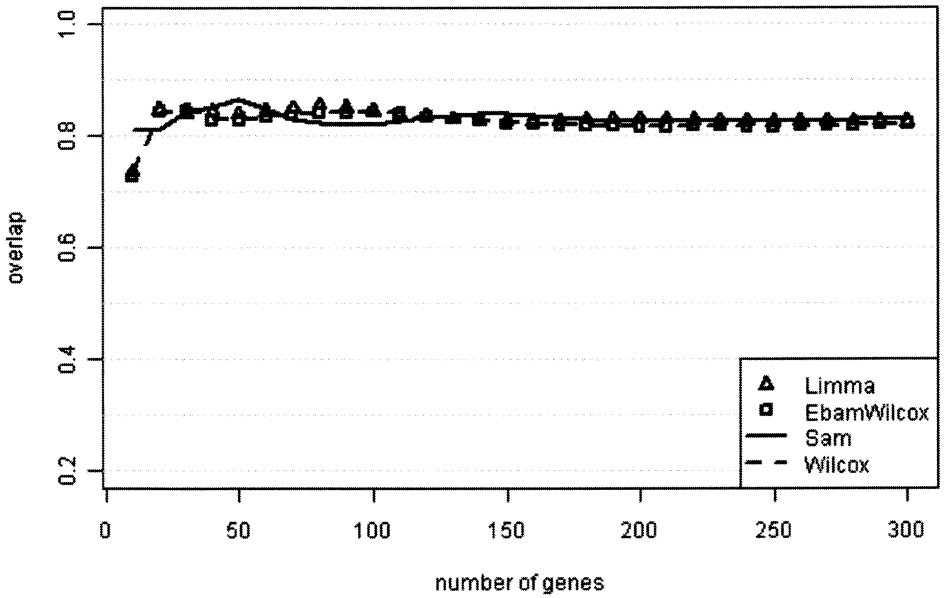


Figure 2. Mean percentage of overlapping genes for *dataset 2*

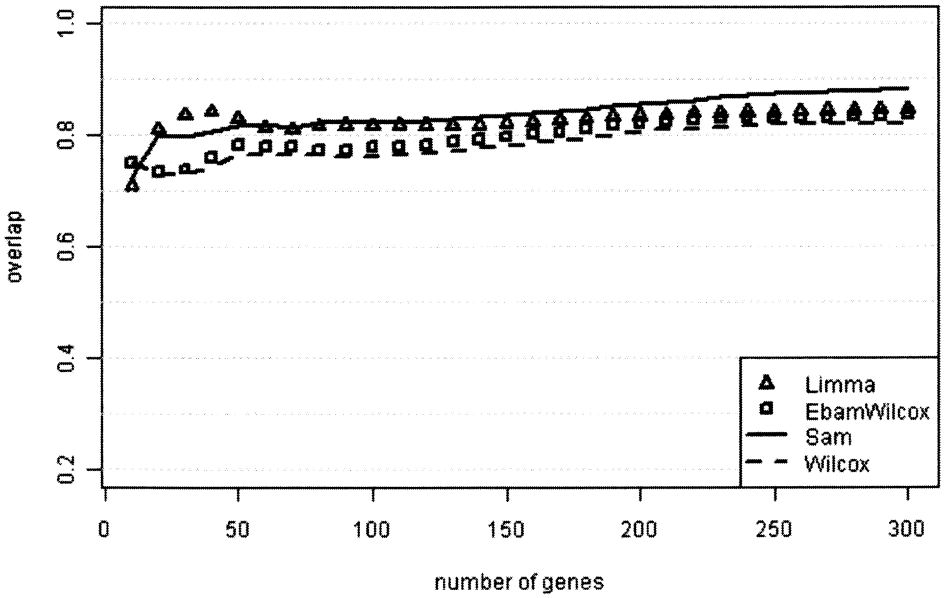


Figure 3. Mean percentage of overlapping genes for *dataset 3*

and *dataset 3* for the same number of genes was from 75 to 80%, depending on the method. The standard errors for the mean values of POG were in the range between 0.1 and 1%.

For *dataset 3*, the Significance Analysis for Microarrays and Parametric Empirical Bayes Method seemed to yield a higher overlap of selected genes than the Wilcoxon rank sum test or the Nonparametric Empirical Bayes Method, however the differences were small. For *dataset 1* and *dataset 2*, the mean overlap of gene rankings was comparable for all considered methods.

Additionally, to illustrate the variability between the top k-lists, the proportion of genes common for all the samples using each method was calculated. Figures 4–6 show the percentages of common genes for the top k-lists obtained for all subsamples. The percentage of overlapping genes varied from 20–25% for the top 100-list for *dataset 1* and 20–40% for *dataset 3* to 35–40% for *dataset 2*. This means that for 100 lists of the top 100 genes, for the analyzed datasets, 20–40 genes, depending on the dataset and the method, might be encountered on each list. For the top ranked list containing 300 genes, the overlap was also between 20 and 40%, so 60 to 120 genes were common for all the rankings.

To assess the similarities of gene rankings yielded by the considered gene selection methods, for each subsample, the POG for pairs of rankings produced by different methods were calculated. Then for each pair of methods the mean values of POG for all subsamples were computed. The mean values of POG were calculated for the consecutive top k-lists, $(k = 10, 20, \ldots, 300)$. To visualize the similarities of the gene lists yielded by various feature selection methods, the CAT-plots were created. Figures 7–9 present the CAT-curves for pairs of methods. The highest percentages

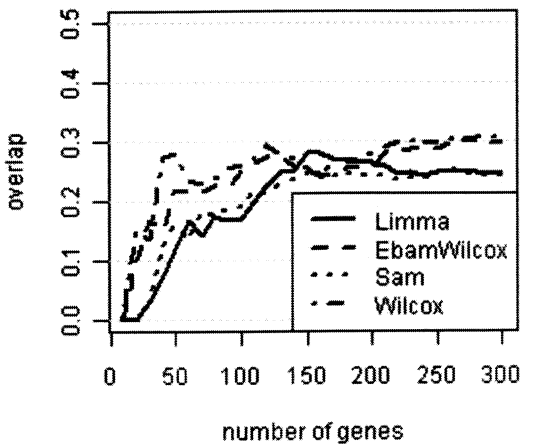


Figure 4. Percentage of overlapping genes for 100 sub-samples of *dataset 1*

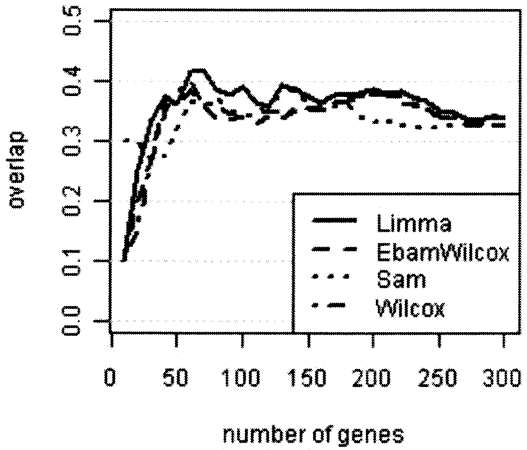


Figure 5. Percentage of overlapping genes for 100 sub-samples of *dataset 2*

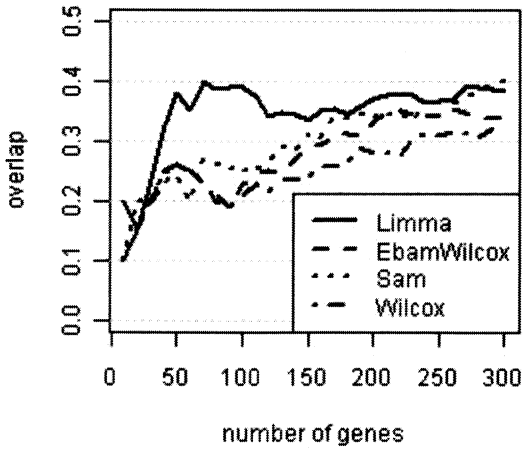


Figure 6. Percentage of overlapping genes for 100 sub-samples of *dataset 3*

of overlapping genes were obtained by using the Wilcoxon rank sum test and the Nonparametric Empirical Bayes Method – based on this test, the percentage was over 90%. Also, the overlap of rankings produced by Parametric Empirical Bayes Method and Significance Analysis of Microarrays was about 85 and over 90% for *dataset 1* and *dataset 2*, respectively, for the list of 100 features, but for *dataset 3*, for the same list size, it was about 50%. For the other methods, the results obtained for different datasets varied depending on the dataset. The highest values of POG for all

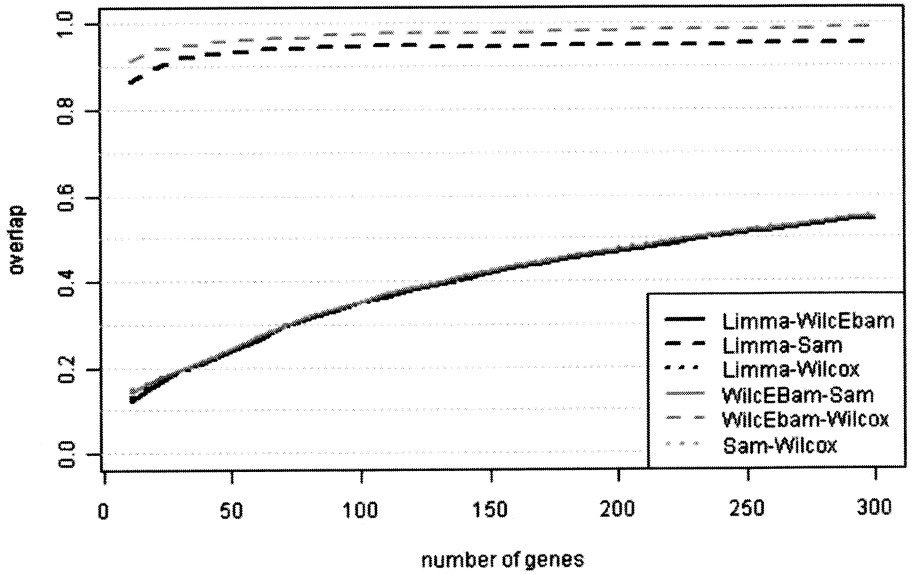


Figure 7. Mean percentage of overlapping genes for pairs of methods for *dataset 1*

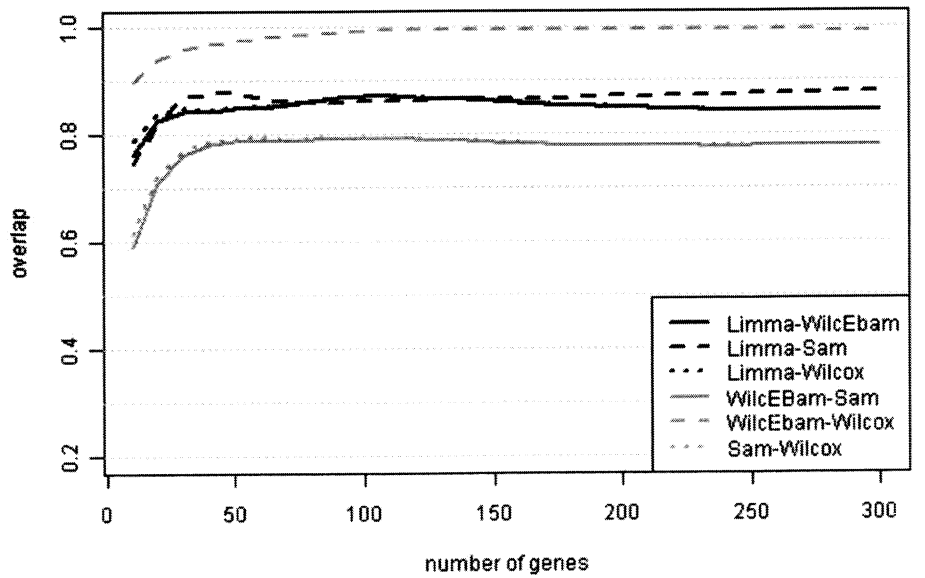


Figure 8. Mean percentage of overlapping genes for pairs of methods for *dataset 2*

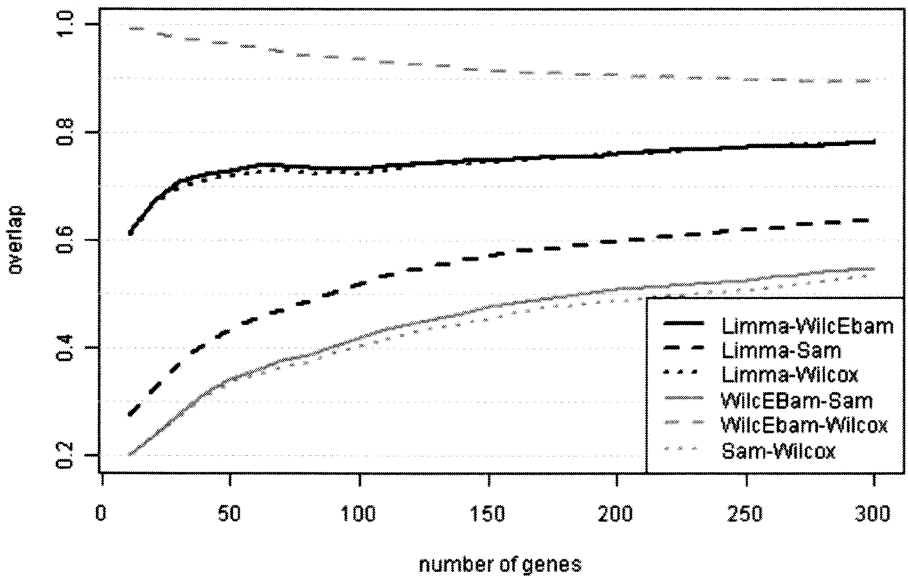


Figure 9. Mean percentage of overlapping genes for pairs of methods for *dataset 3*

methods were received for *dataset 2*, and for the list of 100 genes they varied from 80 to over 90%. This dataset comprised the biggest number of samples (200 arrays). The smallest overlap of gene rankings was obtained for *dataset 3*.

Figures 10–12 show the proportion of genes common for all subsamples for all pairs of considered gene selection methods. The highest overlap was obtained for *dataset 2* (about 25–35% for all pairs of methods for the list of 100 genes) and the lowest overlap was derived for *dataset 1*. For some methods, it was below 10% for the top 100-list.

There is a question regarding which genes should be used in further analyses. The answer is not straightforward and various strategies may be considered, depending on the main aim of the experiment. One of the possible solutions is the aggregation of ranks from multiple rankings.

The analysis of the three datasets shows that the smallest overlap of top k-lists was obtained for *dataset 1*. For this dataset, the comparison of classification based on rankings derived from the whole training set (*ranking 1*) and the rankings returned as a result of aggregation of lists obtained from re-sampling of the training set (*ranking 2*), was performed. The selection of genes was performed with the use of the Parametric Empirical Bayes Method.

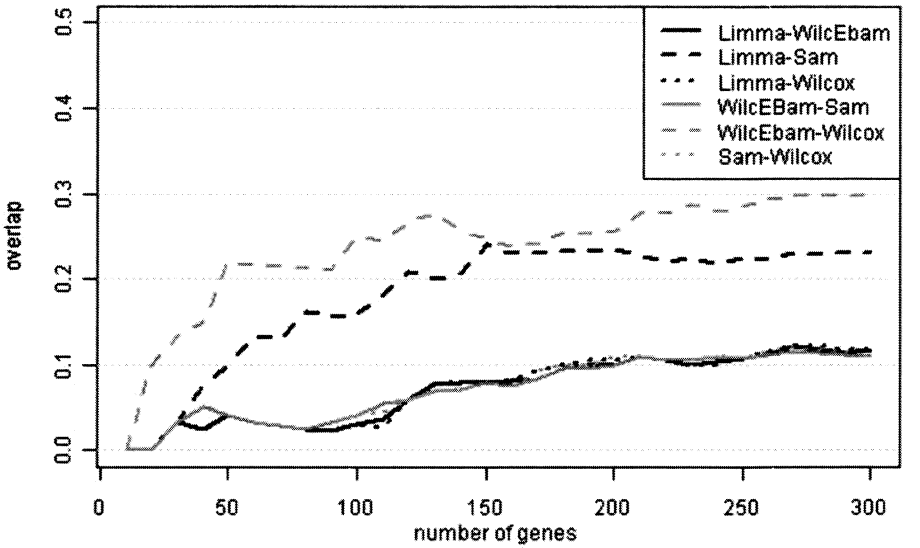


Figure 10. Percentage of overlapping genes for 100 sub-samples of *dataset 1*, for pairs of methods

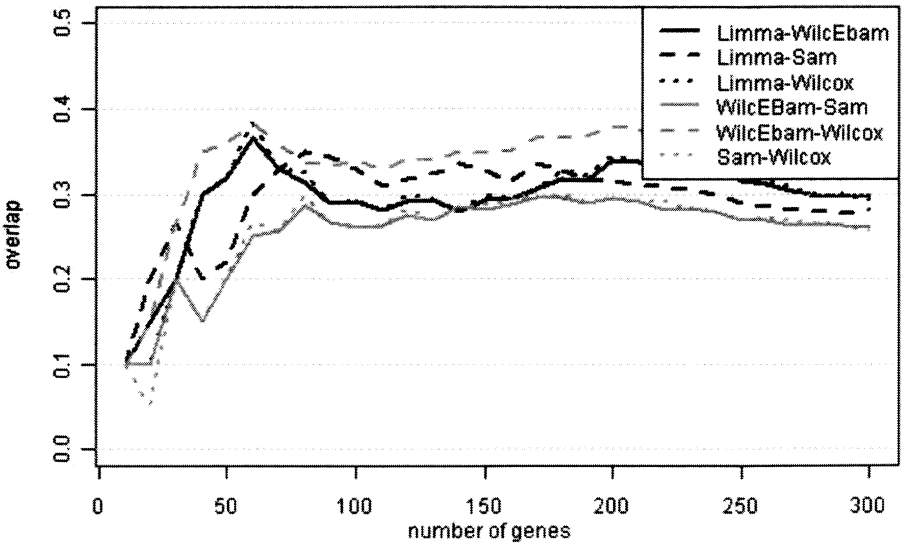


Figure 11. Percentage of overlapping genes for 100 sub-samples of *dataset 2*, for pairs of methods

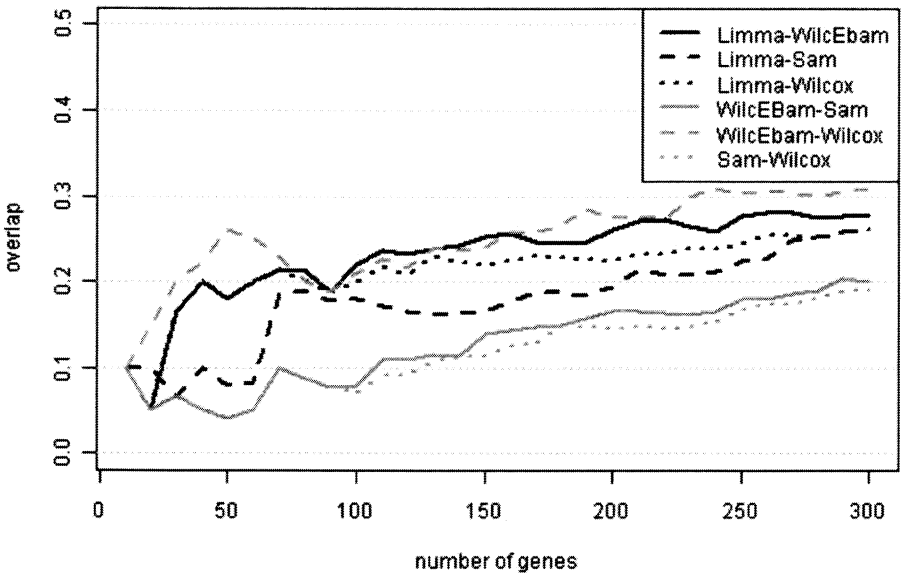


Figure 12. Percentage of overlapping genes for 100 sub-samples of *dataset 3*, for pairs of methods

Figure 13 presents the cross-validation misclassification error rates obtained for the classification based on the two types of rankings (*ranking 1* and *ranking 2*). The classification was carried out for the consecutive subsets of the first $2, 3, \ldots, 100$ genes from both rankings. DQDA, DLDA, SVMl and SVMr methods were used for classification. The lowest misclassification error rates were obtained for SVMr for *ranking 2* – for 33 genes the error was equal to 0.057, while for the same number of genes for *ranking 1*, the error was 0.071. The error rates obtained for the SVMl were comparable to those obtained for radial kernel. The lowest error was obtained for 83 genes and was equal to 0.058 for *ranking 2*, while for *ranking 1*, the error was 0.067. For DLDA, the error rates were in the range between 0.088 and 0.131 for *ranking 2* and, respectively, between 0.096 and 0.140 for *ranking 1*. The highest misclassification error rates were derived for DQDA. The errors varied from 0.116 to 0.151 for *ranking 2* and from 0.109 to 0.163 for *ranking 1*.

The misclassification errors were lower for rankings obtained from aggregation of many lists derived from resampling of the training set (*ranking 2*) for all applied methods and almost all considered subsets of genes.

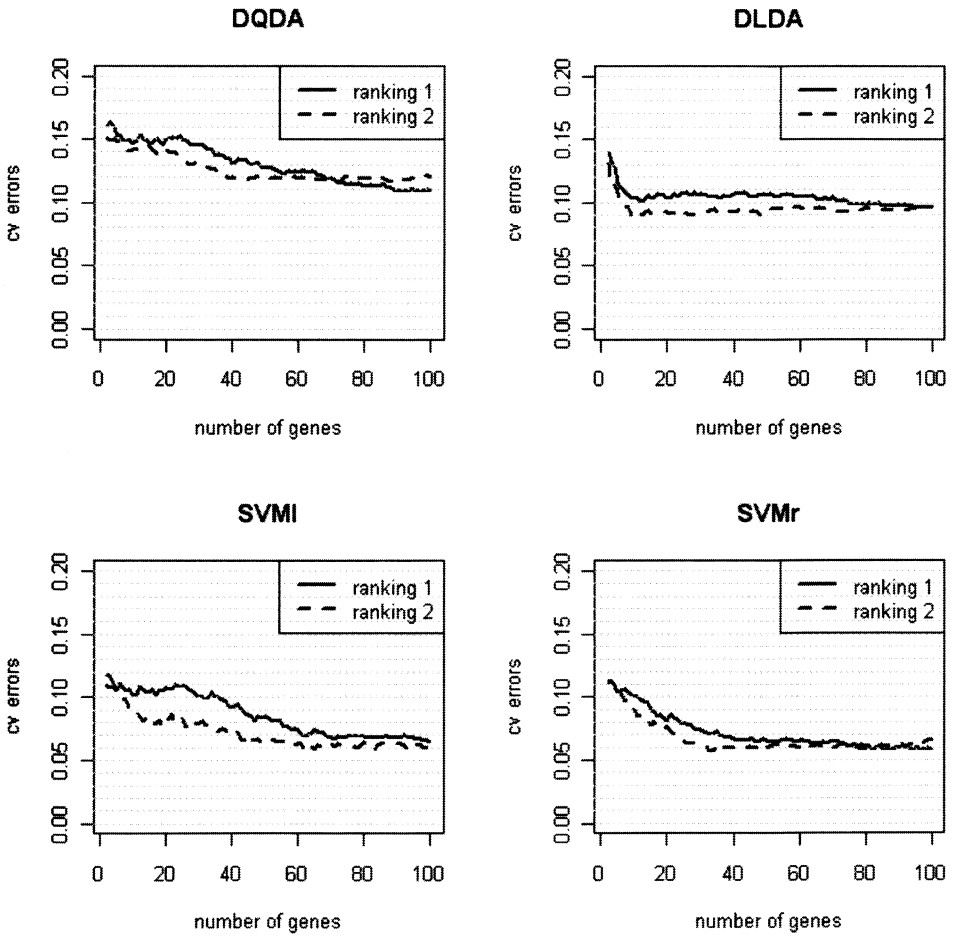


Figure 13. Comparison of cross-validation misclassification error rates for *dataset 1* for pairs of gene rankings: ranking based on the training set (ranking 1) and ranking derived from aggregation of rankings obtained from application of the gene selection method with subsamples of the training set (ranking 2)

Conclusion

The selection of features is a very important stage of elaboration of data from microarray experiments. The assessment of the stability of obtained gene rankings seems to be relevant and the careful analysis and comparison of gene lists obtained for perturbed datasets and/or various gene selection methods may help to get more reliable rankings. Certainly, further investigations in this area are necessary.

REFERENCES

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L. C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., & Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24, 537–544.

Boulesteix, A. L., & Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Brief Bioinformatics*, 10, 556–568.

Boulesteix, A. L., Strobl, C., Augustin, T., & Daumer, M. (2008). Evaluating Microarray-based Classifiers: An Overview. *Cancer Informatics*, 6, 77–97.

Chen, C., Mendez, E., Houck, J., Fan, W., Lohavanichbutr, P., Doody, D., Yueh, B., Futran, N. D., Upton, M., Farwell, D. G., Schwartz, S. M., & Zhao, L. P. (2008). Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. *Cancer Epidemiol Biomarkers Prev*, 17(8), 2152–2162.

Efron, B., & Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol*, 23, 70–86.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martinez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q., & Yu, W. (2005). Multiple-laboratory comparison of microarray platforms. *Nat Methods*, 2, 345–350.

Mas, V. R., Maluf, D. G., Archer, K. J., Yanek, K., Kong, X., Kulik, L., Freise, C. E., Olthoff, K. M., Ghobrial, R. M., McIver, P., & Fisher, R. A. (2009). Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma. *Mol Med*, 15(3–4), 85–94.

Public Repository ArrayExpress. (2013). Transcription profiling by array of whole blood samples from colorectal cancer patients and healthy individuals, accession number E-MTAB-1532. Retrieved from http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1532/.

Public Repository Gene Expression Omnibus. (2009). RMA expression data for liver samples from subjects with HCV, HCV-HCC, or normal liver, accession number GSE1423. Retrieved from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14323.

Public Repository Gene Expression Omnibus. (2011). Gene expression profiling of oral squamous cell carcinoma (OSCC), accession number GSE30784. Retrieved from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30784.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1). DOI: 10.2202/1544-6115.1027.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9), 5116–5121.

Van Sanden, S., Lin, D., & Burzykowski, T. (2008). Performance of gene selection and classification methods in a microarray setting: A simulation study. *Communications in Statistics – Simulation and Computation*, 37(2), 409–424.

Xu, Y., Xu, Q., Yang, L., Ye, X., Liu, F., Wu, F., Ni, S., Tan, C., Cai, G., Meng, X., Cai, S., & Du, X. (2013). Identification and Validation of a Blood-Based 18-Gene Expression Signature in Colorectal Cancer. *Clin Cancer Res*, 19(11), 3039–3049.

Zhang, M., Zhang, L., Zou, J., Yao, Ch., Xiao, H., Liu, Q., Wang, J., Wang, D., Wang, Ch., & Guo, Z. (2009). Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, 25(13), 1662–1668.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0034

Classification Issue in the IVF ICSI/ET Data Analysis: Early Treatment Outcome Prognosis

Paweł Malinowski[1], Robert Milewski[1], Piotr Ziniewicz[1], Anna Justyna Milewska[1], Jan Czerniecki[2], Sławomir Wołczyński[3]

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

[2] Department of Biology and Pathology of Human Reproduction, Institute of Animal Reproduction and Food Research Polish Academy of Sciences, Olsztyn, Poland

[3] Department of Reproduction and Gynaecological Endocrinology, Medical University of Bialystok, Poland

Abstract. Infertility is a serious social problem. Very often the only treatment possibility are IVF methods. This study explores the possibility of outcome prediction in the early stages of treatment. The data, collected from the previous treatment cycles, were divided into four subsets, which corresponded to the selected stages of treatment. On each such subset, sophisticated data mining analysis was carried out, with appropriate imputations and classification procedures. The obtained results indicate that there is a possibility of predicting the final outcome at the beginning of treatment.

Introduction

Infertility is a problem that affects a growing number of couples that wish to have a child. Based on current statistics, approximately 18–20% of the couples in Poland suffer from infertility (Radwan, 2011). Currently, there are many known causes of infertility, including the crucial age of the woman (Milewski et al., 2008). In a significant proportion of cases, a direct cause of infertility cannot be determined, and the results of both women and men are in the norm, a so-called idiopathic infertility.

Depending on the identified causes, there are many treatments for infertility, but in many cases the only way to obtain offspring is by using In Vitro Fertilization (IVF) methods. In spite of constant improvement in Assisted Reproductive Technology (ART) that continues to enhance efficacy of the treatment, the pregnancy rate is still low and remains in the range of 40% (Milewski et al., 2013). Hence, there is a continuing need for an in-depth analysis of the data obtained in the treatment process, to find predictors for

ISBN 978-83-7431-392-6 ISSN 0860-150X

pregnancy and other factors that contribute to the next stages of treatment outcomes.

An extensive database was created in the Department of Reproduction and Gynecological Endocrinology at the Medical University of Bialystok, using dedicated software. The database covers the 6 year period from 2005 to 2010. Earlier studies analyzing the database focused mainly on a whole feature set. Those included classification alone (Milewski et al., 2012), selecting relevant features (Milewski et al., 2010) and using them to help classification (Milewski et al., 2011). This study focuses on classification by using subsets of features available at three selected phases of treatment. For comparison only, the same methodology was applied to the full set of features, available at the end of treatment. The purpose was to explore the possibility of successful classification with information available at the end of each selected phase of treatment. Like in the previous work (Milewskiet al., 2012), final results are shown for data not used in the learning phase, so that the results are as unbiased as possible.

Material and Methods

The analyzed database contains 1445 observations (single cycle of treatment) and 150 features (which include the outcome – pregnancy or no pregnancy), 22% of data is missing. Only about a third of the treatment cycles ended up inpregnancy. This high rate of missingness and relatively high outcome imbalance (2:1) can be linked to the nature of IVF ICSI/ET treatment.

Data analysis was carried out using the R environment (R version 3.0.1 (2013-05-16) "Good Sport"). There were used the packages presented in Table 1.

Table 1. Packages used

Package Name	Version	URL
e1071	1.6-1	http://CRAN.R-project.org/package=e1071
VIM	3.0.3.1	http://CRAN.R-project.org/package=VIM
randomForest	4.6-7	http://CRAN.R-project.org/package=randomForest
missForest	1.3	http://CRAN.R-project.org/package=missForest

On top of classification algorithm a cross-validation procedure was used. The whole dataset was divided randomly (on observations), at a 7:3 ratio,

to create a learning and validation part. This division was retained during further data split according to features. All algorithms were trained on the learning part only using k-fold cross-validation ($k = 10$). In almost all cases learning observations were partitioned into k subsamples. One of them was retained, while the rest were used for the training process for a specific algorithm and its set of parameters. The result of the training was then tested against the retained part, producing an error estimate. This process was repeated k times, and at the end, a single estimate of error was produced by averaging. This basic k-fold cross-validation procedure, implemented in package "e1071" (Meyer et al., 2012), was used for analysis.

For classification, a Random Forest (Breiman, 2001) algorithm was used, which proved to be one of the best in analyzing the database (Milewski et al., 2012). During the learning process, the algorithm builds a set of decision trees, based on available data. Each tree chooses, "gives a vote for", an observation class. Whole forest chooses the class with the majority of votes. Given N observations and M features, each tree is grown on samples of N observations selected with replacement (there are copies of observations). At each tree node some of the features are randomly selected (their number should be much smaller than M, square root by default for classification). The tree node splits data after the best possible split is found, but using previously selected features only as a criterion. By default each tree is built until the next split would result in an empty node (default for classification), but a minimum number of observations can be set in the terminal node.

The following parameters were selected for training random forest:

- number of trees (*ntree*),
- number of features at each split node (*mtry*),
- minimum number of observations in terminal node (*nodesize*).

These parameters were tuned using a grid search and 10-fold cross-validation. Random Forest algorithm, implemented in package "randomForest" (Liaw et al., 2002), which based on the original Fortran 77 implementation by Breiman, was used.

Since information on treatment outcome is not available before classification, unsupervised imputation procedures had to be used. This is the one of the main differences from the previous study (Milewski et al., 2012), where 2 of 3 used imputation methods were, in fact, supervised. Three, single-imputation algorithms were used before classification:

- a "standard" one,
- kNN-based,
- Random Forest based – "missForest".

A "standard" algorithm imputes missing values based on the mean (for numerical features), median (ordinal) or mode (categorical) of all other training observations. The kNN-based algorithm, implemented in the "VIM" package (Templ et al., 2013), tries to fill in missing data in a similar way to the standard algorithm, but utilizes only the k nearest observations – "the neighbors". Distances are calculated by using a version of Gover metric. The number of neighbors (k) is a free parameter.

The "missForest" algorithm (Stekhoven et al., 2012a), implemented in the package with the same name (Stekhoven et al., 2012b), starts with "standard" imputation. Then features are sorted by amounts of missing values, starting with the lowest amount. In the next step, an iterative procedure is used. For each feature, a random forest is trained, treating the selected feature as the dependent one and observations with filled values in that feature as the learning set. After training, missing values are imputed using predictions from the previously trained random forest. Depending on the type of feature (can be both – categorical or continuous), random forest for either classification or regression is used. After all features are imputed, the stopping criterion is evaluated, and depending on it the algorithm stops or continues to the next iteration. The stopping criterion is met as soon as the difference between the newly imputed data matrix and the previous one increases for the first time with respect to both variable types, if present (Stekhoven et al., 2012a).

A standard imputation procedure was used on the whole dataset, using the learning part only as a base. This was implemented manually, due to the lack of such a procedure in R.

For the "missForest" algorithm, the *ntree* and *mtry* parameters were chosen to tune. Note, that it is difficult to simply use a cross-validation on top of "missForest", due to its iterative nature. Instead, a ten step, pseudo cross-validation procedure was used to estimate error of imputation for each subset of tuned parameters. In each step, an additional 5% of the learning data were marked as missing, and the algorithm was trained on such data. After that, imputation error was calculated. For continuous variables, a normalized mean root square error ($NRMSE$) was used to calculate error. For categorical ones, a percent of false classified (PFC) data was used. Those measures were defined as follows (Oba et al., 2003):

$$NRMSE = \sqrt{\frac{mean\left((X_{con,true} - X_{con,imp})^2\right)}{\mathrm{var}(X_{con,true})}} \tag{1}$$

$$PFC = \frac{count(X_{cat,true} \neq X_{cat,imp})}{\#NA_{cat}} \tag{2}$$

where: $X_{con,true}$, $X_{cat,true}$ are original data with continuous (categorical) features; $X_{con,imp}$, $X_{cat,imp}$ are imputed data with continuous features, $\#NA_{cat}$ equals the number of added missing values

Note, that values originally missing are not counted in those measures (that is the missing values in the $X_{con,true}$, $X_{cat,true}$ datasets). Those two values were calculated at each step, and averaged at the end for each pair of *ntree* and *mtry*. To get an optimal pair, each such pair was ranked based on $NRMSE$ and PFC values separately. The pair that minimized the sum of ranks was chosen as the final one. If there were more such pairs, the one with less *ntree*, and at the end, less *mtry* (by minimizing those parameters, the algorithm became more simple, computational and logical) was chosen. After the best parameters were obtained, the whole dataset was imputed based on all data (again, this is the result of the iterative nature of this algorithm). This pseudo cross-validation procedure, and functions used to calculate $NRMSE$ and PFC (where the original dataset contains missing values, and we do not want to count them) were implemented manually due to lack of such procedures in R.

The same pseudo cross-validation procedure was used to tune the number of neighbors in the kNN imputation algorithm. This time, if there were two or more k, which minimizes the sum of ranks, the lesser one was chosen (by minimizing this parameter, the algorithm became more simple, computational and logical).

Data Preparation and Imputation

In further analysis, only features filled in more than 20% were used. Features containing only one value were also removed. After reduction, 108 features remained and only 5% of data were missing. The dataset was then divided into 4 sets corresponding to 4 selected treatment phases:

- medical history (data available before treatment),
- beginning of the treatment,
- gametes selection and moment of fertilization,
- from fertilization to embryo transfer.

Those 4 sets are presented in Figure 1. The first feature is the dependent one, then the features from the next sets follow. Each set is divided from the others by a black line. Missing values are marked as black, other values

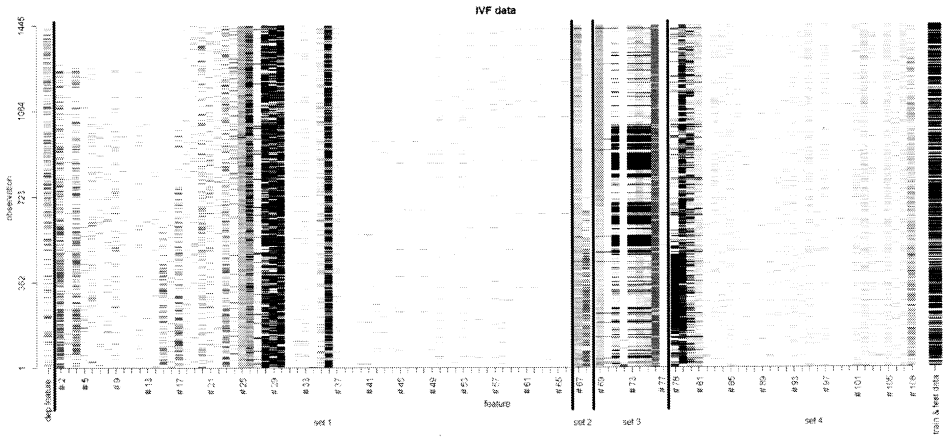


Figure 1. Division of the dataset

are gray-scaled. At the right side, there is a cross-validation indicator – observations marked as black (white) are training (validation) data.

Based on those four sets, four new datasets were built by utilizing the additive rule; that is, each new dataset contained the previous one and some additional features. Table 2 symbolically presents this (+ means, that given subset is present in dataset). This way, each new dataset contains information available at a chosen treatment phase.

Table 2. Datasets creation

New dataset id	Number of features	Sets			
		medical history	begin of the treatment	gametes selection and fertilization moment	from fertilization to embryo transfer
1	66	+			
2	68	+	+		
3	77	+	+	+	
4	108	+	+	+	+

After creation of the four subsets, they were imputed by three previously described procedures. During the pseudo cross-validation phase, the following ranges of parameters were tuned:

- kNN imputation: k (number of neighbors) in 1–100 range,
- missForest imputation: *ntree* in 30–240 by 30 range, *mtry* in 3–14 range.

Ranges for missForest were chosen partially according to recommendations (Stekhoven et al., 2012a). Results for kNN imputation are presented in Figure 2 (dataset 1 and 2) and Figure 3 (dataset 3 and 4). The gray lines present k optimizing specified measures. The black lines present k, which optimizes the sum of ranks – the final ones. Results for missForest imputation are presented in Figure 4 (dataset 1 and 2) and Figure 5 (dataset 3 and 4). Similar to earlier pictures, a gray '+' sign presents coordinates (pair *ntree* and *mtry*) optimizing specified measures (which are printed also). The black '+' presents a pair that optimizes the sum of ranks. If a gray mark (either for kNN or missForest) is not present, then it is the same as the black one (the set of parameters optimizing a particular measure is the same as the set optimizing the sum of ranks).

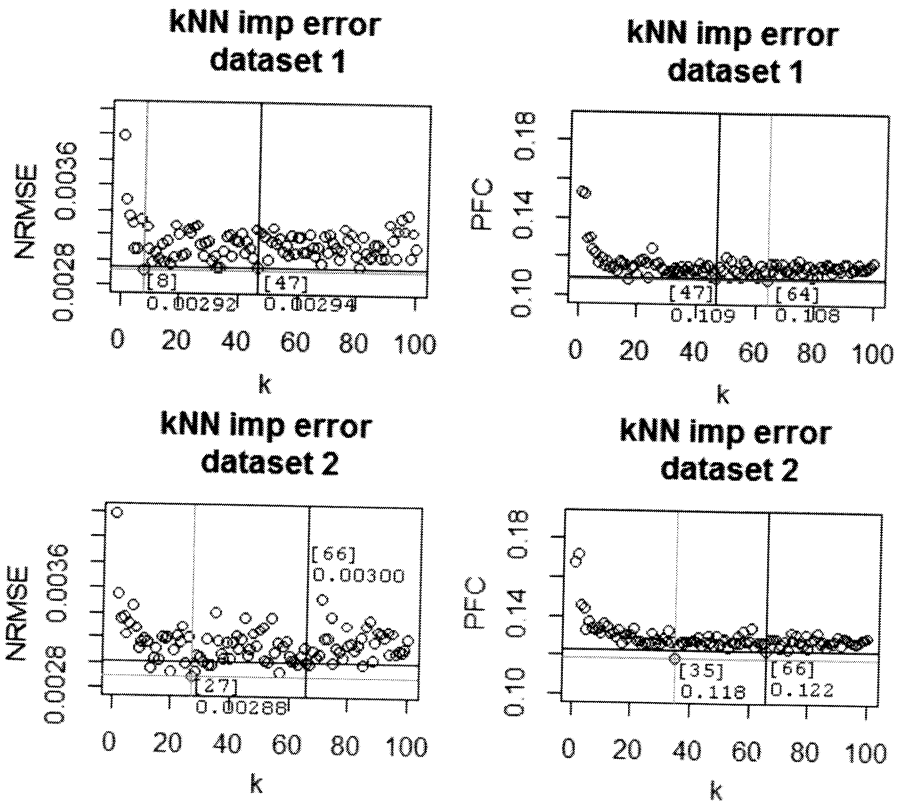


Figure 2. Results from pseudo cross-validation imputation for kNN method; datasets 1 and 2

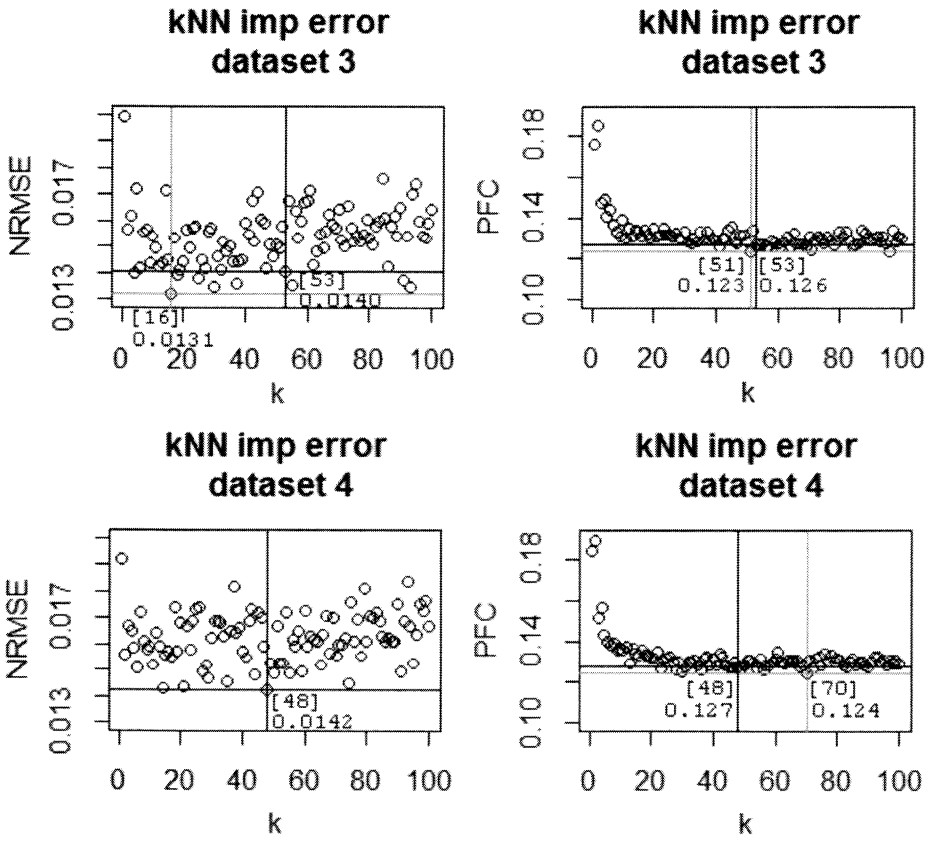


Figure 3. Results from pseudo cross-validation imputation for kNN method; datasets 3 and 4

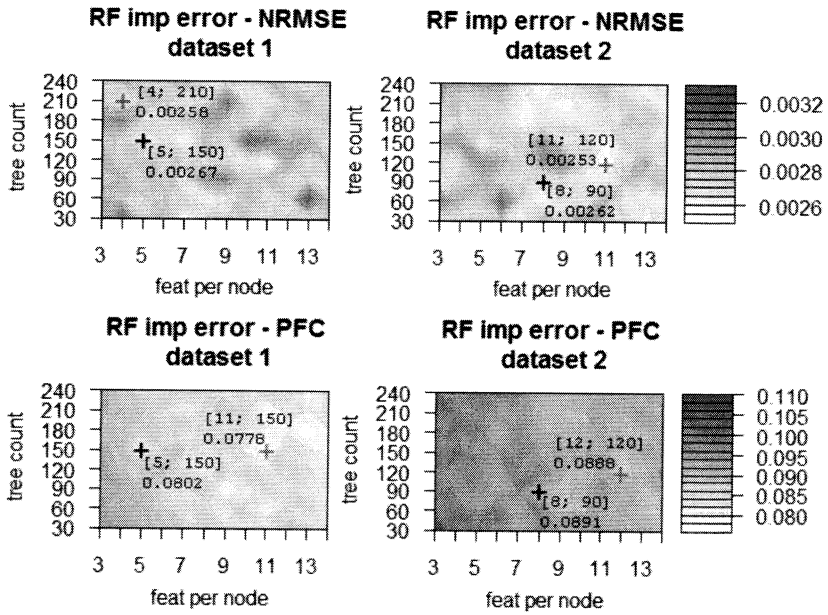


Figure 4. Results from pseudo cross-validation imputation for missForest method; datasets 1 and 2

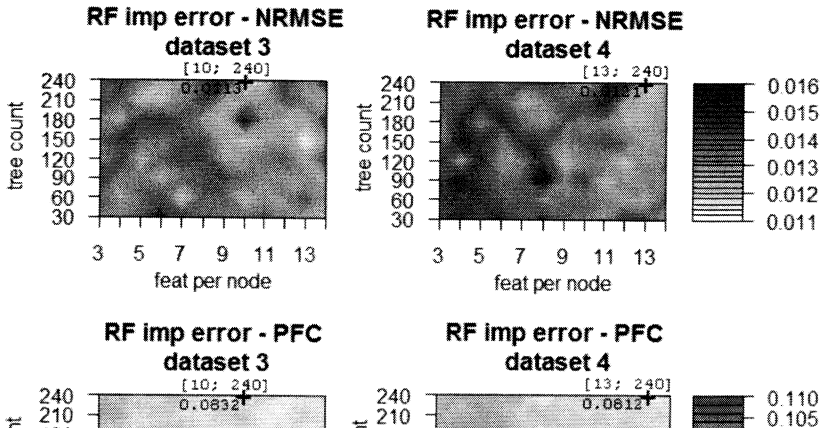


Figure 5. Results from pseudo cross-validation imputation for missForest method; datasets 3 and 4

Classification Results

After imputation, 12 datasets were created (the result of using 3 imputation methods on 4 datasets). Those datasets were classified using the Random Forest algorithm. During the cross-validation phase, the following ranges of parameters were tuned:

- number of trees: from 200 to 3000 by 200,
- number of features per node: from 2 to 20,
- minimum number of observations in node: form 1 to 10.

The specified range of parameters is based on recommendations (Breiman, 2001). The best classifier was tested on the validation dataset to produce an unbiased estimate of classification performance. Because it is difficult to visualize a 3-dimensional parameter set, results for best *nodesize* only are presented. Figure 6 presents results for datasets 1 and 2, which are also presented in Table 3 and Table 4 as full contingency tables for validation datasets and best mix of algorithms for those datasets. Figure 7 presents results for datasets 3 and 4, which are also presented in Table 5 and Table 6 as full contingency tables for validation datasets and best mix of algorithms for those datasets.

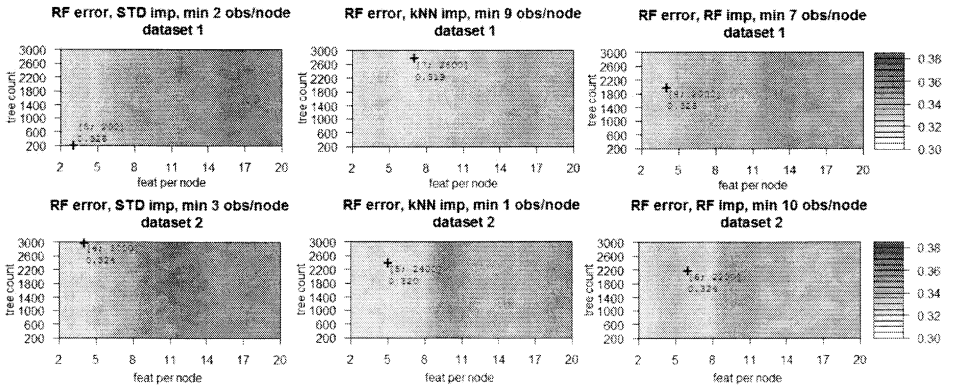


Figure 6. Results from cross-validation classification; datasets 1 and 2 imputed in 3 ways

Table 3. RF accuracy on kNN-imp validation dataset 1

Outcome prediction on kNN-imp validation observations dataset 1		Predicted outcome		Accuracy
		no	yes	
Observed outcome	no	274	7	0.975
	yes	145	8	0.052
Accuracy		0.654	0.533	0.650

Table 4. RF accuracy on kNN-imp validation dataset 2

Outcome prediction on kNN-imp validation observations dataset 2		Predicted outcome		Accuracy
		no	yes	
Observed outcome	no	275	6	0.979
	yes	142	11	0.072
Accuracy		0.658	0.647	0.659

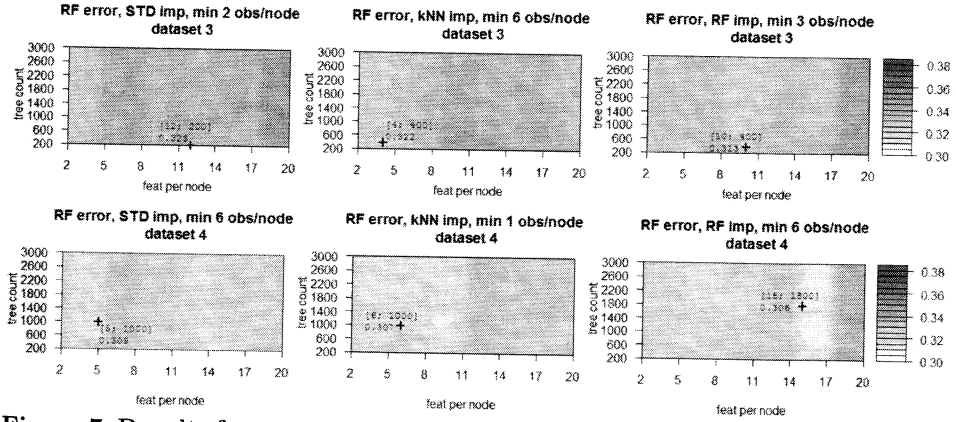


Figure 7. Results from cross-validation classification; datasets 3 and 4 imputed in 3 ways

Table 5. RF accuracy on RF-imp validation dataset 3

Outcome prediction on RF-imp validation observations dataset 3		Predicted outcome		Accuracy
		no	yes	
Observed outcome	no	267	14	0.950
	yes	137	16	0.104
Accuracy		0.661	0.533	0.652

Table 6. RF accuracy on RF-imp validation dataset 4

Outcome prediction on RF-imp validation observations dataset 4		Predicted outcome		Accuracy
		no	yes	
Observed outcome	no	268	13	0.954
	yes	127	26	0.170
Accuracy		0.678	0.667	0.677

Conclusions

Classification results in validation datasets differ from those obtained in k-fold cross-validation by 3–4 percent, which is a very good achievement. Since validation datasets were not used in the training of the classifier, this small difference guarantees error rate stability on future, unknown data. In the second selected phase of the treatment, accuracies for both possible outcomes are almost the same (about 2/3), with is also a very rare result. They are also comparable to results obtained from a full feature set. Treatment results can then be successfully predicted at the beginning of the actual treatment process, which is very important. Pregnancy prediction from medical history alone is far worse for the "yes" response, very near to 50% (the result ofa coin toss). Datasets 1 and 2 differ by only two features, so they should be very relevant to a successful outcome. The 3rd dataset contains a few new features compared to the 2nd, but they actually worse the "yes" response accuracy. The overall accuracies for the four datasets are almost the same, and worse by 12% than in the previous study (Milewski et al., 2012). This should be attributed to the previously mentioned change of imputation algorithms.

The analyzed database again proved to be resistive to successful classification even using state-of-the-art algorithms. However, this study shows that outcomes can be predicted in earlier phases of treatment and that predictions can be consistent in terms of success or failure. When comparing obtained results to previous, the focus should be on imputation algorithms. Since unsupervised and supervised methods are somewhat extreme paradigms, maybe some semi-supervised methods may improve classification results without compromising limited information availability at the beginning of treatment.

REFERENCES

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2 (3), 18–22.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch F. (2012). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-1. Retrieved from http://CRAN.R-project.org/package=e1071.

Milewski, R., Malinowski, P., Milewska, A. J., Ziniewicz, P., Czerniecki, J., Pierzyński, P., & Wołczyński S. (2012). Classification issue in the IVF ICSI/ET data analysis. *Studies in Logic, Grammar and Rhetoric*, 29(42), 75–85.

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*, 25(38), 49–57.

Milewski, R., Malinowski, P., Milewska, A. J., Ziniewicz, P., & Wołczyński, S. (2010). The usage of margin-based feature selection algorithm in IVF ICSI/ET data analysis. *Studies in Logic, Grammar and Rhetoric*, 21(34), 35–46.

Milewski, R., Milewska, A. J., Czerniecki, J., Leśniewska, M., & Wołczyński, S. (2013). 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.

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K., & Ishii, S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16), 2088–2096.

Radwan, J. (2011). Epidemiologia niepłodności. In J. Radwan, & S. Wołczyński (Eds.), *Niepłodność i rozród wspomagany* (pp. 11–14). Poznań: Termedia.

Stekhoven, D. J., & Bühlmann, P. (2012a). MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 1(28), 112–118.

Stekhoven, D. J., & Bühlmann, P. (2012b). missForest: Nonparametric Missing Value Imputation using Random Forest R package version 1.3. Retrieved from http://CRAN.R-project.org/package=missForest.

Templ, M., Alfons, A., Kowarik, A. & Prantner, B. (2013). VIM: Visualization and Imputation of Missing Values. R package version 3.0.3.1. Retrieved from http://CRAN.R-project.org/package=VIM.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0031

Poincaré Plots in Analysis of Selected Biomedical Signals

Agnieszka Kitlas Golińska[1]

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

Abstract. Poincaré plot is a return map which can help perform graphical analysis of data. We can also fit an ellipse to the plot shape by determining descriptors $SD1$, $SD2$ and $SD1/SD2$ ratio to study the data quantitatively. In this paper we show examples of application of Poincaré plots in analysis of various kinds of biomedical signals: RR intervals, EMG, gait data and EHG.

Introduction

In the study of biomedical signals we are always searching for new methods of analysis. The most popular are linear methods, like Fourier transform, but in recent years we have observed increased interest in new, nonlinear methods, like eg. methods originating from chaos theory. In this paper we share the results of the application of an interesting and simple nonlinear method – Poincaré plot.

This method can be used in analysis of non-filtered and also non-stationary data. Some authors (Brennan et al., 2001; Karmakar et al. 2009; Tulppo et al., 1996) claim that it is a valuable method due to its ability to display nonlinear features of the time series.

Our aim is to show how this method works on biomedical signals and what new information can be obtained in this way.

The most common application of Poincaré plot is to ECG data (R-R intervals). In this paper we focus on EMG, gait data, and EHG and only mention R-R intervals because they are extensively studied.

Poincaré Plots – Basics, Descriptors $SD1$ and $SD2$ and Selected Examples

Poincaré plots are return maps in which each result of measurement is plotted as a function of a previous one. It is a simple and effective concept.

ISBN 978-83-7431-392-6 ISSN 0860-150X

The idea is as follows:

Let us denote the data by: $x_0, x_1, x_2, x_3, x_4, \ldots$. The return map will be a plot of the points (x_0, x_1), (x_1, x_2), (x_2, x_3), (x_3, x_4), ...

A shape of the plot describes the evolution of the system and allows usto visualize the variability of time series x_n (Hoshi et al., 2013; Karmakar et al., 2009).

There are standard descriptors used in quantifying Poincaré plot geometry, namely $SD1$ and $SD2$ (Brennan et al., 2001; Piskorski et al., 2007; Tulppo et al., 1996). We can obtain them by fitting an ellipse to the plot shape (Figure 1). Descriptors $SD1$ and $SD2$ represent the minor and the major semi-axes of this fitted ellipse. Brennan et al. (2001) gave a description of $SD1$ and $SD2$ in terms of linear statistics. $SD1$ is the standard deviation of the distances of points from axis 1 and determines the width of the ellipse (short-term variability), $SD2$ equals the standard deviations from axis 2 and length of the ellipse (long-term variability).

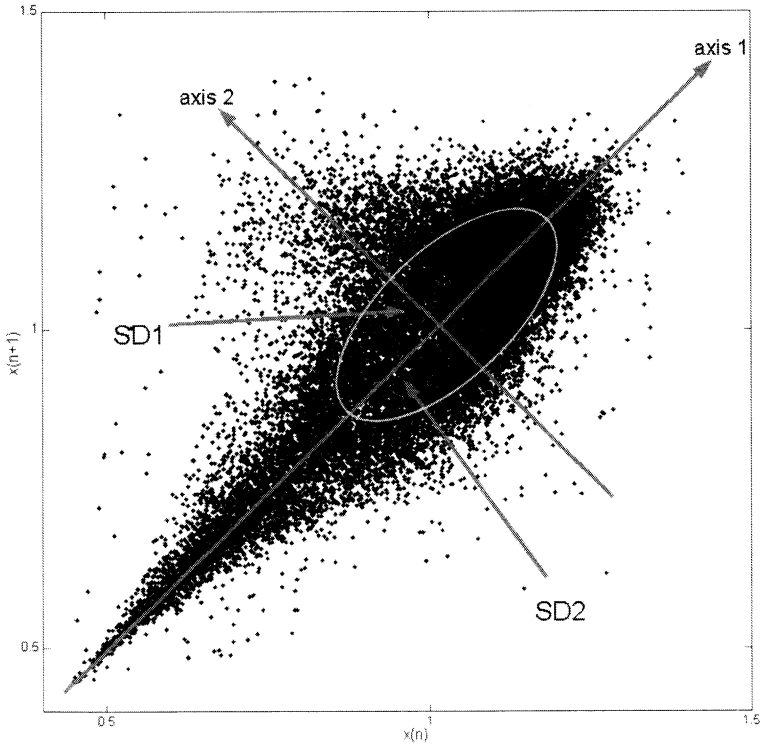


Figure 1. The idea of an ellipse fitted to the Poincaré plot and descriptors $SD1$ and $SD2$

Descriptors $SD1$ and $SD2$ can be defined as:

$$SD1 = \frac{\sqrt{2}}{2} SD(x_n - x_{n+1}) \tag{1}$$

$$SD2 = \sqrt{2SD(x_n)^2 - \frac{1}{2} SD(x_n - x_{n+1})^2} \tag{2}$$

where SD is a standard deviation of the time series.

Uncorrelated noise and its Poincaré plot are shown in Figure 2. Here we have no structures on the Poincaré plot.

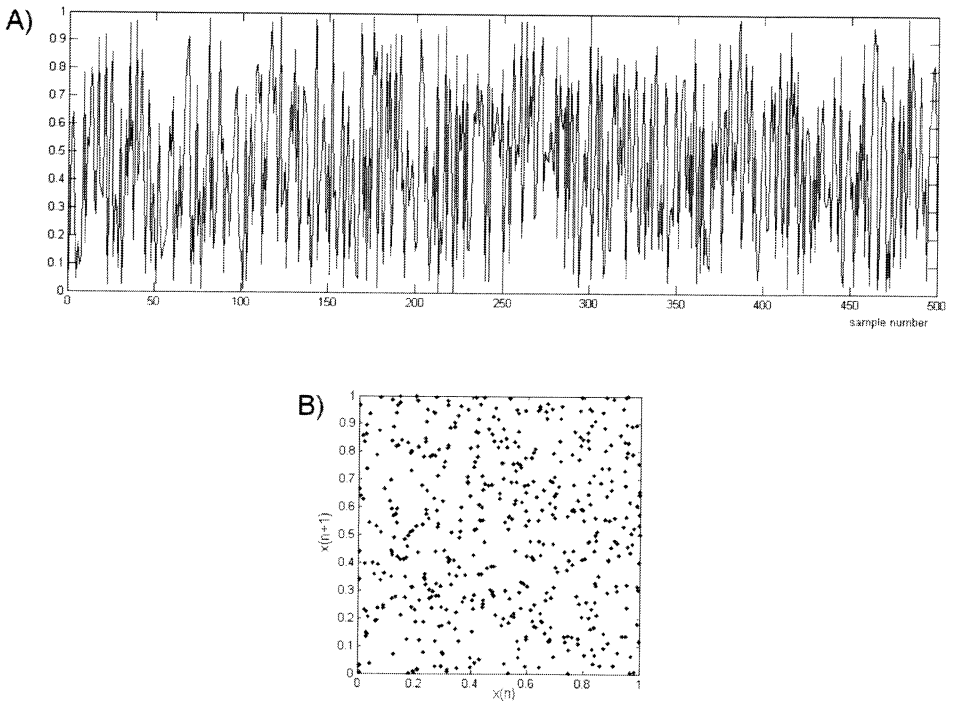


Figure 2. A) Uncorrelated noise and B) its Poincaré plot

Periodic function (here function *cos*) and its Poincaré plot are shown in Figure 3. Here we obtain an ellipse on the Poincaré plot.

As we can see above one can observe various kinds of shapes of Poincaré plots. In the literature, we can find mainly comet-shaped plots, torpedo-shaped plots and ellipse-shaped plots (Brennan et al., 2001; Kitlas et al., 2004; Tulppo et al., 1996). In ECG record studies,the shape of the Poincaré plot can be categorized into functional classes associated with degrees of heart failure (Woo et al., 1992).

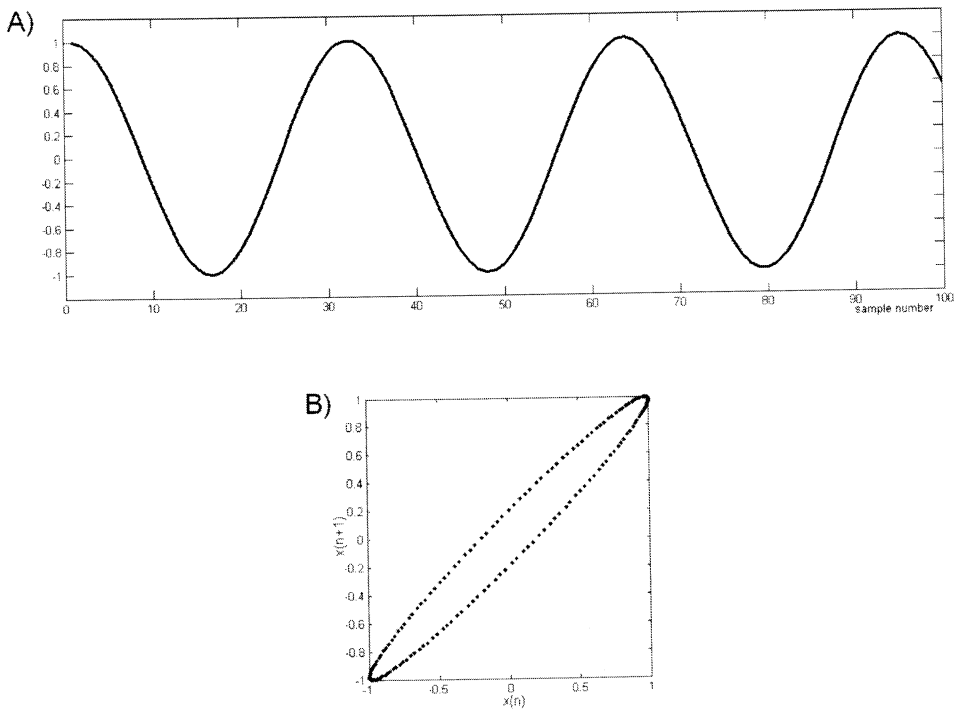


Figure 3. A) Periodic function *cos* and B) its Poincaré plot

Analyzed Signals (R-R Intervals, EMG Signals, Gait Data and EHG Records) and their Poincaré Plots

We analyzed various kinds of signals: RR intervals, EMG, gait data and EHG. All these signals were obtained from Physionet (Goldberger et al., 2000).

As mentioned above, R-R intervals were extensively studied by means of Poincaré plot, so we present only basic information about it. R-R intervals are the series of time intervals between the beats of heart (Brennan et al., 2001).

There are interpretations of $SD1$ and $SD2$ in these kindsof studies. $SD1$ is an instantaneous beat-to-beat variability and $SD2$ acontinuous beat-to-beat variability (Brennan et al., 2001; Kitlas et al., 2004; Piskorski et al, 2007). The $SD1/SD2$ ratio represents the randomness in the heart rate variability time series (Biala et al., 2010).

RR-intervals and Poincaré plot are shown in Figure 4. Values of $SD1$, $SD2$ and their ratio for R-R intervals (normal sinus rhythm) are presented in Table 1.

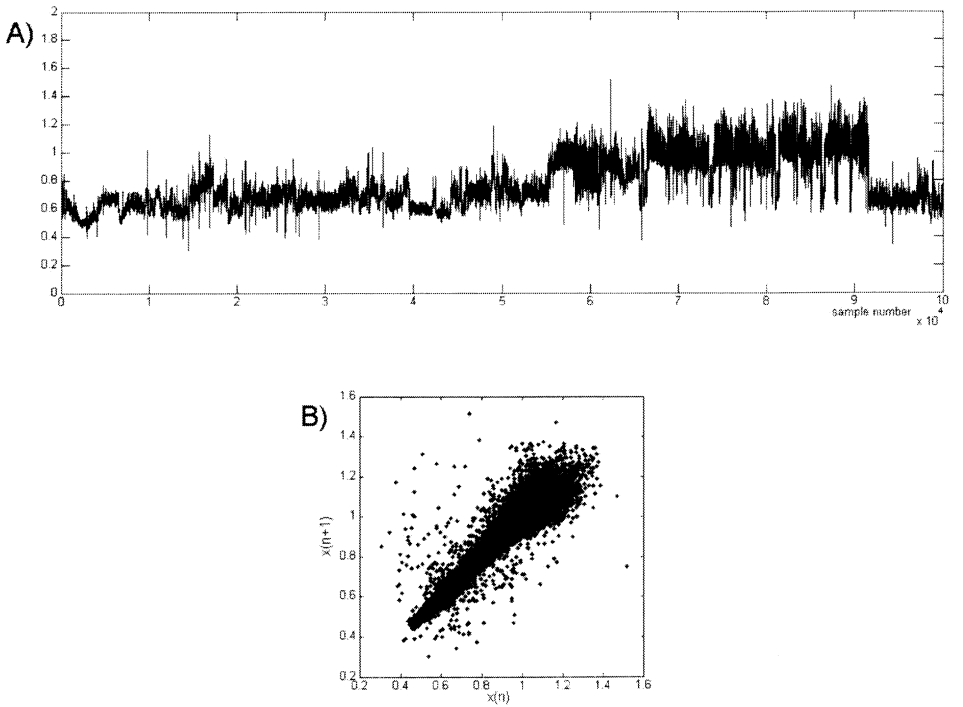


Figure 4. A) R-R intervals from a healthy subject (normal sinus rhythm) and B) Poincaré plot

Table 1. Values of descriptors $SD1$, $SD2$ and $SD1/SD2$ ratio for selected R-R intervals

R-R intervals	$SD1$	$SD2$	$SD1/SD2$ ratio
subject with normal sinus rhythm	0.019	0.189	0.103

Electromyography (EMG) is an important technique for evaluating activity and properties of muscles. Selected EMG signals and their Poincaré plots are presented in Figure 5. These are short EMG recordings from three subjects: one healthy, one with myopathy and one with neuropathy. EMG records were obtained using a 25 mm concentric needle electrode placed in the tibialis anterior muscle. Subjects dorsiflexed the foot gently against resistance and then relaxed (Goldberger et al., 2000). Values of $SD1$, $SD2$ and their ratios are presented in Table 2. One can observe that the lowest $SD1/SD2$ ratio value is for the healthy subject.

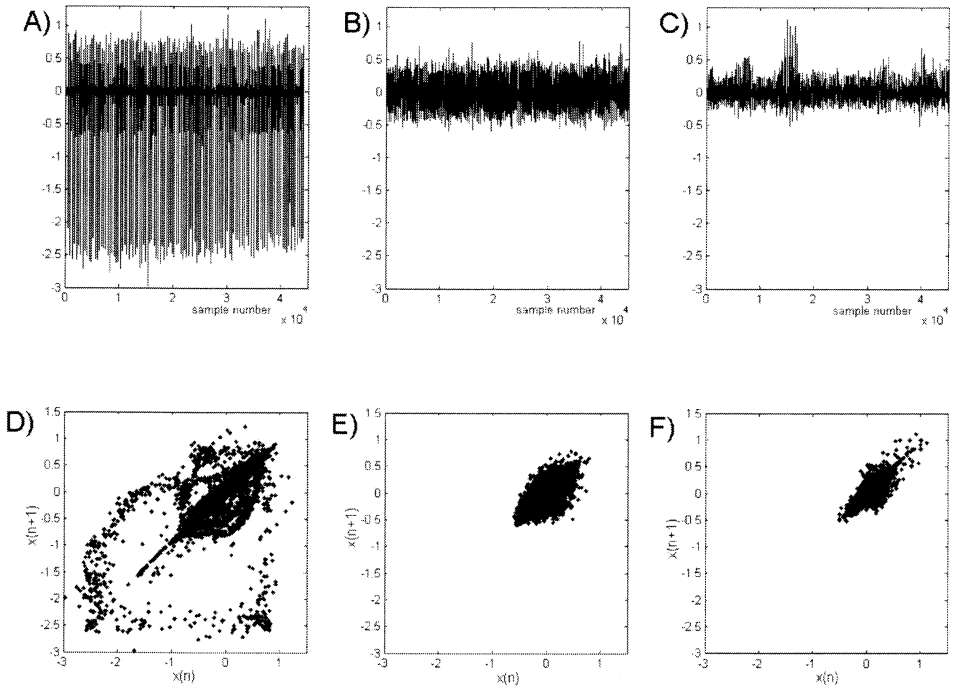


Figure 5. EMG signals from three subjects: A) one with neuropathy, B) one with myopathy and C) one healthy and their respective Poincaré plots D) E) F)

Table 2. Values of descriptors $SD1$, $SD2$ and $SD1/SD2$ ratio for selected EMG signals

EMG signal	$SD1$	$SD2$	$SD1/SD2$ ratio
subject with neuropathy	0.029	0.128	0.230
subject with myopathy	0.037	0.087	0.427
healthy subject	0.016	0.097	0.164

Gait data is the data of human motion. Selected gait data signals and their Poincaré plots are presented in Figure 6. These are short gait data, illustrating the right stride interval from three subjects: one healthy subject, one with Huntington's disease and one with Parkinson's disease. The signals were obtained using force-sensitive resistors, with the output roughly proportional to the force under the foot (Goldberger et al., 2000). Values of $SD1$, $SD2$ and their ratios are presented in Table 3. One can observe that the lowest $SD1/SD2$ ratio value is for the healthy subject.

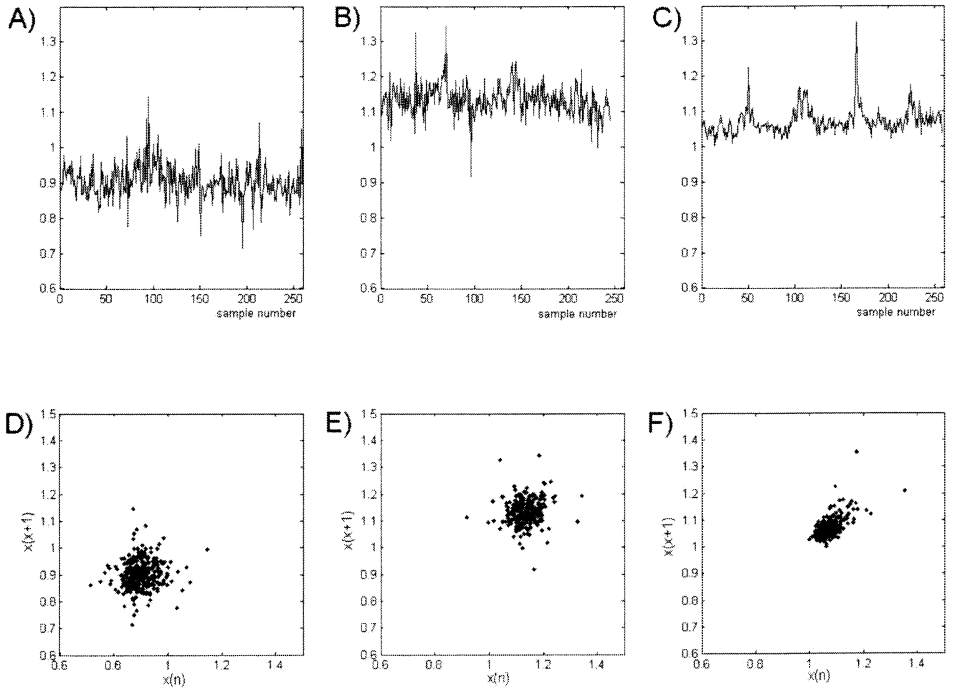


Figure 6. Gait data signals from three subjects: A) one with Huntington's disease, B) one with Parkinson's disease and C) one healthy subject, along with their respective Poincaré plots D) E) F)

Table 3. Values of descriptors $SD1$, $SD2$ and $SD1/SD2$ ratio for selected gait data

Gait data (right stride interval)	$SD1$	$SD2$	$SD1/SD2$ ratio
subject with Huntington's disease	0.045	0.041	1.094
subject with Parkinson's disease	0.037	0.045	0.834
healthy subject	0.017	0.036	0.466

Electrohysterography (EHG) is a technique used for recording changes in electric potential associated with uterine contractions. Selected EHG records and their Poincaré plots are presented in Figure 7. Signals were obtained during pregnancy from two subjects: one with delivery on term and one with preterm delivery (Goldberger et al., 2000). Values of $SD1$, $SD2$ and their ratios are presented in Table 4. One can observe that the lowest $SD1/SD2$ ratio value is for the subject with term delivery.

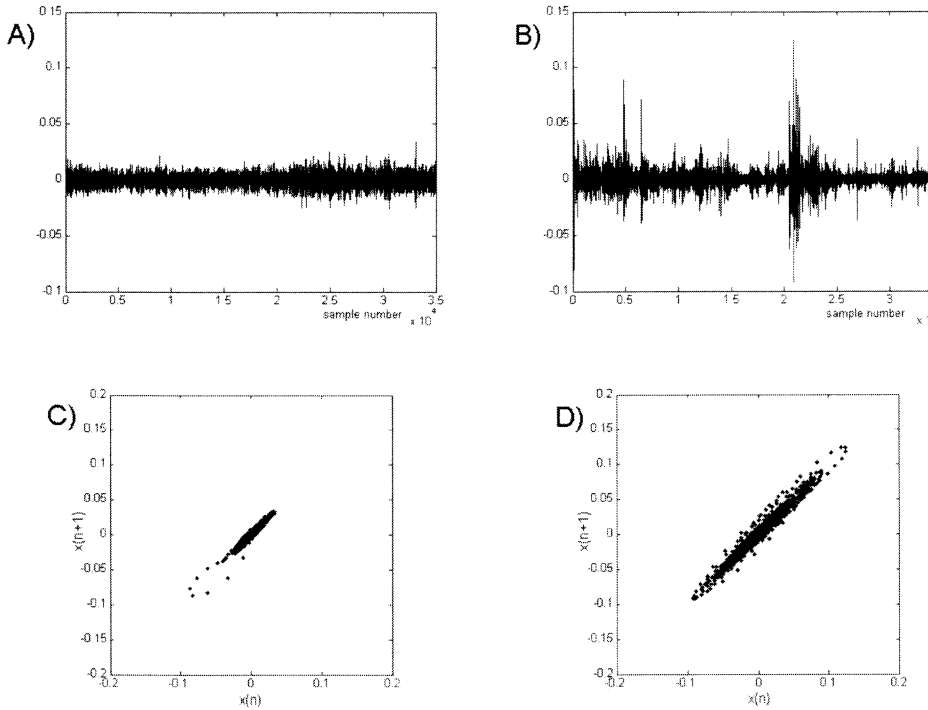


Figure 7. EHG signals (obtained during pregnancy) from two subjects: A) one with preterm delivery, B) one with term delivery and their respective Poincaré plots C) D)

Table 4. Values of descriptors SD1, SD2 and SD1/SD2 ratio for selected EHG signals

EHG signals	$SD1$	$SD2$	$SD1/SD2$ ratio
subject with preterm delivery	0.001	0.008	0.134
subject with term delivery	0.001	0.013	0.081

A Short Literature Overview

There are many papers on the application of Poincaré plots in heart variability studies (mainly R-R intervals), but not in other studies. We mention only afew of them with recent interesting results.

Hoshi et al. (2013) performed the studies to verify whether there is

a correlation between $SD1$, $SD2$, and $SD1/SD2$ ratios and nonlinear indexes (e.g. Lyapunov Exponent, Correlation Dimension, DFA method...) of heart rate variability either in cases of disease or inhealthy conditions. They claim that a high relationship with nonlinear indexes and low relationship with linear indexes suggests nonlinear fractal features.

Biala et al. (2010) studied the effect of maternal smoking habits during pregnancy on healthy and intrauterine growth restricted children (some with asthma). They analyzed ECG records during sleep and suggest that smoking can have adverse implications for the development of the autonomic nervous system of intrauterine growth restricted children, especially those with asthma, who showed a decrease in short-term variability amongst this group.

Karmakar et al. (2009) pointed out that Poincaré plot and descriptors SD1 SD2 have some limitations, so they provided a new descriptor – Complex Correlation Measure – to study the temporal variation of Poincaré plot.

Piskorski et al. (2007) studied the geometry of Poincaré plot of RR intervals and redefined $SD1$, which allowed them to define two new useful descriptors. They have shown that there is an asymmetry in the Poincaré plot and called it heart rate asymmetry. This paper is very interesting from a theoretical point of view.

In our previous studies, we have analyzed heart rate variability using Poincaré plots in cases of children with diabetes type 1 and late vascular complications (Kitlas et al., 2004). There were two groups of children: atest group of 35 children and acontrol group of 50 healthy children. For each subject, a 24-hour Holter ECG was recorded and then divided into two segments: day activity and night activity. $SD1/SD2$ ratios were lower for unhealthy children, which indicates a more regular heart rate. Similar results were obtained by Faust et al. (2012). They studied ECG records from 15 patients with diabetes and from 15 healthy volunteers by means of different methods, including Poincaré plots. They computed descriptor $SD2$ and compared the results – diabetic subjects had lower values of this descriptor than healthy subjects.

Conclusion

Poincaré plot is a nonlinear method that can help us to analyze signals qualitatively and quantitatively. It reflects the variability of data. It seems that too low or too high $SD1/SD2$ ratio values are connected with illness,

with respect to the biomedical signal. This requires further study. We hope that other signals, not only R-R intervals, will be investigated by means of Poincaré plot. Through such investigation, a new interpretation of this method can emerge in biomedical signal studies.

REFERENCES

Biala, T., Godge, M., Schlindwein, F. S., & Wailoo, M. (2010). Heart rate variability using Poincaré plots in 10 year old healthy and intrauterine growth restricted children with reference to maternal smoking habits during pregnancy. In Conference Proceeding: Computingin Cardiology, 26–29 September 2010 (pp. 971–974). Belfast, Ireland.

Brennan, M., Palaniswami, M., & Kamen, P. (2001). Do existing measures of poincare plot geometry reflect nonlinear features of heart rate variability. *IEEE Transactions on Biomedical Engineering*, 48, 1342–1347. DOI: 10.1109/10.959330.

Faust, O., Acharya, U. R., Molinari, F., Chattopadhyay, S., & Tamura, T. (2012). Linear and non-linear analysis of cardiac health in diabetic subjects. *Biomedical Signal Processing and Control*, 7, 295–302. DOI: 10.1016/j.bspc.2011.06.002.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23), e215–e220. Retrieved July 30, 2013, from Circulation Electronic Pages: http://circ.ahajournals.org/cgi/content/full/101/23/e215.

Hoshi, R. A., Pastre, C. M., Vanderlei, L. C. M., & Godoy, M. F. (2013). Poincaré plots indexes of heart rate variability: relationship with other nonlinear variables. *Autonomic Neuroscience*, Retrieved July 30, 2013, from ScienceDirect database on the World Wide Web: http://www.sciencedirect.com. DOI: 10.1016/j.autneu.2013.05.004.

Karmakar, C. K., Khandoker, A. H., Gubbi, J., & Palaniswami, M. (2009). Complex correlation measure: a novel descriptor for Poincaré plot. *BioMedical Engineering OnLine*, 8(17). Retrieved July 30, 2013, from BioMedical Engineering OnLine on the World Wide Web: http://www.biomedical-engineering-online.com. DOI: 10.1186/1475-925X-8-17.

Kitlas, A., Oczeretko, E., Kowalewski, M., & Urban, M. (2004). Poincaré plots in analysis of heart rate variability. *Physica Medica*, XX (Suppl. 1), 76–79.

Piskorski, J., & Guzik, P. (2007). Geometry of Poincaré plot of RR intervals and its asymmetry in healthy adults. *Physiological Measurement*, 28, 287–300. DOI: 10.1088/0967-3334/28/3/005.

Tulppo, M. P., Makikallio, T. H., Takala, T. E. S., & Seppanen, T. V. H. H. (1996). Quantitative beat-to-beat analysis of heart rate dynamics during exercise. *American Journal of Physiology*, 271, H244–H252.

Woo, M. A., Stevenson, W. G., Moser, D. K., Trelease R. B., & Harper, R. M. (1992). Patterns of beat-to-beat heart rate variability in advanced heart failure. *American Heart Journal*, 123(3), 704–710.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0043

The Evaluation of Skeletal Age Based on Computer-Supported Methods in Comparison to the Atlas Method

Anna Predko-Maliszewska[1], **Agnieszka Predko-Engel**[2], **Maciej Goliński**[3]

[1] Department of Medical Informatics, University of Bialystok, Poland
[2] Department of Orthodontics, Palack'y University Olomouc, Czech Republic
[3] Department of Programming and Formal Methods, University of Bialystok, Poland

Abstract. This article describes methods used in estimating skeletal age based both on the evaluation of skeletal maturation of the palm and the wrist (Greulich and Pyle's atlas method) and the Cervical Vertebral Maturation method (CVM). The method of evaluating the skeletal age based on the measurement of cervical vertebrae with equations introduced by A. Machorowska-Pieniążek is also mentioned. The article shows results obtained by computer analysis of the age of cervical vertebrae compared to the results gained from the implemented equations provided by A. Machorowska-Pieniążek and the results obtained from the atlas method.

Introduction

The estimation of human skeletal age is one of the ways of identifying the advancement of human biological development. Other characteristics taken into account while estimating the advancement of human biological maturity are: the development of secondary sexual characteristics, the age of first menstruation, voice change in the case of boys, the mineralization of teeth and their eruption, body mass and height. The estimation of biological maturity is widely used in pediatrics to estimate the child's development against the population and to possibly identify any deviations. It is also used in dentistry, allowing, for example, the right time of an orthodontic treatment to be chosen.

Bone age is not closely connected with the metrical age of a human. Among children of the same metrical age there can be significant differences in height and some of them may already have permanent teeth, while others may still have deciduous teeth. According to the time of maturing, the youth can be divided into early, moderately, and late maturing (Fish-

ISBN 978-83-7431-392-6 ISSN 0860-150X

man, 1987). The average age of the beginning of the maximal growth spurt is about 10 years for girls and 12 for boys, and after 2 years the peak of the growth spurt is reached (Hägg et al., 1980; Taranger et al., 1980). A large difference may also appear in the beginning of the maximal growth spurt, which usually occurs between the ages of 8 and 11 for girls, and 10 and 14 for boys (Fishman, 1987) and lasts about 5 years in both cases (Taranger et al., 1980).

Methods Used in the Estimation of Skeletal Age

The methods used in the estimation of skeletal age can be divided into two groups:

1. Those based on x-ray photos of the palm and the wrist
   - The atlas method by Greulich and Pyle (Greulich et al., 1959)
   - TW1-TW2-TW3 (Tanner et al., 1983)
   - The method of the skeletal age estimation by Björk and Helm (Helm et al., 1971)
   - The method by Hägg and Taranger (Hägg et al., 1980; Taranger et al., 1980)
   - The method Skeletal Maturation Assessment (SMA) by Fishman (Fishman, 1987)
2. Those utilizing analysis of the shape and size of the cervical vertebrae based on the side photos (cephalograms) taken routinely before an orthodontic treatment. An important characteristic of cephalograms is that besides bone tissue, soft tissue is also visible in them. The photo also includes the cervical vertebrae.
   - The method by Lamparski (1972) and its modifications, e.g.:
     - CVMI – cervical vertebrae maturation index (Hassel et al., 1995)
     - The Cervical Vertebral Maturation (CVM) – the estimation of the maturation of the cervical vertebrae by Baccetti, Franchi, and McNamara (Baccetti et al., 2002; 2005)

In this article, only the methods that were used in the examination are presented in detail.

The atlas method by Greulich and Pyle (Greulich et al., 1959) is a method based on the analysis of the development of the bones of the left palm and wrist. It consists of the comparison of an x-ray of the palm with the atlas photos and the choice of the most similar x-ray. This method allows the skeletal age to be determined.

The beginnings of the works on the atlas reach back to the year 1931, and were connected to the research conducted under the direction of Prof. T. W. Todd. The research included a group of healthy children of the white race, and it consisted of making x-rays of the palm and the wrist. During the first year of life, the x-rays were taken every 3 months, after the first year and until the fifth year every 6 months, and then once a year.

The development of the palm and the wrist consists of many chronological changes until the final shape and size are reached. Working on the atlas, the norm for children in the given age was set based on 100 x-ray photos. Standards are separate for boys and girls.

The Cervical Vertebral Maturation method (CVM) developed by Baccetti et al. (2002; 2005) is a modification of the Lamparski method. In the CVM method, only three cervical vertebrae are analyzed: C2, C3, C4.

Baccetti developed the scheme of the changes of the size and shape of the cervical vertebrae. In the initial phase, the C3 and C4 vertebrae are shaped like trapezoids, and the lower edge of the stem is flat. Next, on the lower edge of the vertebrae, a deepening concavity appears – first on C2, then C3, and finally on the C4 vertebra. At the same time, the shape of the vertebrae changes – gradually from a trapezium, through a rectangle with longer horizontal edges, than a square, to a rectangle with longer vertical edges.

Baccetti assigned 6 CVM phases (CS1 to CS6) to the changes in the shape and size of the cervical vertebrae. By identifying the CVM phase, it is possible to estimate the time of the maximal growth spurt. According to Baccetti, the growth spurt begins in the CS3 phase and ends in CS4. Table 1 presents the chronological scheme of the changes of the cervical vertebrae.

Table 1. Chronological scheme of the morphological changes of the C2, C3, and C4 vertebrae

The CVM phase	The C2 vertebra	The C3 vertebra	The C4 vertebra
CS1	– The lower edge is flat	– The lower edge is flat – The vertebra is similar in shape to a rectangular trapezium	– The lower edge is flat – The vertebra is similar in shape to a trapezium
CS2	– A concavity appears on the lower edge	– The lower edge is flat – The vertebra is similar in shape to a rectangular trapezium	– The lower edge is flat – The vertebra is similar in shape to a rectangular trapezium
CS3	– There is an indentation on the lower edge	– A concavity appears on the lower edge – The vertebra is similar in shape to a rectangular trapezium	– The lower edge is flat – The vertebra is similar in shape to a rectangular trapezium

The CVM phase	The C2 vertebra	The C3 vertebra	The C4 vertebra
CS4	– There is an indentation on the lower edge	– There is an indentation on the lower edge – The vertebra takes the shape of a rectangle	– A concavity appears on the lower edge – The vertebra takes the shape of a rectangle
CS5	– There is an indentation on the lower edge	– There is an indentation on the lower edge – Both vertebrae or only one take the shape of a square	
CS6	– There is an indentation on the lower edge	– There is an indentation on the lower edge – Both vertebrae or only one take the shape of a rectangle with longer vertical edges	

A. Machorowska-Pieniążek (2011) undertook research concerned with identifying the sequences of the changes appearing in the cervical vertebrae. She selected a group of 256 boys and girls between the ages of 8 and 18, receiving orthodontic treatment in the years 1993–2003 in the Department of Orthodontics, Department of Dentistry for Children and Adolescents in Zabrze, Medical University of Silesia in Katowice. The research consisted of making vertical and horizontal measurements of the C2, C3, and C4 vertebrae (Figure 1).

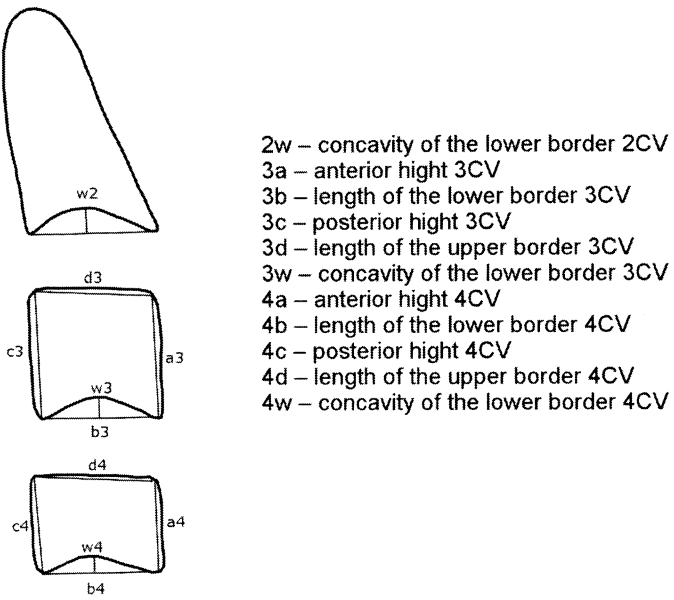


Figure 1. The vertical and horizontal measurements of the C2, C3, and C4 vertebrae

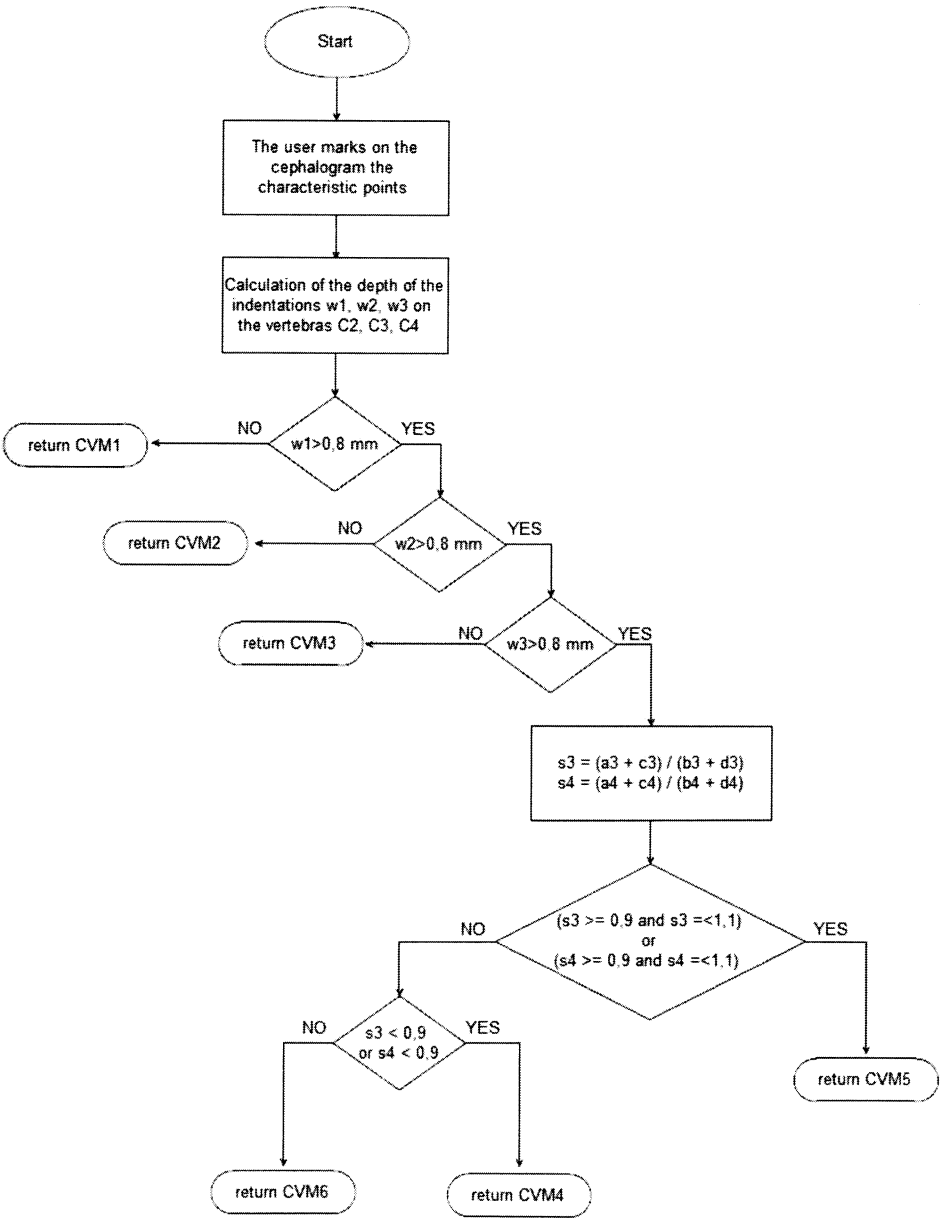


Figure 2. Algorithm identifying the CVM phase

The research showed that the horizontal dimensions of the C3 and C4 vertebrae among girls in the ages of 10 to 14 years, and boys to the age of 16, were larger then the vertical dimensions. Furthermore, the horizontal dimensions showed lower dynamics of growth than the vertical dimensions.

The author, thanks to the undertaken research, derived a formula for the skeletal age for girls (1) and boys (2).

$$CVM(K) = 2.56 + (0.36 \cdot 3a) + (0.26 \cdot SW) + (0.09 \cdot 3D) \quad (1)$$

$$CVM(M) = 3.41 + (0.45 \cdot 4a) + (0.42 \cdot SW) + (0.08 \cdot 3D) \quad (2)$$

where:

- SW – the sum of the lower indentations of the C2, C3, and C4 vertebrae
- $3D$ – the size of the 3C vertebra (3a+3b+3c+3d)

Based on the Baccetti method, an algorithm for identifying the CVM phase was developed (Predko-Maliszewska et al., 2011). In the first step, the user marks the characteristic points on the cephalogram, i.e. the points delineating the lower edge on the C2 vertebra and the indentation on the base of the stem, and on the C3 and C4 vertebrae, the points delineating the upper, front, lower, and back edges of the stem, and also the indentations in the lower edges. Then, the algorithm determines the CVM phase according to Baccetti's procedure.

The scheme (Figure 2) presents the algorithm identifying the CVM phase. The notation refers to Figure 1.

Based on the calculations of the depth of the indentations on the lower edge of the cervical vertebrae and the length and width of the edges of the vertebrae obtained by the CVM phase identification algorithm, formulas (1) and (2) to calculate the skeletal age, provided by A. Machorowska-Pieniążek, were implemented.

Material and Methods

To compare the method of evaluating bone age based on the formulas developed by A. Machorowska-Pieniążek and the atlas method, based on the analysis of the x-ray photos of the palm and the wrist, a group of 44 patients (21 girls and 23 boys), receiving treatment in the Department of Orthodontics LF UP in Olomouc between 1.07.2004 and 11.03.2008, were studied. Table 2 contains the characteristics of the study group.

Every person from the study group was subjected by A. Predko-Engel to:

- the evaluation of skeletal age using the atlas method by Greulich and Pyle based on x-rays of the palm and the wrist,
- the identification of the CVM phase using the computer algorithm to evaluate the skeletal age based on the cervical vertebrae,
- the evaluation of skeletal age using the implemented formulas developed by A. Machorwska-Pieniążek.

Table 2. Characteristic of the study group

Metrical age		Girls		Boys		Total
Age group	Age range	Quantity	Mean age	Quantity	Mean age	
11	⟨10,11)	4	10.5			4
12	⟨11,12)	6	11.5	5	11.6	11
13	⟨12,13)	6	12.3	6	12.5	12
14	⟨13,14)	5	13.7	3	13.1	8
15	⟨14,15)			4	14.3	4
16	⟨15,16)			5	15.3	5
Total		21		23		44

Results

The obtained results allowed us to determine that a correlation between the evaluation using the atlas method and using the formula by A. Machorowska-Pieniążek exists (Table 3). At the same time, both methods show a correlation with the CVM phase using the algorithm for computer identification of the CVM phase.

Table 3. The correlation between used methods

	Atlas method / implemented formulas by A. Machorwska-Pieniążek	Implemented formulas by A. Machorwska-Pieniążek / CVM phase	Atlas method / CVM phase
Correlation – boys	0.884048	0.768216	0.73584
Correlation – girls	0.787521	0.843425	0.694544

However, a difference in the evaluation of the skeletal age exists using both methods. The obtained results allowed us to determine that the atlas method indicates a higher age compared to the age obtained using the method by A. Machorowska-Pieniążek (Figure 3, Figure 4).

The largest differences occurred in the estimation of the skeletal age for girls; smaller differences occurred in the group of boys. Among the girls, the highest noted difference was the value 3.8 years; in the case of the boys, it was 1.8 years.

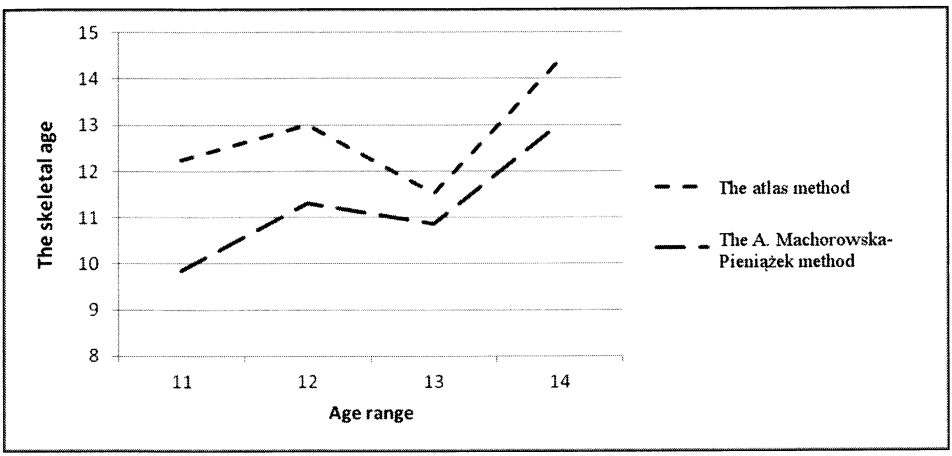


Figure 3. The skeletal age – girls

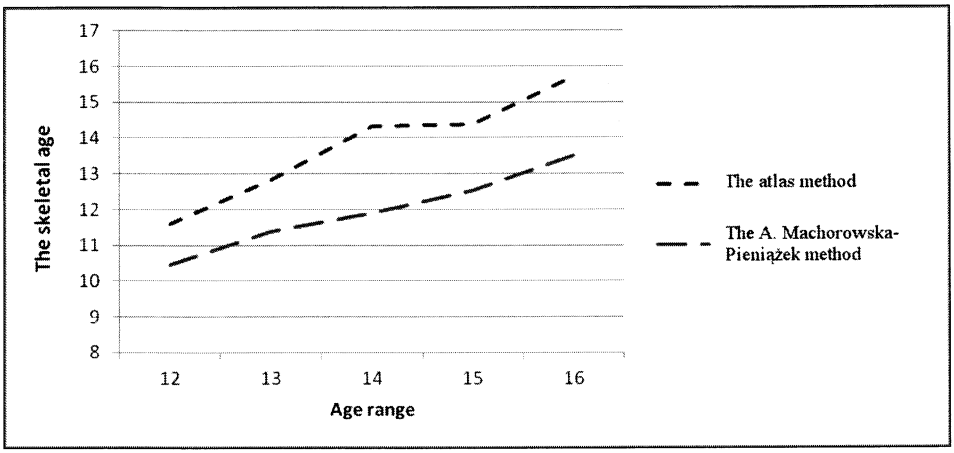


Figure 4. The skeletal age – boys

In the estimation of skeletal age for boys, the greatest differences have a value at 0.5–1 year. For girls, the greatest differences have a value of 1–2 years or more (Table 4).

Table 4. The differences between the atlas method and the A. Machorowska-Pieniążek method in the estimation of skeletal age

	Difference to 0.5 years	Difference between 0.5 and 1 year	Difference between 1 and 2 years	Difference over 2 years
Boys	30.43%	39.13%	30.43%	–
Girls	20.00%	15.00%	35.00%	30.00%

Figure 5. The cephalogram estimated for 11.2 years of skeletal age and the CVM phase 5

The greatest obtained difference in the estimation of the skeletal age in the girl population was found in the x-ray photos of the palm (Figure 6) and in the cephalograms (Figure 5) taken at the age of 11.9 years.

Using the atlas method, the skeletal age was estimated to be 15 years, using the implemented formula it was evaluated to be 11.2 years, while the CVM phase was estimated to be CS5. The difference between those two methods is 3.8 years. The cephalometric photo was an element of research used to determine the repetition of the estimation of the CVM phase (Predko-Engel et al., 2012). The research consisted of the double estimation of the CVM phase made by six doctors in the intervals from 2 weeks

Figure 6. The x-ray of the palm and the wrist estimated for 15 years

to 2 months. The discussed photo was estimated to be in the phase CS1, CS5, CS6 in the first examination, and to be in the phase CS1, CS2, CS6 in the second one. In the first series, two researchers evaluated the photo to be in the CS1, and in the second one, to be in CS6.

Discussion and Conclusions

The analysis of the results allows us to conclude that both methods of evaluating skeletal age generate similar results. However, the implemented method by A. Machorowska-Pieniążek gives a lower result. In the case

of boys, the differences are lower (0.8 years on average), and they are higher in the case of girls (1.4 years on average). The greatest differences between the results of both methods concern boys at the age of 14 years (1.4 years on average) and girls at 11–12 years (2 years on average). It is worth noting that in both the case of girls and boys, the age overlaps the time of the maximal growth spurt.

While analysing the obtained results, one should take into consideration the fact that the estimated skeletal age is important information for a doctor. An evaluation of the skeletal age that is incorrect by 2 years or more may lead to making wrong treatment decisions concerning the patient's plan of treatment and may eventually have a negative effect on the results of a treatment.

The method proposed by A. Machorowska-Pieniążek is based on the size of the cervical vertebrae and the depth of the indentation appearing on the cervical vertebrae. Those same elements are taken into consideration in the method provided by Baccetti with the evaluation of the CVM phase. Therefore, the evaluation using the implemented methods gives a high correlation. However, as it was shown (Predko-Engel et al., 2011), Baccetti's method doesn't correlate highly with Björk's method to estimate the time of the maximal growth spurt.

Also noteworthy is the earlier described case that was evaluated very differently by the researchers using the CVM method. The result provided by A. Machorowska-Pieniążek's method is also different from the evaluation using the atlas method in this case. The obtained difference is significant and is 3.8 years. It is also important that in 30% of the girls' cases, the results obtained by both methods provided differences over 2 years. In this case, A. Machorowska-Pieniążek's method should be recognized as one of the methods of evaluating skeletal age, but it cannot replace other used methods. In the cases where the effect of the treatment depends on the phase of the growth spurt, in which the treatment is undertaken or not, it is recommended to use a few methods of evaluation of the skeletal maturity, because none of the methods are distinctly more precise than the rest.

It should also be taken into consideration that the determination of the characteristic points is not very precise, and even a minimal shift of a point may change the result. However, this is more important in the case of evaluating the CVM phase, than estimating the skeletal age using A. Machorowska-Pieniążek's method. It is also due to the fact that in the developed algorithm of the evaluation of skeletal age using the CVM method, the shift of a characteristic point may result in the change of the proportions of the size of the vertebrae, or reduce or increase the concavity

at the base of the vertebra, which may lead to the photo being qualified as a different CVM phase. In the case of the method by A. Machorowska-Pieniążek, a minimal shift of the characteristic point changes the result only in the decimals.

REFERENCES

Baccetti, T., Franchi, L., & McNamara, J. A. (2002). An improved version of the cervical vertebral maturation (CVM) method for the assessment of mandibular growth. *The Angle Orthodontist*, 72(4), 316–323.

Baccetti, T., Franchi, L., & McNamara, J. A. (2005). The Cervical Vertebral Maturation (CVM) Method for the Assessment of Optimal Treatment Timing in Dentofacial Orthopedics. *Seminars in Orthodontics*, 11(3), 119–129. DOI:10.1053/j.sodo.2005.04.005.

Fishman, L. S. (1987). Maturation patterns and prediction during adolescence. *The Angle Orthodontist*, 57(3), 178–193.

Greulich, W. W., & Pyle, S. I. (1959). *Radiographic Atlas of Skeletal Development of the Hand and Wrist* (2nd ed.). Stanford California: Stanford University Press.

Hägg, U., & Taranger, J. (1980). Skeletal stages of the hand and wrist as indicators of the pubertal growth spurt. *Acta Odontol Scand*, 38(3), 187–200.

Hassel, B., & Farman, A. G. (1995). Skeletal maturation evaluation using cervical vertebrae. *Am J Orthod Dentofacial Orthop*, 107(1), 58–66.

Helm, S., Siersbæk-Nielsen, S., Skieller, V., & Björk, A. (1971). Skeletal maturation of the hand in relations to maximum puberal growth in body height. *Tandlaegebladet (Danish Dental Journal)*, 75 (12), 1223–1234.

Lamparski, D. G. (1972). Skeletal age assessment utilizing cervical vertebrae (Master of Science thesis), University of Pittsburg. In S. J. Stiehl, Die Entwicklung der Halswirbel als Kriterium für die skelettale Reife: vergleich mit der klassischen Methode der Handwurzelaufnahme. Inaugural Dissertation Marburg an der Lahn, (2007).

Machorowska-Pieniążek, A. (2011). Morfometryczna ocena wieku kostnego, sekwencja zmian wzrostowych kręgów szyjnych. *Dent Med Probl*, 48(3), 335–341.

Predko-Engel, A., Kamínek, M., & Langová, K. (2011). Skeletální věk v ortodoncii. *Čes. Stomatol*, 111(6), 154–159.

Predko-Engel, A., Kamínek, M., Langová, K., & Fudalej, P. (2012). Skeletal age according to cervical spine. *Ortodoncie*, 21(4), 218–226.

Predko-Maliszewska, A., & Predko-Engel, A. (2011). Computer estimation of skeletal maturation on the basis of cervical vertebrae maturation, *Studies In Logic, Grammar And Rhetoric*, 25(38), 109–119.

Tanner, J. M., Whitehouse, R. H., Cameron, N., Marshall, W. A., Healy, M. J. R., & Goldstein, H. (1983). *Assessment of skeletal maturity and predictionof adult height (TW2 Method).* London: Academic Press.

Taranger, J., & Hägg, U. (1980). The timing and duration of adolescent growth. *Acta Odontol Scand*, 38(3), 57–67.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0032

Ruby vs. Perl – the Languages of Bioinformatics

Maciej Goliński[1], **Agnieszka Kitlas Golińska**[2]

[1] Department of Programming and Formal Methods, University of Bialystok, Poland
[2] Department of Medical Informatics, University of Bialystok, Poland

Abstract. Ruby and Perl are programming languages used in many fields. In this paper we would like to present their usefulness with regard to basic bioinformatic problems. We concentrate on a comparison of widely used Perl and relatively rarely used Ruby to show that Ruby can be a very efficient tool in bioinformatics. Both Perl and Ruby have a built-in regular expressions (or regexp) engine, which is essential in solving many problems in bioinformatics. We present some selected examples: printing the file content, removing comments from a FASTA file, using hashes, printing nucleotides included in a sequence, searching for a specific nucleotide in sequence and translating nucleotide sequences into protein sequences obtained in GenBank format. It is our belief that Ruby's popularity will rise because of its simple syntax and the richness of its methods. Programs in Ruby are very easy to read and therefore easier to maintain and debug, which are the most important characteristics for a programming language.

Introduction

It is our intent to show that a relatively rarely scientifically-used programming language – Ruby – can be a very efficient tool in the field of bioinformatics, much more so than widely used Perl, and that applications written in Ruby are much easier to read or maintain, and – most of all – easier to write. Ruby, compared to Perl, is a new language, still gaining popularity, while Perl has a well established position as a general-purpose programming language.

The Perl Language

Perl is a programming language developed in 1987 by Larry Wall. It is a dynamic, interpretive, general-purpose language. It incorporates features of other languages including AWK, shell scripting (sh), C, and Lisp (Schwartz et al., 2011).

ISBN 978-83-7431-392-6 ISSN 0860-150X

Perl is sometimes called the hacker language because of its sometimes not easily readable syntax (Foy, 2007). Here is an example of a short, and relatively simple, program which finds the documentation on the *atan2* function and then formats it differently for printing, using a complicated regular expression, a tool which is explained later:

```
#!usr/bin/perl
@lines = 'perldoc -u -f atan2';
foreach (@lines)
{
  s/\w<([^>]+)>/\U$1/g;
  print;
}
```

One of the very important features of Perl languages is the regular expressions they use. Perl is a widely used tool in the field of bioinformatics, especially in the study of the structure and function of genes and proteins.

The Ruby Language

Ruby is a programming language developed in Japan in 1995 by Yukihiro Matsumoto. It is dynamic and reflective, which makes it a very efficient, general-purpose tool. Ruby supports many programming paradigms, including functional, object-oriented and imperative. It is also excellent for metaprogramming, an advanced programming concept. The language was influenced by Perl, Smalltalk, Eiffel and Lisp (Thomas et al., 2009).

One of the most basic ideas for Ruby is that everything is an object, including numbers, classes, and exceptions (Thomas et al., 2009). Thanks to that, a programmer can treat all constructs with a certain universality.

Another important feature of Ruby is a built-in regular expressions handler, which is extremely useful in problems of bioinformatics.

Ruby is very helpful in processing files. It saves the programmer the trouble of remembering to close opened files (which is a very common problem) (Thomas et al., 2009). In addition, it's very easy to manipulate long text files, like those containing nucleotide sequences in FASTA format.

The Regular Expressions

Both Perl and Ruby have a built-in regular expressions (or regexp) engine (Foy, 2007; Thomas et al., 2009), which is essential in solving many problems in bioinformatics. Regexps are an efficient tool in finding parts of

a text (or other sequences of characters) that match a given pattern. To provide a pattern, one should use a special sub-language, created for that purpose. A regexp consists of a number of characters, as well as a few special ones. The pattern is usually placed between slashes "/". Here is a simple example:

```
/gene/
```

This regexp will match a single occurrence of a sequence "gene". This is no different from a natural text searching. To make regular expressions more interesting, we have to introduce a few special symbols.

The most basic symbol is a dot ".", which means "any single character". This means that the regexp:

```
/.at/
```

will match both "rat" and "cat", because they have a single character preceding "at". The pattern:

```
/Ru.y/
```

matches the word "Ruby", but not "Rugby", since there are two characters, where there should be only one.

The next symbol, a vertical line "|", indicates an alternative. The following pattern will match both the words "Perl" and "Ruby":

```
/Perl|Ruby/
```

The characters surrounded by parentheses are grouped together. Grouping and any alternatives often appear together:

```
/(r|c)at/
```

matches the same words as:

```
/(rat)|(cat)/
```

An important note: The following pattern does not mean the same thing as the previous one:

```
/rat|cat/
```

It means the same thing as this one:

```
/ra(t|c)at/
```

There are special characters that signal the beginning and the end of a string. The pattern:

```
/^Ruby/
```

matches the word "Ruby" only if it occurs at the beginning of the analyzed string, and the regexp:

```
/Perl$/
```

will match the word "Perl", only if it is at the end of the string.

Another very important feature of the regular expressions are repetitions. They are a symbol or a set of symbols indicating how many times the previous expression should be repeated in the text.

The question mark "?" matches the preceding element zero or one time. For example:

```
/-?15/
```

matches both "−15" and "15".

The "*" character matches the preceding pattern zero or more times. For example:

```
/10*1/
```

will match eg. "101", "11" or "10000001".

The plus sign "+" denotes one or more repetitions of the preceding pattern. For example:

```
/10+1/
```

will match "101", "100001", "1001", but not "11".

Regular expressions are a very useful tool in the field of bioinformatics, especially in parsing files in the FASTA format.

The FASTA Format and GenBank Format

FASTA format is a text-based format for representing peptide or nucleotide sequences (Baxevanis et al., 2004). In FASTA, amino acids or nucleotides are written in single-letter codes, which makes them easy to process.

A part of the file in FASTA format is presented below (Campylobacter jejuni subsp. jejuni NCTC 11168 complete genome) (National Center for Biotechnology Information, 2006):

```
>gi|30407139|emb|AL111168.1| Campylobacter jejuni subsp. jejuni NCTC 11168 complete
genome
ATGAATCCAAGCCAAATACTTGAAAATTTAAAAAAAGAATTAAGTGAAAACGA
ATACGAAAACTATTTATCAAATTTAAAATTCAACGAAAAACAAAGCAAAGCAG
```

```
ATCTTTTAGTTTTTAATGCTCCAAATGAACTCATGGCTAAATTCATACAAACA
AAATACGGCAAAAAAATCGCGCATTTTTATGAAGTGCAAAGCGGAAATAAAG
CCATCATAAATATACAAGCACAAAGTGCTAAACAAAGCAACAAAAGCACAAAA
ATCGACATAGCTCATATAAAAGCACAAAGCACG
```

In the first line there is a description (comments) of the file and then lines of sequence data. Here we present only 5 lines, although there are many more in this file.

The GenBank (National Center for Biotechnology Information, 2009) is an open access nucleotide and protein sequence database. Files in GenBan format contain an extensive description, nucleotide sequence and its translation to protein sequence (Baxevanis et al., 2004).

A part of the file in GenBank format is presented below (Homo sapiens 43kDa acetylcholine receptor-associated protein (RAPSN) mRNA) (National Center for Biotechnology Information, 2001):

```
/translation="MGQDQTKQQIEKGLQLYQSNQTEKALQVWTKVLEKSSDLMGRFR
VLGCLVAHSEMGRYKEMLKFAVVQIDTARELEDADFLLESYLNLARSNEKLCEFH
KTISYCKTCLGLPGTRAGAQLGGQVSLSMGNAFLGLSVFQKALESFEKALRYAHN
NDDAMLECRVCCSLGSFYAQVKDYEKALFFPCKAAELVNNYGKGWSLKYRAMS
QYHMAVAYRLLGRLGSAMECCEESMKIALQHGDRPLQALCLLCFADIHRSRGDLE
TAFPRYDSAMSIMTEIGNRLGQVQALLGVAKCWVARKALDKALDAIERAQDLAE
EVGNKLSQLKLHCLSESIYRSKGLQRELRAHVVRFHECVEETELYCGLCGESIGE
KNSRLQALPCSHIFHLRCLQNNGTRSCPNCRRSSMKPGFV"

ORIGIN
      1 cccaactggc agcgacagct gcagacgggc tgaaccagct ttgttcccag ggtggcgcct
     61 gctctccatc caggccccat tccggctccc acccgacgct gcttttgttc ccacgtttcg
    121 gggggcagct ggcactgtga ttcctgcccc atgagtgcct agaggcacgg agccaccagg
    181 gatcacccca cgtgggacac agggcttggg gaggatgggg caggaccaga ccaagcagca
    241 gatcgagaag gggctccagc tgtaccagtc caaccagaca gagaaggcat tgcaggtgtg
    301 gacaaaggtg ctggagaaga gctcggacct catggggcgc ttccgcgtgc tgggctgcct
    361 ggtcacagcc cactcggaga tgggccgcta caaggagatg ctgaagttcg ctgtggtcca
    421 gatcgacacg gcccgggagc tggaggatgc cgacttcctc ctggagagct acctgaacct
    481 ggcacgcagc aacgagaagc tgtgcgagtt tcacaagacc atctcctact gcaagacctg
    541 ccttgggctg cctggtacca gggcaggtgc ccagctcgga ggccaggtca gcctgagcat
    601 gggcaatgcc ttcctgggcc tcagcgtctt ccagaaggcc ctggagagct tcgagaaggc
    661 cctgcgctac gcccacaaca atgatgacgc catgctcgag tgccgcgtgt gctgcagcct
    721 gggcagcttc tatgcccagg tcaaggacta cgagaaagcc ctgttcttcc cctgcaaggc
    781 ggcagagctt gtcaacaact atggcaaagg ctggagcctg aagtaccggg ccatgagcca
    841 gtaccacatg gccgtggcct atcgcctgct gggccgcctg ggcagtgcca tggagtgttg
    901 tgaggagtct atgaagatcg cgctgcagca cggggaccgg ccactgcagg cgctctgcct
    961 gctctgcttc gctgacatcc accggagccg tggggacctg gagacagcct tccccaggta
   1021 cgactccgcc atgagcatca tgaccgagat cggaaaccgc ctggggcagg tgcaggcgct
   1081 gctgggtgtg gccaagtgct gggtggccag gaaggcgctg gacaaggctc tggatgccat
   1141 cgagagagcc caggatctgg ccgaggaggt ggggaacaag ctgagccagc tcaagctgca
   1201 ctgtctgagc gagagcattt accgcagcaa agggctgcag cgggaactgc gggcgcacgt
   1261 tgtgaggttc cacgagtgcg tggaggagac ggagctctac tgcggcctgt gcggcgagtc
```

```
1321 cataggcgag aagaacagcc ggctgcaggc cctaccctgc tcccacatct tccacctcag
1381 gtgcctgcag aacaacggga cccggagctg tcccaactgc cgccgctcat ccatgaagcc
1441 tggctttgta tgactcctgg cagcaggcgt gggcttcctc ctcgccactc ctgctctttc
1501 tccactgcac gccagaggcc catttactcc tggggcagct gccaggtcgt cctcaccata
1561 gccaaggcct tggggcctgc ccagggctgc tcccctgggc ccagctcccc tccctgcctc
1621 tttgtacttt gctctttata gaaaaataaa ctgtttgtac ctggtcccag g
```

Selected examples

In this section, we present a few programs very useful in the field of bioinformatics written both in Perl and in Ruby. The purpose of these examples is to present an alternative for the commonly used Perl language, which in our opinion is simpler to write, simpler to read, and simpler to maintain.

The goal of this program is to open a simple text file, and print its contents on the console, line by line.

Perl:

```
open(F, "file.txt");
while($line = <F>)
{
  print "$line"
}
close F;
```

Ruby:

```
File.open("file.txt") do |f|
  while line = f.gets
    print line
  end
end
```

The program in Perl is fairly straightforward. First we open the file, than in a while loop we obtain each line separately, save it in a variable, and print it. The often forgotten part is closing the file, which is both unprofessional and potentially dangerous to the file. The Ruby approach takes care of the last problem automatically by the usage of blocks.

In this program we take a file in a FASTA format, than copy its contents to a second file, omitting the lines containing the comments.

Perl:

```
open (F, "seq.fa");
open (FF,">seq2.txt");
while (<F>)
{
  next if(/^>/);
  print FF;
}
close F;
close FF;
```

Ruby:

```
File.open("seq.fa") do |in|
  File.open("seq2.fa", "w+") do |out|
    while line = in.gets
      out << line unless line =~ /^>/
    end
  end
end
```

Both approaches utilize regular expressions to check if the line begins with a ">" sign. The program in Ruby is shorter, and there is no problem with unclosed files. Also, the part concerning copying the lines is much easier to understand.

The hash is a variation on the table, where instead of just numbers, anything can serve as an index called a key. This program shows the way to use hashes in both languages. The key in the hash is a name of a species, and the value is a gene count. The program prints the names with their gene counts.

Perl:

```
%gene_counts = ("Human" => 31000, "Fruit fly" => 13000,
"Mouse" => 30000, "Chickenpox virus" => 69, "Rice" => 40000,
"Tuberculosis bacteria" => 4000);
while ( ( $key, $value ) = each %gene_counts )
{
    print "$key has $value genes in its genome.\n";
}
```

Ruby:

```
gene_counts = ("Human" => 31000, "Fruit fly" => 13000,
"Mouse" => 30000, "Chickenpox virus" => 69, "Rice" => 40000,
"Tuberculosis bacteria" => 4000)
gene_counts.each_pair {|key, value| puts "#{key}has #{value}
genes in its genome."}
```

Both programs first define the hash. Then, in the Perl approach, we obtain the key-value pair in a while loop, and print the appropriate sentence. The Ruby program is again much simpler, and again, thanks to the use of code blocks.

This example prints the nucleotides that are included in a given sequence. It utilizes both a hash and regular expressions.

Perl:

```
%dict = (A => Adenine, T => Thymine, G => Guanine,
C => Cytosine);
$sequence = 'CTATGCGGTA';
while ( $sequence =~ /./g )
{
    print "$dict{$&}\n";
}
```

Ruby:

```
@dict = {"A" => "Adenine", "T" => "Thymine", "G" => "Guanine",
"C" => "Cytosine"}
sequence = "CTATGCGGTA"
sequence.scan(/./).each {|i| puts @dict[i]}
```

Both programs first define both the hash, which serves as a dictionary for the nucleotides' names, and a fragment of a DNA sequence. The Perl program uses a match operator with an additional "g" modifier, which allows for the scanning of the entire sequence, in order to match patterns to a string. Then, it prints the value corresponding to the letter obtained from the sequence. This method may be a bit difficult to understand. The Ruby approach is simpler thanks to the scan method, which is easier to use than the match operator.

This program is designed to count the occurrences of a specific nucleotide in a given sequence, in this case Adenine (A).

Perl:

```
$sequence="ATGAATCCAAGCCAAATACTTGAAAATTTAAAAAAAGAATTAAGTGAAAAC
GAATACGAAAACTATTTATCAAATTTAAAATTCAACGAAAAACAAAGCAAAGCAGATCTTTT
AGTTTTTAATGCTCCAAATGAACTCATGGCTAAATTCATACAAACAAAATACGGCAAAAAAA
TCGCGCATTTTTATGAAGTGCAAAGCGGAAATAAAGCCATCATAAATATACAAGCACAAAGT
GCTAAACAAAGCAACAAAAGCACAAAAATCGACATAGCTCATATAAAAGCACAAAGCACG";
$sum=0;
@tab=split('', $sequence);
foreach $i (@tab)
{
  $sum++ if $i eq 'A';
}
print $sum;
```

Ruby:

```
@sequence="ATGAATCCAAGCCAAATACTTGAAAATTTAAAAAAAGAATTAAGTGAAAAC
GAATACGAAAACTATTTATCAAATTTAAAATTCAACGAAAAACAAAGCAAAGCAGATCTTTT
AGTTTTTAATGCTCCAAATGAACTCATGGCTAAATTCATACAAACAAAATACGGCAAAAAAA
TCGCGCATTTTTATGAAGTGCAAAGCGGAAATAAAGCCATCATAAATATACAAGCACAAAGT
GCTAAACAAAGCAACAAAAGCACAAAAATCGACATAGCTCATATAAAAGCACAAAGCACG";
@sum=0
@sequence.each_char {|i| @sum+=1 if i == 'A'}
puts @sum
```

The program in Ruby is quite simple. It passes each character of the string into the block, where it is compared with the letter A. The Perl approach is more complicated, since Perl treats strings as a singular value. Therefore, it is impossible to iterate the string. It is necessary to split the string into a table with single characters as elements. This allows for an iteration, and counting of the letter A.

Files in GenBan format contain an extensive description, nucleotide sequence and its translation to protein sequence. How can we obtain this translation? First, we need a genetic code for translation and then we implement programs in Perl and Ruby as one can see below. For this analysis we selected the GenBank: AF111785.1 file (Homo sapiens myosin heavy chain IIx/d mRNA) (National Center for Biotechnology Information, 1998):

Perl:

```
%dict = ("TTT" => "F", "TTC" => "F", "TTA" => "L", "TTG" =>
"L", "CTT" => "L", "CTC" => "L", "CTA" => "L", "CTG" => "L",
```

```
"ATT" => "I", "ATC" => "I", "ATA" => "I", "ATG" => "M", "GTT"
=> "V", "GTC" => "V", "GTA" => "V", "GTG" => "V", "TCT" =>
"S", "TCC" => "S", "TCA" => "S", "TCG" => "S", "CCT" => "P",
"CCC" => "P", "CCA" => "P", "CCG" => "P", "ACT" => "T", "ACC"
=> "T", "ACA" => "T", "ACG" => "T", "GCT" => "A", "GCC" =>
"A", "GCA" => "A", "GCG" => "A", "TAT" => "Y", "TAC" => "Y",
"TAA" => "STOP", "TAG" => "STOP", "CAT" => "H", "CAC" => "H",
"CAA" => "Q", "CAG" => "Q", "AAT" => "N", "AAC" => "N", "AAA"
=> "K", "AAG" => "K", "GAT" => "D", "GAC" => "D", "GAA" =>
"E", "GAG" => "E", "TGT" => "C", "TGC" => "C", "TGA" =>
"STOP", "TGG" => "W", "CGT" => "R", "CGC" => "R", "CGA" =>
"R", "CGG" => "R", "AGT" => "S", "AGC" => "S", "AGA" => "R",
"AGG" => "R", "GGT" => "G", "GGC" => "G", "GGA" => "G", "GGG"
=> "G");
$sequence=uc("atgagttctgactctgagatggccatttttggggaggctgctccttt
cctccgaaagtctgaaagggagcgaattgaagcccagaacaagccttttgatgccaagaca
tcagtctttgtggtggaccctaaggagtcctttgtgaaagcaacagtgcagagcagggaag
gggggaaggtgacagctaagaccgaagctggagctactgtaacagtgaaagatgaccaagt
cttccccatgaaccctcccaaatatgacaagatcgaggacatggccatgatgactcatcta
cacgagcctgctgtgctgtacaacctcaaagagcgctacgcagcctggatgatctacacct
actcaggc");
$goal="MSSDSEMAIFGEAAPFLRKSERERIEAQNKPFDAKTSVFVVDPKESFVKATVQS
REGGKVTAKTEAGATVTVKDDQVFPMNPPKYDKIEDMAMMTHLHEPAVLYNLKERYAAWMI
YTYSG";
$translation="";
while ($sequence =~ /.../g)
{
    $translation .= $dict{$&};
}
print "Success!" if ($translation eq $goal);
```

Ruby:

```
@dict = {"TTT" => "F", "TTC" => "F", "TTA" => "L", "TTG" =>
"L", "CTT" => "L", "CTC" => "L", "CTA" => "L", "CTG" => "L",
"ATT" => "I", "ATC" => "I", "ATA" => "I", "ATG" => "M", "GTT"
=> "V", "GTC" => "V", "GTA" => "V", "GTG" => "V", "TCT" =>
"S", "TCC" => "S", "TCA" => "S", "TCG" => "S", "CCT" => "P",
"CCC" => "P", "CCA" => "P", "CCG" => "P", "ACT" => "T", "ACC"
=> "T", "ACA" => "T", "ACG" => "T", "GCT" => "A", "GCC" =>
"A", "GCA" => "A", "GCG" => "A", "TAT" => "Y", "TAC" => "Y",
```

```
"TAA" => "STOP", "TAG" => "STOP", "CAT" => "H", "CAC" => "H",
"CAA" => "Q", "CAG" => "Q", "AAT" => "N", "AAC" => "N", "AAA"
=> "K", "AAG" => "K", "GAT" => "D", "GAC" => "D", "GAA" =>
"E", "GAG" => "E", "TGT" => "C", "TGC" => "C", "TGA" =>
"STOP", "TGG" => "W", "CGT" => "R", "CGC" => "R", "CGA" =>
"R", "CGG" => "R", "AGT" => "S", "AGC" => "S", "AGA" => "R",
"AGG" => "R", "GGT" => "G", "GGC" => "G", "GGA" => "G", "GGG"
=> "G"};
@sequence="atgagttctgactctgagatggccatttttggggaggctgctcctttcct
ccgaaagtctgaaagggagcgaattgaagcccagaacaagccttttgatgccaagacatca
gtctttgtggtggaccctaaggagtcctttgtgaaagcaacagtgcagagcagggaagggg
ggaaggtgacagctaagaccgaagctggagctactgtaacagtgaaagatgaccaagtctt
ccccatgaaccctcccaaatatgacaagatcgaggacatggccatgatgactcatctacac
gagcctgctgtgctgtacaacctcaaagagcgctacgcagcctggatgatctacacctact
caggc";
@goal="MSSDSEMAIFGEAAPFLRKSERERIEAQNKPFDAKTSVFVVDPKESFVKATVQS
REGGKVTAKTEAGATVTVKDDQVFPMNPPKYDKIEDMAMMTHLHEPAVLYNLKERYAAWMI
YTYSG";
@translation=""
@sequence.upcase.scan(/.../).each
{
    |i| @translation << @dict[i]
}
puts "Success!" if @translation == @goal
```

Both approaches begin by defining a few variables. The first one is a hash serving as a dictionary for the translation. The second one is the given sequence in a GenBank format. The next variable contains the same sequence, but as a natural protein sequence, which is later used to check if the program is valid. The solution in Perl is far more difficult to comprehend than the one in Ruby. The *uc* function in Perl transforms the string into upper case.

A Quick Overview of Literature

There are many books or papers on the application of Perl in bioinformatics (Moorhouse et al., 2004; Tisdall, 2001), but not so many on the application of Ruby (Aerts et al., 2009). However, this is changing every day because Ruby is becoming more and more popular. We couldn't find

papers in which Perl and Ruby were compared for simple and basic use in the field of bioinformatics.

It is worth mentioning that in recent years some scientists and programmers have developed libraries and tools for bioinformatics, molecular biology, genomics and life sciences, namely BioPerl (BioPerl, 2012; Stajich et al., 2002) and BioRuby (BioRuby, 2013; Goto et al., 2009). In our paper, we showed that one can perform basic bioinformatics analyses in Perl and Ruby without downloading and using these special libraries. We also compared programs written in both languages, in our opinion in favor of the Ruby language.

Conclusions

Perl and Ruby are useful tools in the field of bioinformatics. Both languages are general purpose, free to use, and popular. While Perl has a stable position in medical computer science, Ruby is still working its way into the field. It is our belief that Ruby's popularity will rise because of its simple syntax and the richness of its methods. The programs in Ruby are very easy to read and, therefore, easier to maintain, which are the most important characteristics for a programming language.

REFERENCES

Aerts, J., & Law, A. (2009). An introduction to scripting in Ruby for biologists. *BMC Bioinformatics*, 10(221), Retrieved July 30, 2013, from BioMed Central: http://www.biomedcentral.com/1471-2105/10/221. DOI: 10.1186/1471-2105-10-221.

Baxevanis, A. D., & Ouellette, B. F. F. (2004). *Bioinformatics: a practical guide to the analysis of genes and proteins.* USA: Wiley-Interscience.

BioPerl (2012). Retrieved July 30, 2013 from http://www.bioperl.org/.

BioRuby (2013). Retrieved July 30, 2013 from http://www.bioruby.org/.

Foy, B. (2007). *Mastering Perl* (2nd ed.). USA: O'Reilly Media.

Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J., & Katayama, T. (2009). BioRuby: Bioinformatics software for the Ruby programming language. *Bioinformatics*, 26(20), 2617–2619. DOI: 10.1093/bioinformatics/btq475.

Moorhouse, M., & Barry, P. (2004). *Bioinformatics, Biocomputing and Perl: an introduction to bioinformatics computing skills and practice*, USA: Wiley.

National Center for Biotechnology Information, U.S. National Library of Medicine. (1998). *Homo sapiens myosin heavy chain IIx/d mRNA, complete cds.* Retrieved July 30, 2013, from http://www.ncbi.nlm.nih.gov/nuccore/4808814?report=genbank#sequence_4808814.

National Center for Biotechnology Information, U.S. National Library of Medicine. (2001). *Homo sapiens 43kDa acetylcholine receptor-associated protein (RAPSN) mRNA, complete cds.* Retrieved July 30, 2013, from http://www.ncbi.nlm.nih.gov/nuccore/19310212?report=genbank.

National Center for Biotechnology Information, U.S. National Library of Medicine. (2006). *Campylobacter jejuni subsp. jejuni NCTC 11168 complete genome.* Retrieved July 30, 2013, from http://www.ncbi.nlm.nih.gov/nuccore/30407139?report=fasta.

National Center for Biotechnology Information, U.S. National Library of Medicine. (2009). Retrieved July 30, 2013, from http://www.ncbi.nlm.nih.gov/genbank/.

Schwartz, R., & Phoenix, T. (2011). *Learning Perl.* USA: O'Reilly Media.

Stajich, J. E., Block, D., Boulez, K., Brenner, S., Chervitz, S., Dagdigian, C., Fuellen, G., Gilbert, J. Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., & Birney, E. (2002). The BioPerl Toolkit: Perl modules for the life sciences. *Genome Research*, 12(10), 1611-1618. DOI: 10.1101/gr.361602.

Thomas, D., Fowler, Ch., & Hunt, A. (2009). *Programming Ruby 1.9: the pragmatic programmers' guide.* USA: Pragmatic Bookshelf.

Tisdall, J. (2001). *Beginning Perl for bioinformatics.* USA: O'Reilly Media.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0042

Joinpoint Regression Analysis of Potential Years of Life Lost Due to Main Causes of Death in Poland, Years 2002–2011

Michalina Krzyżak[1], Dominik Maślach[1], Martyna Skrodzka[2], Katarzyna Florczyk[2], Anna Szpak[3], Bartosz Pędziński[1], Paweł Sowa[1], Andrzej Szpak[1]

[1] Department of Public Health, Medical University of Bialystok, Poland
[2] Students Scientific Group of Public Health, Department of Public Health, Medical University of Bialystok, Poland
[3] Department of Restorative Dentistry, Medical University of Bialystok, Poland

Abstract. The purpose of the study was to analyse the level and the trends of Potential Years of Life Lost due to the main causes of death in Poland in the years 2002–2011. The material for the study was the information from the Central Statistical Office on the number of deaths due to the main causes of death in Poland in the years 2002–2011. The premature mortality analysis was conducted with the use of the PYLL (Potential Years of Life Lost) indicator. PYLL rate was calculated following the method proposed by J. Romeder, according to which premature mortality was defined as death before the age of 70. Time trends of PYLL rate and the average annual percent change (APC – Annual Percent Change) were assessed using jointpoint models and the Joinpoint Regression Program. In the years 2002–2011, PYLL rate for all-cause deaths decreased by 7.0% among men and 8.1% among women. In 2011, the main reasons for PYLL among men were: external causes (27.6%), cardiovascular diseases (24.2%) and cancers (20.3%). Among women the leading causes were: cancers (41.1%), cardiovascular diseases (19.7%) and external causes (12.5%). PYLL rate increased among men for colorectal cancer, and among women for colorectal and lung cancer. The presented epidemiological situation for premature mortality in Poland shows that in the majority of cases it is caused by preventable deaths, which highlights a need to intensify measures in primary and secondary prevention.

Introduction

Premature deaths in younger age groups influence societies in their social and economic aspects; therefore, reduction of the number of these deaths is an important aim for activities undertaken within the public health sector.

A traditional premature mortality indicator shows the rate of deaths in a population and allows for the analysis of time trends and the comparison

ISBN 978-83-7431-392-6 ISSN 0860-150X

of premature mortality in various populations but it does not include social and economic burdens imposed on a society by premature deaths. Therefore, the Potential Years of Life Lost (PYLL) rate has been introduced, which is an addition to the premature mortality analysis, as it includes the number of deaths resulting from a particular cause as well as the age at death (Gardner et al., 1990; Romeder et al., 1977).

PYLL rate is an indicator that arbitrarily assumes a limit of life expectancy, e.g. in OECD countries, it is set at the age of 70 (OECD, 2011). A death at the age of 30 is accountable for 40 potential years of life lost. Therefore, deaths in younger age groups cause greater social and economic burdens of loss because they are the reasons for more potential years of life lost. In Poland, the use of a synthetic PYLL rate in epidemiological studies is not widespread.

The purpose of the study was to analyse the level and the trends of Potential Years of Life Lost due to the main causes of death in Poland in the years 2002–2011.

Material and Methods

The material was based on the data from the Central Statistical Office of Poland on the number of deaths registered in the years 2002–2011. Causes of death were coded according to the 10^{th} revision of the International Classification of Diseases (World Health Organization, 2004).

PYLL were calculated according to the method proposed by Romeder (1977). The number of deaths in 5-year age groups was used to calculate the PYLL in Poland in years 2002–2011. The calculations were made according to the formula:

$$\text{PYLL} = \sum_{i=1}^{70} d_i \cdot (70 - i)$$

where

- 70 is the cut-off age before the occurrence of death
- i is the average number of potential years of life lost due to causes of death registered for the given age group (e.g. 42.5 years for the age group 25–29)
- d_i is the number of deaths in i age group.

The PYLL rate was calculated as a quotient of the PYLL number and the number of inhabitants in Poland in the age group 1–69. The PYLL rate was calculated per 100 000 people, separately for both sexes.

Time trends for the PYLL rate in the years 2002–2011 in Poland were analysed for general-causes of death and three main categories of death: cancer, cardiovascular diseases, and external causes. Moreover, some causes of more relevant impact on premature mortality were selected from the main causes of death. The categories of causes of death used in the analysis are shown in Table 1.

Table 1. Cause of death categories used in the analysis

Cause of death categories	Classification of diseases according to ICD 10
All causes	A02.0 – Y89.9
Cancer	C00 – C97
Colorectal cancer	C18 – C21
Lung cancer	C33 – C34
Breast cancer	C50
Cervical cancer	C53
Cardiovascular diseases	I00 – I99
Ischemic heart diseases	I20 – I25
Cerebrovascular diseases	I60 – I69
External causes	V01 – Y89
Traffic accidents	V01 – V99
Suicides	X60 – X84

The changes in PYLL rates for each cause were analysed using the joinpoint software, using a log-linear model assuming constant variance. This method is an extension of the linear regression model, in which the time trend is determined by the joined segments (joinpoints) in which changes in time trends occur in a statistically significant way by performing several permutation tests (Kim et al., 2000).

On the basis of the linear regression model, in which the natural logarithm of the PYLL rate was a dependent variable and the calendar year was an independent variable ($y = a + bx$, where $y = \ln(\text{PYLL rate})$, $x =$ calendar year), Annual Percent Change (APC) of PYLL rates for each trend was determined according to the following formula:

$$\text{APC} = 100 \cdot (\exp^b - 1).$$

A 95% confidence interval was set in order to estimate the statistical significance of the APC level in the analysed period. The trends and APC were analysed using the Joinpoint Regression Program, Version 4.0.4 (National Cancer Institute, Statistical Research and Applications Branch, 2013).

Table 2. Rate of Potential Years of Life Lost per 100 000 population and their trends in the years 2002–2011 in Poland

	Men						Women					
	Rate of PYLL			Trend			Rate of PYLL			Trend		
	2002	2011	%Change	Time	APC	95%CI	2002	2011	%Change	Time	APC	95%CI
All causes	8919.4	8301.0	–7.0	2002–2008 2008–2011	+0.6* –3.4*	(0.0;1.2) (–5.0;–1.8)	3374.4	3101.7	–8.1	2002–2008 2008–2011	+0.2 –3.0*	(–0.6;1.0) (–5.2;–0.8)
Cancers	1853.4	1682.8	–9.2	2002–2008 2008–2011	–0.5 –2.1*	(–1.1;0.1) (–3.9;–0.3)	1351.1	1275.5	–5.6	2002–2008 2008–2011	+0.4 –1.3*	(–0.6;1.5) (–2.0;–0.6)
Colorectal cancer	123.7	139.1	+12.4	2002–2011	+1.8*	(1.0;2.6)	89.1	100.8	+13.1	2002–2011	+2.5	(–16.2;25.4)
Lung cancer	584.0	502.6	–13.9	2002–2008 2008–2011	–0.8* –3.2*	(–1.4;–0.2) (–5.0;–1.3)	177.2	215.2	+21.4	2002–2011	+2.7*	(2.0;3.4)
Breast cancer	–	–	–	–	–	–	237.2	226.3	–4.6	2002–2011	–0.7*	(–1.1;–0.3)
Cervical cancer	–	–	–	–	–	–	126.3	95.8	–24.1	2002–2011	–2.7*	(–3.5;–2.0)
Cardiovascular diseases	2170.9	2006.7	–7.6	2002–2008 2008–2011	+0.5 –4.7*	(–0.1;1.0) (–8.5;–0.6)	725.3	612.1	–15.6	2002–2008 2008–2011	–0.9* –4.9	(–1.6;–0.1) (–9.9;0.3)
Ischemic heart diseases	913.9	639.7	–30.0	2002–2008 2008–2011	–2.6* –6.1*	(–3.8;–1.4) (–9.5;–2.5)	194.7	142.3	–26.9	2002–2011	–3.7*	(–4.8;–2.7)
Cerebrovascular diseases	430.8	366.1	–15.2	2002–2008 2008–2011	–0.5 –5.0*	(–1.9;1.0) (–8.9;–0.8)	252.7	174.1	–31.1	2002–2011	–3.9*	(–4.5;–3.3)
External causes	2673.1	2294.2	–14.2	2002–2011	–1.6*	(–2.0;–1.1)	517.9	387.3	–25.2	2002–2011	–3.1*	(–4.3;–1.9)
Traffic accidents	834.6	615.1	–26.3	2002–2011	–4.3*	(–5.6;–3.0)	205.1	130.9	–36.2	2002–2011	–4.9*	(–7.2;–2.5)
Suicides	736.2	720.2	–2.2	2002–2011	–0.2	(–1.7;1.4)	113.9	82.4	–27.7	2002–2011	–2.7*	(–4.3;–1.1)

%Change – decrease or increase of PYLL rate from 2002 to 2011 in percent (of observed values)

* – the APC is statistically significantly different from zero ($p < 0.05$)

Results

As presented in Table 2, in the years 2002–2008, PYLL rate for men for all causes of death increased by 0.6% per year ($p < 0.05$). After 2008, the trend changed and decreased, and the PYLL rate continued to decrease by 3.4% per year ($p < 0.05$) until 2011. Among women the trend also increased by 0.2% per year until 2008 and then continued to decrease by 3.0% per year ($p < 0.05$), as presented in Figure 1. The PYLL rate for all causes of death was almost three times higher among men than women. Among men in 2011, PYLL amounted to $8301.0/10^5$; among women, it amounted to $3101.7/10^5$.

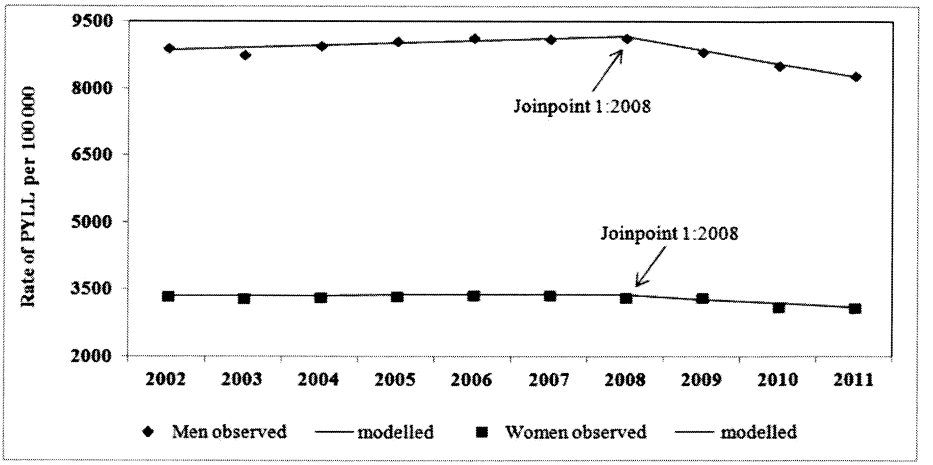


Figure 1. Rate of PYLL for all causes in Poland, years 2002–2011: observed and modeled values

Figures 2 and 3 present the proportion of PYLL due to main causes of death in the years 2002 and 2011, separately for men and women. In 2011, the main causes of premature mortality among men were external causes, responsible for 27.6% of PYLL, followed by cardiovascular diseases – 24.2% and cancer – 20.3%. These three cause of death groups were responsible for 72.1% of PYLL among men. For women, the leading cause of premature mortality was cancer – 41.1% of PYLL, then cardiovascular diseases – 19.7% and external causes – 12.5%. Similarly to men, these three cause of death groups were responsible for 73.3% of PYLL.

As presented in Table 2, in the years 2002–2008, the PYLL rate for men significantly decreased due to lung cancer (APC = 0.8%, $p < 0.05$) and ischaemic heart disease (APC = 2.6%, $p < 0.05$). After 2008, a significant de-

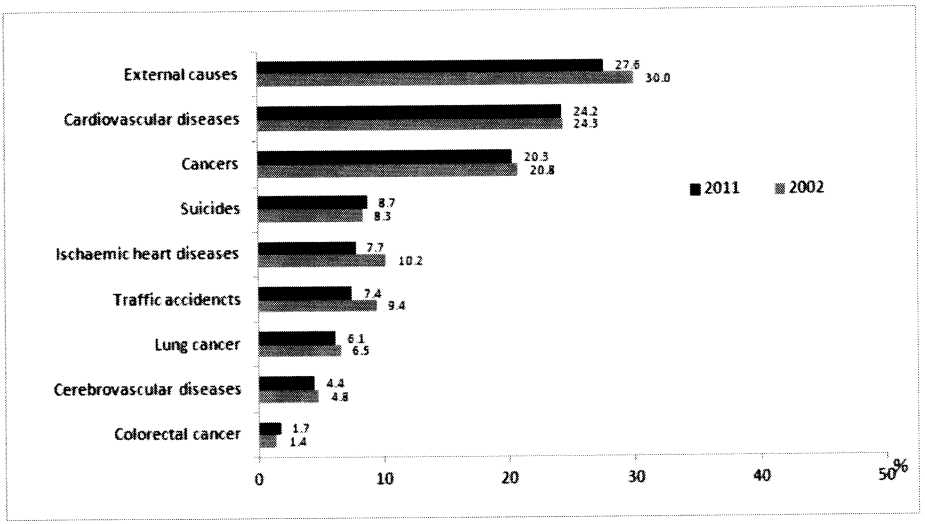


Figure 2. Potential Years of Life Lost by main causes of death among men in Poland

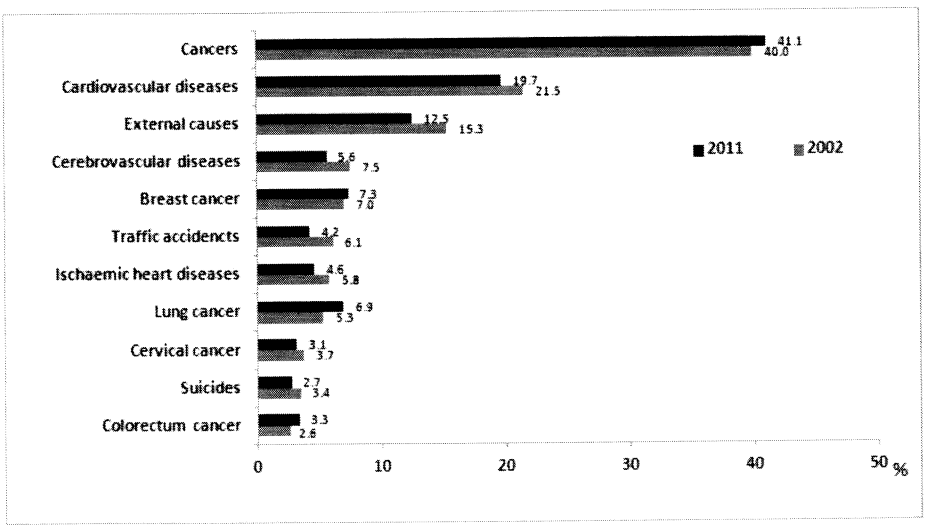


Figure 3. Potential Years of Life Lost by main causes of death among women in Poland

crease in the rate of PYLL was noted for all cancers (APC = 2.1%, $p < 0.05$), lung cancer (APC = 3.2%, $p < 0.05$), cardiovascular diseases (APC = 4.7%, $p < 0.05$), ischaemic heart disease (APC = 6.1%, $p < 0.05$) and cerebrovascular disease (APC = 5.0%, $p < 0.05$). During the analysed period,

the PYLL rate decreased among men for external causes (APC = 1.6%, $p < 0.05$). The PYLL rate also decreased for traffic accidents (APC = 4.3%, $p < 0.05$).

Among women in the years 2002–2008, the PYLL rate significantly decreased only due to cardiovascular disease (APC = 0.9%, $p < 0.05$). After 2008, a significant decrease in the PYLL rate was noted for all cancers (APC = 1.3%, $p < 0.05$). During the analysed period, the PYLL rate among women decreased for breast cancer (APC = 0.7%, $p < 0.05$), cervical cancer (APC = 2.7%, $p < 0.05$), ischaemic heart diseases (APC = 3.7%, $p < 0.05$), cerebrovascular disease (APC = 3.9%, $p < 0.05$), external causes (APC = 3.1%, $p < 0.05$), traffic accidents (APC = 4.9%, $p < 0.05$), and suicides (APC = 2.7%, $p < 0.05$). In the years 2002–2011, an increase in the PYLL rate among women was noted for lung cancer (APC = 2.7%, $p < 0.05$).

Discussion

In Poland, from the mid 1960s to the end of the 1980s, premature mortality among men increased systematically, whereas it remained on the same level among women. After 1991, a slowdown in the trend in both men and women was seen, and then a reverse change and a decrease in premature mortality were observed (Zatoński, 1996). Despite these changes in Poland from 1995–2007, the levels of premature mortality were some of the highest in the European Union (European Commission, 2010). The results of the research show that since 2008, PYLL rate has decreased slightly among men and women for all causes of death, but the tempo of these changes was among the slowest in the EU.

In years 2002–2011 in Poland, the average PYLL rate for all causes of death was three times higher for men than for women. The biggest surplus of PYLL rate between men and women was for suicides, the smallest one – cancers. Higher premature mortality rates for men than for women are common in all parts of the world (Colin et al., 2004). In all EU countries, men die earlier than women, and in 2010, an average surplus in mortality among men in comparison to women amounted to approximately 65%; in Poland it was higher and amounted to 91%. Higher mortality among men in Poland occurred in all age groups, but was the highest in persons above the age of 60 (Wojtyniak et al., 2012).

The results of the research indicate that in Poland, the structure of causes of premature mortality for men and for women was different in comparison to that in the European Union. According to the OECD data in

Europe (OECD, 2007), in 2007, the prime categories of causes for potential years of life lost before the age of 70 among men were external causes (29%), cancers (21%) and cardiovascular diseases (18%). As for women, they were cancers (31%), external causes (17%) and cardiovascular diseases (13%). In Poland in the years 2002–2011, the main causes of PYLL for men were external causes and for women cancers. Cardiovascular diseases were the second cause of premature death for both sexes.

External causes of death are potentially preventable. Traffic accidents and suicides were the leading causes of PYLL within external causes. During the study period, the PYLL rate of traffic accidents decreased in a similar way for men and women. Unfortunately, the PYLL rate for suicides among men did not change significantly and in 2011, premature mortality for this reason was higher than for traffic accidents and amounted to 8.2% of PYLL. Trends in suicides are thought to reflect changing patterns of mental health within populations, so the presented situation regarding premature mortality due to suicide among men calls for the improvement of mental health strategies to include suicide reduction targets.

Premature mortality due to cardiovascular diseases in Poland has declined since 1991 (Zatoński, 1996), but the presented study shows that the rate of PYLL for cardiovascular diseases among men has decreased since 2008. During the years of the studythe rate of PYLL decreased for ischemic heart diseases and cerebrovascular diseases among men and women. Premature mortality from cardiovascular diseases is preventable through healthier lifestyle and timely access to medical treatment.

The results of the research show that the PYLL rate due to all cancers has decreased slightly for men and women since 2008. However, unfavourable changes were prevalent for colorectal cancer in both sexes and lung cancer in women. There was a favourable and statistically significant change for lung cancer in men and for breast and cervical cancer in women.

In Poland, lung cancer is still the leading cause of death; however, the trends are different for each sex. Among men, the general mortality trend decreases slowly, whereas among women it increases rapidly. Similar changes, especially in the female population, have also been noticed in many EU countries (Tyczyński et al., 2004). In other countries, for example: France, Spain, and Italy, the trend is either stable or is decreasing (Bosetti et al., 2012).

Lung cancer is a preventable disease. The results of the research and the ones published for the country show that measures undertaken to fight tobacco smoking have a certain favourable influence among men; however, they are of a small effectiveness among women.

In the last thirty years in Poland, rates of incidence and mortality due to colorectal cancer have increased. Among men, that increase was the fastest as far as cancers are concerned (Didkowska et al., 2011). As the results of our study have shown, the trend in premature mortality was similar. It also increased among women due to colorectal cancer. Even though the trend was not statistically significant among women, it indicates the increase of the threat of this type of cancer in Poland. The risk of death due to colorectal cancer in Poland, the Czech Republic, Slovakia and Hungary belongs to the highest in Europe, and the divergence between Poland and North-Western European countries is systematically growing (Bosetti et al., 2011).

Premature deaths due to colorectal cancer are preventable through intensified primary prevention based on the modification of risk factors related to lifestyle, secondary prevention based on colonoscopy and faecal occult blood tests, and finally, the improvement of standards of diagnosis and treatment. Population screenings carried out in the 90s showed that standards of diagnosis and treatment were rarely followed (Gatta et al., 2010).

Cervical cancer is also a cancer that can be effectively prevented. Poland is a country of mid/high risk of incidence and mortality due to cervical cancer. In Europe, the risk is higher in Romania and Bulgaria. Although general mortality due to cervical cancer is decreasing in Poland, the divergence between Poland and Western European countries is not growing smaller (Arbyn et al., 2009).

The presented results show that the premature mortality trend due to cervical cancer in Poland was similar to the general mortality trend during the years of the study, and was also too slow, which calls for the intensification of prevention efforts and for the improvement of treatment standards.

In the majority of highly developed countries, deaths due to cardiovascular diseases, cancers and external reasons constitute over 70% of potential years of life lost before the age of 70 for both men and women (Savidan et al., 2010). In most cases, these deaths are preventable thanks to measures undertaken in the field of health promotion, social education, early detection of diseases, and effective treatment and rehabilitation (Gromulska et al., 2008). Therefore, knowledge about the causes of and trends in premature death in a population is essential to determining priorities in health care planning and to measuring their effectiveness. Such measures, especially the ones concerning preventive medicine at primary and secondary levels and improvement in availability of optimal treatment, as well as better socio-economic conditions of a population, may contribute to the decrease of potential years of life lost.

Conclusions

Among men after 2008, for all analysed causes of death except for colorectal cancer, favourable changes were noted. Among women, the PYLL rate systematically increased for colorectal cancer and lung cancer. After 2008, favourable changes were noted for all-cause deaths and cancers.

The presented epidemiological situation for premature mortality in Poland shows that in the majority of cases it is caused by preventable deaths, which suggests a need to intensify measures in primary and secondary prevention.

REFERENCES

Arbyn, M., Raifu, A. O., Weiderpass, E., Bray, F., & Anttila, A. (2009). Trends of cervical cancer mortality in the member states. *Eur J Cancer*, 45(15), 2640–2648.

Bosetti, C., Levi, F., Rosato, V., Bertuccio, P., Lucchini, F., Negri, E., & La Vecchia, C. (2011). Recent trends in colorectal cancer mortality in Europe. *Int J Cancer*, 129(1), 180–191.

Bosetti, C., Malvezzi, M., Rosso, T., Bertuccio, P., Gallus, S., Chatenoud, L., Levi, F., Negri, E., & La Vecchia, C. (2012). Lung cancer mortality in European women: trends and predictions. *Lung Cancer*, 78, 171–178.

Colin, M., Thies, B., & Doris, M. F. (2004). *Global burden of Disease: update 2004.* World Health Organisation.

Didkowska, J., Wojciechowska, U., & Zatoński, W. (2011). *Nowotwory złośliwe w Polsce w 2009 roku.* Warszawa: Centrum Onkologii, Instytut im. M. Skłodowskiej-Curie.

European Commission. (2010). *Health trends in the UE.*

Gardner, J., & Sanborn, J. (1990). Years of potential life lost (YPLL) – what does it measure? *Epidemiology*, 1(4), 322–329.

Gatta, G., Zigon, G., Aareleid, T., Ardanaz, E., Bielska-Lasota, M., Galceran, J., Góźdź, S., Hakulinen, T., Martinez-Garcia, C., Plesko, I., Zakelj, M. P., Rachtan, J., Tagliabue, G., Vercelli, M., & Faivre, J. (2010). Patterns of care for European colorectal cancer patients diagnosed 1996–1998: a EUROCARE High Resolution Study. *Acta Oncol.*, 49(6), 776–783.

Gromulska, L., Wysocki, M. J., & Goryński, P. (2008). Lata przeżyte w zdrowiu (Healthy Life Years, HYL) – zalecany przez Unię Europejską syntetyczny wskaźnik sytuacji zdrowotnej ludności. *Przegl Epidemiol*, 62, 811–820.

Kim, H. J., Fay, M. P., Feuer, E. J., & Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Stat Med*, 19(3), 335–351. (correction: 2001; 20: 655)

National Cancer Institute, Statistical Research and Applications Branch. (2013). Joinpoint Regression Program (Version 4.0.4.).

OECD (2007). *Health at a Glance 2007: OECD Indicators.* OECD Publishing. DOI: 10.1787/health_glance-2007-en.

OECD (2011). *Health at a Glance 2011: OECD Indicators.* OECD Publishing. DOI: 10.1787/health_glance-2011-en.

Romeder, J., & McWhinnie, J. (1977). Potential Years of Life Lost Between Ages 1 and 70: An Indicator of Premature Mortality for Health Planning. *Int J Epidemiol*, 6(2), 143–151.

Savidan, A., Junker, Ch., Cerny, T., & Ess, S. (2010). Premature mortality in Switzerland from 1995–2006: causes and trends. *Swiss Med Wkly.*, 140:w13077. DOI: 10.4414/smw.2010.13077.

Tyczyński, J. E., Bray, F., Aareleid, T., Dalmas, M., Kurtinaitis, J., Plesko, I., Pompe-Kirn, V., Stengrevics, A., & Parkin, D. M. (2004). Lung cancer mortality patterns in selected Central, Eastern and Southern European countries. *Int J Cancer*, 109(4), 598–610.

Wojtyniak, B., Goryński, P., & Moskalewicz, B. (2012). *Sytuacja zdrowotna ludności Polski i jej uwarunkowania* (pp. 58–59). Warszawa: Narodowy Instytut Zdrowia Publicznego.

World Health Organization. (2004). *International Statistical Classification of Diseases and Health Related Problems: tenth revision* (2nd ed.). Geneva: WHO.

Zatoński, W. (1996). *Evaluation of health in Poland since 1988.* Warszawa: Centrum Onkologii, Instytut im. M. Skłodowskiej-Curie.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0036

Hospital Statistics as a Tool for Obtaining Data Necessary in the Healthcare Entity Management Process

Aleksandra Sierocka[1], Bożena Woźniak[1], Petre Iltchev[2], Michał Marczak[2]

[1] K. Jonscher Hospital in Lodz, Poland

[2] Health Care Policy Department, Medical University in Lodz, Poland

Abstract. Statistical methods used by healthcare entities enable the collection of various information about the structure and characteristics of treated patients. They are an important source of knowledge, and form a database that plays an important role in entity management theory. In the presented study, we analysed the hospital stays of patients treated in all hospital wards of the 3rd City Hospital in Łódź during 2012. The following, in particular, were taken into account: admittance procedure, discharge procedure, age and sex of hospitalised persons. Patients in over 55% of cases were admitted using the sudden admittance procedure. At the same time, over 3/4 of the stays ended with a referral for further treatment in ambulatory conditions, and death occurred in approx. 5% of hospitalisations. By comparing the discharge procedures, the percentage of deaths in the Anaesthesiology and Intensive Care Wards can be seen clearly (more than 70%). Internal wards are next in turn (10.6 and 16.6%). The biggest differences in the length of hospitalisation between the studied institution and the NFZ data (which are averaged values from all medical entities in Poland) concern the E77, A49, A48, A87, A33, D18, E16, E61 and G37 groups.

Introduction

Statistical methods used by healthcare entities enable the collection of various information about the structure and characteristics of treated patients. They are an important source of knowledge, and form a database that plays an important role in entity management theory. Unfortunately, in most cases, due to a permanent lack of time and an accumulation of current affairs, directors who manage hospitals, clinics or laboratories omit the aforementioned methods in daily operation and when making key decisions. As a consequence, they encounter significant problems in the effective analysis of various aspects of the operation of their institutions (Zieliński, 2013).

It should be mentioned that the awareness and understanding of events occurring in the managed healthcare entity enables the optimal manage-

ISBN 978-83-7431-392-6 ISSN 0860-150X

ment of the given entity, adequate to their capabilities and resources, not only at the given moment, but also in the future due the ability to create forecasts for the future. In the situation where, for the treatment of every patient a specific amount is received (in accordance with the principles established by the NFZ National Health Fund), the issue of correct settlement of services and the possibility of indicating specific reasons, for which one hospitalisation results in losses and another in profits, is the essence of a well-operating entity. Simple statistical analyses provide the answers to basic questions and indicate the reasons for many difficulties which the institutions in question face. They also prove the existence of phenomena, which we may not have been aware of, and act as a bargaining chip in negotiations with supervisory, controlling or founding bodies (Sanak et al., 2010).

As a result of the increasing scope and detail of the data collected by medical entities, we are able to generate reports that a few years ago were out of the realm of possibility. Based on conducted hospital statistics we are able to locate the information that is important from the point of view of individual wards, departments or the management itself. Thus, the variety of conducted analyses is very large. The most known and popular list is a semi-annual and annual analysis of hospital activity, containing data about the number of beds, person-days, the number of treated persons, deaths, and mortality and bed occupancy indicator. Systematic analyses are becoming more and more important, e.g.: reasons of patient hospitalisation in individual hospital wards (Kaczor et al., 2011; Rudnicka-Drożak et al., 2010), deaths (Karwat, 2012; Wróblewska et al., 2008) or the average hospitalisation time.

Material and Methods

All the data concerning the method of treatment of patients admitted to hospital are entered in the institution's IT system. What analyses we have at our disposal and how advanced the lists and comparisons we are able to obtain are dependent on the complexity of a given system and the possibility of extracting the necessary information. In this study, we have analysed the hospital stays of the patients treated at all hospital wards of the 3rd City Hospital in Lodz during 2012. In particular, the following were taken into account: admittance procedure, discharge procedure, and age and sex of hospitalised persons. Pursuant to the settlement data the DRG groups most frequently occurring in the year 2012 were also indicated, and the obtained results were compared with the information provided on the DRG statistics, 2013. For calculations, Excel spreadsheet was used along with

basic statistical functions, showing that such types of lists do not require complicated software and sophisticated knowledge.

Results

The obtained results are presented in Tables 1–3 and Figure 1. Tables 1–3 enable the number of hospitalisations to be established in accordance with procedures of admittance, discharge, sex and age (for the entire hospital) as well as the characterization of specific organisational units, which is more important from the point of view of managers/heads of these organisational cells.

Table 1. Number of hospitalisations in the year 2012 and the percentage of stays divided into procedure of admission and sex of patients

Ward	Number of finished hospitalisations in the year 2012	Admittance procedure						Sex	
		without referral	transfer from another hospital	planned with referral	planned with rights	sudden	sudden, through an EMS team	Woman	Man
Anaesthesiology and Intensive Care	148	81	3	0	0	13	51	85	63
General surgery	1783	48	15	403	1	761	555	832	951
General surgery Przyrodnicza street	110	0	1	20	0	41	48	51	59
Trauma and Orthopaedics Surgery	1647	13	5	888	1	546	194	941	706
Internal Diseases and Cardiology	2048	161	10	190	687	999	1	1212	836
Operative and Conservative Gynaecology	341	0	1	274	5	44	17	341	0
Neurology	1121	111	8	170	2	473	357	622	499
Ophthalmology	1833	2	0	1567	1	224	39	1095	738
Rehabilitation	166	1	0	165	0	0	0	97	69
Neurologic Rehabilitation	325	2	0	323	0	0	0	171	154
Stroke	421	58	1	2	0	116	244	255	166
Internal A	1239	45	2	53	0	383	756	742	497
Hospital Emergency Room	766	0	1	0	0	284	481	482	284
Total	11948	522	47	4055	697	3884	2743	6926	5022
%	100	4.37	0.39	33.94	5.83	32.51	22.96	57.97	42.03

Table 2. Number of hospitalisations divided by discharge procedure, presented according to individual hospital wards

Ward	Discharge procedure										
	ending the therapeutic and diagnostic process	continuing in ambulatory treatment	referral to a psychiatric hospital	continuing treatment – others	continuing in another inpatient facility	continuing in a long term care facility	discharge at own request	death	continuing in the same facility	leaving against medical advice	transfer to another ward
Anaesthesiology and Intensive Care	2	0	0	0	13	0	0	104	0	0	29
General surgery	5	1547	7	0	15	4	63	75	3	2	62
General surgery Przyrodnicza street	14	82	0	0	1	0	6	5	0	0	2
Trauma and Orthopaedics Surgery	0	1578	1	1	18	3	9	3	15	0	19
Internal Diseases and Cardiology	2	1610	4	0	109	21	17	216	16	0	53
Operative and Conservative Gynaecology	331	4	0	0	4	0	0	0	2	0	0
Neurology	489	443	4	2	35	4	45	17	9	4	69
Ophthalmology	5	1814	1	3	1	0	3	0	0	0	6
Rehabilitation	2	156	0	0	0	0	3	0	2	0	3
Neurologic Rehabilitation	4	292	2	0	3	1	8	0	3	0	12
Stroke	129	128	1	0	11	11	10	30	11	0	90
Internal A	0	871	6	1	36	44	28	205	3	0	45
Hospital Emergency Room	1	547	0	0	45	3	9	22	4	1	134
Total	984	9072	26	7	291	91	201	677	68	7	524
%	8.24	75.93	0.22	0.06	2.44	0.76	1.68	5.67	0.57	0.06	4.39

Table 3. Number of hospitalisations by patients' age

All hospital wards	Patients' age (in years)								
	18–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	> 91
Total	97	546	691	826	2012	2357	2859	2339	221
%	0.81	4.57	5.78	6.91	16.84	19.73	23.93	19.58	1.85

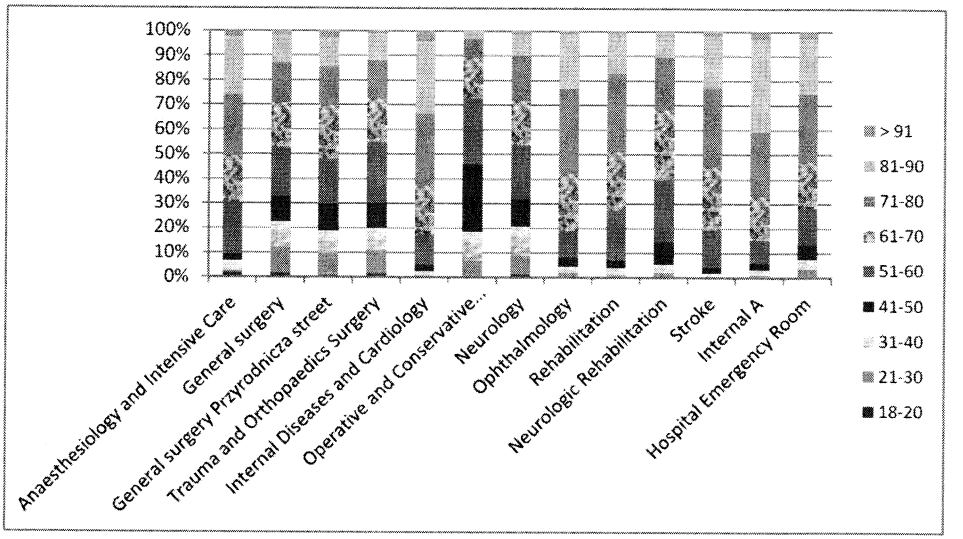


Figure 1. Percentage of hospitalisations by wards, divided by patient's age

Based on the presented data, it is clear that patients in over 55% of cases were admitted using the sudden admittance procedure (Table 1). This is a significant difficulty for the hospital, since this type of admittance has a much higher risk than planned admittance (which makes up only 1/3 of cases). They require higher financial outlays (i.e., necessary diagnostics and laboratory testing, medical advice, longer hospitalisation time, required transfers between wards etc.) and are highly unpredictable (e.g. due to bad health, co-existing diseases or the patient's age). At the same time, over 3/4 of the stays end with a referral for further treatment in ambulatory conditions, and death occurs in approx. 5% of hospitalisations (Table 2).

The age of treated patients also has an enormous impact on the hospital's financial condition (the older the person, the more co-existing diseases, the worse their health and the longer the duration of hospitalisation). The results obtained by the study confirm the generally known fact that the society is getting older, and the demand for geriatric wards will be increasing (the group of patients older than 91 years was almost 2% of the total) (Table 3).

At the same time, the analysis conducted using the criterion of division by wards shows that some of the institution's wards accept most of their patients in a planned manner (e.g. rehabilitation, ophthalmology, gynaecology or orthopaedic) (Table 1). By comparing the discharge procedures, the percentage of deaths in the Anaesthesiology and Intensive Care Wards is

clearly visible (more than 70%) (Wróblewska et al., 2008). Internal wards are next in turn (10.6 and 16.6%) (Kaczor et al., 2011). This type of situation results from the fact that the specifics of each ward vary, in addition to differences in the character of medical services (procedures) provided and the age of hospitalised persons (Figure 1).

The subsequent stage of the analysis was the attempt to indicate DRG groups with the highest significance for the studied entity. Based on the settlement data reported to the NFZ for the year 2012, the most frequent groups were nominated, their percentage share in the total was established, and the median and modal values and average of hospitalisation time were calculated. At the same time, the obtained results were compared with the data provided by DRG statistics, 2013. The complex analysis of the DRG system prepared by the NFZ concerns the services provided within the years 2009–2012 throughout the entire country of Poland. This list includes the number of provided DRG products and the average patient stay time divided by age, sex, admittance and discharge procedure, ICD-10 basic diagnosis and ICD-9 medical procedure for all DRG groups, provided in the years 2009–2012. We will be interested only in the year 2012.

Due to a high diversity of DRG groups listed by the hospital in the year 2012, this list refers only to the first 20 items. The results obtained from the studied institution and data from the NFZ system are presented in Table 4.

Analysing the data presented above, it can be clearly seen that the biggest differences in the length of hospitalisation between the studied institution and the NFZ data (which are averaged values from all medical entities in Poland) concern the E77, A49, A48, A87, A33, D18, E16, E61 and G37 groups. For these services, the length of stay of the patient in the studied hospital is approx. 2 days longer than in other institutions. This suggests it may be worth considering why other hospitals are able to treat the patients in a shorter time, and thus to bear lower expenses (for food, drugs, materials, stay etc.).

A significant source of knowledge on hospitalisations is the information on the average time of hospital treatment (separately for each DRG group), listed along with the hospitalisations above a set number of days, which are financed as a group by the NFZ (in accordance with the catalogue of DRG groups). All stays of this extended type should be analysed for the reasons the patient remained in the hospital, and to indicate possible actions that could improve the proceedings in similar cases in the future (complications, adverse event, lack of access to diagnostics, lack of cooperation in transferring of patients to other institutions, e.g. due to lack of space or logistics difficulties) (Podlaski Urząd Wojewódzki w Białymstoku, 2012).

Table 4. A list of DRG groups most frequently occurring in the studied hospital with the NFZ data (DRG statistics, 2013) for the year 2012

DRG group		Number of occurrences	% of occurrence among all DRG groups	Average hospitalisation time (days)	Modal value (days)	Median value (days)	Data from DRG statistics, 2013	
							Modal value (days)	Median value (days)
B13	Uncomplicated cataract surgery by emulsification with simultaneous lens implantation	1024	10.27	2.61	2	2	0	2
E77	Other cardiovascular diseases > 17 years of age	569	5.71	10.94	8	10	7	6
A76	Head trauma with significant brain damage, treated conservatively	268	2.69	4.54	2	2	2	3
B12	Complicated cataract surgery by emulsification with simultaneous lens implantation	256	2.57	2.96	2	2	2	2
H83	Average procedures on soft tissue	251	2.52	3.53	1	2	0	2
A49	Brain stroke – treatment > 3 days	243	2.44	12.95	8	11	7	8
A48	Complex treatment of brain stroke > 7 days on the stroke ward	227	2.28	15.66	10	12	8	10
H62	Breaks or dislocations of the pelvis or lower limb	216	2.17	9.65	4	10	4	8
H33	Average procedures on lower limb	178	1.79	3.19	2	2.5	2	2
A87	Other nervous system diseases	170	1.71	10.39	7	8	2	5
A33	Balance disorders	169	1.70	7.05	6	6	3	4
D18	Atypical viral pneumonia	166	1.67	13.51	13	12	7	8
B98	Conservative ophthalmological treatment	163	1.64	6.52	1	5	2	3
G34	Endoscopic and percutaneous procedures of bile ducts and pancreas	159	1.60	4.50	2	4	1	4
E61	Heart rhythm disorders > 69 years of age or with complications	155	1.56	7.57	2	7	2	4
G37	Acute pancreatitis	154	1.55	6.47	5	5	6	7
H64	Smaller breaks or dislocations	144	1.44	2.87	1	2	1	1
E16	Acute endomyocarditis > 69 years of age or with complications	140	1.40	10.98	7	10	4	5
F46	Abdominal diseases	139	1.39	5.53	2	4	2	3
G25	Cholecystectomy	127	1.27	3.52	2	2	3	3

Conclusions

In summary, we can see how important statistics are, when correctly and systematically kept at medical institutions. The simplest lists and analyses frequently allow the areas that for some reasons generate additional costs to be pinpointed. The knowledge of the entire facility and of the specifics of individual wards provided by prepared reports is an indescribable source of information in an institution's management process. It is the initial point for making key decisions in order to ensure the correct operation of the entire unit, and it enables immediate reaction in crisis situations (e.g. failure of medical equipment, loss of skilled medical personnel).

A hospital, as a whole, will not be able (in the long run) to keep afloat in such difficult times on the medical services market without a thriving department of statistics or analytics. Knowledge from the scope of provided DRG groups and their participation in the costs of the entire contract for hospital treatment enables monthly limits to be planned and problems with both exceeding and not reaching the financial plan to be avoided.

Information for heads of wards/managers of individual departments which enables the characterisation of hospitalised patients, taking into account the age, sex, procedure of admittance and discharge, and the type of medical procedures conducted during the stay, is especially important to enabling the correct organisation of work of medical personnel, to meeting the demand for materials and medicine, or to ensuring free beds for emergency patients.

A series of articles published in journals (Roszkowska et al., 2002) which contain various types of hospital analyses allow for the establishment of what information is of interest to the decision-makers in other hospitals, as well as the possibility to compare the results of the given entity with the data from other institutions (Narodowy Fundusz Zdrowia, Departament Świadczeń Opieki Zdrowotnej, 2012).

REFERENCES

DRG statistics. (2013). Retrieved from http://prog.nfz.gov.pl/app-jgp/.

Kaczor, I., Lolo, A., Pakieła, O., Wójcik, D., Zajbt, M., & Wełnicki, M. (2011). The most frequent causes of hospitalizations of patients over 90 years old in the internal diseases ward. *Gerontologia Polska*, 19(3–4), 146–149.

Karwat, K. (2012). Analiza przyczyn zgonów w klinice pneumonologii. *Polski Merkuriusz Lekarski*, XXXII(190), 221–224.

Narodowy Fundusz Zdrowia, Departament Świadczeń Opieki Zdrowotnej. (2012). *Analiza świadczeń zdrowotnych w latach 2009–2011 w rodzaju: lecznictwo szpitalne.* Retrieved from http://www.nfz.gov.pl/new/art/5059/2012_08_24_analiza_SZP.pdf.

Podlaski Urząd Wojewódzki w Białymstoku. (2012). *Analiza działalności leczniczej w rodzaju stacjonarne i całodobowe – szpitalne świadczenia zdrowotne w województwie podlaskim w 2011 roku.* Retrieved from http://www.bialystok.uw.gov.pl/NR/rdonlyres/9347A161-8C36-4BE4-AC48-0758FA55C10F/24124/Analizadzia%C5%82alno%C5%9Bciecznejwrodzajustacjonarneic.pdf.

Roszkowska, H., Chańska, M., & Seroka, W. (2002). Hospital morbidity in mazovian Voivodeship in the first year of the new health care system. *Przegląd Epidemiologiczny*, 56, 151–158.

Rudnicka-Drożak, E., Rybojad, B., Jaworska, I., Korecka, R., & Aftyka, A. (2010). Analysis of the causes of hospitalization in the Intensive Care Unit of the Regional Specialist Hospital in Lublin. *Polish Journal of Public Health*, 120(1), 29–32.

Sanak, U., Młynarczyk, A., & Karbarz, K. (2010). *Raport Lecznictwo szpitalne w Małopolsce 2009.* Kraków: Departament Zdrowia i Polityki Społecznej UMWM.

Wróblewska, K., Jaszewski, M., Chilińska, J., Krajewska-Kułak, E., Jankowiak, B., Kowalewska, B., & Gołębiewska, A. (2008). The analysis of the most frequent causes of death of patients in the Cardinal Stefan Wyszynski Provincial Hospital in Łomża – preliminary study. *Problemy Higieny i Epidemiologii*, 89(1), 136–141.

Zieliński, T. (2013). Statystyka w zarządzaniu placówkami medycznymi, Retrieved from http://www.statsoft.pl/czytelnia/medyczne/wzarzadzaniu.html.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0035

Information and Communication Technologies in Primary Healthcare – Barriers and Facilitators in the Implementation Process

Bartosz Pędziński[1,2], Paweł Sowa[1], Waldemar Pędziński[2,3], Michalina Krzyżak[1], Dominik Maślach[1], Andrzej Szpak[1]

[1] Department of Public Health, Medical University of Bialystok, Poland
[2] Lomza Medical Center Ltd., Poland
[3] Department of Clinical Nursing, Higher School of Agrobusiness, Lomza, Poland

Abstract. Despite the great expansion and many benefits of information and communication technologies (ICT) in healthcare, the attitudes of Polish general practitioners (GPs) to e-health have not been explored. The aim of this study was to determine the GPs' perception of ICT use in healthcare and to identify barriers to the adoption of EMR (Electronic Medical Records) in the Podlaskie Voivodeship. Online and telephone surveys were conducted between April and May 2013. Responses from 103 GP practices, 43% of all practices in the region, were analysed. The results showed that 67% of the respondents agreed that IT systems improve quality of healthcare services. In the GP group who declared at least partial EMR implementation, 71.4% see the positive impact of IT on practice staff processes and 66.1% on personal working processes. In this group, more than three-quarters of GPs did not see any positive impact of ICT on the average number of patients treated per day, number of patients within the practice or scope of services. The four most common barriers to EMR implementation were: lack of funds, risk of a malfunction in the system, resistance to change, and lack of training and proper information. Although the use of ICT by Polish GPs is limited, their attitude to e-health is generally positive or neutral and resembles the overall pattern in Europe. Barriers identified by GPs need to be taken into account to ensure the effective implementation of e-health across the country.

Introduction and Objectives

Implementation of new Information and Communications Technology (ICT) solutions in the healthcare systems of the most developed countries has allowed for the development of e-health tools: Electronic Medical Records (EMR), Electronic Health Record (EHR), e-prescription and e-referrals. They improve quality of healthcare and organization of work, and increase cost-effectiveness. A systematic review of 257 studies on the

ISBN 978-83-7431-392-6 ISSN 0860-150X

impact of ICT solutions in outpatient and inpatient settings revealed that quality of healthcare improved because of an increase in guideline-driven care, enhanced surveillance and monitoring and decreased medication errors (Chaudhry et al., 2006). A study of 119 ambulatory healthcare units confirmed that Health Information Technology (HIT) improves clinical outcomes, increases the use of vaccinations and improves medication adherence. Moreover, it has led to cost savings for physicians, improved staff productivity and enriched patient-provider interactions (Police et al., 2010).

On the other hand, some publications indicate that HIT has not always led to a better quality of care (Linder et al., 2007; Romano et al., 2011). It has often prolonged working time (Poissant et al., 2005) and has disrupted workflow causing temporary declines in productivity (Menachemi et al., 2011). Implementation of ICT may be financially beneficial in the long term, but in the early years there are enormous costs of the IT software as well as training and management of the system (Hillestad et al., 2005). A cost-benefit analysis of replacing paper medical records with electronic medical records in a primary care clinic in the US showed that the estimated net benefit for a 5-year period was $86 400 while the total 5-year cost equalled $46 400 per provider (Wang et al., 2003). Negative phenomena related to the digitalisation of healthcare are particularly prominent in the early stages of IT system implementation, with its low level of functionality and its mismatch to the needs of users (Zakaria et al., 2010).

The processes of adaptation of IT solutions in the healthcare sector began in the 1970s, but only a few countries currently use EMR (Protti et al., 2010; Schoen et al., 2012). In Poland, patient data are stored electronically in only 15% of GP practices (Pedzinski et al., 2013) and 8% of hospitals (Najwyższa Izba Kontroli, 2013). Beginning on 1/08/2014, it will be a legal requirement for all Polish healthcare providers to use electronic medical records only (Ustawa z dnia 28 kwietnia 2011 o systemie informacji w ochronie zdrowia, 2011). In a very short time, widespread implementation of e-prescriptions, e-referrals and personal health records (PHR) is planned. Successful adaptation of e-health systems and the implementation of EMR will be highly dependent on the medical staff's attitudes to this process and their perceptions of barriers and facilitators related to it. GPs coordinate the patient in a healthcare system and will determine the success of new solutions such as e-prescription, e-referral for diagnostic tests and e-referral to a specialist. They will be the key element, as they are the first and most frequent point of contact with the patient (Pike, 2010). Referring the patient to the various levels of the healthcare system and providing continuous and long-term care is strongly associated with the information

about the patient that GPs will receive and generate (Kahn, 2004; Peterson, 2012). Family doctors play a key role in the adaptation of EHR and Personal Health Record (PHR) while they are responsible for the majority of data entry. The digitalisation of the health sector in the vast majority of countries started from the primary care sector (Lockhart, 2008). Therefore, it is important to understand the barriers and opportunities for implementation of HIT as perceived by primary care physicians.

Material and Methods

The study was conducted between April and May 2013. Online and telephone surveys were aimed at 237 primary healthcare providers contracted by the National Health Fund in the Podlaskie Voivodeship. In practices with more than one doctor, the one whose surname came first in the alphabet was interviewed. After exclusion of one survey with incomplete data, 103 questionnaires were included in the analysis. They represented 43% of primary healthcare clinics in the voivodeship.

Questions concerning GP attitudes towards ICT were obtained from the Dobrev et al. (2008) study. The questions on barriers to EMR adoption were based on the barriers most often listed in the Polish scientific and gray literature. The questions concerning GP attitudes towards ICT and perception of impact of ICT use were multichotomous (one of three possible answers). Positive impacts of ICT on healthcare were defined as: decreases in the workload of support staff, increases in the scope of services, increases in the average number of patients treated per day, increases in the number of patients within the practice (data presented in Table 1). The questions regarding barriers to EMR implementation included 8 potential barriers and the respondent could choose up to 4. A survey questionnaire for both telephone and online interviews was designed with *Google Forms*. *SPSS Statistics® 20.0* was used for statistical analysis. The chi-square test was used to evaluate differences in perceived barriers among EMR adopters and a non-adopters group with the significance level $\alpha = 0.05$.

Results

In the studied primary healthcare units, from 1 to 11 doctors worked under different forms of employment, while the percentage of institutions with one doctor accounted for 35%, two doctors – 28.2%, and three doctors

Table 1. GPs' attitudes towards ICT use in healthcare in the Podlaskie Voivodeship

Questions	Agree	Don't know	Disagree
the use of software and IT systems improves the quality of healthcare services	67.0%	18.4%	14.6%
the use of software and IT systems in healthcare should be included in the medical education	87.4%	8.7%	3.9%
to really benefit from IT, all health actors have to share clinical information in a network	72.8%	16.5%	10.7%
IT systems would be more used if GPs were provided with more training	72.8%	14.6%	12.6%
my practice would need better support with the maintenance of IT systems	81.6%	8.7%	9.7%
the cost of IT is ultimately the decisive factor on the use of ICT	66.0%	15.5%	18.5%
the use of telemonitoring will in the future allow physicians to treat people with chronic conditions better	67.0%	25.2%	7.8%

– 12.6%. There were 36.9% rural and 63.1% urban practices. The ages of the GPs ranged from 29–71 years (mean 51.1 ± 9.1) and 64.1% were female.

Two thirds of general practitioners (67.0%) agreed that software and IT systems improve the quality of healthcare services. The vast majority of interviewees agreed that the use of software and IT systems in healthcare should be included in the medical education system (87.4%) and that their unit would need better support with the maintenance of an IT system (81.6%). A similar percentage of physicians saw the need to establish a network to share clinical information and training for a wider use of IT systems (72.8% accordingly). 66.0% of physicians agreed that the cost of IT was ultimately the decisive factor for the use of ICT and 67.0% believed that telemonitoring would allow for better care for chronic patients in the future (Table 1).

GPs who declared either partial or full implementation of an EMR system had a positive or neutral attitude to ICT use in healthcare (Table 2). Most GPs recognised the positive impact of ICT on the working processes of practice staff (66.1%) and on a personal level (71.4%). More than three-quarters of GPs did not report any impact on the average number of patients

Table 2. GPs' perceptions of impacts of ICT on healthcare in user groups in the Podlaskie Voivodeship

Impacts of ICT use	positive	none	negative
on personal working processes	71.4%	21.4%	7.1%
on practice staff working processes	66.1%	19.6%	14.3%
on quality of diagnosis and treatment decisions	44.6%	53.6%	1.8%
on workload of support staff	41.1%	25.0%	33.9%
on scope of services	7.1%	92.9%	0.0%
on doctor–patient relationship	25.0%	42.9%	32.1%
on average number of patients treated per day	8.9%	76.8%	14.3%
on number of patients in practice	5.4%	82.1%	12.5%

treated per day (76.8%), number of patients within the practice (82.8%) or scope of services (92.9%). A negative attitude was presented by only one-third of physicians: 32.1% believed that ICT had a negative impact on the patient-doctor relationship and 33.9% – that it had a negative impact on the support staff's workload. 44.6% stated that ICT had a positive impact and 53.6% stated that there was no impact on the quality of diagnosis and treatment.

The respondents were asked to choose a maximum of 4 from a list of 8 potential barriers to implementation of an EMR system. 57.3% identified lack of funds, 48.5% concern of a malfunction in the system, 38.8% resistance to change and 38.2% lack of training and proper information. Other barriers included: privacy and security issues (33.0%), lack of time for system implementation (30.1%), negative impact on doctor–patient interaction (19.4%) and difficulties in finding the right software (11.7%). In the studied group of 103 GPs, there were 57 (55.3%) GPs who declared partial or full implementation of an EMR system (adopters group) and 46 (44.7%) GPs who declared no EMR implementation (non-adopters group). Perceived barriers to implementation of EMR in both groups are presented in figure 1. A significant difference between groups ($p = 0.025$) was observed only for the concern of malfunction in the system. The general attitude to EMR system implementation was analysed only in GPs who declared partial or complete implementation. In this group, there were 27 satisfied, 22 rather satisfied, 3 rather dissatisfied and 4 dissatisfied with the EMR system.

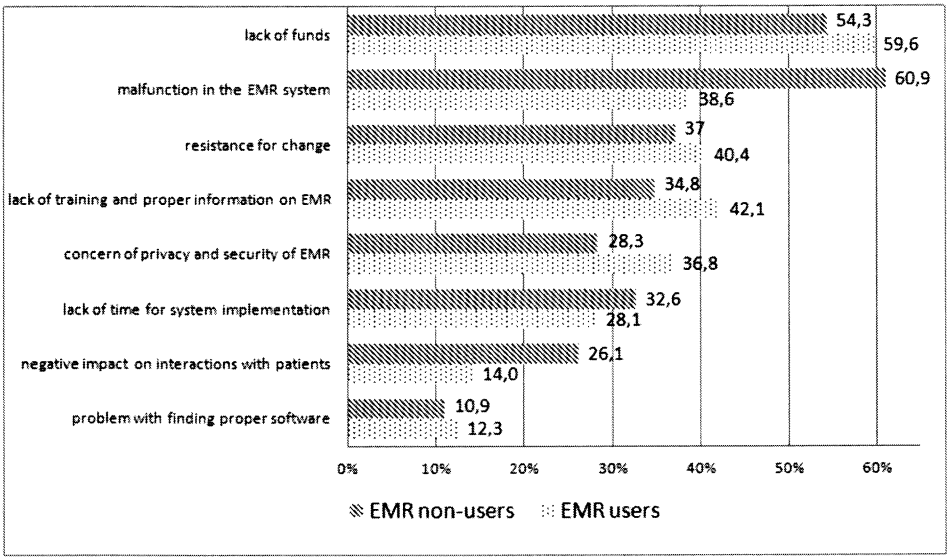


Figure 1. Barriers to EMR system implementation in user and non-user groups

Discussion

Despite a low level of HIT adaptation in Poland, 67% of surveyed GPs believed that the use of software and IT systems improves the quality of healthcare services. This result was similar to the European survey on Benchmarking ICT use in 2008 (Dobrev et al., 2008), which showed that, regardless of the degree of implementation of ICT in the country, most physicians see opportunities to use IT systems to improve the quality of services. In some studies where the attitudes towards ICT use were analysed between EHR adopters and non-adopters, the results have shown that physicians who at least partially implemented the system were more convinced of the positive effects than those who had never worked with it (Jha, DesRoches et al., 2009; Leung et al., 2003; Morin et al., 2005; Sequist et al., 2007).

In the 2008 Benchmarking ICT study (Dobrev et al., 2008; European Commission. Information Society and Media Directorate General, 2007), the most important facilitating factors in Europe and in Poland were as follows: the need for e-health inclusion in medical education, the need for more IT training and a better networking of all healthcare in order to share clinical information. These results are in line with what was found in the Podlaskie Voivodeship. However, when it comes to the potential barriers one significant difference can be noted. In Poland in 2008, cost was a decisive

factor concerning ICT use and was seen as more important than lack of ICT maintenance support, while in our study these relationships were opposite. This observation could be partly explained due to the fact that five years had passed. In the 5-year period between the observations in question, some GPs implemented the EMR systems, so they must have overcome the financial barrier. Nevertheless, cost is still the crucial barrier to ICT use. Generally, 66% of GPs declared that the cost of IT is ultimately the decisive factor on the use of ICT and the highest proportion of GPs (57.3%) identified lack of funds as one of eight potential barriers to EMR system implementation.

When it comes to GPs' perceptions of the impacts of ICT use, the results found in the Podlaskie Voivodeship were strongly in line with the general pattern in Poland and other European countries in 2008 (Dobrev et al., 2008). The GPs who declared partial or entire implementation of EMR were most positive that ICT use improved personal working processes and practice staff working processes. Most of the GPs did not see any positive impact on the workload of support staff, quality of diagnosis and treatment decisions, scope of services offered, doctor–patient relationship, the average number of patients treated per day, or the number of patients in the practice.

The four most common barriers to EMR implementation in the Podlaskie Voivodeship were: lack of funds, risk of a malfunction in the system, resistance to change and lack of training and proper information. The lack of funds as a major barrier to adoption of EMR has been shown in several other studies, particularly in the US (DesRoches et al., 2013; Gans et al., 2005; Jha, DesRoches et al., 2009; Miller et al., 2004). The risk of a malfunction in the system, defined as: slow system speed, system downtime and inadequate ICT resources, were reported by Georgiou et al. (2009), Hier et al. (2005) and Kossman et al. (2008). Resistance to change very often results from lack of training and proper information about EMR; therefore, these factors are strongly connected with a lack of understanding of potential benefits (Hackl et al., 2009; Loomis et al., 2002; Saleem et al., 2005). Other common personal and organisational barriers to EHR adoption were analysed in the CMVH Literature Review 2010 (Cotea, 2010).

In our study, the analysis of EMR users' and non-users' perceptions of barriers revealed a statistically significant difference only in the concern of a malfunction in the system. This concern was more often reported by physicians who had not implemented the system than by those who had implemented the system. This may suggests that the system malfunction risk is less common in real use than it is conceived to be. There is also a risk that physicians who haven't implemented the system tend to exaggerate the

malfunction in the system. Such an attitude may be caused by lack of proper information in medical society or may suggest that some physicians prefer to stress external problems (system malfunction) rather than deal with internal barriers (resistance for change or lack of time). The analysis of other differences in the perceived barriers of EMR between users and non-users is limited due to small sample size and discrepancies between declared and real implementation stages of EMR systems. In the studied group of 103 GPs, 55.3% declared partial or full implementation of an EMR system, while in fact only 14.7% stored complete patient medical histories (i.e., simultaneous collection of data on medical diagnoses, drug prescriptions, medical interviews, physical examinations, anthropometric measurements and diagnostic test results) (Pedzinski et al., 2013).

Conclusions

In the Podlaskie Voivodeship the GP's attitude to ICT in primary healthcare is generally positive or neutral and resembles the overall pattern in Europe. Lack of funds, risk of a malfunction in the system, resistance to change and lack of training as well as proper information about EMR were the most common barriers to EMR implementation. Statistically significant differences between EMR users' and non-users' perceptions of barriers were only shown in the concern of a malfunction in the EMR system, which suggests that this problem might be exaggerated.

It is crucial to take into account the barriers perceived by GPs while their attitudes are a significant factor in the acceptance and efficiency of EMR in practice. Unfortunately, the legal obligation for healthcare providers to implement EMR without any financial or non-financial incentives may undermine the widespread use of e-health.

REFERENCES

European Commission. Information Society and Media Directorate General. (2007). *Benchmarking ICT use among General Practitioners in Europe 2007, Country Profile: Poland.*

Chaudhry, B., Wang, J., & Wu, S. (2006). Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Annals of Internal Medicine*, 144(10), 742–752.

Cotea, C. (2010). Electronic Health Record Adoption: Perceived Barriers and Facilitators. *Research Coordination Unit, CMVH.*

DesRoches, C. M., Painter, M. W., & Jha, A. K. (2013). *Health Information Technology in the United States: Better Information Systems for Better Care.* Robert Wood Johnson Foundation's annual report.

Dobrev, A., Haesner, M., & Husing, T. (2008). *Benchmarking ICT use among General Practitioners in Europe – final report.* Bonn: European Commision, 53–60.

Gans, D., Kralewski, J., Hammons, T., & Dowd, B. (2005). Medical groups' adoption of electronic health records and information systems. *Health Affairs*, 24(5), 1323–1333.

Georgiou, A., Ampt, A., Creswick, N., Westbrook, J. I., & Braithwaite, J. (2009). Computerized Provider Order Entry – What are health professionals concerned about? A qualitative study in an Australian hospital. *International Journal of Medical Informatics*, 78(1), 60–70.

Hackl, W., Hoerbst, A., & Ammenwerth, E. (2009). The electronic health record in Austria: physicians' acceptance is influenced by negative emotions. *Studies in Health Technology & Informatics*, 150, 140–144.

Hier, D. B., Rothschild, A., LeMaistre, A., & Keeler, J. (2005). Differing faculty and house staff acceptance of an electronic health record. *International Journal of Medical Informatics*, 74(7–8), 657–662.

Hillestad, R., Bigelow, J., & Bower, A. (2005). Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Affairs (Millwood)*, 24(5), 1103–1117.

Jha, A. K., Bates, D. W., Jenter, C., Orav, E. J., Zheng, J., Cleary, P., & Simon, R. (2009). Electronic Health Records: Use, Barriers and Satisfaction Among Physicians Who Care For Black and Hispanic Patients. *Journal of Evaluation in Clinical Practice*, 15(1), 158–163. doi: 10.1111/j.1365-2753.2008.00975.x.

Jha, A. K., DesRoches, C. M., Campbell, E. G., Donelan, K., Sowmya, R. R., Ferris, T. G., Shields, A, Rosenbaum, S., & Blumenthal, D. (2009). Use of electronic health records in US hospitals. *The New England Journal of Medicine*, 360(16), 1628–1638.

Kahn, N. B. (2004). The Future of Family Medicine: A Collaborative Project of the Family Medicine Community. *Annals of Family Medicine, 2 (suppl 1)*, s3–s32.

Kossman, S. P., & Scheidenhelm, S. L. (2008). Nurses' perceptions of the impact of electronic health records on work and patient outcomes. *CIN: Computers, Informatics, Nursing*, 26(2), 69–77.

Leung, G. M., Yu, P. L., Wong, I. O., Johnston, J. M., & Tin, K. Y. (2003). Incentives and barriers that influence clinical computerization in Hong Kong: a population-based physician survey. *Journal of the American Medical Informatics Association*, 10, 201–212.

Linder, J. A., Ma, J., Bates, D. W., Middleton, B., & Stafford, R. S. (2007). Electronic health record use and the quality of ambulatory care in the United States. *Archivos de Medicina Interna*, 167(13), 1400–1405.

Lockhart, P. (2008). Two nations achieve a high level of primary care information technology (IT) interoperability: an introduction to a series comparing Denmark and New Zealand's IT and health care. *Informatics in Primary Care*, 16(3), 179–181.

Loomis, G. A., Ries, J. S., Saywell Jr., R. M., & Thakker, N. R. (2002). If electronic medical records are so great, why aren't family physicians using them? *Journal of Family Practice*, 51(7), 636–641.

Menachemi, N. & Collum, T. H. (2011). Benefits and drawbacks of electronic health record systems. *Risk Management and Healthcare Policy*, 4, 47–55.

Miller, R. H., & Sim, I. (2004). Physicians' Use Of Electronic Medical Records: Barriers And Solutions. *Health Affairs*, 23(2), 116–126.

Morin, D., Tourigny, A., Pelletier, D., Robichaud, L., Mathieu, L., Vezina, A., Bonin, L., Buteau, M. (2005). Seniors' views on the use of electronic health records. *Informatics in Primary Care*, 13, 125–133.

Najwyzsza Izba Kontroli. (2013). *Informatyzacja szpitali – informacja o wynikach kontroli*, Raport NIK 2013.

Pedzinski, B., Sowa, P., Kolpak, M., Pedzinski, W., & Szpak, A. (2013). The use of electronic medical records in primary healthcare in the podlaskie voivodeship. *Polish Journal of Public Health*, 123(2), 107–111.

Peterson, K. A. (2012). Essential requirements of information technology for primary care. *Family Practice*, 29(2), 119–120.

Pike, Ch. (2010). An Empirical Analysis of the effects of GP competition. *Cooperation & Competition Panel Working Paper Series*, vol. 1(2).

Poissant, L., Pereira, J., Tamblyn, R., & Kawasumi, Y. (2005). The Impact of Electronic Health Records on Time Efficiency of Physicians and Nurses: A Systematic Review. *Journal of the American Medical Informatics Association*, 12(5), 505–516.

Police, R. L., Foster, T. & Wong, K. S. (2010). Adoption and use of health information technology in physician practice organisations: systematic review. *Informatics in Primary Care*, 18(4), 245–258.

Protti, D., & Johansen, I. (2010). Widespread Adoption of Information Technology in Primary Care Physician Offices in Denmark: A Case Study. *The Commonwealth Fund: 'Issues in international health policy'*, 1379(80).

Romano, M. J., & Stafford, R. S. (2011). Electronic health records and clinical decision support systems: impact on national ambulatory care quality. *Archivos de Medicina Interna*, 171(10), 897–903.

Saleem, J. J., Patterson, E. S., Militello, L., Render, M. L., Orshansky, G., & Asch, S. M. (2005). Exploring barriers and facilitators to the use of computerized clinical reminders. *Journal of the American Medical Informatics Association*, 12(4), 438–447.

Schoen, C., Osborn, R., Squires, D., Doty, M., Rasmussen, P., Pierson, R., & Applebaum, S. (2012). A Survey of Primary Care Doctors in Ten Countries Shows Progress in Use of Health Information Technology, Less in Other Areas. *Health Affairs (Millwood)*, 31(12), 2805–2816.

Sequist, T. D., Cullen, T., Hays, H., Taualii, M. M., Simon, S. R., & Bates, D. W. (2007). Implementation and use of an electronic health record within the Indian Health Service. *Journal of the American Medical Informatics Association*, 14(2), 191–197.

Ustawa z dnia 28 kwietnia 2011 o systemie informacji w ochronie zdrowia. (2011). Dz.U. 2011 vol. 113 item 657.

Wang, S. J., Middleton, B., & Prosser, L. A. (2003). A cost-benefit analysis of electronic medical records in primary care. *The American Journal of Medicine*, 114(5), 397–403.

Zakaria, N., Mohd Affendi, S. Y., & Zakaria, N. (2010). Managing ICT in healthcare organization: culture, challenges, and issues of technology adoption and implementation. In Y. K. Dwivedi, K. Khoumbati, B. Lal, & A. Srivastava (Eds.), *Handbook of Research on Advances in Health Informatics and Electronic Healthcare Applications: Global Adoption and Impact of Information Communication Technologies* (pp. 153–168). Hershey, PA: IGI Global.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0044

The Knowledge of Medical Professionals from Selected Hospitals in the Lubelskie Province about Diagnosis-Related Groups Systems

Petre Iltchev[1], Aleksandra Sierocka[2], Sebastian Gierczyński[1], Michał Marczak[1]

[1] Chair of Health Care Policy, Medical University of Lodz, Poland
[2] K. Jonscher Hospital in Lodz, Poland

Abstract. Health information technology (IT) in hospitals can be approached as a tool to reduce health care costs and improve hospital efficiency and profitability, increase the quality of healthcare services, and make the transition to patient-centered healthcare. A hospital's efficiency and profitability depends on linking IT with the knowledge and motivation of medical personnel. It is important to design and execute a knowledge management strategy as a part of the implementation of IT in hospital management. A Diagnosis-Related Groups (DRG) system was introduced in Poland in 2008 as a basis for settlements between hospitals and the National Health Fund (NHF). The importance and role of a DRG system in management of healthcare entities was emphasized based on a survey of medical professionals from two hospitals in the Lubelskie province. The goal of a survey is to assess the knowledge of medical professionals about the DRG system and how the medical personnel uses the DRG system in order to achieve the strategic goals of the organization. A newly developed survey was used to assess the medical personnel's knowledge of DRG, using 12 closed and 5 open questions. The survey was conducted on 160 medical employees from two hospitals in the Lubelskie province. In conclusion, medical personnel's DRG knowledge unambiguously contributes to reducing hospital costs and increasing profitability. The DRG related knowledge enables personnel to obtain value from data by applying DRG data-driven decisions.

Introduction

Implementation of DRG in Poland forced a change in hospitals to occur. The biggest challenge in implementing DRG reporting is to understand its vast impact on hospital operations, costs, and profitability. DRG directly impacts each medical professional from nurses to coders to physicians. In Poland, the labor costs in the hospital budget are between 40 and 70 percent. Optimization of management of medical staff, as far as their cost culture, can have an impact on a hospital's efficiency and debt. Staff engagement

ISBN 978-83-7431-392-6 ISSN 0860-150X

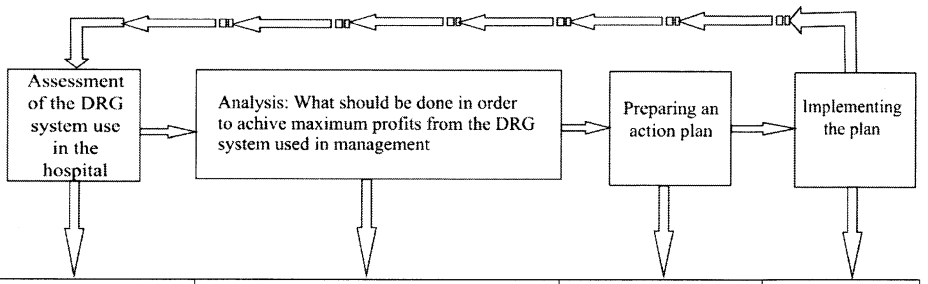
in costs is key for hospital efficiency and profitability. DRG is a way to improve the hospital costs culture. Better collaboration between coders, nurses and physicians, a culture of team work, can overcome barriers hampering hospital efficiency improvements.

Health information technology (IT) in hospitals can be approached as a tool to reduce health care costs and improve hospital efficiency and profitability, increase the quality of healthcare services, and make the transition to patient-centered healthcare. The hospital efficiency and profitability depends on linking IT with the knowledge and motivation of medical personnel. The success of using IT applications is impacted by implementation and the knowledge of users. You cannot separate initiatives designed to increase the knowledge and motivation of the medical personnel from the implementation and use of IT in hospitals. Without proper knowledge investments, medical informatics will not deliver the expected results. It is important to design and execute a knowledge management strategy as a part of the implementation of IT in hospital management. The Diagnosis-Related Groups (DRG) related knowledge survey is the first phase of a project for knowledge management. A DRG system was introduced in Poland in 2008 as a basis for settlements between hospitals and the National Health Fund (NHF). DRG is a combination of medical and cost approaches. Five years after the introduction of the DRG system in Poland, we assume that it should form the basis for auditing and strategically managing a hospital. DRG system knowledge among the medical personnel is important for the implementation of the hospital's strategy and for its financial results. The personnel of public hospitals should be interested in their positive financial results, especially taking article 59 of the Act on Medical Activity (Ustawa z dnia 15 kwietnia 2011 r. o działalności leczniczej, 2011) into account. Pursuant to article 59, "1. An independent public health care institution will cover on its own its negative financial result ... 4. In case the negative financial result is not covered using a method established in clause 2, the entity which created the institution within 12 months will issue a decree, regulation or a resolution to change the organisational and legal form or to dissolve the independent public health care institution".

In accordance with the 80/20 principle, there are a few factors which ensure the effect of using a DRG system in the management of the hospital. The following are examples of key indicators that characterise the DRG system in auditing and financial management of the hospital:

a) profit from a patient,
b) profit from a medical procedure,
c) daily profit from a patient.

Medical professionals who are aware of the role of a DRG system are able to use it better in the hospital management process and to take it into account when making medical decisions. Figure 1 presents the role of the analysis of the medical personnel's DRG system knowledge in the hospital management process.



Analysis of the processes	Analysis of the use of DRG information in the process of making strategic, tactical and	DRG system training	Modelling with the use of DRG
Employee interviews and surveys	Analysis of the needs of decision-makers. Analysis of the necessary changes to IT	Implementing a DRG based controlling system	Pharmaco-economical
Case study, e.g. the introduction of ICD-100, ICD-11 in other countries	Brainstorming - proposals of actions, new sources of information, improvements in organisation and course of processes.	Introduction of a new system for motivating employees	
DRG system knowledge analysis	Establishing the types of training needed by individual groups of medical employees	Providing the financial resources for the projects' implementation	

Figure 1. The place of employees' DRG system knowledge in the hospital management process

Investment in the DRG system training of medical staff can be very profitable for hospitals (Walker et al., 2010). A few examples of case study topics are provided by Busse et al. (2011), Minich-Pourshadi (2011) and Walker et al. (2011). Knowledge analysis related to a DRG system is fundamental for successful DRG implementation and realization of profits from the contract with the NHF. This process can be helped by applying the work of McKenzie et al. (2003). Modelling hospital financial result with the use of DRG can be based on the works of Wike et al. (2011) and HOPE – European Hospital and Healthcare Federation (2006). For an example of pharmaco-economical simulations, the work of Wike et al. (2010) can be consulted. Analysis of the use of information related to DRG in the process of making strategic, tactical and operational decisions, as described by LaValle et al. (2010), is fundamental for giving hospitals a competitive advantage in the new economic environment. Analyses of the necessary changes

in IT systems are crucial for hospital operational effectiveness (ECRI Institute, 2012).

The development of the DRG system should be understood to be an element of a widely understood hospital IT strategy (HealthLeaders Media, 2010). Thus, it is very important to establish how the DRG system support infrastructure will be integrated with the remaining IT systems used in the hospital. It is also important to answer the question of what impact the changes to the DRG system will have on the hospital's finances (Busse et al., 2011, Charlesworth et al., 2012, HealthLeaders Media, 2011).

Data – Survey Results and Interpretation

The study was conducted among the medical personnel in two hospitals in the Lubelskie province – the Bychawa District Hospital and Cardinal Stefan Wyszyński Provincial Specialist Hospital – within the January – April 2013 period. The structure of surveys received from the study participants (160) is as follows in Table 1.

Table 1. Participants in the study, as per profession and medical entity

	Bychawa District Hospital	Cardinal Stefan Wyszyński Provincial Specialist Hospital in Lublin	Total
Head of a Hospital Ward	2	8	10
Physician	8	20	28
Nurse	50	50	100
Medical statistician	3	9	12
Medical secretary	3	7	10

The provided answers to the survey show that in order to improve their financial results in a manner that does not require significant investments, the hospitals should commence by organising training courses. Internal courses are an investment that will bring the highest return. Small financial outlays enable the general DRG system knowledge level for the entire medical personnel to increase as well as its impact on the hospital's financial result. Indicating the role of the system, the sources of information, the connection of efficiency, cost-effectiveness and effectiveness oriented thinking with responsibility, in addition to a motivating system, are the keys

to increasing the operational effectiveness of hospitals. This connection is presented in a graphic manner in the Figure 2.

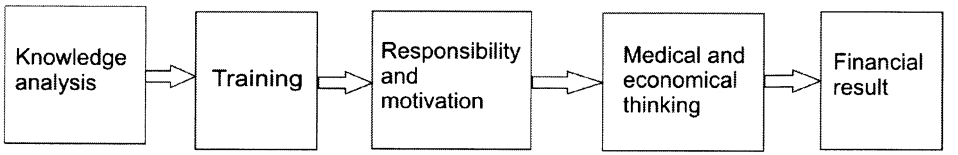


Figure 2. The medical personnel economic awareness increase cycle

For 74% of employees responding to the survey, no answer and the opinion that the DRG system has no impact on their work means that they are not able to assess the economic results of the medical decisions they make. The lack of economic knowledge of the medical personnel in question means that they are not aware that the results of the medical decisions taken may frequently result in a negative financial result for a specific medical case. The sum of financial results of individual cases (patients) creates the financial result of the entire hospital. The change of approach and organisation of work is important, to ensure that the medical personnel do not consider the DRG system to be an obstacle. The awareness of the economic impact of the DRG system on the hospital's financial result, as well as the remuneration of each employee and active search for work improvement methods, may change the approach of persons who consider the use of the system to be an obstacle and make them actively use the system.

The DRG system in Poland is undergoing dynamic changes and constant development. Changes occur in illness groups, individual units, dictionaries of admittance and discharge, and finally, principles of contracting. All these changes justify the need for continuous training instead of single sessions. The answer of close to 1/3 of those surveyed is even more incomprehensible in that light. This shows that the hospital management needs to engender a significant increase in awareness among the medical personnel on the importance of the DRG system for the hospital's operation and positive financial result. In addition, the selected forms of training indicate a preference among medical employees for more modern online training courses, or an on-line connection in a hospital or outside the medical institution, compared to more traditional forms of training.

If we are aiming to achieve a positive financial result in a hospital, each ward should be trying to generate profit. Otherwise, each hospital ward head will think that other wards should generate profits, instead of his or hers. As a consequence of such an approach, the hospital will probably have a negative financial result.

A detailed analysis of responses provided by medical statisticians and key persons responsible for the execution of an NHF contract, that is, heads of hospital wards and physicians, enables better planning of actions intended to increase the use of DRG in hospital management. Table 2 presents their answers, compared to all the surveyed persons.

Table 2. Answers of key persons, compared to all the surveyed persons

Scope of the question	Percentage of correct answers			
	Heads of hospital wards	Physicians	Medical statisticians	All answers
Knowledge concerning the frequency of reporting of execution of contract	100	29	100	25
Knowledge whether the hospital departments are held accountable pursuant to income and costs of execution of NHF contracts in accordance with DRG	100	7	100	27
Knowledge whether the medical personnel knows the contract values of the hospital ward	100	100	0	20
The execution of a contract is monitored pursuant to a monthly reports prepared using medical documentation	100	86	17	24
After execution of a contract fewer procedures are performed	100	54	0	5
Participation in DRG-related courses and training	90	0	67	14
The DRG system improves management of the hospital's finances	60	18	58	16
Introduction of a DRG-based controlling system may improve the hospital's finances management	20	0	0	1
DRG system training for the personnel will improve the hospital's finances management	40	46	25	18
Self-assessment of DRG system knowledge (on a satisfactory level)	100	18	83	18

Scope of the question	Percentage of correct answers			
	Heads of hospital wards	Physicians	Medical statisticians	All answers
Courses and training were the preferred form of DRG skills and knowledge improvement of the surveyed	100	68	83	19
DRG system related activities do not hinder working directly with a patient	50	14	0	15
DRG system related activities are too time-consuming and hinder working directly with a patient	50	43	0	46
The surveyed think that the DRG system is complicated and using it requires specialised knowledge	70	82	100	62
The DRG system is used in the management of a hospital ward when filling out medical documentation and reports	50	14	42	13
The DRG system is not used in the management of a hospital ward	20	14	0	6
The motivation and wages system is not connected with the level of costs	100	93	92	79
The hospital supports new ideas concerning management, work organisation and process improvement	90	11	0	8
Medical personnel needs contract-related information in order to better manage the hospital / ward / patient streams / processes in the area of: number of contracted services, degree of contract execution	100	79	0	28

The data presented in Table 2 show that the analysis of answers from such types of surveys should not be overly generalised. This may lead to incorrect decisions. Analysing the answers of individual groups of employees to individual questions, we may deduce the expected, desirable and actual state of providing the personnel with contract information. Medical personnel know the hospital contracts that have an impact on its operation.

The answer to this question shows that the management of the studied hospitals is convinced that the personnel know the contracts that impact them. Whereas, according to the statisticians, persons processing the contract data, the employees do not have this information in the studied hospitals. From the answers to this question, we may suggest the following direct actions: notifying the employees about the contracts by providing information both from direct superiors, as well as by placing these data in the hospital's IT system (intranet) and showing employees how to search for them.

Proposed Solutions

Before we explain why we talk about the role of the DRG system, we should answer a few questions and consider a few important issues. Before explaining why, let's take a look at a few DRG implementations and ask ourselves a few questions. Firstly, think about your controlling experiences over the years. In theory, the use of a DRG system in management seems simple and obvious. In practice, it is much more complex. We may indicate the following limitations: a) organisational structure of a hospital; b) frequently no implemented controlling system; c) hindered conduction of organisational changes; d) no established methodology of DRG system use in hospital management.

Medical personnel decide on the hospital's financial results by making medical decisions, taking the economic aspect into account. Knowledge of the DRG system, which is the basis of contracting with the NHF and of settling and paying for the services, has a deciding impact on the operational effectiveness of the healthcare entity. Organising and conducting medical personnel training is a non-investment route to increasing the operational effectiveness of the hospital. The conducted survey may be considered the first step to organizing the training. A more detailed survey will enable management to determine what type of training various types of medical personnel require. Individualising the assessment of knowledge and modifying the training using an e-learning platform enables one to adapt the contents and training time to suit the needs and capabilities of each medical employee.

The survey shows that in addition to training and motivation, the implementation of the auditing is important for the use of the DRG system in the financial management of the hospital. Establishing the expectations (again) in relation to the DRG system's place in the financial management

of the hospital should be the initial point in the process of DRG implementation concerning auditing and strategic management of the hospital (redefine what DRG means). Thus, the expectations of the auditing system should be defined according to what services it should deliver and what answers it should provide. Requirements for the auditing system with DRG elements include: a) ease of use of the system; b) automatic data entry (automatic exchange of information between independent IT systems implemented in the hospital), which will save time; c) availability of information; d) supporting the financial decision making process, for example by comparison over time.

Creating a description and presentation of best practices (cases) is one of the ways to propagate solutions tested in one hospital ward to others. Creating a culture of DRG system use in the hospital's financial management will not only improve the financial result, but over a longer time horizon, may also be a factor that ensures a competitive advantage.

Conclusions

In order to obtain maximum benefits from the DRG system, hospitals should educate their medical personnel on the system in question. Training and the process of solving case studies will lead to an increased consciousness and effectiveness of the decisions taken, by raising the economic awareness of the personnel, and will improve the financial results of the hospitals.

The presented study provides information about the DRG system knowledge possessed by employees in the studied hospitals. The method used and the questions posed may be used for the analysis of the state of DRG system knowledge in every other medical institution. Survey results form the basis for establishing a plan for action, and are a perfect reflection of the situation present in most Polish medical institutions. The proposed actions, intended to increase the knowledge of medical personnel concerning the DRG, may be the initial point to start from as every medical institution prepares its own programme.

A challenge for the hospital management is implementing modern DRG based controlling systems in order to increase the responsibility of medical personnel for the financial results of the institution.

Surveys conducted in 2 hospitals are not enough to form a basis for general conclusions concerning DRG system knowledge. This study is intended to establish the state of knowledge in the surveyed hospitals in order to react, to create a program of appropriate actions, which will result in

an increased knowledge and will be used in the management process. An active approach to the DRG system is the road to profit (wider gains in profitability).

We have found that the main cause of disappointment in implementing DRG related IT systems is lack of knowledge about DRG, relatively minor investment in training, and an underfunded training budget. One of the obstacles in the implementation of IT systems in hospitals is the gap between the functional capabilities of the implemented solution and knowledge of hospital personnel. Increasing hospital personnel's knowledge and motivation is a low-cost investment to increase ROI from implemented IT systems in hospitals.

In order to maximise the effects while minimising the costs, and to achieve positive effects as rapidly as possible, the training may be conducted in stages. Heads of hospital wards, physicians and medical statisticians have to be trained first. They will be leaders in implementing changes – initiating new actions intended to increase economic effectiveness and to achieve a permanent positive financial result (profit).

For hospital efficiency, it is important to move away from DRG reporting to cost culture. Cost culture can improve the operations of the hospital and drive down the costs. However, it takes some time to develop cost culture in hospitals. Implementation of DRG in hospitals in Poland may not immediately lead to operational efficiencies and cost reductions.

The survey was conducted after obtaining approval from the Bioethics Committee of the Medical University in Łódź and from the Bychawa and Lublin hospital directors.

REFERENCES

Busse, R., Geissler, A., Quentin, W., & Wiley, M. (Eds.). (2011). *Diagnosis-Related Groups in Europe. Moving towards transparency, effciency and quality in hospitals.* Open University Press. World Health Organization 2011 on behalf of the European Observatory on Health Systems and Policies. Retrieved from http://www.euro.who.int/__data/assets/pdf_file/0004/162265/e96538.pdf.

Charlesworth, A., Davies, A., & Dixon, J. (2012). *Reforming payment for health care in Europe to achieve better value* (Research report). Retrieved from: http://www.kpmg.no/arch/_img/9826485.pdf.

ECRI Institute. (2012). Electronic Health Records: Is your hospital making all the right connections? in: *ECRI Institute's top 10 C-suite watch list. Hospital Technology issues for 2012* (pp. 2–3). Retrieved from: https://www.ecri.org/Forms/Documents/ECRI_Institute_Top_10_C-Suite_Watch_List_Hospital_Technology_Issues_for_2012.pdf.

HealthLeaders Media. (2010). HealthLeaders Media Breakthroughs. *HIT that enables quality, efficiency and value.* Retrived from http://www.healthleaders media.com.

HealthLeaders Media. (2011). HealthLeaders Media Impact Analysis. *ICD-10 skating on thin margins.* Retrived from http://www.healthleadersmedia.com.

HOPE – European Hospital and Healthcare Federation. (2006). *DRG as a financing tool.* Retrieved from http://www.hope.be/05eventsandpublications/docpublications/77_drg_report/77_drg_report_2006.pdf.

LaValle, S., Hopkins, M., Lesser, E., Shockley, R., & Kruschwitz, N. (2010). *Analytics: The new path to value. How the smartest organizations are embedding analytics to transform insights into action.* IBM Global Business Services. Business Analytics and Optimization Executive Report. MIT Sloan Management Review. Retrieved from http://public.dhe.ibm.com/common/ssi/ecm/en/gbe03382usen/GBE03382USEN.PDF.

McKenzie, K., Walker, S., & Lewis, M. (2003). Building the Bridge to E-Coding. In *Proceedings Health Information Management Association of Australia Conference 2003*, Sydney. Retrived from http://eprints.qut.edu.au/7185/.

Minich-Pourshadi, K. (2011). ICD-10 Puts revenue at risk. *HealthLeaders Media Intelligence.* Retrieved from http://content.hcpro.com/pdf/content/268585.pdf.

Ustawa z dnia 15 kwietnia 2011 r. o działalności leczniczej. (2011). Dz.U. 2011 vol. 112 item 654. Retrived from http://isap.sejm.gov.pl/DetailsServlet?id=WDU20111120654.

Walker, S., Eynstone-Hinkins, J., & Schmider, A. (2011). Implementation of ICD-11 perspectives from the Australian Bureau of Statistics. In R. Jakob (Ed.), *World Health Organization Family of International Classifications Network Meeting*, 29 October – 5 November 2011, Southern Sun, Cape Town, South Africa. Retrived from http://eprints.qut.edu.au/46662/.

Walker, S., & Waller, G.. (2010). Experiences in training ICD-10 trainers. In R. Jakob (Ed.), *Proceedings of the WHO-FIC Annual Meeting, Toronto 2010*, World Health Organization. Retrieved from http://eprints.qut.edu.au/46664/.

Wike, M., & Grube, R. (2010). Pharmaco-economic evaluation of antibiotic therapy strategies in DRG-based healthcare systems – a new approach. *European Journal of Medical Research*, 15, 564–570.

Wike, M., Grube, R., & Bodmann, K. (2011). The use of a standardized PCT-algorithm reduces costs in intensive care in septic patients – A DRG-based simulation model. *European Journal of Medical Research*, 16, 543–548.

VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0037

Assessment of the Effectiveness of Medical Education on the Moodle e-Learning Platform

Wiesław Półjanowicz[1], Grzegorz Mrugacz[2], Michał Szumiński[3], Robert Latosiewicz[4], Alina Bakunowicz-Łazarczyk[3], Anna Bryl[3], Małgorzata Mrugacz[3]

[1] Department of Didactics and Modern Technologies in Education, University of Bialystok, Poland

[2] Center for Reproductive Medicine "Bocian", Bialystok, Poland

[3] Department of Pediatric Ophthalmology, Medical University of Bialystok, Poland

[4] Department of Rehabilitation and Physiotherapy, Medical University of Lublin, Poland

Abstract. This paper presents an analysis of learning effectiveness for the courses "Selected issues in visual rehabilitation" and "Ophthalmology and ophthalmic nursing" taught in the years 2009–2011 at the Medical University of Bialystok, Poland. We compared the effectiveness of traditional and distance learning methods; an e-learning platform was implemented experimentally for the purpose of this study. We assessed the usefulness of online learning in terms of organization, knowledge gained and students' satisfaction with the course. The study was conducted among 75 second year master degree students in the nursing field in the academic years 2009/2010 and 2010/2011. The students were divided into two groups. For the study group of 39 persons (52%), lectures and seminars took place on an e-learning platform, while 36 persons (48%) in the control group attended traditional classes. 80% of students in the e-learning group and 89% of students in the traditional group assessed the organization of both forms of courses positively. The fact that the majority of students in both the e-learning (89%) and traditional classes (86%) gave positive feedback indicates that for both forms there was a high level of content and technical preparedness. The mean scores of the final exam for both courses were 82% in the e-learning group and 79% in the traditional group in the years 2009–2011. The above results show that both forms of learning are equally effective.

Introduction

A growing number of people are interested in distance learning. Widespread access to computers and the Internet means that more and more people can participate in the e-learning process. The search for new methods of learning and teaching has caused distance learning to be increasingly used at the university level. E-learning is independent of place

ISBN 978-83-7431-392-6 ISSN 0860-150X

and time, and the teachers determine the rules of conduct and access to classes. E-learning does not intend to replace traditional forms of education, but it is a good way to supplement and widen knowledge transfer (Allan, 2007; Arbaugh et al., 2010; Bramley, 2001; Smith et al., 2008).

The availability of a number of distance learning programs enables the development of modern e-learning courses, which are placed on an accessible platform in accordance with SCORM, AICC, and IMS (Piskurich, 2003; Waćkowski et al., 2007). With these tools, interactive tests, quizzes and tasks to test students' knowledge can be created.

This study of learning effectiveness in the courses "Selected issues in visual rehabilitation" and "Ophthalmology and ophthalmic nursing" aimed at evaluation of the effectiveness of nursing education supplemented by distance learning techniques. The Bioethics Committee of the Medical University of Bialystok approved the study (consent No. R-I-002/338/2009).

Aim

The courses "Selected issues in visual rehabilitation" and "Ophthalmology and ophthalmic nursing" were taught experimentally in a complementary e-learning system during the 2009–2011 academic years. The lectures and seminars were conducted on-line, but practical classes were taught traditionally. Students who participated in the e-learning method had free access to the distance-learning platform (Moodle) and to the courses. Students had continuous and unlimited access to the teaching materials on the platform, but the teacher responsible for the course determined the order of the topics and the period of their availability. The final exam was, however, in "traditional" form, as laid out in the Regulations of the Medical University of Bialystok.

Material and Methods

The study was conducted with a group of 75 second year master degree nursing students enrolled in the courses "Selected issues in visual rehabilitation" and "Ophthalmology and ophthalmic nursing" during the years 2009–2011. The students were divided into two groups. In the control group (36 people), lectures and seminars were taught traditionally. In the study group (39 people), lectures and seminars took place on the e-learning platform MOODLE, ver. 1.9 (Rice, 2010) (Figure 1), where all didactic ma-

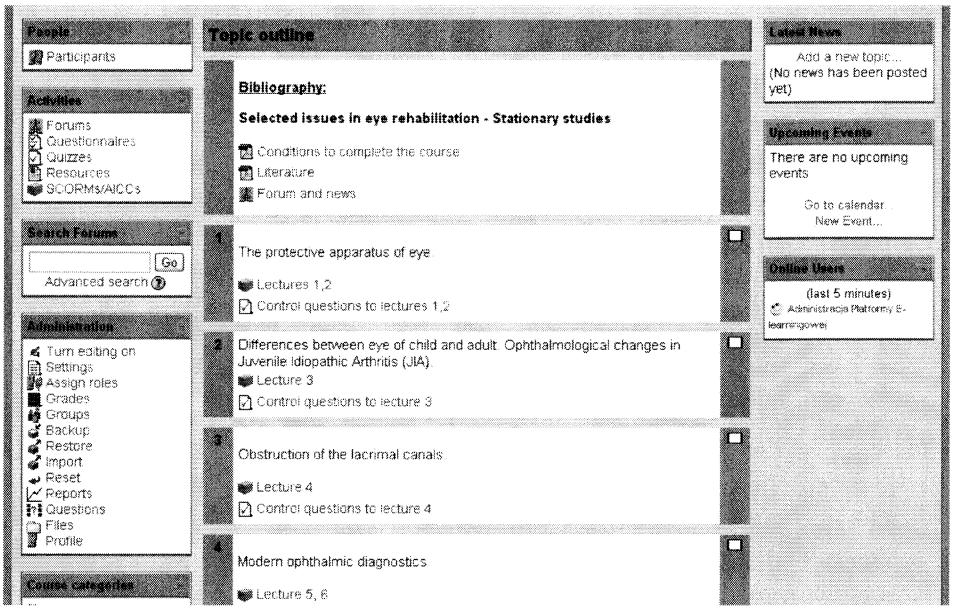


Figure 1. Screenshot of e-learning course "Selected issues in visual rehabilitation"

terials, prepared for distance learning were available. Every student from the study group had 24-hour access to didactic materials, including multimedia, knowledge assessment tests and evaluation forms. Students' activity (e.g. lecture reading, watching multimedia materials or completing an evaluation form) was registered and a log of individual achievements was created. The order of didactic topics and the period of their availability was supervised by academic teachers responsible for particular subjects.

A final assessment of both courses (e-learning and traditional) was conducted traditionally in the form of a multiple-choice test at the same time for both groups. The final exam scores for both groups were compared in order to analyze the level of knowledge gained. At the end of the cycle of classes but before completion of the final exam, students in both groups filled an evaluation questionnaire pertaining to their opinions of the course, the level of their satisfaction with the course, and the organization of classes.

Results and Discussion

This study on learning effectiveness for the courses "Selected issues in visual rehabilitation" and "Ophthalmology and ophthalmic nursing" in-

Table 1. Final exam scores (in %) for the study and the control groups in the course "Selected issues in visual rehabilitation" in the academic years 2009–2011

Academic year	Group	Parameter								
		n	mean score $\bar{x}$	SD	Min.	Q_1	Me	Q_3	Max.	p
2009/2010	Study	22	81.8%	9.1%	65%	75%	80%	90%	100%	**0.004**
	Control	17	71.8%	10.4%	55%	65%	70%	80%	90%	
2010/2011	Study	17	86.5%	8.2%	75%	80%	90%	90%	100%	0.34
	Control	19	83.9%	5.4%	75%	80%	85%	90%	90%	

Table 2. Final exam scores (in %) for the study and the control groups in the course "Ophthalmology and ophthalmic nursing" in the academic years 2009–2011

Academic year	Group	Parameter								
		n	mean score $\bar{x}$	SD	Min.	Q_1	Me	Q_3	Max.	p
2009/2010	Study	22	78.0%	13.2%	47%	71%	79%	89%	95%	0.62
	Control	17	74.9%	14.9%	47%	63%	79%	89%	95%	
2010/2011	Study	17	79.4%	3.3%	72%	78%	78%	83%	83%	**0.04**
	Control	19	81.6%	2.7%	78%	78%	83%	83%	83%	

cluded relatively equal groups of students participating through the use of e-learning and according to the traditional method. 39 persons were (52%) in the e-learning group and 36 students (48%) belonged to the traditional learning methods group. Table 1 presents the mean final exam scores for the course "Selected issues in visual rehabilitation" of the students in the e-learning and traditional groups in the academic years 2009–2011. Table 2 presents the mean final exam scores for the course "Ophthalmology and ophthalmic nursing" in both groups during the same period.

It can be concluded that the mean scores obtained on the final exams during 2009–2011 had an upward trend for both courses in both groups of students.

In the academic year 2009/2010, the mean final exam score for the course "Selected issues in visual rehabilitation" was 81.8% ($\pm$9.1%) in the

e-learning group and 71.8% (±10.4%) in the traditional group. The differences between these mean scores were statistically significant ($p < 0.05$). In 2010/2011 mean scores were 86.5% (±8.2%) and 83.9% (±5.4%) respectively, which was not statistically significant ($p > 0.05$).

In the academic year 2009/2010, the mean final exam score for the course "Ophthalmology and ophthalmic nursing" was 78% (±13.2%) in the e-learning group and 74.9% (±14.9%) in the traditional group. The difference was not statistically significant ($p > 0.05$). In 2010/2011, the mean final exam score for this course was 79.4% (±3.3%) and 81.6% (±2.7%), respectively. Differences between these mean scores were statistically significant ($p < 0.05$).

In the study group (e-learning), the median final exam score for the course "Selected issues in visual rehabilitation" was 80% in 2009/2010 and 90% in 2010/2011. Differences between these scores were not statistically significant ($p > 0.05$). In the control group (traditional form of learning) the median final exam scores for this course during these years were: 70% and 85%, respectively. These differences in scores were statistically significant ($p < 0.001$). The results obtained are presented in Figure 2.

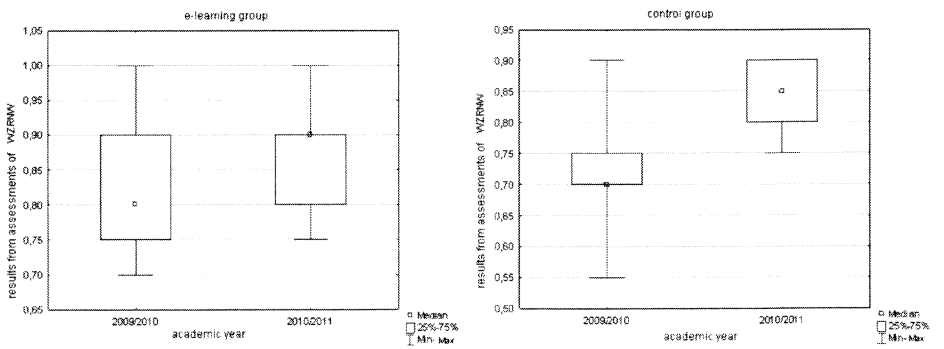


Figure 2. Final exam scores for the course "Selected issues in visual rehabilitation" in the academic years 2009–2011 in the study and control groups

A slightly better median final exam score in the study group may be explained by full-time access to didactic materials on the e-learning platform, which facilitated repeated readings and multiple analyses of didactic materials. In contrast, students from the control group based their studying on their own notes and information from the recommended literature list.

The change in the final exam scores for the course "Ophthalmology and ophthalmic nursing" during 2009–2011 (two academic years) is illustrated

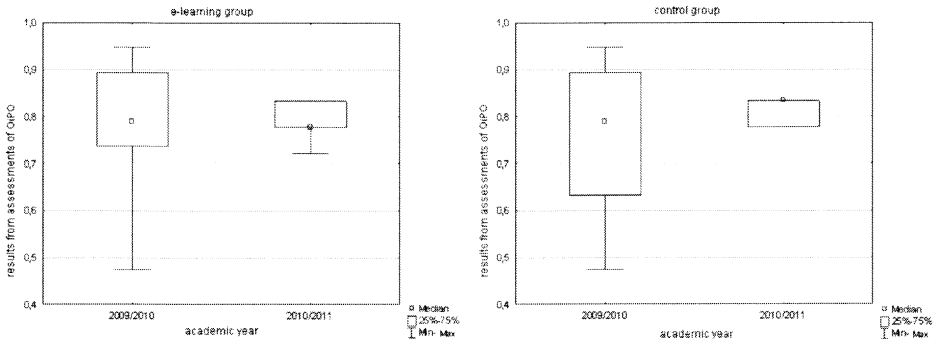


Figure 3. Final exam scores in the course "Ophthalmology and ophthalmic nursing" in the academic years 2009–2011 in the study and control groups

Table 3. Assessment of course organization in the academic years 2009–2011

Academic year	Course	Group	Parameter						
			n	definitely no (1)	rather no (2)	I don't know (I don't have any opinion) (3)	rather yes (4)	definitely yes (5)	*p*
2009/2010	Selected issues in visual rehabilitation	Study	22 100%	0 0%	4 18.2%	2 9.1%	12 54.5%	4 18.2%	0.531
		Control	17 100%	0 0%	3 17.6%	0 0%	12 70.6%	2 11.8%	
	Ophthalmology and ophthalmic nursing	Study	22 100%	0 0%	3 13.6%	2 9.1%	14 63.6%	3 13.6%	0.628
		Control	17 100%	0 0%	3 17.6%	0 0%	12 70.6%	2 11.8%	
2010/2011	Selected issues in visual rehabilitation	Study	17 100%	0 0%	0 0%	0 0%	9 52.9%	8 47.1%	0.138
		Control	19 100%	0 0%	0 0%	2 10.5%	13 68.4%	4 21.1%	
	Ophthalmology and ophthalmic nursing	Study	17 100%	1 5.9%	2 11.8%	2 11.8%	9 52.9%	3 17.6%	0.139
		Control	19 100%	0 0%	0 0%	0 0%	12 63.2%	7 36.8%	

in Figure 3. The similar results prove that distance learning is comparable to the traditional approach.

Evaluation of the two-year study period shows that the organization of classes was rated positively by 31 (80%) of the students in the e-learning groups and 32 (89%) of the students in the traditional groups. Only about 13% of students had a negative opinion on the organization of both forms of courses (Table 3). Differences on the Likert scale between the analyzed groups were not statistically significant ($p > 0.05$).

In the questionnaire the students were asked about course preparation as well as availability and relevance of the didactic content. 36 students from the e-learning group (91%) and 32 students from control group (88%) graded those issues positively (Table 4). In the academic year 2009/2010, differences between the responses in both groups were statistically significant ($p < 0.05$).

Table 4. Students' opinions on the preparation, availability and relevance of the didactic content of the courses in the years 2009–2011

Academic year	Course	Group	Parameter						
			n	definitely no (1)	rather no (2)	I don't know (I don't have any opinion) (3)	rather yes (4)	definitely yes (5)	*p*
2009/2010	Selected issues in visual rehabilitation	Study	22 100%	0 0%	0 0%	0 0%	14 63.6%	8 36.4%	**0.015**
		Control	17 100%	0 0%	1 5.9%	2 11.8%	14 82.4%	0 0%	
	Ophthalmology and ophthalmic nursing	Study	22 100%	0 0%	0 0%	0 0%	14 63.6%	8 36.4%	**0.015**
		Control	17 100%	0 0%	1 5.9%	2 11.8%	14 82.4%	0 0%	
2010/2011	Selected issues in visual rehabilitation	Study	17 100%	0 0%	0 0%	1 5.9%	10 58.8%	6 35.3%	0.78
		Control	19 100%	0 0%	0 0%	2 10.5%	12 63.2%	5 26.3%	
	Ophthalmology and ophthalmic nursing	Study	17 100%	0 0%	2 11.8%	3 17.6%	6 35.3%	6 35.3%	0.06
		Control	19 100%	0 0%	0 0%	0 0%	13 68.4%	6 31.6%	

The students were asked whether the implemented curriculum enriched their knowledge and skills. 34 students in the e-learning group (86%) and 30 in the control group (83%) responded affirmatively (rather yes, definitely yes) (Table 5). Differences between the analyzed groups, however, were not statistically significant ($p > 0.05$).

Table 5. Opinions on the prepared courses in the years 2009–2011 (acquired knowledge and skills)

Academic year	Course	Group	Parameter						
			n	definitely no (1)	rather no (2)	I don't know (I don't have any opinion) (3)	rather yes (4)	definitely yes (5)	p
2009/2010	Selected issues in visual rehabilitation	Study	22 100%	0 0%	1 4.5%	2 9.1%	13 59.2%	6 27.3%	0.122
		Control	17 100%	0 0%	2 11.8%	2 11.8%	13 76.5%	0 0%	
	Ophthalmology and ophthalmic nursing	Study	22 100%	0 0%	1 4.5%	2 9.1%	14 63.3%	5 22.7%	0.186
		Control	17 100%	0 0%	2 11.8%	2 11.8%	13 76.5%	0 0%	
2010/2011	Selected issues in visual rehabilitation	Study	17 100%	0 0%	0 0%	0 0%	7 41.2%	10 58.8%	0.212
		Control	19 100%	0 0%	2 10.5%	2 10.5%	8 42.1%	7 36.8%	
	Ophthalmology and ophthalmic nursing	Study	17 100%	0 0%	2 11.8%	3 17.6%	8 47.1%	4 23.5%	0.081
		Control	19 100%	0 0%	0 0%	0 0%	11 57.9%	8 42.1%	

The obtained results lead to the conclusion that e-learning can be a good way to supplement traditional methods. In some forms, such as lectures or seminars, it can successfully replace traditional classes. Setting e-learning courses available to students in other medical fields is advisable and justified because of the intensity of the different types of activities demanded of these students, i.e. lectures, seminars, practical or lab classes, clinic hours or internships.

The students' opinions indicate a high level of preparation of teaching materials and course organization. The popularity of distance learning may result from interactive access to the information contained in the online course (lesson, quiz, task, forum) or just multiple access to the course material resources. A student can repeatedly refer to the previously analyzed issue, learn it better, acquire knowledge on the topic, and independently test their knowledge on the subject (Douglas et al., 2004; Półljanowicz et al., 2009; Półljanowicz et al., 2010).

As a result of online learning, the university teacher has a new role as a kind of mentor. Using the prepared teaching materials (Figure 4) and tests (Figure 5), the teacher can conveniently check the students' knowledge. Both the teacher and student are able to see test results immediately (Rice, 2010).

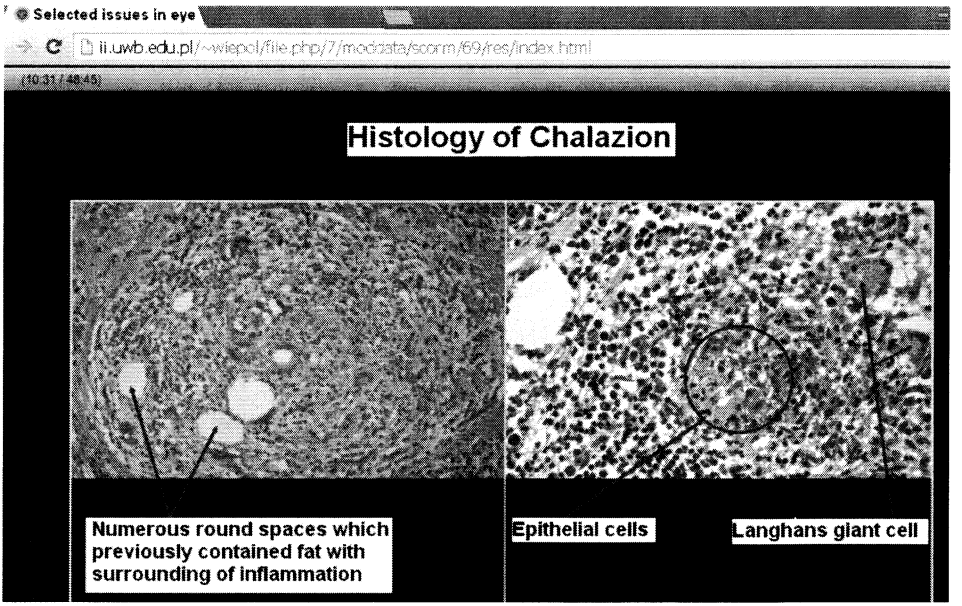


Figure 4. Screenshot of multimedia teaching materials on visual rehabilitation

Pilot studies conducted among students majoring in *computer sciences* at the School of Humanities and Journalism and *medicine* at the Medical University in Poznan in the academic year 2011/2012 to assess the relevance of e-learning materials and electronic examination of students (Roszak et al., 2013a; Roszak et al., 2013b) confirmed a high evaluation of e-learning methods and tools in the teaching process.

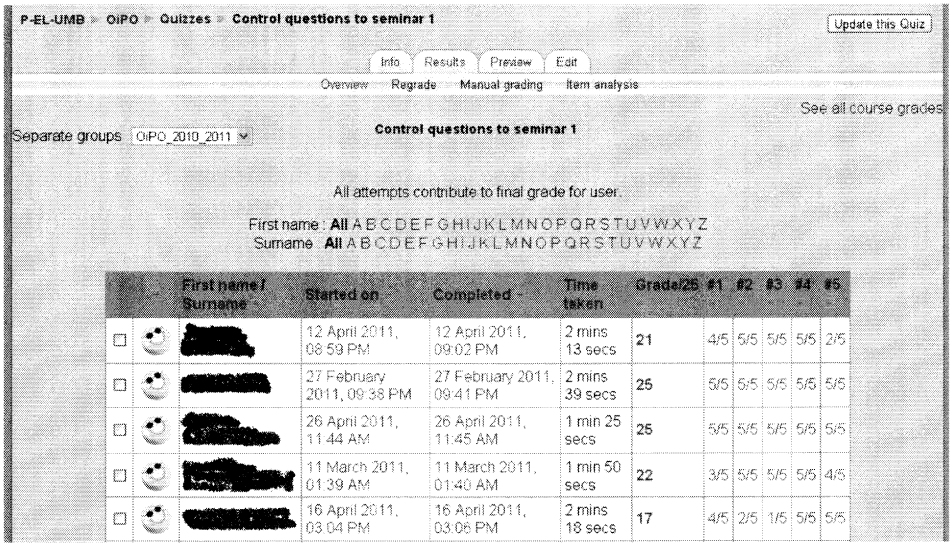


Figure 5. Screenshot of quiz assessing student knowledge in the course "Ophthalmology and ophthalmic nursing"

Studies performed at Maria Curie-Skłodowska University of Lublin on Logopedy with Audiology Faculty (in cooperation with the Institute of Physiology and Pathology of Hearing, Warsaw) confirmed that the majority of students evaluated e-learning systems positively (as very good – 64% and good – 31% of students), regarding quality and accessibility of didactic materials. Approximately 68% of students reported that e-learning is an effective method of education while nearly 86% reported that electronic knowledge assessment tests are a great advantage of distance learning. However, the major advantages of e-learning are: all day access to didactic materials (90%), saving time (98%), and individualization of the teaching process (71%). Moreover, according to academic teachers from the Institute of Physiology and Pathology of Hearing in Warsaw, e-learning allows time spent on educational activities to be reduced without a decrease in learning quality (Bombol-Lagha et al., 2012).

Our study included a group of 75 nursing students enrolled in the courses "Selected issues in visual rehabilitation" and "Ophthalmology and ophthalmic nursing" in the years 2009–2011 using e-learning and traditional learning methods. The positive results produced by a comparison of these groups confirmed the assumption that the use of e-learning would not worsen the state of professional knowledge, student satisfaction, or learning effectiveness compared with students attending traditional classes.

In connection with the development of Internet access and mobile devices, distance education is and will continue to become more and more popular and natural. Students enrolling in virtual courses are not bound to the place or time of their implementation. They can acquire knowledge at their own time and save time on commuting to the university. Virtual systems of consultation with the instructor (forums, FAQ, chat) do not leave the student alone in the jungle of available information. Students have more flexibility in content implementation; however, this is directly connected with more self-discipline, conscientiousness and responsibility on the part of the learners (Douglas et al., 2004; Selvi, 2010; Shroff et al., 2007; Wu et al., 2010). The next step to increase the attractiveness of distance learning is to personalize e-learning courses according to individual learning style. This may increase student satisfaction and knowledge gained.

Conclusions

The results obtained during this study allow us to conclude that e-learning is not inferior to traditional teaching methods for the courses "Selected issues in visual rehabilitation" and "Ophthalmology and ophthalmic nursing" for nursing majors.

The study and statistical analysis allow us to conclude that course organization was rated positively by 80% of the students in the e-learning groups and 89% in the traditional groups. 91% of the students in the e-learning groups and 88% in the traditional groups rated course preparation as well as availability and relevance of the didactic content positively. In addition, 86% of students in the e-learning groups and 83% in the traditional groups rated their acquisition of knowledge and skills positively.

The multimedia, e-learning teaching materials, available to students throughout the semester, enabled more flexible learning opportunities and preparation for the final exam, while simultaneously allowing students to widen and assimilate knowledge of ophthalmology and visual rehabilitation.

REFERENCES

Allan, B. (2007). Time to learn? E-learners' experiences of time in virtual learning communities. *Management Learning*, 38, 557–572.

Arbaugh, J. B., Desai, A., Rau, B., & Sridhar, B. S. (2010). A review of research on online and blended learning in the management disciplines: 1994–2009. *Organization Management Journal*, 7(1), 39–55.

Bombol-Lagha, M., & Śliwa, L. (2012). Wyniki wdrożenia formy blended learning na kierunku Logopedia z audiologią na Uniwersytecie Marii Curie-Skłodowskiej. *E-mentor*, 5(47), 57–61.

Bramley, P. (2001). *Ocena efektywności szkoleń.* Kraków: Dom Wydawniczy ABC.

Douglas, D. E., & van der Vyver, G. (2004). Effectiveness of e-learning course materials for learning database management systems: an experimental investigation. *Journal of Computer Information Systems*, 44(4), 41–48.

Piskurich, G. M. (2003). *The AMA Handbook of E-Learning, Effective Design, Implementation, and Technology Solutions.* New York: AMACOM.

Półjanowicz, W., Latosiewicz, R., Niewiński, A., & Milewski, R. (2009). E-learning in students education in Medical University of Bialystok. *Bio-Algorithms and Med-Systems*, 5(9), 111–115.

Półjanowicz, W., Latosiewicz, R., Kulesza-Brończyk, B., Piekut, K., Kalisz, A., Piechocka, D. I., & Terlikowski, S. J. (2010). Comparative analysis of e-learning and traditional teaching methods in the field of nursing in the Medical University of Bialystok. *The chosen aspects of woman and family's health*, 2, 94–104.

Rice, W. H. (2010). *Tworzenie serwisów e-learningowych z Moodle 1.9.* Gliwice: Helion.

Roszak, M., Kołodziejczak, B., Ren-Kurc, A., Kowalewski, W., & Bręborowicz, A. (2013a). Learning Content Development System (LCDS) jako narzędzie tworzenia materiałów powtórkowych. *E-mentor*, 1(48), 40–46.

Roszak, M., Kołodziejczak, B., Kowalewski, W., & Ren-Kurc, A. (2013b). Standard Question and Test Interoperability (QTI) – ewaluacja wiedzy studenta. *E-mentor*, 2(49), 35–40.

Selvi, K. (2010). Motivating factors in online courses. *Procedia – Social and Behavioral Sciences*, 2(2), 819–824.

Shroff, R. H., Vogel, D. R., Coombes, J., & Lee, F. (2007). Student E-Learning Intrinsic Motivation: A Qualitative Analysis. *Communications of the Association for Information Systems*, 19(12), 241–260.

Smith, G. G., Heindel, A. J. & Torres-Ayala, A. T. (2008). E-learning commodity or community: Disciplinary differences between online courses. *The Internet and Higher Education*, 11(1), 152–159.

Waćkowski, K., & Chmielewski, J. M. (2007). Rola standaryzacji platform w e-learningu. *E-mentor*, 2(19), 25–32.

Wu, W. C., & Hwang, L. Y. (2010). The effectiveness of e-larning for blended courses in colleges: a multi-level empirical study. *International Journal of Electronic Business Management*, 8(4), 301–310.

List of Reviewers for the Journal of Studies in Logic, Grammar and Rhetoric

1. Grzegorz Bancerek – Poland
2. Annalisa Bonomo – Italy
3. Tomasz Burzykowski – Belgium
4. Adam Czarnota – Australia
5. Anna Doliwa-Klepacka – Poland
6. M[a] Elena Gómez Parra – Spain
7. Małgorzata Gos – Poland
8. Piotr Hofmański – Poland
9. Honorata Jakuszko – Poland
10. Adam Jamróz – Poland
11. Ewa Kawalec – Poland
12. Pauline N. Kawamoto – Japan
13. Dariusz Kiełczewski – Poland
14. Jerzy Kopania – Poland
15. Ljubica Kordić – Croatia
16. Cezary Kosikowski – Poland
17. Marcin Koszowy – Poland
18. Wioletta Małgorzata Kowalska – Poland
19. Izabela Kraśnicka – Poland
20. Robert Kublikowski – Poland
21. Zbigniew Kuderowicz – Poland
22. Magdalena Kun Buczko – Poland
23. Monika Łodej – Poland
24. Witold Marciszewski – Poland
25. Roman Matuszewski – Poland
26. Marek Mazurkiewicz – Poland
27. Robert Milewski – Poland
28. Stanisław Zenon Mnich – Poland
29. Ewa Myrczek-Kadłubicka – Poland
30. Adam Naumowicz – Poland
31. Edward Oczeretko – Poland
32. Jerzy Pogonowski – Poland

33. Dorota Potocka – Poland
34. Anna Poznańska – Poland
35. Daniel Rabczenko – Poland
36. Slavian Radev – Bulgaria
37. Piotr Rudnicki – Canada
38. Christoph Schwarzweller – Germany
39. Aleksander Bogusław Stępkowski – Poland
40. Dariusz Surowik – Poland
41. Andrzej Szpak – Poland
42. Andrzej Trybulec – Poland
43. Kazimierz Trzęsicki – Poland
44. Josef Urban – Czech Republic
45. Freek Wiedijk – The Netherlands
46. Rafał Wojciechowski – Poland
47. Iwona Wrońska – Poland

ISBN 978-83-7431-392-6

ISSN 0860-150X