Graphical representation of the relationships between qualitative variables concerning the process of hospitalization in the gynaecological ward using correspondence analysis

Anna Justyna Milewska[1], Dorota Jankowska[1], Urszula Górska[1], Robert Milewski[1], Sławomir Wołczyński[2]

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

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

Abstract. Correspondence analysis is an exploratory method that allows for the analysis of categorical variables. Its aim is to graphically present the relationship between the variables shown in a contingency table, using points in two- or three-dimensional space, with a minimum loss of information about the diversity of rows and columns. Correspondence analysis developed from the 40's of the 20^{th} century, but the name in its current form gained popularity through the work of O. Hill, J.-P. Benzécrie and M. Greenacre, among others. Now, thanks to wide access to statistical packages, carrying out calculations does not create problems for researchers, and the interpretation of the results is simple and generally consistent with earlier intuitive assumptions. Correspondence analysis is used in many fields. The use of correspondence analysis for data describing the hospitalization process shows interesting relationships. It helped to depict issues such as an aging population, shorter hospital stay, or the migration of patients from other provinces.

Introduction

The development of technology causes that more and more often we are dealing with large databases containing many variables. Researchers often do not have full knowledge of the existence of multiple relationships and structures which are hidden in them. The development of modern methods of exploration and increasing performance of computers shortens the time needed to explore even the most complex and invisible at first sight relationships. More and more common is the use of data mining tools and techniques to reduce the variables. Their skilful use allows to explore a variety of databases containing information from different fields of science. It is also worth paying attention to the fact that discovery of dependencies is only the first step of the analysis. Thousands or even millions of variables generate the

ISBN 978–83–7431–336–0 ISSN 0860-150X

existence of hundreds of potential relationships. Selecting the most interesting and valuable can be labour-intensive as well as time consuming. Often their interpretation is cumbersome and requires deep knowledge of the field. Graphical methods designed specifically for the presentation of the results of individual analyzes may be helpful. Correspondence analysis gives this possibility. It allows the transparent and clear presentation of many relationships on one illustration.

Correspondence Analysis

Correspondence analysis is a technique that allows the examination of the relationship between qualitative variables (nominal and ordinal), which are common in medicine. The analysis of such data, which can be found in the contingency table, is started from the verification of the hypothesis showing no relationship between the characteristics using the χ^2 test. This procedure provides information only about the importance of the relationship between the variables. To indicate what is the nature of relations correspondence analysis can be used. On the other hand, it is an exploratory technique of reviewing large data sets. This tool enables the detection of associations between the two features with a graphical representation of the collected data [7]. This is done by creating the so-called correspondence map presenting relationships between variables. It therefore allows defining hypotheses, which can then be verified in a more formal way.

Correspondence analysis has developed since the 40's of the 20th century but the name in its current form gained popularity through the work of O. Hill, J.-P. Benzécrie and M. Greenacre, among others. Its idea is to create maps that graphically illustrate the contingency table which summarizes the collected data, where each row and each column is represented by one point. The main aim of this method is to show the set of points in the space of a maximum three dimensions, with a minimum loss of information on the diversity of rows and columns [27].

Correspondence analysis is carried out according to a certain schema [5]. It will be presented with a fictional example. [Tab. 1] presents information about the smoking status, depending on the level of physical activity which have been collected among 200 respondents.

The analysis of collected data begins with the creation of the correspondence matrix P. It is obtained by converting the contingency table in

Tab. 1. The relationship between smoking status and physical activity level of respondents

Smoking status	Physical activity level			
	(1) none	(2) seldom	(3) regularly	TOTAL
(a) none	12	22	35	**69**
(b) light	6	14	18	**38**
(c) medium	25	16	4	**45**
(d) heavy	28	20	0	**48**
TOTAL	**71**	**72**	**57**	**200**

a matrix of relative frequencies (dividing the number in each cell by the total number):

$$P = \begin{bmatrix} 0.060 & 0.110 & 0.175 \\ 0.030 & 0.070 & 0.090 \\ 0.125 & 0.080 & 0.020 \\ 0.140 & 0.100 & 0.000 \end{bmatrix}$$

The next step is to determine the matrix of row profiles (and column profiles). It is created by dividing the relative frequencies in each row (column) of correspondence matrix by the sum of relative frequencies across the row (column):

$$W = \begin{bmatrix} 0.17 & 0.32 & 0.51 \\ 0.16 & 0.37 & 0.47 \\ 0.56 & 0.36 & 0.09 \\ 0.58 & 0.32 & 0.00 \end{bmatrix} \qquad K = \begin{bmatrix} 0.17 & 0.31 & 0.61 \\ 0.08 & 0.19 & 0.32 \\ 0.35 & 0.22 & 0.07 \\ 0.39 & 0.28 & 0.00 \end{bmatrix}$$

Obtained row (column) profiles can be graphically represented in the space generated by the columns (rows) of the correspondence matrix. Particular frequencies in the profile can be understood as other coordinates in the considered space. In the analyzed example there will be illustrated row profiles in 3-coordinates system corresponding to physical activity categories [Fig. 1]. An analogous presentation of the column profiles is not possible due to the four-dimensional space generated by the rows of the matrix P.

Summed relative frequencies in each row of correspondence table is called the row mass (similarly for columns). This measure provides information about the significance of the rank of each row. An important parameter for the correspondence analysis is the average row (column) profile. It is

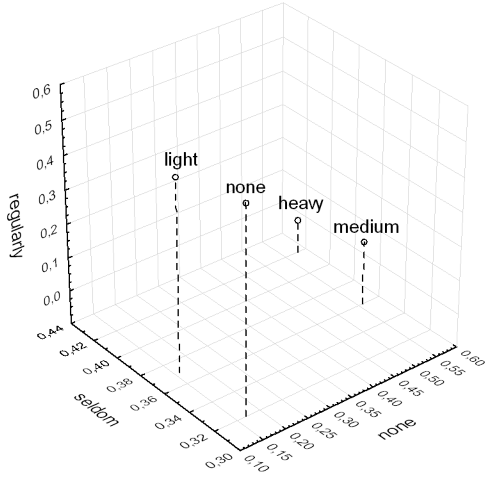


Fig. 1. Points represent row profiles in the space generated by the three levels of physical activity

obtained as a part of a summary row (column) of the contingency matrix in the total number of investigated population. In this example, we have:

– average row profile:

(1)	(2)	(3)
0.335	0.36	0.285

– average column profile:

(a)	0.345
(b)	0.190
(c)	0.225
(d)	0.240

It is obvious that individual coordinates of average row (column) profile are the masses of corresponding columns (rows). The average profile is the center of mass of analyzed profiles.

Comparing and analyzing the profiles is done by determining the distance between them using χ^2 metric:

$$\chi^2 = d^2(p_i, p_i') = \sum_{i=1}^{k} \frac{(p_i - p_i')^2}{\bar{p}_i}$$

where:

k – dimensions of profiles

p_i, p_i' – subsequent profiles coordinates, where the distance is calculated

$\bar{p}_i$ – subsequent average profiles coordinates

It is worth noting that in this way, there can be only compare profiles of different categories of the same variable. Profiles scattering around the average profile is determined by the so-called total inertia which is calculated from the formula:

$$\Lambda^2 = \sum_{j=1}^{w} m_j d_j^2(p_j, \bar{p})$$

where:

w – number of rows (columns)
m_j – mass of j-row (column)
$\bar{p}$ – average row (column) profile
$d_j^2(p_j, \bar{p})$ – distance from row (column) profile to average profile measured with a χ^2 metric

Inertia for rows and for columns are equal. Its maximum value is $\min(w, k) - 1$ where w and k are the number of rows and columns of the contingency table respectively. The small value of inertia indicates slight differences between the profiles and the average profile. In such a situation we have in the analyzed example, because $\Lambda^2 = 0.299505$. In addition, the smaller the inertia is, the smaller the chance of occurrence of statistically significant relationships between the studied characteristics is. This follows from the relation:

$$\chi^2 = \Lambda^2 n$$

where:

χ^2 – value of the χ^2 test statistic
n – the total sample size

Another key step in the algorithm, by which correspondence analysis runs, is the projection of rows and columns profiles matrix for up to a three-dimensional space. At the same time the largest part of the information on the diversity of rows and columns should be maintained. This step is a response to a problem with graphical presentation of large number of dimensions. This is done by the method called singular value decomposition (SVD) [22, 27]. It runs in the following way:

- Symmetrical standardization of correspondence matrix P:

$$P = [p_{ji}] \rightarrow A = [a_{ji}]$$

where: $a_{ji} = \dfrac{p_{ji} - p_{j.}p_{.i}}{\sqrt{p_{j.}p_{.i}}}$

- Presentation of formed matrix A as a product of the following three matrixes:

$$A_{wxk} = U_{wxr} \cdot D_{\lambda_{rxr}} \cdot V_{rxk}^T$$

where:

w, k – numbers of rows and columns respectively
A – matrix with rank equal to r
D_λ – matrix that has the diagonal with nonzero singular values AA^T in non-decreasing order
U – matrix whose columns are the orthonormal eigenvectors corresponding to eigenvalues $\lambda_1^2, \lambda_2^2, \ldots$ of AA^T matrix
V – matrix whose columns are the orthonormal eigenvectors corresponding to eigenvalues $\lambda_1^2, \lambda_2^2, \ldots$ of A^TA matrix

In such distribution the columns of U are orthonormal basis for the columns of the matrix A. Thus they form the principal axes of subspace projection of category stored in columns. Similarly, the columns of the matrix V gives orthonormal basis for the transposed rows of the matrix A and hence generate a subspace projection of the principal axes of categories stored in rows.

There is the following relationship between the eigenvalues of the matrix A^TA and the total inertia:

$$\Lambda^2 = \sum_{i=1}^{r} \lambda_i^2$$

where:

r – $\operatorname{rank} A$, $r = \min(w, k) - 1$
w, k – number of rows and columns in A matrix respectively

This relationship allows to determine how much of the total inertia is explained by the i-th factor, thanks to the determination of the quotient $\frac{\lambda_i^2}{\Lambda^2} \cdot 100\%$. What more it gives the possibility to choose the dimension of space, which will take place in the projection of the analyzed correspondence matrix. It is the minimum n such that:

$$\sum_{i=1}^{n} \frac{\lambda_i^2}{\Lambda^2} \cdot 100\% \geq m$$

where m is a predetermined level of explanation of the total inertia (eg 75% or 80%). Such proceeding provides the best selection of space that will represent the considered correspondence matrix in the fullest manner and will ensure the least loss of information resulting from the reduction of dimension. In the analyzed example $\lambda_1^2 = 0.296857$. It allows to explain 99.12% of inertia. In the case of two-dimensions space it reproduces full value of inertia, as is shown in [Tab. 2].

Tab. 2. The cumulative percentage of inertia explained by the individual eigenvalues

Number of dimensions	Total inertia = 0.29951; $\chi^2 = 59.901$; $df = 6$; $p = 0.0000$				
	Singular Values	Eigenvalues	Percent of inertia	Cumulative percent	χ^2
1	0.544845	0.296857	99.11556	99.11556	59.37130
2	0.051468	0.002649	0.88444	100.00	0.52979

After selecting the dimension of the projection space remains only the computation of new coordinates of rows and columns profiles [4–5]:

– principal coordinates for row profiles $F = D_r^{-1} \cdot U \cdot D_\lambda$

– principal coordinates for column profiles $G = D_c^{-1} \cdot V \cdot D_\lambda$

where:

D_r – diagonal matrix in which coefficients lying on the diagonal are sums of frequencies of appropriate rows of the correspondence matrix P

D_c – diagonal matrix in which coefficients lying on the diagonal are sums of frequencies of appropriate columns of the correspondence matrix P

To create an n-dimensional map of correspondence, the first n columns of the matrix F is used to determine the coordinates of rows and analogously the first n columns of the matrix G to determine the coordinates of columns [8]. Such map in one coordinate system illustrates the best representations of row and column profiles, despite the fact that these profiles exist in different spaces [9]. It allows to analyze and interpret the distance between the point imaging one of the categories of analyzed features and the center of projection or between other points representing the various categories of the same variable. It also provides the ability to draw conclusions about the coexistence of different categories of comparable qualitative features. Row and column profiles lying close to each other create a combinations of category occurring together more often than would result from independence of variables.

2-dimensional map of the correspondence for the analyzed example is shown on [Fig. 2]. From the arrangement of points we may conclude that people practicing sport regularly usually do not smoke. Moreover, among medium smokers are those who show lack of physical activity.

To assess the accuracy of our correspondence map a parameter is used which refers to the quality of individual points. It is defined as the quotient

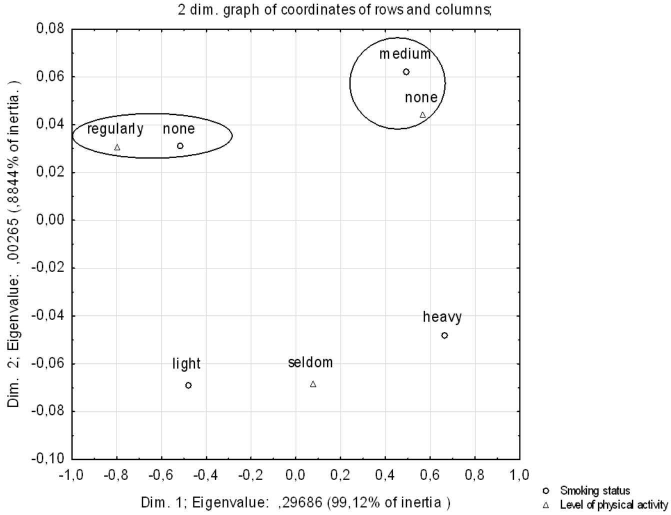


Fig. 2. Correspondence map, which represents the relationship between smoking status and physical activity level of respondents

of the square of the distance between a certain point and the beginning of the selected coordinate system, by the square of the distance between this point and beginning of the coordinate system, with the largest number of dimensions in the considered situation. The quality of representation of row or column profiles in low-dimensional space is perfect when the measure is equal to 1 [27].

Application of correspondence analysis in scientific research

The method discussed in the current work is used to analyze qualitative data. Many scientific disciplines describe the objects of their interest with a nominal scale. It is worth especially noting that medical data describe cases of patients who have multiple units and many of the symptoms of disease at the same time. We can describe trials and populations of this type using contingency tables. However, it appears that many labels of the individual variables (e.g. identified disease entities) and a large number of characteristics lead to the creation of many tables. Their complexity makes it impossible to observe hidden dependencies. Correspondence analysis allows to look at the same time at many aspects. In addition, by reducing dimensions and graphical presentation using two- or three-dimensional space, it is easier to interpret complex relationships. It is worth noting that the structures detected by correspondence analysis take into account several variables at the same time.

Correspondence analysis is used in many fields. Literature provides many examples of the use of this method in marketing research [3, 10, 26]. As one of them can replace the analysis of professional activity by age and gender. The use of correspondence analysis allows to show on one illustration several distinctive groups. It can be observed that age categories are closely related to work full-time, and which are particularly exposed to the problem of unemployment. An interesting idea is also the use of correspondence analysis to test students' knowledge about teaching and learning. Discovering patterns makes it possible to design better and more efficient educational programs [1]. There is also an attempt to use this tool in describing political preferences, as well as in the analysis of Parliament members speeches [2]. It allows to present transcripts containing over 100 million words using correspondence maps. Thanks to them we can see how different the language used in the Sejm and Senate is, as well as in subsequent years, the vocabulary change.

Attention should also be paid to the use of correspondence analysis in medical research. One of the more known works using the method with the correspondence maps shows the relationship between the type of headache and the age of patients [7]. The author also presents another interesting dependencies, such as the relationship between personality type and belonging to different diagnostic groups. In [25] correspondence analysis was used to identify the relationship between the incidence of the back pain symptom and a variety of factors, such as age, BMI, smoking, drinking alcohol. Correspondence maps may present which of these variables has a significant relationship with the occurrence of the disease.

The use of correspondence analysis in medicine on example of hospitalization on the gynaecological ward

Saving information in the process of hospitalization we get a typical example of a large data set. Extremely useful in such cases are data mining methods. Their application in medicine is becoming more popular; such as the use of artificial neural networks to predict the outcome of infertility treatment [15, 20], the use of feature selection algorithms to reduce dimensionality of the original data set of women treated for infertility [16–17], or cluster analysis in the process of recruitment for competitive swimming [24]. Data mining methods are useful even in such issues as the analysis of microarray data [18], or the analysis of the information gathered in the process of deep sequencing [6].

In this paper we present the use of correspondence analysis on data from hospital cards from the Department of Gynaecology. In this case, the database contains more than eight thousand cases and several features describing the treatment. For the analysis were selected 14 qualitative variables such as: age group, the NHF (National Health Fund) district, cause of hospitalization, categorized length of hospital stay, etc. The aim of the analysis was performed through graphical representation of the relationships between qualitative variables referring to the hospitalization process. To perform the analysis, statistical package Statistica 10.0, StatSoft was used.

Because of assumptions of the analysis, numerical data were grouped. Ranges describing the age are left-closed, and the length of treatment are presented by class: 1 day, 2 days, ..., 6 days, 1 week, 1–2 weeks, 2–3 weeks, more than three weeks. Disease entities were coded according to the international statistical classification of diseases and health problems ICD-10 [4]. Patients were also classified into five groups of diagnoses, created on the basis of the main causes of hospitalization: gynaecology, obstetrics, infertility, insemination, IVF ET. Place of residence was defined within the NHF district, which covers the patient.

The first pair of analyzed variables were age and the NHF district. Presentation of the relationship of these variables is presented in two-dimensional space, which explained 98.8% of the total inertia [Fig. 3].

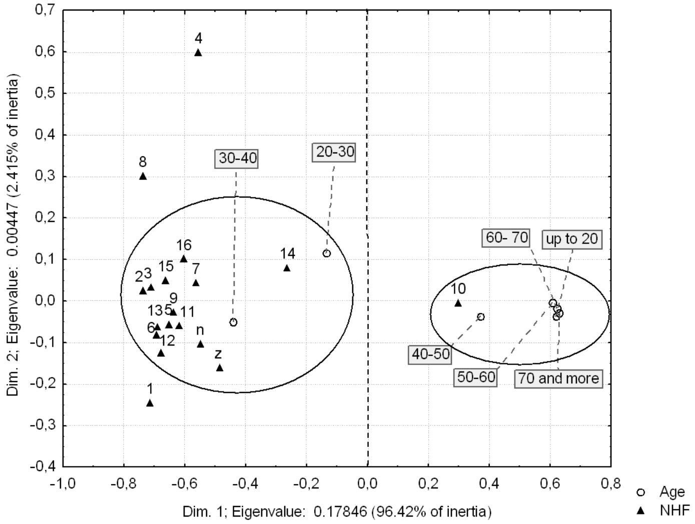


Fig. 3. Graphical representation of the relationship of residence (NHF) and the age of patients

The first dimension clearly distinguished NHF = 10 (Podlaskie Province), which is located on the right of the center axis in relation to the other NHF-s which are on the left side. In the figure we selected two areas of focus points. We note that the older patients (over 40 years) are residents of the 10th NHF, but younger patients (20–40 years old) live in areas outside the Podlaskie Province. Note that the interpretation of the clusters on the maps present the coexistence correspondence characteristics without giving information about the strength of this relationship. Observations obtained in this way need to be confirmed by other statistical methods.

In case of the presentation relationship between length of hospitalization and the place of residence (NHF), two-dimensional correspondence map explains 98.5% of the total inertia [Fig. 4]. The first dimension clearly differentiates between short and long hospitalizations. In this figure, we see the same as in the previous: two areas of focus. This time, a short, one-day hospitalizations coexisted with the place of residence other than the Podlaskie Province and longer hospitalizations were related to patients from the Podlaskie Province (NHF = 10).

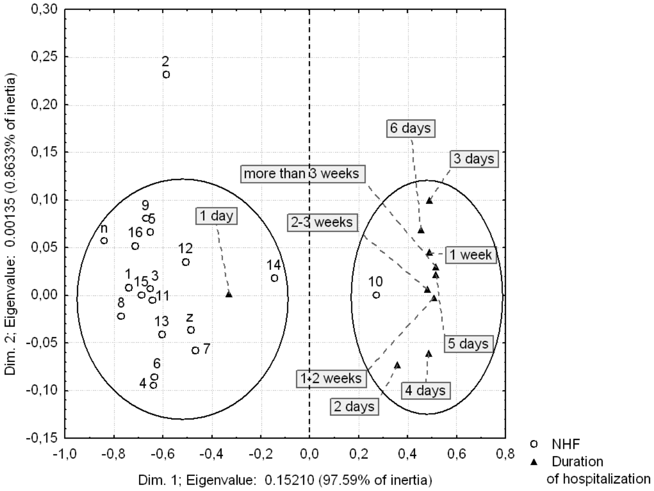


Fig. 4. Graphical representation of the relationship of residence (NHF) and the length of hospitalization

The next graph shows the relationship between characteristics: length of hospitalization and age [Fig. 5]. In this case, the two-dimensional space has been used, which explains 88.3% of the total inertia. We can see three interesting concentrations: short hospitalization (1 day) on young patients (20–40 years old), over one week hospitalization on 60–70 aged women, the longest hospital stays on the oldest patients. These dependencies seem to be natural, because the older the age – the worse the health, and the treatment is often complicated by comorbidities [11].

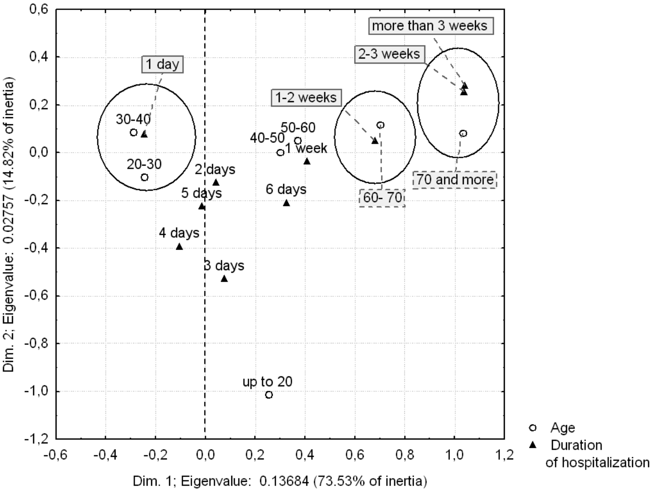


Fig. 5. Graphical representation of the relationship between age of patients and length of hospitalization

The main cause of hospitalization on the gynaecological ward is very diverse. For this reason, patients were assigned to the more homogeneous diagnoses subgroups. Another figure [Fig. 6] clearly shows the relationship between the two groups: the cause of treatment and the age of patients.

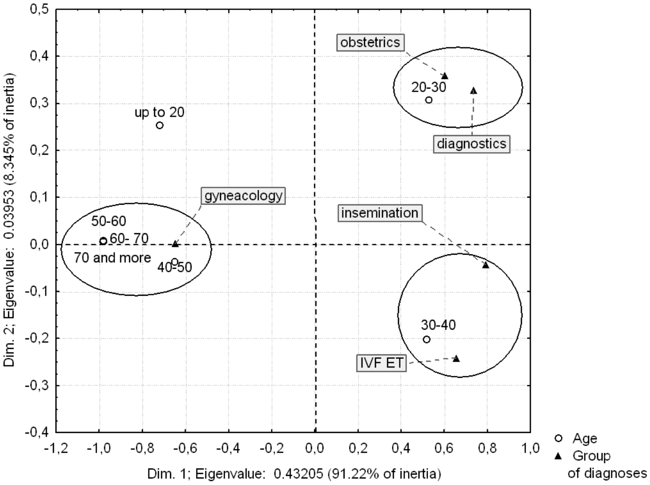


Fig. 6. Graphical representation of the relationship between the age of patients and the cause of hospitalization

Two-dimensional correspondence map explains almost all of the general inertia – 99.56%. The first dimension is stretched by gynaecology on the left side of the axis and the other groups on the right side. The second dimension differentiates patients treated by IVF ET from patients hospitalized with obstetric and infertility diagnostic reasons. The graph shows three areas of focus. The first tells us that the gynaecological patients were over 40. The second presents that the diagnosis and obstetric problems relate to patients aged 20–30 years, and the third focus shows that women aged 30–40 years were treated using IVF ET method or using insemination procedure. This is the age group in which a largest number of women decide to use the assisted reproduction methods in infertility treatment. It is limited from below by the continuously progressive deposition of reproductive decisions, and from above by drastically decreasing the effectiveness of treatment in women over 40 years old [19].

Further analysis of grouped causes of hospitalization showed a clear dependence on the area [Fig. 7]. The two-dimensional space explains 99.17% of the total inertia. The first dimension clearly indicates the focus of the NHF 10 and gynaecological-obstetric causes, which are located on the right of the axis center. On the other hand, we see that the diagnosis and treatment of infertility caused migration of patients from other provinces. These observations were analyzed with traditional statistical methods and presented in [14], as well as developed by basket analysis in [12].

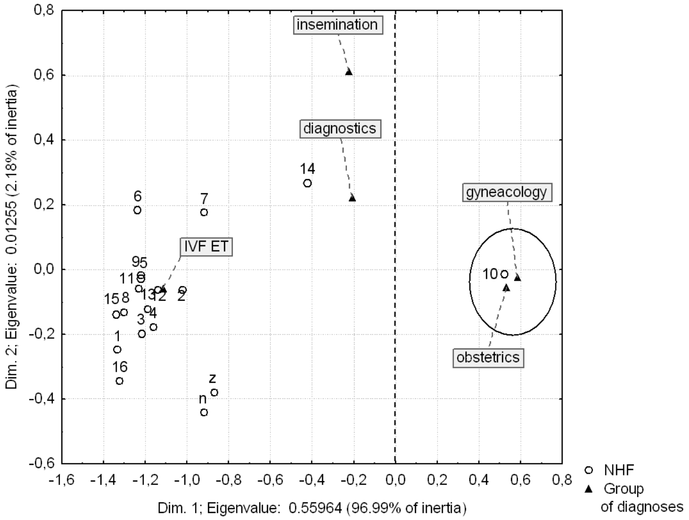


Fig. 7. Graphical representation of the relationship between the residence of patients and the cause of hospitalization

Another analysis of the grouped causes of hospitalization and the length of treatment is shown in a two-dimensional correspondence map, which explains 95.25% of the total inertia [Fig. 8]. The first dimension differentiates one-day hospitalizations from longer hospitalizations. Three areas of focus show the duration of hospitalization depending on the cause of hospitalization. One-day hospitalizations concern the treatment of infertility. Obstetric problems require a 2–5 day stay in the hospital, but the gynaecological problems need the longest hospitalization.

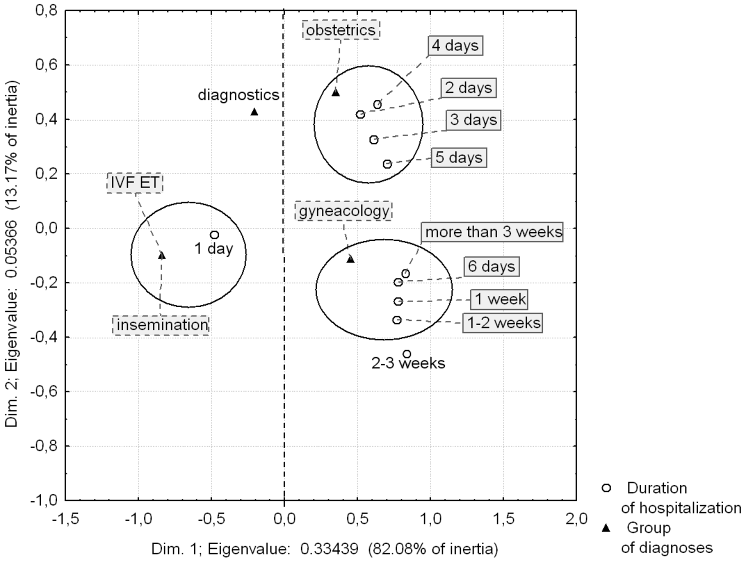


Fig. 8. A graphical representation of relationship between the length of the treatment and cause of hospitalization

Analyzed database concerns patients hospitalized in 1996, 2000 and 2004. The choice of particular years resulted from the subsequent stages of the reform of the health care system. The year 1996 refers to the period in which the health service was still financed directly from the state budget. The year 2000 was the first full year, in which the Health Management Organization was established. The year 2004 was the first full year after creating the institution of the National Health Fund.

Subsequent correspondence maps present relationship between the hospitalization year, the length of treatment and the age of patients. The analysis was performed only for patients whose main cause of hospitalization were gynaecological problems. In both cases the two dimensional space explained 100% of the total inertia, as this is the maximum dimension of the space.

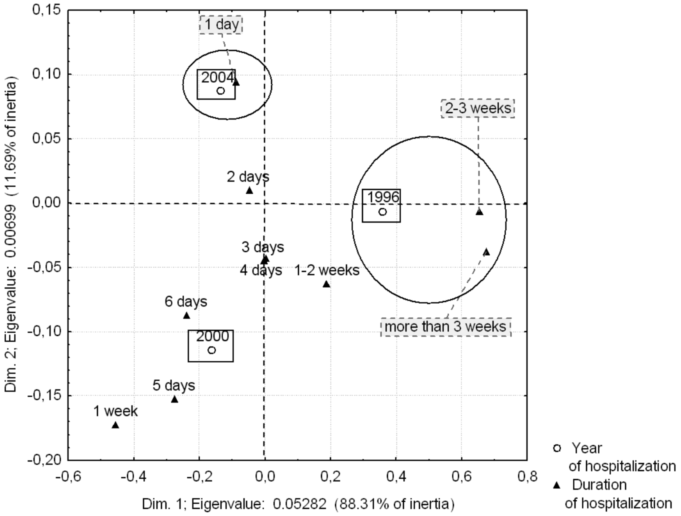


Fig. 9. Graphical representation of the relationship between the length of hospitalization and the year of treatment

[Fig. 9] shows the relationship of particular years, and different lengths of hospitalization. We note that the first dimension differentiates shorter (1–7 days) from longer hospitalizations. The year 1996 lies on the right side of the axis in contrast to the years 2000 and 2004 which are on the left side of axis. We can see that the longest hospitalizations (2–3 weeks, more than 3 weeks) coexists with the year 1996, while the one-day hospitalizations – with the year 2004. This relationship is consistent with the changes which have taken place in the health care system under the influence of the reform. The requirement of the financial account was introduced, which has resulted in a significant shortening of duration of hospitalization times [28]. The development of new treatment methods has also contributed to the reduction of hospitalization time [21, 23].

Another correspondence map shows the relationship between age of patients and the year of hospitalization [Fig. 10]. Over the analyzed years, we can see the effects of aging. It is shown, that the oldest group of patients (more than 70 years old) is closest to the year 2004. The results of the demographic changes observed on the basis of these data were presented in [13].

For patients hospitalized for gynaecological reasons the analysis of coexistence of treatment length and disease entities was performed, which are the cause of hospitalization [Fig. 11]. Two-dimensional space explains 84.2% of the total inertia. The first dimension differentiates the shorter and longer

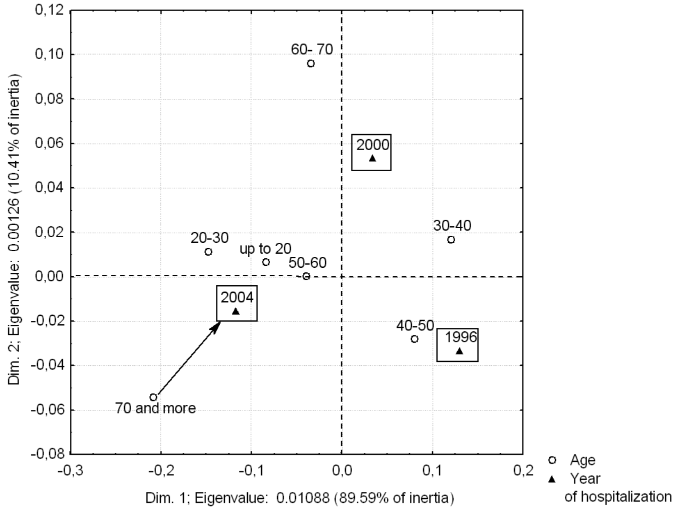


Fig. 10. Graphical representation of the relationship between the year of hospitalization and the age of patients

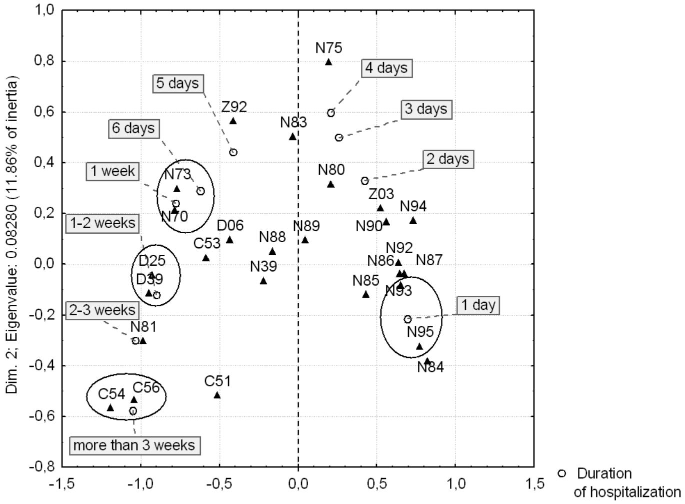


Fig. 11. Graphical representation of the relationship between length of hospitalization and the major diagnosis

hospitalizations. We can distinguish several clusters on the correspondence map:

- one-day hospitalizations relate to patients with diagnoses that require short treatment because of bleeding: N95, N93,

– 6–7 days hospitalizations relate to treatment of inflammation: N70, N73,
– 1–2 weeks hospitalizations relate to patients requiring surgical treatment: D39, D25,
– the longest hospitalizations (more than 2 weeks) relate to patients with cancer: C54, C56.

Conclusions

Correspondence analysis is a method being used increasingly in scientific research. Thanks to access to statistical packages performing calculations does not create problems to researchers, and the interpretation of the results is simple and generally consistent with prior intuition about the expected relationship between the features. Correspondence analysis provides researchers with many possibilities. First, it allows for the discovery of structures and patterns hidden in large databases. Secondly, it can reduce the variables to facilitate interpretation (taking into account the fact that some of the variables are redundant) and creates clear images, called correspondence maps that show related variables in the form of clusters. Observation of the cluster areas often shows the possibility of obtaining interesting results on the basis of observed data in subgroups.

The use of correspondence analysis to data concerning the process of hospitalization showed many interesting relationships. The analysis has outlined problems such as the population aging, shorter hospital stay, or the migration of patients from other provinces. However, the next step is to carry out further analysis using traditional statistical methods.

REFERENCES

[1] Askell-Williams H., Lawson M. J., A Correspondence Analysis of Child-Care Students' and Medical Students' Knowledge about Teaching and Learning, International Education Journal, 5 (2), pp. 176–204, 2004.

[2] Biecek P., Poszukiwanie struktury danych na przykładzie analizy korespondencji, pp. 8–13, Sulejów, 2010 (odczyt „Do czego to się przydaje")

[3] Bendixen M., A Practical Guide to the Use of Correspondence Analysis in Marketing Research, Marketing Bulletin, 14, 2003.

[4] Bartkowski S., Międzynarodowa statystyczna klasyfikacja chorób i problemów zdrowotnych, Rewizja dziesiąta, Tom 1, Vesalius, Kraków, 2006.

[5] Clausen S. E., Applied Correspondence Analysis: An Introduction, SAGE Publications, Thousand Oaks, CA, 1998.

[6] Czerniecki J., Wołczyński S., Deep sequencing – a new method and new requirements of gene expression analysis, Studies in Logic, Grammar and Rhetoric, 25 (38), pp. 41–47, 2011.

[7] Greenacre M., Correspondence analysis in medical research, Statistical Methods in Medical Research, 1, 1992.

[8] Greenacre M., The use of correspondence analysis in the exploration of health survey data, Documentos de Trabajo, 5, 2002.

[9] Greenacre M., Jorg B., Correspondence Analysis in Social Sciences, Academic Press, New York, London, 1994.

[10] Hoffman D. L., Franke G. R., Correspondence analysis: Graphical representation of categorical data in marketing research, Journal of Marketing Research, XXIII, pp. 213–227, 1986.

[11] Leowski J., Polityka zdrowotna a zdrowie publiczne, Wydawnictwo CeDeWu, Warszawa, 2004.

[12] Milewska A. J., Górska U., Jankowska D., et al., The use of the basket analysis in a research of the process of hospitalization in the gynecological ward, Studies in Logic, Grammar and Rhetoric, 25 (38), pp. 83–98, 2011.

[13] Milewska A. J., Milewski R., Mnich S. Z., Karpińska M., Wołczyński S., Wpływ starzenia się społeczeństwa na strukturę chorobowości w ginekologii, Przegląd Menopauzalny, 5, pp. 330–334, 2010.

[14] Milewska A. J., Milewski R., Wołczyński S., Analiza zjawiska migracji pacjentów na Podlasie na przykładzie Kliniki Ginekologii, Polityka Zdrowotna, 7, pp. 71–76, 2008.

[15] Milewski R., Jamiołkowski J., Milewska A. J. et al., 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), pp. 900–906, 2009.

[16] Milewski R., Malinowski P., Milewska A. J., et al., Nearest neighbor concept in the study of IVF ICSI/ET treatment effectiveness, Studies in Logic, Grammar and Rhetoric, 25 (38), pp. 49–57, 2011.

[17] Milewski R., Malinowski P., Milewska A. J., et al., The usage of margin-based feature selection algorithm in IVF ICSI/ET data analysis, Studies in Logic, Grammar and Rhetoric, 21 (34), pp. 35–46, 2010.

[18] Milewski R., Milewska A. J., Czerniecki J., Oligonucleotide microarrays in biomedical sciences – the use and data analysis, Studies in Logic, Grammar and Rhetoric, 25 (38), pp. 29–40, 2011.

[19] Milewski R., Milewska A. J., Domitrz J., et al., In vitro fertilization ICSI/ET in women over 40, Przegląd Menopauzalny, 2 (36), pp. 85–90, 2008.

[20] Milewski R., Milewska A. J., Jamiołkowski J., et al., The statistical module for the system of electronic registration of information about patients treated for infertility using the IVF ICSI/ET method, Studies in Logic, Grammar and Rhetoric, 21 (34), pp. 119–127, 2010.

[21] Neis K. J., Brandner P., Wagner S., Laparoskopische Operationsverfahren in der Gynakologie, Gynakologe, 39, 87–104, 2006.

[22] Nenadić O., Greenacre M., Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package, Journal of Statistical Software, 20 (3), 2007.

[23] Obrzut B., Granice laparoskopii operacyjnej, Ginekologia Praktyczna, 15 (4), pp. 7–11, 2007.

[24] Roczniok R., Zastosowanie analizy skupień w procesie naboru do pływania sportowego, Zeszyty Metodyczno-Naukowe AWF, Katowice, 21, pp. 167–175, 2006.

[25] Sanittham K., Plubin B., Bookkamana P., Correspondence analysis in back pain symptom of employees in the Factory, Lampang Province, Thailand.

[26] Stanimir A., Wykorzystanie analizy korespondencji w badaniach marketingowych, Zastosowanie metod statystycznych w badaniach marketingowych III, StatSoft Polska, pp. 337–346, 2008.

[27] Stanisz A., Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, Tom 3. Analizy wielowymiarowe, StatSoft Polska, Kraków, 2006.

[28] Stelmach W., Olas-Janaszkiewicz E., Maniecka-Bryła I., et al., Średni czas pobytu i wykorzystanie łóżek na oddziałach chirurgii ogólnej w okresie funkcjonowania różnych systemów finansowania ochrony zdrowia, Wiadomości Lekarskie, 60 (7–8), pp. 341–345, 2007.

The practicality of any nonparametric statistical procedure should be confirmed thoroughly with regard to the data distribution under study

Maciej Górkiewicz[1], **Agnieszka Gniadek**[2]

[1] Department of Epidemiology and Population Research, Jagiellonian University Medical College, Poland

[2] Department of Medical and Environmental Nursing, Faculty of Health Sciences, Jagiellonian University Medical College, Poland

Abstract. The present study was motivated by pilot research aimed to examine the aptness of the anti-fungal everyday activity in the hospital settings. The number of uncovered fungi colonies per cubic meter of air was chosen here as the crucial indicator of the quality. The empirical probability distribution functions of this indicator at various hospital's wards showed great variety of their shapes. Nevertheless, from the practical point of view, preferably the comparisons between these functions should be expressed in terms of the mean values. Therefore, here the practical problem arose: how to avoid inappropriate choice of a nonparametric statistical test for equivalence of the mean values given random data sets. In the present study, the review of the nonparametric methods was limited to the most popular ones only: the Box-Cox transformation, the Mann-Whitney rank sum test, and the log-rank approach. The more advanced formal considerations were omitted. The limitations of the Box-Cox transformation and the Mann-Whitney rank sum test were explained with clear examples based on the artificial samples. Practical criterions, helpful to avoid common pitfalls and misunderstandings were recommended. The advantages and the weaknesses of the log-rank approach were demonstrated basing on the real-life data sets.

Introduction

Any statistical procedure was founded on a number of assumptions regarding not only formal features of the data under investigation, but also the attributes of the anticipated area of application of the results of the statistical analysis. In the real world, the departures from these ideal assumptions are unavoidable [1–3]. Thus, in applied research the practicality of the results has two faces. First, the all way of analyses should be correct from a pure mathematical point of view, and then, from the practical perspective, the real meaning of the results should correspond with the anticipated area of their use.

ISBN 978–83–7431–336–0 ISSN 0860-150X

The parametric procedures were constructed under common assumption that the data samples were drawn from a normal distribution. For that reason, if the empirical distribution of the data distinctly differs from ideal normal distribution, then some nonparametric procedure is usually applied. Unfortunately, many non-statisticians are wrongly convinced, that in such a situation, the use of the most suitable nonparametric "ersatz" doesn't change the essence of the conclusions from the calculations [4–5]. For instance, numerous non-statisticians wrongly believe that parametric Student t-test, applied to the data after Box-Cox transformation, gives conclusion regarding data before transformation. Many other non-statisticians accept as true the imprecise supposition that the Mann-Whitney rank sum test examines the relation between medians. On the other side, it is known that the parametric procedures are fairly robust to the moderate departures from normality [6], at least the obtained results need some moderate corrections with regard to estimated skew and kurtosis coefficients [7]. In consequence, other numerous non-statisticians, dealing with the reasonable number of random data, do not take into account the use of any nonparametric procedure. They wrongly believe that for any kind of distributions, the parametric procedures at any circumstances generate reliable conclusions, at least with regard to the relation between the mean values.

All above mistaken beliefs can lead to misinterpretation of the data under examination, subsequently to wrong practical decisions, and as a final result, to depreciation of the statistical methodology in the public opinion. In the literature, the discussed potential causes of this undesirable phenomenon include:

(i) pressure 'publish or perish' on candidate researchers [8];

(ii) wishful thinking instead the critical one [9–10];

(iii) not-user-friendly style and confusing terminology applied in the statistical textbooks, like the misleading idiom 'distribution free' frequently used with regard to some nonparametric procedure [11–12];

(iv) numerous silent assumptions in the statistical instructions, that are obvious for the statisticians, but rather hard to reveal for the others [13–14].

Several fundamental changes in statistical training and practice are recommended in literature, with a general purpose of changing for the better of this situation. It was suggested to make a greater emphasis on the philosophy underlying the statistical methodology [15], with special focus on a common-sense approach and on the possible pitfalls and misinterpretations [16–18]. The wide-ranging use of the exploratory data analysis is postulated, first before starting the usual confirmatory data analysis, with the aim to reveal unexpected or misleading patterns in the data and to foster hypothesis development and refinement, and then after this, with the aim to help one interpret the obtained results [19]. Modern tools for the visual exploration of large databases create an opportunity to discover the outliers and clusters within the data [20] and to confirm the fit to supposed distribution [21], avoiding the false impressions caused by inspection limited to the traditional histograms only [22].

The present study corresponds in general with all the above cited ideas concerning the desirable improvements in the current research practice, but the sphere of interest was strictly limited here to a practical problem: how to assess the aptness of anti-fungal everyday prevention. Consequently, the rest of this paper was organized as follows. First, the problem how to assess the quality of the everyday anti-fungal clinical practice is discussed. This section can be omitted by person non interested in the clinical problems. The review of the nonparametric methods started with the most popular transformations aimed to diminish the influence of non-normality. The limitation of the Box-Cox method was shown on the exemplary data, and some other methods were briefly characterized on the base of literature. Then, the Wilcoxon-Mann-Whitney rank-sum test was examined in aspect to the question if the results of this test can be interpreted in terms of medians and usual arithmetic means. In the next section, the descriptive statistics for the motivating practical question was made in two tables including the main characteristic of the $2 \cdot (45 + 20) = 130$ samples obtained in patients' rooms at two hospital wards. Finally, the classical log-rank methodology was applied in the study. More advanced permutation tests and stochastic modelling procedures were only discussed as the possible subject for further investigations.

How to assess the aptness of clinical everyday anti-fungal practice?

In clinical settings, the primary mode of acquiring a mould infection is inhalation of room air polluted with fungal spore-loaded dust [23]. Therefore, the concentration of the fungi in the air, that is the number of uncovered fungi colonies per cubic meter of air, at various hospital wards was commonly acknowledged as the crucial indicator of the quality of the overall anti-fungal activity [24]. The appropriate isolation of patients from harmful aerosols might be achieved only with combined use of several means [25]. Moreover, the final effectiveness of usual everyday antifungal activity must be under permanent inspection with the aim to put into action routines, like wearing of filtering masks [26]. This is avery challenging problem, from practical as well as theoretical point of view, because the anti-fungal activity represents only a component of the whole very complex system named – good clinical practice. Therefore, there arises a great difficulty in defining elements of the intervention. It may be impossible to single out which particular parts are effective, since one component may not work without another [27–28].

Therefore, in our best knowledge, the problem, how to utilize the measurements of the fungi concentration at the given hospital with the aim to improve the current system of the anti-fungal activity, up to now haven't any acknowledged practicable solution. The general idea of the our solution to this problem is as follows. The anti-fungal practice, carried out at each particular hospital, should be considered primarily at the whole, as a complex intervention, observing the *primum non nocere* principle [29–31]. Initially, it can be advantageous to consider the relatively simple question: are the frequencies of departures of the concentration of the fungi in the air over appropriate levels in the patients' rooms at the hospital under examination are at least as low as the frequencies in the analogous wards at other good hospitals? After that, before starting with any serious modifications, one should get the reliable answers to the two main initial questions:

(i) are the mean values of the concentration of the fungi in the air in various wards at the hospital under examination are at least as low as that concentration measured with use of the comparable methodology in the analogous wards at the other good hospitals?

(ii) are the mean values of the concentration of the fungi in the wards at the hospital under examination decreased gradually from the maximal value in the entry room to appropriate values inside, at the patients' rooms?

In our investigations of the fungi concentration in the air at various hospitals wards, it was proved that there wasn't the significant correlation between measurements made in adjacent moments in the sequences of morning-evening measurements. Therefore, the series of the measurements made at the same place can be considered as random samples of the uncorrelated data [32]. Nevertheless, there occurred two interacted difficulties: the underlying distributions are far from normal, so the use of the nonparametric methodology should be considered; nonetheless, from the practical point of view, the conclusions must refer to mean values (that is to the expected values) directly, without any subsidiary substitution with some other attribute, like median, geometric mean, harmonic mean, trimmed mean and so on. In our investigations data sets corresponded to the Weibull distribution. Therefore, the search for the most acceptable method can be limited to the review of the parametric and nonparametric test for comparing means of the Weibull populations [33]. Nevertheless it seems to be more appropriate not to ignore the most popular universal procedures, like Box-Cox transformation and the Wilcoxon-Mann-Whitney rank-sum test.

Transformations aimed to diminish the influence of non-normality

The real-life samples examined in applied research fairly often showed attributes atypical for a normal distribution, like coefficients of skewness and kurtosis far from zero, outliers, heavy tails. Since the consequences of non-normality for test statistics are difficult to investigate, many studies suggested the use of transformation procedures developed for specific forms of non-normality. The Box-Cox method was developed for distributions intermediate between the normal and the log-normal distributions, with the aim to restore normality of the data. The Box-Cox transformation was defined with formulas (1), (2), (3).

$$Y = (X^C - 1)/C; \quad \text{for } C \neq 0; \tag{1}$$

either

$$Y = \ln(X); \qquad \text{for } C = 0; \tag{2}$$

$$\text{optimal}(C) = C|\text{optimal}(J) \tag{3}$$

were:

X – transformed variable before Box-Cox transformation;
Y – transformed variable after Box-Cox transformation;
C – power parameter;
J – criterion of optimality, assumed correspondingly to anticipated parametric procedure of the further analyses.

[Tab. 1] includes two examples. In the first example, the populations numbered 1, 2, and 3 in [Tab. 1], had the same mean values of variable Y, equal to $m_y = 0$, but different variances V_y. In result, the mean values of variable X in these populations occurred manifestly different, accordingly to known formula (4). In the second example, the exemplary populations numbered 4, 5, and 6 in the [Tab. 1], had the same mean values of variable X, equal to $m_x = 1.65$, but different variances V_y. In result, the mean values of variable Y in these populations occurred manifestly different, accordingly to known formula (4).

$$\ln(m_x) = m_y + V_y/2 \tag{4}$$

where:

m_x – mean value of variable X with log-normal distribution;
m_y, V_y – mean value and variance of variable $Y = \ln(X)$ with normal distribution.

Tab. 1. Relationship between mean values of random variable X with log-normal distribution and variable Y = ln(X) in some exemplary populations

population	1	2	3	4	5	6
$m_y = \ln(Me_x)$	0	0	0	– 0.5	0	0.25
V_y	0.5	1	2	2	1	0.5
m_x	1.28	1.65	2.72	1.65	1.65	1.65

m_x, Me_x – mean value and median of variable X with log-normal distribution;
m_y, V_y – mean value and variance of variable $Y = \ln(X)$ with normal distribution.

It is easy to notice, directly from definition of the log-normal distribution [34], that if the variable X in formula (2) has the ideal log-normal distribution, then the variable Y after Box-Cox transformation obtains the ideal normal distribution. Therefore, without any doubts any parametric procedure may be applied to examine mean values of the variable Y among some populations under examination. Moreover, usually there aren't any serious doubts that the ANOVA parametric procedure can be applied with respect to mean values of the variable Y in random samples drawn from exemplary populations numbered 1, 2, and 3 in [Tab. 1], because ANOVA is fairly robust to moderate heteroscedasticity [6], so use of any known counterpart to ANOVA [35–37], will be to no purpose here. The question is, if the proved relationship between mean values of the variable $Y = \ln(X)$ may be adopted to the actually interested relationship between mean values of the variable X, $X > 0$.

The criterions for the appropriate use of the Box-Cox method can be easily deduced from the examples considered in [Tab. 1]. It was demonstrated that the normality of the variable after Box-Cox transformation doesn't give sufficient reason to make an inference about mean values of this variable before Box-Cox transformation, basing on the relationship among mean values of this variable after Box-Cox transformation, as proved with some parametric test. Thus, except for the normality, the homogeneity of the variances of the variable after Box-Cox transformation must be proved too. If the request of the homogeneity isn't fulfilled than the conclusions relate to some other averages, e.g. to the geometric mean for transformation $Y = \ln(X)$, but not to usual arithmetic mean values. Guo and Luh [38–39], discussed several other transformation, more suitable for non-normal distributions that are affected by heavy tails or outliers. There also the results of a parametric procedure applied correctly to the transformed samples, can be related to some special averages of the variable before transformation, but generally not to the usual arithmetic mean values.

The general conclusion from the above review can be summarised as follows. It can be very advantageous to apply a transformation approach dealing with unusual distributions at pilot studies and exploratory data analyses, because it is a quick and easily computable method. However, it should be implemented with great caution.

Wilcoxon-Mann-Whitney rank-sum test

Wilcoxon-Mann-Whitney rank-sum test, or shortly, Mann-Whitney test, pertains to some statistical comparison of two separate populations given two independent random samples, but it is usually thought as the most reliable nonparametric alternative for 2-sample Student t-test in situations where the data appear to arise from non-normal distributions. For this reason, it is easily available at almost all popular statistical packages. The concise introduction to this test with very intuitive graphics one can find in [40].

Let $X = x$ denotes a single random number drawn from the first population, and $Y = y$ denotes a single random number drawn independently from the other separate population. The Mann-Whitney test has been performed as the non-parametric alternative to the parametric Student t-test, but in essence it examines the null hypothesis (5).

$$H_0: \qquad \Pr(x < y) = \Pr(x > 0) \tag{5}$$

under restriction:

$$\Pr(x = y) = 0 \tag{6}$$

where random variables X and Y are both measured at least on an ordinal scale.

Therefore, many non-statisticians are wrongly convinced that for any distributions of variables X and Y the null hypothesis (5) is perfectly equiponderant with the null hypothesis (7) on equivalence of the medians of these variables, and consequently, that the null hypothesis (5) is perfectly equiponderant with the null hypothesis (8) on equivalence of the mean values of these variables, at least for the symmetrical (non-skewed) distributions.

$$H_0': \qquad Me_x = Me_y \tag{7}$$

$$H_0'': \qquad m_x = m_y \tag{8}$$

with silent (wrong) justification: because hypothesis (5) holds $\Pr(x < y) = \Pr(x > 0)$;
where m_x, Me_x, m_y, Me_y – mean values and medians of variables X, Y respectively.

Both of the above convictions are generally wrong, and in practice often lead either to disadvantageous decisions or at best to absurd conclusions in a particular matter under study. The last statement can be easily supported by a simple example shown in [Tab. 2]. In this table the three exemplary cases were constructed in such a way that the restriction (6) was satisfied in each case. Then, the symmetrical distribution of the variable X, and the shape of the symmetrical distribution of the variable Y, both remain the same at all three exemplary cases, but the distribution of Y is shifted to right in case 2, and is shifted to left in case 3. Therefore, the medians of these distributions, initially the same in case 1, are manifestly different from the others, as well in case 2 as in case 3. In other words, the hypotheses (7) and (8) are fulfilled in case 1 only, but they are manifestly violated in the both two remaining cases. Nevertheless, it is easy to notice that in all three cases the hypothesis (5) is evidently satisfied, because probabilities $\Pr(x < y) = \Pr(x > 0) = \frac{1}{2}$ didn't change from case to case.

Tab. 2. Mann-Whitney test for some exemplary pair of the symmetrical distributions

case	shift	$f(X) > 0$	$m_x = Me_x$	$f(Y) > 0$	$m_y = Me_y$	p_{MW}
1	0	$-1 < X < +1$	0	$Y < -2 \cdot 10^6$ or $Y > +2 \cdot 10^6$	0	0.5
2	$+10^6$	$-1 < X < +1$	0	$Y < -1 \cdot 10^6$ or $Y > +3 \cdot 10^6$	$+10^6$	0.5
3	-10^6	$-1 < X < +1$	0	$Y < -3 \cdot 10^6$ or $Y > +1 \cdot 10^6$	-10^6	0.5

$f(X)$, $f(Y)$ – density of symmetrical distribution of the variable X and Y respectively;
m_x, Me_x, m_y, Me_y – mean values and medians of variables X, Y respectively;
p_{MW} – ideal (expected) value of the significance of the Mann-Whitney rank-sum test.

Let us consider a somewhat more down-to-earth situation of drawing samples from the unknown distributions assumed in the above [Tab. 2]. For instance, let X and Y be an anticipated incomes, expressed in \$, from two kind of businesses. It is easy to notice that in each case in [Tab. 2] the probability $\Pr(y < \min(x)) = \Pr(y > \max(x)) = \frac{1}{2}$. Therefore, if the N random values of the variable Y is drawn independently one from other, then the probability that exactly $N/2$ values of Y will occur beyond $\min(x) = -1$, equal to probability that exactly $N/2$ values of Y will occur over $\max(x) = +1$, will depend only from N, and for instance, for the moderate dimension of a sample, $N = 64$, in average only on one occasion in the 20 experiments the sample will not divided into two exactly the same parts, first one of $N/2 = 32$ x's beyond $\min(x) = -1$, and second one of $N/2 = 32$ x's over $\max(x) = +1$. If the random sample of x's has there a dimension also equal to $N = 32$, then the sums of the ranks of the y's and x's occur the same, equal to $32 \cdot (1 + 128) = 32 \cdot (33 + 96) = 4.128$ as well for variable Y as for variable X. A 'naïve' researcher, believing without any doubts that Mann-Whitney test can be directly related to medians, at almost each time will find a reason for an interpretation that an expected balance equal to zero dollars doesn't differ significantly from an expected income equal to million dollars (case 2 in [Tab. 2]), or that it doesn't differ significantly from an expected loss equal to million dollars (case 3 in [Tab. 2]).

The exemplary distributions of Y, in each case considered in [Tab. 2], are divided into two parts, with a gap of density $f(Y) = 0$ between them. It should be noted that a quite similar manifestation of the relations between hypotheses (5), (7), and (8), can be modelled without this gap, also with distribution with a single mode, under restrictions that ratio $f(Y)/f(X)$ has a pattern either low-high-low or high-low-high, so it isn't an example for the Simpson's paradox [41].

The general conclusion from the above considerations can be summarised as follows. The Mann-Whitney test can be applied without any hesitation to practical problems that can be expressed in terms of the hypothesis (5), without any serious focus on the medians or on the usual arithmetic means. If the problem under study must be related at least to medians, like hypothesis (7), then the ratio $f(Y)/f(X)$ should be proved with respect to its monotonicity. If the problem under study must be related to the usual arithmetic means, then additionally both distributions, the $f(Y)$ and the $f(X)$ should be sincerely symmetrical. As to the last case, it is known, that for symmetrical non-normal distributions, the differences in the power between Student t test and Mann-Witney rank-sum test are so small that the choice is immaterial for practical purposes [42–43].

Motivating example

[Tab. 3] showed the descriptive statistics of the concentration of fungi in the air in patients' rooms at two hospital wards under study, as measured in the morning and in the evening during five consecutive winter days. It is easy to notice the evident departure from normality, with a mean values rather far from the medians, and relative great coefficients as well as for skew as for kurtosis. Therefore, the parametric tests, like ANOVA and Student t-test, seem to be inappropriate here. From practical reasons, the results of comparisons should be related to usual arithmetic means. Thus, the most popular counterparts, like transformations and Mann-Whitney rank-sum test, also seems to be quite inappropriate to apply in the matter.

In such a case, the simulation approach seems to be most suitable [44–45], in particular with respect to easy available on-line calculators [46–47]. On the other hand, in this study the log-rank plots supported supposition that the distribution ofconcentration of the fungi are consistent with Weibull distribution, see [Tab. 4]. Consequently, with the aim to make comparisons between average concentrations of fungi at the different times and sites, the log-rank approach was applied.

Tab. 3. Descriptive statistics of the concentration of fungi in air in the patient's bedrooms during five consecutive days

ID	ward	time	N	mean	SD	median	min.	max	skew	kurtosis
1	HP	morning	45	35.3	47.9	15	0	195	1.78	2.72
2	HP	evening	45	25.0	48.8	0	0	235	2.79	8.26
3	BO	morning	20	73.5	320.5	0	0	1435	4.47	20.00
4	BO	evening	20	17.0	73.7	0	0	330	4.47	20.00

The Weibull distribution is a continuous probability distribution. The probability function $F(X)$ of a three parameter Weibull random variable X is given with the formula (9), where X_0 is the shift parameter, A is the scale parameter, and B is the shape parameter.

$$F(X) = 1 - \exp(-((X - X_0)/A)^B); \quad X \geq X_0; \quad A > 0; \quad B > 0. \qquad (9)$$

It is easy to notice that for the $B = 1$ the equation (9) represents the exponential distribution of random variable X with the mean value equal to the sum of the shift and the scale parameters, equal to $X_0 + A$. Moreover, for the $B = 1$ the log-rank transformation of the X leads to the linear regression between the $F(X)$ and the log-rank of X, given random sample of Xs. It seems, that on the explanatory analyses stage [19], it can be quite enough to put all trust on the log-rank probability plots methodology [21], avoiding the false impressions caused by inspection limited to the traditional histograms only [2, 22].

Tab. 4. The log-rank plots for concentration of the fungi in the patients bedrooms

ID	Ward	N	Log-rank equation	R^2
1	HP	90	Log-rank $= 16.58N_C + 32.824$	0.992
2	BO	40	Log-rank $= 1.510N_C + 34.508$	0.838
2	BO	40	Log-rank $= -0.67N_C^2 + 4.19N_C + 33.16$	0.978

N_c – cumulativenumber of cases, related to successive log-rank of concentration of fungi;
R^2 – coefficient of determination of the estimated log-rank equation.

The log-rank test compares area under curve (AUC) under estimates of the hazard functions for two or more groups along with all diapason from the first to the last observed event [48]. For the Weibull distribution, defined with formula (9), the hazard function is defined with formula (10).

$$h(x) = f(X)/(1 - F(X)) = (B/A) \cdot ((X - X_0)/A)^{B-1} \tag{10}$$

where: X_0 is the shift parameter, A is the scale parameter, and B is the shape parameter.

For the shape parameter equal to $B = 1$ the hazard function is stable and it is equal to reciprocal of a mean value, so the log-rang test can be considered as an exact alternative for other statistical tests applied for testing equality of the mean values [33].

Results obtained with the log-rank test

The log-rank test for differences between morning and evening concentration separately at the each wards under study, showed that both differences were insignificant here, $p = 0.26$ for the HP ward, and $p = 0.68$ for the BO ward. Therefore, it was decided to join the morning and evening data.

[Tab. 5] showed the significance equal to p(chi2) = 0.05; that is just on the borderline. This undecided result of the long-rank test corresponded to the result of comparing the frequencies of the departures over norm 50 CFU/m^3 for fungi concentration in patients' rooms, see [Tab. 6].

Tab. 5. The log-rank test for difference between mean concentration of the fungi

frequency	HP	BO	p(chi2)
observed	90	40	0.05
expected	99.6	30.4	

Tab. 6. Frequency of departures over norm for fungi concentration

ward	time	N	$N\| < 50$	$N\| > 50$	$\%\| > 50$	95%CI	
HP	morning	45	34	11	24.4%	14.2%	38.9%
HP	evening	45	38	7	15.6%	7.5%	29.2%
BO	morning	20	19	1	5.0%	0.0%	25.7%
BO	evening	20	19	1	5.0%	0.0%	25.7%

$N| > 50$; $\%| > 50$ – number (percentage) of departures over norm in total number of N events;

95%CI – confidence interval for $\%| > 50$.

Conclusions and discussion

In this study the highly skewed data from the pilot study on concentration of the fungi in the hospital wards create basis to illustrate the way for searching after appropriate nonparametric statistical procedure aimed to make comparisons between mean values. In general, the thesis that the usefulness of any nonparametric statistical procedures should be confirmed thoroughly with regard to the data distribution, was confirmed with the use of the extremely simple, but clear examples of the inappropriate understanding the aftermaths of the Box-Cox transformation to normality, and then the essence of Mann-Whitney rank-sum test. The log-rank approach was examined using the real-life data sets, obtained at two chosen hospitals. It was demonstrated, once again, that the Box-Cox transformation method can lead to erroneous conclusion even with respect to ideal log-normal populations. Then, it was demonstrated, also once again, that the results of the Mann-Whitney rank-sum test often doesn't correspond to relations neither between medians nor between means. The practical criterions, helpful to recognize the situations allowing to conclude on relations between mean values, were recommended. The real-life data sets, obtained at two chosen hospitals, corresponded to the Weibull distribution. Therefore, the log-rank approach was the primary candidate for the most acceptable method in the matter. For the shape parameter near to $B = 1$ the conclusions from the log-rank test are valid directly to relation between the mean values in the populations under investigation. Moreover, the log-rank plots and log-rank test applied together may provide a deeper insight into essentials of the investigated relationship, than the simple comparisons of the mean values only. For this reasons, the log-rang plots and the log-rank test applied jointly, seem to be quite sufficient to provide trustworthy conclusions about the aptness of the anti-fungal everyday activity in the hospital settings. Thus, the search for the most acceptable method of statistical analysis was shut in this pilot study on the log-rank approach. In case of need, other methodology should be applied with the aim to disclose the causes of the detected insufficiency of the anti-fungal practice, but it lies beyond the scope of this paper.

The present study, as each pilot study, has its typical limitations. Only two hospital wards, and the moderate number of data, $N = 2 \cdot (45 + 20) = 130$, were investigated. Nevertheless, with regard to planning further investigations, this pilot study gave rather decisive support to estimate the sufficient number of the data at each ward under study, near to these applied here, between $N = 2 \cdot 20 = 40$ and $N = 2 \cdot 45 = 90$.

Acknowledgement

The initial version of this paper was presented at the annual summer seminar of the Polish National Group of the International Society for Clinical Biostatistics (ISCB) in Krakow (Poland) on 29.06.2012. Authors are sincerely grateful to all participants in the discussion for providing their inspiring comments to our presentation, what helped us to make this paper much concise and persuasive. However, all views expressed are our.

REFERENCES

[1] Martin M. A, Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties, Computational Statistics & Data Analysis, 51, pp. 6321–6342, 2007.

[2] Callaert H., Nonparametric hypotheses for the two-sample location problem, Journal of Statistics Education, 7 (2), 1999.

[3] Lee E. T., Desu M. M., Gehan E. A., A Monte-Carlo study of the power of some two-sample tests, Biometrika, 62, pp. 425–532, 1975.

[4] Batanero C., Godino J.D., Vallecillos A., et al., Errors and difficulties in understanding elementary statistical concepts, International Journal of Mathematical Education in Science and Technology, 25 (4), pp. 527–547, 1996.

[5] Castro A. E., Vanhoof S. S., Van den Noortgate W., et al., How confident are students in their misconceptions about hypothesis tests?, Journal of Statistics Education, 17 (2), 2009.

[6] Tan W. Y., Tabatabai M. A., Some Monte Carlo studies on the comparison of several means under heteroscedasticity and robustness with respect to departure from normality, Biometrical Journal, 28 (7), pp. 801–814, 1986.

[7] Cressie N. A. C., Whitford H. J., How to use the two sample t-test, Biometrical Journal, 2, pp. 131–148, 1986.

[8] Altman D., Egger M., Gotzsche P., et al., The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies, Plods Med., 4 (10), pp. 296, 2007.

[9] Tyszka T., Pułapki psychologiczne, Psychologia biznesu dla menedżerów, Akademia Leona Koźmińskiego w Warszawie, Warszawa, 2012.

[10] Smallbone T., Quinton S., Increasing business students' confidence in questioning the validity and reliability of their research, Electronic Journal of Business Research Methods, 2 (2), pp. 153–162, 2004.

[11] Lavy L., Mashiach-Eizenberg M., The interplay between spoken language and informal definitions of statistical concepts, Journal of Statistics Education, 17 (1), 2009.

[12] Lovett M. C., Greenhouse J. B., Applying cognitive theory to statistics instruction, The American Statistician, 54 (3), pp. 196–206, 2000.

[13] Chance B. L., Components of statistical thinking and implications for instruction and assessment, Journal of Statistics Education, 10 (3), 2002.

[14] Brewer J. K., Behavioral statistics textbooks: source of myths and misconceptions?, Journal of Educational Statistics, 10 (3), pp. 252–268, 1985.

[15] Krzanowski W., Statistical principles and techniques in scientific and social research, Oxford University Press, New York, 2007.

[16] Barr J., Gould M., Joffe A., Pitfalls in the interpretation of multivariable models in the critical care literature, Chest, 127 (1), pp. 411–412, 2005.

[17] Good P. I., Hardin J. W., Common errors in statistics (and how to avoid them), Wiley Interscience, 2003.

[18] Campbell M., Machin D., Medical statistics. A commonsense approach, John Wiley & Sons, England, 1999.

[19] Behrens J. T., Principles and procedures of exploratory data analysis, Psychological Methods, 2 (2), pp. 131–160, 1997.

[20] Marchette D. J., Jeffrey L., Solka J. L., Using data images for outlier detection, Computational Statistics & Data Analysis, 43 (4), pp. 541–552, 2003.

[21] ReliaSoft Corporation, Probability plotting papers. ©1996–2006. ReliaSoft Corporation, http://www.weibull.com/GPaper/index.htm.

[22] von Hippel P. T., Mean, median, and skew: correcting a textbook rule, Journal of Statistics Education, 13 (2), 2005.

[23] Kelman B., Robbins C., Swenson L., et al., Risk from inhaled mycotoxins in indoor office and residential environments, International Journal of Toxicology, 23 (1), pp. 3–10, 2004.

[24] Maschmeyer G., Prevention of mould infections, Journal of Antimicrobial Chemotherapy, 63, (Suppl. 1), pp. i27–i30, 2009.

[25] Bodey G. P., Freireich E. J., Influence of high-efficiency particulate air filtration on mortality and fungal infection: a rebuttal, The Journal of Infectious Diseases, 194, pp. 1621–1622, 2006.

[26] Pawińska A., Mikrobiologiczne monitorowanie środowiska szpitalnego. In: Profilaktyka zakażeń szpitalnych – bezpieczeństwo środowiska szpitalnego, A. Pawińska, (Ed.), pp. 57–87, α-medicapress, Bielsko-Biała, Poland, 2011.

[27] Cimoca G., A simple algorithm for comparing hospital units efficiency, Appl Med Inform, 8 (1–2), pp. 3–7, 2001.

[28] Campbell M., Fitzpatrick R., Haines A., et al., Framework for design and evaluation of complex interventions to improve health, British Medical Journal, 321 (7262), pp. 694–696, 2000.

[29] Gniadek A., Cytotoxicity of Aspergillus fungi as a potential infectious threat. In: Insight and Control of Infectious Disease in Global Scenario, edited by. Priti Kumar Roy, InTech Rijeka, Croatia, pp. 231–248, 2012.

[30] Gniadek A., Macura A. B., Intensive care unit environment contamination with fungi, Advances in Medical Sciences, 57, pp. 283–287, 2007.

[31] Gniadek A., Skawińska M., Szczypczyk M., et al., Stosowanie klimatyzacji a występowanie grzybów w powietrzu sal bloku operacyjnego, Mikologia Lekarska, 12 (1), pp. 31–36, 2005.

[32] Knoth S., Schmid W., Monitoring the mean and the variance of a stationary process, Statistica Neerlandica, 56 (1), pp. 77–100, 2002.

[33] Watthanacheewakul L., Comparisons of power of parametric and nonparametric test for testing means of several weibull populations, Proceedings of the International MultiConference of Engineers and Computer Scientists IMECS Hong-Kong, vol. II, pp. 1534–1538, 2011.

[34] StatSoft Inc. Electronic Textbook StatSoft, Glossary, Item: Lognormal Distribution. © Copyright StatSoft, Inc., 1984–2011. http://www.statsoft.com/textbook/statistics-glossary/w/?button=0#LognormalDistribution.

[35] Chmiel I., Górkiewicz M., The bootstrap and multiple comparisons procedures as remedy on doubts about correctness of ANOVA results, Applied Medical Informatics, 30 (1), pp. 9–15, 2012.

[36] Górkiewicz M., Using propensity score with receiver operating characteristics (ROC) and bootstrap to evaluate effect size in observational studies, Biocybernetics and Biomedical Engineering, 29 (4), pp. 41–61, 2009.

[37] Van Der Laan P., Verdooren L. R., Classical analysis of variance methods and nonparametric counterpart, Biometrical Journal, 29 (6), pp. 635–655, 1987.

[38] Guo J. H., Luh W. M., Transformation works for non-normality? On one-sample transformation trimmed t methods, British Journal of Mathematical and Statistical Psychology, 54, pp. 227–236, 2001.

[39] Guo J. H., Luh W. M., An invertible transformation two-sample trimmed t statistic under heterogeneity and nonnormality, Statistics and Probability Letters, 49, pp. 1–7, 2000.

[40] Bellera C. A., Julien M., Hanley J. A., Normal approximations to the distributions of the Wilcoxon statistics: Accurate to What N? Graphical Insights, Journal of Statistics Education, 18 (2), 2010.

[41] Bereziewicz W., Górkiewicz M., Jak dużo a priori w a posteriori: poznanie naukowe z zastosowaniem metod statystyki, Cogitatum, 2, pp. 1–9, 2012.

[42] Zimmerman D. W., Comparative power of Student t test and Mann-Whitney U test for unequal sample sizes and variances, Journal of Experimental Education, 55, pp. 171–174, 1987.

[43] Hilgers R., On the Wilcoxon-Mann-Whitney-test as nonparametric analogue an extension of t-test, Biometrical Journal, 24 (1), pp. 3–15, 1982.

[44] Davidson R., MacKinnon J.G., Bootstrap tests: how many bootstraps?, Econometric Rev, 19, pp. 55–68, 2000.

[45] Martis M. S., Validation of simulation based models: A theoretical Outlook, The Electronic Journal of Business Research Methods, 4 (1), pp. 39–46, 2006.

[46] Aksenov S., Confidence Intervals by Bootstrap, Wolfram Research Inc., 2002. http://library.wolfram.com/infocenter/MathSource/4272/.

[47] Siniksaran R., BootStrapPackage: A Package of Bootstrap Algorithms for Mean, Simple Linear Regression Models, and Correlation Coefficient, Wolfram Research Inc., 2001. http://library.wolfram.com/infocenter/MathSource/815/.

[48] Zhou M., Log-rank Test: When does it Fail – and how to fix it, 2006. http://www.ms.uky.edu/∼mai/research/LogRank2006.pdf.

Weighted clustering and ROC analysis in assessment of the quality of life in patients with chronic heart failure

Aleksander Owczarek[1], **Bożena Szyguła-Jurkiewicz**[2], **Michał Cogiel**[3], **Damian Grzechca**[4]

[1] Division of Statistics, Medical University of Silesia, Poland
[2] III Department of Cardiology, Medical University of Silesia, Poland
[3] Student Scientific Group, Medical University of Silesia, Poland
[4] Department of Electronics, Silesian Technical University in Gliwice, Poland

Abstract. Chronic heart failure is increasingly prevalent in population and has a significant impact on the length as well as the quality of patients' life. In Polish population there are no norms for the SF-36 test to assess the health related quality of life (HRQoL). Weighted k-means algorithm has been used to divide the population into 2 groups with better and worse quality of life and then cut-off points have been calculated based on the ROC curves analysis. Vitality has been the best discriminating factor. Poor quality of life was related with higher risk of depression development, MACE (major adverse cardiac event) occurrence and worse clinical parameters.

Introduction

Chronic heart failure (CHF) is associated with high mortality and morbidity regardless of the development in pharmacological treatment [1]. Several risk factors for major cardiac adverse events (MACE: sudden death, hospitalization due to exacerbation of CHF) have been already identified: elevated creatinine plasma level, age, female gender, NYHA class, left and right ventricular function, the peak exercise oxygen reuptake test and brain natriuretic peptide plasma level [2–4]. Depression is another very important risk factor (as it turned out recently). It is highly prevalent in this group of patients and may bias patients' reports of their Health Related Quality of Life (HRQoL) [5]. Such patients are not only more likely to develop depression, but once depressed, they are more likely to experience deteriorating heart disease, need repeated procedures, or die due to MACE.

The HRQoL, which was found to be very important in the assessment of CHF progression and treatment results, may be evaluated with The Short Form (36) Health Survey [6]. Results obtained in such a survey of patient's

ISBN 978–83–7431–336–0 ISSN 0860-150X

health are compared to cut-off points (different for various populations) and patient's well-being and its changes during the treatment are estimated. However, there are no cut-off points marked in the Polish population. These which are in the test instruction refer to American population [6]. In view of great social, cultural and economical differences between Polish and American population, referring the SF-36 results to these norms would not reflect the reality. This makes it difficult to evaluate outcomes yielded with the SF-36 test.

The purpose of the present study was to: 1) assess cut-off points for scales of the SF-36 test, 2) find which scales of the test best discriminate groups according to the HRQoL, 3) compare group with better and worst HRQoL regarding clinical parameters, depression and MACE occurrence.

Material and methods

Material

One hundred and ninety three consecutive patients with chronic systolic heart failure were included in the prospective study. Detailed (medical) description of patient treatment as well as measured clinical parameters one may find in our previous paper [7]. *Inclusion criteria*: 1. symptoms of systolic heart failure for at least 2 years; 2. increased LV end-diastolic diameter (LVEDD > 57 mm) and reduced LV ejection fraction (LVEF $< 45\%$) shown by the ECG; 3. 5-year or longer history of hypertension before the onset of heart failure symptoms (documented at least 2 episodes of systolic blood pressure ≥ 140 mmHg and/or diastolic blood pressure ≥ 90 mmHg); 4. lack of significant ($> 30\%$) narrowing in coronary arteries indicated by the coronary angiogram. *Exclusion criteria*: 1. confirmed coronary artery disease and/or history of myocardial infarction; 2. acquired or congenital valve disease leading to impairment of myocardial function excluding functional mitral and/or tricuspid regurgitation; 3. connective tissue disease and/or neoplasm; 4. infection; 5. endocrine diseases, i.e. diabetes mellitus, hyper- or hypothyroidism, Cushing disease; 6. advanced liver or kidney disease.

HRQoL was measured with the SF-36 test. The SF-36 encloses eight scaled scores, which are sums of the questions (36) in corresponding section. Each scale is directly transformed into a 0–100 scale on the assumption that each question carries equal weight. These eight parts are respectively: Physical Functioning *PF*, Role Physical *RP*, Bodily Pain *BP*, General Health *GH*, Vitality *V*, Social Functioning *SF*, Emotional Role Functioning *RE* and Mental Health *MH*. First four coefficients are related to physical functioning and four consecutive ones to mental health. The presence of depression was diagnosed according to patient's history, clinical observation, the Beck Depression Inventory [8] and the Hamilton rating scale for depression [9]. If depression was suspected patient were consulted by a psychiatrist. The clinical observation of patients began on admission to hospital and lasted for 36 months.

Methods

Weighted k-means clustering: The algorithm proposed in [10] has been used to find a partition of a dataset X, with $M = 193$ records and $N = 8$ features corresponding to the SF-36 test scales, into $k = 2$ clusters. It is a modification of classical k-means clustering, however to identify the importance of different features, a weight is assigned to each feature in the distance calculation. Formally, the minimization of the following objective function is being done:

$$Q(U, Z, W) = \sum_{l=1}^{k} \sum_{i=1}^{M} \sum_{j=1}^{N} u_{i,l} w_j^{\beta} d(x_{i,j}, z_{l,j}) \tag{1}$$

where:

U – an $M \times k$ partition matrix, $u_{i,l} \in \{0, 1\}$,

$Z = \{Z_1, Z_2, \ldots, Z_k\}$ – a set of k vectors representing the k-clusters centers,

$W = [w_1, w_2, \ldots, w_N]$ – a set of weights,

$d(x_{i,j}, z_{l,j})$ – a distance or dissimilarity measure between object i-th and the center of l-th cluster on the j-th feature; in paper we used: $d(x_{i,j}, z_{l,j}) = (x_{i,j} - z_{l,j})^2$,

$\beta > 1$ – a fuzziness parameter.

W-k-means clustering algorithm:

A. Random generation of initial set of weights W^0, $\sum_{j=1}^{N} w_j = 1$ and partitioning matrix U^0; set $t = 0$.

B. In original paper authors suggested to choose randomly an initial set of Z. However, to improve final results, we used a classical k-means algorithm. Calculation of Q.

C. Update of matrix U: $u_{i,j}^{(t+1)} = 1$ if

$$\underset{1 \le t \le k}{\forall} \sum_{j=1}^{N} w_j^{\beta} d(x_{i,j}, z_{l,j}) \le \sum_{j=1}^{N} w_j^{\beta} d(x_{i,j}, z_{t,j}) \tag{2}$$

otherwise for $t \neq l$ $u_{i,j}^{(t+1)} = 0$.

D. Update of matrix Z:

$$\underset{1\leq l\leq k}{\forall}\ \underset{1\leq j\leq N}{\forall} z_{l,j}^{(t+1)} = \left(\sum_{i=1}^{M} u_{i,l}x_{i,j}\right)\cdot\left(\sum_{i=1}^{M} u_{i,l}\right)^{-1} \tag{3}$$

E. Update of matrix W: $w_j^{(t+1)} = 0$ if $D_j = 0$, otherwise

$$w_j^{(t+1)} = \left(\sum_{s=1}^{h} [D_j/D_s]^{\frac{1}{\beta-1}}\right)^{-1} \tag{4}$$

where: h is the number of features with $D_j \neq 0$ and

$$D_j = \sum_{l=1}^{k}\sum_{i=1}^{M} u_{i,j} d(x_{i,j}, z_{l,j}) \tag{5}$$

Then Q recalculation. If $Q^{(t+1)} = Q^{(t)}$ then stop, else go to step C.

In order to refine the optimal β coefficient we used the Bouldin-Davies (DB) index (6) and mean Cohen's effect size (ES) – for all eight scales.

$$DB = \frac{1}{2}\sum_{i=1, i\neq j}^{n} \max\left(\frac{\sigma_1+\sigma_2}{d(z_1,z_2)}\right) \tag{6}$$

where: σ_1 – the average distance of all patterns in the i-th cluster to their cluster center z_i, $d(z_1, z_2)$ – distance of cluster centers z_1 and z_2. Small values of DB correspond to clusters that are compact, and whose centers are far away from each other [11]. ES was also used as a measure of the strength of the particular SF-36 test scales in the relationship between two groups yielded by the w-k-means algorithm. Cohen's ES is defined as the difference between two means divided by a standard deviation for the data [12]. The larger ES, the bigger size of the effect (higher relevance of analyzed factor). Coefficient values are important in interpreting the data, as it is possible to determine, not only the statistical significance but clinically relevant changes (or differences) in the quality of life [13–14].

In order to illustrate the quality of obtained distribution of the population into two groups with different HRQoL we used the Principal Component Analysis. To specify cut-off points for each of the scales ROC curves were plotted and typical performance measures for the confusion matrix were calculated (including Matthews correlation coefficient MCC [15] and normalized mutual information NMI [16]). Relative risk with confidence intervals was calculated to assess the risk of development (or having) depression and MACE. Kaplan-Meier survival curves for both groups were computed and compared with the log-rank test. For clinical data we used: the χ^2 test for categorical data, for interval data (including the SF-36 scales) with normal distribution or after normalization with the Box-Cox transformation, the t-Student test, otherwise the U Mann-Whitney test. Variables distribution was evaluated with the Shapiro-Wilk test. Homogeneity of variances was assessed by the Levene test.

Results

Weighted k-means clustering results

We used the DB-index and mean effect size values to select the β parameter in order to get optimal clustering results. Based on results presented in the [Fig. 1] we chose $\beta = 2.4$.

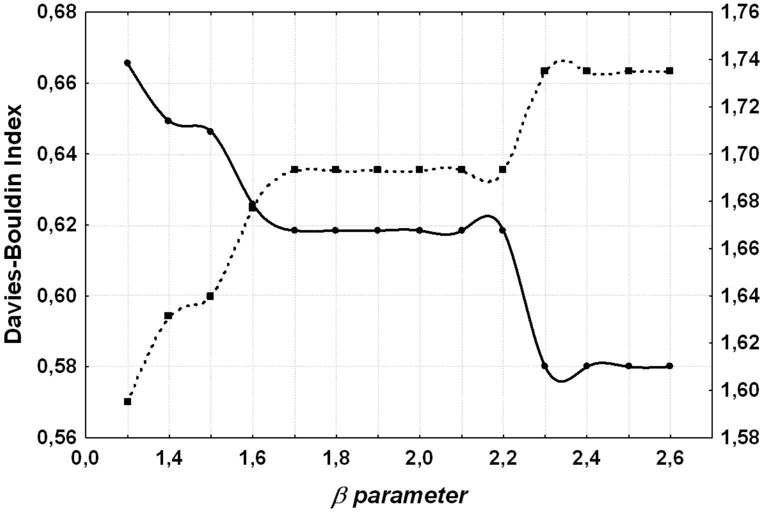


Fig. 1. Values of DB-Index (*solid line*) and mean Effect Size (*dashed line*) according to β parameter in weighted-k-means clustering algorithm

[Fig. 2] and [Tab. 1] present respectively the PCA results and weights yielded by the clustering algorithm as well as ES values for each scale of the SF-36 test. As it can be seen, the obtained groups are well separated from each other. The most important weights proved to be Vitality and the worst one the Role Emotional. Taking into account ES, the best factor was the same, however the worst one was the Physical Functioning.

[Fig. 3] shows comparison of the SF-36 scales between both groups. Statistically significant differences between all eight scales were found ($p < 0.001$).

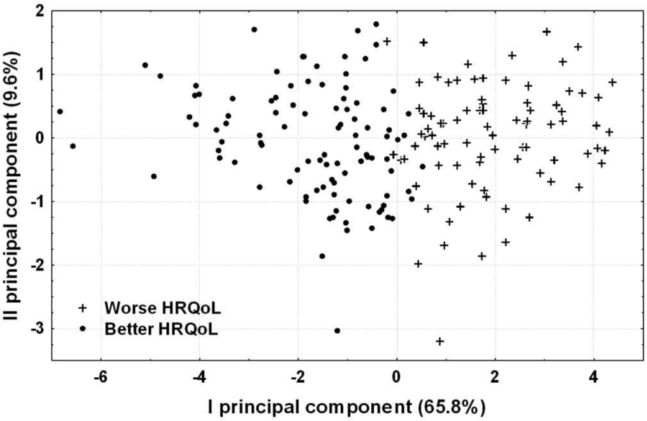


Fig. 2. The PCA projection for the SF-36 test scales

Tab. 1. Weights yielded by the W-k-mean algorithm and corresponding mean Effect Sizes values

SF-36 scale	PF	RP	BP	GH
Weight	0.1028	0.1147	0.1036	0.1526
Cohen's ES	1.2946	1.6044	1.7259	1.9981
SF-36 scale	V	SF	RE	MH
Weight	0.1815	0.1403	0.0820	0.1225
Cohen's ES	2.7457	2.1923	1.5417	1.7762

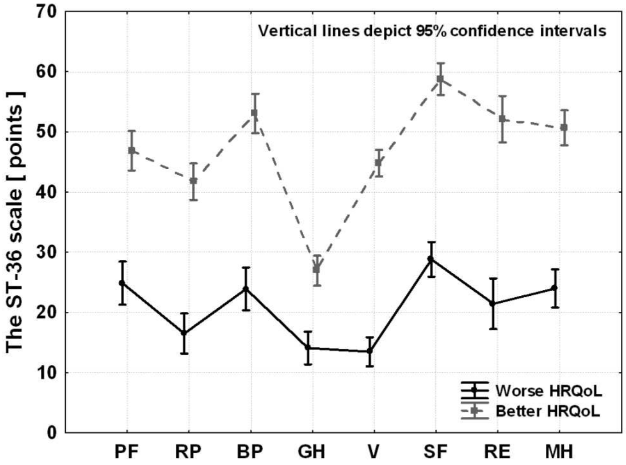


Fig. 3. Comparison of the SF-36 scales between both groups

ROC curves

[Tab. 2] presents results based on ROC curves. For all scales cut-off points were computed. Sensitivity and specificity are one approach to quantify the diagnostic ability of the test. This coefficients measure respectively the proportion of actual positives and the proportion of negatives which are correctly identified. A test with a high specificity has a low statistical significance (α), while a test with a high sensitivity has a low statistical power ($1 - \beta$). In clinical practice, however, the test result is all that is known, so we want to know how good the test is at predicting the disease.

Tab. 2. Results of the ROC analysis

	PF	RP	BP	GH	V	SF	RE	MH	All
	> 40	> 31	> 32	> 15	> 25	> 50	> 33	> 40	> 265
AUC	0.816	0.879	0.888	0.763	0.965	0.942	0.862	0.881	0.994
Se	0.705	0.648	0.933	0.867	0.924	0.924	0.781	0.905	0.981
Sp	0.795	0.932	0.727	0.580	0.932	0.852	0.818	0.761	0.955
ACC	0.746	0.777	0.839	0.420	0.927	0.891	0.798	0.839	0.969
PPV	0.804	0.919	0.803	0.736	0.942	0.882	0.837	0.819	0.963
NPV	0.693	0.689	0.901	0.711	0.911	0.904	0.758	0.870	0.977
FPR	0.205	0.068	0.273	0.420	0.068	0.148	0.182	0.239	0.045
FNR	0.295	0.352	0.067	0.133	0.076	0.076	0.219	0.095	0.019
MCC	0.499	0.594	0.682	0.470	0.854	0.781	0.597	0.678	0.937
NMI	0.883	0.897	0.909	0.880	0.946	0.928	0.896	0.908	0.971

Thus, we have also calculated other parameters, especially positive predictive value and normalized mutual information. Positive predictive value is the proportion of patients with positive test results who are properly diagnosed. It is a key measure of the diagnostic method as it reflects the probability that a positive test corresponds to the underlying condition being tested for, in our case, better HRQoL.

The problem with PPV is that its value depends also on the prevalence of the disease, which of course may vary. In order to deal with this problem it should only be used if the ratio of the number of patients in the disease group and the number of patients in the healthy control group is equivalent to the prevalence of the diseases in the studied population. However, our study on HRQoL in patients with systolic heart failure in Poland is unique and there is no information about the worst quality of life prevalence in CHF population. This led us to normalized mutual information which is interpreted as an amount by which the model reduces our uncertainty about the true state. As it can be seen, the best discriminating value has Vitality (highest area under curve, best accuracy and positive predictive value, Matthews correlation coefficient and mutual information) and then Social Functioning and Mental Health.

Statistical analysis

For the sum of all scales we obtained high value of NMI. Patient with CHF who has more than 265 points is very likely to have good HRQoL. Relative risk (RR) of depression development for a person with less than 265 points is 3.29 (95% CI: 2.04–5.32; $p < 0.0001$). RR for MACE occurrence is 1.83 (95% CI: 1.31–2.55; $p < 0.001$). The Number Needed to Treat which is the number of patients who need to be treated in order to prevent MACE outcome is 3.79, so we need to treat 4 patients to avoid one adverse cardiac event.

[Fig. 4] shows Kaplan-Meier survival curves in both yielded by the w-k-mean algorithm groups. Patients with the worst quality of life statistically significantly more often and earlier underwent adverse cardiac events than the other group. In [Tab. 3] we enclosed the comparison of relevant for the HRQoL clinical parameters between both groups.

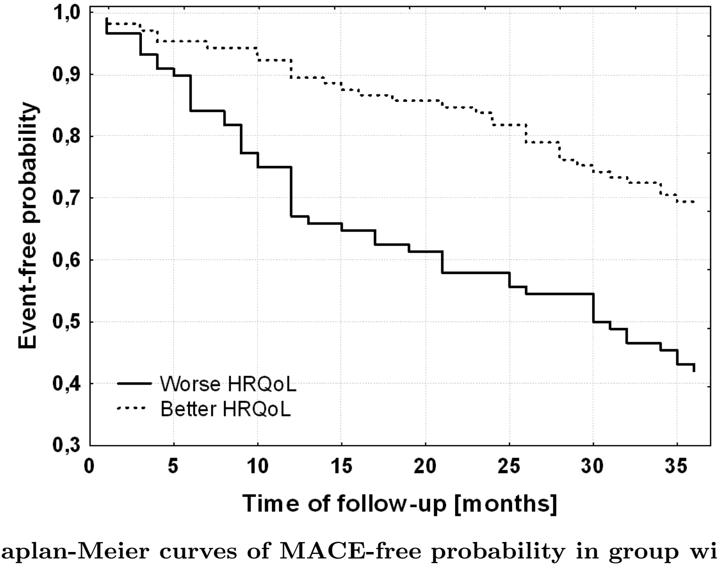


Fig. 4. Kaplan-Meier curves of MACE-free probability in group with better and worst HRQoL ($p_{\text{log-rank}} < 0.001$)

Tab. 3. Comparison of Clinical parameters Between group with better and worse HRQoL

Parameter	+ HRQoL	– HRQoL	p
Death	5 (4.76%)	31 (35.23%)	< 0.001
MACE	33 (31.43%)	51 (57.95%)	< 0.001
PAP>19 [mmHg]	49 (46.67%)	56 (64.64%)	< 0.05
RAP>5 [mmHg]	47 (44.76%)	54 (61.36%)	< 0.05
hs-CRP [mg/l]	1.64/4.85	2.80/4.22	< 0.01
NT-pro BNP [pg/ml]	548/972	1669/3055	< 0.001
Bilirubin [Bmol/l]	16.25/10.22	18.45/12.10	0.0964
Long QT	28 (26.67%)	39 (44.32%)	< 0.05
LVEDD [mm]	65.8±8.3	68.1±7.9	< 0.05
LVESD [mm]	51.0±9.5	53.5±9.6	< 0.05
IVRT [s]	71.0/40.0	60.0/30.0	< 0.01
TAPSE [mm]	21.0/5.0	19.0/10.0	< 0.001
E/A	1.4/1.2	1.8/2.7	< 0.01

Mean±STD or Me/IQR (Interquartile range)

As it can be seen, patients with the worst quality of life have statistically significant elevated pulmonary and right arterial pressure as well as bilirubin, high sensitive C-reactive protein (CRP) and brain natriuretic peptide NT-pro BNP plasma level. They also have almost twice often long QT syndrome (in electrocardiogram). In echocardiography, these patients asserted higher left ventricular (LV) end diastolic and systolic diameter, isovolumetric relaxation time, the E/A ratio of transmitral flow and lower tricuspid annular plane systolic excursion (TAPSE).

Conclusions

1. Application of the weighted k-means algorithm yields two well separated in the SF-36 scales dimension groups with poor and good HRQoL. Weights obtained in the clustering process correspond generally with clinically relevant differences measured with the Cohen's ES and evaluation measurements of ROC curves. The obtained groups differ significantly in all eight scales of the SF-36 test.

2. For all scales cut-off points were calculated and Vitality proved to be the best discriminating SF-36 scale. Vitality corresponds with patient's well-being and "life energy". So, there is less than 6% chance that the patient with more than 25 points has poor HRQoL. 92% of the patients with real good quality of life will be correctly identified by this scale (nevertheless all SF-36 scales should be taken into consideration in assessment of patient's HRQoL).

3. Taking into account the sum of all SF-36 scales, there is less than 4% chance that the patient with more than 265 points has worse HRQoL. 99.4% of the patients who actually have good quality of life will be correctly identified with this test.

4. It is undeniable that poor Health Related Quality of Life (less than 265 points in all scales of the SF-36 test) is associated with higher risk of hospitalization and death occurrence as well as with depression development. Therefore, it is **very important** to perform screening tests for quality of life (and further for depression) in all patients with chronic heart failure, because effective treatment of depression may improve their long-time prognosis.

5. Quality of life is strongly associated with the worst clinical parameters [Tab. 3]. Depression might promote an inflammatory response (represented by CRP and NT-pro BNP) by activating the immune response. Alternatively, the effects of depression on inflammation might be due to its links with psychological stress. Worse echocardiography parameters and higher values of arterial pressure are related with heart remodeling in chronic heart failure. On the other hand, heart insufficiency handicaps patient's physical and social activity.

List of abbreviations

ACC	Accuracy
AUC	Area Under ROC Curve
FNR	False Negative Rate
FPR	False Positive Rate
MCC	Matthews correlation coefficient
NMI	Normalized Mutual Information
NPV	Negative Predictive Value
PPV	Positive Predictive Value
ROC	Receiver Operating Curve
Se/Sp	Sensitivity/Specificity
STD	Standard Deviation
WKM	Weighted K-Means Algorithm

REFERENCES

[1] Jessup M., Brozena S. C., Heart failure, N Engl J Med, 348 (20), pp. 2007–2018, May 2003.

[2] Muntwyler J., Abetel G., Gruner C., Follath F., One-year mortality among unselected outpatients with heart failure, Eur Heart J, 23 (23), pp. 1861–1866, December 2002.

[3] Cohn J. N., Johnson G. R., Shabetai R., et al., Ejection fraction, peak exercise oxygen consumption, cardiothoracic ratio, ventricular arrhythmias, and plasma norepinephrine as determinants of prognosis in heart failure. The V-HeFT VA Cooperative Studies Group, Circulation, 87 (6), pp. 5–16, Juni 1993.

[4] de Groote P., Dagorn J., Soudan B., et al., B-type natriuretic peptide and peak exercise oxygen consumption provide independent information for risk stratification in patients with stable congestive heart failure, J Am Coll Cardiol, 43 (9), pp. 1587–1589, May 2004.

[5] Faller H., Störk S., Schuler M., et al., Depression and disease severity as predictors of health-related quality of life in patients with chronic heart failure – a structural equation modeling approach, J Card Fail, 15 (4), pp. 286–292, May 2009.

[6] Ware J. E., Kosinski M., Dewey J. E., How to score Version 2 of the SF-36® Health Survey (Standard and Acute Forms), Medical Outcomes Trust and QualityMetric, Incorporated 2002.

[7] Szyguła-Jurkiewicz B., Owczarek A., Duszańska A., et al., Long-term prognosis and risk factors for cardiac adverse events in patient with chronic systolic heart failure due to hypertension, PAMW, 118 (5), pp. 280–287, 2008.

[8] Beck A. T., Ward C. H., Mendelson M., et al., An inventory for measuring depression, Arch Gen Psychiatry, 4, pp. 561–571, 1961.

[9] Hamilton M., A rating scale for depression, J Neurol Neurosurg Psych, 23, pp. 56–62, 1961.

[10] Huang Z., Ng M. K., Rong H., Li Z., Automated variable weighting In k-means type clustering, IEEE PAMI, 27 (5), pp. 657–668, 2005.

[11] Halkidi M., Batistakis Y., Vazigriannis M., On clustering validation techniques, J Intell Inf Syst, 17 (2), pp. 107–145, 2001.

[12] Kazis L. E., Anderson J. J., Meenan R. F., Effect sizes for interpreting changes in health status, Med Care, 21, pp. 178–189, 1989.

[13] Osoba D., King M., Meaningful differences In: Fayers P., Hays R. (eds): Assessing quality of life in clinical trials. II Ed., Medical Press, pp. 243–259, 2005.

[14] Sprangers M. A., Moinpour C. M., Moynihan T. J., et al., Assessing meaningful change in quality of life over time: a user's guide for clinicians, Mayo Clin Proc, 77, pp. 561–571, 2002.

[15] Baldi P., Soren B., Chauvin Y., et al., Assessing the accuracy of prediction algorithm for classification, an overview, Bioinfo Rev, 16 (5), pp. 412–424, 2000.

[16] Bush W. S., Edwards T. L., Dudek S. M., et al., Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction, BMC Bioinformatics, 9 (238), pp. 1–17, 2008.

How much credible are the responses obtained from an individual respondent in a non-repeated questionnaire survey: looking for practical methods with a statistical support

Izabela Chmiel[1], **Maciej Górkiewicz**[2]

[1] Department of Medical and Environmental Nursing, Faculty of Health Sciences, Jagiellonian University Medical College, Poland

[2] Department of Epidemiology and Population Research, Jagiellonian University Medical College, Poland

Abstract. All recognised psychometric methodologies, like the Classical Test Theory (CTT), the Item Response Theory (IRT), cognitive approaches, all in their essence are oriented on the population scale of the investigation. However, in daily practice, questionnaire inquiries are administered regularly only among a very limited group of people, frequently by very diverse professionals without any solid statistical background and usually with a clear practical purpose to support a decision on the course of therapy of an individual patient. Authors had some experience in both of the above domains. This paper was intended to remain on the borderline. First, it briefly discusses how the psychometricians put into practise the principle: proper instrument + proper procedure + proper attitude. The main focus was put on the demonstration of how to use several additional items in a questionnaire survey with an aim to verify the credibility of an individual respondent. The common pitfalls were illustrated with examples of analyses with use of easily available statistical procedures, like confidence intervals for proportion and the Friedman test for orderings.

Introduction

Historically, the psychometric and the statistical methodologies evolved in a close association, but without intense involvement with everyday psychological know-how. Many topics that are important for test users don't receive enough attention in psychometrics, so the question, what psychometrics can do for applied psychology, remains open [1–3]. Not a bit less vital can be the inverted question: what the questionnaire users can do for psychometrics [4]. However, the scope of this paper was limited to much simpler, practical question: how to enlarge our trust to the responses obtained from an individual respondent, basing above all on the common sense, supported with some relatively simple statistical procedures.

ISBN 978–83–7431–336–0 ISSN 0860-150X

Classical test theory (CTT) postulates an ideal situation, in which a perfect responder is always entirely eager to give the scores to each item of the reliable questionnaire of the simple linear structure, that can be expressed with an equation as (1) [5–6].

$$Y = \sum(X_i + e_i) = Y\hat{}+ \sum(e_i); \quad i = 1, 2, \ldots, K. \tag{1}$$

where: K – number of items in a single-scale questionnaire; $Y\hat{} = \sum(X_i)$ – actual result of the measurement, an estimate of the aggregate true score Y; X_i – responder's score to a particular i-th item; e_i – random error component, from definition of the expected value equal to 0.

Thus, in the frame of CTT, the prime attempt should be put on building a reliable questionnaires [7]. Nevertheless, the basic notions and indices of the CTT techniques, in these the factorial structure of the item set and Cronbach's alpha index of the internal consistency, became standard tools in every psychologist's tool kit [8].

Item Response Theory (IRT) generally didn't resign from the perfect respondent assumption, but it postulated more complex models then CTT. Unidimensional Rasch approach seemingly didn't modify the CTT formula (1), so an aggregate score is estimated in the same way, as the sum of the scores obtained by the items in the questionnaire. The difference here has a structural nature for it was assumed that true scores of items are put in linear order [9]. It is easy to notice that with respect to ordered chance to be censored, the error components of the particular items have their expected values generally not equal to 0. Nevertheless, if the postulated ordering of the true scores exist in the real world, they can be easy estimated, because of the relatively small probabilities of the contradictory actual scores [10]. In a doubtful case it seems to be more appropriate to resign from the Rasch modeling, than to adjust the essentials of a model to the actual data [11]. The Structural Equations Models (SEM) technique can connect, in various ways, several equations like (1). Moreover, SEM created an opportunity to include into a joint model many variables, latent as well as manifest, like features of the responders, and attributes of a survey [12–14].

The cognitive approaches, contrary to CTT and IRT, gave attention to the psychological aspects of the questionnaire inquires, before all in terms of responder's ability and readiness to provide the honest answers to the questionnaire items [15–16]. Practical recommendations how to organize a questionnaire survey are based here on real-world observations [17], with correspondence to known psychological approaches, such as the false memory phenomenon [18], the cognitive-affective models of goal-setting [19], the model of planned behaviour [20], or the attitude-social influence-efficacy (ASE) paradigm [21].

The professional developers of the questionnaires addressed to broad target must create a balanced amalgam of some practical modus operandi. For instance, a potential user of the known SF-36 questionnaire, besides detailed handbook how to carry out a survey and all subsequent calculations, can get a lot of authorized and very useful information [22], on the factorial structure of SF-36 [23], on Rasch models for SF-36 [24–25], on path models for SF-36 [26]. In addition, the numerous disinterested scientific reports are available, among other with regard to nonstandard subjects [27], or to nonstandard procedure [28]. The adaptation of a standard questionnaire to other (nonstandard) population needs a special prudence [29]. However, this seems to be less risky for the researcher than trying to create ad hoc ones own new questionnaire, especially without proper psychological background [30].

The rest of this paper was organized as follows. Brief examples on standard adaptation and validation procedures from the authors own studies are given in the chapter: The ground rule: proper instrument + proper procedure + proper attitude. In the next chapter: Statistical supports for credibility of an individual respondent, the method based on comparisons between responder's opinion versus a corresponding pattern is proposed. The two exemplary patterns, that is the typical relationship between the inclination to guess Yes-or-No answer versus the level on the disagreement in the matter [31], and the standard ordering of the items of the physical functioning (PF) scale in SF-36 questionnaire were obtained from our previous studies of Polish groups [32].

The ground rule: proper instrument + proper procedure + proper attitude

The developers of the wildly used questionnaires have made a vital attempt to be in a reasonable agreement with all recognized approaches to psychometrics. With respect to the applied surveys the common recommendations for the researchers, how to enlarge their chance to obtain valid results, can be summarized as follows [33]:

1. draw a representative sample of responders from a population under study;
2. use a standard questionnaire;

3. apply a standard procedure if possible;
4. confirm a sufficient similarity between standard population and a population under study.

In spite of a broad use, the notion of a standard questionnaire hasn't any formal definition. The usual obligatory demands for a standard questionnaire included: available detailed scientific report from a large scale confirmatory survey, authorized handbook with instructions how to carry out justifiable surveys with use of this questionnaire, and desirably, at least several research reports from surveys made with use of this questionnaire in various circumstances and by different research teams. It should be emphasized once again, that the practicality and efficacy of any standard questionnaire and of the standard procedure of its use, both were proved jointly by their developer with respect to some accurate specified standard population. Thus, if an actual survey was made at some other population, then the sufficient similarity between standard population and the population under study must be proved thoroughly [34]. In case of need, it should be proved that the all identified differences were irrelevant in the matter, for instance for the fertility behaviour [35]. However, a great number of the potentially influential variables makes very likely the occurrence of the Simpson paradox [36].

The authorized handbook [22] provided the detailed recommendations how to carry out a survey with the standard SF-36 questionnaire, and then, how to make calculations and interpret the results. All 36 items of the SF-36 questionnaire produce only 9 variables (health-related quality of life domains): GH – general health; HT – change in health; VT – energy/vitality; MH – mental health; RP – role limitation-physical; SF – social functioning; BP – bodily pain; RE – role limitation-emotional, and PF – physical functioning. The raw SF-36 data should be standardized with a range of 0–100% separately for each the above 9 scales. For the purposes of the confirmatory analyses the two standard populations where characterised with their estimates of the mean values and standard deviations for the three domains: PF = 83.29±23; RP = 82.51±25; VT = 58.31±20 for the USA general population, and PF = 83.9±11.6; RP = 72.4±5.1; VT = 64.5±5.7 for the Finnish general population.

The authorized recommendations [22] were respected rigorously at the thesis stage [33], however some additional analyses were applied. The fundamental presumption that the study group can be considered as representative, at least for Polish convalescents after successful clinical therapy against *acute pancreatitis,* was supported by several arguments. The initial sample included all of the 422 patients hospitalised for *acute pancreatitis* at the 1st Department of General Surgery at the Jagiellonian University of Krakow (Poland) from 2000 to 2006. The only four exclusion criteria were used: age: < 18 years or > 70 years (66 excluded); death (34 excluded); non complete clinical data (20 excluded), complication with other illness (36 excluded). The standard procedure for the mail survey was applied with proper thoroughness. The standard Polish version of SF-36 questionnaire with standard instructions was mailed to all of the 266 non-excluded survivors. A covering letter accompanying each questionnaire included also the explanation of the survey purpose and of the possible health benefits for the respondent. A phone consultation in completing the form, if needed, was offered. Nevertheless, the $N = 124$ participants didn't return an answer, but $N = 142$ survivors (81 men and 61 women) returned acceptably completed forms. The three clinical types of disease were represented at the study sample with appropriate proportion: 61:41:40. The response rate RR $= 142/266 = 53.4\%$ was acknowledged as sufficient for the mail survey. Moreover, the clinical and demographic data for non-responders and responders were quite similar, so the adjusting for non-response was unnecessary. The assumptions of normality of the scores for the 9 particular health-related quality of life domains, as measured with the SF-36 questionnaire at the study group, were supported with moderate values of skew and ranged from skew $= -0.52$ to skew $= 0.24$, and of the kurtosis varied between kurtosis $= -1.12$ and kurtosis $= 0.43$. Consequently, the confirmatory analyses of the data reliability and validity were executed predominantly at the frame of the classical test theory (CTT) with the use of the parametric procedures. The proper correlation structure of the raw SF-36 data was confirmed for each domain separately not only under the criterion that Cronbach's alpha > 0.7 but also under the criterion that each item is in a quite strong correlation with the summary score of its domain, but is in relatively weaker correlation with any other item. The concurrent validity was confirmed on the base of estimates for the study group: PF $= 64.5 \pm 27.1$; RP $= 59{,}0 \pm 30{,}9$; VT $= 52.5 \pm 16.8$; and as well, on the base of estimates for the Finnish convalescents also after *acute pancreatitis*: PF $= 83.0 \pm 21.6$; RP $= 69.4 \pm 27.8$; VT $= 60.4 \pm 23.4$ [37]. It was easy to notice, that the study group didn't differ significantly with respect to PF, RP and VT domains from any of the above groups, because all considered differences between mean values were less than their standard deviations of the study group. Beyond the obligatory recommendations, the study data were reanalysed at several parallel studies with the use of somewhat more advanced methodology [22]. The postulated linear ordering of the items of PF scale in the study sample was confirmed with the Rasch methodology, and then used in comparative analyses with the aim to confirm the concurrent validity of the study data [32]. In the study data the significant regression was detected between mean scores of the SF-36 domains and their standard deviations, SDˆ = −0.043 + 0.524*mean; $R = 0.705$ with statistics $F = 6.9$; $p = 0.03$. In such a situation, the multiple comparisons procedures [38], and the bootstrap [39], were used with the aim to get an additional support to previous conclusions in the matter, based on parametric procedures. The suitability of the applied clinical classification of the patients menaced with *acute pancreatitis* was confirmed also in terms of the propensity score [40]. The informative links between age and gender on one side, and the chosen SF-36 domains on the other side were estimated [41–42].

The analyses (cited in this chapter) provided strong support to conclude that the study group is representative, at least for Polish convalescents after *acute pancreatitis*, that in general the members of the study group gave trustworthy scores to items of the SF-36 questionnaire. These findings raise the possibility that the data obtained in this group with other, nonstandard questionnaires, can be considered as a source of valid information. Basing on this conviction, the health behaviours of convalescents after *acute pancreatitis* were classified [43], recommendations on needed psychoeducational intervention for convalescents were proposed [44–45], and the proposal to school's health education were suggested [46].

Quite analogous approach, following scrupulously the recommendations of the developers of the standard questionnaire, but not neglecting the parallel analyses with other methodologies, were applied in our studies on adopting the known CES-D questionnaire [47], and in adopting the physicians' career satisfaction questionnaire [48]. The forward-backward translation procedure was applied with special attention to a dogma, that the more the respondents are emotionally invested in the item, the more likely those emotions will influence their scores. Concurrent validity of CES-D was proved by comparison with scores obtained through the well-known Beck Depression Inventory. The responders from the sample included 3544 permanent residents of Krakow (Poland), recruited from the HAPIEE Study (Health, Alcohol and Psychosocial factors In Eastern Europe, url=http://www.ucl.ac.uk/easteurope/hapiee.html). Besides, the two above studies the concurrent cross-cultural validity was confirmed by the similarity between standard factorial structure and the one estimated for the Polish version of the adopted questionnaire.

The endeavour to introduce a novel questionnaire creates new fundamental challenges for developers. In such case, the structural equation modelling (SEM) in the frame of item response theory (IRT) should be preferred. This allows exploration in compound of the truly multivariate models, where multiple independent variables can influence multiple intermediate variables in the prediction of the final effects of the antecedent variables. In the study [49], two simple unidimensional scales, as described by the equation (1), were examined. First, the questionnaire uses only three items to measure the latent variable named motivation-to-work, and the other one uses eleven items to measure the latent variable named attitude-to-patients. The SEM model linked one latent variable to the other. In result, the unidimensionality of both questionnaires under study, and the significant correlation between the considered latent variables, $R^2 = 0.78$; $p < 0.001$ was confirmed simultaneously. In continuation [50] the use of some easy available data of a candidate nurse as a substitute to the questionnaire review with above questionnaires was considered. However, it should be emphasized that this technique seems to be useful for managers looking for suggestions how to develop patient-friendly staff in a rather long perspective, but not to evaluate an individual candidate. In the study [51] on the false memory, the latent variable named Model, expressed an inclination of the respondent to invent the “ad hoc” explanations in spite of insufficient information. It was proved that a simple linear model like (1) should be rejected from consideration. In the final quasi-linear SEM model the six paths, each significant, at least on the level of $p < 0.034$, connected four independent variables (gender, ratio of the true answers, and ratios of two kinds of the wrong answers) and the latent variable Model into a complex relationship, and visibly different from a simple model (1).

Statistical supports for credibility of an individual respondent

The question as such, had been a vital issue in almost all real-life domains. However, in medicine the problem, of how far patient’s opinions may be trusted, has its special significance. Generally, there is an agreement between regulatory authorities and the research community that patient-reported outcome (PRO) assessment in health care should proceed from a strong conceptual basis, with rationales clearly articulated in advance concerning what is to be measured and how this is to be accomplished, with greater awareness to recall bias and degrees of psychometric validation [52–55]. It should be recognised that patients’ and their care-providers’ views can show some discrepancy, especially with regard to the course of rehabilitation and other long-time care, therefore, the interviews carried out by other persons from outside seem to be indispensable here [46, 56–57]. Before choosing a validation method the two crucial prerequisites should be considered thoroughly:

(i) the hypothetical source of false answers: unplanned random answering versus intentional (maybe: to some extent unconscious) play-acting or pretending;

(ii) the anticipated meaning of a patient's opinion: expert's report versus subjective conviction or impression.

As to the first of the above issues (i), the strategy of random answering can be easy modelled with the use of some commonly applied standard distribution. The strategy of inventing fictitious self-image is generally more difficult to unveil, especially without any clear concept of a possible pattern or a scale of a self-worth underlying this strategy. In this study we attempted to disregard this problem by using nonparametric statistical procedures.

As to the second issue (ii), expert's opinion can be verified directly by comparison with real world occurrences, and with opinions of other experts. This problem has great practical relevance, and plentiful literature on the matter, nevertheless, it wasn't included in the scope of this paper. The dishonesty of subjective conviction is generally more difficult to reveal. In this paper we suggest the use of a characteristic pattern, that is a typical relationship between variables measured in a questionnaire survey at some postulated populations. The proposed validation techniques were aimed to recognise a responder either as outsider or as a member of these populations. The practical difficulties with the use of the two patterns, that is the typical relationship between the inclination to guess Yes-or-No answer versus level on the disagreement in the matter [31], and the standard ordering of the items of the physical functionning (PF) scale in SF-36 questionnaire [32], were explained in this paper with the exemplary statistical calculations.

The method for evaluating an individual inclination to guess Yes-or-No answers [31], was originally developed to examine the members of a small group of experts. In the experiment the role of anonymous experts played $N = 84$ graduated nurses. The questionnaire included mixture of controversial items with different levels of disagreement in literature. The two kinds of items with dichotomous decisive Yes-or-No answer were used: the $K = 31$ items allowed apparently only decisive answer, but $L = 44$ items permitted explicitly also the third I-don't-know option of a answer. In such a way the questionnaire created series of two seemingly equivalent decision situations. In the first situation the participants, aimed to avoid a decisive answer, giving neither Yes nor No answer. But in the second situation they can choose freely an additional option I-don't-know. It was proved that, even in the anonymous survey, the same participants avoided the decisive Yes or No answer, significantly less often in the first situation, only 3 times at $N \cdot K = 84 \cdot 31 = 2604$ answers, what leads to a proportion $Pr_1 = 3/2604 = 0.0012$; 95%CI: $Pr_1 < 0.003$; than in the second situation, 608 times at $N \cdot L = 84 \cdot 44 = 3696$ answers, what leads to much greater proportion $Pr_2 = 608/3696 = 0.165$; and confidence interval 95%CI: $0.153 < Pr_2 < 0.177$. The odds ratio $OR = Pr_2/Pr_1 = 142.8$ with a confidence interval $74.0 < OR < 275.7$. Moreover, the strong log-linear relationship (2), $p < 0.01$; between the proportion of Yes versus No answers and the frequency of the I-don't-know answers at the $L = 44$ items with this option was observed. It is easy to notice in equation (2), that odds ratio $OR_{\text{I-don't-know/I-know}}$ obtained its maximal value for $\text{Ln}|OR_{\text{Yes/No}}| = 0$; that is in situation when (frequency of Yes) = (frequency of No).

$$\text{Ln}(OR_{\text{I-don't-know/I-know}}) = 0.526 - 0.943 \cdot \text{Ln}|OR_{\text{Yes/No}}| \qquad (2)$$

where:

$OR_{\text{I-don't-know/I-know}} =$
$= \text{(frequency of I-don't-know)/(frequency of either Yes or No)};$

$OR_{Yes/No} = \text{(frequency of Yes)/(frequency of No)}.$

It seems that the design of a verifying experiment should be limited here to three binary variables only, that is: level of agreement in a standard population with regard to choice between Yes versus No answer (Agreement = low vs. Agreement = high), encouragement to I-don't-know answer (Option I-don't-know = offered vs. Option = hidden), answer = decisive versus answer = ambiguous. In result the set of $N = 60$ respondent answers to the verifying items can be summarised as a 3D table of frequencies, like [Tab. 1]. The null hypothesis that the respondent provided his answers independently from item's values of variables Agreement and Option can be easily proved with calculator for Fisher exact test, available on-line [58]. It should be noted that in spite of relatively large number $N = 60$ of the verifying items, the estimated significance of the null hypothesis was quite near to $p = 0.05$. Thus, the use of the discussed method for evaluating an individual inclination to guess Yes-or-No answers, seems to be useful only in a situation if a researcher is truly interested in the viewpoint of a respondent on almost all of the verifying items.

Tab. 1. Exemplary data on an individual inclination to guess Yes-or-No answers

Option I-don't-know	Yes:No Agreement = low	Yes:No Agreement = high	total
Offered	$N_{\text{ambiguous}}/N_{\text{decisive}} = 3/7$	$N_{\text{ambiguous}}/N_{\text{decisive}} = 3/12$	6/19
Hidden	$N_{\text{ambiguous}}/N_{\text{decisive}} = 1/14$	$N_{\text{ambiguous}}/N_{\text{decisive}} = 0/20$	1/34
Total	$N_{\text{ambiguous}}/N_{\text{decisive}} = 4/21$	$N_{\text{ambiguous}}/N_{\text{decisive}} = 3/32$	7/53

N_{decisive} – number of either Yes or No answers;
$N_{\text{ambiguous}}$ – number of other answers;
Yes:No Agreement = low
if in standard population Probability(Yes) ≈ Probability(No);
Yes:No Agreement = high
if in standard population |Probability(Yes) – Probability(No)| > $\frac{1}{2}$;
Null hypothesis H_0:
the same probabilities in each cell $Pr(N_{\text{ambiguous}}/N) = 7/(7+53) = 7/60$;
Fisher exact test:
two-sided mid-significance $p = (0.037 + 0.033)/2 = 0.035$; reject H_0;
Pearson χ^2 test don't valid:
$\chi^2 = 7.28$; *df* $= 3$; significance $p = 0.065$; don't reject H_0.
On-line calculator: http://www.quantitativeskills.com/sisa/statistics/fiveby2.htm

The other, a strongly ordered pattern of J objects, $O_1 < O_2 < \ldots < O_J$, for recognising a responder either as an outsider or as a member of some assumed population can be easily constructed basing on established standard ordering, for instance on ordering of ten items of the physical functioning in the SF-36 questionnaire [32]. Several known procedures can be used to prove the level of concordance between estimated responder's ordering versus ordering assumed in a pattern [59]. However in this paper, for significant reasons it was suggested to apply the procedure of pair-wise comparisons with some other fixed object O_x from inside a pattern, with further use of the Friedman test [60]. The first reason is that, under analogous exertion for a respondent, the comparisons usually lead to more reliable estimates than rankings [61]. Moreover, the needed sample size for the Friedman test begins here from $J = 7$ objects in an ordered pattern and only two or three other objects O_x [60]. Thus, the proposed way of verification of a respondent can be made with no more than 21 verifying items, added with this purpose to the core questionnaire. The logic and computational details of the Friedman test are described in [60]. All computations are straightforward, the formulas (3–5) can be easily implemented in any universal spreadsheet, in case of necessity with the use of only basic arithmetical operations. Moreover, the host [60] submits on-line access to the user-friendly calculator for $K = 3$ and $K = 4$ initial rankings, which performs automatically all further calculations of the Friedman test.

In [Tab. 2] for each object from a given pattern in the section named raw initial ranks there were shown results of the three separate evaluations, named X, Y, Z, obtained with the 9-level Likert scale, from score = 1, by step = 1, up to score = 9. For instance, object O_1 obtained scores $x_1 = 1$, $y_1 = 2$ and $z_1 = 3$; object O_2 obtained scores $x_2 = 2$, $y_2 = 3$ and $z_2 = 4$; and so on, up to object O_7 with its scores $x_7 = 7$, $y_7 = 8$ and $z_7 = 9$.

With the aim to verify hypothesis H'_0, that the evaluations X, Y and Z didn't differ with respect to this pattern, at the beginning separately for each object its raw evaluations were changed with their relative ranks. For instance, the ranks of the object O_1 were transformed into relative ranks $x'_1 = 1$, $y'_1 = 2$ and $z'_1 = 3$; because its raw ranks are ordered: $x_1 < y_1 << z_1$. Analogously, for each other object, its minimal raw evaluation was transformed into relative rank = 1; its intermediate raw evaluation into relative rank = 2; and its maximal raw evaluation into relative rank = 3. The Kendall's coefficient of concordance W was estimated with formula (3) as $W = 1$. The test statistic Q was estimated with formula (4) as $Q = 14$. Because the data sample was sufficiently large, the distribution of the test statistic Q can be considered as a close approximation of the chi-square distribution with degree of freedom equal to $df = K - 1$ [60]. Therefore, the significance of the null hypothesis H_0 was estimated with formula (5) as $p = 0.0009$, manifestly smaller than $p = 0.05$. Thus, the null hypothesis H'_0 should be rejected without any serious doubt. It should be concluded that the evaluations X, Y and Z did differ significantly with respect to a pattern under investigation.

$$W = 12 \cdot \sum_k \left(\sum \text{relative.rank}_j|k\right)^2 / J^2 \cdot K \cdot (K^2 - 1)) - 3 \cdot (K+1)/(K-1), \quad (3)$$

$$Q = J \cdot (K - 1) \cdot W, \quad (4)$$

$$p(Q) = p(\chi^2 = Q)|(df = K - 1) \quad (5)$$

where: $j = 1, 2, \ldots, J$; J – number of objects under evaluation; $k = 1, 2, \ldots, K$; K = number of ways of evaluation; relative.rank$_j|k$ – relative rank of j-th object under k-th way of evaluation.

With the aim to verify somewhat different hypothesis H''_0, that the evaluations X, Y and Z were generated by the same latent ordering of the compared objects, at least with an insignificant random error, at the beginning each raw initial evaluation X, Y and Z separately should be transformed into standardized ranks. For instance, object O_1 got standardized rank $y''_1 = 1$ because its raw score $y_2 = 2$ was a minimal Y score, but object O_7 got standardized rank $y''_7 = 7$ because its raw score $y_2 = 8$ was a maximal Y score among the all $J = 7$ of Y scores under investigation. The standardized ranks of the remaining objects and the remaining ways of scoring were defined as usual. Subsequently, the standardized ranks were processed in the same manner as the raw evaluations formerly. For instance, the standardized ranks of the object O_1 were transformed into relative ranks $x'_1 = 2$, $y'_1 = 2$ and $z'_1 = 2$; because its standardized ranks are just the same: $x_1 = y_1 = z_1 = 2$. Thus, the Kendall's coefficient of concordance W was estimated with formula (3) as $W = 0$. The test statistic Q was estimated with formula (4) as $Q = 0$. The significance of the null hypothesis H_0 was estimated with formula (5) as $p \approx 1.0$, manifestly greater than $p = 0.05$. The null hypothesis H''_0 should be accepted without any serious doubt. It should be concluded that the scores X, Y and Z were generated by the same latent ordering of the compared objects.

Tab. 2. Friedman test for exemplary data of K = 3 orderings without ties

pattern	raw initial ranks			relative row ranks			standardized ranks			relative row ranks		
object	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
1	1	2	3	1	2	3	1	1	1	2	2	2
2	2	3	4	1	2	3	2	2	2	2	2	2
3	3	4	5	1	2	3	3	3	3	2	2	2
4	4	5	6	1	2	3	4	4	4	2	2	2
5	5	6	7	1	2	3	5	5	5	2	2	2
6	6	7	8	1	2	3	6	6	6	2	2	2
7	7	8	9	1	2	3	7	7	7	2	2	2
sum	–	–	–	7	14	21	–	–	–	14	14	14

$B = J_2 \cdot K \cdot (K^2 - 1)) = 7 \cdot 7 \cdot 3 \cdot (3 \cdot 3 - 1) = 1176;$
$C = 3 \cdot (K + 1)/(K - 1) = 3 \cdot (3 + 1)/(3 - 1) = 6;$
for raw initial ranks:
$A = 12 \cdot \sum(\text{sum}^2) = 12 \cdot (7 \cdot 7 + 14 \cdot 14 + 21 \cdot 21) = 8236$; $W = A/B - C = 1$;
$Q = J \cdot (K - 1) \cdot W = 7 \cdot (3 - 1) \cdot 1 = 14$; $df = K - 1 = 2$; $p(\chi^2) = 0.0009$;
conclusion: raw initial ranks of the $J = 7$ objects from a pattern differ significantly;
for standardized initial ranks:
$A = 12 \cdot \sum(\text{sum}^2) = 12 \cdot (14 \cdot 14 + 14 \cdot 14 + 14 \cdot 14) = 7056$; $W = A/B - C = 0$;
$Q = J \cdot (K - 1) \cdot W = 0$; $df = K - 1 = 2$; $p(\chi^2) \approx 1.0$;
conclusion: standardized initial ranks of the $J = 7$ objects from a pattern don't differ significantly.

It should be emphasized that the manifestly opposite conclusions for the above null hypotheses H'_0 and H''_0 were both obtained in the approved manner with the same Friedman test on the base of the exactly the same raw data. This occurrence exemplified the first devious trap that was covered in the Friedman test methodology: a researcher should distinguish the real meaning of comparing the raw evaluations versus comparing the standardized ranks of these evaluations.

In the proposed procedure all elements of a pattern, $O_1 < O2 < \ldots < O_J$, $J = 7$, are presented separately, in random sequence, and a responder is asked to compare a presented object with also separately presenting the three other fixed objects O_i; $i = X, Y, Z$; from inside a pattern, using 5-level Likert scale, score 1: $O_i << O_j$; score 2: $O_i < O_j$; score 3: $O_i \approx O_j$; score 4: $O_i > O_j$; score 5: $O_i >> O_j$; were: relation $<<$ denotes a judgement "definitely less ..."; $<$ denotes "rather less ..."; $\approx$ denotes "rather not different ...". Because number J of objects is greater than the number of a Likert scale levels, the occurrence of ties (the same scores for some objects) is inevitable here.

The real-life exemplary data were shown and analysed in [Tab. 3]. As above, in [Tab. 2], also the two different null hypotheses were verified in [Tab. 3]:

H_0': The ranks of the objects assumed in the pattern and all three raw initial ranks didn't differ significantly;

H_0'': The ranks of the objects assumed in the pattern and all three standardized ranks didn't differ significantly.

The significance of the null hypothesis H_0 was estimated here with the formula (5) as $p = 0.011$, manifestly smaller than $p = 0.05$. For that reason, the null hypothesis H_0' should be rejected without any serious doubt. It should be concluded that the raw initial ranks cannot be generated by the same latent ordering of the compared objects as is assumed in the pattern. The significance of the null hypothesis H_0'' was estimated with formula (5) as p 0.99, manifestly greater than $p = 0.05$. For that reason, the null hypothesis H_0'' should be acknowledged without any serious doubt. It should be concluded that all three standardized ranks can be generated by the same latent ordering of the compared objects as is assumed in the pattern.

It should be emphasized that the manifestly opposite conclusions for the above null hypotheses H_0' and H_0'' both were obtained in the approved manner with the same Friedman test on the base of exactly the same raw data. This occurrence exemplified the second devious trap that was covered in the Friedman test methodology: a researcher should distinguish the real meaning of the comparing the raw evaluations defined with the various Likert scales versus comparing the standardized ranks of these evaluations defined with exactly the same Likert scales (that is the same origin and the same number of levels at all used Likert scales).

Tab. 3. Friedman test for exemplary data of K = 4 orderings with some ties

pattern	raw initial ranks			relative row ranks				standard ranks			relative row ranks			
object	X	Y	Z	patt	X	Y	Z	X	Y	Z	patt	X	Y	Z
1	2	1	1	2	4	2	2	1.5	1.5	1.5	1	3	3	3
2	4	2	2	2	4	2	2	3.5	3.5	3.5	1	3	3	3
3	2	1	1	4	3	1.5	1.5	1.5	1.5	1.5	4	2	2	2
4	5	4	3	2.5	4	2.5	1	6	5.5	5.5	1	4	2.5	2.5
5	4	2	2	4	3	1.5	1.5	3.5	3.5	3.5	4	2	2	2
6	5	4	3	4	3	2	1	6	5.5	5.5	3.5	3.5	1.5	1.5
7	5	5	4	4	2.5	2.5	1	6	7	7	3	1	3	3
sum	–	–	–	22.5	23.5	14	10	–	–	–	17.5	18.5	17	17

$B = J^2 \cdot K \cdot (K^2 - 1)) = 7 \cdot 7 \cdot 4 \cdot (4 \cdot 4 - 1) = 2940$;
$C = 3 \cdot (K + 1)/(K - 1) = 3 \cdot (4 + 1)/(4 - 1) = 5$;
for raw initial ranks:
$A = 12 \cdot \sum(\text{sum}^2) = 12 \cdot 1354.5 = 16254$; $W = A/B - C = 0.529$;
$Q = J \cdot (K - 1) \cdot W = 7 \cdot (4 - 1) \cdot 1 = 11.1$; $df = K - 1 = 3$; $p(\chi^2) = 0.011$;
conclusion: raw initial ranks of the $J = 7$ objects from a pattern differ significantly;
for standardized initial ranks:
$A = 12 \cdot \sum(\text{sum}^2) = 12 \cdot 1226.5 = 14718$; $W = A/B - C = 0.006$;
$Q = J \cdot (K - 1) \cdot W = 0.13$; $df = K - 1 = 3$; $p(\chi^2) \approx 0.99$;
thus, standardized initial ranks of the $J = 7$ objects from a pattern don't differ significantly.

Discussion and conclusions

The patient is the primary recipient of treatment, so it is an urgent need to recognize the patient's own perspective on the illness experience and the effects of therapy, as necessary and unique complement to all professional's evaluations. Therefore, in a daily medical practice, questionnaire inquiries are administered regularly with clear practical purpose to support a decision on the course of therapy in very limited groups of patients. This study was focused on how to make results of the questionnaire examinations more reliable and easily understandable to health workers and other professionals with a limited background in the psychometric and statistical methodology.

Generally, this paper proposed an intuitive, yet statistically precise approach to applied questionnaire examinations. The first topic, 'The ground rule: proper instrument + proper procedure + proper attitude', corresponded to typical simple way of reasoning: we can trust in the data obtained from an individual respondent, because these data are only a fragment of the whole data set from questionnaire survey of the proved reliability and validity. This approach can fail, particularly in situation if an individual respondent under examination gave the false answers, but the final scores of these answers satisfiedthe usual formal criterions. Therefore, in the frame of the second topic, ‘Statistical supports for credibility of an individual respondent’, a fresh and innovative approach to task of recognising an individual respondent either as a typical member or as an unusual member of the homogenous group is suggested. The proposed methodology corresponded to somewhat more sophisticated way of reasoning: we can trust in all data obtained from an individual respondent, because the answers of this respondent to a set of the verifying items are in a general agreement (or: in a close agreement) with the acknowledged pattern. The use of known Friedman test is then recommended. The Friedman test can be considered as a nonparametric two-way analysis on ranks. In spite of all its advantages, in practice the Friedman test was not often used, maybe because of the two devious traps that lurk there for an inexperienced researcher. For this reason, the Friedman test procedure, and the associated common misunderstandings were thoroughly explained in this paper with the exemplary data showed in [Tab. 2] and [Tab. 3]. The two exemplary patterns, that is the typical relationship between the inclination to guess Yes-or-No answer versus level on the disagreement in the matter, and the standard Rasch ordering of the items in an applied questionnaire, were based on Authors’ own previous studies in Polish groups.

REFERENCES

[1] Sijtsma K., Future of Psychometrics: Ask What Psychometrics Can Do for Psychology, Psychometrika, 77 (1), pp. 4–12, 2012.

[2] Sijtsma K., Reliability Beyond Theory and Into Practice, Psychometrika, 74 (1), pp. 169–173, 2009.

[3] Borsboom D., The Attack of the Psychometricians, Psychometrika, 71 (3), pp. 425–440, 2006.

[4] Gimeno-Santos E., Frei A., Dobbels F., et al., Validity of instruments to measure physical activity may be questionable due to a lack of conceptual frameworks: a systematic review, Health and Quality of Life Outcomes, 9 (86), 2011. http://www.hqlo.com/content/9/1/86

[5] StatSoft, Inc. Reliability and Item Analysis, in: StatSoft, Inc. (2012). Electronic Statistics Textbook, Tulsa, OK: StatSoft,
WEB: http://www.statsoft.com/textbook/.2012

[6] de Klerk G., Classical test theory (CTT), In Born M., Foxcroft C. D. & Butter R. (Eds.), Online Readings in Testing and Assessment, International Test Commission, 2008. http://www.intestcom.org/Publications/ORTA.php

[7] Streiner D. L., Norman G. R., Health measurement scales a practical guide to their development and use, Oxford University Press, Inc., New York, 1989.

[8] Sijtsma K., On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha, Psychometrika, 74 (1), pp. 107–120, 2009.

[9] Fischer G. H., Molenaar I. W., Rasch Models – Foundations, Recent Developments, and Applications, Springer-Verlag, Berlin, 1995.

[10] Masters G. N., A Rasch model for partial credit scoring, Psychometrika, 47, pp. 149–173, 1982.

[11] Tennant A., Penta M., Tesio L., et al., Disordered Thresholds: An Example from the Functional Independence Measure, Rash Measurement Transactions, 17 (4), pp. 945–948, 2004. http://www.rash.org/rmt/rmt174a.htm,

[12] Chang H. H., Wang C., Book Review [M.D. Reckase (2009) Multidimensional Item Response Theory. New York: Springer], Psychometrika, 76 (3), pp. 504–506, 2011.

[13] Willse J. T., Goodman J. T., Comparison of Multiple-Indicatorrs, Multiple-Causes – and Item Response Theory-Based analyses of Subgroup Differences, Educational & Psychological Measurement, 68 (4), pp. 587–602, 2008.

[14] StatSoft, Inc.Structural Equation Modeling., in: StatSoft, Inc. (2012). Electronic Statistics Textbook, Tulsa, OK: StatSoft.
WEB: http://www.statsoft.com/textbook/

[15] Collins D., Pretesting survey instruments: An overview of cognitive methods, Quality of Life Research, 12, pp. 229–238, 2003.

[16] Mislevy R. J., Verhelst N., Modeling item responses when different subjects employ different solution strategies, Psychometrika, 55, pp. 195–215, 1990.

[17] Rimm E. B., Stampfer M. J., Colditz G. A., et al., Effectiveness of various mailing strategies among nonrespondents in a prospective cohort study, Am J Epidemiol, 131, pp. 1068–1071, 1990.

[18] Gerrie M. P., Belcher L. E., Garry M., Mind the gap: false memories for missing aspects of an event, Applied Cognitive Psychology, 20 (5), pp. 689–696, 2006.

[19] Siegert R. J., McPherson K. M., Taylor W. J., Toward a cognitive-affective model of goal-setting in rehabilitation: is self-regulation theory a key step?, Disabil Rehabil, 26 (20), pp. 1175–1183, 2004.

[20] Ajzen I., The theory of planned behavior, Organ Behav Hum Dec Proc, 50, pp. 179–211, 1991.

[21] De Vries H., Dijkstra M., Kuhlman P., Self-efficacy: the third factor besides attitude and subjective norm as a predictor of behavioral intentions, Health Educ Res, 3, pp. 273–282, 1988.

[22] Ware J. E., Kosinski M., Dewey J. E., How to Score Version 2 of the SF-36 Health Survey, Lincoln, RI, Quality Metric Inc., 2000.

[23] Ware J. E. Jr, Kosinski M., Gandek B., et al., The factor structure of the SF-36 Health Survey in 10 countries: results from the IQOLA Project. International Quality of Life Assessment, J Clin Epidemiol, 51 (11), pp. 1159–65, 1998.

[24] Martin M., Kosinski M., Bjorner J. B., et al., Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale, Quality of Life Research, 16, pp. 647–660, 2007.

[25] Bjorner J., Ware J., Kosinski M., The potential synergy between cognitive models and modern psychometric models, Quality of Life Research, 12, pp. 261–274, 2003.

[26] Keller S. D., Ware J. E. Jr, Bentler P. M., et al., Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: results from the IQOLA Project. International Quality of Life Assessment, J Clin Epidemiol, 51 (11), pp. 1179–1188, 1998.

[27] Hayes V., Morris J., Wolfe C., Morgan M., The SF-36 health survey questionnaire: is it suitable for use with older adults?, Age Ageing, 24 (2), pp. 120–125, 1995.

[28] Lyons R. A., Wareham K., Lucas M., SF-36 scores vary by method of administration: implication for study design, J PublHlth Med, 21, pp. 41–45, 1999.

[29] Hambleton R. K, Patsula L., Increasing the Validity of Adapted Test: Myths to be Avoided and Guidelines for Improving Test Adaptation Practices, J Appl Testing Technology (JATT), 1 (1), pp. 1–30, 1999.

[30] Gimeno-Santos E., Frei A., Dobbels F., et al., Validity of instruments to measure physical activity may be questionable due to a lack of conceptual frameworks: a systematic review, Health and Quality of Life Outcomes, 9 (1), 86, 2011. http://www.hqlo.com/content/9/1/86

[31] Chmiel I., Górkiewicz M., Method for evaluating an individual inclination to guess Yes-or-No answers in case of a diversity of opinion at group of trustworthy responders, in: Bobrowski L., Burzykowski T., Doroszewski J., Enachescu C. (eds). 114-th ICB Seminar – VIII-th International Seminar: Statistics and Clinical Practice, Warszawa, pp. 59–61, 2011.

[32] Górkiewicz M., Chmiel I., Applying Rasch approach to comparative analysis of the of quality life measurements made with Polish version of the SF-36 questionnaire. in: Wybrane Determinanty Pielęgniarstwa, Część II, Sienkiewicz Z., Fidecki W., Wójcik G. (red.), Warszawski Uniwersytet Medyczny, Warszawa, pp. 128–136, 2010.

[33] Chmiel I., Determinants of quality of life following acute pancreatitis, Dysertacja doktorska, promotor: Antoni Czupryna. Uniwersytet Jagielloński, Wydział Nauk o Zdrowiu, Kraków, 2011.

[34] Scott K. M., Sarfati D., Tobias M. I., Haslett S. J., A challenge to the cross-cultural validity of the SF-36 health survey: factor structure in Maori, Pacific and New Zealand European ethnic groups, Soc Sci Med, 51 (11), pp. 1655–1664, 2000.

[35] Georgiadis K., Anthropological demography in Europe. Methodological lessons from a comparative study in Athens and London, Demographic Research, 17 (1), pp. 1–22, 2012. http://www.demographic-research.org/volumes/vol17/1/

[36] Bereziewicz W., Górkiewicz M., How much a priori in a posteriori: scientific recognition with use of the statistical methodology, Cogitatum, 2, pp. 1–9, 2012. on-line: http://filozof.uni.lodz.pl/knf/cogitatum/numer2/cog2bereziewicz.pdf

[37] Halonen K., Pettila V., Leppaniemi A., et al., Long-term health-related quality of life (HRQL) in survivors acute pancreatitis, Intensive Care Med, 29, pp. 782–786, 2003.

[38] Chmiel I., Górkiewicz M., Czupryna A., Brzostek T., Multiple comparisons procedures in analysis of health-related quality of life outcomes., in: Balcerar-Nicolau H., Bobrowski L., Doroszewski J., Kulikowski C. (eds). Lecture Notes of the VII-th ICB Seminar: Statistics and Clinical Practice, Warszawa, pp. 62–67, 2008.

[39] Chmiel I., Górkiewicz M., The Bootstrap and Multiple Comparisons Procedures as Remedy on Doubts about Correctness of ANOVA Results, Applied Medical Informatics, 30 (1), pp. 9–15, 2012.
http://ami.info.umfcluj.ro/index.php/AMI/article/view/352

[40] Górkiewicz M., Using propensity score with receiver operating characteristics (ROC) and bootstrap to evaluate effect size in observational studies, Biocybernetics and Biomedical Engineering, 29 (4), pp. 41–61, 2009.

[41] Chmiel I., Górkiewicz M., Czupryna A., Brzostek T., Vitality and feeling of happiness versus age, gender, physical functioning, and limitations in social role due to physical problems among convalescents after acute pancreatitis, Rocznik Naukowy, 19, Akademia Wychowania Fizycznego i Sportu w Gdańsku, Gdańsk, pp. 79–84, 2009.

[42] Chmiel I., Górkiewicz M., Czupryna A., Brzostek T., Age and gender as predictors of the physical ability among convalescents after acute pancreatitis., in: Fidecki W., Wysokiński M. (eds.) Selected problems of the aging population, Radomska Szkoła Wyższa, Radom, pp. 279–291, 2009.

[43] Chmiel I., Czupryna A., Górkiewicz M., Brzostek T., Health behaviours of covalescents after acute pancreatitis, in: Zdrowie, Kultura Zdrowotna, Edukacja. Czerwiński J., Demel M., Frołowicz T., et al. (eds.), Akademia Wychowania Fizycznego i Sportu w Gdańsku, Gdańsk, 2, pp. 145–150, 2008.

[44] Chmiel I., Czupryna A., Brzostek T., et al., Educational needs of patients after acute pancreatitis (preliminary report), Pielęgniarstwo XXI wieku, 28 (3), pp. 51–56, 2009.

[45] Chmiel I., Czupryna A., Górkiewicz M., Brzostek T., The causes of acute pancreatitis and the range of psychoeducational intervention for convalescents, Medical Studies, 11, pp. 51–56, 2008.

[46] Górkiewicz M, Chmiel I., Dutes of contemporary education of children and youth from perspective of health rehabilitation at convalescents after hard disease, in: Augustyn A., Bodanko A., Niestolik N. (red.n.) Dylematy współczesnego wychowania i kształcenia, Wyd. Akademii Humanistyczno-Ekonomicznej w Łodzi, pp. 113–118, Łódź, 2011.

[47] Dojka E., Górkiewicz M., Pająk A., Psychometric value of CES-D scale for the assessment of depression in Polish population, Psychiatria Polska, 37 (2), pp. 281–292, 2003.

[48] Peña-Sánchez J. N., Domagala A., Górkiewicz M., et al., Adapting a tool in Poland for the measurement of the physicians' career satisfaction, Problemy Medycyny Rodzinnej, 12 (1), pp. 58–65, 2011. http://pmr.org.pl/

[49] Wilczek-Rużyczka E., Czabanowska K., Walewska E., et al., Motivation to work fortifies attitude to motivating patients: Evidence from Leonardo da Vinci program on motivational skills training in health social care., in: Balcerar-Nicolau H., Bobrowski L., Doroszewski J., Kulikowski C. (eds). Lecture Notes of the VII-th ICB Seminar: Statistics and Clinical Practice, Warszawa, 113–119, 2008.

[50] Wilczek-Rużyczka E., Górkiewicz M., Decision variables of psychological model of nurse's to patients, in: Człowiek i jego decyzje, Kłosiński K.A., Biela A. (eds.), Wydawnictwo KUL, Lublin, 179–186, 2009.

[51] Górkiewicz M., Kreiner D.S., Gender Differenes in Creating False Memory under the DRM Paradigm, European Epi-Marker, 11 (2), pp. 6–12, 2007.

[52] Suhonen R., Leino-Kilpi H., Välimäki M., Development and psychometric properties of the Individualized Care Scale, Journal of Evaluation in Clinical Practice, 11 (1), pp. 7–20, 2005.

[53] Rothman M. L., Beltran P., Cappelleri J. C., et al., Patient-reported outcomes: conceptual issues, Value Health, 10 (2), pp. S66–S75, 2007.

[54] Bottomley A., Jones D., Claassens L., Patient-reported outcomes: assessment and current perspectives of the guidelines of the Food and Drug Administration and the reflection paper of the European Medicines Agency, Eur J Cancer, 45, pp. 347–353, 2009.

[55] Wiklund I., Assessment of patient-reported outcomes in clinical trials: the example of health-related quality of life, Fundamental & Clinical Pharmacology, 18 (3), pp. 351–363, 2004.
http://www.ncbi.nlm.nih.gov/pubmed/15147288

[56] Chmiel I., Górkiewicz M., The scope of acceptance by patients motherly and friendly style of nurse's supporting behaviour in palliative care, Problemy Pielęgniarstwa, 18 (4), pp. 11–17, 2010.

[57] Gniadek A., Kozicka M., Górkiewicz M., Unexpected, indirect and seeming associations between factors of the quality of life in elderly individuals, in: Wybrane Determinanty Pielęgniarstwa, Część II. Sienkiewicz Z., Fidecki W., Wójcik G. (eds.), Warszawski Uniwersytet Medyczny, Warszawa, pp. 118–127, 2010.

[58] SISA. On-line Calculators for Scientists. GraphPad Software, Inc., 2002–2012. http://www.quantitativeskills.com/sisa/

[59] StatSoft, Inc.How to Analysis Data with Low Quality or Small Samples, Nonparametric Statistics., in: StatSoft, Inc. Electronic Statistics Textbook, Tulsa, OK: StatSoft, 2012.
http://www.statsoft.com/textbook/nonparametric-statistics/

[60] Lowry R., The Friedman Test for 3 or More Correlated Samples., in: Concepts and Applications, 1998–2012. http://vassarstats.net/textbook/index.html

[61] Böckenholt U., Comparative Judgments as an Alternative to Ratings: Identifying the Scale Origin, Psychological Methods, 9 (4), pp. 453–465, 2004.

Classification issue in the IVF ICSI/ET data analysis

Robert Milewski[1], Paweł Malinowski[1], Anna Justyna Milewska[1], Piotr Ziniewicz[1], Jan Czerniecki[2,3], Piotr Pierzyński[4], Sławomir Wołczynski[4]

[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 of Polish Academy of Sciences in Olsztyn, Poland

[3] Department of Cytobiochemistry, Institute of Biology, University of Bialystok, Poland

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

Abstract. The effectiveness of infertility treatment using IVF ICSI/ET method depends on many different factors. Their identification and classification of individual cases remains a difficult task, despite the use of advanced statistical methods. This article presents the application of Random Forest and SVM classifiers, to analyze the data of patients undergoing the infertility treatment process.

Introduction

There are many methods referring to the general term "data mining methods" which can be used for the analysis of medical data, for example in [5] basket analysis was used on data of patients hospitalized in the gynecological ward. These methods are especially useful in the treatment of outcome prediction, like in [6, 11], were artificial neural networks have been used to predict the success of IVF ICSI/ET treatment. This article focuses on algorithms for classification in order to generate decision rules. Generated rules allow predicting the target class observations, also for new data. Since medical data are analyzed, those rules should have high efficiency and resistance to accidental errors and over-fitting. Therefore, only state-of-the-art classifiers are used: SVM and Random Forest, using their R language implementations. In contrast to [9], no feature selection algorithm is used. To counter over-fitting, cross-validation meta-algorithm was extensively used. Finally, three different imputation methods were used to fill missing data and make classifier work easier.

ISBN 978–83–7431–336–0 ISSN 0860-150X

The medical problem

Infertility is a social disease, which despite the intensive development of medical knowledge and advanced treatment techniques, still affects a significant percentage of couples. One of the contributing factors is postponing parenthood [12]. The chance for getting pregnant decreases with woman's age, mainly due to the decrease in the number and quality of oocytes. The effectiveness of infertility treatments, including the most advanced called In Vitro Fertilization with Intracytoplasmic Sperm Injection and Embryo Transfer (IVF ICSI/ET), is also correlated to woman's age, with success rates averaging at 10–15% pregnancies per treatment cycle in women of 40 and more years of age [10]. Then predictive methods allowing individual prognosis are needed. They could allow to select the best possible treatment approach and reduce the risk of complications.

Material and methods

Data for analysis were collected using the system of electronic registration of information about patients treated for infertility [7], with the statistical module based on artificial neural networks [11]. The system was designed to collect the specialist data, which significantly increased the accuracy and precision of the collected data, and led to increasing the number of recorded features. More recently, such systems have become more popular – they are dedicated to the specificity of the chosen medical unit, such as for instance the system to support clinical-research-teaching unit [14–15].

IVF ICSI/ET data were analyzed using methods which are accessible from the R software (http://www.R-project.org), an open source implementation of the computer language S – either by a native implementation or an interface to existing libraries. Some additional methods were implemented manually. In [Tab. 1] the corresponding R packages along with their version numbers are listed. R version 2.14.2 was used for the analysis.

Tab. 1. Used R packages and their versions

Package	Version	URL
e1071	1.6	cran.r-project.org/web/packages/e1071
randomForest	4.6–6	cran.r-project.org/web/packages/randomForest
VIM	3.0.1	cran.r-project.org/web/packages/VIM

The most frequently used meta-algorithm was cross-validation. The whole dataset was divided randomly to learning and validation part at 7:3 ratio. In order to learn a specified algorithm, a k-fold cross-validation ($k = 10$) was performed on learning data. Learning data was randomly partitioned into k subsamples. Of the k subsamples, a single subsample was retained as the test data for the model, and the remaining $k - 1$ subsamples were used as training data. The cross-validation process was then repeated k (folds) times, with each of the k subsamples used exactly once as the test data. The k results from the folds were then averaged to produce a single estimation.

Two classification algorithms were used in order to predict the outcome:

- Support Vector Machine (further referred as SVM)
- Random Forest (further referred as RF)

The SVM method [1] tries to build a hyperplane in parameter space that separates observations that belong to different classes. It is achieved by maximizing the margin, i.e. distance of hyperplane to nearest training observation of any class. Sometimes such hyperplanedoes not exist. SVM algorithm allows some violation of linear separation by using additional C (cost) parameter. SVM can be also modified to create a non-linear classifier via kernel trick, transforming original parameter space to other. This transformation may be nonlinear and resulted transformed space high dimensional; thus though the classifier is a hyperplane in the high-dimensional (even infinite-dimension) feature space, it may be nonlinear in the original input space. For analysis the Gaussian kernel was chosen with one parameter γ. Those two parameters – C and γ – were selected using grid search over wide range of values and 10-folf cross-validation to find the best-ones. SVM in R language is implemented in package "e1071" [3], and it is an interface to libsvm (version 2.6), and this implementation was used.

Random Forest algorithm [2], proposed by Leo Breiman and Adele Cutler, builds a set of decision trees based on learning data. Prediction of each tree is used as a sort of vote. Whole forest chooses class with majority of votes. Let N be the number of observation in training set, and M – the number of features. Each tree is grown as follows:

- pick up a sample of N observations at random with replacement – selected tree will be trained on this sample only
- at each node pick up *mtry* features (a number much smaller than M – a square root by default for classification) at random, and find the best split using those *mtry* features only
- grow each tree to full extent (this is also recommended for classification, but can be changed).

When full forest is grown, a version of distance matrix, called proximities, can be computed based on it. All the data are put down in each tree. If two observations are in the same terminal node, then their proximity is increased by one. The final result is normalized by dividing them by the number of trees.

Among many algorithm parameters which can be set for further tuning, the following were selected:

- number of trees,
- number of features at each node split,
- minimum number of observations per final node.

Those parameter were tuned using again grid search and 10-fold cross-validation. Random Forest in R language is implemented in package "randomForest" [4], based on original Fortran 77 implementation by Breiman, and this implementation was used.

Three algorithms were used for data imputation:

- a "standard" one
- kNN-based
- proximity-based

A "standard" algorithm imputes missing value based on mean (for numerical features), median (ordinal) or mode (categorical). It was partially implemented manually due to lack of cross-validation friendly version of this procedure in R language. The kNN-based algorithm tries to fill data in similar way to the standard algorithm, but utilizing only the part of observations. In given observation missing data are filled with values based on values from its k nearest neighbors only. Those neighbors are found by using version of Gover distance. This algorithm in R language is implemented in "VIM" package [13]. Proximity-based algorithm is also similar to the "standard" one. Algorithm, implemented in package "randomForest", runs as follows:

- fill missing data using "standard" method,
- run random forest on such data to find proximities,
- correct previously filled data based on calculated proximities. Use weighted mean (for numerical feature; weights are proximities) or the category with the largest average proximity (for categorical or ordinal data),
- calculate step 2 and 3 chosen (20) number of times.

Although standard imputation procedure is fairly simple, the next two have free parameters and required further tuning. For this purpose, 10-fold cross-validation procedure was applied on for each set of parameters. At each step additional 5% of filled test data were randomly marked as missing, and imputation algorithm with given set of parameters tried to fill it, based on train data observations. The objective was to minimize mean (relative) error of missing values prediction, based on 10 folds. After the best set of parameters was found (for second and third algorithm), the whole dataset was imputed based on train and test observations only, using those three methods.

Data preparation

The dataset has contained 1445 observations and 150 features. About 22% of original data was missing. Features containing more than 80% missing data were removed from dataset. Further investigation revealed features with only 1 level, which were also removed. Resulted dataset contained 108 features and only 5% of the data were missing. This dataset is symbolically depicted on [Fig. 1]. In the main part of [Fig. 1] black color means missing data and colors from white to gray means different levels of given feature. The first feature is the dependent one – treatment outcome. On the right side there is a barcode-like indicator of learning and validation division of the dataset. Black color indicates observation which was taken to the learning set, while white color is reserved for validation data.

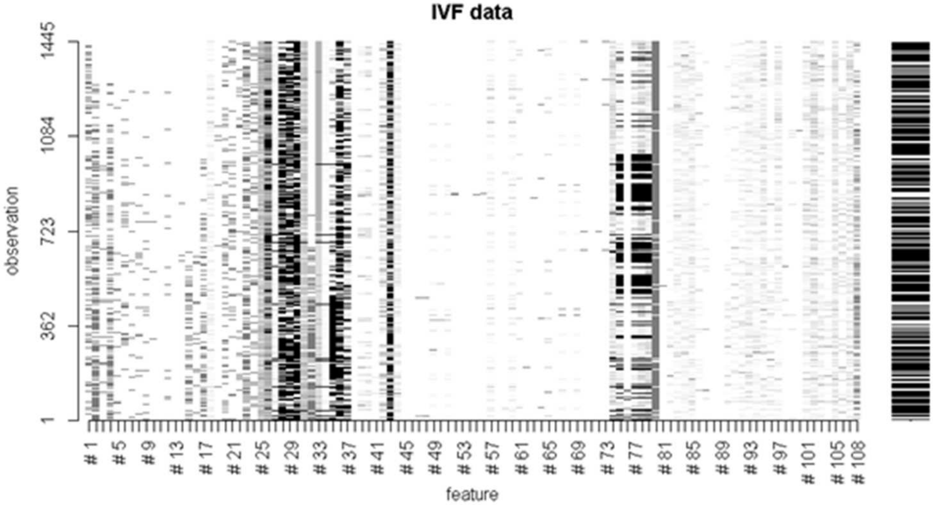


Fig. 1. Analyzed dataset with validation division

The next step was imputation of missing data. Three algorithms were chosen:

- “standard”one, further referred as “STD imp”
- kNN-based, further referred as “kNN imp”
  - k in 20–65 range
- proximity-based, further referred as “RF imp”
  - tree count in 1000–3000 by 200 range
  - *mtry* in 5–24 range

[Fig. 2] presents 10-fold cross-validation mean (relative) prediction error for kNN imp and RF imp procedures. Best found k equals to 42 at mean error slightly less than 30%. For proximity-based imputation procedure, the best set of parameters includes 2000 trees and 11 features at each split node at mean error around 25%.

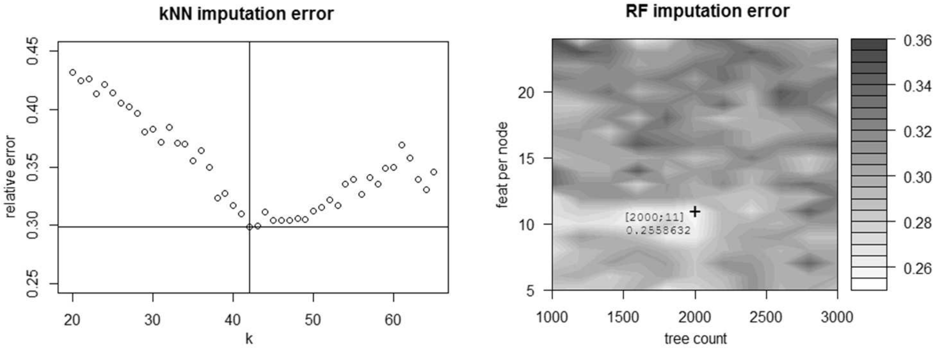


Fig. 2. kNN imputation cross-validation error

kNN-based imputation procedure had almost one and a half greater error range than proximity-based. After the best set of parameters was found, whole dataset was imputed based on learningdata only. This created 6 datasets, which were used for further analysis.

Data Classification

Two parameters of chosen SVM classifier – C and γ were tuned using grid search. Range of search was the same for both parameters:

$$range = \{2^{-20}, 2^{-19.8}, 2^{-19.6}, \ldots, 2^{0} = 1, \ldots, 2^{19.6}, 2^{19.8}, 2^{20}\}$$

SVM classifier was trained with each combination (out of 40201) of parameters using 10-fold cross-validation on learning datasets generated by 3 different imputation methods. In [Fig. 3] ean classification error is presented for those datasets in specified range of parameters with best model marked with (+).

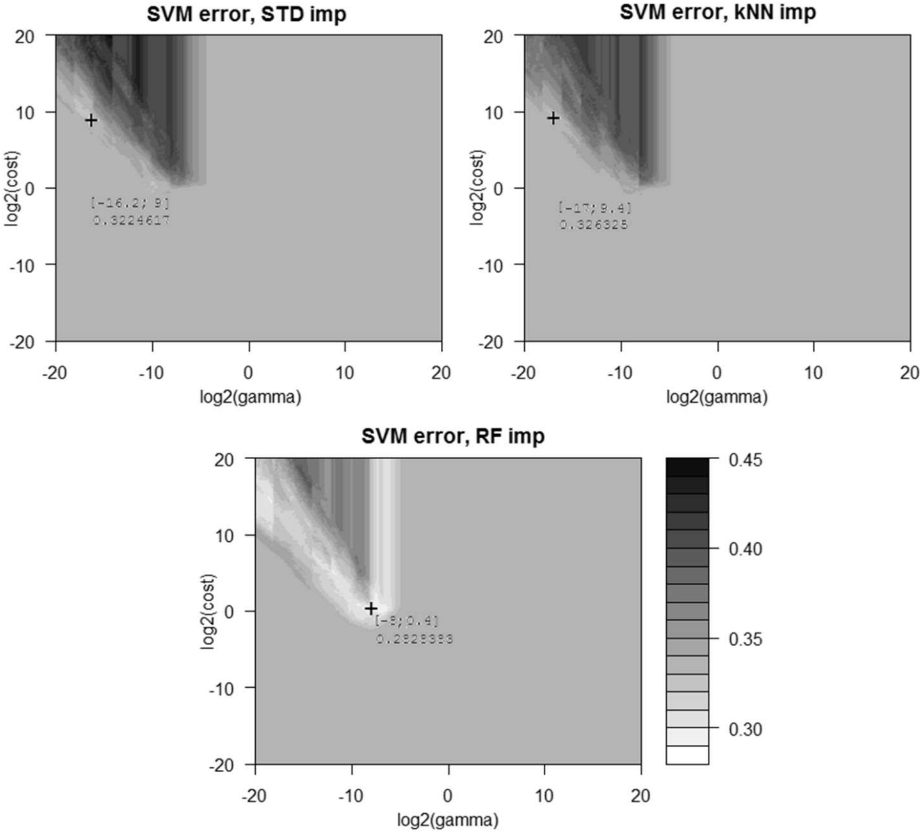


Fig. 3. SVM classifier cross-validation error

In wide range of parameters classification error was around 33%. This corresponds to outcome ratio in the whole dataset. In fact, for most of the parameters in the studied range, trained SVM classifier has predicted lack of pregnancy for all cases. This includes the "best" results on STD and kNN imputed datasets. Results on validation observations confirmed that behavior. Only on RF imputed dataset SVM yield different and superior result, which is shown in [Tab. 2].

Random Forest algorithm was trained using grid search on following parameters range

- number of trees
  $ntree = \{1000, 1200, 1400, \ldots, 2600, 2800, 3000\}$
- number of used features at each split node
  $mtry = \{5, 6, 7, \ldots, 22, 23, 24\}$
- minimal number of cases in tree terminal node
  $nodesize = \{1, 2, 3, 4, 5, 6, 7\}$.

Tab. 2. SVM accuracy on RF-imp validation dataset

Outcome prediction on RF-imp validation observations		Predicted outcome		Accuracy
		no	yes	
Observed outcome	no	261	27	90.6%
	yes	97	49	33.6%
Accuracy		72.9%	64.5%	71.4%

Again, 10-fold cross-validation procedure was used which each combination (out of 1540) of parameters on learning datasets generated by 3 different imputation methods, to find the best one. Because it is difficult to visualize 3-dimension parameter space, results for only two *nodesize* values will be presented per each imputation method, which gave minimum mean error. [Fig. 4] presents result for STD-imp learning dataset, which are also presented in [Tab. 3] as full contingency table for validation dataset.

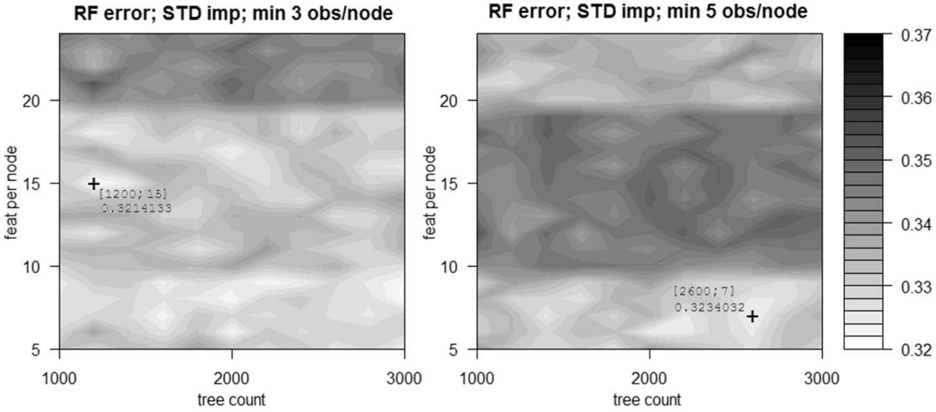


Fig. 4. RF cross-validation error on STD-imputed dataset

Tab. 3. RF accuracy on STD-imp validation dataset

Outcome prediction on STD-imp validation observations		Predicted outcome		Accuracy
		no	yes	
Observed outcome	no	266	22	92.3%
	yes	124	22	15.1%
Accuracy		68.2%	50%	66.4%

[Fig. 5] presents results for kNN-imp learning dataset, which are also presented in [Tab. 4] as full contingency table for validation dataset.

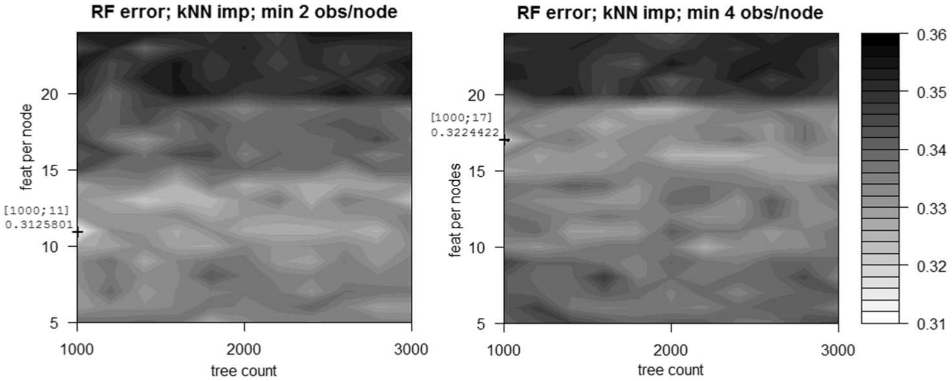


Fig. 5. Random forest mean error on kNN-imputed dataset

Tab. 4. RF accuracy on kNN-imp validation dataset

Outcome prediction on kNN-imp validation observations		Predicted outcome		Accuracy
		no	yes	
Observed outcome	no	265	23	92.0%
	yes	128	18	12.3%
Accuracy		67.4%	43.9%	65.2%

[Fig. 6] presents results for RF-imp learning dataset, which are also presented in [Tab. 5] as full contingency table for validation dataset.

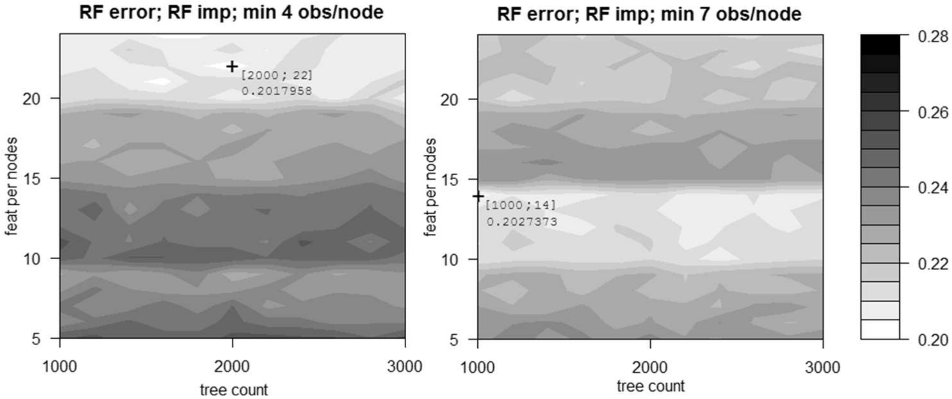


Fig. 6. Random forest mean error on proximity-imputed dataset

Tab. 5. RF accuracy on RF-imp validation dataset

Outcome prediction on RF-imp validation observations		Predicted outcome		Accuracy
		no	yes	
Observed outcome	no	227	61	78.8%
	yes	30	116	79.5%
Accuracy		88.3%	65.5%	79.0%

Conclusions

Proximity based imputation algorithm clearly outperforms other methods, despite relative error of prediction only 5% less than kNN-based one. It was only the imputation method, which yield sensible result with SVM classifier. SVM classifier performance was disappointing. For wide range of parameters this method predicted only lack of pregnancy. Although error rates for RF classifier on STD and kNN imputed datasets was similar to those obtained by SVM, the first algorithm actually tries to distinguish outcomes of observations. Cross-validation means errors for RF classifier was almost the same over the whole range of checked parameters. Relative differences reached 5 or 8 percent only. Change of *nodesize* parameter only slightly changed this error. It is worth noting that the RF classifier preferred higher *mtry* value on RF-imp dataset. Finally, the use of the RF classifier and RF-based imputation procedure leads to superior result: almost 80% accuracy on learning (note that this is mean accuracy based on 10 folds) and 79% on validation dataset. This error is unequally distributed among negative and positive outcome on the validation dataset. When the algorithm predicts lack of pregnancy, there is ~88% probability, that this answer is a correct one, but for success this probability is only 65.5%. This behavior is consistent with previous studies [6, 8] on this dataset.

Further studies are required to find better algorithms for classification and imputation on IVF data. This article did not include feature selection algorithms; including them in analysis may also yield better results.

REFERENCES

[1] Boser B. E., Guyon I. M., Vapnik V. N., A training algorithm for optimal margin classifiers, In Haussler D. (editor); 5th Annual ACM Workshop on COLT, Pittsburgh, PA, ACM Press, pp. 144–152, 1992.

[2] Breiman L., Random Forests,Machine Learning, 45(1), 2001.

[3] Dimitriadou E., Hornik K., Leisch F., et al., Misc Functions of the Department of Statistics, TU Wien, R package version 1.6., 2011.
http://CRAN.R-project.org/package=e1071

[4] Liaw A. and Wiener M., Classification and Regression by randomForest, R News, 2 (3), pp. 18–22, 2002.

[5] Milewska A.J., Górska U., Jankowska D., et al., The use of the basket analysis in a research of the process of hospitalization in the gynecological ward, Studies in Logic, Grammar and Rhetoric, 25 (38), pp. 83–98, 2011.

[6] Milewski R., Jamiołkowski J., Milewska A. J., et al., 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), pp. 900–906, 2009.

[7] Milewski R., Jamiołkowski J., Milewska A. J., et al., The system of electronic registration of information about patients treated for infertility with the IVF ICSI/ET method, Studies in Logic, Grammar and Rhetoric, 17 (30), pp. 225–239, 2009.

[8] Milewski R., Malinowski P., Milewska A.J., et al., Nearest neighbor concept in the study of IVF ICSI/ET treatment effectiveness, Studies in Logic, Grammar and Rhetoric, 25 (38), pp. 49–57, 2011.

[9] Milewski R., Malinowski P., Milewska A.J., et al., The usage of margin-based feature selection algorithm in IVF ICSI/ET data analysis, Studies in Logic, Grammar and Rhetoric, 21 (34), pp. 35–46, 2010.

[10] Milewski R., Milewska A.J., Domitrz J., et al., In vitro fertilization ICSI/ET in women over 40, Przegląd Menopauzalny, 2(36), pp. 85–90, 2008.

[11] Milewski R., Milewska A.J., Jamiołkowski J., et al., The statistical module for the system of electronic registration of information about patients treated for infertility using the IVF ICSI/ET method, Studies in Logic, Grammar and Rhetoric, 21(34), pp. 119–127, 2010.

[12] te Velde E.R., Pearson P.L., The variability of female reproductive ageing, Human Reproduction Update, 8 (2), pp. 141–154, 2002.

[13] Templ M., Alfons A., Kowarik A. and Prantner B., VIM: Visualization and Imputation of Missing Values. R package version 3.0.1.
http://CRAN.R-project.org/package=VIM, 2012.

[14] Ziniewicz P., Malinowski P., Milewski R., et al., Clinical department information system's internal structure, Studies in Logic, Grammar and Rhetoric, 25 (38), pp. 191–200, 2011.

[15] Ziniewicz P., Malinowski P., Mnich S. Z., et al., Clinical department information system development, Studies in Logic, Grammar and Rhetoric, 2 (34), pp. 129–142, 2010.

Ordinal logistic regression for the analysis of skin test reactivity to common aeroallergens

Dorota Citko[1], **Anna Justyna Milewska**[1], **Jolanta Wasilewska**[2], **Maciej Kaczmarski**[2]

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

[2] Department of Pediatrics, Gastroenterology and Allergology, Medical University of Bialystok, Poland

Abstract. Clinical research is commonly focused on searching for potential factors and illustrates how they affect a patient's condition. In many epidemiological research studies the response variable is categorical, with more than two categories when there is a natural order among the response categories. In these cases, the ordinal logistic regression models may be employed. In practice, the mostly used type of model is a proportional odds model. The model makes assumptions about the nature of the relationship between the response variable and the prognostic factors. If the proportional odds assumption is violated, generalized ordered logit models may be an option, for example, the partial proportional odds model. The study uses the proportional odds model to examine the dust mite sensitization, whereas the partial proportional model is to examine grass pollen sensitization in children's population. The explanatory variables are: year of a test performance, gender, age and season of birth. The considered models have revealed significant factors that influence dust mite and grass pollen sensitization.

Introduction

Facing an enormous complexity and multidimensional nature of the surrounding reality, it is hard to notice cases when a singular variable describes and explains a particular phenomenon. In the case when we consider more than one explanatory variable, a statistical analysis could be supported by various multiply methods. The application of multiple regression models in medical research has greatly increased in recent years [1–2], especially the use of multiply linear regression for continuous response, logistic regression for binary response, and Cox's proportional hazards model [3] for censored response. These models allow to analyze simultaneously the effect of several explanatory variables on a response variable [1].

ISBN 978–83–7431–336–0 ISSN 0860-150X

Logistic regression is used to model the binary response variable. Generalization of the logistic regression forms categorical responses with more than two categories. When there is no natural order among the response categories, nominal logistic regression models are used. When response categories are ordered then the logits can utilize the ordering [4]. In many epidemiological studies the response variable is ordinal, for example severity of disease, quality of life in interval scale, health condition indicator [5]. In these cases the ordinal logistic regression models should be employed. This results in models having simpler interpretations and potentially greater power than the nominal logistic regression models.

There are several ordinal logistic models, such as: cumulative logit model, proportional odds model, continuation ratio logit model, adjacent category logit model. Nevertheless, these models have been rarely utilized in biomedical and epidemiological research [5–8].

In practice, the mostly used type of ordinal logistic regression model is the proportional odds model because of the simplicity of its interpretation [1, 4, 8–10]. However, the proportional odds model makes assumptions about the nature of the relationship between the response variable and the prognostic factors. If the proportional odds assumption is violated, the results of this regression can be misleading or have no meaning at all, and generalized ordered logits models are an option. However, the goodness-of-fit verification of regression models is rarely used in medical research [1].

The study applies the proportional odds model for the analysis of the dust mite allergy and partial proportional odds model for a grass pollen allergy.

Ordinal logistic regression

Suppose that response Y has J categories and the probability for category i is given by $P(Y = i) = \pi_i$ for $i = 1, \ldots, J$. Also consider explanatory variables $x_1, \ldots, x_p$. Sometimes, there may be a latent continuous variable Y for which the cutpoints $C_1, \ldots, C_{J-1}$ define J ordinal categories with associated probabilities $\pi_1, \ldots, \pi_J$ (with $\sum_{j=1}^{J} \pi_i = 1$) [10].

A cumulative probability for Y is the probability that Y falls at or below a particular point. For outcome category j, the cumulative probability is $P(Y \leq j) = \pi_1 + \ldots + \pi_j$, $j = 1, \ldots, J$, where $P(Y \leq 1) \leq P(Y \leq 2) \leq \ldots \leq P(Y \leq J) = 1$. The logits of the cumulative probabilities, called cumulative logits, are

$$\log it[P(Y \leq j)] = \ln\left[\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right] = \ln\left[\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}\right],$$

$$j = 1, \ldots, J - 1.$$

The cumulative logit model is given by

$$\log it[P(Y \leq j)] = \ln\left[\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}\right] = \alpha_j + \beta_{j1}x_1 + \ldots + \beta_{jp}x_p,$$

$$j = 1, \ldots, J - 1.$$

If the intercepts α_j depend on the category j, but the other regression coefficients for explanatory variables do not depend on j, then the model is

$$\ln\left[\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}\right] = \alpha_j + \beta_1 x_1 + \ldots + \beta_p,$$

$$j = 1, \ldots, J - 1.$$

This is called the proportional odds model. It is based on the assumption that the effects of the covariates $x_1, \ldots, x_p$ are the same for all categories, on the logarithmic scale. The proportional-odds assumption is also called the parallel lines assumption. This assumption must be tested for each covariable separately and in the final model, using for example the Brant test.

When the proportional odds model fits well, it requires a single parameter for x_i rather than $J - 1$ parameters to describe the effect of x_i. If it does not, the partial proportional odds model is recommended [11–12]. This model allows some covariables with the proportional odds assumption to be modelled, but for the covariables failed to perform the proportional odds assumption, it is augmented by a coefficient (γ), which is the effect associated with each j'th cumulative logit, adjusted by the other covariables [9]. In the partial proportional odds model some of the β coefficients can be the same for all categories, while others can differ [13].

Statistics for goodness-of-fit for ordinal regression models are:

1. Chi-square statistic

$$X^2 = \sum_{i=1}^{N} r_i^2 = \frac{(o_i - e_i)^2}{e_i},$$

where $r_i = \frac{o_i - e_i}{\sqrt{e_i}}$ are the Pearson chi-squared residuals; o_i and e_i are the observed and expected frequencies for $i = 1, \ldots, N$; N is J times the number of distinct covariate patterns;

2. Deviance $D = 2[(l(b_{\text{max}}) - l(b)]$, where $l(b)$ is the maximum values of the log-likelihood function for the fitted model and $l(b_{\text{max}})$ is the maximum values of the log-likelihood function for the maximal model;
3. Likelihood ratio chi-square statistic $C = 2[l(b) - l(b_{\text{min}})]$, where $l(b_{\text{min}})$ is the maximum values of the log-likelihood function for the minimal model;
4. Pseudo $R^2 = \dfrac{l(b_{\text{min}}) - l(b)}{l(b_{\text{min}})}$.

If the model fits well then both X^2 and D have, asymptotically, the distribution $\chi^2(N - p)$ where p is the number of parameters estimated. C has the asymptotic distribution $\chi^2[p - (J - 1)]$ [10].

A proportional odds model for the analysis of dust mite sensitization in children

The processed data belong to Department of Pediatrics, Gastroenterology and Allergology at the Medical University of Bialystok. Skin prick tests results performed in 2779 children patients were discussed. The population of 1239 patients derives from 1998' and 1540 of them have been diagnosed in 2008'. Skin prick testing was performed using the most common aeroallergens: dust mite and grass pollen. A recorded wheal size for each aeroallergen was treated a desirable result of skin test reactivity. Reactions were considered positive if the wheal was at least 3 mm and was further classified as mild sensitization ⟨3 mm, 6 mm), moderate sensitization ⟨6 mm, 9 mm), severe sensitization ($\geq$9 mm) [14–15]. The results were analyzed separately by: year of a test performance, gender, age, and season of birth. Year of a test performance, gender and age were significantly associated with the degree of skin reactivity to dust mites [Tab. 1]. Gender, age and season of birth were significantly associated with the degree of skin reactivity to grass pollen [Tab. 2].

The study uses the proportional odds model to examine the dust mite sensitization, and next the partial proportional odds model was used to verify the grass pollen sensitization in children's population.

For the first model as a response variable, the dust mite reactivity is classified into four categories: 0 = negative, 1 = mild, 2 = moderate, 3 = severe. A model is constructed using three explanatory variables: year of test performance (1998, 2008), gender (female, male), and a quantitative variable age (in years). The proportional odds model results are shown in [Tab. 3].

Tab. 1. Dust mite sensitization in relation to year, gender, season of birth and age

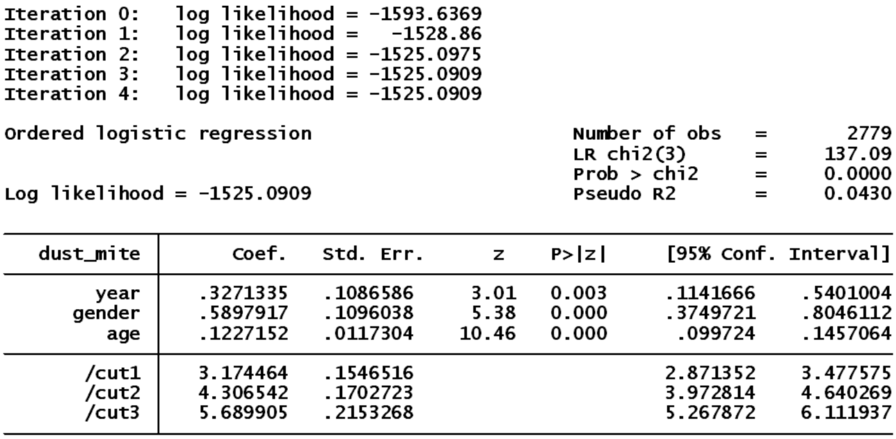
Co-variable		Dust mite				p-value
		Degree of skin test reactivity				
		Negative n (%)	Mild n (%)	Moderate n (%)	Severe n (%)	
Year	1998 2008	1070 (86.4) 1272 (82.6)	111 (9) 162 (10.5)	40 (3.2) 80 (5.2)	18 (1.4) 26 (1.7)	0.026
Gender	F M	1152 (87.5) 1190 (81.4)	112 (8.5) 161 (11)	39 (2.9) 81 (5.6)	15 (1.1) 29 (2)	0.000
Season of birth	Spring Summer Autumn Winter	643 (84.6) 615 (84.9) 542 (84) 537 (83.3)	72 (9.5) 66 (9.1) 67 (10.4) 68 (10.5)	36 (4.7) 33 (4.6) 25 (3.9) 26 (4)	9 (1.2) 10 (1.4) 11 (1.7) 14 (2.2)	0.885
Age	mean/median	6.9/6	8.8/8	9.3/9.5	10.6/10.5	0.000

Tab. 2. Grass pollen sensitization in relation to year, gender, season of birth and age

Co-variable		Grass pollen				p-value
		Degree of skin test reactivity				
		Negative n (%)	Mild n (%)	Moderate n (%)	Severe n (%)	
Year	1998 2008	1030 (83.1) 1307 (84.8)	144 (11.6) 157 (10.2)	39 (3.2) 56 (3.5)	26 (2.1) 20 (43.5)	0.198
Gender	F M	1146 (87.1) 1191 (81.5)	117 (8.9) 184 (12.6)	39 (2.9) 56 (3.8)	16 (1.1) 30 (2.1)	0.001
Season of birth	Spring Summer Autumn Winter	616 (81.1) 628 (86.7) 552 (85.6) 536 (83.1)	103 (13.5) 67 (9.3) 63 (9.8) 68 (10.5)	31 (4.1) 22 (3) 20 (3.1) 22 (3.4)	10 (1.3) 7 (1) 10 (1.5) 19 (3)	0.019
Age	mean/median	7/6	7.7/7	9.1/8	10.3/10	0.000

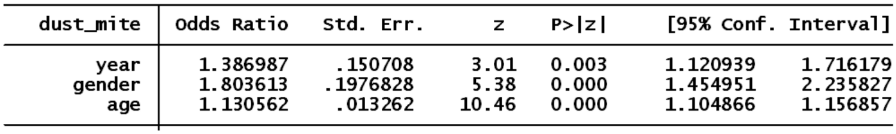
The maximum value of the log-likelihood function for the null model is -1593.64 [Tab. 3] and for the fitted model is -1525.09, giving the likelihood ratio chi-squared statistic $C = 137.09$. The p-value for the Likelihood Ratio Chi-Square test (< 0.0001) showing the overall importance of the explanatory variables. All covariates: year, gender and age are found significant [Tab. 3]. That proves all the explanatory variables used in the model have significantly influenced the dust mite sensitization. The positive coefficients for covariates mean that the likelihood of the dust mite sensitization did increase in time, for boys and older children.

Tab. 3. Results of the proportional odds model according to dust mite sensitization



```
Iteration 0:   log likelihood = -1593.6369
Iteration 1:   log likelihood =   -1528.86
Iteration 2:   log likelihood = -1525.0975
Iteration 3:   log likelihood = -1525.0909
Iteration 4:   log likelihood = -1525.0909

Ordered logistic regression                    Number of obs   =       2779
                                               LR chi2(3)      =     137.09
                                               Prob > chi2     =     0.0000
Log likelihood = -1525.0909                    Pseudo R2       =     0.0430

------------------------------------------------------------------------------
   dust_mite |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .3271335   .1086586     3.01   0.003     .1141666    .5401004
      gender |   .5897917   .1096038     5.38   0.000     .3749721    .8046112
         age |   .1227152   .0117304    10.46   0.000      .099724    .1457064
-------------+----------------------------------------------------------------
       /cut1 |   3.174464   .1546516                      2.871352    3.477575
       /cut2 |   4.306542   .1702723                      3.972814    4.640269
       /cut3 |   5.689905   .2153268                      5.267872    6.111937
------------------------------------------------------------------------------
```

Tab. 4. Odds ratios of the proportional odds model according to dust mite sensitization



```
------------------------------------------------------------------------------
   dust_mite | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   1.386987    .150708     3.01   0.003     1.120939    1.716179
      gender |   1.803613   .1976828     5.38   0.000     1.454951    2.235827
         age |   1.130562    .013262    10.46   0.000     1.104866    1.156857
------------------------------------------------------------------------------
```

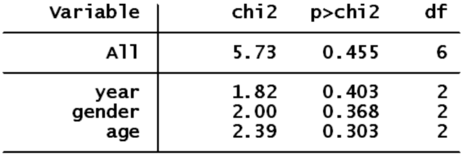
The proportional odds ratios with the 95% confidence intervals for the ordered logit model are given in [Tab. 4]. For children whose tests were performed in 2008', the odds of a severe skin reactivity to dust mites versus the combined categories such as: moderate, mild or negative skin reactivity are 1.39 (95% CI: 1.12 – 1.72) times greater than for children involved in tests in 1998', given the other variables are held constant in the model. Likewise, for the children of 2008', the odds of severe and moderate skin reactivity to dust mites versus the combined categories: mild and negative skin reactivity are 1.39 (95% CI: 1.12 – 1.72) times greater than for the children of 1998'. For the children of 2008', the odds of positive skin reactivity to dust mites (severe, moderate and mild) versus the negative skin reactivity are 1.39 (95% CI: 1.12 – 1.72) times greater than for the children of 1998'.

For boys, the odds of severe skin reactivity to dust mites versus the combined categories: moderate, mild and negative skin reactivity are 1.8 (95% CI: 1.45 – 2.24) times greater than for girls. Also, for boys, the odds of severe and moderate skin reactivity to dust mites versus the combined categories: mild and negative skin reactivity are 1.8 (95% CI: 1.45 – 2.24) times greater than for girls. For boys, the odds of positive skin reactivity to dust mites (severe, moderate and mild) versus the negative skin reactivity are 1.8 (95% CI: 1.45 – 2.24) times greater than for girls.

For one unit increase in age, the odds of severe skin reactivity to dust mites versus the combined categories like moderate, mild and negative skin reactivity are 1.13 (95% CI: 1.1 – 1.16) times greater. Similarly, for one unit increase in age, the odds of severe and moderate skin reactivity to dust mites versus the combined categories mild and negative skin reactivity are 1.13 (95% CI: 1.1 – 1.16) times greater. For one unit increase in age, the odds of positive skin reactivity to dust mites (severe, moderate and mild) versus the negative skin reactivity are 1.13 (95% CI: 1.1 – 1.16) times greater, given the other variables are held constant in the model.

Brant Test is conducted to check the parallel regression assumption [Tab. 5]. The test yielded p = 0.455 indicating that we have not violated the proportional odds assumption and the model is appropriate for this data.

Tab. 5. Brant test of parallel regression assumption for dust mite sensitization



```
    Variable |      chi2   p>chi2    df
-------------+--------------------------
         All |      5.73    0.455     6
-------------+--------------------------
        year |      1.82    0.403     2
      gender |      2.00    0.368     2
         age |      2.39    0.303     2
----------------------------------------
```

The values of the goodness of fit statistics [Tab. 6], the deviance for the fitted model D = 221 and the chi-squared statistics $X^2 = 189.99$, compared to the distribution $\chi^2(213)$ indicates that the model provides a good description of the data.

Tab. 6. Goodness of fit statistics

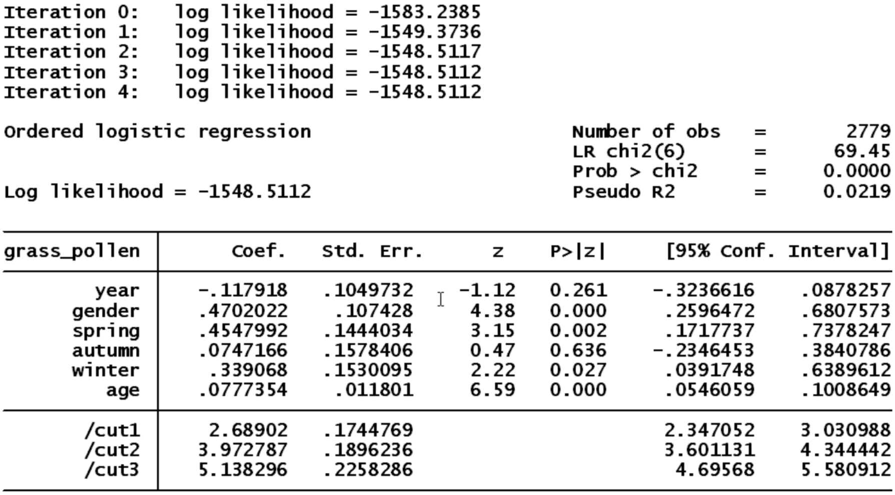
Statistic	Df	Statistic	Statistic/Df
Deviance	213	221	1.037539
Pearson Chi-square	213	189.99	0.891963

A partial proportional odds model for the analysis of grass pollen sensitization in children

Now, as a dependent variable, a grass pollen sensitization inducted wheal size classified into four categories was considered: 0 = negative, 1 = mild, 2 = moderate, 3 = severe. A model was constructed using four explanatory variables: year of test performance (1998, 2008), gender (female,

male), a quantitative variable age (in years) and a variable with more than two categories: season of birth (spring, summer, autumn, winter). For the season of birth variable indicator variables were created, considering summer as a point of references. The proportional odds model results for grass pollen are shown in [Tab. 7].

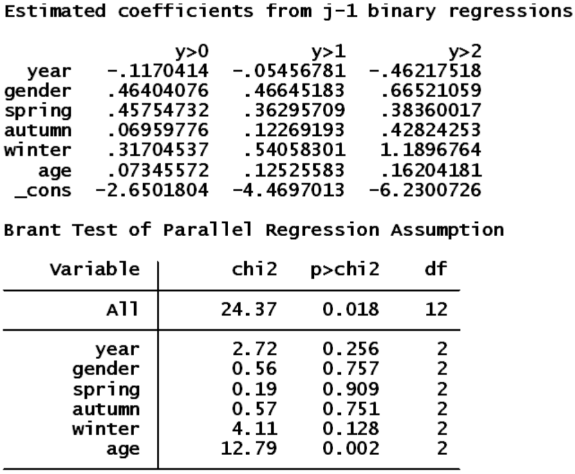
Tab. 7. Results of the proportional odds model according to grass pollen sensitization



```
Iteration 0:   log likelihood = -1583.2385
Iteration 1:   log likelihood = -1549.3736
Iteration 2:   log likelihood = -1548.5117
Iteration 3:   log likelihood = -1548.5112
Iteration 4:   log likelihood = -1548.5112

Ordered logistic regression                       Number of obs   =       2779
                                                  LR chi2(6)      =      69.45
                                                  Prob > chi2     =     0.0000
Log likelihood = -1548.5112                       Pseudo R2       =     0.0219

------------------------------------------------------------------------------
 grass_pollen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
         year |   -.117918   .1049732    -1.12   0.261    -.3236616    .0878257
       gender |   .4702022    .107428     4.38   0.000     .2596472    .6807573
       spring |   .4547992   .1444034     3.15   0.002     .1717737    .7378247
       autumn |   .0747166   .1578406     0.47   0.636    -.2346453    .3840786
       winter |    .339068   .1530095     2.22   0.027     .0391748    .6389612
          age |   .0777354    .011801     6.59   0.000     .0546059    .1008649
--------------+---------------------------------------------------------------
        /cut1 |    2.68902   .1744769                      2.347052    3.030988
        /cut2 |   3.972787   .1896236                      3.601131    4.344442
        /cut3 |   5.138296   .2258286                       4.69568    5.580912
------------------------------------------------------------------------------
```

In the beginning the results of Brant test of parallel regression assumption should be performed.

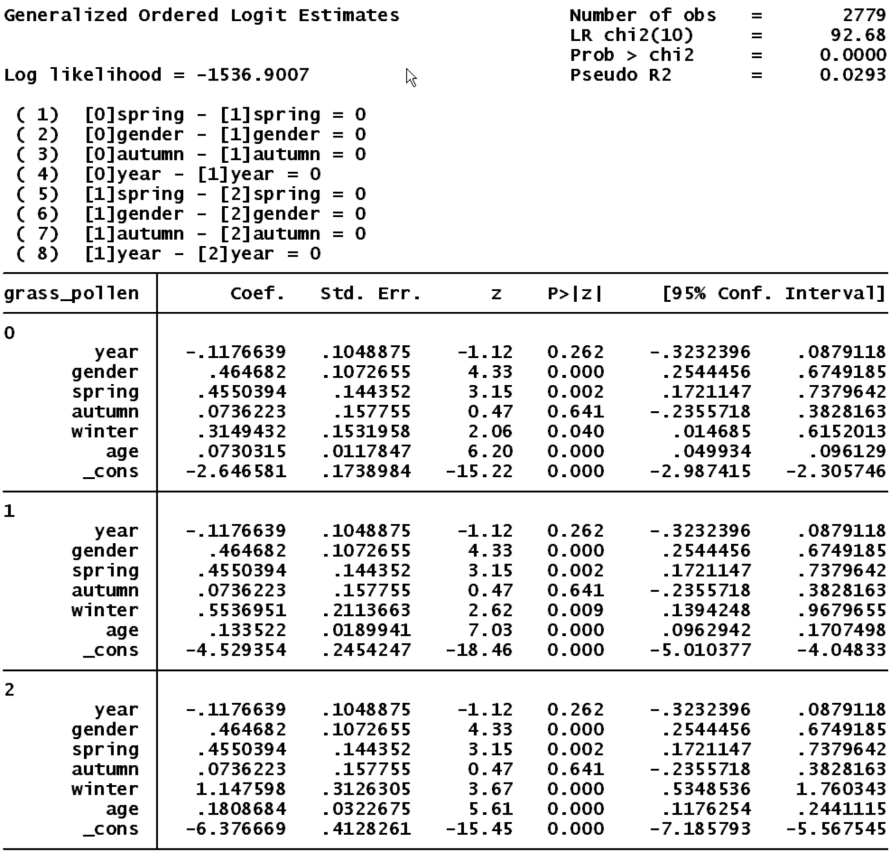
Tab. 8. Brant test of parallel regression assumption for grass pollen sensitization



```
Estimated coefficients from j-1 binary regressions

               y>0          y>1          y>2
  year   -.1170414  -.05456781  -.46217518
gender   .46404076   .46645183   .66521059
spring   .45754732   .36295709   .38360017
autumn   .06959776   .12269193   .42824253
winter   .31704537   .54058301   1.1896764
   age   .07345572   .12525583   .16204181
 _cons  -2.6501804  -4.4697013  -6.2300726

Brant Test of Parallel Regression Assumption

    Variable |      chi2   p>chi2    df
-------------+--------------------------
         All |     24.37    0.018    12
-------------+--------------------------
        year |      2.72    0.256     2
      gender |      0.56    0.757     2
      spring |      0.19    0.909     2
      autumn |      0.57    0.751     2
      winter |      4.11    0.128     2
         age |     12.79    0.002     2
----------------------------------------
```

The Brant test [Tab. 8] yielded p = 0.018, shows that the assumption of the parallel-lines model are violated, but the main problems seem to be with the variable age (p = 0.002). Because the assumptions of the parallel-lines model are violated, the partial proportional model was performed.

Tab. 9. Results of the partial proportional odds model according to grass pollen sensitization

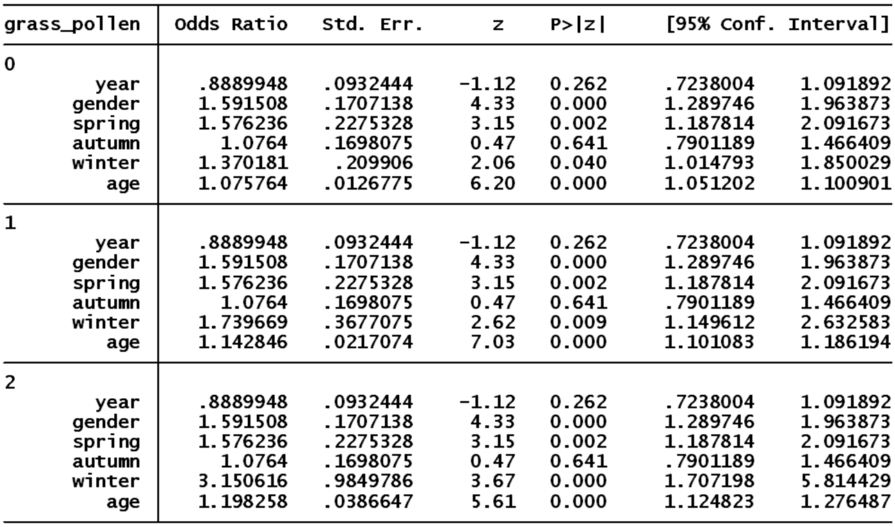


```
Generalized Ordered Logit Estimates                 Number of obs   =       2779
                                                    LR chi2(10)     =      92.68
                                                    Prob > chi2     =     0.0000
Log likelihood = -1536.9007                         Pseudo R2       =     0.0293

 ( 1)  [0]spring - [1]spring = 0
 ( 2)  [0]gender - [1]gender = 0
 ( 3)  [0]autumn - [1]autumn = 0
 ( 4)  [0]year - [1]year = 0
 ( 5)  [1]spring - [2]spring = 0
 ( 6)  [1]gender - [2]gender = 0
 ( 7)  [1]autumn - [2]autumn = 0
 ( 8)  [1]year - [2]year = 0
------------------------------------------------------------------------------
grass_pollen |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0            |
        year |  -.1176639   .1048875    -1.12   0.262    -.3232396    .0879118
      gender |    .464682   .1072655     4.33   0.000     .2544456    .6749185
      spring |   .4550394    .144352     3.15   0.002     .1721147    .7379642
      autumn |   .0736223    .157755     0.47   0.641    -.2355718    .3828163
      winter |   .3149432   .1531958     2.06   0.040      .014685    .6152013
         age |   .0730315   .0117847     6.20   0.000      .049934     .096129
       _cons |  -2.646581   .1738984   -15.22   0.000    -2.987415   -2.305746
-------------+----------------------------------------------------------------
1            |
        year |  -.1176639   .1048875    -1.12   0.262    -.3232396    .0879118
      gender |    .464682   .1072655     4.33   0.000     .2544456    .6749185
      spring |   .4550394    .144352     3.15   0.002     .1721147    .7379642
      autumn |   .0736223    .157755     0.47   0.641    -.2355718    .3828163
      winter |   .5536951   .2113663     2.62   0.009     .1394248    .9679655
         age |    .133522   .0189941     7.03   0.000     .0962942    .1707498
       _cons |  -4.529354   .2454247   -18.46   0.000    -5.010377    -4.04833
-------------+----------------------------------------------------------------
2            |
        year |  -.1176639   .1048875    -1.12   0.262    -.3232396    .0879118
      gender |    .464682   .1072655     4.33   0.000     .2544456    .6749185
      spring |   .4550394    .144352     3.15   0.002     .1721147    .7379642
      autumn |   .0736223    .157755     0.47   0.641    -.2355718    .3828163
      winter |   1.147598   .3126305     3.67   0.000     .5348536    1.760343
         age |   .1808684   .0322675     5.61   0.000     .1176254    .2441115
       _cons |  -6.376669   .4128261   -15.45   0.000    -7.185793   -5.567545
------------------------------------------------------------------------------
```

The partial proportional odds model results are shown in [Tab. 9]. The likelihood ratio chi-squared statistic C = 92.68. The p-value for the Likelihood Ratio Chi-Square test (< 0.0001) showing the overall importance of the explanatory variables. The grass pollen sensitization significantly depends on gender, age, season of birth [Tab. 9]. The year effect is insignificant. The positive coefficients for gender and age mean that the likelihood of the grass pollen sensitization did increase for boys and older children. In Spring born and in Winter born children the likelihood of the grass pollen sensitization increases compared to children born in Summer. In Autumn born children related to Summer born children the seasonal effect is insignificant.

This model is only slightly more difficult to interpret than the earlier proportional odds model. Effects of the constrained variables (year, gender, spring, autumn) can be interpreted much the same as they were previously. For these variables regression coefficients do not depend on response categories. Winter and age variables represent three different regression coefficients per each variable. The coefficients depend on the response categories.

Tab. 10. Odds ratios of the partial proportional odds model according to grass pollen sensitization



```
grass_pollen | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
0            |
        year |   .8889948   .0932444   -1.12   0.262     .7238004    1.091892
      gender |   1.591508   .1707138    4.33   0.000     1.289746    1.963873
      spring |   1.576236   .2275328    3.15   0.002     1.187814    2.091673
      autumn |     1.0764   .1698075    0.47   0.641     .7901189    1.466409
      winter |   1.370181    .209906    2.06   0.040     1.014793    1.850029
         age |   1.075764   .0126775    6.20   0.000     1.051202    1.100901
-------------+----------------------------------------------------------------
1            |
        year |   .8889948   .0932444   -1.12   0.262     .7238004    1.091892
      gender |   1.591508   .1707138    4.33   0.000     1.289746    1.963873
      spring |   1.576236   .2275328    3.15   0.002     1.187814    2.091673
      autumn |     1.0764   .1698075    0.47   0.641     .7901189    1.466409
      winter |   1.739669   .3677075    2.62   0.009     1.149612    2.632583
         age |   1.142846   .0217074    7.03   0.000     1.101083    1.186194
-------------+----------------------------------------------------------------
2            |
        year |   .8889948   .0932444   -1.12   0.262     .7238004    1.091892
      gender |   1.591508   .1707138    4.33   0.000     1.289746    1.963873
      spring |   1.576236   .2275328    3.15   0.002     1.187814    2.091673
      autumn |     1.0764   .1698075    0.47   0.641     .7901189    1.466409
      winter |   3.150616   .9849786    3.67   0.000     1.707198    5.814429
         age |   1.198258   .0386647    5.61   0.000     1.124823    1.276487
------------------------------------------------------------------------------
```

For boys, the odds of severe grass pollen sensitization versus the combined categories: moderate, mild and negative are 1.59 (95% CI: 1.29 – 1.96) times greater than for girls [Tab. 10]. Likewise, for boys, the odds of severe and moderate grass pollen sensitization versus the combined mild, negative categories are 1.59 (95% CI: 1.29 – 1.96) times greater than for girls. For boys, the odds of positive skin reactivity to grass pollen (severe, moderate and mild) versus the negative skin reactivity are 1.59 (95% CI: 1.29 – 1.96) times greater than for girls.

For one unit increase in age, the odds of severe grass pollen sensitization versus the combined categories: moderate, mild and negative are 1.08 (95% CI: 1.05 – 1.1) times greater. Also, for one unit increase in age, the odds of severe and moderate skin reactivity to grass pollen versus the combined categories: mild and negative are 1.14 (95% CI: 1.1 – 1.19) times greater.

For one unit increase in age, the odds of positive skin reactivity to grass pollen versus the negative skin reactivity are 1.2 (95% CI: 1.12 – 1.28) times greater, given the other variables are held constant in the model.

In the Spring born group, the odds of grass pollen sensitization versus the combined categories: moderate, mild and negative are 1.58 (95% CI: 1.19 – 2.09) times greater than in the Summer born group. In the Spring born group, the odds of severe and moderate skin reactivity to grass pollen versus the combined categories mild and negative are 1.58 (95% CI: 1.19 – 2.09) times greater than in the Summer born group. In the Spring born children the odds of positive skin reactivity grass pollen versus the negative skin reactivity are 1.58 (95% CI: 1.19 – 2.09) times greater than in the Summer born children.

In the Winter born children, the odds of positive skin reactivity (severe, moderate and mild) to grass pollen versus the negative skin reactivity are 1.37 (95% CI: 1.01 – 1.85) times greater than in the Summer born group. In Winter born group, the odds of severe and moderate grass pollen sensitization versus the combined categories mild and negative are 1.74 (95% CI: 1.15 – 2.63) times greater than in Summer born. In the Winter born group, the odds of severe grass pollen sensitization versus the combined categories: moderate, mild and negative are 3.15 (95% CI: 1.71 – 5.81) times greater than for the children born in Summer.

Conclusions

Before the most popular among ordinal regression models – the proportional odds model is applied, make sure the proportional odds assumption is satisfied. Otherwise, the results can not be credible. If the assumption is not satisfied the generalized logit models (e.g. partial proportional odds model) should be developed. Application of the ordinal logistic regression models lets us reveal the critical factors that influence dust mite and grass pollen sensitization. In the dust mite case these factors are: year of a test performance, gender and age. In the grass pollen case we find gender, age and year of a test performance significant.

REFERENCES

[1] Bender R., Grouven U., Ordinal logistic regression in medical research, Journal of the Royal College of Physcians of London, 31 (5), pp. 546–551, 1997.

[2] Altman D. G., Statstics in medical journals: development in the 1980s, Stat Med, 10 (12), pp. 1897–1913, 1991.

[3] Harrell F. E. Jr, Lee K.. L., Matchar D. B., Reichert T. A., Regression models for prognostic prediction: advantages, problems, and suggested solutions, Cancer treat Rep, 69 (10), pp. 1071–1077, 1985.

[4] Agresti A., An introduction to Categorical Data Analysis, Wiley, New York.

[5] Abreu M. N. S., Siqueira A. L., Caiaffa W. T., Ordinal logistic regression in epidemiological studies, Rev Saude Publica, 43 (1), 2009.

[6] Amstrong B. G., Sloan M., Ordinal regression models for epidemiologic data, American Journal of Epidemology, 129 (1), 1988.

[7] Bender R., Benner A., Calculating ordinal regression models in SAS and S-Plus, Bometrical Journal, 42 (6), pp. 677–699, 2000.

[8] Ananth C. V., Kleinbaum D. G., Regression models for ordinal responses: a review of methods and applications, Int J Epidemiol, 26 (6), pp. 1323–1333, 1997.

[9] Hosmer D. W., Lemeshow S., Applied logistic regression, 2. Ed., New York, John Wiley & Sons, 2000.

[10] Dobson A. J., An introduction to generalized linear models, 2. Ed., New York, Chapman & Rall/CRC, 2002.

[11] Peterson B., Harrell F. E. Jr, Partial proportional odds models for ordinal response variables, Applied Statistics, 39 (2), pp. 205–217, 1990.

[12] Bender R., Grouven U., Using binary logistic regression models for ordinal data with non-proportional odds, J Clin Epidemiol, 51 (10); pp. 809–816, 1998.

[13] Williams R., Generalized ordered logit/partial proportional odds models for ordinal dependent variables, The Stata Journal, 6 (1), pp. 58–82, 2006.

[14] Marszałkowska J., Gutowska J., Samoliński B., Częstość występowania dodatnich testów skórnych na alergeny pokarmowe w specjalistycznej poradni alergologicznej, Alergia Astma Immunologia, 12 (3), pp. 160–164, 2007.

[15] Arcimowicz M., Samoliński B., Zawisza E., Rapiejko P., Analiza częstości występowania dodatnich testów skórnych na wybrane alergeny pochodzenia roślinnego, Pyłki i pyłkownica: Aktualne Problemy, Instytut Medycyny Wsi w Lublinie, 47, 1995.

Concept of system allowing non invasive detection of uterine contractions in women undergoing In Vitro Fertilization – Embryo Transfer treatment

Robert Milewski[1], Piotr Pierzyński[2], Anna Justyna Milewska[1], Monika Zbucka-Krętowska[2], Sławomir Wołczyński[2]

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

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

Abstract. Uterine contractile activity in women undergoing advanced fertility treatments such as IVF-ET (In Vitro Fertilization – Embryo Transfer) might constitute one of the factors influencing embryo implantation rates, and by that – the success rates of the treatment. In early 1990's it was confirmed that IVF-ET patients experiencing active uterine contractions had up to 3-fold lower success rates as compared to the ones with silent uteri. However, even though it occurs in about one third of patients, exaggerated uterine contractions are not a subject of routine diagnostic tests or any treatment. One of the reasons for that is the lack of appropriate, reliable tools allowing identification of such patients. This paper presents a system enabling easy, non invasive identification of uterine contractions in non pregnant women. Application of the method described in this publication could be helpful in identifying IVF-ET patients with active uterine contractions, who could benefit from additional treatment which could potentially increase their chances for conceiving.

Introduction

Detection, recognition and analysis of signals is inadvertent part of biomedical research projects using multimedia data types such as images, film or sound sequences. In general, the process of signal analysis starts from acquiring appropriate data at the time of patients' examination. Subsequently, following the analog to digital conversion, data is placed in a specially designed framework, for instance a multimedia database.

Digitally recorded signals can be reviewed whenever needed, however, further analysis requires extraction of information which could be presented in quantitative measures, such as length, width, volume or changes of object's shape in time. Analysis of film sequences is relatively complicated as it comprises of both image analysis and the analysis of temporal changes

ISBN 978–83–7431–336–0 ISSN 0860-150X

of such. One of the basic methods applied in analysis of film sequences are analysis of separate film frames with subsequent string analysis of results. Another method focuses on temporal analysis of specific structure of an image. When analyzing film sequences with unstable position of a region of interest, a specific correction against the stable benchmarks needs to be used.

Our paper is presenting a method of analysis of ultrasound image of non pregnant human uteri allowing identification of the contractile activity. The analysis procedure comprises of several stages – namely – setting an observed section of an image, generation of a graph of temporal changes of image parameters and detection of a region of uterus which is the most indicative for uterine contractions (it is the so called endometrial interface). The current method's novelty is an ability of an automatic identification of the endometrial interface. None of the previously published reports enabled the above which was potentially a source of bias [2, 4].

Significance of uterine contractile activity in fertility treatments

Uterine contractile activity is an important component of its receptivity, affecting the process of implantation of embryos [15]. Uterine peristalsis was found to be much higher in the IVF-ET cycles than in the natural ones [16]. Uterine contractions are negatively correlated to the implantation rates in women undergoing embryo transfer, a final stage of IVF–ET treatment [2]. Notwithstanding that increased uterine contractions are found in one third of IVF-ET patients, elevated uterine contractile activity is currently not a subject of any routine diagnosis or treatment [1].

Uterine contractions can be objectively measured by placing the intrauterine pressure transducer and recording the pressure changes [3]. Such an approach is however not acceptable in patients who are about to have Embryo Transfer procedure as it is related to endometrial trauma which severely decreases the chances for successful embryo implantation.

Consequently, non invasive methods of detection and analysis of uterine contractions should be applied. Ultrasound scan is the easiest non invasive method of collection of images of uteri. Within the uterus, one can describe two most distinctive layers – myometrium (outer, hypoechogenic) and endometrium (inner, hyperechogenic). The border between the two is called junctional zone [5]. Changes in junctional zone are reflective of uterine contractions [4–5].

Analysis of ultrasound scans of non pregnant uteri has been a subject of a number of reports which focus rather on the observer counting uterine contractions or analysis of very local (single point) changes of endometrial interface [1–2, 4]. More global approach allowing multi-point analysis of uterine peristalsis could allow not only detection but also determination of direction and power of contractions. Apart from being an interesting clinical research tool, such a method could be more adequate in the identification of women in risk of unsuccessful embryo implantation. It could increase their chances for successful treatment which is especially important in groups with poorer prognosis such as women of above 40 years old age group [10].

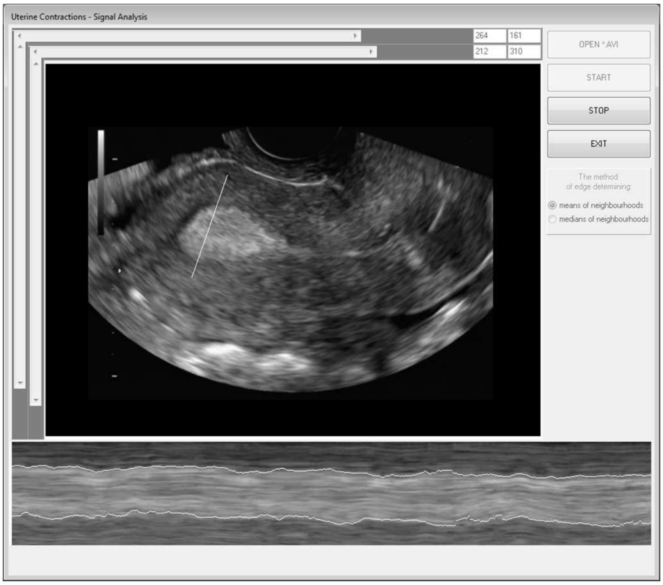
Application for automatic detection of uterine contractions

The ultrasound scans of non pregnant uterus have been performed on consented patients of the Department of Reproduction and Gynaecological Endocrinology, Medical University of Bialystok, being prepared for the Embryo Transfer procedure (a final step of IVF-ET). For the scans, the GE Voluson Expert 730 scan system equipped with Sony VRD–MC6 recorder was used. Film sequences have been stored on DVD disks and converted to AVI format for further usage.

The very first stage of the project was the preparation of a purpose built application for the analysis of ultrasound scan images named Scan Studio. It has been created in Delphi based programming environment with the use of Embarcadero RAD Studio XE2 package.

Upon opening, the Scan Studio application allows uploading the AVI file with ultrasound scan recording and displays its very first frame. Subsequently, with a use of side scrollbars it is possible to position user defined gate identifying the transsection along which the image analysis is going to be performed. The coordinates of the gate's endpoints are displayed and can be used for exact replication of an analysis if needed. The properly adjusted gate is perpendicular to the long axis of the uterine body and is symmetrically covering the transsection of an endometrium and adjacent margin of myometrium [Fig. 1]. This stage is dependant on the user, however, it is relatively simple and it is unexpected that it might be a source of a significant bias.

When the AVI file is played, the Scan Studio automatically detects endometrial interface by using two alternative, user defined methods – means of neighbourhoods or medians of neighbourhoods. The result of the detection is displayed on a time axis in a lower part of the screen, below the image of the uterus [Fig. 1]. For improved visualization, the length of time axis is fixed and equal to the referred screen section. It allows constant inspection of the whole signal (no additional scrolling is required to review the resulting image).



The method
of edge determining:
● means of neighbourhoods
○ medians of neighbourhoods

Fig. 1. Upper section of the figure – optimal setting of the gate on the uterine sagittal cross section. Lower section – image resulting from automatic detection of endometrial interface (marked by bright lines)

The changes of pixels alongside the used defined gate (region of interest) are displayed below the image of an uterus. The analysis and automatic detection of the endometrial interface – lower section of [Fig. 1] – stops when the last frame of an AVI file is reached or it can be halted by user anytime, and the resulting JPG file is recorded.

Algorithms of detection of endometrial interface

Detection of endometrial interface is a key stage of operation of Scan Studio. As endometrium is distinctively more echogenic (brighter) as compared to surrounding myometrium, the application focuses on assignment of borders between darker and brighter regions within user defined cross section (the gate). The border between the endometrium and myometrium is not always explicit as there might be more echogenic (brighter) areas in the myometrium or less echogenic (darker) areas within the endometrium. When the scan image of uterus is observed from a distance, endometrial interface is easily identified, however, when zooming in, one can find that the border is not unequivocal and that the image is noisy in that region.

One of the methods for detection of such a border is averaging of the signal according to the surrounding of each reference points, which is reducing minor deviations and results in finding the value which is dominant in the examined region. In Scan Studio two such methods are implemented which are based on the calculation of means and medians of neighbourhoods. The former (means of neighbourhood method) calculates the mean values in the neighbourhood of a given point alongside the chosen transsection. Subsequently, the application calculates maximal and minimal values for the whole transsection in each separate timepoint. It is assumed that the brightest point of endometrium is located within the uterine cavity and the position of the darkest points are set down separately for upper and lower endometrial interface. Basing on the distance between the marginal brightness values, the cut off point is determined, which represents the border between the endometrium and myometrium – an endometrial interface.

The alternative algorithm is constructed similarly to the described above, and it uses median values instead of means. Both algorithms produce similar results, though the means of neighbourhood method seems to be more exact [Fig. 2]. The weakness of both is that in noisy images the detection of endometrial interface might be biased and might result in sudden shift in the detected endometrial interface occurring in the timepoint of images of poorer quality.

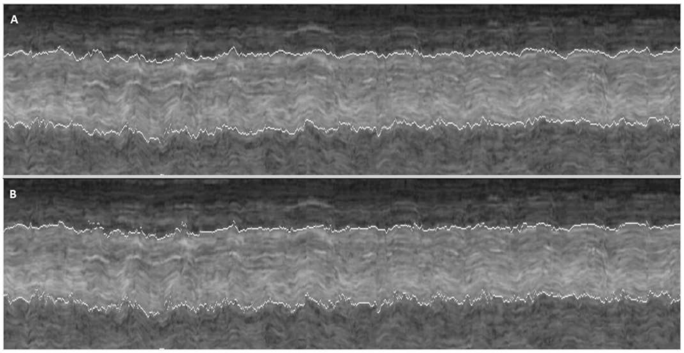


Fig. 2. Uterine contraction graphs produced by application of A – algorithm basing on means of neighbourhoods method, B – algorithm basing on medians of neighbourhoods method

Possibly more effective algorithm for the detection of endometrial interface could be employing the analysis of gradient of signal, when endometrial interface would be set in place of maximal increase or decrease of signal. It could be also associated with setting the medians or means in the surroundings which could – to some extent – eliminate the minor noise.

Another, also interesting approach, which might give even more adequate results could be applying the approximation of a value of the signal alongside the transsection of a sum of logistic functions. In such a case, the cutoff point could be determined in the point of extreme values of derivative functions of both logistic curves (maximum for increasing curve and minimum for decreasing one). It seems that in this approach the system could be more noise resistant.

Conclusions and future plans

Scan Studio provides an easy tool for delineating uterine contractile activity. Unique entity of the application is automatic detection of endometrial interface which is very helpful in detecting the uterine contractions. In its current form, the application can be used for detection of the frequency of uterine contractions and can identify the IVF-ET patients with elevated uterine contractility. Such patients contractions could be effectively treated with medications from the oxytocin antagonists group such as atosiban, which was shown to decrease contractions and promote embryo implantation [12, 14–15].

Further stages of development of the application involve implementation of additional algorithms for the detection of endometrial interface. Subsequent to that, we plan to incorporate the automation of detection of uterine contractions allowing the determination of their frequency and strength. Advanced data analysis could be used to determine the direction of contractions [13].

The completed and operational application can be incorporated within existing system of electronic registration of information about patients treated for infertility using IVF ICSI/ET method [7] with the statistical module [11] and the predictive module, based on the technology of artificial neural networks [6]. The whole system can be also merged with modules based on other advanced data-mining methods for analysis of IVF patients data [8–9].

REFERENCES

[1] Fanchin R., Ayoubi J. M., Righini C., et al., Uterine contractility decreases at the time of blastocyst transfers, Human Reproduction, 16 (6), pp. 1115–1119, 2001.

[2] Fanchin R., Righini C., Olivennes F., et al., Uterine contractions at the time of embryo transfer alter pregnancy rates after in-vitro fertilization, Human Reproduction, 13 (7), pp. 1968–1974, 1998.

[3] Kitlas A., Oczeretko E., Swiatecka J., et al. Uterine contraction signals-application of the linear synchronization measures, European Journal of Obstetrics & Gynecology and Reproductive Biology, 144, Supplement 1, pp. S61–S64, 2009.

[4] Lesny P., Killick S. R., The junctional zone of the uterus and its contractions, BJOG, 111 (11), pp. 1182–1189, 2004.

[5] Lesny P., Killick S. R., Tetlow R. L., et al., Embryo transfer and uterine junctional zone contractions, Human Reproduction Update, 5 (1), pp. 87–88, 1999.

[6] Milewski R., Jamiolkowski J., Milewska A. J., et al., 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), pp. 900–906, 2009.

[7] Milewski R., Jamiolkowski J., Milewska A. J. et al., The system of electronic registration of information about patients treated for infertility with the IVF ICSI/ET method, Studies in Logic, Grammar and Rhetoric, 17 (30), pp. 225–239, 2009.

[8] Milewski R., Malinowski P., Milewska A. J., et al., Nearest neighbor concept in the study of IVF ICSI/ET treatment effectiveness, Studies in Logic, Grammar and Rhetoric, 25 (38), pp. 49–57, 2011.

[9] Milewski R., Malinowski P., Milewska A. J., et al., The usage of margin-based feature selection algorithm in IVF ICSI/ET data analysis, Studies in Logic, Grammar and Rhetoric, 21 (34), pp. 35–46, 2010.

[10] Milewski R., Milewska A. J., Domitrz J., et al., In vitro fertilization ICSI/ET in women over 40, Przegląd Menopauzalny, 2 (36), pp. 85–90, 2008.

[11] Milewski R., Milewska A. J., Jamiołkowski J., et al., The statistical module for the system of electronic registration of information about patients treated for infertility using the IVF ICSI/ET method, Studies in Logic, Grammar and Rhetoric, 21 (34), pp. 119–127, 2010.

[12] Moraloglu O., Tonguc E., Var T., et al., Treatment with oxytocin antagonists before embryo transfer may increase implantation rates after IVF, Reproductive Biomedicine Online, 21 (3), pp. 338–343, 2010.

[13] Oczeretko E., Swiatecka J., Kitlas A., et al., Visualization of synchronization of the uterine contraction signals: running cross-correlation and wavelet running cross-correlation methods, Medical Engineering & Physics, 28 (1), pp. 75–81, 2006.

[14] Pierzynski P., Gajda B., Smorag Z., et al., Effect of atosiban on rabbit embryo development and human sperm motility, Fertility and Sterility, 87 (5), pp. 1147–1152, 2007.

[15] Pierzynski P., Oxytocin and vasopressin V1A receptors as new therapeutic targets in assisted reproduction, Reproductive Biomedicine Online, 22 (1), pp. 9–16, 2011.

[16] Zhu L., Li Y., Xu A., Influence of controlled ovarian hyperstimulation on uterine peristalsis in infertile women, Human Reproduction, 27 (9), pp. 2684–2689, 2012.

Detrended Fluctuation Analysis (DFA) in biomedical signal processing: selected examples

Agnieszka Kitlas Golińska[1]

[1] Department of Medical Informatics, Institute of Computer Science, University of Bialystok, Poland

Abstract. Detrended Fluctuation Analysis (DFA) quantifies fractal-like autocorrelation properties of the signals. It is useful for analyzing biomedical signals which are mostly complex and non-stationary. In this paper we review selected examples of application of the DFA method in cardiology, neurology and other studies. We also present our findings – some of our original work. We conclude that using the DFA method we can determine which signal is more regular and less complex (in practice to distinguish healthy from unhealthy subjects).

Introduction

Biomedical systems exhibit complexity and nonlinear structure and this complexity is present in measured signals, such as ECG or EEG [4]. It is generally accepted that the remarkable complexity of biological signals is a result of two factors [6]: high complexity of systems (many degrees of freedom) and their susceptibility to environmental factors. Chaotic systems exhibit characteristics of stochastic systems, but can be described using only a few variables (in some cases only one). Also biological signals are difficult to analyze because they are mostly non-stationary [4, 9].

Classical methods of signal analysis work well mostly on stationary signals, so we need a new solution – new methods. Nonlinear dynamics (more precisely in this case – chaos theory) provides many new ways of analyzing signals, such as fractal methods. Some of these methods determine the scaling exponent of the signal which indicates the presence or absence of fractal properties (self-similarity) [9]. DFA is a scaling analysis method that provides a simple quantitative parameter to represent the autocorrelation properties of a signal [4]. It is also known for its robustness against non-stationarity [9].

ISBN 978–83–7431–336–0 ISSN 0860-150X

Detrended Fluctuation Analysis (DFA)

Detrended Fluctuation Analysis is an interesting method for scaling the long-term autocorrelation of non-stationary signals. It quantifies the complexity of signals using the fractal property [12, 14]. DFA was first proposed by Peng et al. in 1995 [12]. This method is a modified root mean square method for the random walk. Mean square distance of the signal from the local trend line is analyzed as a function of scale parameter. There is usually power-law dependence and interesting parameter is the exponent. In many cases the DFA scaling exponent can be used to discriminate healthy and pathological data [15].

DFA algorithm

We will illustrate the DFA algorithm on 1-dimensional signal $B(i)$, $i = 1, \ldots, N$ [11–12]. First, we compute the integrated signal according to the formula

$$y(k) = \sum_{i=1}^{k} (B(i) - B_{\text{avg}}) \tag{1}$$

where B_{avg} is the mean value of the signal. Next we divide the data into segments of length n and find the linear approximation y_n using least squares fit in each segment separately (representing the trend in a given section).

The average fluctuation $F(n)$ of the signal around the trend is given by this formula:

$$F(n) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (y(k) - y_n(k))^2} \tag{2}$$

The calculations are repeated for all considered n. We are interested in the relation between $F(n)$ and size of segment n. In general $F(n)$ will increase with the size of segment n.

Next, we create a plot – double logarithmic graph ($\log F(n)$ vs $\log n$). The linear dependence indicates the presence of self fluctuations and the slope of the line $F(n)$ determines the scaling exponent α [1, 9, 12, 15–16]:

$$F(n) \sim n^{\alpha} \tag{3}$$

For example, recent studies have shown that DFA functions of different R-R series (from ECG) are approximated by power-law [1, 12, 15] as well as synchronization signals EEG [9, 16].

Scaling exponent α

The parameter α (scaling exponent, autocorrelation exponent, self-similarity parameter) represents the autocorrelation properties of the signal [2, 4, 9, 13, 15–16]:

1. $\alpha < 0.5$ anti-correlated signal
2. $\alpha = 0.5$ uncorrelated signal (white noise)
3. $\alpha > 0.5$ positive autocorrelation in the signal
4. $\alpha = 1$ 1/f noise
5. $\alpha = 1.5$ Brownian noise or random walk

Gifani et al. [4] claim, that using scaling exponent α one should be able to completely describe the significant autocorrelation properties of the biomedical signals. Often computed separately exponent for low and high n can describe short-range scaling exponent (or fast parameter) α_1 and long-range scaling exponent (or slow parameter) α_2 for time scales [3].

Example of the DFA method

In [Fig. 1] and [Fig. 2] we present an example of application of the DFA method. We selected R-R intervals signal (from ECG). Original signal was integrated and detrended – presented in [Fig. 1]. Next, double logarithmic plot was created and scaling exponents were calculated.

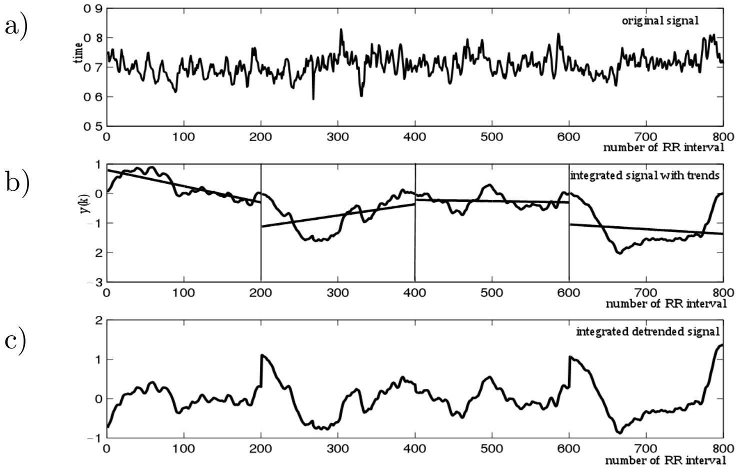


Fig. 1. DFA method: a) selected original signal (R-R intervals from ECG), b) integrated signal with local trends estimated in each section, c) detrended integrated signal

In [Fig. 2] double logarithmic graph $\log F(n)$ vs $\log n$ is shown. The slope of the line determines the scaling exponent (short-range scaling exponent α_1 and long-range scaling exponent α_2).

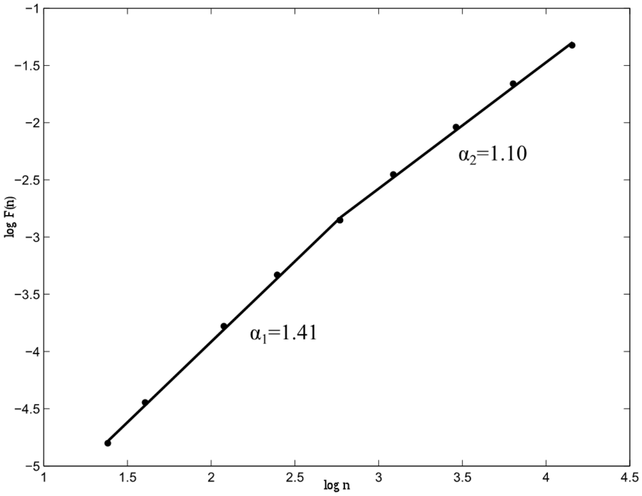


Fig. 2. Scaling exponent (short-range scaling exponent α_1 and long-range scaling exponent α_2) for an example RR intervals signal

Application in biomedical processing

Selected examples in cardiology, neurology and other studies

In biomedical signals analysis DFA is mostly used in ECG studies [1–3, 14–15] and EEG studies [4, 9–10, 16].

Acharya et al. [2] classified certain disease using DFA in ECG studies. DFA was also used in the analysis of atrial signal during adrengenic activation in atrial fibrillation [3]. Pikkujamsa et al. [14] studied cardiac inter-beat interval dynamics from childhood to senescence. They claim that the loss of complexity and alterations of fractal organization related with aging (also apparent in many diseases) may be associated with the reduced ability to adapt to physiological stress. The DFA method can also help to diagnose heart failure [1]. In this paper DFA was applied to R-R intervals studies and differences are observed between scaling exponent α of healthy and unhealthy subjects. Rodriguez et al. [15] showed that significant differences in scaling of intra-beat dynamics can be observed with time series of about 5–30 min. This could make intra-beat scaling analysis potentially applicable to real clinical data. Also intra-beat dynamics displays differences in the scaling behavior of healthy and unhealthy subjects.

Gifani et al. [4] claim, that using DFA, they can describe the dynamics of brain during anesthesia. They found the optimum fractal-scaling exponent by selecting the best domain of box sizes, which have meaningful changes with different depth of anesthesia. Lee et al. [9–10] analyzed the EEG in sleep apnea and long-range autocorrelations by calculating its scaling exponents. The scaling exponents of the apnea were lower than those of the healthy subject. Stam et al. [16] examined the hypothesis that cognitive dysfunction in Alzheimer's disease is associated with abnormal spontaneous fluctuation of EEG synchronization levels during an eye-closed resting state.

Phinyomark et al. [13] claim that DFA's scaling exponent is an efficient parameter in practical surface EMG controlled prostheses. The studies show that scaling exponent in various hand motions have the significant difference value and small experimental variation. The authors think that DFA could be considered as an element of multifunction myoelectric control system.

Selected examples from our studies

In our work [7–8] we applied the DFA method to heart rate variability studies (RR intervals). We have analyzed two groups of children: children with diabetes type 1 with microalbuminuria and healthy children. For each child 24 hours ECG (R-R intervals) was recorded. Then we divided these records into two segments: day (6.00–22.00) and night (22.00–6.00) respectively.

The DFA method showed statistically important differences between studied groups of children and also differences between night and day [7–8]. We obtained scaling exponent for healthy and unhealthy children near 1.0, which is consistent with studies performed by Yeh et al. [17] and Pikkujamsa et al. [14]. Also values of scaling exponents were higher for unhealthy subjects than for healthy, which suggest more regular, less complex signals for unhealthy children. This could indicate too regular heartbeat. In these cases heart could not rest, works like in an athlete, which is very dangerous.

We concluded that using nonlinear dynamics methods (DFA) we could quantitatively and qualitatively study the heart rate variability and distinguish healthy from unhealthy subjects.

Here we present our recent findings – analysis of EMG signals. These signals were obtained from Physionet [5]. We concentrated on short EMG recordings from three subjects: healthy, one with myopathy and one with neuropathy (presented in [Fig. 3]). EMG records were obtained using 25 mm concentric needle electrode placed in tibialis anterior muscle. Subjects dorsiflexed the foot gently against resistance and then relaxed.

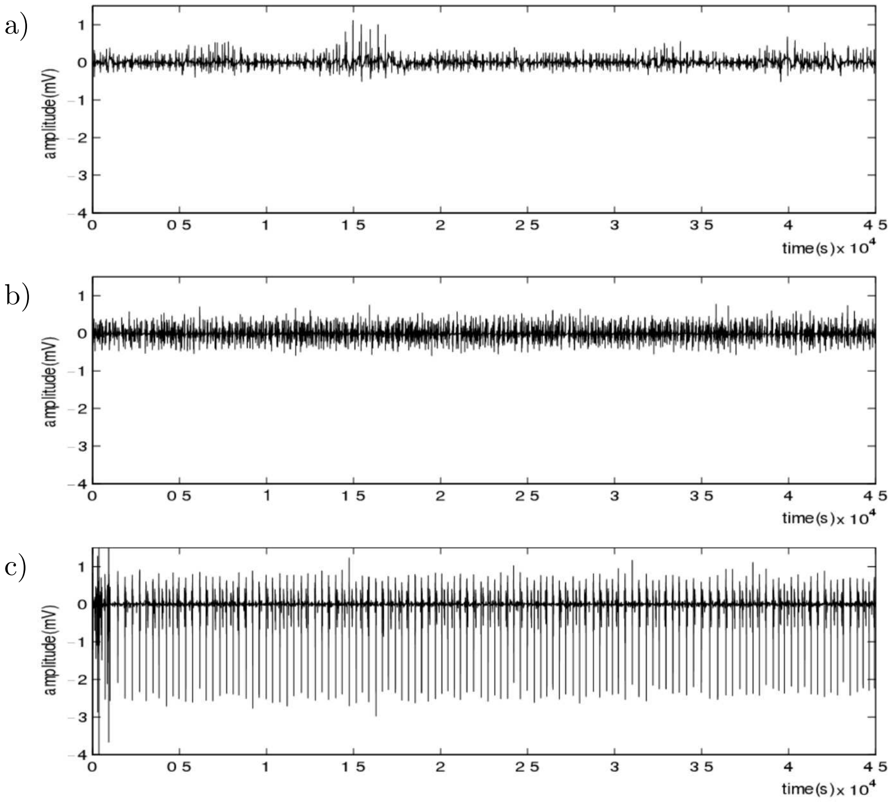


Fig. 3. Selected EMG signals from: a) healthy subject, b) subject with myopathy, c) subject with neuropathy

In [Fig. 4] we present values of scaling exponents and the slope of the line $F(n)$ on double logarithmic plot obtained by using the DFA method for studied signals. All results are between 0.52 and 1.41, which suggest self-similarity properties of these signals. We can also observe differences between values of short-range scaling exponent α_1 and long-range scaling exponent α_2. For unhealthy subjects values are lower than for the healthy one, especially values of α_2. So signals from unhealthy subjects are less regular and more complex than from the healthy one. Differences in values of α_2 are consistent with work by Pikkujamsa et al. [14] – alterations of long-range (fractal, self-similarity) organization related with disease.

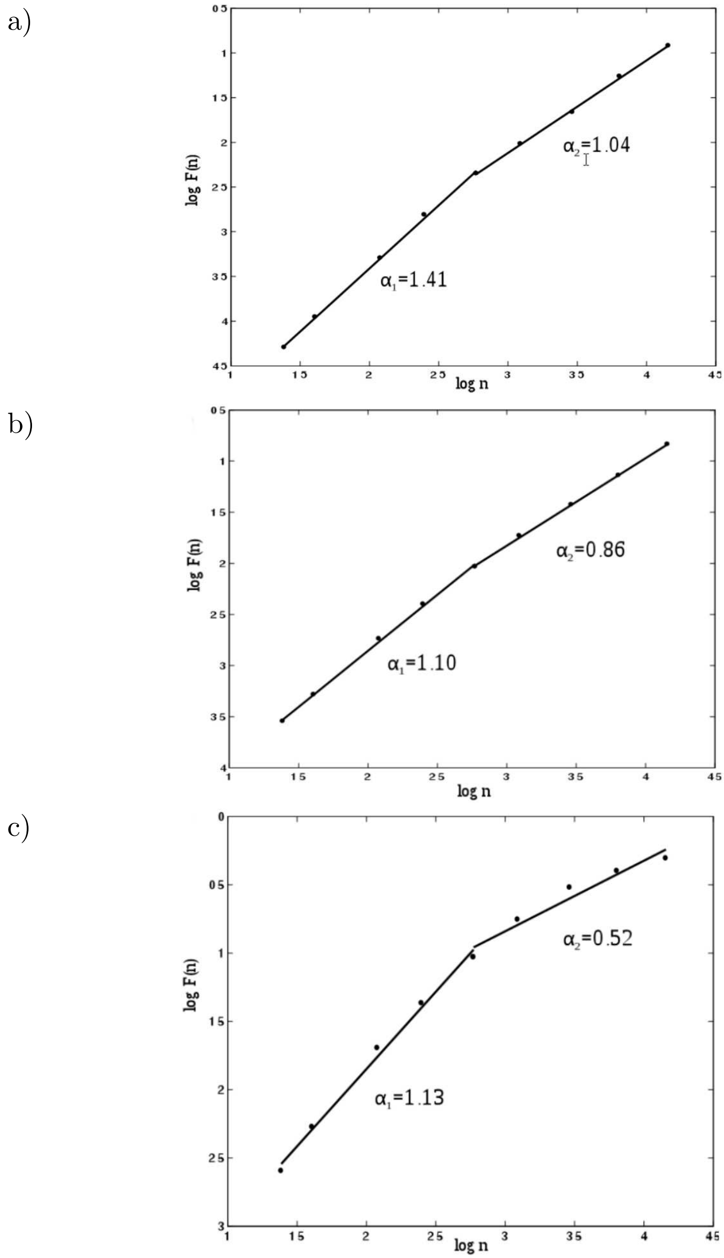


Fig. 4. Scaling exponents for: a) healthy subject, b) subject with myopathy, c) subject with neuropathy

Conclusions

Using the DFA method we can distinguish healthy from unhealthy subjects. Also we can determine which signal is more regular and less complex – useful for analyzing biomedical signals. We concluded that using nonlinear dynamics methods, like the DFA method we could quantitatively and qualitatively study physiological signals.

REFERENCES

[1] Absil P. A., Sepulchre R., Bilge A., Gerard P., Nonlinear analysis of cardiac rhythm fluctuations using DFA method, Physica A, 272, pp. 235–244, 1999.

[2] Acharya R. U., Lim C. M., Joseph P., Heart rate variability analysis using correlation dimension and detrended fluctuation analysis, ITBM-RBM, pp. 333–339, 2002.

[3] Corino V. D. A., Ziglio F., Lombardini F., et. al., Detrended Fluctuation Analysis of atrial signal during adrengenic activation in atrial fibrillation, Computers in Cardiology, 33, pp. 141–144, 2006.

[4] Gifani P., Rabiee H. R., Hashemi M. H., et al., Optimal fractal-scaling analysis of human EEG dynamic for depth of anesthesia quantification, Journal of the Franklin Institute, 344, pp. 212–229, 2007.

[5] Goldberger A. L., Amaral L. A. N., Glass L., et al., PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation, 101 (23), pp. e215–e220, 2000. [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]

[6] Jaśkowski P., Zastosowanie metod dynamiki nieliniowej do analizy sygnału EEG człowieka, Current Topics in Biophysics, 19, pp. 4257, 1995.

[7] Kitlas A., Analiza zmienności rytmu serca (interwały R-R) w 24-godzinnym zapisie EKG u dzieci za pomocą wybranych metod dynamiki nieliniowej (praca magisterska), Uniwersytet w Białymstoku, 2003.

[8] Kitlas A., Oczeretko E., Kowalewski M., et al., Nonlinear dynamics methods in the analysis of the heart rate variability, Advances in Medical Sciences (Annales Academiae Medicae Bialostocensis), 50 (Suppl 2), pp. 46–47, 2005.

[9] Lee J. M., Kim D. J., Kim I. Y., et al., Detrended fluctuation analysis of EEG in sleep apnea using MIT/BIH polysomnography data, Computers in Biology and Medicine, 32, pp. 37–47, 2002.

[10] Lee J. M., Kim D. J., Kim I. Y., et al., Nonlinear analysis of human sleep EEG using detrended fluctuation analysis, Medical Engineering & Physics, 26, pp. 773–776, 2004.

[11] Peng C. K., Buldyrev S. V., Havlin S., et al., Mosaic organization of DNA nucleotides, Physical Review A, 49, pp. 1685–1689, 1994.

[12] Peng C. K., Havlin S., Stanley H. E., Goldberger A. L., Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series, Chaos, 5, pp. 82–87, 1995.

[13] Phinyomark A., Phothisonothai M., Limsakul C., Phukpattaranont P., Detrended fluctuation analysis of electromyography signal to identify hand movement, The 2nd Biomedical Engineering International Conference (BMEiCON), pp. 324–329, 2009.

[14] Pikkujamsa S. M., Makikallio T. M., Sourannder L. B., et al., Cardiac interbeat interval dynamics from childhood to senescence. Comparison of conventional and new measures based on fractals and chaos theory, Circulation, 100 (4), pp. 393–399, 1999.

[15] Rodriguez E., Echeverria J. C., Alvarez-Ramirez J., Detrended fluctuation analysis of heart intrabeat dynamics, Physica A, 384, pp. 429–438, 2007.

[16] Stam C. J., Montez T., Jones B. F., et al., Disturbed fluctuation of resting state EEG synchronization in Alzheimer disease, Clinical Neurophysiology, 116, pp. 708–715, 2005.

[17] YehR. G., Shieh J. S., Chen G. Y., Kuo C. D., Detrended fluctuation analysis of short-term heart rate variability in late pregnant women, Autonomic Neuroscience: Basic and Clinical, 150, pp. 122–126, 2009.

The Ruby Language in biomedicine: a short review and selected examples

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

[1] Department of Medical Informatics, Institute of Computer Science, University of Bialystok, Poland

Abstract. Ruby is a dynamic programming language – both in relation to the typing discipline and in the increasing number of fields of usage. It may be very useful in biomedical data studies. In this paper we present some facts about the Ruby language together with the example of how to use Ruby in EMG signal analysis. We also review selected applications of the Ruby language in biomedical studies and present our original work. Our goal was to check if in Ruby one can write programs for basic signal analysis (power spectrum based on Fourier transform and energy of signal), which could give accurate results. We also compare computation times using Ruby language and MatLab. We conclude that the Ruby language gives the same results as MatLab and although computation times are worse than in MatLab, we propos using the Ruby language. It is a simple, efficient tool and in contrast to the expensive MatLab – free of charge.

Introduction

Software plays an important role in biomedical studies. There are many programs, which are used, but they are mostly very expensive. One of the widely used programs is MatLab, language of technical computing, developed by MathWorks. We propose the use of the Ruby language – simple, efficient and free of charge tool [10, 13].

Some facts about Ruby language

The Ruby language was designed in Japan in 1995 [13]. Its creator, Yukihiro Matsumoto, attempted a merge of his favourite programming languages, i.e. Perl, Smalltalk, Eiffel, Ada, and Lisp [4]. His objective was a language with a focus on the ease of use by humans, not by computers, which means that it is a language with a very high level of abstraction. Dynamic typing model, in this case "duck typing", makes writing even simpler. The language contains a lot of syntactic sugar [5].

ISBN 978–83–7431–336–0 ISSN 0860-150X

Ruby's main paradigm is object-oriented programming, which is currently the most popular way to write applications. It may also serve as a bridge to functional programming, supposedly the next major programming paradigm. It is one of the few languages that implements full object-oriented paradigm, which means that even numbers and classes are objects [3]. Some people, mostly from the Java community, say it's too much of a generalization, but it is just another property, that makes the language easy to use and general-purpose.

One of the core strengths of Ruby is its reflectivity [13]. All classes are open and can be modified at any moment by a programmer. Ruby supports single inheritance, but programmers can include modules to expand classes' behavior. This style of programming is called a mixin. Metaprogramming is a popular and efficient way to write applications, and it's supported by Ruby [11]. It increases productivity and encourages the programmers to re-use their code. The language is widely used as a scripting language, which helps in OS administration, XML processing or loading CSV files. Ruby can also be embedded in HTML code. It allows creating dynamic, server-side-evaluated web pages. Ruby on Rails is the so-called killer app for Ruby [4], which allows creating functional database-based web applications in a matter of minutes. Ruby may not be the fastest language to compute, but it is time-saving for the programmers.

A short review of application of Ruby language in biomedicine

The Ruby language is very useful in biomedical informatics. The position [1] is an introduction to Ruby for biomedical researchers. It contains a discussion of many applications covering the most common computational tasks in the field of biomedicine.

While browsing many scientific article databases, we could only find that Ruby is generally used as a helper tool, not the main one.

In the paper [12] a Ruby script is used to load 5880 anonymized magnetic resonance in studies (almost 2 million images).

Lim in his article [9] introduces a novel method of introducing bioinformatics to students in a Programming Languages class. Ruby allowed the teacher to present basic concepts of string manipulation in a way that is familiar to programmers.

In the paper [7] the authors present BioRuby, an open-source bioinformatics library as a functional way to simplify work for bioinformatics researchers. The article is an introduction to BioRuby by demonstrating a few key features of the library, eg. handling FASTA file format, fetching a KEGG graph with the KEGG API, using the interactive environment and availability for all platforms and Java Virtual Machine.

So far, the Ruby language is not often used in the field of medicine. There are a few applications in signal analysis, so we can expect more in the future.

Basics of the Ruby language – how to use Ruby in signal analysis

Ruby is a fully object-oriented language. It means that everything is an object: numbers, strings, nil (empty value), or even classes themselves (contrary to eg. C++ or Java). Thanks to this, it is possible to call methods for any variable. To find out, what is the class of any given object, the method "`class`" can be called. To get a list of methods available for an object, the method "`methods`" may be used.

Examples:

```
number = 5
number.class       #returns the class of a variable, Fixnum here
number.methods     #returns a list of method available for
                                                  the object
nil.class          #returns NilClass
"text".class.clas #returns the class of the String class - Class
```

To print something on screen, we can use one of three instructions: "`print`", "`puts`", or "`p`":

- `print` – simply print a string on screen,
- `puts` – print and add the end of line symbol,
- `p` – print a more detailed information about an object (ie. it calls the "`inspect`" method) and adds the end of line symbol.

Examples:

```
print "Hello!\n"
puts "Hello!"
p "Hello!"
```

The method "`gets`" is used for input. The plus sign is used for concatenation of strings.

Example:

```
print "Input text: "
print "Inputed: " + gets
```

The variables' names in Ruby can consist of letters, numbers, and underline, but must begin with a letter. To assign a value to a variable, the "`=`" sign is used. There is no need to declare the type of a variable in Ruby.

Example:

```
print "What's your name?"
name = gets
puts "Hello, " + name
```

The variables can be used in operations.

Example:

```
x = 7
y = 3
sum = x + y
puts "The sum: " + sum.to_s
```

To concatenate a string with a number, we must call the "`to_s`" conversion method.

In Ruby, a table can contain objects of different classes.

Example:

```
table = [3, "a", 2.5, ["x", "y", "z"]]
```

To add another object to a table, the "`<<`" operator can be used.

Example:

```
table << "object"
```

This operator will append an object at the end of the table.

Code blocks

A code block is an unnamed function. It can be passed as a parameter to a method. A code block is a fragment between curly braces or the keywords "`do`" and "`end`". The local variables of a block are declared between two vertical lines. One of the most frequently used method that accepts a block is "`each`", which is used for iteration of a collection.

Example:

```
tab = [1, 2, 3, 4, 5, 6, 7, 8, 9]
sum = 0
tab.each {|i| sum += i}
puts "The result: #{sum}"
```

Another usage of code blocks is repetition.

This example prints the string three times:

```
3.times {puts "Hello, Ruby!"}
```

Blocks allow code to be loop-free, which makes debugging much easier.

Code block is an easy way of handling files, which in most languages is quite tiresome. Using traditional technique would look something like this:

```
file = File.open("file_name", "r+")
p file
file.close
```

With code blocks it would be much easier and shorter:

```
File.open("file_name") {|file| p file}
```

There is an additional advantage to this method. We don't have to remember to close the file, because Ruby will do it for us. To get a line of the contents of a file, the method "`gets`" may be called. This method, however, returns a string, so we have to convert it to a float using "`to_f`". Then the "`while`" loop would be useful to get the entire contents, line by line.

Application of code blocks in our work

In our programs we used code blocks to load the signals:

```
data=[]
File.open("signal.txt") do |f|
  while line=f.gets
    data<<line.to_f
  end
end
```

We have the signal from file signal.txt in a table data. Now it's only a matter of passing this table to the method that performs the discrete Fourier transform. The result is written to another file, in the following way:

```
File.open("signalout.txt", "w+") {|out| out<<dft(data)}
```

Methods in EMG signal analysis using the Ruby language and MatLab

We used two basic, widely used methods in signal analysis – power spectrum based on the Fourier transform (frequency domain analysis) and energy of signal (time domain analysis). We wrote programs for these methods in Ruby (version 1.9) and MatLab (version 2006).

Power spectrum is based on the Fourier transform. In MatLab we used built-in "`fft`" function (fast Fourier transform algorithm) for the Fourier transform, which is implemented on environment level. In the Ruby language there aren't any built-in functions for signal processing and analysis (only basic mathematical functions), so we wrote a program for Fourier transform ourselves using definition of Fourier transform – our written function "`dft`" and also – to compare computing time – "`fft`" function (fast Fourier transform algorithm). Both in Matlab and in Ruby we calculated energy using simple iterators and arithmetic operators.

Power spectrum

Power spectrum is defined as [2, 14]:

$$P_{xx}(\omega) := |\hat{x}(\omega)|^2 = \hat{x}(\omega)\overline{\hat{x}(\omega)} \tag{1}$$

where x is complex conjugate of x and

$$\hat{x}(\omega) := \int_{-\infty}^{\infty} x(t)e^{-i\omega t}dt \tag{2}$$

is the Fourier transform. It is very important transform in signal processing and analysis, because it yields the information about frequencies occurring in signals and we can compute the dominant frequency for signals using this method.

In our experiment we used a MatLab function "`fft`", which calculates the Fourier transform using the fast Fourier transform algorithm:

```
s=load('e2.txt');
x=fft(s);
```

and a Ruby function "`dft`" (written by us) to calculate the discrete Fourier transform:

```
File.open("signal.txt", "w+") {|out| out << dft(data)}
```

Then (in both languages) we calculated the absolute value squared, using built-in functions.

Signal energy

Signal energy is also an important parameter in signal processing and analysis. It is defined as in [14]:

$$E_x := \int_{-\infty}^{\infty} x^2(t)dt \tag{3}$$

Using energy of signal we can classify signals and this is a basic parameter of signal.

In the Ruby application, the energy of a signal was calculated by writing a method "`energy`" which uses simple iterators and arithmetic operators:

```
def energy(signal)
  signal2=signal.map {|x| x*x}
  signal2.inject(0) {|sum, x| sum+x}
end

data=[]
File.open("signal.txt") do |f|
   while line=f.gets
     data<<line.to_f
   end
end
puts energy(data)
```

MatLab is a language created for matrix-handling, so the code was easy to write:

```
s=load('signal.txt');
sum(s.^2)
```

Both, with the signal energy and the power spectrum, the signals were loaded from a file, and the result was written to a file.

Selected EMG signals

We selected three EMG recordings for our studies: from a healthy subject, one with myopathy and one with neuropathy. EMG records were obtained using 25 mm concentric needle electrode placed in tibialis anterior muscle. Subjects dorsiflexed the foot gently against resistance and then relaxed. All signals were obtained from Physionet [6]. Every signal had 8192 samples.

Results of EMG signal analysis using the Ruby language and MatLab

We obtained exactly the same results using MatLab and the Ruby language for dominant frequency (see [Tab. 1]). Also dominant frequency for all three EMG signals was very low. We couldn't distinguish between these signals using this Fourier based method. So Fourier transform based methods aren't perfect, especially for these biomedical signals, where we have frequently nonlinearity, nonstationarity and low frequency. In modern literature on biomedical signal analysis there are voices that suggest new methods, based on nonlinear dynamics [8]. Nevertheless, Fourier transform based methods are frequently first selected methods in any signal analysis, so we also want to present our results. In the future we plan to used some of the nonlinear methods (among others from chaos theory and fractal geometry) to these EMG signals to find out more about their nature.

Tab. 1. Values of dominant frequency for selected EMG signals

Selected EMG signal	Dominant frequency (Hz)
Healthy subject	0.063
Subject with neuropathy	0.062
Subject with myopathy	0.058

In [Tab. 2] we present obtained values of signal energy (exactly the same results using MatLab and Ruby). The highest values are for a signal from the subject with neuropathy, the lowest – for a signal from a healthy subject. Using signal energy we may distinguish a healthy subject from unhealthy subjects. Here we can see that in our pathological cases the signal is strengthened.

Tab. 2. Values of signals energy for selected EMG signals

Selected EMG signal	Values of signal energy
Healthy subject	21.485
Subject with neuropathy	417.070
Subject with myopathy	39.672

As see above, we have obtained some results which may suggest that Fourier transform based methods are not adequate to these signals and that in our pathological states energy of the biomedical signal is rising. Of course, these results need confirmation on larger groups of subjects, therefore we are very careful in our conclusions. Here, we have focused on the application of the Ruby language, but in the future we plan to investigate these matters more thoroughly.

Comparison of computation times using Ruby language and MatLab

In [Tab. 3] and [Tab. 4] we present a comparison of computation times of the Fourier transform and signal energy programs executed in MatLab and implemented in the Ruby language.

Tab. 3. Comparison of computation times of the Fourier transform using MatLab built-in "`fft`" function and our written "`fft`" function in the Ruby language (our signal length presented in bold font)

Signal Length (samples)	Computation time (s)	
	Ruby	MatLab
0	0.000	0.00000065
256	0.015	0.00027823
512	0.034	0.00000660
1024	0.081	0.00001123
2048	0.156	0.00002081
4096	0.329	0.00004239
8192	**0.699**	**0.00013816**
16384	1.526	0.00043267
32768	3.311	0.00108538
65536	7.066	0.00314686

Tab. 4. Comparison of computation times of signal energy using MatLab and the Ruby language (our signal length presented in bold font)

Signal Length (samples)	Computation time (s)	
	Ruby	MatLab
0	0.000000	0.0000013
256	0.000094	0.0000035
512	0.000187	0.0000045
1024	0.000421	0.0000054
2048	0.000952	0.0000099
4096	0.001591	0.0000134
8192	**0.002948**	**0.0000234**
16384	0.005616	0.0000446
32768	0.011092	0.0001032
65536	0.024242	0.0003448

Computation times dependency is nearly linear in all considered cases for commonly used signal lengths. Although Ruby's execution time seems poor in the comparison, it should be remembered that it is an interpreted language, while MatLab is compiled and highly optimized. And last, but not least, Ruby is free of charge, while MatLab is an expensive tool.

Conclusions

We have concluded that the Ruby language gives the same results as MatLab and although computation times are worse than in MatLab, we propose using the Ruby language. It is a simple, efficient and – in contrast to expensive MatLab – free of charge tool. The fact, that it is a second slower (for the common signal sizes) is not that much of a problem for scientific usage. The delay is negligible. In the future we plan to create Biomedical Signal Analysis Library in Ruby and we expect more applications of Ruby in biomedicine studies generally.

REFERENCES

[1] Berman J. J., Ruby Programming for Medicine and Biology, Jones & Bartlett Pub, 2007.

[2] Challis R. E., Kitney R. I., Biomedical signal processing (in four parts). Part 3: the power spectrum and coherence function, Medical & Biological Engineering & Computing, 29, pp. 225–241, 1991.

[3] Fitzgerald M. J., Learning Ruby, O'Reilly Media, Sebastopol, 2007.

[4] Flanagan D., Matsumoto Y., The Ruby Programming Language, O'Reilly Media, Sebastopol, 2008.

[5] Fulton H., The Ruby Way, Second Edition: Solutions and Techniques in Ruby Programming, Addison-Wesley Professional, Boston, 2006.

[6] Goldberger A. L., Amaral L. A. N., Glass L., et al., PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, Circulation, 101 (23), pp. e215–e220, 2000. [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]

[7] Goto N, Prins P., Nakao M., et al., BioRuby: bioinformatics software for the Ruby programming language, Bioinformatics, 26 (20), pp. 2617–2619, 2010.

[8] Klonowski W., Application of new non-linear dynamics methods in biosignal analysis, Proceedings of the World Medical Conference, Prague, Czech Republic, 26–28.09.2011, pp. 180–187, 2011.

[9] Lim D., A Ruby in the rough: using VHLLs in bioinformatics, Journal of Computing Sciences in Colleges, 21 (6), pp. 108–116, 2006.

[10] Olsen R., Eloquent Ruby, Addison-Wesley Professional, Boston, 2011.

[11] Perrotta P., Metaprogramming Ruby: Program Like the Ruby Pros, Pragmatic Bookshelf, 2010.

[12] Rascovsky S. J., Delgado J. A., Sanz A., et al., Informatics in Radiology: Use of CouchDB for Document-based Storage of DICOM Objects, Radiographics, 32 (3), pp. 913–927, 2012.

[13] Thomas D., Fowler Ch., Hunt A., Programming Ruby 1.9: The Pragmatic Programmers' Guide, Pragmatic Bookshelf, 2009.

[14] Zieliński T. P., Cyfrowe przetwarzanie sygnałów: od teorii do zastosowań, Wydawnictwa Komunikacji i Łączności, Warszawa, 2005.

The use of DRGs in hospital management

Petre Iltchev[1], **Aleksandra Sierocka**[2], **Michał Marczak**[1]

[1] Department of Health Care Policy, Faculty of Health Science, Medical University of Lodz, Poland

[2] Barlicki Hospital in Lodz, Medical University of Lodz, Poland

Abstract. Four years after the implementation of a payment system based on diagnosis related groups (DRGs) by the National Health Fund (NHF) in Poland, little research has been done on the use of DRGs in strategic management, controlling and managing hospital finances. Today's reality of managing health facilities forces their managers to take DRGs into account. This paper presents the possible use of DRGs in hospital management. The first part of the paper describes the nature of DRGs, while the second discusses best practices in the use of DRGs in hospital management and controlling. The NHF's policy of frequently modifying dictionaries describing DRGs and the way the NHF presents data on its web site hinder the application of DRGs in the strategic and operational management of hospitals. This paper is based on a case study of DRG use in the management of the Barlicki Hospital in Lodz (a clinical hospital of the Medical University of Lodz).

Introduction

Health care reforms as well as the restructuring and conversion of hospitals into non-public health care facilities (NZOZ) make it necessary to introduce new management methods, a new approach to effectiveness, labor efficiency, costs and performance, and DRGs should be part of these new tools. "Apart from their use in reimbursement systems, case-mix systems such as DRG were designed for planning, budgeting, management and financing inpatient care" [7]. Currently, the DRG system benefits mainly (only) the NHF. Hospitals are struggling to implement DRG-based planning, budgeting, and management.

Four years after the NHF implemented the DRG system, DRG data have yet to be applied in hospital management. DRGs may have a significant impact on hospitals' contracts and financial position. With proper use of data from the NHF DRG Statistics web service, hospitals can achieve a competitive advantage and increase their effectiveness. There is a huge potential for improving hospital management and profitability using DRG

ISBN 978–83–7431–336–0 ISSN 0860-150X

data. The key to success is to combine internal (hospital) and external (NHF) data. Few hospital managers have a vision of how to use DRGs for effective management and very few hospitals in Poland have incorporated DRGs into controlling and strategic management processes. The concept of including DRGs in hospital strategic management faces many challenges. Managers of public hospitals in Poland are rarely able to plan long term, or to see beyond the term of the hospital's contract with the NHF.

Hospital managers lack the knowledge and examples of best practices on how to implement DRGs at each level of management – from strategic to operational. The profitability of a hospital as a whole depends on the profits of each individual hospital unit (of each contract and process). The implementation of an appropriate level of detail in the measurement of hospital performance requires external data as a basis for benchmarking. Hospitals should introduce financial monitoring and controlling at the level of hospital departments and units. An increase in liability for medical and financial results at the lowest level of organizational structure leads to higher profitability.

This paper begins a discussion series aimed at extending management "beyond traditional 20th century hospital management." Over the years, controlling has been "the next big idea" in management theory, but few hospitals have put it into management practice.

The application of DRGs is presented as part of the process of building a hospital data warehouse. Some examples are also shown of how managers can compare costs on the basis of DRGs. DRG-based decision-making ideas and models may be used for the improvement of hospital management.

The history and nature of DRGs

DRGs were introduced in the 1960s, when Robert Fetter started a project aimed to compare the quality of medical services. The challenge he faced was to eliminate the impact of the state of patients' medical complications on the performance of health care facilities. Subsequently, DRGs were used to analyze the costs of medical services. The DRG system is used in most OECD countries. The first attempts to introduce DRGs in Poland were made in the years 2000–2003 by the Lower Silesian Sickness Fund, but the liquidation of Sickness Funds and the establishment of the NHF halted the above-mentioned efforts. Prior to July 1, 2008 the NHF used the Catalogue of Hospital Services to determine payments for hospitals. It was not until July 1, 2008 that hospitals around the country started

DRG-based reporting. Introducing DRGs, the NHF made the following assumptions:

a) an active influence on the costs of services;
b) a possibility for comparing the performance of hospitals.

The new way of determining payments has given hospital managers the opportunity to:

a) plan their budgets;
b) develop variants of contracts for negotiation with the NHF and other payers;
c) monitor, control and actively manage costs in accordance to DRGs;
d) link resource planning (medical personnel, medicinal products, hospital infrastructure-equipment, beds, etc.) with DRG-based contracts.

The Polish DRG system is based on only one principal diagnosis, and its distinguishing feature is that it also includes time spent in hospital and separate valuation of scheduled and emergency hospitalization [6]. The Polish DRG system contains 16 major categories and 519 groups. The basic dictionary includes: age, sex, mode of admission and discharge, and international classification of medical procedures.

Methodology

The methodology used for DRG-based hospital management in the various areas comprises of:

a) general data warehouse theory;
b) data warehouse design for hospitals;
c) controlling, benchmarking, strategic management.

The role of controlling

Controlling in hospitals can be considered part of managerial control. The purpose of controlling is to examine budget implementation and deviations from plans, and to calculate the costs and financial results for different units in the organization's structure. The DRG system is an additional dimension, specific to controlling processes in hospitals. The factors affecting the organization and frequency of controlling activities include:

a) the complexity and turbulence of the economic environment;
b) the value of the contract with the NHF (initially one should focus on controlling contracts with the highest value);

c) the organizational level at which goals are assigned, plans developed and responsibility for financial results delegated.

Many hospital mangers find it difficult to combine medical objectives with economic ones. Economic objectives affect the manner of management, which is "focused on the market of medical services, the rational use of resources and the rendering of services in compliance with the practice of medicine" [10]. Most public hospitals which have been converted into non-public hospitals may not expect a rapid increase in revenues. In the initial phase after conversion the NHF remains the basic source of revenues. This means that the survival and development of these facilities depends on their contracts with the NHF. After signing a contract, managers should focus on the effective management of costs and resource utilization [3]. The costs of contracts with NHF can be considered a typical optimization problem. At this stage, however, optimization does not seem to be used to a sufficient extent in hospital management. The issues of optimization of contracts signed with the NHF can be seen in terms of two basic strategies:

- maximizing the value of contracts and financial results given the available resources;
- minimizing the use of resources at a given value of the contract, with active management of costs.

While developing a model for optimizing the financial results of a contract, the function describing the costs should be constructed in the following way:

- it should be built on a multi-dimensional cost model which includes the mode of admission and discharge, sex and age of the patients;
- it should be taken into account that costs are not always proportional to the length of stay; costs are often the highest in the first few days, and gradually decrease along with improving patient health;
- it should be remembered that "The estimated values of c all being positive could be the result of the overall costs mainly being driven by the costs made for non-survivors, which is not surprising given the well known fact that dying patients are, on average, far more cost consuming than surviving patients." [3].

It is generally believed that the costs of medical services are undervalued in NHF contracts. This in turn means that contractor bears higher costs than are reimbursed by payments received from the NHF. The hospital may take steps to carry out the contract on a larger scale in order to avoid losses. Nevertheless, smaller facilities are left in a situation in which revenues from the contract will be lower than the costs. The only thing they can do is reduce costs. In the case of hospitals, it is important for the costs of particular DRGs to be lower than those posted on the NHF DRG Statistics web site. Hospitals may reduce costs by:

- reducing the costs of daily patient hospitalization and treatment;
- shortening the length of stay in hospital; if the hospital does not perform scheduled operations on Saturday and Sunday, then admitting a patient on Friday means increased costs and therefore it should be avoided as long as the patient's life is not at risk.

Cost analysis is based on expenses per patient admission. Costs include expenses incurred from admission to discharge [8]. The prerequisite for cost reduction is that income / financial result maximization must not deteriorate the quality of care provided. The purpose of analysis is to determine whether it is more cost effective to use expensive technology which would shorten the length of stay or use traditional therapy with longer stays. Daily expenses include all patient expenses (expenses which a patient has generated) per day. These expenses are usually the greatest during the first day after admission, but there are some exceptions and therefore separate models should be developed for each element of the multidimensional analytical space.

The use of a cost reduction strategy may collide with the other strategy, that is, shortening the time of stay. For example, if a less potent medical product is administered, the patient may stay longer in hospital and vice versa. Thus, we are faced with a dilemma: Which costs should be minimized? What should the cost function contain? How often should the cost function be updated? High costs may result from the low labor productivity of medical personnel or from insufficient use of expensive specialized equipment. The right path leading to lower costs is to increase the use of resources by gaining more clients who will complement contracts with the NHF.

A problem with the implementation of this strategy is the Polish employment law which prevents dynamic changes in the number of employees according to needs. Civil contracts are much more flexible in this respect, but hospital managers who exclusively use this form of employment risk that highly specialized doctors might easily change jobs. The employment of highly skilled staff should be governed by contracts of employment. Specialist equipment that may not be effectively used because of reduced NHF contracts is yet another problem. Under the circumstances, active management of operating costs to maximize financial results is a partial solution for hospital mangers. Active cost management requires the implementation of a controlling system and an appropriate change in the organizational structure of the health care facility.

How to create value with DRGs?

The key to ensuring the financial success of a hospital by using DRG-based payments is managing the profitability of contracts at the DRG level. This means that a hospital obtains higher revenues than costs for a particular DRG. Given the fact that in Poland there exist both public and non-public hospitals, it is easy to predict the winners and losers. The non-public ones will be in a better position because in public health care institutions medical goals take precedence. DRGs, if used appropriately, may be an asset for the organization. Therefore, DRG data should be used adequately in hospital management. DRG analysis is similar to balance sheet analysis:

- change of data over time characterizing a given DRG and the major category;
- change of proportion of particular major categories over time;
- change of proportion of a given DRG in a given major category and in overall DRGs.

Having data concerning the performance of a contract in a given hospital at the DRG level, one can compare these results with data from the NHF DRG Statistics web service. This shows where a given hospital stands relative to other hospitals in terms of costs, length of hospitalization, ICD-9 and ICD 10 medical procedures used, and mode of patient admission and discharge.

In order to efficiently utilize DRG information, one needs to have detailed historical data concerning contracts: a) from the hospital's IT system; b) from the NHF DRG Statistics web service for all hospitals. The managers theoretically have detailed information on DRGs obtained from the hospital's IT system. What is important in analysis of the potential of DRG use in the process of hospital management is to have data on hospital costs broken down by DRG, wards, cost centers, patients, and medical procedures. A hospital's IT system may combine both of the abovementioned sources of data to use them in management. The following stages of DRG analysis may be distinguished:

a) extracting data from the hospital's IT system;
b) importing NHF data;
c) combining data from these two sources;
d) data analysis – comparing, developing models, describing data, and making forecasts on their basis. New analyses using DRGs and new data from the NHF DRG Statistics data web service make it possible to compare hospitals in particular regions and by hospital type. Thus,

one can identify the best, average, and the worst hospitals. An example here is a comparison of clinical hospitals across Poland.

The role of data warehouses

Data warehouses are often used as a platform supporting strategic management processes. In Poland it has not been until now that hospital managers find it necessary to use such solutions. A data warehouse, being an analytical platform, facilitates such analyses as:

a) monitoring the achievement of medical and financial goals by a hospital;
b) analysis of a hospital's efficiency and benchmarking vs. all hospitals/hospitals of the same kind;
c) analysis of profitability by DRG;
d) analysis of tendencies in use of medical services;
e) analysis of cost levels;
f) analysis of cost influencing factors and their change over time (e.g. length of hospital stay);
g) analysis of the structure of a medical service by age, sex, and mode of admission and discharge.

Typical questions that may be addressed by analysis of processed and aggregated data from the NHF DRG Statistics web service include:

- What is the share of 10 DRGs with the highest value in the hospital's contract with the NHF?;
- Where do those DRGs come in the NHF ranking?;
- What are the factors that determine differences between the hospitalization time in a given hospital and the average hospitalization time as given by the NHF for a particular DRG?;
- The share and cost of a given DRG in a given major category;
- The share and cost of a given DRG against all DRGs;
- Average daily costs for a given DRG;
- Which DRGs are characterized by the highest/lowest costs?;
- Patients with which DRGs are hospitalized for the longest/shortest time?

The development of a data warehouse in a hospital must bring economic benefits. The managers must learn how to create business value using the implemented data warehouse. In this context, it would be useful to address questions such as: What new business value and benefits can be gained using a data warehouse? Where does the hospital stand relative to its competitors? This question can be answered only if the hospital uses benchmarking to compare its results with those of other hospitals. The role of DRG analysis increases in a fast-changing environment with substantial changes in the costs of medical technologies, labor, and entry of new market players (competitors).

Of utmost importance is the manner of extracting, transforming, cleaning and combining external and internal data. From the business point of view it is important to have a documented approach to combining DRG data from the NHF and from the hospital. It is better to combine dictionary categories in the process of analysis rather than while building a data warehouse. In this way, one can ensure consistency with NHF data.

Before one can commence data analysis, it is necessary to develop a data model. The key issue here is to determine which data will be obtained from which source. Modeling data for DRG use in hospital management is an element of a larger project – developing a data warehouse for the hospital. The data model was designed using the free application BizAgi Process Modeler. The model employs the main functional components which are crucial for the hospital's controlling. The first data warehouse model was expanded by reengineering data obtained from the NHF DRG Statistics web site. The objective was for the application to create value for the hospital's managers as early as in the initial phase of development.

Hospital managers do not wait until the full comprehensive solution is in place. During the implementation of the project, management consultants train the hospital managers on how to use DRGs in controlling and strategic decision making. Seeing and assessing the real benefits and business value connected to DRG data, the hospital's managers will support the subsequent part of the project, which is aimed to increase the number of dimensions and the level of detail of available data.

The following analytical tools for storing and processing multidimensional data may be identified:

a) spreadsheets with pivot tables;
b) databases;
c) Google Fusion Tables.

Each of these tools has its strengths and weaknesses. Spreadsheets can be comfortably used for importing NHF data concerning several DRGs and their comparison with a hospital's performance. However, with greater volumes of data the advantage of database management systems is evident. In turn, Google Fusion Tables are a useful tool for data presentation and visualization and ensuring good data accessibility.

Comparative analysis of DRGs in the Barlicki Hospital (a clinical hospital in Lodz) with the NHF DRG statistics

Hospital managers may find it difficult to extract value from DRGs. To achieve this, they need to compare internal DRG data with external data. The first step is to import data from the NHF DRG Statistics web service. Subsequently, data from the NHF service need to be combined with the hospital's data. After this process has been completed, it will be possible to analyze and compare the performance of the hospital with that of other hospitals. Such an analysis may equip the hospital's managers with information that may be used for making strategic choices concerning the direction of developing the hospital's medical activity. To draw business benefits from DRGs, the hospital's managers need to have some knowledge in such areas as controlling, performance management and strategic management. The necessary data must be prepared by analysts and the IT personnel. This team needs to have expertise on databases, data warehouses, multidimensional modeling and OLAP technology.

Some of the challenges related to importing data from the NHF web site with Excel include:

a) data are not available in a spreadsheet, but are presented on web pages;

b) there are no clear rules for generating top-level page addresses for DRGs or second-level page addresses for NHF regions and hospital types;

c) data may not be readily imported form the NHF web site to a spreadsheet; if a NHF branch does not have any patients with a given DRG, that branch does not show the value "0"; rather, it is not displayed on the web page at all.

Data imported from the NHF DRG Statistics web service are stored in 13 sheets:

- the first one contains all data from tables 1 to 5;
- data in the second tab are data from table 6 of the web service – information on ICD-9;
- the third tab presents data from table 7 – ICD-10 codes;
- the other tabs contain detailed data concerning NHF branches and hospital types.

In order to decrease the volume of data imported from the DRG Statistics web service, one should decide which DRGs are most vital to the hospital. This can be done according to contract values, number of patients or man-days.

Some of the practical problems related to data analysis that must be solved in the process of modeling include the lack of DRG dictionaries and frequent modifications of DRG dictionaries.

To provide an example of data analysis, this paper presents data concerning those DRGs that are most important to the Barlicki Hospital. It was assumed that dictionaries describing the data are to incorporate those elements that were present throughout all the studied years. [Tab. 1] presents a dictionary of admissions and [Tab. 2] a dictionary of discharges. These dictionaries will be expanded in the future as the NHF adds new elements.

Tab. 1. Dictionary of admission modes used in 2009–2012

Description of admission mode	Applied in DRG in the year			
	2009	2010	2011	2012
Scheduled admission	Yes	Yes	—	—
Scheduled admission based on referral	—	Yes	Yes	Yes
Emergency admission with referral from the emergency	Yes	—	—	—
Emergency admission resulting from transfer by a medical emergency team	—	Yes	Yes	Yes
Emergency admission with referral other than from the emergency	Yes	—	—	—
Emergency admission without referral	Yes	Yes	—	—
Emergency admission – other cases	—	Yes	Yes	Yes
Admission of a newborn as a result of childbirth in this hospital	—	Yes	Yes	Yes
Scheduled admission of a person who benefited from health care services out of turn under a privilege afforded her by the law	—	Yes	Yes	Yes
Transfer from another hospital	—	—	—	Yes
Admission of a person subjected to mandatory treatment – admissions related to the implementation of statutory compulsory treatment set out in art. 26 of the Act of 26 October 1982 on upbringing in sobriety and counteracting alcoholism and art. 33. 1 and art. 34.1 of the Act of 5 December 2008 on preventing and fighting infections and infectious diseases in humans	—	—	—	Yes
Forced admission – forced admission in connection with the statutory obligation to submit to hospitalization referred to in art. 35.1 of the Act of 5 December 2008 on preventing and fighting infections and infectious diseases in humans, art. 21.3, art. 23, 24 and 29 of the Act of 19 August 1994 on the protection of mental health, art. 30 and 71.1 and 3 of the Act of 29 July 2005 on counteracting drug addiction, art. 94, 95a and 96 of the Act of 6 June 1997 on Penal Code, art. 203 and 260 of the Act of June 6, 1997 on Code of Criminal Procedure and art. 12 and 25a.2 of the Act of 26 October 1982 on proceedings in juvenile cases	—	—	—	Yes

Source: Based on DRG data from the NHF site: "DRG Statistics". Available at https://prog.nfz.gov.pl/APP-JGP/KatalogJGP.aspx, accessed on 12.06.2012.

Tab. 2. Dictionary of discharge modes used in 2009–2012

Discharge mode	Applied in DRG in the year			
	2009	2010	2011	2012
Completion of a therapeutic or diagnostic process	Yes	Yes	Yes	Yes
Referral for further treatment in an outpatient clinic	Yes	Yes	Yes	Yes
Referral for further treatment – other cases	Yes	Yes	—	—
Discharge at the patient's own request	Yes	Yes	Yes	Yes
Death of patient	Yes	Yes	Yes	Yes
Referral for further treatment in another hospital	—	Yes	Yes	Yes
Referral for further treatment in a stationary care facility	Yes	—	—	—
Referral for further treatment in a stationary care facility other than a hospital	—	Yes	Yes	—
Referral for further treatment in a long-term care facility	Yes	—	—	—
The person treated left a stationary care facility without formal discharge before the completion of a therapeutic or diagnostic process	—	Yes	Yes	—
Discharge under art. 22.1.3 of the Act of August 30, 1991 on health care facilities	—	Yes	Yes	—
Referral for further treatment at a unit (other than a hospital) of a health care facility offering therapeutic medical services such as stationary and 24 h medical care;	—	—	—	Yes
The person treated left a unit of a health care facility offering therapeutic medical services such as stationary and 24 h medical care without formal discharge before the completion of a therapeutic or diagnostic process	—	—	—	Yes
The person treated, admitted with code "9" or "10", left the hospital without formal discharge	—	—	—	Yes
Discharge under art. 29.1.3 of the Act of 15 April 2011 on medical activity	—	—	—	Yes

Source: Based on DRG data from the NHF site: "DRG Statistics". Available at https://prog.nfz.gov.pl/APP-JGP/KatalogJGP.aspx.

What are the transformation rules for dictionaries describing admissions and discharges? These rules may only be determined on the basis of analyzing and mapping medical conditions. Without appropriate mapping of dictionaries used by the NHF in particular years, it would be impossible to compare hospitals' activity in respect of a given DRG in a detailed and accurate way. A comparison that takes into account patients' age and sex, medical procedures used, and mode of admission and discharge may form the basis for comparing one hospital with others in terms of both their medical and economic performance.

Another problem with DRG data analysis is the lack of grouping rules, or a so-called grouper, which is no longer made available by the NHF.

Tab. 3. Analysis of DRG groups with the highest values at the Barlicki Hospital in 2010 using data from the "DRG Statistics" service

		Barlicki Hospital			Main data for selected DRG groups			NHF Lodz region			Clinical hospitals			Municipal, county, city hospitals			District hospitals		
DRG	**Description of DRG**	N	L	W	N	L	W	N	L	W	N	L	W	N	L	W	N	L	W
A11	Comprehensive intracranial treatment	0.27	9	1.7	6.6	11	1.4	0.43	11	1.4	4.3	11	1.4	0.3	15	1.0	2.0	13	1.2
A23	Major operations on the spinal cord and spinal canal	0.23	8	1.4	11.0	8	1.4	1.2	8	1.4	4.5	8	1.4	1.4	6	1.9	4.5	8	1.4
B13	Uncomplicated cataract surgery by emulsification with simultaneous lens implantation	0.93	2	1.6	106.0	2	1.5	8.2	2	1.5	16.1	2	1.5	18.3	2	1.5	31.8	2	1.5
B12	Complicated cataract surgery by emulsification with simultaneous lens implantation	0.62	2	1.8	54.6	2	1.7	3.1	2	1.8	14.4	2	1.8	9.8	2	1.8	16.0	2	1.8
Q01	Endovascular aortic aneurysm repair	0.03	6	10.3	1.7	7	8.9	0.16	6	10.3	1.1	7	8.9	0.2	6	10.3	0.3	7	8.7
L94	Kidney Transplant – category II	0.04	4.5	9.5	0.7	19	2.3	0.08	22	2.1	0.5	19	2.3			0	0.2	19	2.4
F11	Comprehensive gastric and duodenal surgery	0.17	4	3.1	3.7	13	0.9	0.3	8	1.6	1.3	8	1.5	1.0	15	0.8	1.2	14	0.9
G34	Endoscopic and percutaneous procedures on biliary tract and pancreas	0.48	3	1.4	19.5	4	1.0	1.3	4	1.0	5.0	4	1.0	6.2	4	1.0	5.7	5	0.8

Notes: N – Number of hospitalizations (000); L – Length of stay, median (days); W – Average price for 1 day of hospitalization (PLN 000)

NordDRG is an example of a DRG system that gives access to information about the grouper [9].

The present study compared the number of patients, costs and length of stay in the Barlicki Hospital with other Polish clinical hospital. We focused on those DRGs with the greatest share in the hospital's budget or the largest number of patients or man-days.

[Tab. 3] presents 8 top DRGs in terms of contract value or number of patients or man-days. Furthermore, it is shown why one should use detailed data in medical-economic analyses. As it can be easily seen from the table, comparisons of the Barlicki Hospital's performance with overall NHF data and with data for clinical hospitals lead to very different results. The difference in costs and average length of hospital stay (median) between particular hospitals in terms of the selected DRGs may influence the comparison results.

Conclusions

The DRG system in Poland should not be limited to contracting, reporting to the NHF and determining NHF payments to hospitals.

In the process of DRG implementation in hospital management we gain experience and test various approaches to data collection, cleaning and aggregation. The managers voice their opinions about reports and identify potential future improvements. Further development of data warehouses should focus on tapping external data sources. Real benefits from data warehouses may be gained when they are used in combination with dashboards in the process of management. Polish experiences in terms of employing DRGs in hospital management are particularly relevant for countries which have yet to implement a DRG system or which have introduced it only recently.

REFERENCES

[1] BizAgi Process Modeler, [http://www.bizagi.com/index.php?option=com_content&view=article&id=126& Itemid=127&dwl=3b17460c2172fa142a8add7a95e9b283&lang=en]

[2] Enterprise Warehouse Solutions, Inc, World-Class Data Warehousing Models: Healthcare – Clinical, Enterprise Warehouse Solutions, Inc, 2009. www.EWSolutions.com (10.02.2010 r.)

[3] Hanna V., Sethuraman K., The Diffusion of Operations Management Concepts into the Health Care Sector, 2005. http://www.mbs.edu/download.cfm?DownloadFile=951E3EB1-123F-A0D8-42DC3582CE6ECFE6 (21.12.2011)

[4] Imhoff C., Galemmo N., Geiger J., Mastering Data Warehouse Design, Wiley, 2003.

[5] Jegers M., Applying cost minimization techniques to hospitals: A comment, European Journal of Operational Research, 197, pp. 828–829, 2009.

[6] Kozierkiewicz A., Jednorodne grupy pacjentów. Przewodnik po systemie, Narodowy Fundusz Zdrowia, Centrala, Warszawa, 2009.

[7] Serden L., Lindqvist R., Rosen M., Have DRG-based prospective payment systems influenced the number of secondary diagnoses in health care administrative data?, Health Policy, 65, pp. 102, 2003.

[8] Suthummanon S., Omachonu V., Cost minimization models: Applications in a teaching hospital, European Journal of Operational Research, 186, pp. 1175–1183, 2008.

[9] The NordDRG Manual documents the grouping rules of the NordDRG system, http://www.nordcase.org/eng/norddrg_manuals/

[10] Zbroja A., Decyzyjny rachunek kosztów w szpitalu – konieczność czy alternatywa?, w: Zarządzanie finansami placówek medycznych, Instytut Przedsiębiorczości i Samorządności, Warszawa, 2001. http://www.emedyk.pl/artykul.php?idartykul_rodzaj=8&idartykul=1 (21.12.2011).

[11] Ustawa z dnia 15 kwietnia 2011 r. o działalności leczniczej, Dz. U. Nr 112, poz. 654.

Identification and analysis of adverse events on the example of SP ZOZ in Swidnica

Jacek Domejko[1], **Aleksandra Sierocka**[2], **Adam Rybicki**[3], **Mariusz Piechota**[4], **Michał Marczak**[5]

[1] Gynecological and Obstetric Hospital in Walbrzych, Poland

[2] Barlicki Hospital in Lodz, Medical University of Lodz, Poland

[3] Hospital District Fürstenland Toggenburg, Switzerland

[4] Department of Anaesthesiology and Intensive Therapy, Military University Hospital in Lodz, Poland

[5] Department of Health Care Policy, Faculty of Health Science, Medical University of Lodz, Poland

Abstract. Identification of a series of various complications which happen during hospitalisation and then their classification enables understanding how many factors may influence their occurrence. The described solution shows an example of adverse event analysis along with an indication of preventive and prophylactic activities. The research material enabled finding 187 out of 1285 case histories in which an adverse event occurred, and also to establish its form (type). Then, as a result of general analysis directed at the total elimination or partial reduction of the identified events, 6 general recommendations were made.

Introduction

Adverse events may occur at every stage of health service provision by a medical entity, especially by a hospital services provider. We define them as: "harm to the patient's health caused during the diagnostic and/or treatment, not related to the natural course of the illness or the patient's condition, and also the risk of its occurrence" [3].

In order to identify and then establish a directory of the abovementioned events, hazards should be taken into account resulting not only from the activity of the medical personnel, the used devices, but also the drugs and medical treatments used in the course of therapy, and the diagnostic, therapeutic, nursing and rehabilitation procedures.

The development of current knowledge, science and technology, as well as the rapid development of medicine means that we can effectively cure

ISBN 978–83–7431–336–0 ISSN 0860-150X

illnesses that were previously incurable. Unfortunately this requires a series of invasive tests or a series of complicated surgical procedures. These procedures frequently significantly increase the risk of making a mistake.

Methods

The "black spots"[1] method used in this work was created by the team of prof. Michał Marczak [5, 8]. It consists of risk identification, ordering of threats, indication of possible causes and proposing preventive measures. Using the analysis of selected hospital stays as an example, an attempt was made to model patient safety measures. When constructing models, a series of various factors was taken into account, related e.g. to the specifics of the procedures, that is the performed procedures (in accordance with ICD-9), hospitalization time, amount of days with vascular catheter, number of surgical procedures, number of infections, number of bedsores, duration of surgery and critical events which have occurred during hospitalization, resulting in death/disability/longer stay etc.

Material

Research material including 1285 case histories of patients treated in the hospital in 2010 was obtained from SP ZOZ "Latawie" in Swidnica. Analysis was performed for selected stays at the following departments: Anaesthesiology and Intensive Care, Gynaecology and Obstetrics, Neurology, General Surgery, Neonatal Physiology and Pathology, Trauma and Orthopaedics Surgery and Cardiology. The collected data was additionally verified by cross-analysis of epidemiologic nurse reports with results of microbiological test cultures. All unclear issues and doubts were settled by the prof. Marczak's team through an additional, detailed analysis of discharge abstracts only or the entire case histories.

In order to establish which of the incidents which occurred may be considered to be adverse events, the obtained results were consulted with both

[1] It is a set of various methods and partial analyses (expert analyses, epidemiological monitoring etc.), frequently with various methodology, to which other components may be connected as needed. It is cyclical and has a block algorithm structure, which means that after removing the most dangerous "black spots" that are removable (taking into account the economical and organisational state of the unit) in another iteration on the subsequent level of hierarchy the risk analysis, selection of "black spots" etc. are performed again.

internal experts (hospital employees) and external experts (from outside the unit).

Results and discussion

The collected data enabled the establishing of all adverse events, which have occurred during the provision of analysed medical services. The obtained results can be presented as follows [Tab. 1–7]:

Tab. 1. Number and types of complications which happen during selected hospitalisations – Anaesthesiology and Intensive Care

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
30	11	supraventricular tachycardia	3	10
		gastrointestinal bleeding	1	3.33
		tracheotomy wound bleeding requiring surgical correction	1	3.33
		atrial fibrillation	2	6.67
		sudden cardiac arrest	9	30
		unplanned extubation	2	6.67
		obturation of the intubation tube	1	3.33
		pneumothorax	1	3.33
		acute respiratory failure	7	23.33
		infection of upper respiratory tract	2	6.67
		post-operative wound infection caused by *Acinetobacter baumanii*	1	3.33
		otitis media	1	3.33

source: own research

Tab. 2. Number and types of complications which happen during selected hospitalisations – Trauma and Orthopaedics Surgery

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
164	38	removal of thigh bone screw impossible	1	0.6
		infection of upper respiratory tract	1	0.6
		colliquative necrosis of tissue near post-surgery scar caused by MRSA[1]	1	0.6

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
		colliquative nercrosis of fatty tissue after tissue contusion	1	0.6
		skin necrosis above metal fasteners	1	0.6
		arm skin necrosis	1	0.6
		localised skin necrosis	1	0.6
		post-surgery ischaemia	9	5.5
		radial nerve paresis	1	0.6
		unsuccessful attempts to place the plug screw at the top of femoral nail	1	0.6
		dyspnoea with tachycardia	1	0.6
		bed sore in the area of sacral bone	1	0.6
		intra-gluteal area bed sore	1	0.6
		sacral area bed sore	1	0.6
		heel bed sore	1	0.6
		serum blisters on the lower legs	4	2.4
		fracture of the femoral bone (in the greater trochanter area) when implanting the endoprosthesis	1	0.6
		repeated hospitalisation caused by post-surgical wound infection after a previous procedure, caused by *Staphylococcus aureus*	1	0.6
		post-surgical respiratory and circulatory failure	1	0.6
		post-operative respiratory failure after general anaesthesia	1	0.6
		post-surgical wound dehiscence caused by MRCNS[2], MLSB[3]	1	0.6
		jumping out of the window and breaking both heel bones	1	0.6
		hip injury as a result of a fall	1	0.6
		serum and blood discharge from the post-surgical wound	1	0.6
		serum and pus discharge from the post-surgical wound	1	0.6
		serum discharge from the post-surgical wound	5	3
		urinary tract infection	1	0.6
		bone infection caused by *Streptococcus pyogenes*	1	0.6
		post-operative wound infection caused by *Acinetobacter baumanii*	1	0.6

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
		pulmonary embolism	2	1.2
		ischemic myocardial infarction	1	0.6
		post-puncture syndrome	2	1.2
		shedding of epidermis in both antecubital spaces	1	0.6

source: own research

[1] MRSA – methicyllin-resistant Staphylococcus aureus
[2] MRCNS – methicillin resistant coagulase negative Staphylococcus
[3] MLSB – Macrolide-Lincosamide-Streptogramin B

Tab. 3. Number and types of complications which happen during selected hospitalisations – Neonatal Physiology and Pathology

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
109	20	catarrhal infection of the nose and throat	2	1.8
		eye infection caused by *Escherichia coli*	1	0.9
		sudden cardiac arrest	1	0.9
		urinary tract infection caused by *Enterobacter cloace*	1	0.9
		urinary tract infection caused by *Enterococcus faecalis*	5	4.6
		urinary tract infection caused by *Enterococcus faecium* HLAR[1]	2	1.8
		urinary tract infection caused by *Escherichia coli*	4	3.7
		urinary tract infection caused by *Klebsiella pneumoniae*	1	0.9
		urinary tract infection caused by *Proteus mirabilis*	1	0.9
		urinary tract infection caused by *Staphylococus epidermidis* MRCNS, MLSB	2	1.8
		respiratory tract infection caused by *Pseudomonas aeruginosa*	1	0.9
		blood infection caused by *Morganella morgani*	1	0.9
		blood infection caused by *Staphylococcus epidermidis MRCNS*	1	0.9
		eye infection caused by *Haemophilus influenzae*	1	0.9

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
		ear infection caused by methicillin-susceptible coagulase-negative *Staphylococus*	1	0.9
		infection of the navel area caused by *Enterococcus faecalis*	1	0.9
		infection of the navel area caused by *E. coli*	1	0.9

source: own research

[1] HLAR – high-level aminoglycoside resistance

Tab. 4. Number and types of complications which happen during selected hospitalisations – General surgery

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
219	31	haematoma in the wound	1	0.46
		gastrointestinal haemorrhage	1	0.46
		sudden cardiac arrest	4	1.83
		respiratory failure	4	1.83
		psychotic symptoms	1	0.46
		post-surgical respiratory failure	3	1.37
		re-amputation	1	0.46
		repeated surgery due to the necessity of excising necrotic tissue around the post-surgical wound	1	0.46
		post-surgical wound abscesses	1	0.46
		retroperitoneal space abscess	1	0.46
		bile discharge	3	1.37
		bile discharge around the Kehr tube	1	0.46
		significant ileus	1	0.46
		post-operative wound infection caused by *Candida spp*	2	0.91
		post-operative wound infection caused by *Enterococcus faecalis*	1	0.46
		post-operative wound infection caused by *Escherichia coli*	2	0.91
		post-operative wound infection caused by *Klebsiella oxytoca*	1	0.46
		post-operative wound infection caused by *Pseudomonas aeruginosa*	1	0.46

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
		post-operative wound infection caused by *Staphylococcus aureus*	2	0.91
		post-operative wound infection caused by methicillin-resistant coagulase-negative *Staphylococcus*	1	0.46
		post-operative wound infection caused by *Streptococcus gr.* G	1	0.46

source: own research

Tab. 5. Number and types of complications which happen during selected hospitalisations – Gynaecology and Obstetrics

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
300	10	diarrhoea	1	0.33
		stomach aches, fluid in the abdominal cavity	1	0.33
		infection	1	0.33
		post-surgical wound bleeding requiring repeated surgery	1	0.33
		oedema of lower limbs	1	0.33
		paralysis of the femoral nerve	1	0.33
		vomiting and pain in the area of the pubic symphysis and epigastrium	1	0.33
		electrolyte and protein disorders	1	0.33
		post-surgical wound infection	2	0.67
		post-operative wound infection caused by *Escherichia coli*	1	0.33

source: own research

Tab. 6. Number and types of complications which happen during selected hospitalisations – Cardiology

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
325	48	anaemia	1	0.31
		3rd degree AV block with cardiac arrest – implantation of an electrode for temporary cardiac pacing	1	0.31

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
		serious cardiac decompensation	2	0.62
		ventricular tachycardia	1	0.31
		ventricular tachycardia with blood pressure decrease	1	0.31
		tachycardia with QRS complexes with a left bundle branch block morphology	1	0.31
		heart failure decompensation (IV NYHA[1]) accompanying severe aortic stenosis	2	0.62
		Cx dissection in PCI[2]	1	0.31
		short term consciousness loss episode	1	0.31
		hypokaliaemia	1	0.31
		hypotonia	1	0.31
		hypoglycaemia	1	0.31
		infection of the glans	1	0.3
		nosebleed	1	0.31
		haematuria	1	0.31
		heparine-induced thrombocytopenia	1	0.31
		sudden cardiac arrest	4	1.23
		ischemic cerebral stroke with right side paresis	1	0.31
		acute heart failure	1	0.31
		NYHA 3rd degree heart failure	1	0.31
		NYHA 4th degree heart failure	5	1.54
		decompensated diabetes	1	0.31
		hypotonia and low cardiac output symptoms	1	0.31
		pulmonary oedema	1	0.31
		waiting for equipment for the haemodynamic lab	1	0.31
		oliguresis	1	0.31
		acute respiratory and circulatory failure	1	0.31
		acute post-contrast kidney failure with hypotonia	1	0.31
		acute post-contrast kidney failure	5	1.54
		acute duodenal ulcer with haemorrhage	1	0.31
		coronary vessel perforation with bleeding into the pericardial sac and loss of arterial blood pressure	1	0.31

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
		psychomotor agitation of the patient (resulting from the removal of the catheter and IV cannula)	1	0.31
		superficial injury of the right forearm due to the patient slipping under the shower	1	0.31
		zoster	1	0.31
		technical problems with the angiograph	1	0.31
		revision of coronary vessels due to the dissection of the RCA proximal section[3] and constriction of the left posterior descending artery	1	0.31
		diabetic foot with tissue necrosis, hypoalbuminemia	1	0.31
		false aneurysm in the place the right femoral artery is punctured for coronarography	1	0.31
		TIA[4] during PCI	1	0.31
		vomiting	1	0.31
		infection of the central cathether insertion place by *Staphylococcus epidermidis*	1	0.31
		infection of the central cathether insertion place by *Staphylococcus aureus*	1	0.31
		wound infection caused by *Proteus vulgaris*	1	0.31
		pneumonia caused by *Klebsiella pneumoniae*	1	0.31
		pneumonia caused by *Neisseria spp.*	1	0.31
		pneumonia caused by *Pseudomonas aeruginosa*	1	0.31
		pneumonia caused by *Staphylococcus aureus*	2	0.62
		fainting without loss of consciousness	1	0.31
		significant hypercholesterolemia	1	0.31

source: own research

[1] NYHA – New York Heart Association
[2] PCI – percutaneous coronary interventions
[3] RCA – right coronary artery
[4] TIA – transient ischemic attack

Tab. 7. Number and types of complications which happen during selected hospitalisations – Neurology

Number of tested patients	Number of hospitalizations with complications	Type of complication/event	Number of events	Event frequency (%)
138	19	hyponatremia	1	0.72
		gastrointestinal bleeding	3	2.17
		massive oedema of limbs	2	1.45
		sudden cardiac arrest	2	1.45
		ischemic cerebral stroke	1	0.72
		heel bed sore	1	0.72
		sacral area bed sore	1	0.72
		acute haemodialysis caused by the increase of Ca, ionised Ca, creatinine and urea levels	1	0.72
		acute kidney failure in the course of infection	1	0.72
		ulceration of the bed sore in the area of sacral bone and buttocks	1	0.72
		tarry stool	1	0.72
		respiratory tract infection caused by *Staphylococcus haemolitycus*	1	0.72
		urinary system infection caused by *Acinetobacter baumanii*	1	0.72
		urinary system infection caused by *Escherichia coli*	2	1.45
		urinary system infection caused by *Klebsiella pneumoniae*	1	0.72
		urinary system infection caused by *Staphylococcus haemolitycus*	1	0.72
		urinary system infection caused by coagulase-negative *Staphylococcus*	1	0.72
		pneumonia	1	0.72
		pneumonia caused by *Acintetobacter baumanii*	1	0.72
		pneumonia caused by *Proteus mirabilis*	1	0.72

source: own research

The adverse events listed above (their occurrence) may be mostly prevented, taking into account in the further activity of the medical entity, the recommendations below.

General recommendations:

1. Improving the cooperation between non-surgical wards personnel and the anaesthesiology and intensive care ward personnel [7].

Justification:

In order to decrease the mortality indicator on non-surgical wards most important factor is to identify the patients with a high risk of complications or death early on. Appropriate, early, proper treatment of these patients enables avoiding many hazardous complications or even death. Good cooperation between the non-surgical wards personnel and the anaesthesiology and intensive care personnel enables appropriately early transfer of critically ill patients, especially in worsening clinical condition from a non-surgical ward to the anaesthesiology and intensive care ward, before the most significant complications occur, e.g. cardiac arrest, severe respiratory failure or circulation failure.

Advanced life support techniques available at Anaesthesiology and Intensive Care Wards (and not available at other wards) are intended to ensure temporary assistance to the basic vital functions of the patient, which were significantly disturbed in the process of a potentially reversible illness.

2. Improving the cooperation between the surgical wards personnel and the anaesthesiology and intensive care ward personnel [1, 7, 10].

Justification:

In order to decrease the mortality indicator on surgical wards the most important factor is to implement for the patients at risk appropriate preventive measures before the surgery, early identification by an anaesthesiologist of patients with an increased risk of death or post-surgical complications, and subjecting these patients to specialised anaesthesiological care directly after the surgery. The proposed solution is particularly justified by the conclusions from scientific research presented below: many post-surgery complications may be prevented by early risk identification and therapy, first 48 hours after the surgery is a critical period for high risk patients, planned transfer of high risk patients directly to intensive care wards should be seriously considered, since this could significantly decrease post-surgical mortality, among patients treated at a hospital in a non-optimum manner before accepting them at the intensive care ward an increased mortality was established, the longer the patient was present in the hospital before being accepted into intensive care, the higher was the mortality, many surgical patients could benefit from being transferred to an intensive care ward, but they are not given this opportunity, appropriate monitoring and proper treatment of surgical patients may significantly decrease the mortality, some high risk patients should be assigned to intensive care ward in the post-surgery period, planned acceptance at the intensive care ward may significantly decrease mortality during the post-surgery period, elderly patients should be sent directly to intensive care wards after surgical procedures. For the elderly, the National Confidential Enquiry into Perioperative Deaths of 1999 recommends better cooperation between surgeons, anaesthesiologists and doctors with specialist knowledge on the care of elderly.

Appropriate cooperation between the surgical wards and the intensive care ward should not be confined only to patients qualified for surgical procedures (undergoing surgical procedures). It should also apply to patients undergoing diagnostic procedures, treated conservatively, or sick in days after surgeries (not under special supervision due to a surgery which they underwent). In such case the justification presented in item 1 is still valid for this group of patients.

3. Perform internal audit concerning the identification, monitoring and treatment of patients in a serious condition [2].

Justification:

In accordance with the RESUSCITATION GUIDELINES 2010, "early recognition of the patient's condition and prevention of circulatory failure form the first link of the survival chain. In case of a circulatory failure inside the hospital less than 20% of patients survive until they are discharged from the hospital. Circulatory failure occurring in patients on wards without monitoring is not a sudden or unpredictable event, nor caused by primarily cardiologic reasons. In this group of patients a slow and progressing deterioration of the general condition, including hypoxemia and hypotension, which remain unnoticed by medical personnel, or are diagnosed, but are not adequately treated. With many of these patients an unmonitored heart failure occurs, and the rhythm which causes it is usually not suitable for defibrillation. In the medical documentation of patients with whom UHF occurs, or who have suddenly required acceptance at the Intensive Care Ward (ICW) there is frequently evidence showing lack of diagnosis or lack of treatment of occurring respiratory and circulatory disorders. Insufficient care frequently includes: infrequent, late or incomplete assessment of basic vital signs; no knowledge concerning their proper values; the design of observation charts is not good enough; low frequency and specificity of "track and trigger" systems; insufficient number of medical personnel, and thus lack of ability to monitor the patients and provide them with better care. A frequent problem is the inefficient treatment of respiratory tract patency disorders, circulatory and respiration disorders, inappropriate use of oxygen therapy, weak communication, lack of teamwork."

The audit should provide an answer to the following questions: Is the medical personnel provided with training concerning the symptoms of the worsening general condition of the patient and is there a need for rapid action in order to improve the existing situation? Do hospital wards use proper and regular monitoring of the patient's basic vital signs? Are there clear guidelines at the hospital wards, helping the medical personnel in early detection of the patient condition worsening? Is there a simple, unified system for calling for help at the hospital? Do severely ill patients receive proper and timely help?

4. Enter into the documentation of case history a serious condition identification chart ("early warning scale") [6, 9].

Justification:

Currently at many hospitals, in order to identify the patients requiring enhanced monitoring, treatment or specialised consulting, early warning scales are used, or criteria for calling a resuscitation team. Various scales (point systems) are used to assess the clinical condition of the patient, in which the help of the members of such a team is required. They include, among others, Early Warning Scoring System, Modified Early Warning Score. The action of the teams is based on both an alarming value of a single parameter and on obtaining an appropriate amount of points in a complex point system. Meeting these criteria calls the appropriate personnel to the patient's bed. The original Early Warning Scoring System was created on the base of the APACHE scale, which is used to assess the condition of patients at intensive care wards. The system is based on physiological changes, such as heart rate, breathing rate or blood pressure value.

5. Create a "rapid response team" in the hospital [4, 12].

Justification:

Studies conducted in many countries have shown that serious adverse events occur in 15–20% patients sent to the hospital. Up to 80% of adverse events is preceded by physiological or biochemical disorders (or irregularities), which occur a few hours, or even a few days earlier. Introducing this type of supervision over hospitalized patients has resulted in a reduction of the amount of sudden cardiac arrest cases at hospital wards, and sometimes also a reduction of mortality. An increased survivability of patients undergoing major surgery was shown. The National Institute for Health and Clinical Excellence has recommended the "placement of rapid reaction teams" in the hospitals as one of the twelve actions used to prevent (reduce) the sickness and mortality rates as a part of 100 000 Lives Campaign.

6. In the hospital medical those procedures should be performed, in which the personnel has a significant experience and which are performed in large amounts [11].

Justification:

Size and type of the hospital may be important when performing some surgical procedures. In many studies the risk of death was lower and the length of hospital stay shorter in large clinical hospitals when compared to small clinical hospitals or non-clinical hospitals. The size of the hospital was the most important (the highest difference in survivability) in case of elderly and high-risk patients. The difference in mortality between a hospital with a small amount of beds, and a hospital with a high amount of beds amounted to more than 5% in case of lung or oesophagus resection, 2–5% in case of stomach resection, urinary bladder excision, surgery on a non-ruptured abdominal aortic aneurysm, replacement of mitral or aortic valve, and below 2% in case of colectomy, lobectomy or nephrectomy.

Conclusions

The described solution shows an example of adverse event analysis along with an indication of preventive and prophylactic activities.

Identification of a series of various complications which happen during hospitalisation and then their classification enables understanding how many factors may influence their occurrence. This factor may be both e.g. a team of specialists or a unit manager without appropriate qualifications, as well as defective medical equipment.

Unfortunately, gaining full knowledge about adverse events is not simple. Collecting all necessary information is a very difficult and arduous process. It requires not only appropriate experience, but also comprehensive knowledge about the organisational structure of a given facility, its specifics and character. However, it forms a necessary and required part of the process enabling not only the identification of the source of potential risks, but also indicate those, that cause the most damage and are important for the correct operation of the health care facility, including the safety of the patients.

The research material enabled finding 187 out of 1285 case histories in which a adverse event occurred, and also to establish its form (type). Then, as a result of general analysis directed at the total elimination or partial reduction of the identified events, 6 general recommendations were made. It should be also noted, that the next stages of the study, not covered by this publication, include: ordering of threats, typing of black spots, creating the map of black spots, creating guidelines and detailed procedures applying to specific types of events classified as black spots, in the order they were assigned.

REFERENCES

[1] Bennett-Guerrero E., Hyam J. A., Shaefi S., et al., Comparison of P-POSSUM risk-adjusted mortality rates after surgery between patients in the USA and the UK, Br J Surg, 90, pp. 1593–1598, 2003.

[2] Gamil M., Fanning A., The first 24 hours after surgery. A study of complications after 2153 consecutive operations, Anaesthesia, 46, pp. 712–715, 1991.

[3] Hajdukiewicz D., Risk reduction in a hospital – programme for action, Menedżer Zdrowia, 5, pp. 52–59, 2004.

[4] Jones D., Bellomo R., Devita M. A., Effectiveness of the Medical Emergency Team: the importance of dose, Crit Care, 13, pp. 313, 2009.

[5] Marczak M., Active methods of risk management, systems diagnostics and determinants of the Polish health care system, in: Logical, Statistical and Computer Methods in Medicine, edited by Milewski R., Surowik D., Medical University in Bialystok, pp. 121–143, 2011.

[6] Morgan R. J. M., Williams F., Wright M. M., An early warning scoring system for detecting developing critical illness, Clin Intensive Care, 8, pp. 100, 1997.

[7] National Confidential Enquiry into Perioperative Deaths, UK: HMSO, London, pp. 712–715, 1999.

[8] Sierocka A., Adverse event risk management, using an example of one of the clinical hospitals in Łódź, PhD thesis, Health Care Policy Department of the Medical University in Lodz, 2010.

[9] Stenhouse C., Coates S., Tivey M., et al., Prospective evaluation of a modified Early Warning Score to aid earlier detection of patients developing critical illness on a surgical ward, Br J Anaesth, 84, pp. 663, 2000.

[10] Turner M., McFarlane H. J., Krukowski Z. H., Prospective study of high dependency care requirements and provision, J R Coll Surg Edinb, 44, pp. 19–23, 1999.

[11] Urbach D. R., Baxter N. N., Does it matter what a hospital is “high volume” for? Specificity of hospital volume-outcome associations for surgical procedures: analysis of administrative data, BMJ, 328, pp. 737–740, 2004.

[12] Wachter R. M., Goldman L., The emerging role of “hospitalists” in the American health care system, N Engl J Med, 335, pp. 514–517, 1996.

Social sciences methodology vs medical approach to quality in health care

Jacek Michalak[1]

[1] Department of Quality of Services, Procedures, and Medical Standards, Medical University of Lodz, Poland

Abstract. The search of lists of medical specialties, scientific disciplines and requirements of medical and social science journals was performed to find the convergences between social science and medical approach to study quality in health care. 14 databases were also searched with 3 sets of key words. 420 top-listed articles were hand-searched to disclose 34 reports fulfilling inclusion criteria. The results indicated that the number of medical specialties exceeded 50, the number of scientific disciplines is continuously increasing and the medicine and some social sciences are put together into “applied sciences”. However, the multidisciplinary research seems to be promising to overcome the difficulties in comparison of quality in health care in different countries. Team work requires more precise definitions used in social and medical sciences to avoid the problems resulting from the different understanding of the same terms in different disciplines and specialties.

Introduction

Methodology is a guideline system for solving a problem, with specific components such as phases, tasks, methods, techniques and tools. According to Merriam-Webster dictionary, methodology is defined as:

– a body of methods, rules, and postulates employed by a discipline
– a particular procedure or set of procedures
– the analysis of the principles or procedures of inquiry in a particular field [http://www.merriam-webster.com/dictionary/methodology]

Referring to “classic” division of science one can disclose methodology of:

– Exact science (e.g. physics)
– Natural science (e.g. biology)
– Social science (e.g. economics)

There are different classifications of scientific methodologies. It can be easily seen that medicine – recognized as the art and science of healing – cannot be attributed neither to social nor natural science. That implicates that medical methodology is something different from methodology

ISBN 978–83–7431–336–0 ISSN 0860-150X

of natural science. However, the Vancouver protocol has been generally accepted by all biomedical journals since 1978 and approved by the National Library of Medicine in 1979. According to that protocol the Uniform Requirements for Manuscripts Submitted to Biomedical Journals is now a must for authors intending to publish an article in any biomedical journal [3, 16]. The structure of a biomedical article should reflect the methodology of the biomedical research. According to current version of Uniform Requirement the format of an article encouraged for original research articles is:

a) Introduction (a brief, logical lead-in to the subject, hypothesis, and objectives),
b) Materials and Methods,
c) Results, and
d) Discussion.

However, this format does not include several other important and integral parts of an article [3].

No similar "uniform requirements" have been implemented in social sciences. The authors representing social sciences, even those investigating health problems, do not need to obey such rigid requirements. The structure of an article ("main text") may be quite different from that of a biomedical journal [17]. The British Sociological Association supports the following structure/elements of a typical scientific project or paper:

a) Conception or design.
b) Data collection and processing.
c) Analysis and interpretation of the data.
d) Writing substantial sections of the paper (e.g. synthesizing findings in the literature review or the findings/results section).

Starting from the conception (hypothesis to be verified) or from the background (towards the proven hypothesis) seems to be the most important difference between "medical" and "social science" approaches. The question arises whether the differences in the structure of an article may be attributed to the different methodology of different scientific disciplines. Moreover, it may be important to establish whether the differences of methodologies would lead to different results and conclusions.

The quality of health care is an excellent example to study. Considering this problem as a medical quality, social understanding of quality and, last but not least, managerial approach to quality may result in different methodology and significantly different opinions. The European Health Consumer Indices is an example of such mixture of different methods [9]. EHCI combining different types of indicators (epidemiology, public opinion survey and different requirements and expectations) is used to compare the health care systems in European countries:

- Patient rights and information (12 indicators)
- Accessibility – waiting times for treatment (5 indicators)
- Outcomes (8 indicators)
- Prevention/ Range and reach of services provided (10 indicators)
- Pharmaceuticals (7 indicators)

However, it is difficult to prove how, for example the indicator called "layman pharmacopeia[1]" is a factor measuring the quality of health care system. The same problems are with, for example "undiagnosed diabetes". How can one count the undiagnosed entity without diagnosis? EHCI must not be mismatched with ECHI (European Community Health Indicators) [10] containing 88 health indicators which describe and measure:

- demographic and socio-economic situation
- health status, health determinants
- health interventions: health services
- health interventions: health promotion

It seems that differences between EHCI and ECHI (as well as other sets of indicators) result from different methodologies used, so different comparisons on the quality in health care may provide even conflicting opinions. This study was aimed at the methodological aspects of research on the quality in health care to disclose whether significant methodological differences and their consequences exist between social sciences and medical sciences.

Material and methods

The following searches were performed to compile lists of medical specialties, classification of scientific disciplines, and top-rated articles in the scientific databases. The lists of medical specialties were taken from the official documents and legal regulations in the European Union, the USA and Poland. The classification of scientific disciplines was based on monographs, legal regulations in Europe and the USA, and research papers found by scrutinizing 5 databases: three multidisciplinary bases (Science Direct,

[1] Pharmacopeia is highly specific, pharmacists addressed, and it contains the procedures and descriptions of medicinal compound. It is necessary to be educated in pharmacy to understand such information. The information directed to the patient must be simplified and adjusted to an average patient's knowledge. That is an important difference.

OvidSP/MEDLINE/Embase, EBSCO), and two medical bases: HighWire Press and PubMed. Search of Google and Google – Scientific Reports was performed as well. Search of Science Direct was performed in Springer journals and Elsevier journals. The latter were searched in the following categories: economics; business, management, accounting; mathematics; medicine; social science.

The key words used in following combinations:

1. social sciences methodology medical quality health care
2. social sciences methodology quality health care
3. medical methodology quality health care

Ten top listed articles in each of the search were based on the best matches, relevance, percentage of hits. Hand search was also performed to clarify and to verify the information. Ten top listed articles in each out of 42 searches (14 bases, 3 combination of key words) has been reviewed. The following criteria of exclusion were used: methodology not sufficiently described, quality not precisely defined. Only the articles referring directly to the methodology of social science and medical methodology were included.

Results and Discussion

The most comprehensive and covering almost all medical specialties in one document is the Directive 2005/36/EC[2]. It contains 54 medical specialties [Tab. 1]. Some of the Polish specialties (e.g. forensic medicine or hypertensiology) are not mentioned here. Currently in Poland there are 37 "basic" specialties and 30 "detailed" medical specialties. On the other hand, the number of subspecialties may be different in different countries. In the USA the following 12 specialties are recognized inside "neurology": behavioral neurology, clinical neurophysiology, geriatric neurology, headache medicine, neuromuscular medicine, neurodevelopmental disabilities, neuro-oncology, neuroradiology, vascular neurology, hospice and palliative medicine, pain medicine, sleep medicine. Above mentioned abundance of disciplines and specialties make it difficult to compare different specialties in different countries.

[2] Note that there is substantial overlap between some of the specialties and it is likely that, for example "clinical radiology" and "radiology" or "dental, oral and maxillo-facial surgery" and "maxillo-facial surgery" refer to a large degree to the same pattern of practice across Europe. The rest of medical professions e.g. specialties in pharmacy, dentistry, nursery, laboratory diagnostics, orthoptics etc. are not discussed here.

Tab. 1. Medical specialties in EU Directive

Allergology	Anaesthetics	Biological hematology
Cardiology	Child psychiatry	Clinical biology
Clinical chemistry	Clinical neurophysiology	Clinical radiology
Dental, oral and maxillo-facial surgery	Dermatology	Dermato-venerology
Endocrinology	Gastro-enterologic surgery	Gastroenterology
General hematology	General surgery	Geriatrics
Immunology	Infectious diseases	Internal medicine
Laboratory medicine	Maxillo-facial surgery	Microbiology
Nephrology	Neurology	Neuro-psychiatry
Neurosurgery	Nuclear medicine	Obstetrics and gynecology
Occupational medicine	Ophthalmology	Orthopaedics
Otorhinolaryngology	Paediatric surgery	Paediatrics
Pathology	Pharmacology	Physical medicine and rehabilitation
Plastic surgery	Podiatric Medicine	Podiatric Surgery
Psychiatry	Public health and Preventive Medicine	Radiology
Radiotherapy	Respiratory medicine	Rheumatology
Stomatology	Thoracic surgery	Tropical medicine
Urology	Vascular surgery	Venerology

For example, “Anesthetics[3]” is not the same as “Anesthesiology”. So, it can be accepted that medicine itself uses different methodologies depending on the subject of studies (specialty), and the term “medicine” should not be used as a general description of all medical activities. The term “quality” in case of forensic medicine must have quite different meaning as compared with internal medicine. Public health researchers often use queries, like in social sciences, nevertheless the Vancouver protocol is obeyed in public health journals. Unfortunately, the social sciences use the term “medicine” or “medical” in the broad meaning, that leads the misunderstandings.

[3] The term “anesthetics” refers usually to medical products used in anesthesiology.

Also the classification of scientific disciplines is generally difficult, and seems to be a never-ending process. Continuous development of different disciplines, emerging new interdisciplinary sciences as well as overlapping of the fields of interests, result in difficulties in delineating the borders between scientific disciplines. Abbott proposed to manage such "chaos" by placing 44 disciplines into 5 branches [1]:

1. Humanities (history, linguistics, literature, performing arts, philosophy, religion, visual arts)
2. Social sciences (anthropology, archaeology, area studies, cultural and ethnic studies, economics, gender and sexuality studies, geography, political science, psychology, sociology)
3. Natural sciences (space science, earth sciences, life sciences, chemistry, physics)
4. Formal sciences (computer sciences, logic, mathematics, statistics, systems science)
5. Professions and applied sciences (agriculture, architecture and design, business, divinity, education, engineering, environmental studies and forestry, family and consumer science, health science[4], human physical performance and recreation, journalism, media studies and communication, law, library and museum studies, military sciences, public administration, social work, transportation) [1].

On the other hand, the number of disciplines in social science may be increased up to 11 (in such a case linguistics is a social science) or even 19 (including e.g. communication studies and public administration). Pieter identified nine types of scientific methods applicable to all scientific disciplines: observation, intuitive method, source criticism, survey, critical analysis, experimental method, monographic method, case study, diagnostic survey [20]. No specific "social science" or "medical" methodology was disclosed. One can discuss whether this list is complete, but the alternative is the 44 methodologies, at least one for each discipline. Moreover, Abbott's list does not cover the next 27 "applied science", one of them is medicine[5], and the number of medical specialties excesses 50. The search for articles gave surprising results [Tab. 2].

[4] Medicine is understood here as one type of health science.

[5] Medicine – "science and art of healing. It encompasses a variety of health care practices evolved to maintain and restore health by the prevention and treatment of illness in human beings" This definition is obviously too narrow, it does not cover e.g. public health, occupational medicine, pathology, forensic medicine etc.

Tab. 2. The search results of articles and web sites obtained by the use 14 databases

Database	Key words: ***social sciences*** ***medical sciences*** methodology quality health care	Key words: ***social sciences*** methodology quality health care	Key words: ***medical sciences*** methodology quality health care
OVID SP/MEDLINE/ Embase	5 136	5 467	6 773
Science Direct Springer	2	7	75
Science Direct Elsevier	18 904	26 202	34 750
Business, management, accounting	815	2 141	1 003
Economics	685	1 568	813
Mathematics	177	286	336
Medicine	14 351	16 293	27 077
Social Science	3 764	6 211	4 094
HighWire	21 350	29 570	46 996
PubMed	10 446	257 979	455 933
HighWire + PubMed	21 420	29 786	48 853
EBSCO	151 984	217 001	322 815
Google	11 300 000	3 710 000	11 800 000
Google – Scientific reports	703 000	1 170 000	2 110 000

It can be easily seen that the number of retrieved articles is very high, except in one case (Science Direct – Springer). Even in the category "mathematics" the search yielded 177 to 336 items. It can be noted that similar percentages of "social science" and "medical" articles are found in top-listed articles. However, the numerous, important reports have not been retrieved in such a way. The design of this study was based on the assumption that the search in 14 databases would provide the most relevant papers which match the sets of key words. A prediction was made that the same most relevant articles would be retrieved from different databases and the social sciences methodology will be more often used than "medical" methodology in health care. However, the results were astonishing, as in each database the top-listed articles were different and no single paper articles were found twice in different databases. It seems that, for example HighWire and EBSCO use different methods of retrieving data and it is possible to obtain different results. When Science Direct-Elsevier was used – the search engine covers journals of one publisher only – the top-listed articles were different in different searches. First of all, no single paper was found in which the social sciences methodology and medical methodology were applied to the same problem, e.g. multidisciplinary approach to quality in health care.

The results of hand search yielded 34 articles out of 420 searched. This is a better proportion than in other medical investigations. For example, Smetana et al. accepted only 49 articles – out of 2459 citations found in MEDLINE [21]. Even in one type of search, data from different sources may differ significantly. Machlin et al. in an economic study compared four sources on ambulatory care in the USA: Medical Expenditure Panel Survey (MEPS), National Health Interview Survey (NHIS), National Ambulatory Medical Care Survey (NAMCS) and National Hospital Ambulatory Medical Care Survey (NHAMCS) [18]. The numbers of visits to hospital emergency departments were estimated as 46.3 million (MEPS), 52.4 million (NHIS) or up to 90.3 million (NHAMC).

The content of the top-listed articles (sorted according to their relevance) revealed that the type of methodology was not limited to "social science" or medical journals. Some publications in journals devoted to public health use the same methodology as social science. But there were some important differences. The first problem resulted from the terminology of medical disciplines and specialties, as mentioned above. Next – the problem of precision in defining terms and attributes. It is well illustrated by the results of Delphi studies by Haggerty et al. [12]: only 5 out of 25 attributes (accessibility, first-contact, continuity-relational, family-centered care, intersectoral team, population orientation) were recognized as specific to primary care. Though the article was published in a medical journal, the social science methodology was used. However, it remained to be clarified how to distinguish and quantify the "more intersectoral" from "less intersectoral" team. The term "social capital" a corner-stone in health promotion can be described as a metaphor, rhetoric or science. The latter is an association between social and economic factors and health, but the authors have suggested that "the concept of social capital may add little and may perhaps even act to dilute social health initiatives already in place (under the various names of community health promotion, community development, empowerment and capacity building)" [13].

Totally different approach to the problem of quality in health care is based on the quantitative indicators. Asch et al. constructed aggregate scores from 439 indicators of the quality of care for 30 chronic and acute conditions and for disease prevention [2]. It is noteworthy that the authors limited their activities to selected conditions and tried to measure as much as it is possible, but not generalized to the general population. Such ”medical” approach narrowed the scope of the study, but increased its specificity. Sometimes the economic terms implemented into medical investigation may lead to serious consequences. Efficacy studies of the drug/procedure should not be mismatched with effectiveness studies, as efficacy refers to “optimal circumstances” and effectiveness to “usual circumstances”. It means that each type of studies covers different population (eligible subjects vs. any subjects), different intervention (fixed regimen/forced titration vs. flexible regimen) compliance (high vs. low) and outcomes (condition-specific vs. comprehensive) [4]. The results of efficacy study when applied to hospital-acquired infection may indicate quite different way of management as compared to the results of effectiveness study [19]. Deccache et al. also noted the need of combining public health and social sciences methodology to evaluate the quality of health promotion [7]. On the other hand, Camargo et al. stressed that the scientific medicine is not necessarily good medicine, also indicating the need of interdisciplinary approach [5]. Fielding – a sociologist – postulated to overcome the bipolar academic and applied research settings as distinct spheres and to broaden mixed methods research (MMR) applied to research on social aspects of health and illness [11]. From the medical point of view, Wensing advocated implementing social science methods into health sciences research [22]. However, there are some tendencies to generalize and to create models met in social science publications. The comparison

Tab. 3. Managerial and medical cultures: points of divergence according to Davies [2000]

	Managerial culture	Medical culture
Structure:	Bureaucratic	Collegial
Group loyalty:	Low	High
Job security:	Low/medium	High
Disciplinary base:	Social sciences	Natural sciences
Evidence base:	Case studies on organisations	Clinical studies on patients
Focus:	Patients as groups	Patients as individuals
Skills:	Managerial/human relations	Biomedical/technical
Allegiance:	Organisation/corporate goals	Patient/professional
Success measure:	Efficiency	Effectiveness
Quality emphasis:	Consumer rated quality	Technical quality
Performance review:	Public	Confidential
Public trust:	Low	High (but vulnerable)

of managerial with medical cultures by Davies contains some simplifications [Tab. 3], but it may indicate more convergences then divergences in both types of applied sciences [6].

Simplifications cause misunderstandings. For example, managerial and "medical" skills are both based on human relations; technical quality is only an element of quality of health care from the medical point of view; efficiency and effectiveness – depending on their definitions used – are equally important in managerial and medical culture [6]. The above mentioned "efficacy" in medical studies has not necessary the same meaning as the "efficacy" in economic analyses. Nevertheless, tendencies to combine different methodologies indicate the way of future development of social and medical sciences.

Conclusions

1. Methodology of social sciences and medical sciences still differs significantly one from another, but increasing tendencies in combining them can be seen in the last decade.

2. Multidisciplinary research seems to be promising to overcome the difficulties in comparison of quality in health care in different countries.

3. Team work requires more precise definitions used in social and medical sciences to avoid problems resulting from different understanding of the same terms in different disciplines and specialties.

REFERENCES

[1] Abbott A., Chaos of Disciplines, University of Chicago Press, 2001.

[2] Asch S. M., Kerr E. A., Keesey J., et al., Who Is at Greatest Risk for Receiving Poor-Quality Health Care?, N Engl J Med, 354, pp. 1147–56, 2006.

[3] Barron J. P., The Uniform Requirements for Manuscripts Submitted to Biomedical Journals Recommended by the International Committee of Medical Journal Editors, CHEST, 129, pp. 1098–1099, 2006.

[4] Bombardier C., Maetzel A., Pharmacoeconomic evaluation of new treatments: efficacy versus effectiveness studies? Ann Rheum Dis, 58 (Suppl I), 182–185, 1999.

[5] Camargo K., Coeli C. M., Theory in Practice: Why "Good Medicine" and "Scientific Medicine" are not Necessarily the Same Thing, Advances in Health Sciences Education, 11 (1), pp. 77–89, 2006.

[6] Davies H. T. O., Nutley S. M., Mannion R., Organizational culture and quality of health care, Quality in Health Care, 9, pp. 111–119, 2000.

[7] Deccache A., Evaluating quality and effectiveness in the promotion of health: approaches and methods of public health and social sciences, Promot Educ., 4 (2), 10–15, 1997.

[8] Directive 2005/36/EC of the European Parliament and of the Council of 7 September 2005 on the recognition of professional qualifications. L 255/22 EN Official Journal of the European Union 30.9.2005.

[9] Euro Health Consumer Index 2012.
http://www.healthpowerhouse.com/files/Report-EHCI-2012.pdf

[10] European Community health indicators.
http://ec.europa.eu/health/indicators/echi/index_en.htm

[11] Fielding N., Mixed methods research in the real world, International Journal of Social Research Methodology, 13 (2), pp. 127–138, 2010.

[12] Haggerty J., Burge F., Lévesque J. F., et al., Operational Definitions of Attributes of Primary Health Care: Consensus Among Canadian Experts, Ann Fam Med, 5 (4), pp. 336–344, 2007.

[13] Hawe P, Shiell A., Social capital and health promotion: a review, Social Science & Medicine, 51, pp. 871–885, 2000.

[14] Health Consumer Powerhouse, Euro Health Consumer Index, 2012.

[15] http://www.merriam-webster.com/dictionary/methodology
(date of access June 20th, 2012).

[16] International Committee of Medical Journal Editors ICMJE: Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Writing and Editing for Biomedical Publication, Updated April 2010,
http://www.icmje.org/urm_full.pdf (date of access June 20th, 2012)

[17] Kumar S., Calvo R., Avendano M., et al., Social support, volunteering and health around the world: Cross-national evidence from 139 countries, Social Science & Medicine, 74, pp. 696–706, 2012.

[18] Machlin S. R., Valluzzi J. L., Chevarley F. M., Thorpe J. M., Measuring ambulatory health care use in the United States: A comparison of 1996 estimates across four federal surveys, Journal of Economic and Social Measurement, 27, pp. 57–69, 2001.

[19] Michalak J., Orszulak-Michalak D., Farmakoekonomiczne aspekty postępowania w zakażeniach szpitalnych. w: Denys A. (red.) Zakażenia szpitalne, wybrane zagadnienia, ABNC Wolters Kluwer Business, pp. 320–347, Warszawa, 2012.

[20] Pieter J., Ogólna metodologia pracy naukowej, Ossolineum, Wrocław, 1967.

[21] Smetana G. W., Landon B. E., Bindman A. B., et al., A Comparison of Outcomes Resulting From Generalist vs Specialist Care for a Single Discrete Medical Condition. A Systematic Review and Methodologic Critique, Arch Intern Med, 167, pp. 10–20, 2007.

[22] Wensing M., Research methods from social science can contribute much to the health sciences, J Clin Epidemiol., 61 (6), pp. 519–520, 2008.

Identification of adverse events with the use of the black spots method – own example

Anna Staszewska[1], **Aleksandra Sierocka**[2], **Adam Rybicki**[3], **Michał Marczak**[1]

[1] Department of Health Care Policy, Faculty of Health Science, Medical University of Lodz, Poland

[2] Barlicki Hospital in Lodz, Medical University of Lodz, Poland

[3] Hospital District Fürstenland Toggenburg, Switzerland

Abstract. For hospitals, which provide medical services, the opportunity to identify their specific risks would not only be beneficial for the patients and personnel, but would also help ensure smooth operations of the hospitals themselves. The specification of adverse events and various errors and their classification makes one realize how many factors and elements may affect their incidence. Errors may be caused not only by inadequate qualifications of specialists, by a lack of cooperation between the staff taking direct part in the diagnostic and therapeutic processes (e.g. doctors, nurses, technicians, laboratory technicians, etc.), or by the head of the facility, but also by the medical equipment used in the facility. Gaining a full understanding of adverse events is a necessary and indispensable part of the process of identifying the sources of possible risks and indicating those that cause the most damage and are important for the proper functioning of the entire healthcare facility, and also for the safety of the patient. The black spot method discussed in the paper is one of the most important and most widely used risk management method.

Introduction

The importance of the issues on this topic is especially important nowadays, with the very rapid development of medicine, universal and diverse medical and nursing care, demographic changes which cause an increase of the average life duration, which in consequence translates to the increase in the number of patients. A diverse scope of medical services unfortunately also causes an increased risk of occurrence of adverse events.

Risk management consists of appropriate implementation of appropriate procedures and the use of optimum methods of estimating and controlling the level of adverse events which may occur in specific conditions [11]. In the risk management process it is important to analyse and estimate significant risk groups, which occur in a hospital, and also the creation of

ISBN 978–83–7431–336–0 ISSN 0860-150X

strategies and methods enabling for their management and minimisation. One should remember, that some risk groups are present in some medical entities regardless of specifics, other than in established conditions, since they are determined by the type of activity and the scope of provided services. It is thus impossible to create a universal list of adverse events[1], which occur in a hospital. It is however possible to create their general list which will enable the ordering of individual events and assessment of their results. Risk areas present in a hospital may be classified in the following categories [3–4]:

I group: Human factor – causes threats and is subject to them. Hazardous events may be caused by medical errors, by making bad decisions. It should be added that interpersonal communication (patient – patient, personnel – patient, personnel – personnel) may also cause an increase of the risk level.

Risk factors:

- individual characteristics (such as sex, age, character) – influence the risk both on the risk increasing factors and the factors on which it depends whether a given person (both among the personnel and the patients) will be resistant to a given risk or not,
- medical errors
- making of incorrect decision (e.g. discharging oneself from the hospital may lead to an increase of a risk of illness and loss of health)
- improper communication between people, no understanding
- professional stress, stress of patients and visitors
- no will to cooperate between the personnel and patients.

II group: Hospital infections[2] – both patients, personnel and persons with even a singular contact with the hospitals are susceptible. The period in which the infection has revealed itself is important (it was arbitrarily decided that it is 48 hours from signing into or out of a hospital, and in case of long incubation period (HBV, HCV, HIV, tuberculosis) a period no longer than its incubation period).

[1] In the article the term "adverse event" covers a wider scope than "medical events", which in accordance with the amendment of the Patient Rights and Patient Rights Spokesman Act include the infection of a patient with a biological pathogenic factor, damage of the body, health disorder or death.

[2] Hospital infection: "is an infection, which occurred as a result of provision of health services, when the illness: a) was not during its incubation period when the health services were provided, or b) occurred after the provision of health services during a period not exceeding its incubation period" (5 December 2008 Infection and Infectious Disease Prevention and Treatment Act; Dz. U. 2008 No, 234 item 1570).

Risk factors:

- microorganisms present in a hospital environment,
- resistance of exposed persons,
- infection routes and entries
- sterilisation and disinfection technology.

III group: Occupational hazards

Risk factors:

- working time and discipline,
- OSH and ergonomics,
- working space,
- availability, freedom of movement,
- lighting,
- heating, ventilation,
- sanitary unit,
- access to drinking water,
- mess room.

IV group: Manually performed work

Risk factors:

- handling, transporting patients,
- procedures requiring long and precise use of tools,
- loads of joints.

V group: Dangerous substances – when identifying and analysing individual risk groups it should be noted, that in accordance with the 20 April 2004 Medical Products Act, there are four classes of medical products: I, IIa, IIb and III, which indicate a risk related with the use of a medical product. Classification was made, taking into account the criteria of invasiveness of the product, the place of contact and the place of use. The product class is established by the manufacturer, taking into account the anticipated use of the product[3].

Risk factors:

- chemical compounds – lead, asbestos,
- reagents, drugs,
- bodily fluids,
- poisonous waste,
- compounds hazardous due to other reasons – flammable, chemical catalysts, explosives etc.

[3] 20 April 2003 Medical Products Act, chapter 3 – Classification and qualification of medical products (Dz. U. no. 93 of 30 April 2004).

VI group: Medical equipment and tools

Risk factors:

- life support equipment,
- rehabilitation equipment,
- diagnostics and prevention equipment,
- imaging equipment,
- sterilization and autoclaving equipment,
- ionizing radiation,
- the use of screens and displays,
- vending machines,
- devices connected to the current,
- periodical inspections and maintenance.

VII group: Accidents

Risk factors:

- injuries requiring first aid,
- fires,
- hazardous spills,
- choking on food, drugs,
- cuts, damages to the skin.

VIII group: Other

Risk factors:

- violence, threats,
- working in dangerous places,
- noise,
- stress influencing the mental condition.

In order to effectively manage the risk, individual risk groups should be not only precisely identified, but also a strategy should be created, enabling a systematic analysis and assessment of the risk.

Materials and methods

A team of specialists[4], which undertook an innovative risk management enterprise, initially at a hospital in Swidnica, has created a map of "black spots" (areas with a significant concentration of adverse events), that is established "especially dangerous places, which correspond to a point event (...)" [1].

[4] The team consists of: Director of the Swidnica hospital – Jacek Domejko, assistant professor Mariusz Piechota – anaesthesiologist from the University Clinical Hospital in Lodz, prof. Michał Marczak, Aleksandra Sierocka PhD.

In the health care system medical entities, the following places are considered black spots [7]:

- especially dangerous, which are point events (ward, operational unit, central sterilisation point of the hospital, pharmacy etc.) or specific medical procedures which, when performed incorrectly, contribute to the occurrence of complications,
- special concentration of adverse events, the occurrence of which is more frequent than the average number of events aggregated in accordance with the accepted measurement scale.

The black spot method uses an event tree analysis and fault tree analysis method[5] [10]. It qualifies an adverse event as a black spot, and then establishes its risk levels. This method is repeatable and has an algorithmic block structure, that is after eliminating previously specified and ordered black points (among these which may be removed), one "moves onto" another level of ordering, on which a repeated risk analysis is performed [10].

The method of black spots used for the research[6] forms a whole in connection with other risk management methods (event tree analysis, fault tree analysis, expert analysis, epidemiological monitoring) [7]. Its first stage (initial results and conclusions will be analysed and presented in this article) is the identification of adverse events, which occurred during hospitalization. Subsequent stages will include detailed ordering (up to three levels), statistical analysis of medical events, and then presenting remedial actions [7].

Three departments of one of the Lodz clinical hospitals were included in the research (ophthalmology, cardiology and rheumatology and intensive care). Research material includes case histories and other medical documentation of patients treated in the period since 2011. Until today, 243 of complete case histories were analysed. Special care was paid to: type of stay at the ward (planned/emergency), basic and coexisting diagnosis, basic and additional procedures, injections/insertions of a needle (observation chart of peripheral intravenous needle insertions), presence of a fever and other

[5] Event tree analysis (ETA) and Fault tree analysis (FTA) methods form a graphical representation of causal dependencies, concerning accidental events. The are used as tools for the analysis of economic projects, organisational, business and manufacturing investments, and also for assessing hazards for personnel present at an area, at which the occurrence of hazards to their health and life is present. These methods take into account various adverse event results.

[6] Empirical part of the studies and detailed description of the method will be presented in a PhD thesis under the direction of prof. Michał Marczak, entitled "Adverse event risk management using the example of Maria Konopnicka University Clinical Hospital in Lodz".

complications, data from the anaesthesiological protocol and survey, drugs taken).

Consultations were also made with specialists, in order to verify and establish which identified events should be qualified as adverse events.

Analysis of test results was performed by taking into account detailed description of adverse event risk factors, such as:

- cause of the event: character and complexity of the surgery, invasive diagnostics, nursing care, doctor actions, hospitalisation conditions: bed occupancy, number of hired medical personnel, existence of procedures and standards; environmental conditions: sanitary and epidemiological condition of the rooms, sterility of equipment and medical materials,
- human conditions: age and sex of the patient, health and coexisting illnesses, personnel qualifications, obeying the existing procedures and standards.

It is important that with every identified adverse event the cause of this event is analysed individually.

Results

After analysing 98 medical charts from the cardiological department, 5 medical events were noticed, of which the most frequent was nausea [Tab. 1].

Tab. 1. Medical events at the cardiology ward

Type of medical event	Number of events
nausea	2
38.6° fever	1
infection	1
rash	1
Total	**5**

Source: own work on the basis of research material forming the empirical part of a PhD thesis

At the ophthalmology ward during the analysis of 98 case histories, 14 medical events were identified [Tab. 2].

At the intensive care ward after an analysis of 47 case histories, 62 medical events were identified. It should be noted that with one patient multiple events may occur simultaneously [Tab. 3].

Tab. 2. Medical events at the ophthalmology ward

Type of medical event	Number of events
inflammatory reaction after intravenous insertion of a needle	4
eyelid oedema	1
inflammatory reaction in the vicinity of the eye	1
vomiting	5
raised temperature	1
infection of upper respiratory tract	1
allergic rash in the vicinity of wrists	1
Total	**14**

Source: own work on the basis of research material forming the empirical part of a PhD thesis

Tab. 3. Medical events at the intensive care ward

Type of medical event	Number of events
bradycardia	14
apnoea episodes	7
enlarged, distended stomach	5
retention of mucus in the respiratory tract	5
vomiting	4
tachycardia	4
green stool	2
blood leaks, erosion around the probe pipe	2
skin allergy after giving paracetamol	1
desaturation at the pressure of: 115/80	1
a lot of thick secretion in the intubation tube	2
suppurative discharge from the wound	1
hiccups	1
metabolic acidosis	1
presence of fluid in the pleura	1
difficulty in wound healing	1
serum and blood contents of the wound	3
large amount of partially digested food discharges through a fistula	1
probe in the duodenum and in the stomach drains brownish-bloody contents	1
rash over the entire body	1
allergic rash, mainly on the face	1
serum and blood contents of the drainage tube	1
redness and oedema at the needle insertion site	1
dysplastic and inflammatory changes of the lungs	1
Total	**62**

Source: own work on the basis of research material forming the empirical part of a PhD thesis

Discussions and conclusions

The actions described above were indented to identify adverse events, in order to enable further analysis using the black spots method. The initial isolation and grouping of the events, after the analysis of documentation from three wards of a Lodz clinical hospital, it shows that from 81 events, which could have an adverse influence on the course of patient's hospitalisation, the lengthening of the hospital stay, worsening of health, 76.5% of all events are incidents which occurred at the intensive care ward (62). It is related mainly to the profile of the ward (i.e. a higher probability of hospital infections, complications, more specialised procedures), patient's health (resistance), longer hospitalisation stay (average duration of patient's stay at the intensive care ward – 9.8 day, at cardiology – 3.8 day, ophthalmology – 2.7). The most frequently occurring medical events included: bradycardia (14 times – the probable reason may be: congenital heart defect, drugs provided, heart surgery complications, electrolyte imbalance and other), apnoea episodes (7) and vomiting (5).

When constructing risk management and hospital safety improvement programmes (not only for patients) using the black spots method, it is possible not only to lower the costs borne by a given medical entity, but what's most important, to reduce the amount of deaths [6–7].

The method described above has high efficiency, on the condition, that the remedial procedures will be created on the basis of specific medical events existing at individual wards – the individual approach enables to increase the effectiveness of the described method.

Effective identification, classification and assessment of adverse event sources, establishing their range and influence on the conducted activity, and then undertaking actions to minimise their influence reduces the level of risk at the institution.

Every medical entity should introduce a risk management system or at least should use effective tools enabling the minimisation of the amount of adverse events, including medical ones, not only to improve patient's safety but to reduce the costs or improve logistics. The main goal should be to reduce the amount of complications and patient mortality [1]. When creating safety improvement and risk management programmes in medical entities the specifics of a given institution and ward should be taken into account and a plan of action should be created, adapted to its needs and possibilities.

It should be remembered that the entire personnel working at the facility should be involved in this process. The ability to reduce and manage the risk will reduce the amount of adverse risks, improve patient's treatment quality and even reduce the costs of the hospital.

REFERENCES

[1] Adamska-Golińska N., We may prevent thousands of deaths, Menedżer Zdrowia, 1, 2011.

[2] Adamska-Golińska N., Topography of black spots in hospitals, Termedia, 06.02.2011, http://www.termedia.pl/Topografia_czarnych_punktow_szpitalnictwa-2729 (accessed on: 29.05.2012).

[3] Kaustch M., Whitfield M., Surowiec J., Quality management, in: Kautsch M., Whitfield M., Klich J. (ed.), Management in health care, Wydawnictwo Uniwersytetu Jagiellońskiego, pp. 342, Kraków, 2001.

[4] Marczak M. (ed.), Risk management in the health care system – methodology and chosen examples, Łódź, 2008. T.

[5] Marczak M., General methods of risk management, Risk Management in Health Care System – Methodology and Chosen Examples, Edited by: Marczak M., Łódź, 2008.

[6] Piechota M., Influence of operative care on mortality of patients treated on general surgery wards, habilitation thesis, Łódź, 2010.

[7] Sierocka A., Adverse event risk management, using an example of one of the clinical hospitals in Łódź, PhD thesis, Health Care Policy Department of the Medical University in Łódz, 2010.

[8] Sierocka A., Management of hospital infections risk, Risk Management in Health Care System – Methodology and Chosen Examples, Edited by: Marczak M., Łódź, 2008.

[9] Staniec I., Zawiła-Niedźwiecki J., Surgical risk management, Wyd. C.H. Beck, Warszawa, 2008.

[10] Staszewska A., Ilness and health loss risk management for patients, as assessed by the District Specialised Neuropsychiatric Care Unit personnel, graduate thesis under the supervision of prof. Michał Marczak, PhD, Medical University in Łódź, 2009.

[11] Szumlicz T., Social aspects of the insurance market development, Oficyna wydawnicza SGH, Warszawa, 2010.

The results of teaching of subject "Obstetrics, gynecology and gynecological and obstetric nursing" with the use of e-learning platform at the Faculty of Health Sciences, Medical University of Bialystok in 2009–2012

Wiesław Półjanowicz[1], **Robert Latosiewicz**[2], **Sławomir J. Terlikowski**[3]

[1] Department of Applied Informatics in Education, Institute of Informatics, University of Bialystok, Poland

[2] Department of Rehabilitation and Physiotherapy, Medical University of Lublin, Poland

[3] Department of Obstetrics, Gynecology and Maternity Care, Medical University of Bialystok, Poland

Abstract. The aim of the study was to compare the method of distance education (b-learning) with the traditional method of teaching of the subject "Obstetrics, gynecology and gynecological and obstetric nursing" held at the Medical University of Bialystok. The study was conducted among 220 third-year students of bachelor level in three academic years 2009–2012. The research group (115 people) participated in e-learning course, while other 105 students participated in a traditional manner. For distance education an LMS/LCMS-Moodle class system was implemented. The effectiveness of both forms of training was compared (including assessment of control tests, assessment of practical exercises and the final exams results). After completing a series of lectures anonymous survey was carried out in both groups. The questionnaire included questions about the organization of classes, learning effectiveness, student satisfaction with the activities carried out and the interest of students in distance learning. Mean ratings of practical classes held after the series of lectures were almost the same in both groups (4.59 in the e-learning group and 4.56 in traditional group). The average final exam grades were 3.55 ± 0.50 in the group of distance learning and 3.49 ± 0.54 in the conventional group. 93% students of the e-learning group and 83% students of conventional group positively rated the organization of classes. A high percentage (98%) of positive feedback about the classes conducted distantly and in traditional manner (86%) suggests a high level of technical content and preparation of both of these forms of activity. Based on the survey it can be concluded that both forms of education are equally effective.

Introduction

In the contemporary information society the frequency and variety of Internet activity creates new forms of contacts between the teacher and the student. The traditional methods of teaching are in many cases replaced or supplemented by distant methods. They are increasingly used in education

ISBN 978–83–7431–336–0 ISSN 0860-150X

at different educational levels. Distant learning is gaining more recognition in this process and thus is more and more immersed in our lives. At present, students are accustomed to unlimited access to the information on the Internet. Teachers are also looking for new forms of knowledge transfer and new forms of evaluation. The contents of the lectures are enriched by teachers with elements such as interactive multimedia presentations, audio-video clips and computer animation [1–3, 7, 16].

Educational courses related to the medical disciplines, such as nursing, require students to master large amount of medical knowledge. The acquisition becomes much more effective if the knowledge is transmitted by means of modern forms and media in a wide spectrum of time. The traditional form of teaching does not fulfill such conditions as it is usually limited to specified time unit held in the classroom. On the contrary, e-learning platform usually does not put the time and space constraints in the delivery and the use of the knowledge content [8, 12, 14].

However, despite facilities posed by distance learning (opportunity to learn at any place and any time), it will never "crowd out" traditional teaching. Personal contacts between the teacher and the student are, in fact, priceless. A proposal for a combination of both forms of learning is mixed learning (*blended-learning*, *b-learning*), which offers great opportunities for both students and teachers [3, 7–8, 17]. This method produces good results for those who encounter difficulties with understanding of the learning material. With mixed-mode learning, the student has the opportunity to repeat and consolidate multiple individual issues that are implementing in the e-learning platform [11, 17]. One of the advantages of this method of teaching is also relatively "mild" timing schedule – studying of the individual modules can be done at a convenient time for the student.

Aim of the study

The aim of this study was to evaluate the effectiveness of teaching the subject "Obstetrics, gynecology and gynecological and obstetric nursing" to students of the Medical University of Bialystok. The research was approved by the Bioethics Committee of the Medical University of Bialystok (Resolution No. R-I-002/338/2009). The subject was conducted as an experiment during the period of academic years 2009–2012 in the system of complementary learning (blended learning). The lectures were conducted in two manners: as distant (on-line) learning and in a traditional way (*ex cathedra*). Other forms of education, eg. seminars, practical exercises and self-study, were carried out using the traditional method only. The choice of form of lectures (e-learning or traditional) was left to the students (the primary condition was access to the Internet). Students, who had chosen e-learning methods, were given free access to the educational platform on which the lectures were placed. These lectures were prepared in the form of interactive presentations embedded in the Moodle system. It was usually the "lesson" supported by video showing the elements of practical activities in the area of the thematic issue. The sequence of topics and period of their availability were set by the academic teacher responsible for the conducted subject. Within the period set, students had continuous and unlimited access to learning materials. Each individual lecture was accompanied by a set of tests which helped the acquisition of knowledge. While solving tests students had no access to the learning materials placed on the virtual platform. Immediately after completing the tests students received the results, which served for better understanding of the content of the lectures.

Final examinations were held in the "traditional form" according to the principles set out in Regulations of Full-time Study at the Medical University of Bialystok [13].

Material and methods

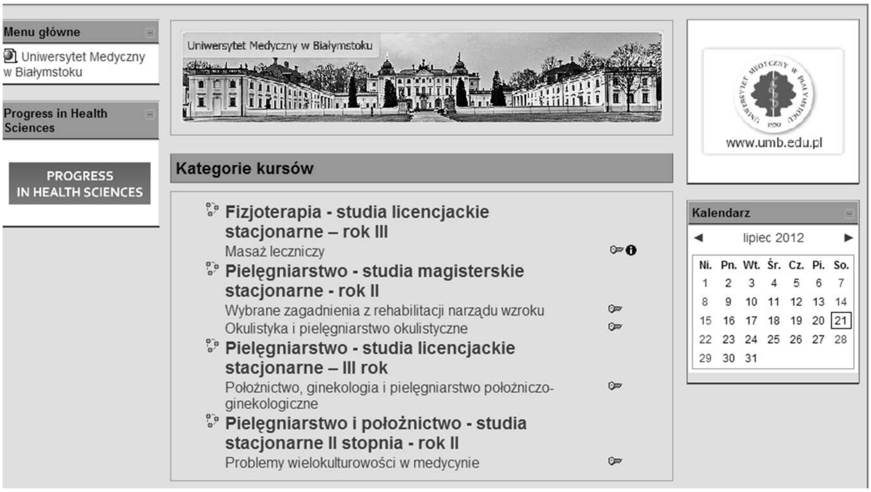
The study was conducted among 220 third-year students of bachelor level in three academic years 2009–2012. The number of participating students in the consecutive years is shown in [Tab. 1].

Tab. 1. Number of students participating in the learning of the subject "Obstetrics, gynecology and gynecological and obstetric nursing" in the years 2009–2012

Academic year	No. of students	E-learning method		Traditional method	
2009–2010	83	43	52%	40	48%
2010–2011	48	25	52%	23	48%
2011–2012	89	47	53%	42	47%
Total	**220**	**115**	**52%**	**105**	**48%**

own source

In the study group of 115 people (52%) lectures were held in the form of e-learning. Specially implemented system LMS/LCMS-Moodle class was used [8, 15] [Fig. 1]. In the control group of 105 students (48%) lectures were held in the traditional manner.


Menu główne
Uniwersytet Medyczny w Białymstoku
Progress in Health Sciences
PROGRESS IN HEALTH SCIENCES
Uniwersytet Medyczny w Białymstoku
www.umb.edu.pl

Kategorie kursów

Fizjoterapia - studia licencjackie stacjonarne – rok III
Masaż leczniczy

Pielęgniarstwo - studia magisterskie stacjonarne - rok II
Wybrane zagadnienia z rehabilitacji narządu wzroku
Okulistyka i pielęgniarstwo okulistyczne

Pielęgniarstwo - studia licencjackie stacjonarne – III rok
Położnictwo, ginekologia i pielęgniarstwo położniczo-ginekologiczne

Pielęgniarstwo i położnictwo - studia stacjonarne II stopnia - rok II
Problemy wielokulturowości w medycynie

Kalendarz

◄ lipiec 2012 ►

Ni.	Pn.	Wt.	Śr.	Cz.	Pi.	So.
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

Fig. 1. The launch window of e-learning platform Moodle (http://elearning-umb.pl)
own source

In the e-learning group students tested their knowledge by solving tests of multiple choice addressing the range of the materials contained in the e-learning platform. The results were recorded in the log and included in the rating system. Students were able to check the results of their educational activities. In the control group (traditional learning) tests were conducted at the end of the lecture series. The results of tests in both groups were expressed as percent. After completing lectures (in chosen form) all of the students had practical exercises. The results were scored in points with maximum value of 12. For all the students of the entire year final examination was held at the same time in the form of a traditional multiple choice test. The results of the examination were expressed in grades from 2 to 5 with 0.5 increment (where 5 means very good).

After completing the full series of lectures the survey of own authorship was conducted in both groups. The questionnaire contained three parts: sociodemographic data, opinions about the activities conducted and opinions about the effectiveness of education, level of student satisfaction with the activities carried out, the interest of students using distance learning and the organization of classes in both forms of education.

Results

The study was conducted on a relatively equinumerous groups of students: e-learning was attended by 115 students (52%), while the traditional

Tab. 2. Ways of using of internet in groups of students in the years 2009–2012 (n = 220)

	E-learning method (n = 115)		Traditional method (n = 105)	
	No. of persons	%	No. of persons	%
Domicile				
city above 80 th. inh.	53	46%	44	42%
city under 80 th. inh.	28	24%	21	20%
rural area	34	30%	40	38%
Permanent access to Internet				
yes	110	96%	92	88%
no	5	4%	13	12%
How often (regularly) do you use the Internet?				
everyday	90	78%	74	70%
almost everyday	17	15%	18	17%
regulary, 2–3 times a week	8	7%	4	4%
once a week	0	0%	2	2%
once a month	0	0%	0	0%
several times a month	0	0%	5	5%
occasionally	0	0%	2	2%
do not use	0	0%	0	0%
For what purposes you primarily use the Internet?				
web browsing	102	89%	89	85%
e-mailing	93	81%	78	74%
downloading files	57	50%	26	25%
chat, SMS sending	42	37%	28	27%
e-learning	52	45%	1	1%
on-line shopping	29	25%	20	19%
e-banking	41	36%	26	25%

own source

method by 105 students (48%). Both groups were dominated by students from cities of more than 80 thousand inhabitants – 53 (46%) and 44 (42%) respectively. Permanent access to the Internet had 110 students (96%) in e-learning group and 92 people (88%) in traditional group. Daily use of the Internet declared more than 70% of students in both groups, while occasional use declared only two students (2%) in the traditional group [Tab. 2]. The main activity of the students on the web was to browse web-sites: 102 students (89%) in the e-learning group and 89 students (85%) in traditional group, and received e-mails: 93 (81%) and 78 (74%) students respectively. Students devote the least amount of time for shopping on-line: 29 persons (25%) in the study group and 20 (19%) in the control group.

In order to compare the effectiveness of teaching the average results obtained by students from both groups after a series of lectures, after completion of practical training and the final examination of the evaluated subject were compared. Students the of e-learning method scored higher in testes evaluating knowledge contained in lectures. The average score was 92% as compared to 77% in the traditional method of teaching [Tab. 3]. One of the reasons could be the fact that the materials of the lectures was available in e-learning platform for a longer period of time (2–3 days) and could be accessed several times. Traditional group had solely own notes made in the course of a lecture held in the auditorium.

Tab. 3. Results of tests lectures of the subject "Obstetrics, gynecology and gynecological and obstetric nursing" in the academic years 2009–2012

Academic year	E-learning method		Traditional method	
	No. of persons	Score in percent	No. of persons	Score in percent
2009–2010	43	91%	40	85%
2010–2011	25	94%	23	87%
2011–2012	47	91%	42	60%
Mean		**92%**		**77%**

own source

The next issue was to check how the material acquired by students during the course of lectures in both forms (e-learning and traditional) was transferred into the knowledge used in practice. Credits (points) earned by students during the practical classes conducted in clinical environment (at bed-side) were compared. Observations made during the three years show that the average assessment of practical skills in e-learning group was 4.59 and in traditional group was 4.56 (max. 5.0) This seems to speak in favor of e-learning methods [Tab. 4]. Still, it must be emphatically stated that both forms of learning rate good.

Average rating of final examination for three consecutive years of teaching was 3.55±0.50 in the group with the method off distance learning, and 3.49±0.54 in the group with the traditional method of teaching [Tab. 5].

Tab. 4. Results of practical training of the subject "Obstetrics, gynecology and gynecological and obstetric nursing" in the academic years 2009–2012 (n = 220)

Academic year	E-learning method		Traditional method	
	score (max. 5.0)	points (max. 12.0)	score (max. 5.0)	points (max. 12.0)
2009–2010	4.61 ± 0.45	10.65 ± 1.4	4.65 ± 0.33	10.75 ± 1.1
2010–2011	4.54 ± 0.42	10.44 ± 1.3	4.63 ± 0.45	10.78 ± 1.4
2011–2012	4.60 ± 0.60	10.35 ± 1.5	4.45 ± 0.64	9.71 ± 1.8
Mean ± SD	**4.59 ± 0.51**	**10.49 ± 1.4**	**4.56 ± 0.51**	**10.36 ± 1.5**

own source

Tab. 5. The results of the final examination of the subject, "Obstetrics, gynecology and gynecological and obstetric nursing" in the academic years 2009–2012 (n = 220)

Academic year	E-learning method		Traditional method	
	No. of persons	Score in degrees (max. 5.0)	No. of persons	Score in degrees (max. 5.0)
2009–2010	43	3.49 ± 0.59	40	3.26 ± 0.66
2010–2011	25	3.74 ± 0.39	23	3.72 ± 0.39
2011–2012	47	3.43 ± 0.51	42	3.49 ± 0.57
Mean ± SD		**3.55 ± 0.50**		**3.49 ± 0.54**

own source

In the end-of-class survey students answered the questions which concerned, inter alia:

– review of the organization of classes carried out distantly and traditionally [Tab. 6], the availability and suitability of these classes [Tab. 7] and the possibility of acquiring new knowledge [Tab. 8],

– interest in participating in the method of classes conducted distantly during the study [Tab. 9], and hypothetically, upon completion [Tab. 10].

Organization of classes was positively rated by 106 respondents (93%) in the e-learning group and 86 students (83%) in the traditional group. In both groups 4 students (4%) negatively referred to the organization of both forms of classes [Tab. 6]. Strongly negative opinion about the classes in the group of traditional teaching expressed one person only.

Tab. 6. The reviews on the organization of distant and traditional classes in 2009–2012 (n = 220)

"Do classes were organized well?"				
	e-learning group (n = 115)		traditional group (n = 105)	
definitely no	0	0%	1	1%
probably not	4	4%	4	4%
I do not know (have no opinion)	5	4%	13	12%
rather	72	63%	58	55%
definitely yes	34	30%	29	28%
Total	**115**	**100%**	**105**	**100%**

own source

The question "Was the content of teaching the classes were well prepared, available and useful?" positively commented 103 respondents (98%) of the e-learning form and 88 students (83%) of the traditional way of teaching [Tab. 7]. Opposite view had one students in the e-learning group and 8 students (8%) of the traditional group.

Tab. 7. The reviews on usefulness of classes in 2009–2012 (n = 220)

"Was the content of teaching the classes were well prepared?"				
	e-learning group (n = 115)		traditional group (n = 105)	
definitely no	0	0%	0	0%
probably not	1	1%	8	8%
I do not know (have no opinion)	1	1%	9	9%
rather	70	61%	55	52%
definitely yes	43	37%	33	31%
Total	**115**	**100%**	**105**	**100%**

own source

The respondents were also asked the following question: "Does realization of the lessons enriched your knowledge and skills?". 111 respondents (97%) of the e-learning and 92 traditional group (88%) gave a positive answer. Answers "definitively no" was given by one student of the an e-learning group and 5 patients (5%) of the traditional [Tab. 8].

Tab. 8. Reviews on the degree of acquisition of knowledge and skills in the prepared course in 2009–2012 (n = 220)

"Does realization of the lessons enriched your knowledge and skills?"				
	e-learning group (n = 115)		traditional group (n = 105)	
Definitely no	0	0%	0	0%
Probably not	1	1%	5	5%
I do not know (have no opinion)	3	3%	8	8%
Rather	59	51%	56	53%
Definitely yes	52	45%	36	34%
Total	**115**	**100%**	**105**	**100%**

own source

During the 3-year follow-up e-learning training method was accepted by 111 students (96%) of the e-learning and 43 of the traditional group (43%). Opposite view presented 3 students (3%) of the e-learning group. As many as 27 students (26%) of the traditional methods did not accept e-learning, including one student who strongly opposed it. Summary of the results is presented in [Tab. 9].

Tab. 9. Reviews of willingness to use the on-line courses during the study in 2009–2012 (n = 220)

If there was a possibility of using on-line classes during studies, would you use this form of education?				
	e-learning group (n = 115)		traditional group (n = 105)	
definitely no	0	0%	1	1%
probably not	3	3%	26	25%
I do not know (have no opinion)	1	1%	35	33%
rather	45	39%	37	35%
definitely yes	66	57%	6	6%
Total	**115**	**100%**	**105**	**100%**

own source

The vast majority, e.g. 99 respondents (86%) in the e-learning positively responded to the possibility of extending education distantly (on-line) after completion of studies. 64 students (61%) participating in the traditional form of education have expressed their interest in this form of education in the future. Almost 1/3 of the students in this group (29%) had no opinion [Tab. 10].

Tab. 10. Reviews on the possibility of using on-line courses after graduation (at work)

If there was a possibility of using on-line classes after studies (eg. at work), would you use this form of education?				
	e-learning group (n = 115)		traditional group (n = 105)	
definitely no	0	0%	1	1%
probably not	1	1%	10	10%
I do not know (have no opinion)	15	13%	30	29%
rather	55	48%	52	50%
definitely yes	44	38%	12	11%
Total	**115**	**100%**	**105**	**100%**

own source

Discussion

Distance learning in higher education is a more and more challenging problem. It should be widely discussed especially in a view of possible extension of e-learning opportunities for a wider range of students studying different fields of medicine. Such studies are of particular intensity and different types of classes (lectures, seminars, exercises, practical work, professional practice). Increased popularity of the e-learning form in the field of medicine may result from interactive access to the knowledge contained in a course on-line. Various forms of activity (interactive lessons, quiz, chat, forums, etc.) are different from classical ones. This helps the acquisition of knowledge because students can repeatedly track previous issues and problems, which helps in better understanding their different aspects. Students are also able to instantly check their knowledge in a given area [6, 12, 14].

Recently, Kalinowska-Przybyłko *et al.* reported the results of research on online education conducted among the students of the Medical University of Warsaw, Faculty of Health Sciences, majoring in Obstetrics [9]. The results confirmed high scores of on-line classes used in the teaching process. Students participating in the remote learning viewed this form of teaching as very interesting. The authors concluded that the effectiveness of teaching on-line had a significant impact on the use of various forms of multimedia technology, as well as the length of the seminar. They also noted that e-learning is equally or more effective than traditional teaching.

The results of this study support our views resulting from our previous studies [12–14]. The development of distant education can be a very good form of supplement to traditional education. In some forms of activities, such as lectures or seminars, it can successfully replace traditional classes. The results also confirm the assumption that e-learning methods are as good as traditional method regarding the effectiveness of teaching and student's satisfaction. The last one can be in part an effect of high level of technical content and preparation of lessons.

It seems that the only matter of time is the change of the role of the university teacher who runs the system of on-line learning. Using previously prepared tests he/she is able to conveniently check the knowledge of students. Both the lecturer and the student can "see" the results immediately after completion [8, 15]. This increases the efficiency of education, and the costs of preparing and implementing activities in the form of e-learning are large only in the initial stage of the development of the teaching materials, but not much more than traditional [4–5, 10].

Conclusions

1. Distant education was seen by students as slightly easier in the course of the learning process due to continuous access to the educational materials.

2. Properly prepared teaching materials placed on virtual platform increases the opportunity to prepare students for the credits and final examinations in the respective field of knowledge.

3. The organization of classes in both forms (e-learning and traditional) was highly rated by students participating in the study.

4. Vast majority of students who have undergone the process of distant education expressed desire to continue this form of learning in the future.

5. The obtained results allow to conclude that the e-learning method can be judged as equivalent to the "traditional" method of teaching of the subject "Obstetrics, gynecology and gynecological and obstetric nursing" at the Medical University of Bialystok.

REFERENCES

[1] Aggarwal A. K., Adlakha V. G., Quality management applied to web-based courses, Total Quality Management, 17, pp. 1–19, 2006.

[2] Allan B., Time to learn? E-learners' experiences of time in virtual learning communities, Management Learning, 38, pp. 557–572, 2007.

[3] Arbaugh J. B., Desai A., Rau B., Sridhar B. S., A review of research on online and blended learning in the management disciplines: 1994–2009, Organization Management Journal, 7, pp. 39–55, 2010.

[4] Bramley P., Ocena efektywności szkoleń, Dom Wydawniczy ABC, Kraków, 2001.

[5] Dąbrowski M., Analiza pomiaru efektywności kosztowej procesów e-learningowych, e-mentor, 5, pp. 18–26, 2008.

[6] Douglas D. E., van der Vyver G., Effectiveness of e-learning course materials for learning database management systems: an experimental investigation, Journal of Computer Information Systems, 41–48, Summer 2004.

[7] Hornik S., Sanders C.S., Li Y., et al., The impact of paradigm development and course level on performance in technology-mediated learning environments, Informing Science, 11, pp. 35–57, 2008.

[8] Hyla M., Przewodnik po e-learningu, Oficyna Ekonomiczna, pp. 217–222, Kraków, 2005.

[9] Kalinowska-Przybyłko A., Kowalczyk-Nowakowska J., Baranowska B., Dmoch--Gajzlerska E., On-line seminars in the education of Warsaw Medical University students, Progress in Health Sciences, 2, pp. 101–106, 2012.

[10] Kirkpatrick D. L., Ocena efektywności szkoleń, Wydawnictwo Studia Emka, Warszawa, 2001.

[11] Półjanowicz W., Citko U., Wykorzystanie b–learningu w kształceniu studentów informatyki Uniwersytetu w Białymstoku, Fenomen Internetu, pp. 568–574, Szczecin, 2008.

[12] Półjanowicz W., Latosiewicz R., Niewiński A., Milewski R., E-learning in students education in Medical University of Bialystok, Bio-Algorithms and Med-Systems, Medical College, Jagiellonian University, 5, pp. 111–115, 2009.

[13] Półjanowicz W., Latosiewicz R., Kulesza-Brończyk B., et al., The effectiveness of education with the use of e-learning platform at the Faculty of Health Sciences, Medical University of Bialystok, Studies in Logic, Grammar and Rhetoric, 25, pp. 159–172, 2011.

[14] Półjanowicz W., Latosiewicz R., Kulesza-Brończyk B., et al., 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, pp. 94–104, Bydgoszcz, 2010.

[15] Rice W.H. IV, Tworzenie serwisów e-learningowych z Moodle 1.9., Helion, pp. 155–272, Gliwice, 2010.

[16] Smith G. G., Heindel A. J., Torres-Ayala A. T., E-learning commodity or community: Disciplinary differences between online courses, The Internet and Higher Education, 11, pp. 152–159, 2008.

[17] Wu Wen-Chieh C., Hwang Lan Yin., The effectiveness of e-learng for blended courses in colleges: a multi-level empirical study, International Journal of Electronic Business Management, 8, pp. 301–310, 2010.