LOGICAL, STATISTICAL AND COMPUTER METHODS IN MEDICINE

Series: STUDIES IN LOGIC, GRAMMAR AND RHETORIC 21(34)

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

LOGICAL, STATISTICAL AND COMPUTER METHODS IN MEDICINE

edited by
Robert Milewski
Dariusz Surowik

University of Białystok
Białystok 2010

Refereed by Edward Oczeretko, Kazimierz Trzęsicki, Marcin Koszowy

Series: STUDIES IN LOGIC, GRAMMAR AND RHETORIC 21(34)
http://logika.uwb.edu.pl/studies/

Edited by Halina Święczkowska
University of Białystok, Faculty of Law, Section of Semiotics

in collaboration with Kazimierz Trzęsicki
University of Białystok, Faculty of Mathematics and Physics
Chair of Logic, Informatics and Philosophy of Science – logika@uwb.edu.pl

This issue has been created in collaboration with Medical University of Białystok

Guest-Editors:

Robert Milewski
Medical University of Białystok

Dariusz Surowik
University of Białystok

Editorial Assistants:

Katarzyna Doliwa
University of Białystok

Dariusz Surowik
University of Białystok

This issue has been financed by the Medical University of Białystok

Cover design: Krzysztof Tur
Type-setting: Stanisław Żukowski

ISBN 978–83–7431–266–0
ISSN 0860–150X

WYDAWNICTWO UNIWERSYTETU W BIAŁYMSTOKU
15-097 Białystok, ul. Marii Skłodowskiej-Curie 14, tel. 0857457059
http://wydawnictwo.uwb.edu.pl, e-mail: ac-dw@uwb.edu.pl

Druk i oprawa: PPHU TOTEM s.c., Inowrocław

CONTENTS

Dorota Jankowska
Anna Justyna Milewska
Urszula Górska
Department of Statistics and Medical Informatics,
Medical University of Bialystok

APPLICATIONS OF LOGIC IN MEDICINE

Abstract: This paper presents the importance of logic in the medical field. Efficient and proper medical work is difficult without the knowledge of the rules of logic. Therefore, the paper will consider ways of implementing both classical logic and non-classical approach, e.g. temporal and fuzzy logic. The thesis will be supported by numerous examples illustrating how indispensable is the cognition of logic and showing how applying logic can effectively improve work in medicine.

Introduction

An essential element of the medical profession is making numerous decisions. In this process doctors rely on gained knowledge and experience. However, it seems necessary for them to have the ability to think logically, to use reasoning, to infer, to precisely and clearly express their thoughts and justify the assertions made. Even when their actions are be based on certain algorithms or standards, they have to logically model the situation. Lack of knowledge concerning the rules of logic can lead to dangerous errors and may result in continuous failures in performance flowing from faulty reasoning processes. This paper will present the importance of logic in the medical field. The precursors of this view were Tytus Chałubiński and Edmund Biernacki in the second half of the 19^{th} century [1]. However, the implementation of logic in medical sciences is ascribed to the Polish general practitioner and philosopher Władysław Biegański (1857–1917). He was the author of an excellent work titled "Logika medycyny". His achievements have been used by equally distinguished professor of history and philosophy of Władysław Szumowski, who lived in the years 1876–1954. W. Szumowski was the autor of such works as "Logika dla medyków", "Filozofia medycyny"

ISBN 978–83–7431–266–0 ISSN 0860-150X

and "Historia medycyny filozoficznie ujęta". He deepened the issues raised in the work of W. Biegański putting much more emphasis on the practical approach. Moreover, he discussed the most logical errors in medicine.

Classical Logic

Logic is the science of reasoning, it deals with the formulation of general rules that one may properly carry out by conducting proof, checking or inferring. Its effects are visible in everyday situations. Without logic it is difficult to imagine efficient functioning of the world of medicine. A doctor who is obliged to take decisions arising from many different factors should control his/her actions in accordance with the principles of logical reasoning. Classical logic provides a variety of tools making it possible. The most useful for non mathematical sciences is inference. It consists of identifying the implications arising from available indications and justification of new knowledge on the basic of knowledge already available [2].

There are two types of inferences – certain and uncertain [3]. Among certain inferences are mainly deductive inference. According to it, true reasons always lead to true, uncontested conclusions [4]. Moreover, schemas and relationships, on which it is based, are logical rules. Formerly deductive inference was defined only as a transition from general to specific [5]. The example of such reasoning may be presented as follows:

	All asthmatics are allergic.
	John has asthma.
Therefore:	*John is allergic.*

However, now it is considered that such approach is too narrow [5]. Among the inferences the whole reasoning is contained, such that from the conditions a conclusion results. For example, the entailment which is based on the so-called law of detachment modus ponendo ponens $\frac{\alpha \to \beta, \alpha}{\beta}$ may look as follows:

As per [6]:	*Cervical cancer is the only human organ cancer which if detected at an early stage can be 100% cured.*
	In a patient cervical cancer was detected early in the form of the so-called preinvasive cancer (0 degree).
Therefore:	*This cancer will be cured completely.*

An example of inference based on hypothetical syllogism rule $\frac{\alpha \to \beta, \beta \to \gamma}{\alpha \to \gamma}$, which is built according to the law of transitivity, may have the following form:

	If during the bacterial infection the temperature of the patient's body rises, then the number of white blood cells in his/her blood will considerably increase.
	Increased leucocyte count causes much faster absorption of bacteria by phagocytosis.
Therefore:	*Increased temperature in the course of infection is the immune response of the body and prevention of further development of the disease through phagocytosis.*

As an example the inferences based on modus tollendo tollens rule $\frac{\alpha \to \beta, \neg\beta}{\neg\alpha}$, which is built on the law of contraposition may be considered as follows:

In accordance with [7]:	*If the patient has been infected with varicella virus causing chickenpox then in the period up to three weeks on his/her body will appear blisters surrounded by red border.*
	The doctor suspects that the patient who hasn't so far suffered from chickenpox may have had contact with a person infected by varicella virus. He wants to determine whether the patient was infected. After three weeks rash was not visible on the patient's body.
Therefore:	*The patient wasn't infected with the chickenpox virus*

In opposition to the rules presented above, the uncertain inference is that the proposal does not flow logically from the premises; true premises do not prejudge the truth of the conclusion, however it allows for the acceptance of the proposal with some probability. This category includes inductive incomplete inference and reductive inference. The first type relies on drawing conclusions about totality from observations of the details. The inductive incomplete inference is when the trait which was observed in a study group is assigned to the whole community. On the other hand, the reductive inference is applied to explain the experimental facts. It is the

opposite to the direction of inference to the previously presented models – it occurs when "proposals result from premises". It is characterized by formulating the causes of some facts and finding the reasons for certain corollaries. For example, if the ECG entry examination presents changes which take the form of a ST segment elevation, forming the so-called Pardee wave, it may be concluded that the patient who has been tested had a heart attack. However, a serious problem is the fact that the evaluation of logical correctness of uncertain inferences is the topic for discussion. A consistent and widely accepted theory which would systematize such reasoning hasn't been yet developed.

The knowledge of the rules of inference is very important in medicine. It is useful both in contact with the patient and during conducting research. It allows one to logically draw final conclusions. Moreover, persons applying it are able to distinguish which of the reasoning is irrefutably correct and which is only probable. In addition, the ability of inference gives the possibility to make some generalizations, to explain the correctness of the decision and enables recognition as true beliefs. Above all, it enables posing hypotheses and thus formulation of the courts and conjecture. Their verification needs a proof, which also inevitably involves the inference. The proof of the theorem is nothing else but a demonstration that the checking argument is true using premises of theorem, axioms, principles of logic and formally carried out reasoning. Broadly speaking, we may distinguished two ways of proof – direct proof and proof by contradiction (indirect proof). Direct proof, also called classic, consist of showing a sentences as a thesis from the available, true premises and initial for the logical systems theorems (axioms) using inference rules. Whereas proof by contradiction is known as ad absurdum proof, by reduction to the absurd. It is based on contradictions between the assumptions and the negation of the thesis. It allows to conclude that if the negation of the thesis is false then that thesis is true [8, 9]. Such reasoning can be carried out in two ways. On the one hand it can take place in accordance with the rules of inference, on the other by showing a counterexample. An example of proof by contradiction may take the following form:

It will be shown that:	*If a man with haemophilia A and healthy woman, so free from hereditary genetic disorders determine by recessive allele, have a daughter then she is a carrier.*
From the human genetics rules, it is known that:	*(∗) Healthy woman is always the child of a healthy man.*

Then suppose that:	*(1) The patient is daughter of healthy woman and a man suffering from haemophilia A.*
	(2) She isn't a carrier. (It is the negation of the thesis)
From (2) under the law (∗) indicates that:	*The patient is daughter of a healthy man.*
It makes contradiction with (1).	
Therefore:	*Daughter of a man with haemophilia A and a healthy woman is a carrier.*

In addition to the previously described, undeniable benefits of using the principles of classical logic in medicine there are now widely developed ways to use also non-classical logic, such as: modal, temporal, epistemic, deontic, multi-valued and fuzzy logic. These which seem to be the most interesting will be presented below.

Temporal logic

The term "temporal logic" is wide. It is used to describe any independent of each other systems which formalize expressions containing time phrases, often in different ways [10]. This subject began to be discussed at the end of the first half of the 20^{th} century. Currently, as temporal logic we generally consider the Tense Logic. Formally it is derived from modal logic. Its creator was Arthur Norman Prior. Building systems of this logic is an admission of a specific model of physical time. In addition, it is characterized by temporal operators that define different tenses [12]:

- Hφ – it has always been the case that φ,
- Pφ – it was sometime the case that φ,
- Gφ – it will always be the case that φ,
- Fφ – it will sometime be the case that φ.

In order to formalize expressions of everyday language Chronological Logic was developed. It was founded by the Polish logician Jerzy Łoś. It contains the following binary functors.

- R t φ – φ is realized at time t,
- U $t_1 t_2$ – time t_1 is before time t_2.

In temporal logic also the von Wright's systems is included – "And Next" and "And Then". These are characterized by the conjunction of binary time operator $\varphi \mathrm{T} \psi$, defined as [11]:

- φ and in the next moment ψ – in "And Next" system (time structure is discrete here),
- φ and later ψ – in "And Then" system (this system involves only the linearity of time).

Quite recently there were formed systems which use the concept of time in computer programs. The authorship of the first of them is assigned to R. M. Burstall and A. Pnueli. They introduced functors relating only to the present and the future:

- $\Box_1\varphi$ – φ is in all conditions,
- $\Diamond_1\varphi$ – φ is in at least one condition,
- $\mathrm{X}\varphi$ – the next condition than the present it will have been φ,
- $\varphi\mathrm{U}\psi$ – at some point it will have been ψ; by that time there is φ.

Over time, this systems was joined by analogous functors describing the past:

- $\Box_2\varphi$ – φ was in all conditions,
- $\Diamond_2\varphi$ – φ was in at least one condition,
- $\mathrm{X}^{-}\varphi$ – on the previous condition than the present had been φ,
- $\varphi\mathrm{S}\psi$ – at some point it happened ψ; previously it had been φ.

The first three of the above mentioned logical systems are used in natural sciences. The last is widely used in computer sciences.

Temporal logics may also be divided taking into account the structure of the time they assume. Then, two groups are distinguished:

– Linear tense logics – depending on the systems; they assume that a preceding relation $<$ is transitive, linear on both sides, without the initial and final time [10],
– Branching tense logics (their authors were N. Rescher and A. Urquhart) assume transitivity and backwards linearity preceding relation (branching of time chain in the future) [10, 13].

The diversity of logics constituting temporal logics is due to the fact that their creation was a response to the need to be used in different fields of knowledge. The most interesting application in medicine has been the so-called bitemporal logic. It defines two types of time [14]:

- VT – the time when the data is valid, the so-called "Valid Time". It is for recording the time when an event takes place in the reality represented by the database.
- TT – the time of the transaction data, the so-called "Transaction Time". It specifies how long the data about an event are held in the database.

Such approach to time allows one to build databases extremely useful for medical purposes. These databases allow storage in addition to the current

condition of data also the information relating to both the past and the future. They show how data changes over time. Such databases prevent losing data resulting from certain modifications or upgrades and it is also useful in planning. Now let us consider an example of how the bitemporal database works.

The following Table 1. presents course treatment of patients of the hospital.

Table 1

The example of bitemporal database

	Patient ID	Medicine	VT	TT
1.	1	A	6.04–17.04	5.04–∞
2.	2	C	8.04–∞	8.04–9.04
3.	2	A	11.04–∞	10.04–∞
4.	3	B	10.04–∞	7.04–12.04
5.	3	C	13.04–19.04	13.04–∞
6.	3	A	20.04–22.04	13.04–∞
7.	4	A	17.04–∞	15.04–∞
8.	4	B	20.04–23.04	17.04–∞

The attribute VT represents the duration of validity. In this example, VT is the time when the patient was treated with specific medicine. Transaction Time (TT) is the time when the decision to conduct a specific treatment was stored in the database. This attribute gives information about when the record was inserted and whether it was deleted or not.

This bitemporal database presents numerous information. Firstly, on the 5^{th} of April the doctor using the database decided that Patient 1. will be receiving medicine A from 6 to 17 April. Symbol ∞ indicates that this decision is still in force.

Instead, the patient's 2. course of treatment is as follows: 8.04 the doctor caring for the patient made a decision about the immediate initiation of dosing medicine C. Symbol ∞ informs that he didn't specify how long this treatment will be continued. On 10.04 it turned out that the patient hadn't responded to the treatment. So the doctor decided to apply to him from 11 April medicine A. This has resulted in the relevant entry in the database. Comment requires only the value of TT attribute in row 2, which is "8.04–9.04". It is so because the concept about the start of the treatment with medicine C was made on 10.04 and the previous one, about dosage C, had been stored in the database until 9.04 (included).

As it is registered in the database, medicine B was administered to Patient 3. from 10.04. The decision about it was taken already on the 7th of April. But on 13.04 the doctor changed the treatment. He planned that the patient will be receiving medicine C from the 13 to the 19 and medicine A in the following days between the 20th and the 22th. TT attribute in rows 5 and 6 has the same value in both cases because decisions were made on 13.04, and from that day are stored in the database.

In addition, on 15th of April the doctor ordered the treatment with medicine A to Patient 4. In this case the end data of treatment hasn't been decided. The concept depended on treatment effects. Moreover on 17.04 the doctor decided additionally on using medicine B to this patient. Thus, patient 4. was treated with two medicines throughout this period.

The database created in such a way gives a wide possibility. First of all, it guarantees that the data will not be lost during update. It allows to view both the current and historical information. It makes it possible to quickly determine the status of data for a specific moment in time – both present and freely chosen [15]. For example, it can be easily seen which records were stored in the database on 9.04. [Table 2].

Table 2

State of data for the day 9.04 of TT

	Patient ID	Medicine	VT	TT
1.	1	A	6.04–20.04	**5.04**–∞
2.	2	C	8.04–∞	**8.04–9.04**
4.	3	B	10.04–∞	**7.04–12.04**

It gives information about treatments taken on 23.04. (which patient is treated and which medicine is used). It is presented in Table 3. (There are both records deleted and stored.).

Table 3

State of data for the day 23.04 of VT

	Patient ID	Medicine	VT	TT
2.	2	C	**8.04**–∞	8.04–9.04
3.	2	A	**11.04**–∞	10.04–∞
4.	3	B	**10.04**–∞	7.04–12.04
7.	4	A	**17.04**–∞	15.04–∞
8.	4	B	**20.04–23.04**	17.04–∞

Moreover, it allows for searching information about treatment taken on 17.04 contained (not deleted) in the database at 21.04. [Table 4.].

Table 4

State of data for the day 17.04 of VT and 21.04 of TT

	Patient ID	Medicine	VT	TT
1.	1	A	**6.04–17.04**	**5.04–∞**
3.	2	A	**11.04–∞**	**10.04–∞**
5.	3	C	**13.04–18.04**	**13.04–∞**
7.	4	A	**17.04–∞**	**15.04–∞**

Data collected in the database may be presented in form of a coordinate system where the axes mean VT and TT [14]. In preceding example this representation is showed on the Figure 1.

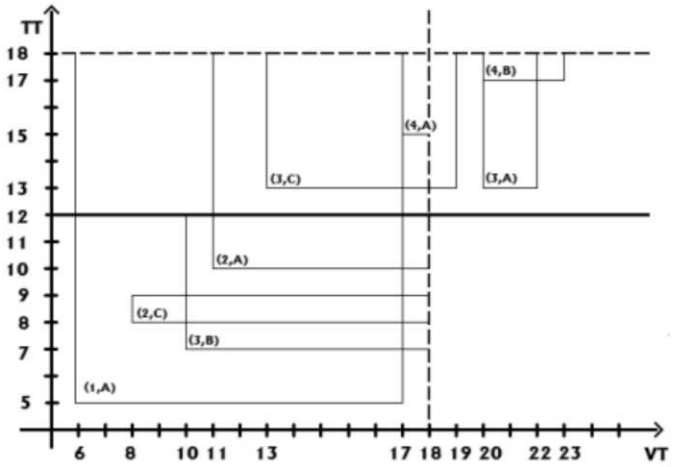


Fig. 1. Graphical representation of bitemporal database from Table 1

As it can be seen it is convenient to represent the database graphically, it gives numerous benefits. Figure 1. is easy to use. It is possible to quickly determine the state of data at any time. For example, the situation for the day 12.04 of TT is marked with a thick solid line on Figure 1. Rectangles, which are in contact with the line or intersect with the line correspond to the records stored in the database on that day. As easy as before it can be read on the Figure 1. which of the patients are treated and which medicine is administered to them on a chosen day. Moreover, assuming that today

is the 18th of April, the dashed horizontal line shows data which actually are stored in the bitemporal database and dashed vertical line represents the current events. The place where this dashed lines intersect designates information in the database which is actually taking place, also current with respect to both VT and TT.

Fuzzy logic

Classical logic assumes that each sentence is either true or false. But this standard creates a problem when describing ambiguous, inexact phenomena and formalizing the intermediate situations. For example, let us consider RBC – the index which means the number of red blood cells in blood morphology. How can it be determined whether the value of RBC, which is 6,4 mln/mm^3, is high? If the person who has been tested is an adult woman then this situation is evident – the value of RBC is high. For a 30-year-old man who every day does hard physical work it is a value classified as standard. By contrast, if the patient is an infant, it arises doubts. Then it is a problem to say what we think about it. It would be the safest to use an expression that this ratio is rather below standard. Interpretation is dependent on the specific situation and its context. Such problem was observed by Plato, who lived in the years 427–327 BC. First attempt to resolve it was taken by the Polish scientist Jan Łukasiewicz, who formed three-valued logic. In this system a value is formulated "possible" between true and false.

However, the real breakthrough was the work of Lotfi A. Zadeh entitled "Fuzzy sets" published in 1965. Here it was defined that fuzzy sets differ from classical approach, which assume that an item belongs to the set or not. They do not have sharp, clearly defined border. Each element belongs to the fuzzy set to a certain extent and their attachment may be expressed as a number in the range [0,1]. Such classification is like the human process of thinking, reasoning and interpreting occurrence. It allows an individual approach to each circumstance as well as formalized situation described in a natural language in which are the determinations of the type: little, medium, small, very, quite [18]. Such vision, although seeming to be simple, has revolutionized the approach to sets [16]. So Lotfi A. Zadeh proposed in 1973 a fuzzy logic. Some statements in this system may be false (0), true (1) or in some part true. There are concepts of half-truths, almost false and practically true and there are admissible logical values in the range [0,1]

Fuzzy sets are found whenever there is some ambiguity or subjectivity, and thus imprecisely worded conditions to belong to the set. Degree of belonging of item x to set A specifies the so-called membership function which will be recorded as $\mu_A(x)$. It is defined in an arbitrary way, but the image of function has to be from the interval [0,1] [17]. Its shape or model may be decided by expert knowledge, or a neural network.

For example, let us consider the problem presented earlier, "Is the value of RBC equal 6.4 mln/mm^3 high?". Then membership function may take the form as in Figure 2.

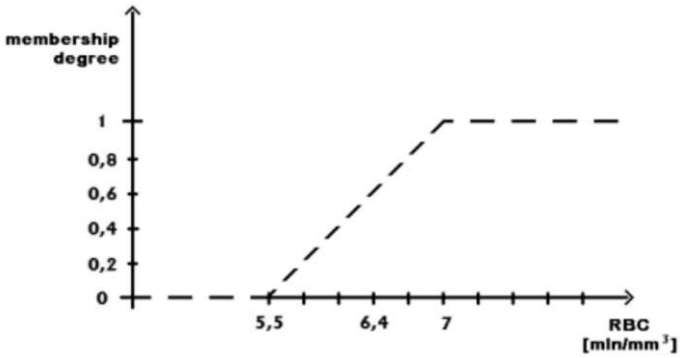


Fig. 2. The membership function of measurement results RBC belonging to a set of indicators of high

or like that:

$$\mu(x) = \begin{cases} 0 & \text{for } x \leq 5,5 \text{ mln/mm}^3 \\ \dfrac{x - 5,5}{1,5} & \text{for } 5{,}5 \text{ mln/mm}^3 < x \leq 7 \text{ mln/mm}^3 \\ 1 & \text{for } x > 7 \text{ mln/mm}^3 \end{cases}$$

Thus, in this interpretation, the number of red blood cells of 6,4 as being 0,6 high value is defined. Furthermore, the indicator which is 5,5 mln/mm^3, does not belong to the set of high values because membership degree is 0. On the other hand, each of the results of over 7 mln/mm^3 is blindly recognized for the high – membership degree is 1. The membership function gives twofold information. It designates areas where there is no doubt with assigning certain element to a particular set but also shows how to define the degree of fulfilled criteria of belonging to the set in the interval in which there is some confusion.

If the task would be to qualify the results of red blood cells measurement to one of three groups: high, normal or low, the membership function as presented in Figure 3 may be used.

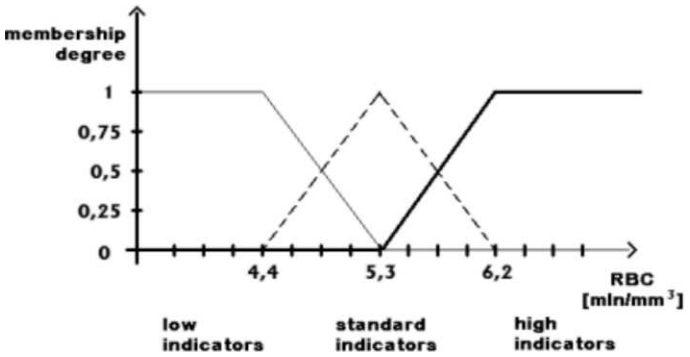


Fig. 3. Membership functions to sets of different types of RBC measurement

Then, as Figure 3. shows, the value of rate which is 5,525 mln/mm^3 belongs to high indicators in 0,25 and at the same time it belongs to the set of standard indicators in 0,75. On the other hand, it isn't contained in the set of low indicators as evidenced by the zero membership degree.

LA Zadeh also defined operations on fuzzy sets, analogously like it was done in classical logic, and the corresponding degrees of membership [18]:

- by the sum of the fuzzy sets A and B, it is meant the smallest fuzzy set containing both set A and set B. Membership degree to the set $A \cup B$ is defined as follows:

$$\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))$$

- intersection of fuzzy sets A and B is the largest fuzzy set belonging simultaneously to both sets. Membership degree to the set $A \cap B$ is modeled as:

$$\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))$$

- complement of the fuzzy set A is defined as the maximum fuzzy set of elements does not belong to set A. Membership degree to the set $\neg A$ is given by formula:

$$\mu_{\neg A}(x) = 1 - \mu_A(x)$$

Let us consider membership functions presented in Figure 3. which allocate RBC measurement into 3 categories according to the value.

– Membership degrees to the sum of set of low indicators and set of standard indicators are specified by the function from Figure 4.
It shows how much the item belongs to one of the sets – low indicators or standard indicators.

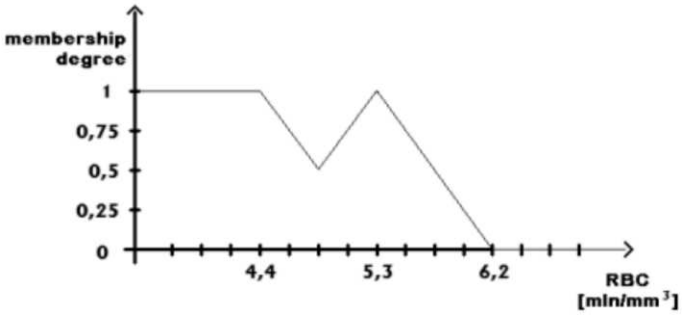


Fig. 4. The membership function of the sum of set of low indicators and set of standard indicators

– Membership degrees to intersection of sets of red blood cells measurement in standard and high are presented by the Figure 5.

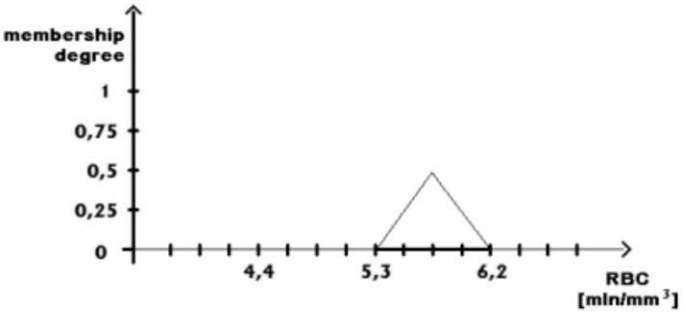


Fig. 5. The membership function of the intersection of set of standard indicators and set of high indicators

This function demonstrates how much the item belongs to both sets simultaneously.

– Membership degrees to complement of set of RBC measurement in standard are presented on Figure 6.

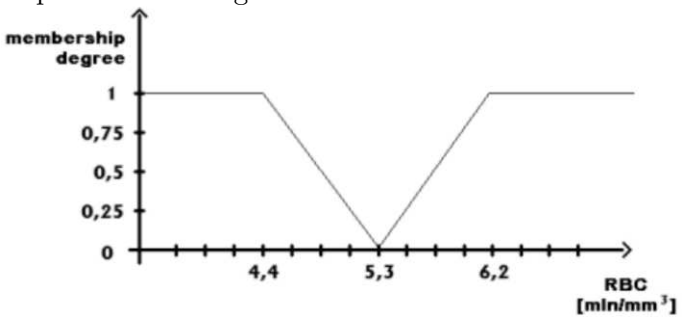


Fig. 6. The membership function of complement of set of standard indicators

It defines how much the item does not belong to the set.

Let us now imagine the situation of a doctor who has to make a decision of starting treatment with a specified antibiotic. This medicine has a number of restrictions to apply. It can be used only to patients over 12 years of age whose weight is at least 40 kg.

The doctor after conducting an appropriate interview obtains all the necessary information. Suppose that he is guided by the logic of the classical meaning. Then the decision made by him is categorical and immediate – if the patient fulfills all required conditions, the treatment is started; but if at least one of the requirements does not occur, the treatment with this antibiotic is no longer taken into account. Fuzzy logic gives the doctor many more opportunities. It allows an individual approach to each patient and it gives a chance to take a subjective decision requiring consideration of several and sometimes many factors. Moreover, it determines the degree in which the patient meets each criterion separately but also both together.

Let us suppose the membership function to set of people aged 12 years looks like it is presented in Figure 7.

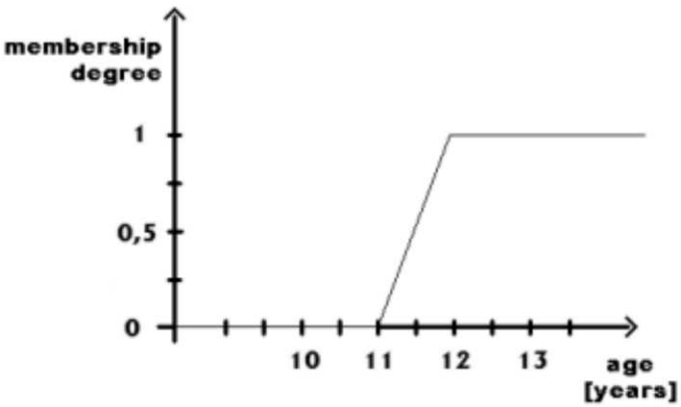


Fig. 7. The membership function to set of people aged 12 years

Furthermore, membership degree to set of people weighing at least 40 kg can be read from the graph in Figure 8. Now, if the patient is a 12 year old boy weighing 38.5 kg then, in accordance with the classical approach, one of requirements to apply specified antibiotic isn't fulfilled and the doctor admittedly rejects the possibility to treat with it. Fuzzy logic allows the doctor to determine the extent to which there are satisfied criteria previously formulated. Let A be the set of those over 12 years and B be the set of

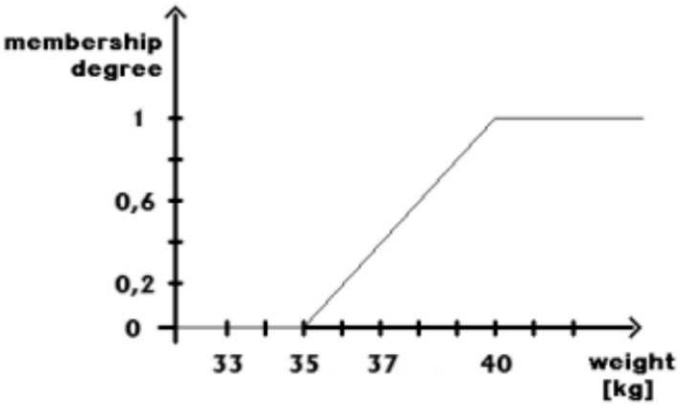


Fig. 8. The membership function to set of people weighing at least 40 kg

people who weigh is at least 40 kg. Then, in pursuance of functions shown in Figures 7 and 8 it is:

$$\mu_A(x) = 1$$
$$\mu_B(x) = 0,7$$

Moreover:

$$\mu_{A\cap B}(x) = \min(1; 0,7) = 0,7$$

It means that in the case of this patient the restriction on the use of the medicine are met in a relatively high degree of at 0.7. It allows the doctor a personal assessment whether this value is sufficient to decide about starting the treatment. Also it gives an opportunity to consider whether, in this situation, the benefits of cure outweigh the risk of possible adverse effects.

Fuzzy logic opens up the possibility for users to formalize the rules of everyday life, as described in colloquial language, which could not be expressed in the framework of the classical approach. What is more, it gives a chance of reasoning consistent with human process of thinking. It is available through the so-called fuzzy inference. This operation consists of a series of stages that can be carried out in many different ways. There are processes such as defuzzyficaton (blurring), inference, fuzzyfication (sharpening) [19, 20]. Furthermore, numerous fuzzy models were created to increase the accuracy of fuzzy reasoning or its simplification as, for example such models as: Mamdani, Takagi-Sugeno, relational, local, global, multimodels [19]. In order to use a fuzzy inference it has to be known, defined by an expert, membership functions of input values to fuzzy sets, base of rules of the form "If ... then ... " and there should be chosen the inference mechanism. An excellent way to cope with these restrictions is to use neural networks [20, 21]. They select all relevant indicators in the way of learning

"under the supervision" and they model the necessary functions or solve various problems of estimation. These are currently one of the fastest developing methods of artificial intelligence.

Fuzzy logic enables to model in a formal way the surrounding world and it is used especially to describe vague and subjective situations. It is an alternative wherever classical logic is no longer sufficient, where this logic is not able to contrive with certain ambiguities. Thus, it has an extremely wide applications. It serves mainly to choose the method of action. It is also used to solve various decision problems. There are systems based on fuzzy logic applied in economics, sociology, electronics, technology, industry, informatics, meteorology, ecology and spatial economy. They are used in refrigeration, air-conditioning equipment and car driving control systems, in image processing systems and water treatment devices, in elevators, cameras with auto focus and household appliances like kitchen dishwashers, washing machines, refrigerators, in decision-making processes relating to trading activities of enterprises, to credit risk assessment, in systems controlling the rolling stock or ventilation of underground tunnels, to solve the problem of traffic jams, and even in the production of Japanese sake alcohol. Above all, fuzzy logic has numerous medical applications, among others, in fields such as cardiology, oncology, endocrinology, pediatrics, intensive care, anesthesiology. It supports the processes of making diagnoses or determining the dose of medicine. It participates in decisions concerning treatment and arising from a number of factors, it may be used also to predict patient length of stay in hospital. It is used in many medical devices, for example, in systems controlling pacemakers or blood pressure and blood sugar, in cancer diagnostics, equipment for insulin dosage, warning systems for heart disease, osteoporosis and arthritis.

Conclusion

Learning, understanding and applying the laws of logic, both classical and non-classical, is necessary to work efficiently in the world of medicine. It allows one to effectively, consistently and without mistakes carry out all the reasoning and to prove the hypothesis in a formal way. It causes that we see the need to justify opinions through the means of argumentation. It gives certainty to the correctness of the formulated views. It informs how to recognize which of the justifications are indisputably certain and which are only to some extent likely. Thus, logic is useful both in the contact with the patient, in the process of diagnosis and therapy planning as well as in

laboratory studies. In addition, its principles are the basis of the activity numerous machinery, application, technologies, systems and computer programs which improve work. So actually, logic is applied in each aspect of medical action. Simultaneously, intensive work on developing ways to use more "alternative logics" for the purposes of the medical field is carried out. Applications of fuzzy logic and bitemporal logic presented in the paper shows areas in which it can be expected beyond these studies. It is believed that the scope of using logic in medicine will soon expand.

REFERENCES

[1] Z. Domosławski: Logika w kształceniu akademickim w koncepcji Władysława Szumowskiego – aktualność problemu na co dzień; Sztuka Leczenia, tom IX, nr 2, 2003.

[2] A. Grzegorczyk: Zarys Logiki Matematycznej, Państwowe Wydawnictwo Naukowe, Warszawa, 1975.

[3] K. Ajdukiewicz: Pragmatic Logic, Reidel, Dordrecht, 1965.

[4] Z. Ziembiński: Logika Praktyczna, Państwowe Wydawnictwo Naukowe, 2009.

[5] K. Ajdukiewicz: Zarys logiki, Polskie Zakłady Wydawnictw Szkolnych, Warszawa, 1955.

[6] Z. Słomko: Ginekologia, Państwowy Zakład Wydawnictw Lekarskich, Warszawa, 2008.

[7] R. L. Cecil, R. F. Loeb: Choroby Wewnętrzne, Państwowy Zakład Wydawnictw Lekarskich, Warszawa, 1957.

[8] H. Rasiowa: Wstęp do matematyki współczesnej, Państwowe Wydawnictwo Naukowe, Warszawa, 1975.

[9] W. Marciszewski: Logika formalna – zarys encyklopedyczny z zastosowaniem do informatyki i lingwistyki, Państwowe Wydawnictwo Naukowe, Warszawa, 1987.

[10] A. Kozanecka: O rodzajach logik temporalnych; Roczniki Filozoficzne, tom LV, nr 1, 2007.

[11] A. Kozanecka, M. Leszczyńska: O wyrażalności niektórych relacji czasowych i własności czasu w języku systemów logiki temporalnej G. H. Von Wrighta; Roczniki Filozoficzne, tom LV, nr 2, 2007.

[12] E. Hajnicz: Reprezentacja logiczna wiedzy zmieniającej się w czasie, Akademicka Oficyna Wydawnicza PLJ, Warszawa 1996.

[13] N. Rescher, A.Urquhart: Temporal Logic, New York, 1971.

[14] M. Giero, R. Milewski: Storing and Retrieving information on the treatment of infertility with the use of the bitemporal database and temporal logic, Studies in Logic, Grammar and Rhetoric, vol. 17 (30), 2009.

[15] M. Giero: Application of bitemporal databases containing medical data, Studies in Logic, Grammar and Rhetoric, vol. 17 (30), 2009.

[16] E. Januszewski: Logiczne I filozoficzne problemy związane z logika rozmytą, Roczniki Filozoficzne, tom LV, nr 1, 2007.

[17] M. Brown: An Introduction to Fuzzy and Neurofuzzy Systems, 1996.

[18] A. Łachwa: Rozmyty świat zbiorów, liczb, relacji, faktów, reguł i decyzji, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2001.

[19] A. Piegat: Modelowanie i sterowanie rozmyte, Akademicka Oficyna Wydawnicza EXIT, Warszawa, 1999.

[20] K. Rykaczewski: Systemy rozmyte i ich zastosowania, 2006.

[21] J. Łęski: Systemy neuronowo-rozmyte, Wydawnictwo WNT, 2008.

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 21 (34) 2010

Roman Matuszewski
University of Bialystok
Bialystok, Poland

Hanna Sovalat
Institut de Recherche en Hématologie
et Transplantations, Mulhouse, France

TAXONOMICAL CLASSIFICATION OF HEMATOPOIETIC CD34+ CELL SUBSETS FROM DIVERSE ORIGIN[1]

Abstract: Taxonomy is one of the numerous methods of classifying empirical data. In this work the *Wroclaw Taxonomy* has been chosen (this method has been previously implemented in the classification of bacteria *yersinia pestis*) [1, 2]. The aim is to group 5 different sources of HPC ($j = 5$ of the studied objects): *normal* and *mobilized BM*, *normal PB*, *leukapheresis products*, *cord blood*, depending on 13 common features which describe the objects (i.e. $i = 13$ measurement variables). The objects will be compared between themselves but not in reference to the optimal model.

Acknowledgements: Sections of the work by Roman Matuszewski have been supported by the Polish State Scientific Research Committee grant 3 T11F 011 30 "*Temporal representation of the knowledge and their implementation in the medical systems*". We also thank prof. Andrzej Trybulec for his valuable suggestions.

Taxonomy is the practice and science of classification. Originally, the term taxonomy referred to the science of classifying living organisms; however, the term is now applied in a wider, more general sense and now may refer to the *classification* of objects, as well as to the *principles* underlying such classification.

Taxonomies, or taxonomic schemes, are composed of taxonomic units or kinds of objects that are arranged frequently in a hierarchical structure. A taxonomy might also be a simple organization of kinds of objects into groups. Mathematically, a hierarchical taxonomy is a tree structure of classifications for a given set of objects.

The empirical data has been taken from the paper [3]. We have two series of data:

[1] This work was presented in the poster session at the International Symposium *Bioengineering and Regenerative Medicine*, Sept. 24–26, 2007, Mulhouse, France.

ISBN 978–83–7431–266–0 ISSN 0860-150X

TABLE 1.

Frequency of *early cell subsets* and *cellular adhesion molecules expression* within the CD34+ cell population derived from five sources of Hematopoietic Progenitor Cells – **HPC** (% of total CD34+ cells).
nBM = normal bone marrow; **mBM** = mobilized bone marrow;
nPB = normal peripheral blood; **LKP** = leukapheresis product;
CB = cord blood.

It creates the table $\boldsymbol{a}_{i,j}$ of 65 empirical measurements:

	j→	1	2	3	4	5
i ↓	CD34+ cell subsets	nBM (n=13) (steady state)	mBM (n=16)	nPb (n=13) (steady state)	LKP (n=29)	CB (n=20)
1	$CD38^-$	3.47± 0.06	3.51± 4.40	1.40± 1.87	2.10± 0.90	15.59± 7.84
2	$HLA\text{-}DR^-$	4.77± 1.60	2.91± 0.50	5.71± 4.15	0.80± 0.58	8.21± 3.99
3	$CD90^+$	13.28± 8.60	12.96± 7.83	ND	19.10±12.89	ND
4	$CD117^+$	11.12± 9.91	49.20± 6.20	8.12± 4.58	61.48± 8.90	16.31± 4.80
5	$PgP170^+$	2.06± 0.78	3.03± 0.20	2.33± 0.16	3.46± 1.62	5.49± 2.54
6	CD11a+	64.54+25.59	88.55+ 3.77	75.54+11.08	90.88+ 4.48	89.43+ 8.57
7	CD11b+	8.27+ 2.40	4.17+ 2.02	3.99+ 1.85	3.75+ 1.52	2.87+ 0.35
8	CD49d+	94.24+ 9.64	97.00+ 5.10	ND	68.45+ 9.96	88.74+ 7.85
9	CD49e+	48.75+ 7.88	45.70+ 5.87	ND	38.90+ 5.42	ND
10	CD54+	11.29+16.12	26.27+ 5.16	14.89+ 9.70	50.80+14.80	24.89+11.03
11	CD58+	67.73+20.60	94.40+ 9.74	75.82+11.66	100	74.85+19.80
12	CD44+	90.61+11.49	89.00+11.30	85.96+ 6.83	100	97.27+ 2.25
13	CD62L+	53.13+12.02	51.37+15.80	57.86+15.26	78.91+ 6.19	73.47+14.09

The data express the mean ± standard deviation. ND = not determined due to cell subset barely detectable.

TABLE 2.

Antigen density of *early marker* and *cellular adhesion molecules* on CD34+ cells derived from five sources of Hematopoietic Progenitor Cells – **HPC** (Nb. of ABC x 10^3 molecules / cell).
nBM = normal bone marrow; **mBM** = mobilized bone marrow;
nPB = normal peripheral blood; **LKP** = leukapheresis product;
CB = cord blood.

It creates the table $\boldsymbol{a}_{i,j}$ of 65 empirical measurements:

	j→	1	2	3	4	5
i ↓	CD34+ cell subsets	nBM (n=13) In steady state	mBM (n=16)	nPb (n=13) In steady state	LKP (n=29)	CB (n=20)
1	$CD38^+$	56.60 ± 15.50	48.97 ± 18.31	35.00 ± 9.80	31.30 ± 16.00	37.00 ± 16.20
2	$HLA\text{-}DR^+$	196.25 ± 8.92	135.42 ± 25.62	89.16 ± 12.14	79.06 ± 8.13	57.28 ± 21.30
3	CD90 $(Thy\text{-}1)^+$	11.04 ± 5.50	11.00 ± 6.18	ND	9.40 ± 1.71	ND
4	CD117 $(c\text{-}kit)^+$	16.74 ± 3.48	8.62 ± 3.12	9.01 ± 4.20	4.96 ± 1.42	11.99 ± 1.43
5	$PgP170^+$	18.62 ± 6.00	18.05 ± 1.12	17.38 ± 2.65	17.76 ± 4.26	36.83 ± 2.34
	Integrin family					
6	CD11a	31.88+ 6.19	22.49+ 8.03	18.27+ 5.15	9.44+ 2.35	14.34+ 9.30
7	CD11b	13.15+ 5.75	11.79+ 4.74	13.04+ 8.45	10.20+ 9.35	12.56+ 1.30
8	CD49d	26.61+ 9.76	17.66+ 3.55	ND	8.23+ 1.39	23.60+ 5.13
9	CD49e	30.15+ 8.42	33.62+ 9.30	ND	18.32+ 4.71	ND
	Ig super family					
10	CD54	10.03+ 1.81	7.75+ 4.32	9.10+ 4.26	3.98+ 1.11	8.49+ 5.72
11	CD58	17.25+ 7.08	20.98+13.69	15.75+ 6.91	9.49+ 2.03	18.41+ 9.04
	Homing associated family					
12	CD44	144.89+24.09	124.93+24.89	103.23+28.32	96.58+17.77	160.36+42.64
13	CD62L	25.09+13.72	24.65+ 7.74	15.31+ 8.67	8.87+ 4.33	60.04+29.10

The data express the mean ± standard deviation. ND = not determined due to cell subset barely detectable.

The aim is to group 5 different sources of **HPC** in every series of data: $j = 5$ of the studied objects, depending on 13 common features which describe the objects: $i = 13$ measurement variables. The 5 objects will be compared between themselves but not in reference to the optimal model.

The concept enabling reciprocal comparison of objects through the analysis of their variables is similarity. The similarity of objects in taxonomy is measured by the means of distance. If the distance is smaller, objects differ between themselves less (they are closer to each other). The matrix $\boldsymbol{a}_{i,j}$ of empirical data consisting of 65 measurements has been standardized through the arithmetic mean for every variable: $s_{i,j} = \dfrac{a_{i,j}}{\overline{a_i}}$.

The *city metric* (Manhattan, Hamming distance) has been implemented which gives similar results as the Euclidean metric. Smaller is here, however, the influence of singular large detached differences. Properties of the metric $\boldsymbol{d}$:

1. identity of indiscernibles: $d_{m,n} = 0$ iff $m = n$,
2. symmetry: $d_{m,n} = d_{n,m}$,
3. triangle inequality: $d_{m,n} \leq d_{m,p} + d_{p,n}$.

Matrix of distances $d_{m,n} = \sum_{k=1}^{i} |s_{k,m} - s_{k,n}|$, where: $m, n = 1, \ldots, j$, $d_{m,n}$ – the distance between a pair of objects m and n, is a synthetic measurement of all variables.

Results for TABLE 1.

Matrix of distances $\boldsymbol{d}_{m,n}$

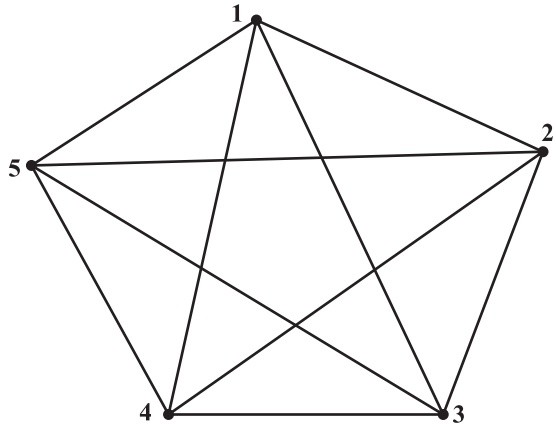
j → / ↓	1	2	3	4	5
1	0	4,33337	6,85084	8,41278	10,1635
2	4,33337	0	8,18558	4,33525	9,66091
3	6,85084	8,18558	0	10,3664	6,98268
4	8,41278	4,33525	10,3664	0	11,9020
5	10,1635	9,66091	6,98268	11,9020	0

This gives us a set of 10 distances $\dfrac{j!}{r!(j-r)!} = 10$, where $j = 5$, $r = 2$ (pair).

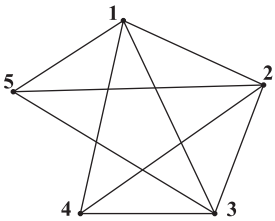
In arithmetic order

$$d_{1,2} = 4,33337$$
$$d_{2,4} = 4,33525$$
$$d_{1,3} = 6,85084$$
$$d_{3,5} = 6,98268$$
$$d_{2,3} = 8,18558$$
$$d_{1,4} = 8,41278$$
$$d_{2,5} = 9,66091$$
$$d_{1,5} = 10,1635$$
$$d_{3,4} = 10,3664$$
$$d_{4,5} = 11,9020$$

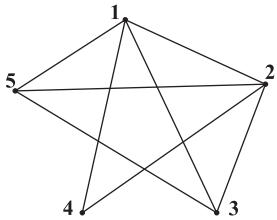
Next, a complete graph $\mathbf{K}_5$ of mutual connections is constructed where the nodes are the objects j, and edges are the distances $d_{m,n}$.



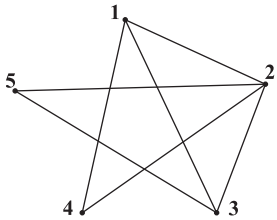
The third phase is the search of the disconnected graphs through successive elimination of edges which are of the largest distances.



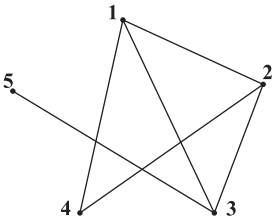
$d_{4,5} = 11,9020$



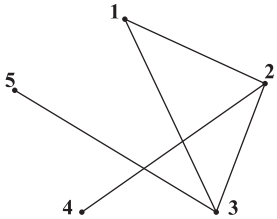
$d_{3,4} = 10,3664$



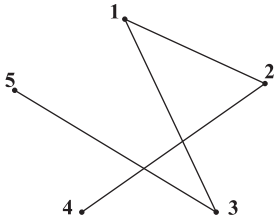
$d_{1,5} = 10,1635$



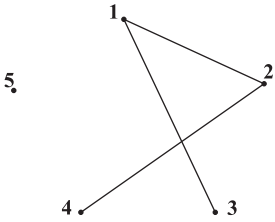
$d_{2,5} = 9,66091$



$d_{1,4} = 8,41278$

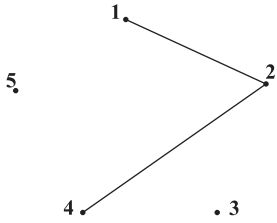


$d_{2,3} = 8,18558$



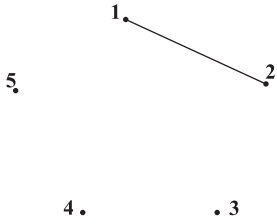
$d_{3,5} = 6,98268$

we have first object 5, in most far distance from others



$d_{1,3} = 6,85084$

object 3, next in distance



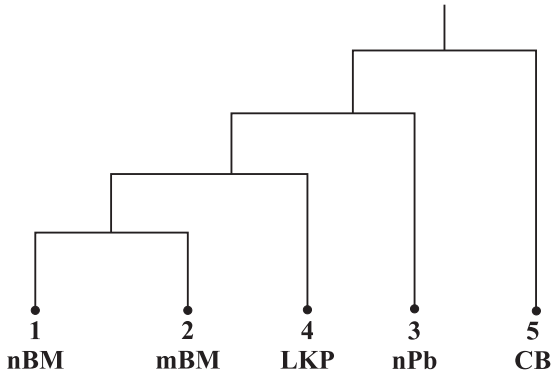
$d_{2,4} = 4,33525$

objects 1 and 2 are closest

The resulting incoherent graphs are groups of mutually disjoint subsets containing single linkage. This may be presented as a tree with clustered objects.

Resulted classification for TABLE 1:

> "Frequency of *early cell subsets* and *cellular adhesion molecules expression* within the CD34+ cell population derived from five sources of **HPC** (% of total CD34+ cells)"



nBM = normal bone marrow; **mBM** = mobilized bone marrow; **nPB** = normal peripheral blood; **LKP** = leukapheresis product; **CB** = cord blood

Figure above demonstrates that **LKP** is closest to **nBM** and **mBM** – these theoretical results confirm our expectations in implementing these three sources of **HPC** into transplantation.

Results for TABLE 2.

Matrix of distances $\boldsymbol{d}_{m,n}$

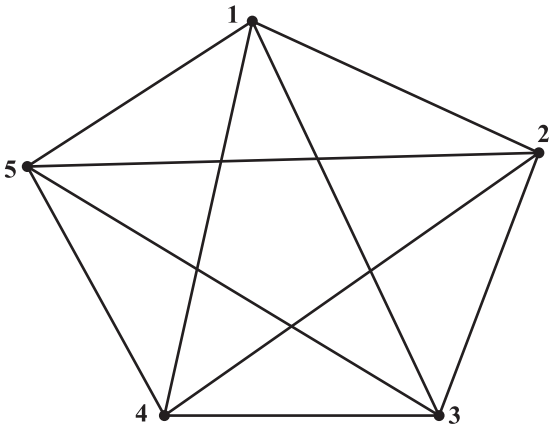
j → / ↓	1	2	3	4	5
1	0	3,66282	9,25248	8,67528	9,46347
2	3,66282	0	7,09278	5,89071	8,69142
3	9,25248	7,09278	0	5,76077	5,67728
4	8,67528	5,89071	5,76077	0	9,49803
5	9,46347	8,69142	5,67728	9,49803	0

This gives us a set of 10 distances $\frac{j!}{r!(j-r)!} = 10$, where $j = 5$, $r = 2$ (pair).

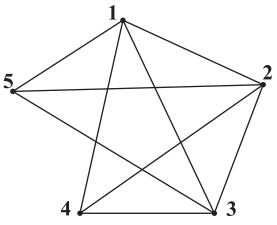
In arithmetic order

$$d_{1,2} = 3,66282$$
$$d_{3,5} = 5,67728$$
$$d_{3,4} = 5,76077$$
$$d_{2,4} = 5,89071$$
$$d_{2,3} = 7,09278$$
$$d_{1,4} = 8,67528$$
$$d_{2,5} = 8,69142$$
$$d_{1,3} = 9,25248$$
$$d_{1,5} = 9,46347$$
$$d_{4,5} = 9,49803$$

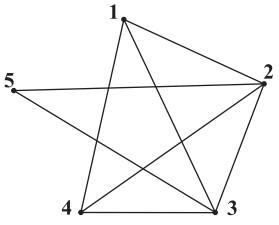
Next, a complete graph $\mathbf{K}_5$ of mutual connections is constructed where the nodes are the objects j, and edges are the distances $d_{m,n}$.



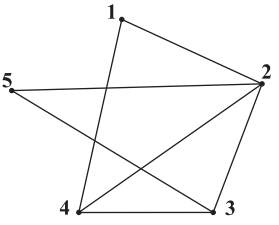
The third phase is the search of the disconnected graphs through successive elimination of edges which are of largest distances.



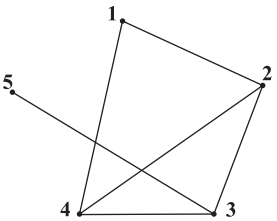
$d_{4,5} = 9,49803$



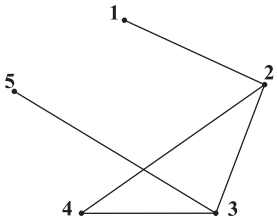
$d_{1,5} = 9,46347$



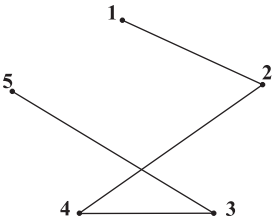
$d_{1,3} = 9,25248$



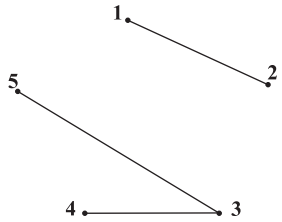
$d_{2,5} = 8,69142$



$d_{1,4} = 8,67528$

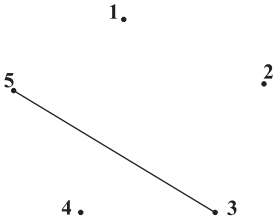


$d_{2,3} = 7,09278$



$d_{2,4} = 5,89071$

we have 2 disconnected graphs, it means that object 1 is closest to object 2



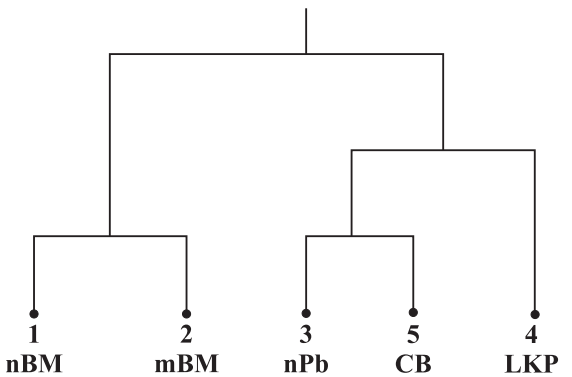
$d_{3,4} = 5,76077$

in next group object 3 is closest to object 5

The resulting incoherent graphs are groups of mutually disjoint subsets containing single linkage. This we may be presented as a tree with clustered objects.

Resulted classification for TABLE 2:

> "Antigen density of *early marker* and *cellular adhesion molecules* on CD34+ cells derived from five sources of **HPC** (Nb. of ABC x 10^3 molecules / cell)"



nBM = normal bone marrow; **mBM** = mobilized bone marrow; **nPB** = normal peripheral blood; **LKP** = leukapheresis product; **CB** = cord blood

Figure above demonstrates that **LKP** is the furthest from **nBM** and **mBM**, but it is at the same time in the group of circulated cells. The examination of features of this group enables to differentiate circulated cells vs. cells to stay in **BM**.

Results calculated with respect to standard deviation are identical as shown above.

REFERENCES

[1] Matuszewski R., Trybulec A., *An algorithm for clustering in metric spaces*, Papers of the Warsaw University in Bialystok, Vol. 5, pp. 117–123, 1977.

[2] Giero M., Matuszewski R., *Lower Tolerance. Preliminaries to Wroclaw Taxonomy*, Formalized Mathematics, Vol. 9, Nr. 3, pp. 597–603, 2001.

[3] Sovalat H., Racadot E., Ojeda M., et al., *CD34+ cells and CD34+ CD38 – subset from mobilized blood show different patterns of adhesion molecules compared to those from steady-state blood, bone marrow and cord blood.* Journal of Hematotherapy and Stem Cell Research, 2003, Vol. 12:5, pp. 473–489.

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 21 (34) 2010

Robert Milewski
Paweł Malinowski
Anna Justyna Milewska
Piotr Ziniewicz
Department of Statistics and Medical Informatics,
Medical University of Bialystok

Sławomir Wołczyński
Department of Reproduction
and Gynecological Endocrinology,
Medical University of Bialystok

THE USAGE OF MARGIN-BASED FEATURE SELECTION ALGORITHM IN IVF ICSI/ET DATA ANALYSIS

Abstract: In the case of infertility treatment, successful classification will facilitate understanding of various factors affecting the success of the process. Classification itself is an important data mining problem. Many classifications and constructions of the classifier algorithms are not able to cope with the analysis of the huge amount of factors associated with this process. Feature selection allows to significantly reduce the volume of analyzed data, while maintaining the classifier prediction quality. This leads to the rejection of nonessential measurements and time reductions.

Keywords: IVF ICSI/ET, infertility treatment, data mining, feature selection, margin, nearest neighbor classifier

Introduction

Infertility treatment is a process whose effectiveness depends on many different factors [1]. Specially designed and developed system to collect data of patients treated for infertility using the IVF ICSI/ET method was introduced at the Department of Reproduction and Gynecological Endocrinology, Medical University of Bialystok [2]. Although conducted research in recent years and some successful treatment predictors identification – all factors which significantly affect the final treatment results – still cannot be fully listed. Also, the prediction for the final result is not likely to be made possible. Therefore, research teams analyzing the infertility treatment effectiveness refer to the increasingly sophisticated statistical methods and bioinformatics. Some results in the prediction of treatment failure were ob-

ISBN 978–83–7431–266–0 ISSN 0860-150X

tained after the neural networks application [3]. These results offer hope to find more advanced methods for predicting treatment effectiveness and convincing us to further utilization of the possibilities opened up by bioinformatics.

The issue of observation classification is one of the most important problems in data mining. In the context of supervised classification, it is to draw a conclusions from data based on gathered earlier results. In the process of the so-called conclusion-drawing (classifier building) chosen algorithm analyzes this data to search patterns. As a result of this a process set of rules (classifier) is created, which allow proper label assignment for a new observation. A good example is the patient's prognosis analysis relying on identified symptoms. Good classifier should be primarily characterized by high speed of execution and correctness of results, also for new observation (so called bias). Intuitively, this imposes a requirement of its intrinsic simplicity – fewer (possibly biased) rules make smaller computational costs and result in lower bias of the entire classifier. Simplicity of the final classifier makes it easier to interpret. It is a classic realization of Ockham's razor.

One of the major problems of effective conclusion-making process is the curse of the dimensionality phenomenon. Input data can have huge number of features (large dimensionality), but only some of them are significant in the classification process. Excessive features obscure regular patterns in the data, decreasing the signal-to-noise ratio. Time cost of conclusion-drawing and the target classification can grow very rapidly with the increasing dimensionality of the input. A very good example of multidimensional data are medical data, which usually contain a lot more features than observations (the ratio seeking up to 10^5 for the gene expression analysis). To cope with the curse of dimensionality, various techniques exist, such as "regularization" (boosting [4], bagging [5]), kernel methods [6], and others. One of them is feature selection and extraction, which represents dimensionality reduction approach.

Feature selection

Feature selection is a process of choosing proper feature subset from all such possible subsets. Feature selection is close to the feature extraction task, often these terms are used interchangeably in literature. In this article it is assumed, that feature extraction consists of the feature construction and feature extraction phase. Feature construction is a process of linking and transforming low level features into higher one. An example of such

technique is the PCA [7], or picture conversion from color to grayscale. This article will focus only on the feature selection phase, assuming that high level features are already constructed.

As a result of the feature selection process a certain feature subset is chosen, which satisfies certain criteria, captures relevant properties of the data, and is also useful in the context of used classifier. Commonly used concepts of relevance and usefulness of a feature subset were established on set theory and probabilistic ground. Feature subset $c' \subseteq C$ is called optimal (1), when accuracy $acc(*)$ of W_{cl} classifier for $\boldsymbol{X}/c'$ set (dataset formed the basis of $\boldsymbol{X}$, in which all data related to features not belonging to subset c' where removed) will be maximum among all such $\boldsymbol{X}/c$ subsets. The feature is called useful, when it belongs to optimal subset.

$$\forall_{c \subseteq C}\ acc(W_{cl}(\boldsymbol{X}/c')) \leq acc(W_{cl}(\boldsymbol{X}/c)) \tag{1}$$

$$\forall_{c \subseteq C-\{F_i\}}\ P(K|c \cup \{F_i\}) = P(K|c) \tag{2}$$

$$P(K|C) = P(K|C - \{F_i\}) \wedge \exists_{c \subset C-\{F_i\}}\ P(K|c \cup \{F_i\}) \neq P(K|c) \tag{3}$$

$$P(K|C) \neq P(K|C - \{F_i\}) \tag{4}$$

$$\exists_{c \subset C-\{F_i\}}\ P(C - c - \{F_i\}, K|c \cup \{F_i\}) = P(C - c - \{F_i\}, K|c) \tag{5}$$

where: c' optimal feature subset; $\boldsymbol{X}$ data set; C set of all features from $\boldsymbol{X}$ data set; c feature subset; W_{cl} classifier related to final conclusion-making; $acc(*)$ classifier accuracy; F_i feature; $P(*)$ probability; K class distribution.

In terms of relevance, features were divided into redundant (2), weak relevant (3) and strong relevant (4). All probabilities listed in (2)–(4) formulas are conditional probabilities of class distribution against certain feature subset. Presence of strong relevant features changes such distribution (4). Presence of irrelevant features does not affect this distribution (2). In case of the weak relevant feature, certain subsets of features exist for which presence of this feature changes the conditional distribution (3).

Depending on other features, a weak relevant one can be locally strong relevant, locally irrelevant. Taking into account local feature relevance phenomena, feature set can be divided in 4 subsets. Figure 1 presents a review of those subsets (1 – strong relevant feature subset, 2 – weak relevant, but locally relevant feature subset, 3 – weak relevant, but locally irrelevant feature subset, 4 – irrelevant feature subset). Optimal feature subset should contain all strong relevant features and weak relevant, but locally relevant ones (feature subsets 1+2 from Figure 1). Collection of irrelevant and weak

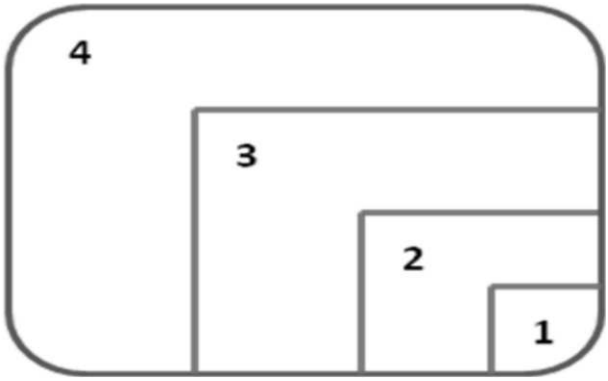


Fig. 1. Set of features division according to their relevancy

relevant, but locally irrelevant features make redundant the feature set (feature subsets 3+4 from Figure 1). Feature F_i is redundant, if equation (5) is satisfied. All these definitions were presented in [8].

The selection of a relevant feature can be accomplished in two ways: by feature ranking or subset selection. In case of feature ranking, score is assigned for each feature, according to certain criterion. The next step are rejection features below chosen score threshold. In the subset selection approach, each possible subset of features is scored by criterion function according to three different models: filter, wrapper or embedded.

In the filter model, space of feature subsets is searched, and in each search step simple filter is used to measure subset score. This filter is usually a statistical measure, or is based on entropy, or the heuristic approach. In a special case, when single feature is treated as a subset, filter model reduces to earlier mentioned feature ranking process. In the wrapper model, space of feature subsets is also searched but in each step of searching final learning algorithm is launched. Based on its results, feature subsets comparison is possible. Embedded models include a group of algorithms, which are characteristic for a process of conclusion-drawing. These are similar to the wrapper models, the difference is, that the process of conclusion-drawing itself directs search and evaluation of different subsets of features.

Margin-based feature selection algorithm

Margin is a geometrical measure of certainty and generalization abilities given by the classifier. Many conclusion-making algorithms and classifiers (e.g. SVM [6]) make use of the margin concept. Articles [9, 10] proposed

margin as a quality measure of feature subset that generate this margin. When searching for an optimal feature subset the algorithm tends to increase such margin. This article suggests the SIMBAF algorithm, which is a modification of the SIMBA [10] algorithm. It is a subset selection method type implementing the filter model. Algorithm accepts data which features:

1. numeric – taking values from real number field
2. ordinal – features which values equivalent to natural numbers (by article convention with 0), order is maintained, not necessarily mutual distance. For such features the distance d between 0 and 1 values not necessarily the same as f. e. distance between 1 and 2 values, but condition $d(0,1) < d(0,2)$ always occurs, and "1" value is between "0" and "2" values.
3. categorical – features which values equivalent to natural numbers (by article convention with 0), order is not maintained, neither mutual distance.

$$\Delta(\boldsymbol{x}_1, \boldsymbol{x}_2) = \Delta_{num,p}(\boldsymbol{x}_1, \boldsymbol{x}_2) + \Delta_{cat}(\boldsymbol{x}_1, \boldsymbol{x}_2) + \Delta_{ord}(\boldsymbol{x}_1, \boldsymbol{x}_2) \tag{6}$$

$$\Delta_{num,p}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left\{ \sum_i \left[\alpha(t_i) \phi_{num}(x_{1i}, x_{2i}) \right]^p \right\}^{1/p} \tag{7}$$

$$\Delta_{cat}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_j \alpha(t_j) \phi_{cat}(x_{1j}, x_{2j}) \tag{8}$$

$$\Delta_{ord}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_k \alpha(t_k) \phi_{ord}(x_{1k}, x_{2k}) \tag{9}$$

where: $\boldsymbol{x}_1, \boldsymbol{x}_2$ observation; $\Delta(*,*)$ observation distance measure, low index means parts: numerical $\Delta_{num,p}(*,*)$, categorical $\Delta_{cat}(*,*)$, ordinal $\Delta_{ord}(*,*)$; i, j, k indexes; p metric order; t_i, t_j, t_k "hidden" parameters, $\alpha(*)$ – specific function, x_{1*}, x_{2*} value of feature $*$ for given observation; $\phi_*(*,*)$ measures of difference between single feature values.

Derivation of target margin form requires definition of the distance between compared observations. It was defined according to (6). Proper parts of expression (6) are related to the following features: numerical (7), categorical (8) and ordinal (9). Indexes i, j, k iterate respectively over selected features types. Introduced later index o will iterate over all features. Proposed algorithm breaks with using modified Euclidean metrics for numerical features [10], replacing it with the Minkowski metric of any order p (7). It should be noted, that for $p < 1$, related formula (7) is not formally metric anymore, because triangle inequity axiom is not satisfied (for such cases opposite direction inequality occurs). As it will be later discussed, change of this parameter leads to interesting and important conclusions and gene-

ralizations. For categorical and ordinal features, due to their nature, metric cannot be introduced as such, and therefore special functions of the similarity are introduced to replace it.

$$\alpha(t_o) = \frac{1}{\pi}\left(\text{arctg}(t_o) + \frac{\pi}{2}\right) = \frac{\text{arctg}(t_o)}{\pi} + \frac{1}{2} \qquad t_0 \in \Re; \quad \forall_{t_o}\ \alpha(t_o) \in (0,1)... \tag{10}$$

$$\delta_o = \max_{b,d} |x_{bo} - x_{do}|; \quad 0 \le \varepsilon_{o1} \le \varepsilon_{o2} \le 1 \tag{11}$$

$$\phi_{num}(x_{1i}, x_{2i}) = \min\left(1, \max\left(0, \frac{|x_{1i} - x_{2i}| - \delta_i \varepsilon_{i1}}{\delta_i(\varepsilon_{i2} - \varepsilon_{i1})}\right)\right) \in [0,1] \tag{12}$$

$$\phi_{cat}(x_{1j}, x_{2j}) = \begin{cases} 0 & \iff x_{1j} = x_{2j} \\ 1 & \iff x_{1j} \neq x_{2j} \end{cases} \tag{13}$$

$$\phi_{ord}(x_{1k}, x_{2k}) = \min\left(1, \max\left(0, \frac{|x_{1k} - x_{2k}| - \delta_k \varepsilon_{k1}}{\delta_k(\varepsilon_{k2} - \varepsilon_{k1})}\right)\right) \in [0,1] \tag{14}$$

where: δ_o value range of o-th feature; ε_{o1} i ε_{o2} cut-off parameters for o-th feature (there are only 2 such parameters)

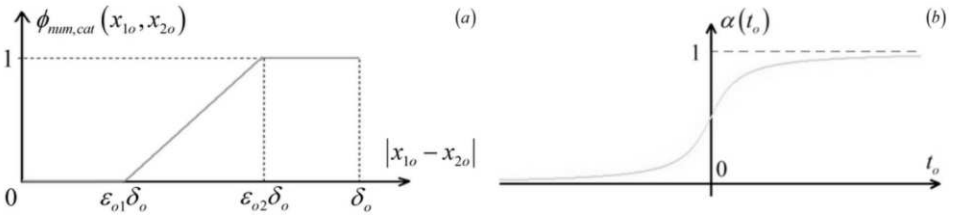


Fig. 2. (a) distance measure for numerical and order features (b) factor $\alpha(t_o)$

Factor $\alpha(t_o)$ plays an important role, it is a scale factor and also calculated weight of o-th feature. In proposed algorithm it is a function (10) of optimized "hidden" parameter t_o (it is used internally). Figure 2b presents the plot of this function. The use of proper function allows to regulate calculated weights. SIMBAF algorithm is not limited to this single function, actually any other sigmoid function normalized to (0,1) can be applied. Functions $\phi(*,*)$ determine the measure of distance between values of a given feature compared observation. They were identified separately for the numerical features (12), categorical (13) and ordinal (14). All these distances were also normalized to the range [0,1] in order to normalize their impact on the target margin. For numeric and order features, the so-called

cut-off parameters ε_{o1} and ε_{o2} are introduced, which are designed to filter out extreme values directly modifying the distance function. They were taken from article [11]. The behavior of measurement ϕ_{num} and ϕ_{ord} is shown on Figure 2a.

$$m = \sum_{\boldsymbol{x} \in \boldsymbol{X}} [\Delta(\boldsymbol{x}, miss(\boldsymbol{x}, u)) - \Delta(\boldsymbol{x}, hit(\boldsymbol{x}, u))] \tag{15}$$

where: m margin; $\boldsymbol{x}$ observation; $hit(\boldsymbol{x}, u)$ u-th nearest same class neighbor of observation $\boldsymbol{x}$; $miss(\boldsymbol{x}, u)$ u-th nearest different class neighbor of observation $\boldsymbol{x}$;

Equation (15) presents final margin form. Parameter u specifies, which nearest neighbor should be taken into consideration. The increase of this parameter decreases the algorithm's sensitivity for outliers, and increases generalization abilities. It is worth noting that the margin in this form is a function of features weights $\alpha(t_o)$. It is important, because it allows weight optimalization. Similar to SIMBA algorithm the gradient ascend method is used. For randomly chosen observation $\boldsymbol{x}$, all t_o parameters are updated according to formula (16) and (17).

$$dt_o = \frac{\partial \Delta(\boldsymbol{x}, miss(\boldsymbol{x}, u))}{\partial t_o} - \frac{\partial \Delta(\boldsymbol{x}, hit(\boldsymbol{x}, u))}{\partial t_o} \tag{16}$$

$$t_o = t_o + dt_o \tag{17}$$

$$\frac{\partial \Delta(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_o} = \begin{cases} \dfrac{\partial \Delta_{num,p}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_i} & \text{(a)} \\ \dfrac{\partial \Delta_{cat}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_j} & \text{(b)} \\ \dfrac{\partial \Delta_{ord}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_k} & \text{(c)} \end{cases} \tag{18}$$

$$\frac{\partial \Delta_{num,p}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_i} = \phi_{num}(x_{1i}, x_{2i}) \left[\frac{\alpha(t_i) \cdot \phi_{num}(x_{1i}, x_{2i})}{\Delta_{num,p}(\boldsymbol{x}_1, \boldsymbol{x}_2)} \right]^{p-1} \frac{\partial \alpha(t_i)}{\partial t_i} \tag{19}$$

$$\frac{\partial \Delta_{cat}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_j} = \phi_{cat}(x_{1j}, x_{2j}) \frac{\partial \alpha(t_j)}{\partial t_j} \tag{20}$$

$$\frac{\partial \Delta_{ord}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\partial t_k} = \phi_{ord}(x_{1k}, x_{2k}) \frac{\partial \alpha(t_k)}{\partial t_k} \tag{21}$$

$$\frac{\partial \alpha(t_o)}{\partial t_o} = \frac{1}{\pi(1 + t_o^2)} \tag{22}$$

where: dt_o adjustment for o-th feature

Appropriate partial derivative (18) should be chosen depending on the type feature: numerical (a), categorical (b) or ordinal (c). Proper formulas for derivatives have been given in equations (19)–(20). Equation (22) defines a derivative of the chosen weight function depending on the free hidden parameter.

Presented margin form draws very interesting conclusions. Equation (22) shows, that for increasing absolute values t_o algorithm adjustments are declining resulting in stabilization of weights $\alpha(t_o)$ near extreme values 0 or 1 while preserving numerical stability of the solution. Moreover, given the algorithm includes, as special cases, the following algorithms:

1. For $u = 1$, $p = 1$ special case of ReliefF algorithm is get
2. For $\varepsilon_{o1} = 0$, $\varepsilon_{o2} = 1$, $u = 1$, $p = 1$ Relief algorithm is get
3. For $\varepsilon_{o1} = 0$, $\varepsilon_{o2} = 1$, $u = 1$, $p = 2$ SIMBA algorithm is get

Factor p plays special role and it is a form of features mixing factor. Clearly, by increasing p, partial derivative (19) is decreasing. Factor in rectangle bracket (19) powered to $p - 1$ is responsible for redundant features reduction in relation to Relief. Therefore, increasing p to reasonable values should decrease redundancy.

Additional details: multiclass problems, missing values, nearest neighbor choose and final algorithm form

Relief algorithm [11] can also identify important features in case where data set has more than two classes and in case of incomplete data. The same method can be used directly in the presented algorithm. Slightly modifying the formula (16) SIMBAF algorithm can be adapted for many classes. A modified form of the optimized hidden parameter adjustment shows equation (23).

$$dt_o = \sum_{c \neq c(\boldsymbol{x})} \frac{P(c)}{1 - P(c(\boldsymbol{x}))} \frac{\partial \Delta(\boldsymbol{x}, miss(\boldsymbol{x}, u, c))}{\partial t_o} - \frac{\partial \Delta(\boldsymbol{x}, hit(\boldsymbol{x}, u))}{\partial t_o} \quad (23)$$

where: $P(c)$ probability of class c; $c(\boldsymbol{x})$ class of observation $\boldsymbol{x}$; $miss(\boldsymbol{x}, u, c)$ u-th nearest class c neighbor of observation $\boldsymbol{x}$.

Another issue is the nearest neighbor choice. Nearest neighbor selection algorithm is sensitive to the curse of dimensionality. For numeric data it has been shown [12], that it might be better to abandon Euclidean metric in favor of Minkowski metrics, even with a fractional order, which better retains the distinction of data points in high-dimensional space. The distance

measure between points is calculated according to formula (24). It should be noted that the parameter p' in the formula below does not correspond to the parameter p from equation (7). Sometimes p' can be as low as p^{-1}.

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \left[\sum_i \left(\frac{|x_{1i} - x_{2i}|}{\delta_i}\right)^{p'}\right]^{1/p'} + \sum_j \phi_{cat}(x_{1j}, x_{2j}) + \sum_k \frac{|x_{1k} - x_{2k}|}{\delta_k} \quad (24)$$

where: $d(*,*)$ distance between two observations; p' metric order.

In case of missing data the following rules were used:

1. When for the first and second observation for a given feature both values are missing, distance and t_o parameter adjustment is set to 0
2. When numerical value is missing, it is replaced by mean value among this feature values
3. When categorical value is missing, distance is a probability that the category is different than in compared observation
4. When ordinal value is missing, it is replaced by median value among this feature values

Listing 1: SIMBAF Algorithm

```
Input data:
    X;                  /*data*/
    N;                  /*feature count*/
    i_max;              /*iteration count*/
    u;                  /*neibour to analize*/
    eps1[N], eps2[N];   /*cutoff parameters*/
    p, p';              /*metric parameters*/
Output data:
    w[N];               /*final weights*/
Variables:
    i,j;
    x;                  /*observation*/
    dt[N], t[N];
Algorithm:
    for j=1 to N
        t[j]=0;
    for i=1 to i_max
        pick up random obserwation x from X
        for j=1 to N
            calculate dt[j] using formula (23)
```

```
        t[]=t[]+dt[]
        for j=1 to N
            calculate w[j] using formula (10)
    return w[];
```

Empirical results and conclusions

The algorithm described above was used to analyze the effectiveness of infertility treatment by the IVF ICSI/ET method on the data set obtained at the Department of Reproduction and Gynecological Endocrinology, Medical University of Bialystok. Specifically designed for this purpose, the application was used to collect these data [2]. The input data set contained a description of 1445 treatment cycles. Each cycle of treatment was represented by 149 independent features (including 107 numerical features, one ordinal feature and 61 categorical features) and one dependant feature – treatment result (pregnant or not). Of course, because of potential treatment process failure at an earlier stage, there were cases of missing data (particularly in describing the characteristics of the final treatment stages). The lower cut-off parameters (ε_{o1}) for numeric and ordinal data was set to 0.1, while the upper (ε_{o2}) to 0.9. Orders of metrics from the equations (7) and (24) were set accordingly to 2.5 (p) and 0.5 (p'). As a result of the described algorithm execution on the collected data, the following features set was obtained (in order, starting with the largest weight):

- The type of treatment protocol (protocols types described in [2])
- Mucus during ovulation,
- The type of anesthesia at the puncture of ovarian follicles,
- Fallopian factor as a cause of infertility,
- Pain during ovulation,
- Male factor as a cause of infertility,
- Temperature increase during ovulation,
- The number of embryos transferred in the third day of culture,
- Semen preparation – twice washing,
- Hyperprolactinemia in medical history,
- Polycystic ovary syndrome as a cause of infertility,
- Blood spotting during ovulation,
- Type B sperm motility,
- The number of embryos transferred in the second day of culture,

- Type A sperm motility
- Endometriosis as a cause of infertility,
- Age of patient,
- The number of expanding blastocysts in the fifth day of culture.

In the generated as a result features set are those, which had been identified long ago as having a significant impact on the effectiveness of infertility treatment. The addition of other features that previously did not seem to have a significant effect on the treatment result may have a significant impact on improving the quality of prediction based on the so-constructed set of features. This fact confirms the desirability of data collection based on the greatest possible number of features, even potentially non-essential to the treatment process.

The analysis of the above data confirmed also the fact, that the concept of margin plays an important role in the process of building the artificial intelligence mechanisms. Algorithms, that can effectively reduce the number of features to be analyzed, were able to be built using a relatively simple mathematical mechanism, improving the generalizing properties of many classifiers and certainly, the speed of data analysis.

The issue of weight optimization in the presented algorithm requires further study. There are better known methods of global optimization algorithms than used here, e.g. the gradient ascend. The function of the margin itself is highly nonlinear and the algorithm may have a tendency to be stuck in local maximum. Another issue is the form of the optimized margin function itself. In the presented algorithm, it is simply a modified measure of distance (similarity) between observations. The introduction of another margin function could improve the optimization properties. An interesting issue requiring further research are other functions of hidden parameter usage in the role of feature weight.

The next step after identifying a set of significant features is the construction of classifiers from the resulting data set using various methods available. There are many state–of–the–art classifiers whose performance should be checked and the results compared with each other. The aim of further studies will be the development of a series of more–improved predictive models of the effectiveness of infertility treatment using the IVF ICSI/ET method.

REFERENCES

[1] Radwan J. (ed.) *Niepłodność i rozród wspomagany.* Termedia, Poznań 2005.

[2] Milewski R., Jamiołkowski J., Milewska A. J., Domitrz J., Wołczyński S. 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), 2009.

[3] Milewski R., Jamiołkowski J., Milewska A. J., Domitrz J., Szamatowicz J., Wołczyński S. Prognozowanie skuteczności procedury IVF ICSI/ET – wśród pacjentek Kliniki Rozrodczości i Endokrynologii Ginekologicznej – z wykorzystaniem sieci neuronowych. Ginekologia Polska, 80 (12), 2009.

[4] Schapire R. E. The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, B. Yu, ed., *Nonlinear Estimation and Classification.* Springer, 2003.

[5] Breiman L. Bagging predictors *Machine Learning*, 24(2), pp. 123–140, 1996.

[6] Boser B. E., Guyon I., and Vapnik V. A training algorithm for optimal margin classifiers. *Proceedings of the 5th Annual Workshopon Computational Learning Theory*, pp. 144–152, ACM Press, 1992.

[7] Jollifee I. T. Principal component analysis. Springer Varlag, 1986.

[8] Kohavi R., John G. Wrapper for feature subset selection. *Artificial Intelligence*, 97, pp. 273–324, 1997.

[9] Kira K., Rendell L. A practical approach to feature selection. *Proceedings 9th International Workshop on Machine Learning* pp. 249–256, 1992.

[10] Gilad-Bachrachy R., Navot A., Tishby N. Margin Based Feature Selection – Theory and Algorithms. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

[11] Kononenko I., Estimating attributes: analysis and extensions of Relief. Bergadano F., De Raedt L. ed, *Proceedings European Conference on Machine Learning*, 1994.

[12] Aggarval Ch. C, Hinneburg A., Keim D. A. On the Surprising Behavior of Distance Metrics in High Dimensional Space. *Lecture Notes in Computer Science*, 1973/2001, Volume, pp. 420–434, 2001.

Małgorzata Ćwiklińska-Jurkowska
Department of Theoretical Backgrounds of Biomedical Science
and Medical Informatics, Collegium Medicum in Bydgoszcz,
Nicolaus Copernicus University

EXPLORATORY DATA ANALYSIS FOR THE HEMATOLOGICAL FEATURES. PART I. METHODOLOGY

Abstract: Part I of the work characterizes the data of human blood parameters. It describes examined biomedical data set and used multivariate statistical exploratory methods like PCA, FA, MDS and clusters analysis. To the factor analysis methods belongs the biplots visualization method. Biplots are simply the scatterplots with the superimposition of the variables. The work compares different other alternative multivariate exploratory data analysis procedures from methodological point of view. The described multivariate ordination methods are applied in the Part II of the work "Application".

Key words and phrases: exploratory data analysis, graphical data visualization, principal components, biplots, MDS, cluster analysis

Introduction

In exploratory data analysis (Larose 2005), one usually has not the a priori idea of expected relations between variables. Exploratory analysis allows to discover interesting relationships between variables. It also enables to identify interesting subsets of a data set and to develop initial ideas of possible connections between features and perhaps with the dependent or classifying variable. Graphical exploratory methods in statistics can explore even unknown phenomenon.

Ordination, i.e. geometrical representation of multivariate data as a low dimensional arrangement of points is a required procedure in many applied research problems. A picture is often needed which can provide meaningful interpretation of the data. Hopefully, the picture will supply helpful information about the relationships between the individuals. When one has a priori grouping of the individuals and searches a low dimensional illustration of the data, highlighting differences between them means one can apply e.g. canonical variate analysis or its generalization. Following the application of

ISBN 978–83–7431–266–0 ISSN 0860-150X

the explorative data analysis in a research domain one wants to know what are the important configurations or characteristics that can be seen in this field and how much of the variability is clarified by them.

Very common statistical technique is to plot a scatter diagram showing the pattern of relationships between a set of samples for only two or three original variables. The next step for such scatter diagrams may be looking for trends (e.g. regression lines), clusters, outliers, collinearities or other regularities. The multivariate extensions are needed to represent p-variate set of observations. To display multivariate data in a two dimensional plot, Bartkowiak et. al (1998) applied methods of the grand tour and hierarchical visualization.

In multidimensional statistical problems one wants to reduce a data set containing many variables to a data set containing considerably smaller number of variables, but that still corresponds to a large part of the variability embodied in the original data set. After applying explorative data analysis in a research domain, one wants to know what are the important patterns that appear in this field and how much of the variance is explained by them.

One of the frequent possibilities include, in the first step, the approximation of p-variate space into s-variate space, saving as much information derived by the data as possible. Common methodology is the Principal Component Analysis (PCA). In the principal component analysis p-dimensional scatter of these cases is approximated by their scatter in an r-dimensional sub-space, obtained by orthogonal projection of R_p onto R_s, chosen as this mininimizing the sum of squares of the residuals orthogonal to R_s. The approximation of the bipolot variables is given by the biplot axes.

Useful method of data ordination, including the information of both cases and variables are biplots. Biplots are simply the scatter plots with the superimposition of the variables. Thus, by using this method of data ordination, including the information of both cases and variables, the relationship between the hematological observations and variables may be investigated.

During interventions with extracorporeal circulation, such as hemoperfusion, different blood parameters are changing. The interdependences between different hematological parameters of human blood are interesting. In multivariate statistical problems there is an obvious need for graphical visualization of the data. Simple biplot diagnostic modeling may be applied. Investigation of the structure of the hematological data by studying scatterplot matrices and biplots with compared additional multivariate techniques is the aim of this work. The examined data set includes measurements of blood when hemoperfusion was done with polymer sorbent.

Material

Experiments in vitro with polymeric sorbents and interdependencies between variables describing parameters of human blood are considered. Worse hematological parameters are frequent complications of interventions with extracorporeal circulation, so eleven hematological variables, reported for each of the 36 experiments in vitro, performed in Colleqium Medicum of Nicolaus Copernicus University were examined, where also the feature of the perfusion experiment time was considered. Hemoperfusion of human blood was done with the polymer sorbent.

Table 1

Names and analyzed variables' description

Feature (variable)	Description
ADHEVIN	Level of platelets' adhesion
TIME	TIME trwania perfuzji / Time of perfusion
CZFIVIN	TIME fibrynolizy / Time of fibrinolise
ERYTVIN	Liczba erytrocytów / Number of erythrocytes
FIBGVIN	Fibrinogen concentration
KAOLVIN	Kaolin-kefalin time
KEFAVIN	Kefalin time
LEUKVIN	Leukocytes number
PROTVIN	Prothrombin time
STYPVIN	Stypven-kefalin time
TROMVIN	Thrombocytes number

The description of the examined hematological features is given in Table 1.

"ADHEVIN" – Level of platelets' adhesion; "TIME" – time of perfusion ("CZAS"); "CZFIVIN" – Time of fibrinolise; "ERYTVIN" – Number of erythrocytes; "FIBGVIN" – Fibrinogen concentration; "KAOLVIN" – Kaolin-kefalin time; "KEFAVIN" – Kefalin time; "LEUKVIN" – Leukocytes number; "PROTVIN" – Prothrombin time; "STYPVIN" – Stypven-kefalin time and "TROMVIN" – Thrombocytes number.

Methods

In multivariate analysis looking simultaneously at many variables is difficult. The most advanced and complex methods do not exceed five origi-

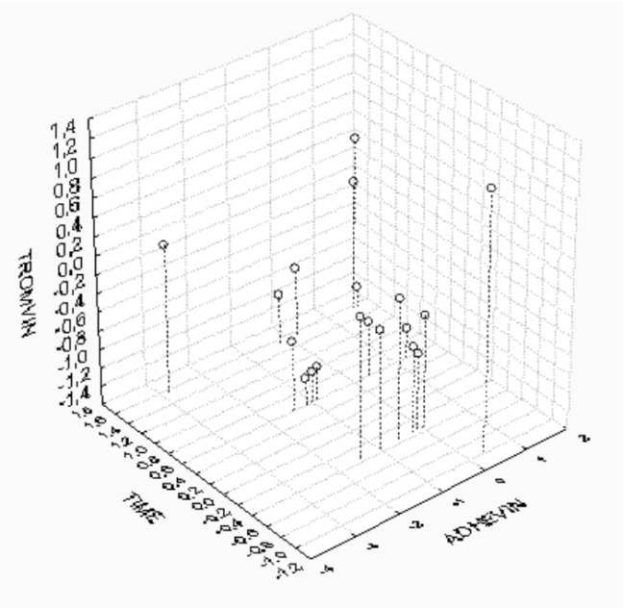


Fig. 1. Scatterplot of three dimensional space based on original variables ADHEVIN, TIME, TROMVIN

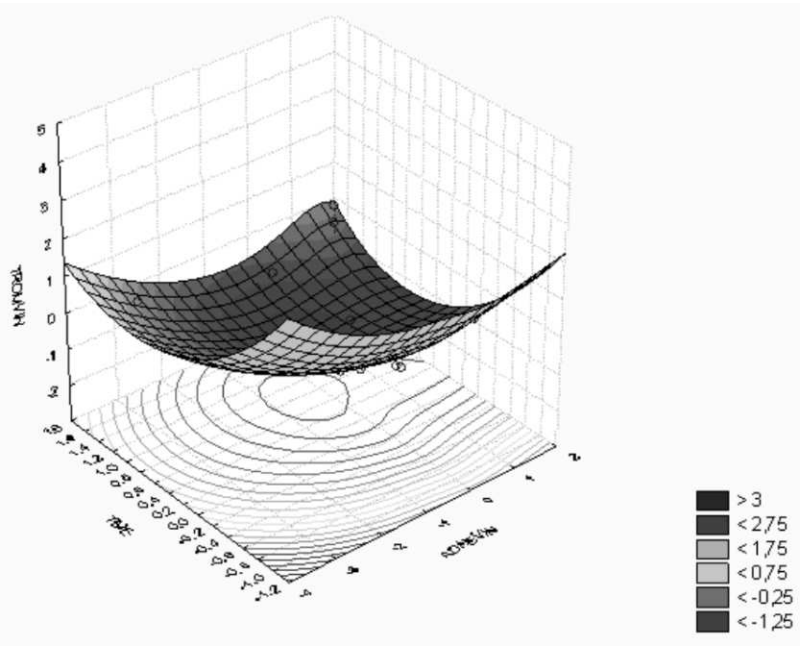


Fig. 2. Scatterplot of three dimensional space based on original variables ADHEVIN, TIME, TROMVIN with overlaid smoothed surface of the relationship between the three variables

nal variables dimensionality. For example, we can look at three-dimensional scatterplot for all three subsets of variables. Fig. 1 presents such three-dimensional scatterplot for applied hematological data set. Fig. 2 contains the same scattterplot with overlaid smooth surface of relationship between the three variables.

A scatterplot matrix permits simultaneous looking at relations between pairs of variables by observing two-dimensional scatterplots presenting the values for fixed pairs of variables. The multivariate extensions (for higher than two or three-dimensional data set) are useful to represent p-variate set of observations. One of the possibilities includes in the first step the approximation of p-variate space into s-variate space, saving as much information derived from the data as possible. Common methodology is the Principal Component Analysis (PCA). In principal, component analysis p-dimensional scatter of these cases is approximated by their scatter in an r-dimensional sub-space, obtained by orthogonal projection of R_p onto R_s, chosen as minimizing the sum of squares of the residuals orthogonal to R_s. Biplots can provide diagnostic aid in analysis.

Biplots are useful procedures in exploratory data analysis and visualization of data sets. The rules of constructing a biplot are given by papers of Gabriel (1971, 1981, 1990) and also can be found in the book written by Krzanowski (1988, 1995).

Principal components are uncorrelated linear combinations of the variables and with variances $\lambda_1, \lambda_2, \ldots, \lambda_p$. Principal components are helpful for reducing the number of variables by finding linear combinations that explain the biggest part of the variability. The coefficients of each principal component are determined by eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of covariance matrix Σ. The principal components are sorted by descending (non increasing) order of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$, which are equal to the variances of the components. Each principal component is a linear combination of the original variables, with coefficients equal to the eigenvalues of the covariance matrix Σ.

The eigenvectors are as a rule taken with unit length. The second component is orthogonal to the first and so on. The coefficients are customarily normalized to 1. The consequence of orthogonality of principal components is a summarizing of variances of all principal components λ_i to overall variance. The usefulness of the principal component is measured by the value of the variance which this component explains.

Using s first principal components (with the highest variances) summarizing most of the variability in the data, one can approximate high-dimensional data in a lower-dimensional linear subspace (Morrison 1976, Mardia

et al. 1979). In this way the dimension can be reduced from p to s. The coordinate system is rotated to make parallel first s coordinates $Y_1, \ldots, Y_s$ with the first s eigenvectors corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_s$ of the common sample covariance matrix Σ.

Connections between similar ordination multidimensional methods like canonical variate analysis, principal component analysis and multidimensional metric scaling are recapitulated below in Tab. 2.

Table 2

Relations between methods of multivariate data ordination based on correlations

	Continuous (quantitative) variables	*Mixed quantitative-qualitative variables*	*Qualitative variables*
A priori grouping into k populations	Canonical variate analysis (Rao 1973, Mardia et al. 1979)	Metric scaling for matrix (*pxp*) of inter-population distances (+add a point technique (Krzanowski 1994)	Correspondence analysis (Hill 1974)
No a priori grouping (one population)	Principal Component Analysis	Metric scaling for matrix (*nxn*) of inter-object distances (Mardia et al. 1979)	Correspondence analysis (Hill 1974)

The fundamental tool for obtaining biplots is the PCA. Principal component analysis (PCA, Hotteling 1936) yields usually reduced plots of multivariate observations. A biplot is a graphical representation of information given in *nxp* matrix $\boldsymbol{X}$. Biplots commonly use two or three dimensions (Gabriel 1971, Bartkowiak et al. 1996). The term "biplot" is not concerned with the dimension of representation, but means that this is a dual representation, variables and individuals on the same plot. Thus tree-dimensional biplots are also possible, or even higher dimensional, though not clear.

Biplots are constructed on the base of the fact that any cases matrix $\boldsymbol{X}$ (nxp) can be expressed as the product of two matrices $\boldsymbol{G}$ (nxr) and $\boldsymbol{H}$ – (rxp), i.e. $\boldsymbol{X} = \boldsymbol{GH}$, where $r = \text{rank(X)}$. It follows that for any multivariate observation, $\boldsymbol{x}_{ij} = \boldsymbol{g}_i' \boldsymbol{h}_j$ ($i = l, \ldots, n$; $j = 1, \ldots, p$), where $\boldsymbol{g}_i'$; and $\boldsymbol{h}_j$ are the r-dimensional rows of $\boldsymbol{G}$ and columns of $\boldsymbol{H}$, respectively. In particular, the $\boldsymbol{g}_s'$ signify the observations and the $\boldsymbol{h}$'s the variables. Thus, in a result biplot a simultaneous representation of data and variables is obtained. For $r = 2$, the biplot can be displayed as a two-dimensional scatterplot of $n + p$ points (n = number of individuals, p = number of variables). Since

$r = \text{rank}(\boldsymbol{X})$, the biplot is only an approximation if $r > 2$. For $r = 3$, the biplot can be displayed as a three-dimensional scatterplot of $n + p$ points. Again, the biplot is only an approximation if $r > 3$.

Biplots give us the possibility of simultaneous graphical representation of both observations (usually marked as points) and variables (usually marked as vectors). These graphs are the scatter plots with the superimposition of the variables. Thus the relationships between the data and variables can be investigated.

Matrices of ranks higher than two cannot be represented exactly by a biplot. However, if a matrix can be acceptably approximated by a rank two matrix, the biplot may allow useful approximate visual inspection of a given matrix. To approximate any rectangular *nxp* matrix of rank r by a *nxp* matrix of lower rank, one may use the singular value decomposition (Gabriel 1971). The goodness of fit of biplot is $\frac{\lambda_1,\ldots,\lambda_s}{\lambda_1\ldots,\lambda_n}$, where s is the number of chosen eigenvectors (the dimensionality of the representation. As we extract consecutive factors, they give an explanation for less and less variability. The decision of when to stop adding consecutive factors mainly depends on when there is only very small rest variability left. The character of this decision may be subjective; on the other hand, various strategies have been developed (Kaiser, 1966; Cattel, 1960). Some authors suggest to take the number of components that ensure obtaining the above value higher than 0.75. Other criterions are Cattel (1966) scree plot principle and Kaiser (1960) criterion (the number of eigenvalues λ_i bigger than 1). The most widely used Kaiser criterion means that if a factor does not extract at least as much as the equivalent of one original variable variability, we drop it.

Biplots have been developed and applied for few decades (Greenacre, 2010; Gower, 2003; Gower at. al 1996, 2010 a,b). An overview of generalizations of the classical linear biplots (based on principal component analysis) is given by Krzanowski (1995). One of the special cases of this generalization is the correspondence analysis (Krzanowski 1995 chap. 12). Biplots may be applied to the principal component analysis (PCA) and the canonical correlation analysis (CCA). Robust biplots have been elaborated by Daigle (2008). In generalized biplots (Gower 1992), both continuous and categorical variables are permitted. The special cases of generalized biplots are classical linear biplots, nonlinear biplots (where the linear axes are replaced by nonlinear trajectories) and, for categorical variables, a new case for multiple correspondence analysis.

Biplots may be also examined in the correspondence analysis (for not measurable data) (Greenacre 1993, Demey et. al. 2003). For classification tasks biplots are also very useful. Interpolative biplot was proposed by Alves

and Oliveira (2003). They showed that while predictive biplots are the best option for interpretation purposes, interpolative biplots are very helpful for classification of new cases (instances) that were not applied for the construction of the principal component or canonical dimension axes. Biplots were generalized also for usage in discriminant analysis (Gardner S. and le Roux 2005). Besides, modern tasks coming from micrarray data sets, biplots are also used, though high dimensionality is met in those problems (Demey et.al 2008). Improved biplot techniques ("better biplots" – Blasius, et. al 2009) are useful when there are many points (perhaps several thousand) and the entire graphical effect of typical biplot can be very confusing. In such situation (Blasius, et. al 2009) propose a number of procedures. For example, the density representation of the points may be applied. Choosing more than one centre of concurrency or the use of colour is also useful. Another possibility is, while respecting the calibrations, moving the axes to new positions more distant from the points, and possibly jointly rotating axes and points. Graphical aspects of biplots in classical and advanced biplot techniques are provided by Gower (2003, 2004).

Biplots may be also performed for higher dimensional representation by only two or three axes, however for graphical interpretation usually two or three dimensional plots are performed (if two or three dimensions represent sufficient variability).

The classical biplot technique makes no provision for missing values on any of the variables to be analyzed. The problem with missing values can be solved by deleting a not complete variable or case or by imputing missing values by mean, median or the value obtained by more elaborated special techniques, e.g. expectation-maximization or propensity score.

Other visualization methods are also applied for comparison purposes. Very close to PCA is the principal factor analysis. The term "factor analysis" includes both principal components and principal factors analysis. In factor analysis, the similarities between objects (e.g., variables) are expressed in the correlation matrix. Factor analysis requires that the underlying data is distributed as multivariate normal, and that the relationships are linear.

PCA and principal factor analysis serve different aims. The principal components analysis is rather a method dimension of dimensionality and principal factors and are often preferred when the goal of the analysis is aimed at the structure recognition. The purpose of PCS is to identify linear orthogonal linear combinations of variables, used for description or for substituting original variables by the smaller number of uncorrelated components. On the other hand, principal factors supply the model of the data and therefore are more complex. In principal components analysis it

is assumed that all variability is supposed be used in the study, whilst in principal factors analysis just the variability common with the other items is employed. Generally, the two methods from factor analysis domain as a rule supply very similar outcomes.

Another applied ordination method is the MDS – multidimensional metric scaling (Tab. 2). The aim of MDS procedure is to rearrange a map from a table of distances between observations points on the plane. Classical multidimensional scaling (Cox & Cox 1994) of a data matrix is also known as the principal coordinates analysis (Gower, 1966). Multidimensional scaling for a set of dissimilarities (or distances) yields returns a set of points such that the distances between the points are approximately equal to the dissimilarities (Digby 1986, Krishnaiah 1977, Mardia et. al 1079). MDS method applies a function minimization procedure that assesses different configurations with the objective of maximizing the goodness-of-fit. Goodness-of-fit statistic (also called "stress measure") based on primal and resulting distances checks how well the distances between objects can be reproduced by the new configuration. The difference with PCA is that in factor analysis, the similarities between objects are expressed in the correlation matrix. The axes from the MDS analysis are arbitrary, and can be rotated in any direction.

MDS identifies important primary dimensions that give the possibility to explain observed similarities or dissimilarities (distances) between the examined objects. With MDS, any kind of similarity or dissimilarity matrices can be considered, among them also the covariance matrix. MDS derives scatterplots of the objects in the different two-dimensional planes, though even three-dimensional MDS plots are possible. The attractiveness of MDS is that we can analyze any kind of distance or similarity matrix. Multidimensional scaling may be also applied for only rank-ordering of distances (or similarities) in the matrix (non-metric multidimensional scaling Mardia et. al 1979). Factor analysis often derive more dimensions than MDS and as a consequence, MDS often provides more clear, interpretable results. In MDS, only the distance is necessary for calculations, in factor analysis the origin data set is needed. Thus, MDS methods are appropriate to a wide range of examinations, because the distance measures can be constructed in many manners.

Cluster analysis

Grouping is often used in many scientific research domains, e.g. in taxonomy. Cluster analysis is a form of unsupervised learning (without the

a priori knowledge of groups). It is used to see if natural groupings are present in the data. A cluster is a group of objects that are similar to one another, and dissimilar to objects in other clusters. In cluster analysis (belonging to the set unsupervised classification procedures), in contrast to supervised classification, there is no target variable for clustering. Thus, cluster analysis can not result in estimate or prediction of the value of the target variable and therefore the cluster procedures' search for the division of the whole objects data is set into homogeneous objects subsets.

Cluster analysis (Hartigan 1975) is an exploratory data analysis technique which classifies (by different unsupervised classification methods) different objects into groups in a manner that the similarity between two clustered is maximal if they belong to the same group and minimal otherwise (Cutsem 1994). The division is conducted into mutually non–overlapping groups. Clustered variables show the similarities between variables. The clustering method applies the dissimilarities (distances) or similarities between objects when creating the clusters. Those measures are used to define the criteria for grouping or separating objects. Different dissimilarity or distance measures may be applied in the cluster analysis (Cutsem 1994), but usually the Euclidean, City-block (Manhattan) distance or Chebychev distance are used.

Cluster analysis is a set of methods of the data exploration. It can be the first step before applying other exploratory analysis procedures. For example, clustering is useful in the selection of best diagnostic features. Variables very close (in the same clusters) may suggest the reduction of features and in the results one may obtain reduced dimensionality. The clustering method has been used lately for gene expression clustering, where very large quantities of genes may exhibit similar behavior. This may reduce the analysis of many genes into important ones.

Hierarchical methods are most popular. If two groups are selected from different partitioning (at different levels) then either these groups are disjoint or one is included in the other. Hierarchical methods may be divided into agglomerative or divisive ones.

The assumptions of linearity and normality, though are often met in multidimensional methods, are not necessary in cluster analysis. However, the representativeness is needed and the collinearity may create unreal picture of clusters.

When the researcher expects that both observations and variables concurrently contribute to the recognition of an important configuration of clusters, the two-way clusters may be applied. In two-way clustering the similarities between different clusters of observations may be induced by

somewhat different subsets of variables. Thus, the resulting arrangement of clusters is by nature not homogeneous. This method offers a powerful exploratory data analysis tool and is often applied in solving bioinformatics problems where gene expression data sets in the so called "heat maps". Genes are clustered in one direction and on perpendicular axe. The studied cases are then grouped, which next can be considered according to the kind of diseases or prediction criterion.

It is worth noting that clustering is often the first step before the applying other, more advanced multidimensional methods from the exploratory data analysis domain, for example in supervised learning. In such case, the first step is clustering, which may be used for the selection or grouping of redundant variables.

Concluding remark

Different ordination methods have different advances and drawbacks and may have different possibilities such as inspecting the data set from different point of views

R E F E R E N C E S

[1] Alves M. R., Oliveira M. B. Interpolative biplots applied to principal component analysis and canonical correlation analysis. Journal of Chemometrics; Volume 17, Issue 11, pages 594–602, November 2003.

[2] Bartkowiak A. Liebhart J. Szustalewicz A. 1996. Visualizing the correlation structure by a biplot extended to 3 dimensions. 34th International Center of Biocybernetics. Seminar. Statistics and Clinical Practice. Warsaw, 24–28 June, 1996. pp. 50–52.

[3] Bartkowiak A. Lisp-Stat. Narzędzie eksploratywnej analizy danych. In Polish. Uniwersytet Wrocławski. 1995.

[4] Bartkowiak A. Szustalewicz A. 1998. Watchings steps of a grand tour implementations. Machine Graphics & Vision, 7(3), pp. 655–680.

[5] Blasius, J., Eilers, P. H. C., Gower, J. C. (2009) Better Biplots, CSDA, 53, 3145–3158.

[6] Cattel R. B. (1966) The scree test for the number of factors. Mutivariate Behavioral Research, 1, 245–276.

[7] Cuadras C. M, Rao C. R. 1993. Recent advances in biplot methodology. Multivariate analysis: future directions 2, 1993 – North Holland.

[8] Cox, T. F. and Cox, M. A. A. (1994) Multidimensional Scaling. Chapman and Hall.

[9] Cutsem B. 1994. Classification and Dissimilarity Analysis. Springer-Verlag, New York.

[10] Daigle G. 1992. A robust biplot. Canadian Journal of Statistics. Volume 20, Issue 3, pages 241–255.

[11] Demey, J. R., Vicente-Villardón, J. L., Galindo-Villardón, M. P. and Zambrano, A. Y. (2008). Identifying molecular markers associated with classification of genotypes by External Logistic Biplots.

[12] Digby P. G. N., Gower J. C. Ordination and classification. Le Presses de L'Universite de Montreal 1986.

[13] Gabriel K. R. 1971. The biplot graphics display of matrices with application to principal component analysis. Biometrika 1971 (58), 453–467.

[14] Gabriel K. R. 1981. Biplot Display of Multivariate Matrices for Inspection of Data and Diagnostics, Interpreting Multivariate Data, Ed. V. Barnett London: John Wiley & Sons.

[15] Gardner S., le Roux N.J. Extensions of Biplot Methodology to Discriminant Analysis.. Journal of Classification 22:59–86 (2005).

[16] Gower, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325–328.

[17] Gower J. C. and Harding S. (1988) Non-linear biplots. Biometrika, 73, 445–55.

[18] Gower J.C. (1990). Three-dimensional biplots. Biometrika 1990 77(4):773–785.

[19] Gower J.C. (1992) Generalized biplots. Biometrika 79, 475–493.

[20] Gower J. C. (2003). Unified biplot geometry. Ii: Developments in Applied Statistics. Eds: Ferligoj A, Mrvar A. Metodoloski zvezki, 19, Ljubljana: FDV.

[21] Gower, J. C. (2003) Visualisation in Multivariate and Multidimensional Data Analysis. Bulletin of the International Statistical Institute 54, 101–104.

[22] Gower J. C. (2004). The geometry of biplot scaling. Biometrika 91(3): 705–714.

[23] Gower, J. C., van der Velden, M. and Groenen, P. J. F. (2010). Area Biplots. Journal of Computational and Graphical Statistics, 19. 46–61.

[24] Gower, J. C. Hand, D. J. (1996) Biplots. London: Chapman and Hall.

[25] Gower, J. C., Lubbe, S. and LeRoux, N, D. J. (2010) Understanding Biplots. Chichester, John Wiley.

[26] Greenacre, M. (2010). Biplots in Practice. BBVA Foundation, Madrid, Spain.

[27] Greenacre, M. J. 1993. Biplots in correspondence analysis. Journal of Applied Statistics 20: 251–269.

[28] Hartigan J.A. 1975. Clustering Algorithms, Wiley, New York.

[29] Hill, M. O., 1974. Correspondence analysis: a neglected multivariate method. Appl. Statist. 23: 340–354.

[30] Hotelling, H., 1933. Analysis of a Complex of Statistical Variables Into Principal Components, Journal of Educational Psychology, volume 24, pages 417–441 and 498–520.

[31] Kao W. J., Sapatnekar S., Hiltner A., Anderson J. M.: Complement mediated leukocyte adhesion on poly(etherurethane-ureas) under shear stress in vitro. J. Biomed. Mater. Res. 1996, 32 (1): 99–109.

[32] Kaiser H. F. (1960) The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141–151.

[33] Krzanowski W. J., (1988). Principles of Multivariate Analysis, A User's Perspective. Oxford Univ. Press: Clarendon.

[34] Krishnaiah P.R. (ed.) 1977. Multivariate Analysis II. North Holland Vol 2. p. 595.

[35] Krzanowski W. J. (1995). Recent advances in descriptive multivariate analysis. Oxford University Press, New York 1995.

[36] Lane D. A., Bowry S. K. The scientific basis for selection of measures of thrombogenicity. Nephrol. Dial. Transplant. 1994, 9: 18–28.

[37] Larose D. T. 2005. Discovering Knowledge In Data. An Introduction to Data Mining. Wiley and Sons.

[38] Lim F., Yang C. Z., Cooper S. L.: Synthesis, characterization and ex vivo evaluation of polydimethylsiloxane polyurea urethanes. Biomaterials 1994, 15 (6): 408–416.

[39] Mardia K. V., Kent J. T., Bobby J. M. (1979). Multivariate Analysis, Academic Press, London.

[40] Rao C. R. (1973). Linear Statistical Inference and Its Applications, Wiley, New York.

Małgorzata Ćwiklińska-Jurkowska
Department of Theoretical Backgrounds of Biomedical Science
and Medical Informatics, Collegium Medicum in Bydgoszcz,
Nicolaus Copernicus University

EXPLORATORY DATA ANALYSIS FOR THE HEMATOLOGICAL FEATURES. PART II. APPLICATION

Abstract: Part II contains numerical and graphical exploratory analysis results and elements of their medical interpretation for the set of the hematological observations and hematological variables. On the basis of correlation matices, matrix of scatterplots with overlaid regression lines and two and three-dimensional biplots relationships between parameters of blood during hemoperfusion is examined. For comparison purposes also MDS and one-way and two-way cluster analysis are performed. Usefulness of applied methods of multivariate data ordination to inspect, e.g. variables' interdependencies was assessed. Applied methods gave very close results and the medical interpretation of the results confirm some physiological clotting ideas. The practical results confirm some hypothesis describing polymer-blood interactions. Additionally, the results of principal factor analysis and multidimensional metric scaling with cluster analysis are concordant. The variety of applied exploration data methods confirm results and give the possibility of looking at data from different point of views.

Introduction

The aim of the study was an investigation of the structure of the hematological data by examining some multidimensional exploratory data analysis methods such as scatterplot matrices and biplots (Gabriel 1971, 1981, 1990, Krzanowski 1988, 1995) and exploratory data analysis like factor analysis (mainly by PCA), MDS and cluster analysis.

A scatterplot matrix enables examination of relations between pairs of variables by graphical representation of a data matrix, while classical linear biplots are based on principal component analysis. Biplots are simply the scatter plots of multidimensional data into two or three dimensions with the superimposition of the variables. Thus, the relationship between the hematological data and variables can be investigated. The methodology of biplots and ither multivariate ordination methods is presented and the application on hematological data set is performed in Part I. "Methodolgy".

ISBN 978–83–7431–266–0 ISSN 0860-150X

Possible multivariate associations of numerical variables are examined in the work.

Aim

Exploring multivariate relationships between hematological data set is examined in the paper containing an interpretation from the medical point of view.

For drawing the plots and obtaining numerical results the program Statistica for Windows, SPPS and package R were used.

Illustration of multidimensional dependencies for hematological variables

The description of the examined hematological features is the following: "ADHEVIN" – Level of platelets' adhesion; "TIME" – time of perfusion ("CZAS"); "CZFIVIN" – Time of fibrinolise; "ERYTVIN" – Number of erythrocytes; "FIBGVIN" – Fibrinogen concentration; "KAOLVIN" – Kaolin-kefalin time; "KEFAVIN" – Kefalin time; "LEUKVIN" – Leukocytes number; "PROTVIN" – Prothrombin time; "STYPVIN" – Stypven-kefalin time and "TROMVIN" – Thrombocytes number.

Exploratory data analysis methods were applied (Bartkowiak et. al, 1996, Bartkowiak, 1995, Krzanowski W. J., 1988, 1995, Krishnaiah 1977, Larose D. T. 2005). Visualizing of interdependencies between variables describing parameters of human blood in experiments in vitro are presented in this section. The diagnostic tools for model of two-way tables are applied. Two kinds of exploratory data analysis are performed: firstly, scatterplot matrices and then biplots. Examining the data using scatterplot matrices is not in fact multidimensional analysis (only combing two-dimensional analysis), but can provide the insight into multidimensional data, if the dimension is reasonable.

The full matrix (Tab. 1) of correlations between the considered variables was calculated and the corresponding scatterplot matrix was obtained (Fig. 1). From Person correlations coefficients computed for all possible pairs of 11 variables we confirm the scatterplot matrix presented on Fig. 1. Looking at the obtained scatter matrices for all possible pairs of 11 variables we can verify that there is a positive significant ($p < 0.05$) Pearson correlation between the variables ADHEVIN and STYPVIN (0.41), TIME and PROTVIN (0.6), ERYTVIN and TROMVIN (0.77), FIBGVIN and TROMVIN (0.56), KAOLVIN and KEFAVIN (0.90), KAOLVIN and

Table 1

Linear correlation coefficients between pairs of variables

	ADHEVIN	TIME	CZFIVIN	ERYTVIN	FIBGVIN	KAOLVIN	KEFAVIN	LEUKVIN	PROTVIN	STYPVIN	TROMVIN
ADHEVIN		−0.0494	−0.0640	0.0056	0.2123	−0.3046	−0.0698	**−0.4893**	0.1382	**0.4099**	0.2254
TIME	−0.0494		0.0541	**−0.6061**	**−0.5317**	−0.1603	−0.2084	**−0.6659**	**0.5994**	0.0662	**−0.7435**
CZFIVIN	−0.0640	0.0541		−0.1619	0.0724	0.1764	0.2772	−0.0493	−0.2072	0.0082	−0.1210
ERYTVIN	0.0056	**−0.6061**	−0.1619		0.1501	**−0.4541**	**−0.4322**	0.1696	−0.2368	0.3114	**0.7685**
FIBGVIN	0.2123	**−0.5317**	0.0724	0.1501		0.1882	0.3240	0.2867	**−0.3966**	−0.1388	**0.5610**
KAOLVIN	−0.3046	−0.1603	0.1764	**−0.4541**	0.1882		**0.8992**	**0.4616**	**−0.4732**	−0.1933	**−0.4380**
KEFAVIN	−0.0698	−0.2084	0.2772	**−0.4322**	0.3240	**0.8992**		0.293	**−0.5567**	−0.1996	−0.3174
LEUKVIN	**−0.4893**	**−0.6659**	−0.0493	0.1696	0.2867	0.4616	0.2930		−0.3160	**−0.5511**	0.2289
PROTVIN	0.1382	0.5994	−0.2072	−0.2368	**−0.3966**	**−0.4732**	**−0.5567**	−0.316		−0.1274	−0.3574
STYPVIN	**0.4099**	0.0662	0.0082	0.3114	−0.1388	−0.1933	−0.1996	**−0.5511**	−0.1274		0.1342
TROMVIN	0.2254	**−0.7435**	−0.1210	**0.7685**	**0.5610**	**−0.4380**	−0.3174	0.2289	−0.3574	0.1342	

In bold: typed significant correlations on level 0. 05

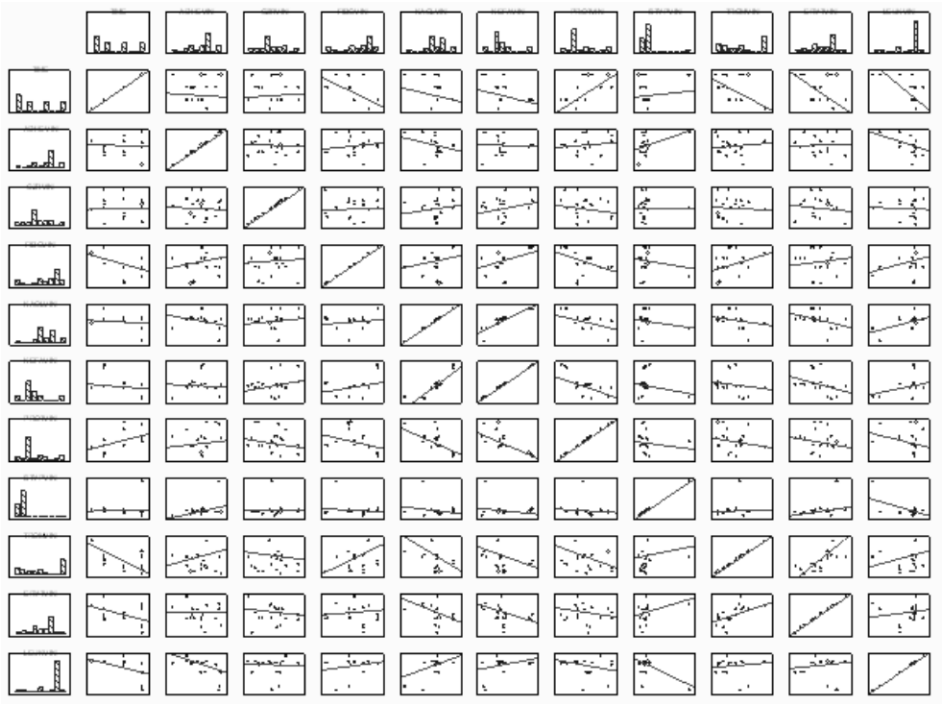


Fig. 1. Scatterplot matrix for all of variables' pairs with overlaid regression lines

LEUKVIN (0.46). Additionally there is a negative significant ($p < 0.05$) correlation between the pairs of variables ADHEVIN and LEUKVIN (– 0.49), CZAS and ERYTVIN (– 0.61), CZAS and FIBGVIN (– 0.53), TIME and LEUKVIN (– 0.67), CZAS and TROMVIN (– 0.74), ERYTVIN I KAOLVIN (– 0.45), ERYTVIN and KEFAVIN (– 0.4322), FIBGVIN and PROTVIN (– 0.40), KAOLVIN and PROTVIN (– 0.47), KAOLVIN and TROMVIN (– 0.44), KEFALVIN and PROTVIN (– 0.56), LEUKVIN and STYPVIN (– 0.55), LEUKVIN and PROTVIN (– 0.32) and finally KEFAVIN and PROTVIN (– 0.56).

The problems of in-vitro blood procedures are generally caused by changes of thrombocytes in time (TIME significantly negatively correlated with TROMVIN, $r = -0.74$), so the row connected with variable TROMVIN (number of thrombocytes) is most interesting. The highest correlated pair is visible in scatterplot matrix (TROMVIN and ERYTVIN, $r = 0.77$).

For pairs of variables correlated significantly positively (ADHEVIN and STYPVIN, TIME and PROTVIN, ERYTVIN and TROMVIN, FIBGVIN and TROMVIN, KAOLVIN and KEFAVIN and KAOLVIN with LEUKVIN) some of these relationships are self-evident. However, the other can find explanation from a point of view of clotting physiology. For example, they suggest an activation of clotting factors during the experiment (variables KAOLVIN and KEFAVIN), activation of clotting factors (variables KAOLVIN and KEFAVIN), consumption of fibrinogen (FIBGVIN) and segmentation of blood cells on polymer (variables TROMVIN and ERYTVIN). Significant positive correlation of a durations of experiment (TIME) and prothrombin time (PROTVIN) perhaps shows on catching on sorbent surface some clotting factors, active in extrinsic clotting pathway. Universally, one thinks that the prothrombin time changes insignificantly during a contact of blood with polymers, but by reason of considerable development of sorbent surface these interactions can be more clear. It can point to the advisability of taking into account the prothrombin time in estimation of hemocompatibility of polymers (foreign surfaces), which are planned to apply in clinical practice.

Next, for the following different pairs correlated significantly negatively (ADHEVIN and LEUKVIN, TIME and ERYTVIN, TIME and FIBGVIN, TIME And LEUKVIN, TIME and TROMVIN, ERYTVIN and KAOLVIN, ERYTVIN and KEFAVIN, FIBGVIN and PROTVIN, KAOLVIN and PROTVIN, KAOLVIN and TROMVIN, KEFALVIN and PROTVIN, LEUKVIN and STYPVIN and KEFAVIN with PROTVIN) some explanation from the physiological clotting point of view can be given. Those results for the times assessing intrinsic system of clotting (KAOLVIN and KEFAVIN) may point

to the activation of clotting factors, taking participation in the beginning of the activation. Simultaneous with the activation decreases also the number of plates of blood (thrombocytes)-TROMVIN correlated significantly negative with KAOLVIN).

Obtained results and the medical interpretation confirm some hypothesizes about interactions of polymeric sorbent with blood (Kao W. J et. al 1996, Lane et. al 1994, Lim et. al 1994).

Examining of the structure of the data by factor analysis

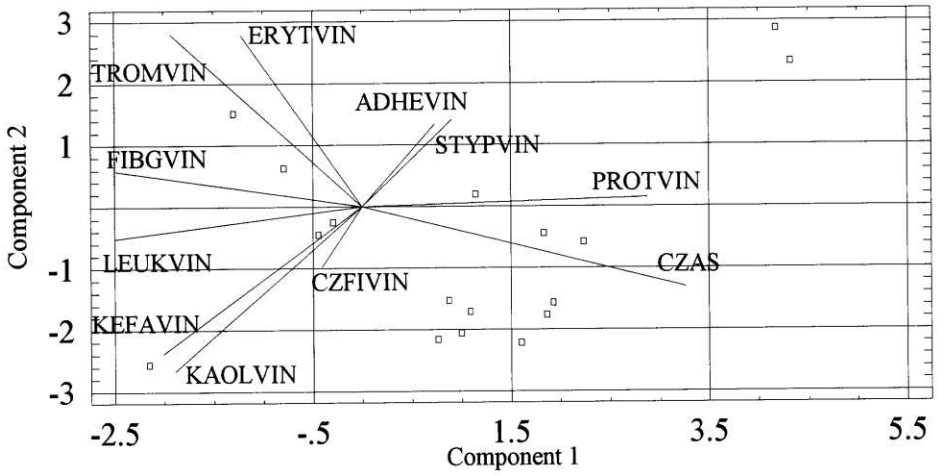


Fig. 2. Two-dimensional biplot for 11 variables

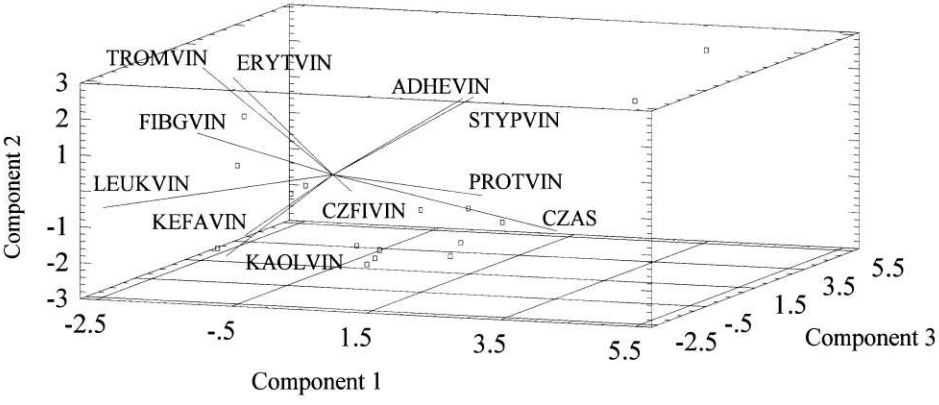


Fig. 3. Three-dimensional biplot for 11 variables, truly reproducing surface, on of which a two-dimensional biplot is presented

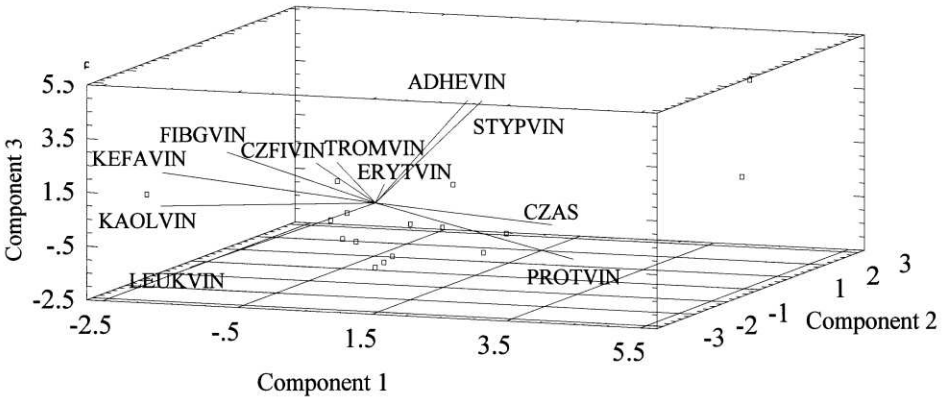


Fig. 4. Eigenectors-principal component coefficients for 11 original variables

The statistical graphics of biplots represent in the same plane both the variables and the cases. Eleven or six variables are represented by arrows (lines), while points represent cases. See the examples in Fig. 2–7. Graphs show visual inspection of main information from the data set. Presented at the biplots with principal components ((Fig. 2–7) vectors (visible as segments from the centroid) represent the original variables. The length of each vector is proportional to the contribution of the corresponding variable in the principal components. The angle between any two vectors is closely related to the correlation between presented variables. Cosine of this angle is a correlation in the case of total variance representation on biplot – then a high positive correlation is achieved for small angles (close to 0 degrees) and the high negative correlation, for angles between the vectors representing the variables, which are close to 180 degrees. Based on this angle, you can draw reliable conclusions about the correlation, if the original variables are well represented on the biplot (goodness of fit). If the vectors are short, they can not be applied to conclude about the correlations (Bartkowiak 1995).

In Fig. 2 and Fig. 5 the two-dimensional biplots constructed from the examined variables are presented. However, the biplot shown in Fig. 2 is based on all variables (11); the biplot shown in Fig. 5 was constructed using 6 chosen variables. Both groups of biplots were constructed from correlation matrices. Commonly, the points and the vectors in the biplot plane represent projections from the multivariate space onto the plane of the first two or three principal components. Points at biplots mean the individual observation boxes ($n = 32$), some of them overlap. From biplots for all 11 variables (Fig. 2–4 and Tab. 1) it can be noted that many subsets of variables

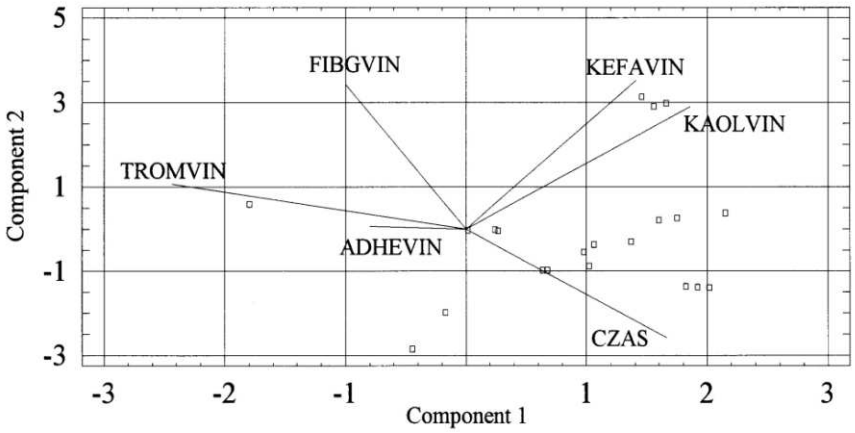


Fig. 5. Two-dimensional biplot for 6 selected variables

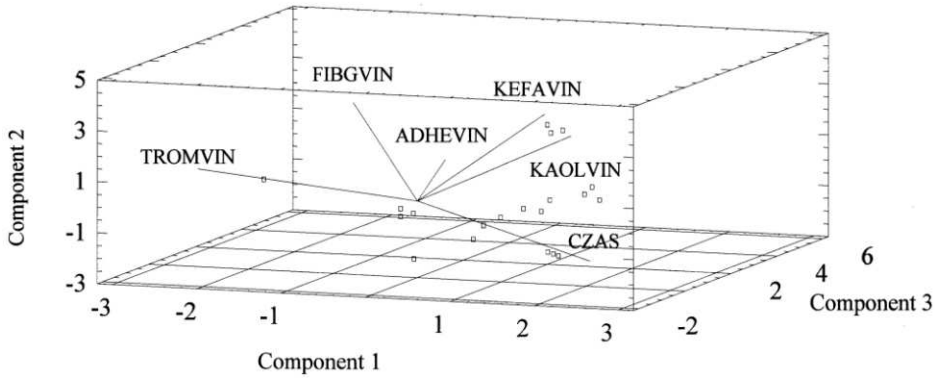


Fig. 6. Three-dimensional biplot for 6 selected variables, truly reproducing surface, on which two-dimensional biplot is presented

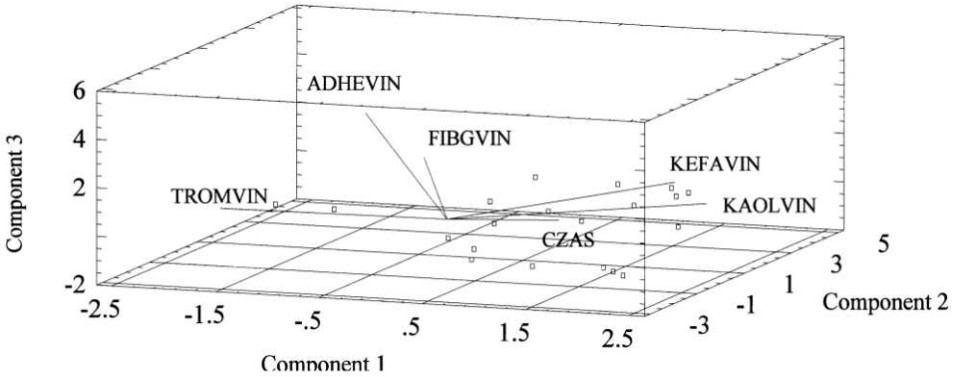


Fig. 7. Three-dimensional biplot for 6 variables, presenting deviation of vectors from plane of two first principal components

Table 2

Principal component analysis – set of 11 original variables

Number of principal component	Eigenvalue	Total variance percentage	Cumulated variance percentage
1	3.39581	30.871	30.871
2	3.06543	27.868	58.738
3	1.66496	15.136	**73.875**
4	1.04753	9.523	83.397
5	.88566	8.051	91.449
6	.42797	3.891	95.340
7	.29461	2.678	98.018
8	.15528	1.412	99.430
9	.04106	.373	99.803
10	.01675	.152	99.955
11	.00495	.45	100.000

Table 3

Eigenvectors-principal component coefficients for 11 original variables.

Variables	Eigenvectors			
	1	2	3	4
ADHEVIN	.106616	.22402	.50364	.49271
TIME	.474063*	–.22017	.03030	–.01168
CZFIVIN	–.0582577	–.16444	.28605	–.32623
ERYTVIN	–.176689	.46696*	–.11107	–.33431
FIBGVIN	–.361119	.09811	.20069	.48914
KAOLVIN	–.2748	–.44697*	.13184	–.05471
KEFAVIN	–.292155	–.39982*	.29908	.07581
LEUKVIN	–.408985*	–.09740	–.42429	.01627
PROTVIN	.417988*	.02436	–.27440	.33963
STYPVIN	.131751	.23607	.49885	–.41092
TROMVIN	–.280475	.47030*	–.00180	.08017

Variables with biggest contribution to eigenvectors are marked by asterisks

are highly correlated with each other. The question arises whether you can choose a smaller and more representative set of characteristics, perhaps dropping the features strongly correlated. In many publications as the most important promoter of hemocompatibility (compatibility with the blood) the effect on platelets is given. Therefore, as the first into a created subset

Table 4

Principal component analysis – set of 6 selected variables

Number of principal component	Eigenvalue	Total variance percentage	Cumulated variance percentage
1	2.37832	39.639	39.639
2	2.00856	33.476	**73.115**
3	1.02162	17.027	**90.142**
4	.438124	7.302	97.444
5	.107607	1.793	99.237
6	.045767	.763	100.00

Table 5

Eigenvectors-principal component coefficients for set of 6 selected variables

Variables	Eigenvectors		
	1	2	3
ADHEVIN	–.202265	.0105427	.915152
TIME	.418215	–.405896	.316225
FIBGVIN	–.251145	.538354*	.191984
KAOLVIN	.466875*	.457417*	–.023692
KEFAVIN	.354357	.554912*	.087172
TROMVIN	–.614468*	.167794	–.132212

Variables with biggest contribution to eigenvectors are marked by asterisks

of variables, the variable TROMVIN (thrombocytes) was selected. Next, the set has additionally been extended by five other features which found a relatively high representation at biplots (Table 4) – the cumulative percentage of variance (compared to the other six sets of original variables which contain variable TROMVIN). So the analysis of principal components and also for those biplots based on subset of six features was developed (Tab. 4, 5, Fig. 5–7). Biplots if Fig. 5–7 represent both all observations and variables subset containing 6 original features subset of a full multivariate data set on the same plot.

In the column "*Eigenvalue*" in Tables 2 and 4, the eigenvalues equal to the variances on the consecutive factors may be obtained. In the second column "Total variance percentage ", these values are expressed as a percent of the total variance (sum of eigenvectors). As we can notice from Tab. 2,

factor 1 accounts for 31% of the variability, factor 2 for 28%, factor 3 for 15% and so on. The third column consists of the cumulative variability extracted.

From Tab. 4 we can see that the first axis for 6 original variables explains 36% of variability of the whole multidimensional data set.

Analyzing the eigenvalues in Tab. 2 it can be stated that the presentation of data on a matrix of the first three principal components reproduces 73% of the total variation of the original six variables. Further, in Tab. 4 can see the better (in comparison to Tab. 2) result – 90% of the variability of representation after a three dimensional projection of the whole multidimensional data set into principal component subspace. It can be explained by a lower dimensionality of the reduced input data set (maybe some information is lost by dropping a number of original variables, from eleven to six-which almost half of them). From the tables of coefficients for principal components (Tab. 2 and 4) it can be easily determined which features have the most influence on another principal component. This is reflected in the illustration in the appropriate biplot (Fig. 2–7).

Scatterplot in the Fig. 5 is a projected multivariate scatter onto a plane but on the subset of 6 variables. The lines represent six original variables. Each case is represented by one point (individual data point) on the same principal variables axes.

For both 11 original variables and 6 original variables the scree Cattel criterion coming from showing subsequent eigenvalues indicate to bigger number of dimensionality than the Kaiser criterion, i.e. the number of eigenvalues bigger than 1 (which is equal to four in the case of 11 input variables and three in the case of 6 input variables).

According to the Kaiser criterion we an see that 4 variables have eigenvalues bigger than 1.

The approximation of the biplot variables is given by the biplot axes. The vectors labeled by names represent the considered variables. Table 4 shows the equations of the principal components. The most important variables are marked by characters "*". For example, the first principal component has the following equation

$$0.106616\ \text{ADHEVIN} + 0.474063\ \text{CZAS} - 0.0582577\ \text{CZFIVIN} - 0.176689\ \text{ERYTVIN} - 0.361119\ \text{FIBGVIN} - 0.2748\ \text{KAOLVIN} - 0.292155\ \text{KEFAVIN} - 0.408985\ \text{LEUKVIN} + 0.417988\ \text{PROTVIN} + 0.131751\ \text{STYPVIN} - 0.280475\ \text{TROMVIN}$$

where the values of the variables in the equation are standardized by subtracting their means and dividing by their standard deviations. For the principal components based on the set with all eleven variables the first

principal component can be defined as related to the activation of the extrinsic system. Important contribution to the value of this component also contends that the number of platelets, leukocytes and blood contact time with the sorbent. The second principal component is related to the activity and the number of thrombocytes and coagulation parameters intrinsic (Tab. 3).

Table 5 shows the equations of the principal components. For example, the first principal component has the following equation

$$0.202265\ \text{ADHEVIN} + 0.418215\ \text{CZAS} - 0.251145\ \text{FIBGVIN} + 0.466875\ \text{KAOLVIN} + 0.354357\ \text{KEFAVIN} - 0.614468\ \text{TROMVIN}.$$

However, in the case of 6 selected features, the above first principal component is associated with the number of platelets, the other components of the plasma coagulation. In this case, for both first principals the important component of variability is the time of the experiment (Table 5). Principal component analysis confirms the suitability of the number of platelets as an important parameter to measure the impact of polymer on the blood. The different signs in the equation (signs of coefficients of loadings) mean distinct contribution into the principal component. It is worth noting a difference for biplots (either two or three-dimensional and) in a configuration relative to each other vectors representing eleven features and only six selected variables characteristics. It is of course the fact that both principal component analyses are taken into account other data matrix X: respectively *nx6* and *nx11* matrix ($n = 32$). Thus biplots from Fig. 5–7 contain only some columns of the matrix *nx11*. In addition to the configuration vectors (features), also the configuration of points (observations) on biplots for 11 variables (Fig. 2–4) and six variables (Fig. 5–7) are different, since they reproduce on biplots significantly different percentage of variability (Table 2 and 4).

Comparing pairs of correlated variables (Tab. 1) with biplots for all 11 variables, an adequacy of biplots for these variables to a certain extent can be observed, though only on a three-dimensional biplot the total variation is represented satisfactorily, in 74% (Table 3). The adequacy of the correlation matrix is more obvious for 6 variables, because in this case reconstruction of variability on biplots is 73% and 90% for two-dimensional and for three dimensional, respectively (Tab. 4). It is worth noting that in the case of a large percentage of the representing of the of variation (like fig. 5–6), a large absolute value of negative correlation is related to the angle between vectors denoting variables close to 180 degrees, and the large positive correlation value means the angle close to 0 degree.

Comparing the biplot 2 and 3-dimensional for the 11 variables it may be observed that the worst representation in the two-dimensional biplot has the CZFIVIN variable (Fig. 2 and 3), and for the six variables has ADHEVIN feature (Fig. 5 and 6). These vectors' characteristics significantly deviate from the plane of the first two principal components. It is seen by selecting in displaying the biplot the third major component parallel to the edge of the graph (Fig. 4 and 7). This is confirmed by the analysis of the third column in Tables 3 and 5. In these tables the high coefficients are found for the relevant variables in comparison with the coefficients in the same row for the primary and last components (first and second column compared with third and/or fourth).

Despite the small angle between variables and TROMVIN ADHEVIN on biplot for the original six variables (Fig. 5) and a small angle between variables and KAOLVIN CZFIVIN biplot for the original 11 variables (Fig. 2), correlations between these variables are not large ($r = 0.23$ and 0.18 respectively), because – as stated above – the representation of the CZFIVIN and ADHEVIN variables is weak. More generally, for biplot devised for standardized features, namely the correlation matrix, no correlations can be inferred if the vectors' features are shorter (Bartkowiak 1995).

Comparing of both cases (i.e. for 11 and 6 variables) on two and three-dimensional biplots (Fig. 2 and 3 and 5 and 6), the closer representation of actual data matrix X of a three-dimensional biplots by the two-dimensional biplots on the plane it can be held for six variables. The diminishing of variability representations from matrix for 11 and 6 variables is 15% and 17%, respectively.

In plot on Fig. 8 we can see which variables have smaller length and therefore are nor well represented in the plane (e.g. CZFVIN, ADHEVIN, STYPVIN).

It is interesting if inspecting the biplots and the interpretation of factor analysis results give comparable results. Other multivariate data ordination, namely principal factor analysis is also performed. The principal factor analysis for the whole input data set is presented graphically in Fig. 8. In the plot of factor loadings in Fig. 8 eleven variables were reduced to two specific factors. Fig. 9 shows similar results with the nonessential difference, coming from the arrangement of points. Next, Fig. 10 is obtained with Varimax rotation – difference with Fig. 10 is caused by the rotation. Additionally, the variables except the points are visualized in comparison with biplots. Corresponding Cattel scree plot is visible in Fig. 11. In graphical method for the *scree* Cattell test we can see the eigenvalues shown in a simple line plot, the values are numerically presented also in the first column of Tab. 2.

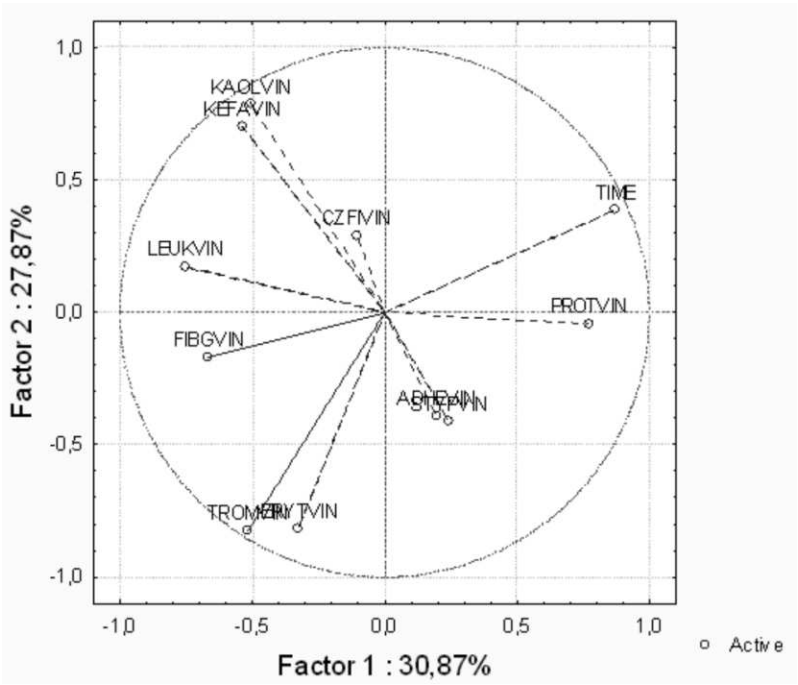


Fig. 8. Projection of variables on Factor1xFactor2 plane

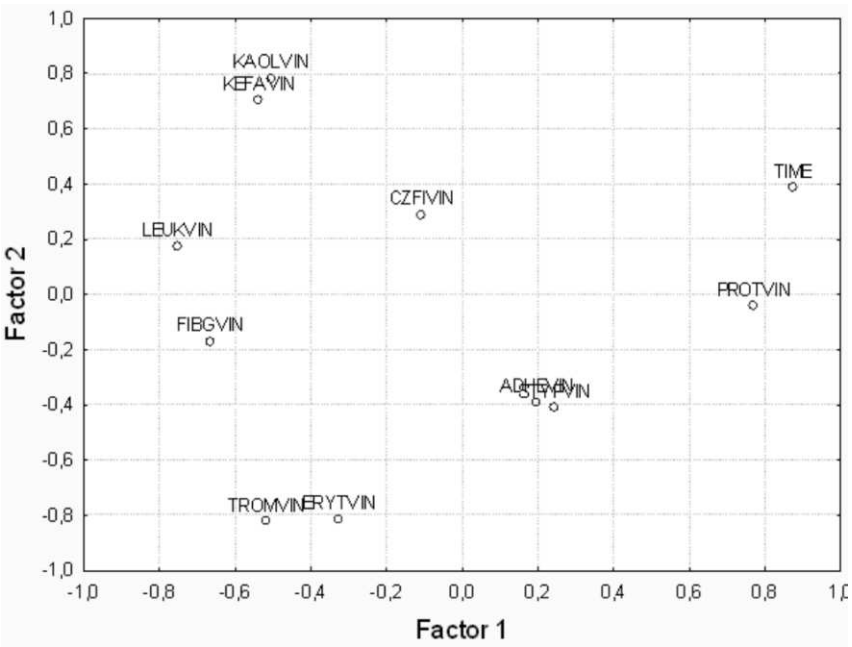


Fig. 9. Principal components without rotation

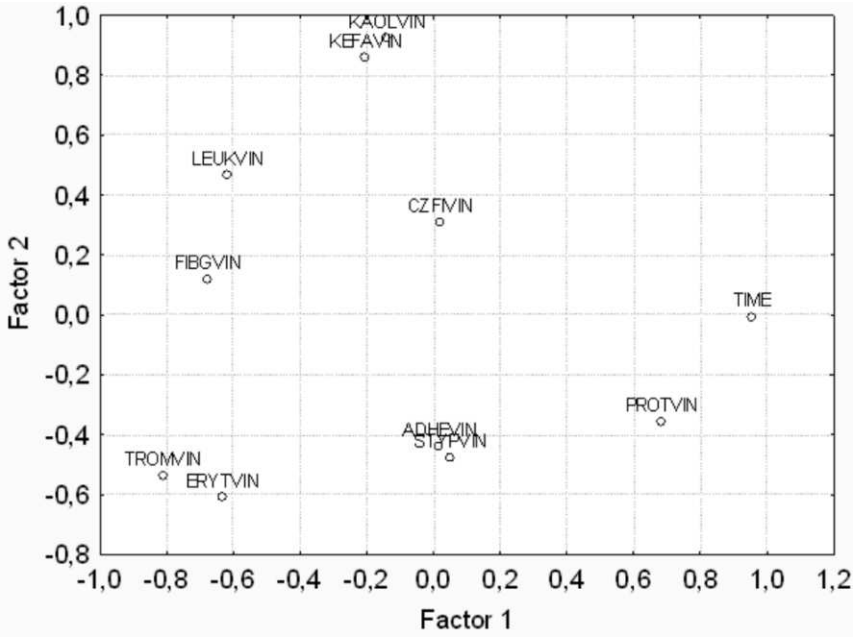


Fig. 10. Principal components with Varimax rotation

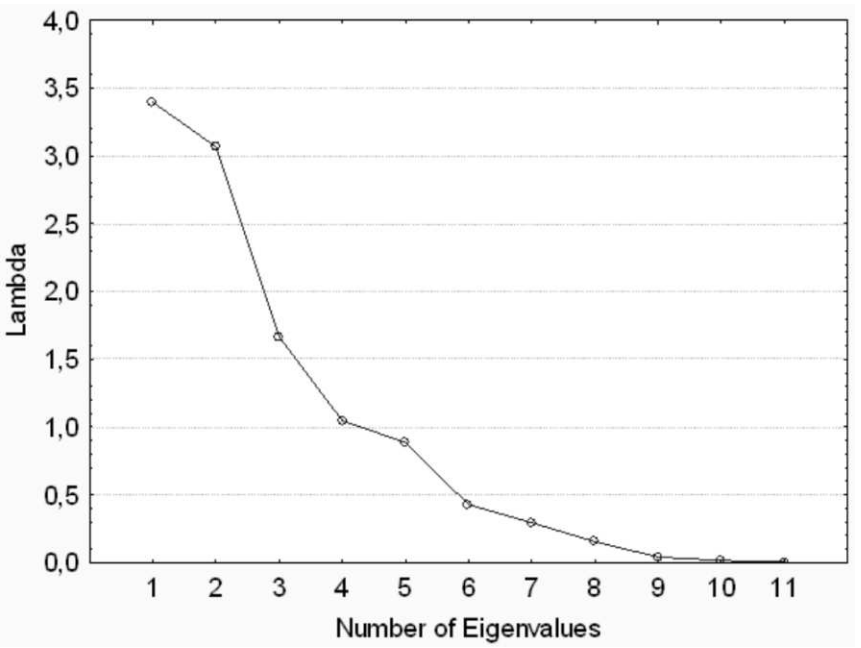


Fig. 11. Cattel scree plot

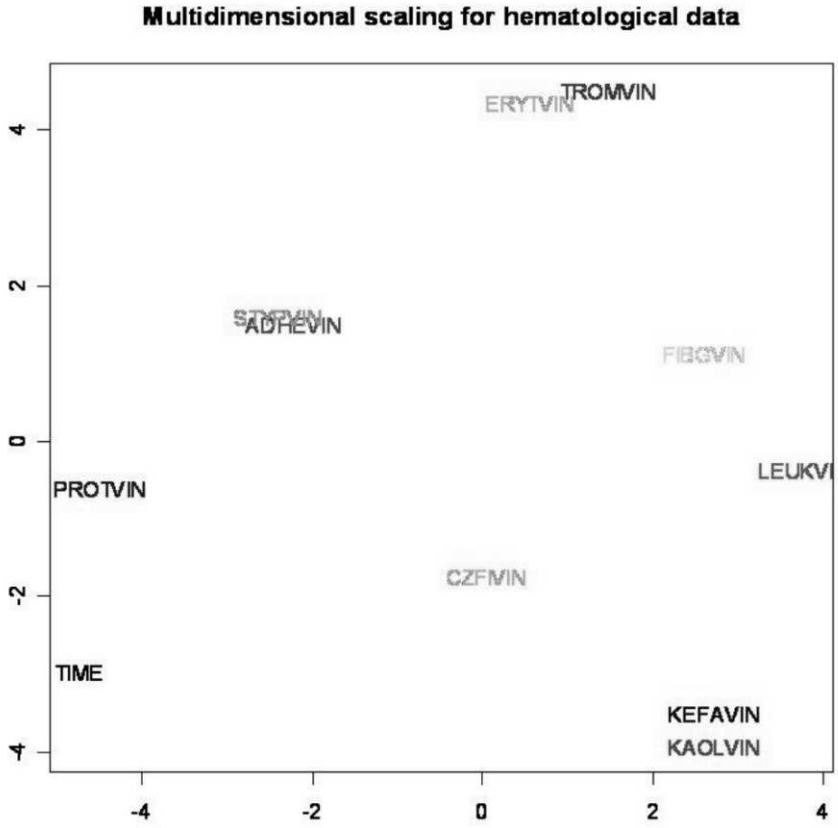


Fig. 12. MDS for all 11 variables

The Cattel scree plot showing subsequent eigenvalues indicate to bigger number of dimensionality representation (equal to 6) than the Kaiser criterion, i.e. the number of eigenvalues bigger than 1 (which is equal to 3).

Another multidimensional data ordination method is visible in Fig. 12 and 13. It is the multidimensional metric scaling result for the Euclidean distance.

In Fig. 12 the "rearranged" hematological features in a proficient way are presented. The obtained configuration best approximates the observed Euclidean distances. For other applied distances or for correlation dissimilarity measure the MDS results (not presented in the paper) are nearly the same as for the Euclidean distance. The diminished set of 11 original variables into its subset of 6 variables, the same as in Tab. 4 and 5 and on Fig. 5–7, are arranged by MDS into the plot in Fig. 13.

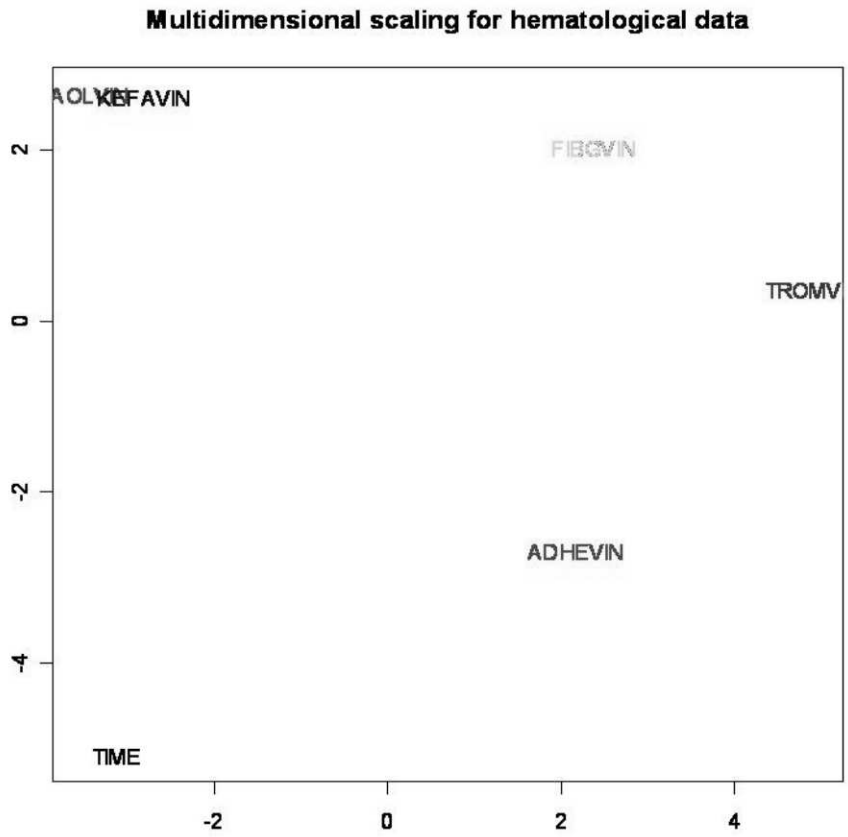


Fig. 13. MDS for 6 variables: "TIME" "ADHEVIN" "FIBGVIN" "KAOLVIN" "KEFAVIN" "TROMVIN"

The arrangement of variables in MDS for 11 and 6 variables is very close to those obtained earlier and described above. MDS is an alternative to factor analysis. However, MDS and factor analysis are basically different methods, though the type of research tasks to which these two techniques can be applied are similar. For example, MDS does not require normality and linearity such restrictions. Moreover, MDS can be applied to any kind of distances or similarities, while factor analysis is based on the covariance matrix.

To discover structures on the basis of distances between variables, grouping of records into groups of similar objects is performed by cluster analysis. In the obtained subsets (in clusters) the similarity of the records is maximized and the similarity of the records in other clusters is minimized.

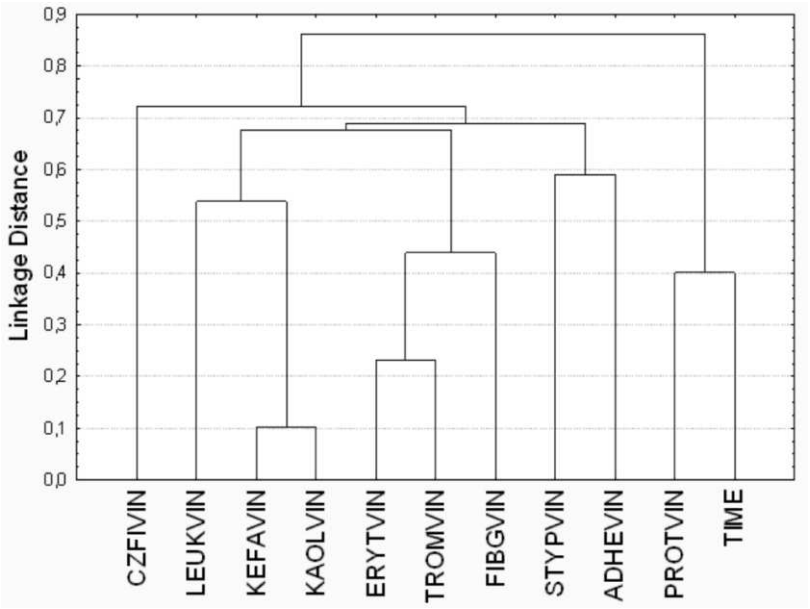


Fig. 14. Hierarchical single linkage clustering by Euclidean distance for whole data set

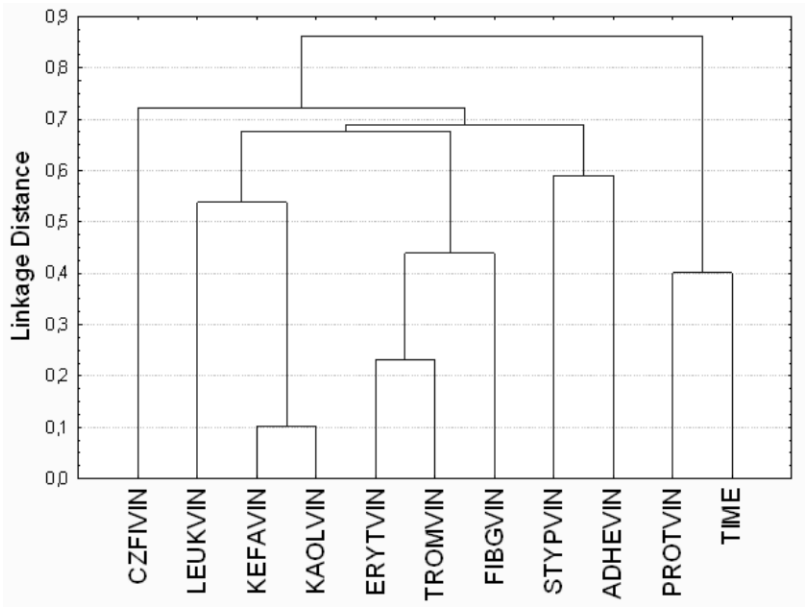


Fig. 15. Hierarchical single linkage clustering by Pearson coefficient for whole data set

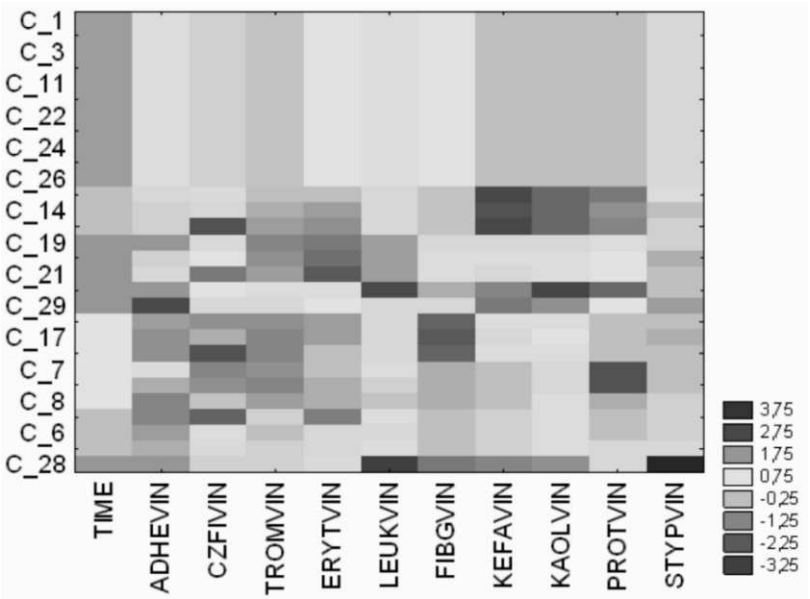


Fig. 16. Cluster two-way joining for hematological features and observations

The applied hierarchical cluster analysis for complete linkage and Euclidean distance is presented on Fig. 11. Other distances gave the same results (for example dissimilarity equal to Person correlation coefficient – Fig. 12). The clustering of variables shows the similarity to other multidimensional reduction methods results. Clusters for two-way joining are presented in a "heat map" (Fig. 13). The dissimilarity of variables is close – the columns in two-way joining plot are strictly related to the tree coming from one-way hierarchical clustering.

Different multivariate methods applied, like the factor analysis (PCA and principal factors), correlations, MDS and clusters gave similar or complementary results.

In all applied methods a low-dimensional graphic representation of the hematological data set is obtained. Relationships obtained by the different ordination methods confirm the dependencies obtained by each other with only subtle differences.

Final remarks

In biomedical problems the question arises how to categorize observed data into meaningful structures. Generally, medical problems are characte-

rized by high complexity. When one wants to describe medical phenomenon, large number of variables is needed. However, with high dimensional tasks the interpretation is difficult. Therefore, the reduction of the dimensionality is needed. Two groups of dimensionality exists (selection or extraction), here we applied the extraction reduction of dimensionality. Grouping of the variables is obtained by similar methods of PCA and factor components methods. If, after dimensionality reduction, one has obtained only two or three dimensions, a physician can interpret the problem using appropriate illustration.

In exploratory data analysis we can examine higher dimensional data sets, where the relationships or trends are difficult to see. The insight into multidimensional data is possible looking at the same time at many graphs using scatteplot technique. However, for bigger number of variables constituent plots are smaller. For p-dimensional data sets one can not simply visualize the whole information coming from data, so the need of representative ordination is needed. The possibility of analyzing relationships between many variables on only one plot gives the two or three-dimensional biplot technique. However, then the representative features in the reduction of dimensionality connected with the eigenvalues of the matrix in principal component is needed.

The research question of interest is usually expressed in terms of both cases (observations) and variables. For example, the biplots or two-way clusters may be applied. Applied explanatory data analyses methods can discover structures in data, however this does not automatically supply an explanation or interpretation.

Conclusion

The relationship between the hematological data and variables can be investigated by graphic representation: scatterplot matrices and biplots. Two-dimensional or three-dimensional biplots is a method of dimensionality reduction giving the possibility of observing simultaneously both variables and observations, so observations may be also visualized in the context of many variables. Useful maps are obtained by multidimensional metric scaling, which is also the ordination multivariate data method, which do not need the assumptions related to PCA. The results confirm some hypothesis describing polymer-blood interactions and may suggest new unknown facts.

The results of factor analysis, cluster analysis and two-way clusters and multidimensional scaling are all concordant.

REFERENCES

[1] Bartkowiak A. Liebhart J. Szustalewicz A. 1996. Visualizing the correlation structure by a biplot extended to 3 dimensions. 34th International Center of Biocybernetics. Seminar. Statistics and Clinical Practice. Warsaw, 24–28 June, 1996. pp. 50–52.

[2] Bartkowiak A. Lisp-Stat. Narzędzie eksploratywnej analizy danych. In Polish. Uniwersytet Wrocławski. 1995.

[3] Krzanowski W. J., (1988). Principles of Multivariate Analysis, A User's Perspective. Oxford Univ. Press: Clarendon.

[4] Krzanowski W. J. (1995). Recent advances in descriptive multivariate analysis. Oxford University Press, New York 1995.

[5] Krishnaiah P. R. (ed.) 1977. Multivariate Analysis II. North Holland Vol 2. p. 595.

[6] Kao W. J., Sapatnekar S., Hiltner A., Anderson J. M.: Complement mediated leukocyte adhesion on poly(etherurethane-ureas) under shear stress in vitro. J. Biomed. Mater. Res. 1996, 32 (1): 99–109.

[7] Larose D. T. 2005. Discovering Knowledge In Data. An Introduction to Data Mining. Wiley and Sons.

[8] Lane D. A., Bowry S. K. The scientific basis for selection of measures of thrombogenicity. Nephrol. Dial. Transplant. 1994, 9: 18–28.

[9] Lim F., Yang C. Z., Cooper S. L.: Synthesis, characterization and ex vivo evaluation of polydimethylsiloxane polyurea urethanes. Biomaterials 1994, 15 (6): 408–416.

Anna Justyna Milewska
Robert Milewski
Urszula Górska
Dorota Jankowska
Department of Statistics and Medical Informatics,
Medical University of Bialystok

STATISTICAL METHODS IN POLISH MEDICAL PUBLICATIONS

Abstract: Conducting research in the field of medicine today requires knowledge of statistical tools. For various reasons their correct selection is often a difficult task. This paper summarizes the most commonly used statistical methods in Polish medical journals published in 2009. We studied whether the choice of statistical tools and the methods of their implementation is connected with the number of points awarded for particular journals by MNiSW.

Introduction

Currently there are almost 10 thousands titles on the bulleted list of journals of the Ministry of Science and Higher Education (MNiSW). Nearly 250 of them refer to medical aspects and are published in Poland. Only 16 of them are characterized by the Impact Factor indicator, that highlighting shall entail an evaluation of at least 10 points. Most journals are assigned 4 points (104), 2 points (67) and 6 points (45). The score is determined on the basis of the opinion of experts appointed for this purpose by the Minister of Science and Higher Education. Next to this classification there are many other lists evaluating the quality of journals. It is worth mentioning the ISI Master Journal List, where at the beginning of 2010 were 247 Polish titles, including 55 medical journals. One may encounter an incorrect belief that journals on this list are those that have Impact Factor, which is not the rule. However, it is a collection of the best journals in their field. It is observed that each year the number of Polish journals in this segment is growing [1]. An interesting list is also published on the Internet platform Index Copernicus. Journals which are included here are also subject to eva-

ISBN 978–83–7431–266–0 ISSN 0860-150X

luation, which also involves granting of points. We found 685 Polish entries, 481 of them being medical journals.

Scientific investigators should take into consideration the fact that they will eventually want to publish their findings. Unfortunately, not many take this into account and only when all the results of the statistical analysis are confirmed they note that the study could have been carried out otherwise. These problems may be associated with too small a sample, incorrectly chosen and unrepresentative sample, ambiguously worded questions, etc. Even before conducting research, the investigator must take into account what tools can used to analyze the results. One should begin by setting a specific purpose for which the research is to be used. The examination should be planned in most detail and clarity [2, 3]. The researcher should remember that "there are only a handful of ways to do a study properly but a thousand ways to do it wrong" [4]. It is clear that practically we are not able to determine the target group we are interested in. Therefore, we should choose a sample which will be a subject of the research. This group should be representative and selected at random. Information about each of its elements should be collected in the form prepared by us on the computer. It is noteworthy that even at this early stage of the examination we should use the knowledge of statistics [5]. Mistakes in this phase are difficult to fix later. Inexperienced researchers repeatedly formulate unclear questions. Sometimes a respondent does not give an answer to a question because he/she does not understand the idea of the author. This may lead to the rejection of the questionnaire from the analysis. If this situation occurs repeatedly and the question is important for the researcher, the examination becomes useless. When you select features that will be analyzed you should choose in what scale they will be described. This affects the choice of statistical tests used later. After preparing research and collecting information about the observation we proceed to develop statistical material. We organize gathered information using the statistical series, then the data are prepared for statistical analysis. Two groups of methods must be distinguished: descriptive statistics and statistical inference. Now, taking the used scale, sample size and other factors, the investigator has to chose an adequate tool from those available. This choice is virtually never obvious. Although the computer performs calculations, the researcher has to indicate the appropriate statistical methods. Interpretation of results also belongs to the person conducting the survey. At present, during analysis, an appropriate computer program is essential. Optimistic is the fact that access to commercial statistical packages is spreading, leading to increased standards of statistical studies [6].

However, lack of statistics knowledge makes useless even the best computer programs.

Authors of scientific publications present the results of their experimental work with more or less complex statistical procedures. Unfortunately, an average doctor, for whom mathematics and statistics are not close, when reading ignores parts of the text which are not important according to his/her opinion. This results in a vicious circle. The person reading a scientific paper who is not paying attention to applied methods, writing his/her own paper does not select them with the due diligence. Often one meets scientific publications which at first sight use certain statistical procedures, however, when the reader wants to learn what methods were used in this particular situation it turns out that is often impossible. Authors often include information suggesting the use of a statistical test, writing for example "statistically significant differences were found at $p < 0,05$". However, it is not specified how it has been designated. The only information in the text referring to the use of statistical methods is the name of the computer program. It should also be noted that one can still find articles where conclusions are not supported by applying methods of statistical inference.

One may meet with the opinion that you can use statistics to prove any hypothesis which the author of a study puts. The fact is that if we do not use statistical tools properly, for example miss important assumptions, we can get surprising conclusions, but not always true. Impartially conducted statistical inference does not lead to such errors. To make correct analysis one should plan in advance the whole investigation. It is important that the group which is the subject of research has to be selected at random and its size can not be too small. Important is also the choice of tests which are used. If one reads a number of publications one can have a feeling that some authors have their favourite methods and apply them regardless of the situation. Particular attention should be paid here to the frequently used Student's t-test and ANOVA [7]. Textbooks describing these methods give an assumption that data may be modified in such a way that it is possible to correctly carry out the analysis. One of them is the normality of distribution of examined variables. If this condition is not met there is the necessity to use alternative statistical methods. Unfortunately, not all authors are announcing whether they considered assumptions of the used tests [8, 9]. You can also meet interesting situations, but absolutely incorrect, where the assumption of normal distribution is actually checked, but another important condition for abundance is missed. Compatibility with normal distribution is checked using an adequate test and the abundance of the sample is less than 10. With such a small number of observations checking this assump-

tion is not appropriate. Applying Student's t-test in this situation should not take place.

Today scientific research and the publication of articles requires knowledge of at least basic statistics. It seems illogical to expend funds and spend time on research to interpret the results in an incorrect way. For many years, in numerous publications, attention has been paid to the problem of improper use of statistical methods. Already in 1994 Douglas Altman in the article "The scandal of poor medical research", wrote that very often available techniques are used wrongly. In result, this leads to many methodological errors. Despite the passage of time we can observe that this situation is still up-to-date.

Purpose of study

The objective of this study was the presentation of statistical methods most frequently used to analyze medical data and to analysis errors in statistical inference. Assessment of the above issues had been made in the context of the points awarded for journals by MNiSW.

Material and methods

For statistical analysis 10 Polish medical journals were chosen at random. As a result of the draw it turned out that they were in the following groups: 3 journals were rated by MNiSW for 2 points, 4 for 4 points, 2 6 points and 1 for 10 points. We analysed all research articles which appeared in these journals in 2009 (n = 248). Articles assessed for 2 points were n = 27, 4 points n = 104, 6 points n = 90 and 10 points n = 27. Breakdown of the percentage of articles in this study considering the MNiSW score presents Figure 1.

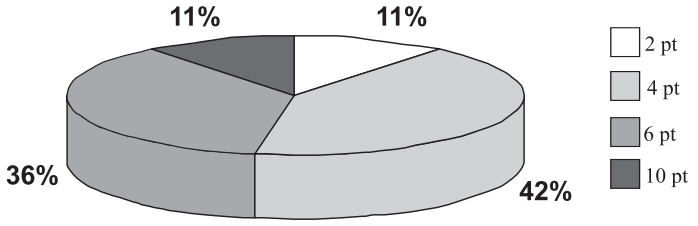


Fig. 1. Breakdown of the percentage of articles based on the score of the journal

In the present analysis for a small research or control group called for a group size of less than 30 items, and other of size 30 and more was large. Figure 2 presents the distribution of the sum of number of elements in the research and control group.

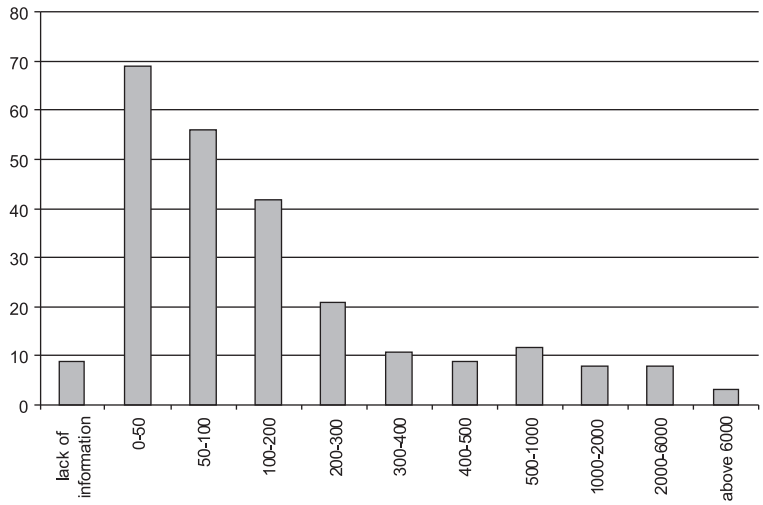


Fig. 2. The number of elements including research and control group

In statistical calculations to verify research hypotheses for qualitative variables the Chi-square test for independence was used. To the ordinal variables analysis and quantitative variables without normal distribution the following test were used: the Mann-Whitney U test, Kolmogorov-Smirnov test and Wald-Wolfowitz test for comparing two groups and non-parametric ANOVA Kruskal-Wallis test, median test and Post-Hoc multiple comparisons of average rang for comparing more groups.

The results were statistically significant at $p < 0,05$. For the statistical analysis the program Statistica 8.0 (StatSoft Inc.) was used.

Results and discussion

During the analysis an alarming phenomenon of lack of statistical inference in the research articles was noticed (Table 1). No inference applies to 25% of all articles, the biggest problem exists in the journals for 4 points, there is no inference performed until 43% of the articles. By contrast, in journals for 10 points, always (100%) statistical inference was carried out. There is a significant statistical dependence between the fact of using me-

Table 1

General characteristics of the analyzed research articles divided into MNiSW scores

Characteristic	2 pt MNiSW n (%)	4 pt MNiSW n (%)	6 pt MNiSW n (%)	10 pt MNiSW n (%)	Total n (%)	Statistical significance
Analyzed research Works	27 (11%)	104 (42%)	90 (36%)	27 (11%)	248 (100%)	–
Works, in which applied statistical inference	22 (81%)	59 (57%)	78 (87%)	27 (100%)	186 (75%)	$p < 0,001$
Research group						
– Small	1 (4%)	17 (16%)	13 (14%)	13 (48%)	44 (18%)	$p < 0,001$
– Large	23 (85%)	82 (79%)	76 (84%)	14 (52%)	195 (79%)	
– Lack	3 (11%)	5 (5%)	1 (1%)	0 (0%)	9 (4%)	
Control group						
– Small	1 (4%)	4 (4%)	12 (13%)	0 (0%)	17 (7%)	NS
– Large	2 (7%)	9 (9%)	21 (23%)	0 (0%)	32 (13%)	
– Lack	24 (89%)	81 (88%)	57 (63%)	27 (100%)	199 (80%)	

thods of statistical inference in the articles of the journals and the number of MNiSW points that this journal has ($p < 0.001$). Median values in the group scoring MNiSW which used statistical inference is 6 points. And in the group in which these methods are not used is 4 points. Fromm [10] based on research conducted in the years 1982–1993 suggested that 13% of the articles was characterized only by using descriptive statistics (lack of statistical inference).

The analysis of the size of research groups indicates a significant relationship between group size and the number of points awarded by MNiSW for journals ($p < 0.001$). Small groups are used mainly for journals for 6 points (median), and large groups in journals at 4 points (median). Z. Sych in his work [11] observed a significant increase in the incidence of large samples (between 1988–1990). The inverse of this situation has been observed in the present study which explains better planning of medical research. The better the journal the better articles are published in it. This is often associated with more expensive research. A small study group appears in the 48% of articles for 10 points, and only 4% for the articles for 2 points. In turn, a large research group is up in 84–85% of the journals for 2 and 6 points.

The control group was found only in 20% of the analyzed articles. Most of the articles for 6 points – 36% of articles used in the control group. However, no control group was used in the articles for 10 points.

The presentation of the results using descriptive statistics occurred in almost all research papers (96%). The most common parameter used was the mean 41%–96%, standard deviation 29%–81%, range 7%–21% and a median 0%–20% (Table 2). There is a relationship between the use of mean ($p < 0.001$) and standard deviation ($p = 0.001$) and the points obtained by the journal from MNiSW. The median in the group using these parameters is 6 points, and the group is not applying the 4 points.

In 8% (n = 19) of the articles does not appear the name of the statistical test and in the results are given p-value. In this case, significant dependence on the MNiSW scoring is not observed, however, in articles for 10 points this problem does not appear.

Already in 1980, S. A. Glantz has observed in his study that the Student's t-test is the most popular and most widely used statistical test in medical articles [7]. In this study, it is used in 11%–48% of articles (Table 2) and significantly more frequently in journals bulleted higher ($p = 0.001$) – median 6 pkt. D. G. Altman in his work [9, 12] draws attention to the mistakes made in popular statistical analysis. In the case of Student's t-test the problem is usually that the data are not in line with the statistical assumption that both sets come from a population with normal distribution and have the same variance. In the present study, information about checking the normal distribution when applying the Student's t-test appeared only in 33% (n = 16) of analyzed articles. Checking the normal distribution should be done also in case of the analysis of variance and Pearson's correlation. Table 2 shows the frequency of checking of this assumption.

The analysis of variance (ANOVA) was used in 15% of the articles. There is a statistical relationship ($p < 0.001$) between the frequency of the used analysis of variance and MNiSW points. It has been used frequently in journals with higher scores (median 6 points).

Non-parametric tests were used in 43% of all articles, in particular groups their number ranged from 8 (which represents 30% of the 10 points group) to 45 times (representing 50% of the 6 points group). Although the application of nonparametric tests should be used in position location measurement (quartiles), median was given only in 18% (n = 19) of articles. However, most often the mean (61%) and standard deviation (54%) was presented, which should be presented when using parametric tests. L. Zaborski also notes that authors often do not examine the characteristics of normal distribution, or completely without knowing it, give to characterize the study population mean and standard deviation [8].

One of the most commonly used tests, in addition to Student's t-test, is the Chi-square test for independence [10]. In the present study it was used

Table 2

The statistical methods used in analyzed research articles including MNiSW scoring

Statistical method	2 pt MNiSW n (%)	4 pt MNiSW n (%)	6 pt MNiSW n (%)	10 pt MNiSW n (%)	Total n (%)	Statistical significance
Descriptive statistics:						
– Mean	18 (67%)	43 (41%)	55 (61%)	26 (96%)	142 (57%)	$p < 0,001$
– Standard deviation	16 (59%)	30 (29%)	48 (53%)	22 (81%)	116 (47%)	$p = 0,001$
– Median	3 (11%)	5 (5%)	18 (20%)	0 (0%)	26 (10%)	NS
– Scope	2 (7%)	15 (14%)	19 (21%)	4 (15%)	40 (16%)	NS
– Quartile	0 (0%)	1 (1%)	6 (7%)	0 (0%)	7 (3%)	NS
– OR	1 (4%)	2 (2%)	8 (9%)	0 (0%)	11 (4%)	NS
– SEM	0 (0%)	0 (0%)	3 (3%)	5 (19%)	8 (3%)	$p < 0,001$
No test name, and there is p-value	2 (7%)	6 (6%)	11 (12%)	0 (0%)	19 (8%)	NS
Student's t-test	5 (19%)	11 (11%)	20 (22%)	13 (48%)	49 (20%)	$p = 0,001$
ANOVA	2 (7%)	9 (9%)	13 (14%)	12 (44%)	36 (15%)	$p < 0,001$
ANOVA + post hoc	0 (0%)	4 (4%)	5 (6%)	6 (22%)	15 (6%)	$p = 0,004$
Pearson correlation	5 (19%)	4 (4%)	8 (9%)	9 (33%)	26 (10%)	NS
Check for normal distribution:						
– Student's t-test	1 (20%)	1 (9%)	11 (55%)	3 (23%)	16 (33%)	NS
– Analysis of variance	2 (100%)	0 (0%)	5 (38%)	2 (6%)	9 (25%)	NS
– Pearson correlation	3 (60%)	0 (0%)	1 (13%)	2 (22%)	6 (23%)	NS
Nonparametric tests	16 (59%)	38 (37%)	45 (50%)	8 (30%)	107 (43%)	NS
Descriptive parameters used in the nonparametric tests						
– Mean	13 (81%)	15 (39%)	30 (67%)	7 (88%)	65 (61%)	NS
– Standard deviation	11 (69%)	12 (32%)	28 (62%)	7 (88%)	58 (54%)	NS
– Median	3 (19%)	2 (5%)	14 (31%)	0 (0%)	19 (18%)	NS
Mann-Whitney U test	5 (19%)	8 (8%)	25 (28%)	0 (0%)	36 (15%)	NS
Chi-square test for independence	7 (26%)	26 (25%)	20 (22%)	2 (7%)	55 (22%)	NS
– Chi-square with the amendment	1 (14%)	6 (23%)	10 (50%)	0 (0%)	17 (7%)	NS
– Lack of necessary amendments	5 (71%)	14 (54%)	6 (30%)	1 (50%)	22 (9%)	NS
Wilcoxon test	2 (7%)	1 (1%)	6 (7%)	3 (11%)	12 (5%)	NS
Kruskal-Wallis test	3 (11%)	7 (7%)	6 (7%)	0 (0%)	16 (6%)	NS
Spearman correlation	1 (4%)	7 (7%)	8 (9%)	3 (121%)	19 (8%)	NS

in 22% of the articles. Important is that this test with small abundance in subgroups has some modifications: Yates amendment, Fisher's exact test and the V-square test. In the current study, 9% of the articles did not apply "corrections" to the test despite the small abundance in subgroups [13].

Low levels of statistical methods used in the published articles observed in this study have long been a subject of particular concern. The basic condition for improving this situation is to improve the level of statistics and increase the accountability of journals.

D.G. Altman in his article "Improving the quality of statistics in medical journals" [9] put forward concrete proposals to raise the level of publications. First of all, he believes that all the research using statistical methods should be reviewed by the statistics. Subsequently, the revised articles should be returned to the same reviewer to re-evaluate. An interesting solution is the suggestion that the journal should supply statistical guidelines for the authors which should be the standard of all research and should also contain a separate paragraph devoted to statistical methods implemented.

Conclusions

Journals with higher MNiSW scoring have a more valid statistical methods and more often use parametric methods such as the analysis of variance and Student's t-test. Large research groups are more often used in the articles for a smaller number of points MNiSW. Greater emphasis should be put on teaching biostatistics in medical studies and related fields to achieve improvements in the application of statistical methods. Journal editorial teams should employ statistical reviewers with experience.

REFERENCES

[1] Racki G. Dwuznaczny urok listy czasopism punktowanych. W: Bibliograficzne bazy danych: kierunki rozwoju i możliwości współpracy. Ogólnopolska konferencja naukowa z okazji 10-lecia bazy danych BazTech. Bydgoszcz, 27–29 May 2009.

[2] Krzych J. Ł. Interpretacja wyników analizy statystycznej danych. Kardiochirurgia i Torakochirurgia Polska, 2007; 4 (3): 315–321.

[3] Moczko J. A., Brębowicz J. A., Tadeusiewicz R. Statystyka w badaniach medycznych. Wydawnictwo Springer Warszawa 1998.

[4] Sacket D. L. Rational therapy in the neurosciences: the role of the randomized trial. Stroke 1986; 17: 1323–1329.

[5] Gellerstedt M. Istotne znaczenie statystyki w badaniach medycznych. Alergia Astma Immunologia, 2003; 8 (1): 25–32.

[6] Zejda J. E. Medyczny artykuł naukowy. Zasady dobrej praktyki publikacji. Ann. Acad. Med. Siles. 2006; 60,4: 323–329.

[7] Glantz S. A. Biostatistics: how to detect, correct and prevent errors in the medical literature. Circulation, 1980; 61, 1.

[8] Zaborski L. Zastosowanie metod statystycznych w badaniach i błędy we wnioskowaniu (kłamstwa statystyczne). Medycyna Pracy XLVI, 1995; 3, sup. 4: 81–92.

[9] Altman D. G. Statystyka i etyka w badaniach naukowych. W: Statystyka w praktyce lekarskiej. Wydawnictwo PWN, Warszawa 1997.

[10] Fromm B. S., Synder V. L.: Research design and statistical procedure used in the Journal of Family Practice. J. Farm. Pract. 1986; 23: 564.

[11] Sych Z. Ocena stosowania metod statystycznych w wybranych krajowych czasopismach medycznych. Roczniki PAM w Szczecinie, 1996; 42: 67–84.

[12] Altman D. G. The scandal of poor medical research. BMJ 1994; 308: 283–284.

[13] Stanisz A. Przystępny kurs statystyki z zastosowaniem STASISTICA PL na przykładach z medycyny. Tom 1. Statystyki podstawowe. Wydawnictwo StatSoft. Kraków 2006.

Magdalena Wietlicka-Piszcz
Małgorzata Ćwiklińska-Jurkowska
Department of Theoretical Backgrounds of Biomedical Science
and Medical Informatics, Collegium Medicum in Bydgoszcz,
Nicolaus Copernicus University

PERFORMANCE OF CLASSIFICATION METHODS FOR DIFFERENTIATION BETWEEN CIRRHOTIC TISSUES AND CIRRHOTIC TISSUE WITH CONCOMITANT HEPATOCELULAR CARCINOMA. CLASSIFICATION OF LIVER TISSUES

Abstract: This paper presents the comparison of various discriminant methods for differentiation between cirrhotic tissue and cirrhotic tissue with concomitant hepatocelular carcinoma on the basis of oligonucleotide microarray dataset. Four methods of dimensionality reduction by selection of features (genes) subsets: linear models with empirical Bayes methods [Smyth 2004], SAM, PAM and Wilcoxon statistic were implemented. For studied subsets of genes ranked by these methods the performance of seven discriminant procedures was estimated by test error, 10-fold CV and bootstrap 0.632 with 95% confidence intervals. The best performance was obtained for SVM, bootstrap aggregating trees as well as adaptive boosting trees.

Key words: classification, hepatocelular carcinoma, microarrays

Introduction

Microarray technology is a new promising tool allowing simultaneous investigation of expression levels of thousands or tens thousands of genes. The elaboration of data from microarray experiments involves the use of methods which enable application when the number of features (genes) is much larger than the number of samples (microarrays). The issue of microarray data classification had been reported in many studies (see, e.g., Boulesteix et al., 2008, Dupuy and Simon 2007, Dudoit et al., 2002; Lee et al., 2005; Statnikov et al., 2005; Van Sanden et al. 2008). The results indicate that, although a few methods like random forests or support vector machines seem to perform better than others, there is no single method that would be suitable for all applications.

ISBN 978–83–7431–266–0 ISSN 0860-150X

Because of the very high number of genes in one microarray the pre-selection of features-genes for inclusion into the classification rule may be an important issue. Lee et al. (2005) mention that various methods of active gene selection applied to the same set of microarray data may give different sets of genes and consequently lead to different discrimination results.

In the present paper the data concerning liver tissue with HCV infection are investigated. The purpose of this study was to investigate the performance of various statistical discrimination methods for classification of liver tissues such as: cirrhotic tissue and cirrhotic tissue from patients with hepatocellular carcinoma.

Materials and methods

Data from high density oligonucleotide arrays (Mas et al. 2009, public repository Gene Expression Omnibus, accession GSE1423) were used for investigation of differentially expressed genes in liver tissue samples.

The total number of 58 samples from liver tissue with HCV infection were examined. 41 cirrhotic tissue samples were from patients without hepatocellular carcinoma (Cirrhosis) and 17 cirrhotic tissue samples were from patients with hepatocellular carcinoma (CirrhosisHCC). Each microarray consisted of 2227 probe sets. All microarrays were divided into a learning set and a testing set, of respectively 40 and 18 microarrays.

The investigated data were previously preprocessed, so for each probe expression summaries were available. The analysis was performed with the use of R and Bioconductor package.

Genes Selection Methods

To identify genes differentially expressed, i.e. genes that exhibit a statistically significant difference across the examined tissue types, fours methods of gene selection were applied. T-statistic is widespread in assessing differential expression in microarray experiments, therefore two methods basing on t-statistic were used: a method based on linear models and moderated t-statistic reported as (*called*) Linear Models for Microarray Data (LimmaBH) introduced by Smyth (2004) and Significance Analysis of Microarrays (SAM) proposed by Thuser et al. (2001). Also a nonparametric test, a Wilcoxon rank sum test was applied to identify differentially expressed genes. Another approach to gene selection is presented by Prediction Analysis of Microarrays (PAM) introduced by Tibshirani et al. (Tibshirani, 2002). It is based on the nearest shrunken centroid classifier and provides a set of genes that best characterizes each class.

LimmaBH is a method based on the linear model and the empirical Bayesian method (Netwon et.al 2001) applied to estimate fold changes of differential expression and on moderated t-statistics to rank genes according to the probability of differential expression. The tests are also adjusted for multiplicity by using the Benjamini & Hochberg method (Benjamini and Hochberg, 1995) to control false discovery rate (the expected proportion of type I errors).

The results of a series of microarray experiments can be represented as a $m \times n$ matrix where m represents the number of genes and n the number of microarrays ($n = n1 + n2$). Let y_{gij} denote the expression level of gene g in array i from group j.

Significance Analysis of Microarrays generates a list of genes ranked according to the probability of differential expression on the basis of the modified t-statistic.

In this method the genes are ranked according to the probability of differential expression on the basis of the modified t-statistic:

$$t_g = \frac{\bar{y}_{1_g} - \bar{y}_{2_g}}{s_g + c}$$

where

$g = 1, \ldots, m$;

s_g is the standard deviation of repeated expression measurements.

This modified t-statistic has t-Student distribution with $n - 1$ degrees of freedom. When the variability measured by s_g is close to 0, the values of classical t-statistic can become too large. So s_g in the denominator is augmented by a small positive constant c. Its value is chosen to minimize the coefficient of variation of the test statistic. This constant ensures that the variance of the score t is independent of gene-expression.

All gene selection methods were applied for the learning set (LS), which consisted of 40 arrays.

Classification methods

The four sets of genes obtained from the gene selection methods were used in the construction of discrimination rules by applying different discrimination methods. The following methods were considered: support vector machines (SVM), diagonal linear discriminant analysis (DLDA), diagonal quadratic discriminant analysis (DQDA), k nearest neighbour (k-NN) and classification trees. Additionally, the methods based on ensemble classifiers such as adaptive boosting and bagging trees were applied (Duda et al. 2001, Webb 2002, Dettling 2004).

For each of the four sets of genes the discrimination methods were applied to subsequently enlarged sets of genes, which included $2, 3, 4, \ldots, 100$ of the highest-ranked genes.

Parametric discriminant methods DLDA and DQDA are special cases of classical linear and quadratic discriminant functions, created by the assumption that the features are independent within each class, so the within-class covariance matrix is diagonal. In nonparametric k-NN method the parameter k (number of neigbours) was chosen for each subset of genes according to the criterion of minimization of CV error.

Support Vector Machines (SVM) method [1] separates 2 data groups by the hyperplane defined on the basis of the criterion that maximize the margin between groups. The Support Vector Machines classification technique is based on mapping the data to represent observations in high dimensional space – usually much higher than the original feature space.

The main advantage of the Support Vector Machines is that complexity of the classifier is determined by the number of support vectors – observations lying on the margin – rather than the dimensionality of the transformed space. As a consequence, SVM have less often problems with overfitting than many other methods and is appropriate for high dimensionality.

The SVM method (Cortes and Vapnik 1995) was applied with the regularizing constant equal to 0.5. The best performance of SVM was obtained for linear kernels, so we present only linear kernel results. The outcomes are consistent with the conclusions of other authors. However, some authors like Noble (2004), find other outcomes – he stated that the SVM using third-degree polynomial kernel was the best performing method.

Combining classifiers based on resampling (randomly generating learning sets)

The single classifiers built on relatively small learning data set are often biased. Then, the methods based on randomly generating subsets of training data joined with combining classifiers built on them may be useful and may improve the performance.

Bagging (Bootstrap AGGregatING) Breiman [1996] is an ensemble based on bootstrap samples created by drawing n times from the learning set with possible replacements, where n means the size of the learning set. The classifier is trained on each bootstrap subsample. Resulting classifiers are then combined, e.g. by the averaging the posterior probability or the unweighted majority vote.

Boosting is also the method based on resampling the learning data set; however boosting is a deterministic procedure, because the selection of

subsequent subsamples depends on the results of combined classifier performance achieved in previous loops. In sequentially generated learning sets the weights of misclassified cases are increased so the ensemble creates the improved classifiers.

"Boosting" the performance of weak classifiers is originated from Freund & Schapire [1996] ARCing-Adaptive Resampling and Combining. The most popular boosting method is Adaptive Boosting (AdaBoost). AdaBoost allows the designer to continue adding classifiers until some desired low training error is achieved. Bagging and boosting are the methods most often used for weak base classifiers, and trees as constituent classifiers are used.

Errors evaluation

In microarray experiments the appropriate classification error estimation is a very important issue. Although in the case when the amount of data is relatively small, which is typical for microarray experiments, the estimation of error rate is not straightforward. The evaluation of the error rate by the test sample error can be biased. Therefore, the assessment of the constructed discrimination procedures was performed by the error estimation on the test set, by 10–fold cross-validation (CV) [Mc Lachlan 1992] applied to the whole dataset and by the bootstrap 0.632 method [Efron 1983] based on 100 randomly generated samples. For 10 fold CV and bootstrap 0.632 mean error rates the 95% confidence intervals were estimated additionally.

Results and discussion

The identification of differentially expressed genes was performed by the use of the four features selection methods mentioned above and consequently four sets of differentially expressed genes were obtained. In each set genes were ordered according to the appropriate test statistic.

The first highest-ranked 100 genes produced by each of the gene selection methods were taken into further consideration, because the increased number of genes over the first 100 most important ones does not result in the improved classification performance. Therefore, further analysis was restricted to the first 100 genes obtained from each of the considered gene selection methods.

Fig. 1. presents the Venn diagrams illustrating all pair wise comparisons among the four genes sets. 27 genes are common across all the subsets. The highest overlapping of genes present sets produced by SAM (in further analysis called set2) and PAM (set3) methods (96 genes). The remaining

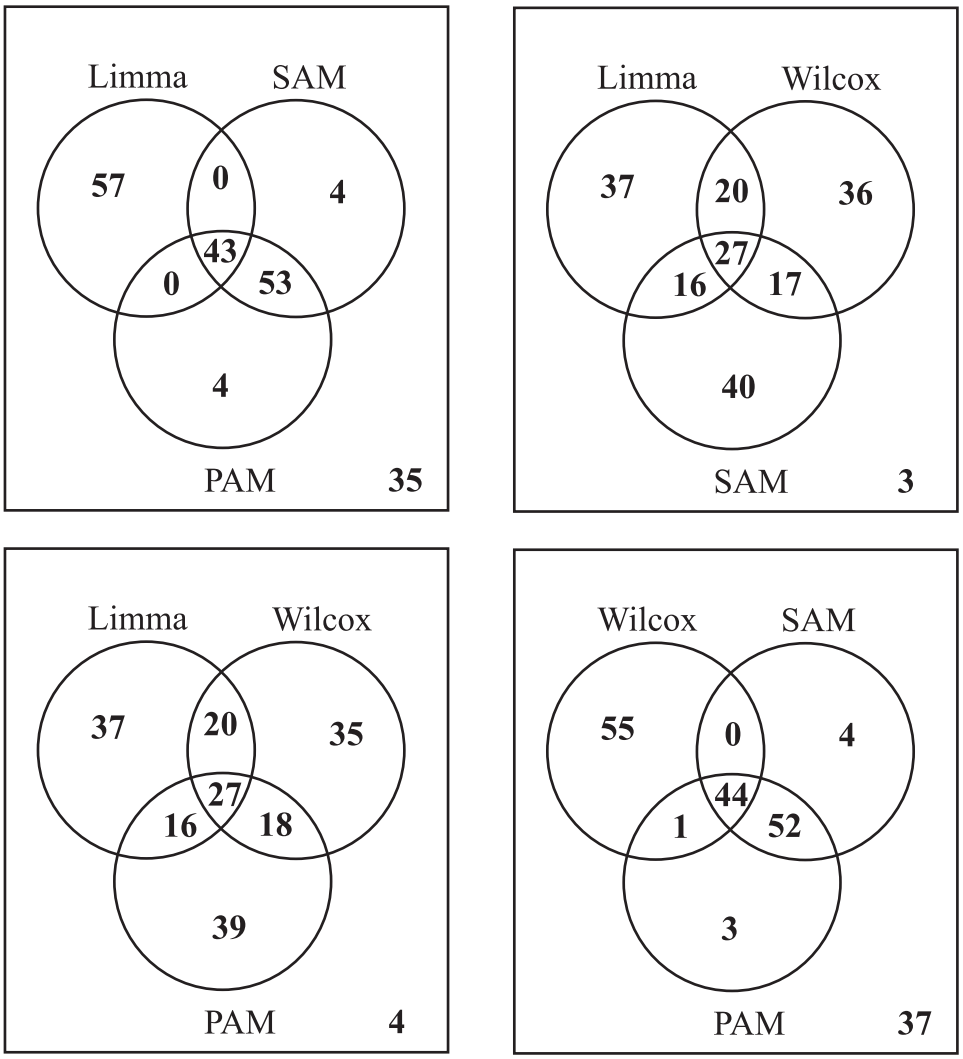


Fig. 1. Venn diagrams illustrate all pair wise comparisons of overlapping among the four gene sets produced by considered gene selection methods like LimmaBH, SAM, PAM and Wilcoxon

pair wise comparisons show overlapping of about half of the total amount of genes. Because of the high overlapping of genes subsets obtained from SAM and PAM methods, the classification results will be shown just for the SAM method. The set of genes obtained by LimmaBH method and ranked by Benjamini-Hochberg procedure will be called set1, and set4 will denote the set of genes ordered according to the nonparametric Wilcoxon statistic.

For each of the four gene sets the misclassification errors were estimated for subsequently enlarged gene sets, which included $2, 3, 4, \ldots, 100$ of the highest-ranked genes.

For each of the selection methods the results are presented in pairs of two figures: the first for the methods not connected with classification tress and the other one for trees.

Figures 2–3 show the misclassification errors estimated by using the test dataset for different discrimination methods applied to sequentially enlarged set of genes of *set1 (LimmaBH)*. Figures 4–5 show the misclassification error rates estimated by 10–fold cross-validation (CV) for different discrimination methods applied also to set 1 and fig 6–7 present for the same genes sets the error rates estimated by the bootstrap 0.632 method.

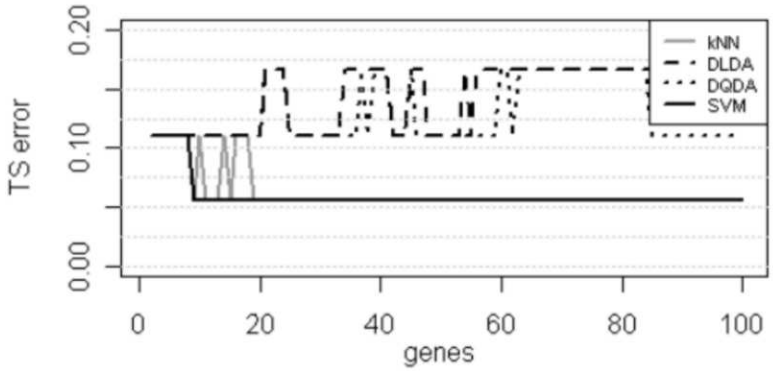


Fig. 2. Classification test errors of discriminant methods: k-NN, DLDA, DQDA, SVM for ascending subsets of *set1 (Limma)*, from 2 to 100 genes

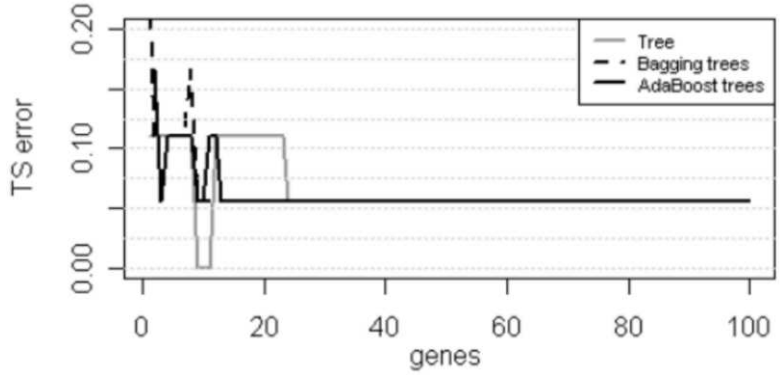


Fig. 3. Classification test errors of discriminant methods: trees, bagging trees and adaptive boosting trees (AdaBoost) for succeeding subsets of *set1* (Limma), from 2 to 100 genes

Test sample errors are between 0.056 and 0.167 for KNN, DLDA, DQDA and SVM (fig. 2). The smallest test errors equal 0.056 were obtained for both SVM and kNN for subsets containing 20 or more genes. Test errors for both bagging and AdaBoost ensemble tree and for single tree classifiers are also equal to 0.056 for more than 20 genes (fig. 3). For bagging and boosting no optimal subset of genes may be pointed. Though for tree classifier the minimum test error reached zero for about 15 genes, the tree is unstable and can be overtrained, so CV errors and bootstrap 0.632 do not confirm this result (fig. 5, 7).

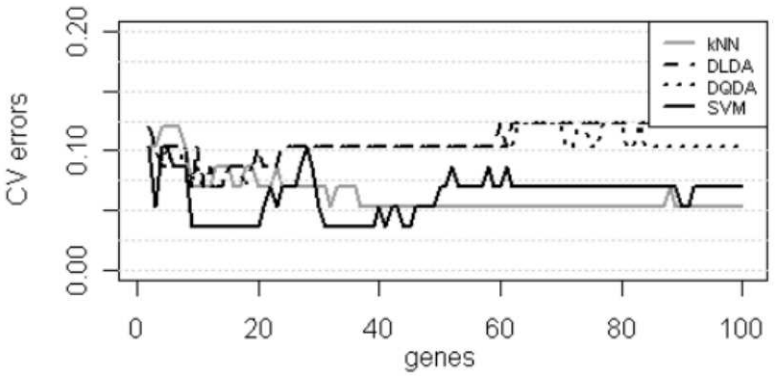


Fig. 4. Classification cross-validation errors of discriminant methods: k-NN, DLDA, DQDA, SVM for ascending subsets of *set1 (Limma)*, from 2 to 100 genes

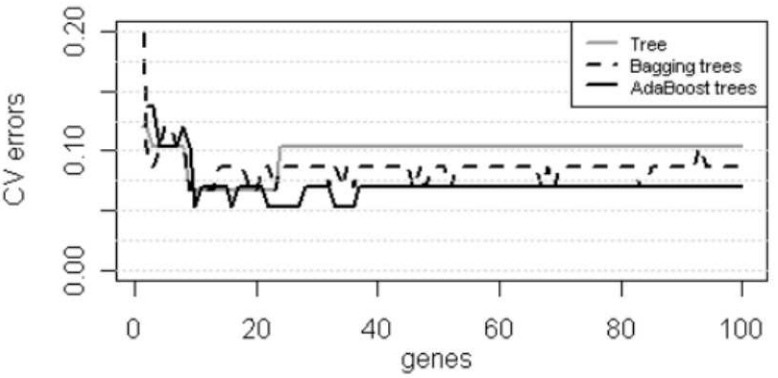


Fig. 5. Classification cross-validation errors of discriminant methods: trees, bagging trees and adaptive boosting trees (AdaBoost) for succeeding subsets of *set1 (Limma)*, from 2 to 100 genes

The representative set of genes, indicated by the smallest errors of SVM classifier, was between 8 and 20 and from 30 till 40 genes (fig. 5) – where the CV error reached the value equal to 0.037. For more then 50 probe sets the CV error for SVM is increasing, in the contrary to kNN classifier (Fig. 4). DLDA and DQDA gave higher CV errors.

CV errors of boosting trees are between 0.037 and 0.1 and for bagging are from 0.053 to 0.137. Bagging and boosting measured by bootstrap 0.632 error gave, similarly to CV error, better results (*smaller errors*) than single tree (comp. fig. 5 and 7). From Fig 7 the optimal subset for ensemble tree classifiers cannot be pointed.

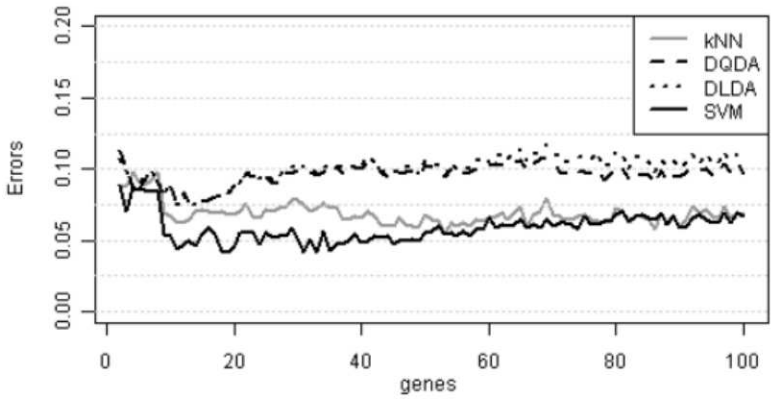


Fig. 6. Bootstrap 0.632 classification errors of discriminant methods: k-NN, DLDA, DQDA and SVM for ascending subsets of *set1* (Limma), from 2 to 100 genes

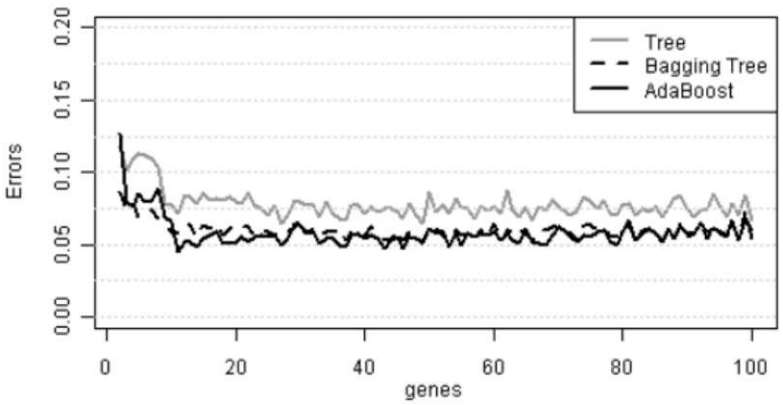


Fig. 7. Bootstrap 0.632 classification errors of discriminant methods: trees, bagging trees and adaptive boosting trees (AdaBoost) for succeeding subsets of *set1 (Limma)*, from 2 to 100 genes

Bootstrap 0.632 errors show smaller fluctuations than for CV errors across the varying number of genes and the values are between 0.04 and 0.12.

For each classification method and for each subset of set1 (LimmaBH) the standard errors were calculated. Table 1 presents the minima, maxima and mean values of standard errors for CV and bootstrap estimation of classifiers for all the subsets of set1 (LimmaBH).

Table 1

The standard errors for the two methods of error assessment

classification method	CV-10			bootstrap 0.632		
	min	max	mean	min	max	mean
kNN	0,027307	0,038233	0,030806	0,002390	0,004054	0,003246
DQDA	0,028306	0,037663	0,035850	0,002021	0,003269	0,002545
DLDA	0,028306	0,037663	0,036003	0,002094	0,003077	0,002641
SVM	0,024570	0,051926	0,028482	0,002384	0,003997	0,003332
Bagging Tree				0,002325	0,004959	0,003445
Tree				0,002708	0,005653	0,003991
AdaBoost				0,002727	0,005680	0,003543

Cross-validation errors (fig. 4–5) seem to suggest that the addition of only one or a few genes can significantly alter the performance – it can improve or worsen the performance. The CV error difference after adding only one or a few genes reaches even 0.1. The CV error can increase even two times. Such differences are not confirmed by a more stable method as bootstrap 0.632 error estimate (Fig. 6–7). On the basis of 95% confidence intervals for CV errors we can not conclude that any examined classification method significantly outperforms the other. For the bootstrap 0.632 assessments the errors have much smaller fluctuations than for CV–10 errors (Fig. 6–11). Therefore, for the comparison of the genes selection methods and classifiers in further analysis we will use bootstrap 0.632 error assessments. The ranges of standard errors for mean error rates evaluated by bootstrap 0.632, presented in Table 1, indicate much smaller diversity than CV error, so the bootstrap error assessment is more precise. Also Braga-Neto and Dougherty (2004) concluded that the cross-validated estimators show high variance and large outliers. The large outliers produced by the cross-validation estimators can cause that significantly inaccurate conclusions can be reached for a considered data set. The authors also concluded that the bootstrap estimators, in particular the bootstrap 0.632 estimator, display the best overall performance [Braga-Neto, 2004].

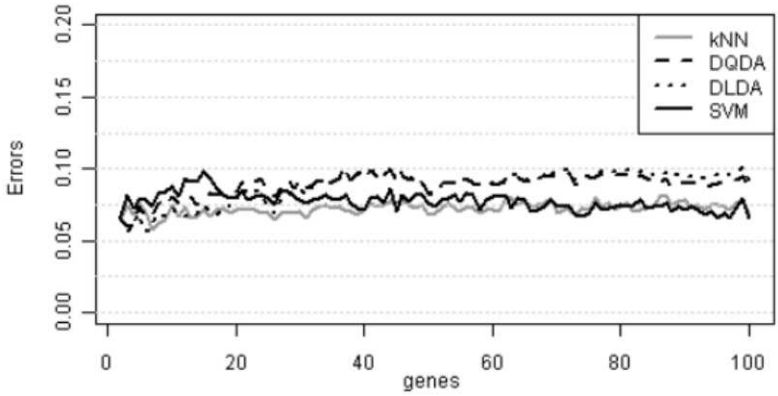


Fig. 8. Bootstrap 0.632 classification errors of discriminant methods: k-NN, DLDA, DQDA and SVM for ascending subsets of *set2* (SAM), from 2 to 100 genes

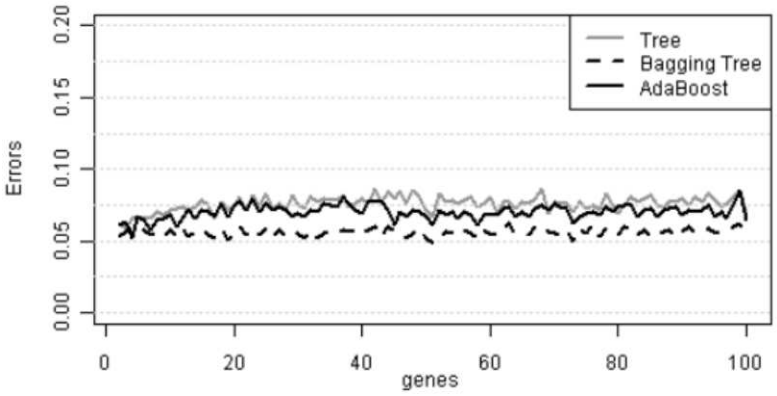


Fig. 9. Bootstrap 0.632 classification errors of discriminant methods: trees, bagging trees and adaptive boosting trees (AdaBoost) for succeeding subsets of *set2* (SAM), from 2 to 100 genes

For set1 (LimmaBH) the smallest bootstrap 0.632 assessment of error rates were reached for SVM method, for the number of genes from 18 till 45. Figures 8–11 show the misclassification errors estimated by the bootstrap 0.632 method for set2 (SAM) and for set 4 (Wilcoxon).

Bootstrap 0.632 errors for DLDA and DQDA for number of genes bigger than 40 are relatively high. From Figure 9 and 10 and using appropriate SE from Table 1 we can conclude that significantly smaller results, measured by the confidence interval of bootstrapping 0.632 error, were obtained for ensemble tree than for other examined classifiers.

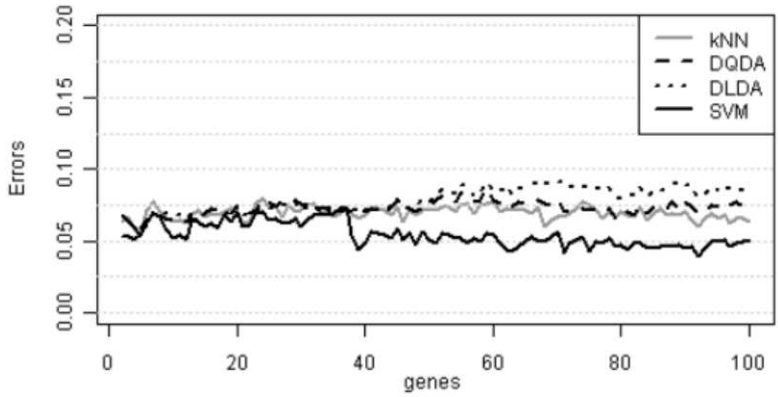


Fig. 10. Bootstrap 0.632 classification errors of discriminant methods: k-NN, DLDA, DQDA and SVM for ascending subsets of *set4* (Wilcox), from 2 to 100 genes

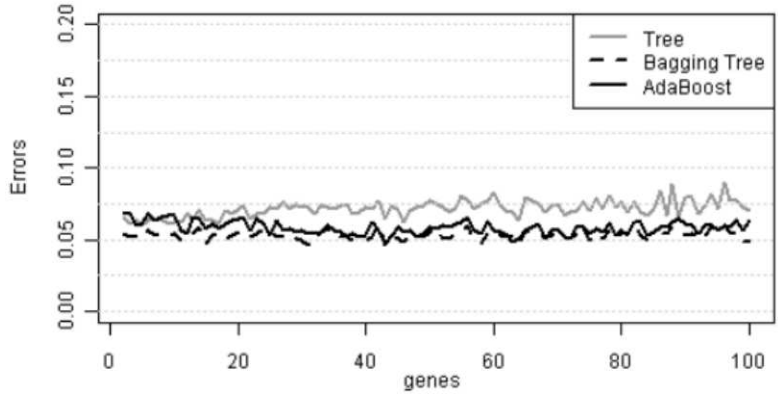


Fig. 11. Bootstrap 0.632 classification errors of discriminant methods: trees, bagging trees and adaptive boosting trees (AdaBoost) for succeeding subsets of *set4* (Wilcox), from 2 to 100 genes

The analysis of the bootstrap 0.632 error rates on Fig. 7, 9, and 11 for three sets: set1 (LimmaBH), set2 (SAM) and set3 (Wilcoxon) show that the differences between error rates are relatively small. However, for set1 and the SVM the bootstrap 0.632 obtain smallest values close to 0.05 (with small variability from 0.041–0.059) for the range between 10 and 45 probe sets.

For Wilcoxon statistic criterion selection, we obtained significantly smallest errors for SVM, however for number of genes bigger than 40 (Fig. 11).

For resampling classifiers we can not point to any optimal subset of genes for adaptive boosting and bagging while for other methods we can see the significant differences between bootstrap errors for consecutive subsamples, for example for SVM.

For set1 (LimmaBH) and set2 (Wilcoxon) the adaptive boosting and bagging tree errors oscillate around 0.05, for set3 bagging tree is also 0.05 but Adaboost is 0.07. Thus for ensemble classifiers set1 and set3 seem to have similar effectiveness. The smallest bootstrap errors – obtained for SVM and for set1 – are about 0.05 for gene sets from 10 to 40 genes, while for set3 errors are about 0.05 for 40 or more genes. For SAM and PAM all bootstrap errors are for all considered classifiers over 0.05.

The comparison of our results with results obtained by Mas et al. 2009 is not straightforward. Those authors performed the selection of genes on the whole data set (all 58 microrarys), not on the smaller subset (learning set), so the classification results based on their selection can be optimistically biased [Braga-Neto and Dougherty 2004]. This issue was raised by many researchers [e.g. Ambroise et al. 2002, Wood et al. 2007]. In our work the differentially expressed genes were selected from the learning set which consisted of 40 microarrays, so the results [*from Mas et al. and presented in current work*] are hard to compare. The authors obtained for random forest out-of-bag error equal to 0.089 for one subset containing fifteen genes selected by Gini index. They also applied logistic regression, where two variables were chosen, obtaining resubstitution error equal 0.036, though resubstitution error is known as optimistically biased.

Conclusion

For considered data set (Mas et al. 2009) concerning discrimination between cirrhotic tissue and cirrhotic tissue with concomitant hepatocellular carcinoma the SVM method, adaptive boosting and bagging gave the best classification results. However, this conclusion could not be valid for another dataset, because the differences in error rates estimates are relatively small and sometimes not significant. For microarray data classification problems the application of several classification methods coming from different sources can be useful. The comparison of considered gene selection methods seems to be a difficult task because they give similar error rates. Basing on SVM classifier results LimmaBH method for genes' selection can be recommended. However, for gene set obtained by the Wilcoxon test the same level of error was reached, although for a bigger number of genes. It is certain that further investigations in this area are necessary.

REFERENCES

[1] Alon, U. et al. (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. PNAS, Vol 96, p 6745–6750.

[2] Ambroise C. Mc Lachlan G. 2002 Selection bias in gene extraction on the basis of microrray gene expression data. PNAS vol 99 No 10 pp 6562–6566.

[3] Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series* B, **57**, 289–300.

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

[5] Breiman L. (1996). "Bagging predictors". Machine Learning 24 (2): 123–140.

[6] Cortes C., Vapnik V. Support-vector network. Machine Learning, 1995, 20, 1–25.

[7] Dettling M. (2004): BagBoosting for tumor classification with gene expression data. Bioinformatics, 20 (18), 3583–3593.

[8] Duda O. R, Hart P. O., Stork D. G.: Pattern Classification. Wiley & Sons. 2001.

[9] Dudoit S., Fridlyand J., and Speed T. P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 2002, 98, 77–87.

[10] Dupuy, A. and Simon, R. 2007. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. Journal of the National Cancer Institute, 99:147–57.

[11] Efron, B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.*, 78, 316–331.

[12] Golub et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, Vol. 286: 531–537.

[13] Lee J. W., Lee J. B., Park M., and Song S. H.: Extensive comparison of recent classification tools applied to microarray data. Computational Statistics and Data Analysis. 2005, 48, 869–885.

[14] Mas VR, Maluf DG, Archer KJ, Yanek K et al. (2009) Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma. Mol Me,d Mar-Apr; 15 (3–4): 85–94.

[15] McLachlan, G. J. (1992). Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley.

[16] Newton M. A., Kendziorski C. M., Richmond C. S., Blattner F. R., Tsui K. W. (2001). On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data. Journal of Computational Biology, 8 (1): 37–52.

[17] Noble W. S. Support vector machine applications in computational biology In: Kernel methods in computational biology. Ed: Scholkopf et al. 2004 Massachusetts Institute of Technology.

[18] Pomeroy, S. et al. (2002) Prediction of Central Nervous System Embryonal Tumor Outcome Based on Gene Expression. Nature, Vol 415, p 436–442.

[19] Smyth, G. K. (2004): Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology, Vol. 3, No. 1, Article 3.

[20] Statnikov A., Aliferis C.F., Tsamardinos I., Hardin D., and Levy S. (2005): A comprehensive evaluation of multicateogry classification methods for microarray gene expression cancer diagnosis. Bioinformatics, 21, 631–643.

[21] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of Webb A. R. Statistical Pattern Recognition (2002), New York: Oxford University Press.

[22] Tusher, V., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America, 98, 5116–5121.

[23] Ulisses M. Braga-Neto, Edward R. Dougherty (2004), Is cross-validation valid for small-sample microarray classification?, Bioinformatics, v. 20 n. 3, p. 374–380, February 2004.

[24] Van Sanden, S., Lin, D., and Burzykowski, T. (2008): Performance of gene selection and classification methods in a microarray setting: A simulation study. Communications in Statistics – Simulation and Computation, 37, 418–433.

[25] Webb A. R. Statistical Pattern Recognition (2002), New York: Oxford University Press.

[26] Wood I Vissher P. Mengerstom K. L. (2007) Classification based upon gene expression data: bias and prediction of error rates. Bioinformatics Vol 23 No 11 pp. 1363–1370.

[27] Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30** (4): e15.

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 21 (34) 2010

Piotr Ziniewicz
Paweł Malinowski
Stanisław Zenon Mnich
Department of Statistics and Medical Informatics,
Medical University of Bialystok

THE APPLICATION OF THE JSP METHOD TO THE SYSTEM DESIGNED FOR THE MANAGEMENT OF A CHOSEN MEDICAL UNIVERSITY DEPARTMENT

Abstract: Creation of a system for managing the activity of a Medical University department presents a great challenge for software developers. Such system should reflect many aspects of the activity of such department, including circulation of documents, technical documentation, and patient data. Considerable diversity of data imposes high complexity of the created system. JSP is one of the many methods which is able to cope with this complexity and at the same time can coordinate programmers' team work.

Introduction

Dynamic advances in medical computer science and growing demands of computer users lead to the creation of novel computer systems of increased complexity. Applications designed to realize single tasks are slowly becoming relics of the past, being replaced by more complex systems designated to manage in an organized and guided manner the work of an individual. Such systems contain a number of co-operating components which are located on many machines and communicate with one another using various methods. Then, not only the requirements of project realization, but also proper communication within the entire project should be defined. An additional problem involves changes in the system specifications that may alter its structure. Taking the above into consideration, it can be concluded that the lack of experience in modern methods of software design is one of the most frequent reasons for the failure of the entire process [4].

ISBN 978–83–7431–266–0 ISSN 0860-150X

Software construction process

The basic problem that appears during software construction is its complexity [1]. When the system becomes too complex, the designer ceases to "reign" over all of its aspects. The general understanding of such a complex system in all of its aspects exceeds human possibilities. The division (structuralization) of the system according to the defined criteria seems to be the solution to the problem. However, the problem described by the system has to meet the basic assumption underlying its decomposition, otherwise structuralization of the system will not be possible.

The decomposition means that every element can be described by a sequence of more detailed elements (sub-systems) which also can be split to even simpler elements. This process should be continued until the obtained elements are able to be implemented. As a result of this process, the hierarchical structure is achieved. It is assumed that in such sub-systems, inner-group communication is more dynamic than in the outer one. Expanding the functionality is the natural tendency of computer systems. Decomposition should be conducted in the way that enables easy implementation of a wide range of changes and other modules to the already existing code.

While constructing the computer system it is assumed that the problem to be solved with its use can be decomposed. The software construction is divided into 5 stages:

- Analysis
- Design
- Implementation
- Testing
- Maintenance

The analysis stage includes formalization of the set of system requirements to meet the construction of the functional model of this system, building a model of system interactions with the external world (users, machines) and creation of the information flow control model between the system and its modules. The design stage translates the logical description defined in the previous stage into the description of the physical structure of the system. Based on the physical structure, elements of hardware may be defined as: computers, servers, network media; software: applications, objects, functions and their allocation to equipment components. The implementation stage means coding the algorithms and data using a specific programming language to create a system structure that matches the earlier physical description. In the testing stage software components have to be physically allocated to the hardware units described by the system's struc-

ture. At this stage, proper functioning of the software is tested in the user's environment. Any bugs that appear are fixed. The ultimate user can have access to the system during this phase. Finally, the maintenance stage begins from the moment of the system's delivery to the user and involves: continuous elimination of the bugs that come out during exploitation, expanding the system's functionality according to the user's suggestions and improving the overall performance. Each of these stages should be carried out by an individual team. The stages may overlap or can be implemented consecutively.

The analysis and design of the system (hereinafter referred to design) are the key stages in the process of its creation. Properly conducted, they can significantly reduce the time of software construction, minimize errors, increase reliability, and simplify testing and servicing of the created system. Over the last years, system design has been in the centre of interest among specialists in this field. As a result, many methods that allow for systematic and analytic approach to software design have been worked out. JSP is one of them.

Jackson Structured Programming Method

The JSP method (Jackson Structured Programming, further called the Jackson method) is a well documented and proven method of software design. It is completely independent of the language used. The method was created in 1976 by Michael Jackson and has become a widely used method of design, especially in Europe. In the 70s and 80s of the 20th century it was considered the standard software specification by WHO and the government of the United Kingdom [8]. This method is most frequently used to deal with sequential problems, particularly when the sequence is arranged in time.

The basic principle of the Jackson method is that the design starts with the analysis and modeling of part of the reality in the context of which the system is to work and which is to include. However, no data structures or methods that the system has to perform are specified. The system created with the JSP method contains direct simulation (model) of the reality which has to be specified prior to any designing [5].

Creating a model of reality is generally easier than taking decisions in the design stage. This is because the model frequently exists in reality and is well known to potential users of the system. When modeling a real problem, the software developer should grasp the user's point of view on that problem. Consulting the user may help avoid confusion and numerous

errors in the future. A clear picture of the model defines possible methods and objects of the proposed system.

Another principle of the Jackson method is that an appropriate model of sequential problem is also sequential. The model consists of a sequence of processes that communicate with one another. The sequence should reflect sequentiality of the actual problem. The model should be implemented through specification transformation into an efficient and useful set of processes, adapted to the available hardware. Special attention should be paid to the correct schedule of the processes, and particularly to the fact that a potentially small number of available processor units have to be shared by relatively many separate processes.

The Jackson method-based design process consists of three stages:

- Analysis
- Specification
- Implementation

The first two stages are responsible for the design of appropriate models of the system. The first stage is responsible for creating the data structure. In the second stage, system specification model and program structure model are created. The final stage means encoding of the existing models using a chosen programming language. Sometimes system implementation models are also created; this, however, will not be described here.

Data structure model construction is of major significance in the Jackson method. The design of this model outstrips all other designs, according to the assumption that it is the data structure model that defines the potential space of the system's function models. This is also consistent with the basic design principle to treat the system as a pattern transforming input data into the output data, with very restrictive assumptions as to the form of data structures. These structures can only be represented as a hierarchical tree and are presented graphically as rectangles which can be interconnected by a relationship denoting inclusion. There are three basic types of the relationship: sequence, selection and iteration. They have been presented in Fig. 1.

The sequence (Fig. 1a) means that the structure A consists of components B, C and D. The choice (Fig. 1b) means that the structure A consists of one of the components B, C or D. The iteration (Fig. 1c) means that the structure A potentially contains multiple components C (may not include any of them). It should be emphasized that these basic types of tree-like structures can be mixed with one another to fully reflect the structure complexity. Additionally, each component structure may itself contain multiple smaller parts. This is a consequence of the decomposition rule mentioned

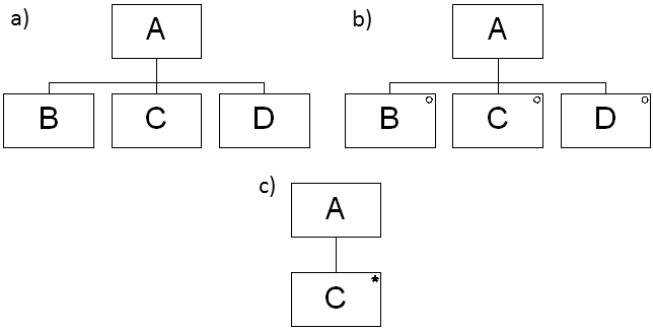


Fig. 1. Tree construction model: (a) sequence, (b) choice, (c) iteration

above. Having developed the full data structure, the application structure model and system specification model should be designed. At this stage, it is assumed that the program structure model inherits from data structure model [6] [7]. This assumption is usually met. If the information consists of separate elements, the procedure processing this information can be divided into appropriate subroutines that process the respective elements. Likewise, processing an alternative means choosing a specific subroutine depending on which component has to be processed. The iterative procedure consists of multiple calls to the same subroutine for each of the components. The appropriate diagrams look exactly the same as in Fig. 1, except that they no longer represent data structures, but the program structure that processes them.

System input data structure

The Jackson method perfectly suits to design a complex system involving a number of interrelated modules. An example of such a system is the application used to support management of educational and research activity of a clinical department. The system operates on a variety of datasets, performing many operations specific to a particular scenario. The scenarios in this case depend on the posts occupied by the system users. Having in mind the enormous complexity of the system logic it is necessary to prepare earlier a detailed documentation in order to avoid complications at later stages of its construction.

To apply the method to design the system that supports management of a clinical department, the following datasets should be specified first [3]:

- all data (entities) inputted to the system (documents that are entered into the system),

- all data (entities) outputted from the system (documents that are presented to the user by the system),
- set of auxiliary data (entity) (used internally as system information store).

Complex information system is split into modules prepared for individual scenarios of system use. The scenarios are related to the posts at which the system is used and so is the list of data ranges used by the information system that supports the clinical department. General data can be divided into 4 categories:

- administrative and financial,
- scientific activity,
- didactic activity,
- medical.

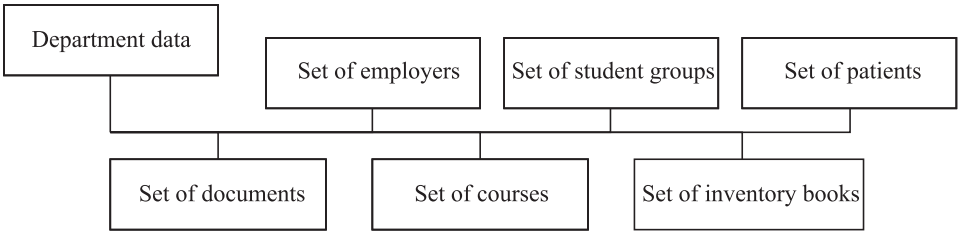


Fig. 2. Overall input data structure of the system

As we can see on the Fig. 2, all of the department data can be divided into six sets. Set of students and courses represents the most of didactic activity of the department and will be described in another article. In this article we will focus on the rest of the sets.

Set of documents consist of many documents data that are used during the department's daily activity. Generally, they may be divided into incoming and outgoing documents. They are represented on the Fig. 3. by "Incoming mail" and "Outgoing mail". It also happened that some documents should be forwarded from one employee to the other (e.g. a draft article can be forwarded to the contributor to implement adjustments). There is an "Internal Mail" on the Fig. 3. for this case. Any mail has a "Header" which gives the date of the mail, its source and destination, document identifier and other common data.

Let us now focus on incoming document data. The most popular mails were split into separated data blocks. They are organized with characteristic for each of them data fields that may be used via dedicated functions (e.g. there may be a function that displays all of the conference invitations with a given subject or destination). There is also a possibility to

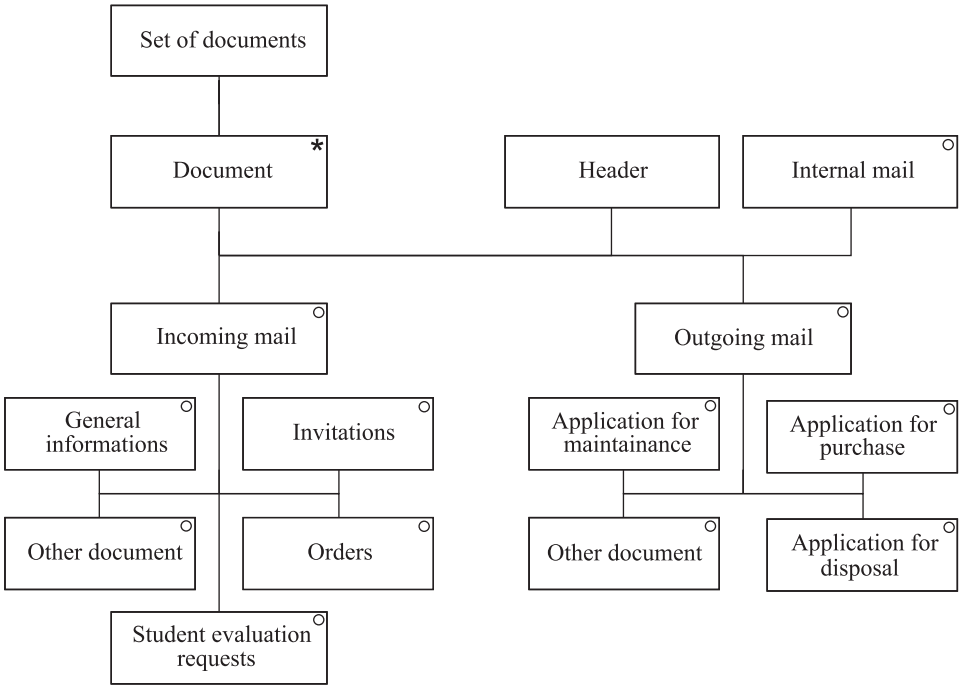


Fig. 3. Input data structure of the documents set

link a "Student evaluation requests" with students from the set of student groups. Then the teacher can run a special function that will display only the requests related with his students. The outgoing mail is frequently used for sending an application for maintenance, purchase or disposal. All of them have to be related to the asset and can modify inventory books. In both cases, besides incoming and outgoing mail, we also have "Other document" data block which includes a common textual memo field which contains the document's content.

In every department there is a necessity to keep the assets record. There is a special data container for this purpose called "Inventory book". On the Fig. 4. one can see an input data structure diagram of a set of such containers. Generally two types of assets can be distinguished: fixed assets and inventories. For every type a separate inventory book is dedicated. Every book has a "Creation time" data field which stores the date when the first record was made in it.

Each inventory book consist of operations set. There can be three operation possibilities: "Acceptance", "Disposal" and "Partial disposal". Depending on the type of asset, different operations are possible and optional data fields can be accessed. Operation "Acceptance" is more universal and

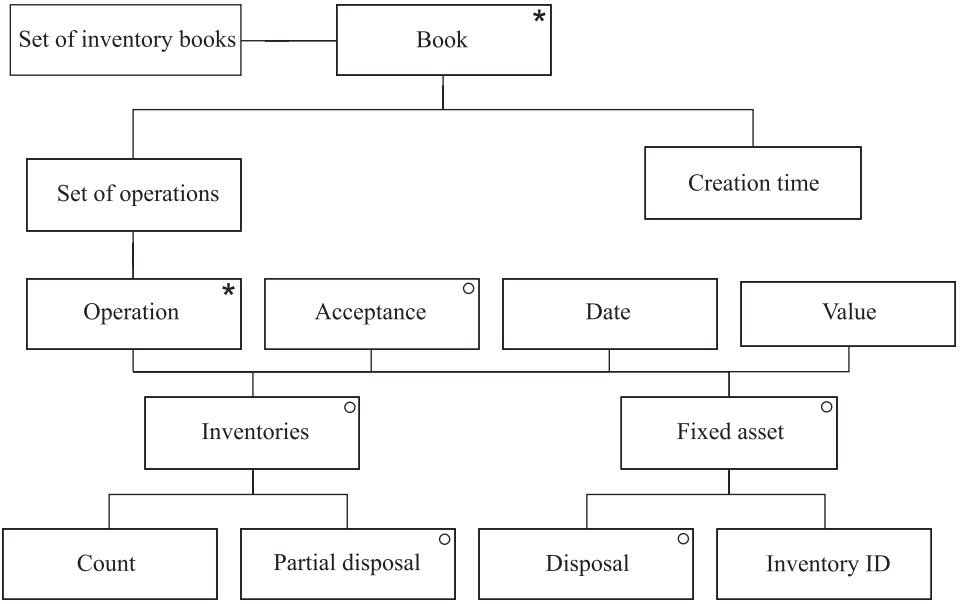


Fig. 4. Input data structure of the inventory books set

appears in both asset type books. Data blocks "Date" and "Value" are general also. First of them denoted the date of the operation and the second one determines the asset value.

In case of the Inventories Book optional operation "Partial disposal" is possible. One inventory can consist of many pieces, therefore operation that dispose only few of them is necessary. Such operation can be also used when one needs to dispose all of the pieces because data field "Count" appears when we operate on the Inventories book. When one substitutes the actual count of the inventory as a count to be disposed, operation of full disposal is made. Data field "Count" is also accessible during "Acceptance" operation made in the Inventories book. In this case it means the count of pieces to be accepted.

During operations on the Fixed assets book operation "Disposal" can be used. There is no need to specify the count because a fixed asset can consist of one element only. In special cases it can be a set of elements of different type (e.g. computer unit). In this case we have a data field named "Inventory ID" which stores unique char string which labels a fixed asset. Operation of "Partial disposal" is not possible here.

One of the most complex data structures of the system is a Set of employers presented on Fig. 5. In this structure not only personal and social data but also all medical, scientific, didactical and administrative data re-

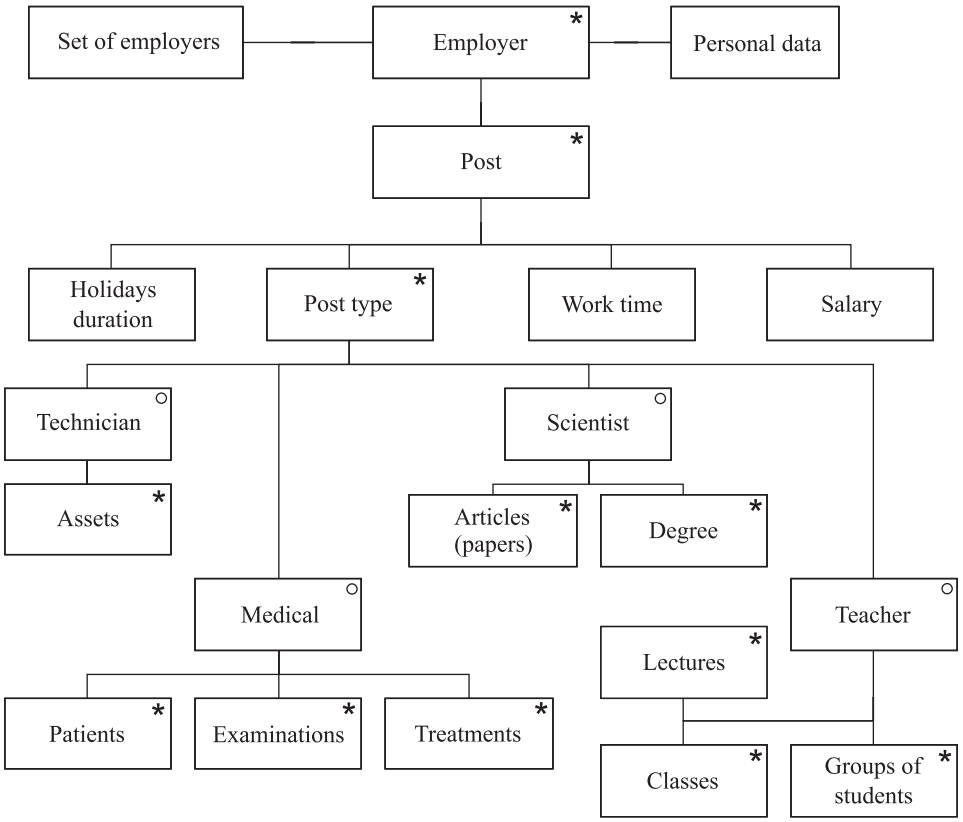


Fig. 5. Input data structure of the employers set

lated to the employee are stored. The most basic data element of this set is the Employer. Every employer has his own "Personal data" block which stores basic data like name, sex, birth date, addresses etc. "Employer" data block determines a person that is employed at the specific post or posts. As one can see on the diagram, every employee can be employed on many posts. The "Post" data block describes a specific post like: secretary, laboratory technician, lecturer, research assistant, cleaner etc. A situation in which one person is employed at many posts is rare but possible.

"Post" data block is bounded to the blocks like "Work time", "Salary" or "Holidays duration", which determines the work conditions at the specified post. Every post is also related to the set of activities bounded to it. It is represented by the "Post type" data block which can be related to one of the four types: "Technician" (administrative activity), "Medical" (medical activity), "Scientist" (scientific activity), "Teacher" (didactic activity). Worth mentioning is the fact that every "Post" can be bounded to several

"Post types". The Post of a secretary should be bounded to technician activity only, but research assistant post must be bounded to medical, scientific and didactic activities simultaneously.

There are several data blocks related to the specific post type, which describes a specified activity. "Medical" data block is related to the set of patients, examinations and treatments that was made or has to be made by an employee. Didactic activity marked as "Teacher" data block on the diagram is bound with "Groups of students", "Classes" and "Lectures" data blocks. Such relation enable the teacher to determine which student belongs to his group and when this group will have classes or lectures. "Groups of students" data block also holds student evaluation and attendance data. Under the "Scientist" data block the information about scientific degrees of an employee can be found. Such information includes obtaining the date of the degree and full article content that was used for. Articles content also is held in such system. "Articles (papers)" data block is designated for this. Such capability enable users to instantaneous access to the scientific production of the whole department. Storing all articles makes the preparation of any scientific survey a simple, and at the same time, a fast task. "Technician" data block is related to the assets that given employee takes care of. It enables the user to check the value of an asset and the date of purchase so that he can determine whether it is still under warranty or no longer.

Patient data stored on the system only includes medical and scientific data, as is presented on Fig. 6. These data do not include management of medicines, bed or insurance billings. Every patient in the "Set of patients" is bounded with a "Personal data" block which contains a simple data like name, sex, birth date, addresses etc. This data determine the person that the patient actually is. Every patient can also have multiple "Contact data" (e.g. to him/her or to the family members). Contact data consist of the name of the person, relationship with the patient, address, phone numbers, e-mail, etc. Most important data block called "Medical history" holds many blocks corresponding to the patient's hospitalizations. Every hospitalization has its "Terms" data block which stores the dates related to it (hospitalization start date, hospitalization end date, pass dates and durations). "Doctor" data block determine which employee takes care of the patient during hospitalization. "Diagnosis" data block stores information about the diagnosis that the patient was admitted to the hospital. During his hospitalization the patient could have been prescribed many different diets. They are all stored in "Diet" data block with information about terms they were in force. Hospitalization bounds to numerous examinations and treat-

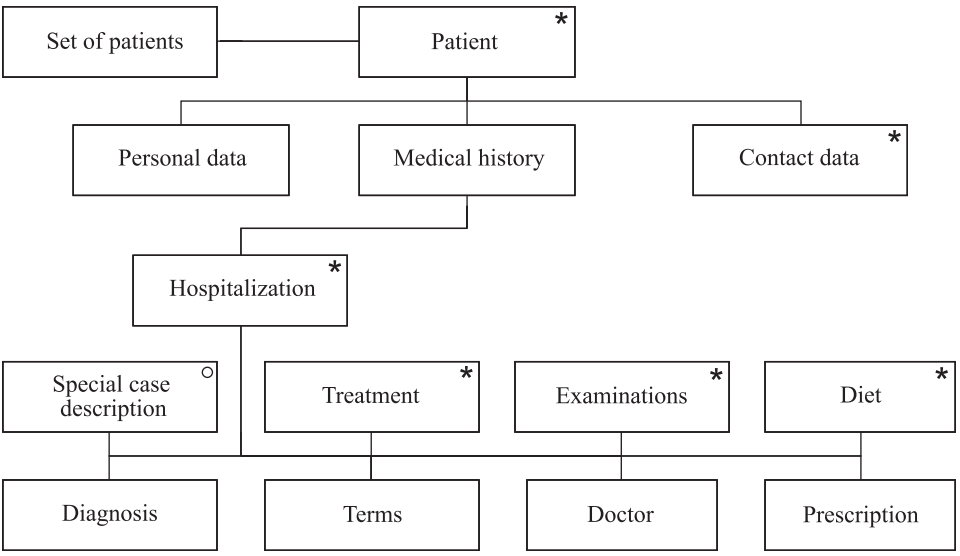


Fig. 6. Input data structure of the patient set

ments. All of them, with their description and results data are stored in the appropriate data block "Examination" or "Treatment". This data draw out all data set of medical history needed for scientific and didactical purposes. There is one more optional data block named "Special case description". It is used to mark a special case that may be interesting for scientist and students. It holds some keywords and case description. Such data are very useful for teachers and scientist and enable them to easily search for the patient's hospitalization that they actually need.

Conclusions

The computer system based on the JSP method ensures compliance with the assumptions made at the stage of analysis and allows tracking its further efficient expansion. The organization of input data structures allows their smooth conversion into elements of the application code responsible for their processing.

The complexity of the system that supports management of a clinical, teaching and research departments makes it impossible to draw up the entire diagram in the current publication. However, the presented fragments prove that with proper effort the presentation of a complete diagram can be feasible.

The elements of the data structure diagram presented above constitute only a small part of the whole project. After correct analysis and decomposition of the problem, the diagram can be completed with new items. Such necessity may occur while discussing the system's functionality with its future users. In such situations, it frequently appears that additional data can increase work efficiency, which should be one of the key benefits of the system implementation [2]. Definition of the diagram for the output and auxiliary data will clarify the tasks of the processes mediating data transformation.

REFERENCES

[1] Adamczewski P., *Zintegrowane systemy informatyczne w praktyce*, Wydawnictwo MIKOM, Warszawa 1998.

[2] Bennett C. J., Computers, personal data and theories of technology: Comparative approaches to privacy protection in 1990s, "Science, Technology & Human Values" 1991, vol 16, no 1, 51–69.

[3] Beynon-Davies P., *Inżynieria systemów informacyjnych Wprowadzenie*, Wydawnictwa Naukowo-Techniczne, 2004.

[4] Flasiński M., Wstęp do analitycznych metod projektowania systemów informatycznych, Wydawnictwa Naukowo-Techniczne, Warszawa 1997.

[5] Jackson M. A., *A System Development Method*; "Tools and Notions for Program Construction", pages 1–26; D Neel ed; Cambridge University Press 1982.

[6] Jackson M. A., *Principles of Program Design*; Academic Press, 1975.

[7] Jackson M. A., *JSP in Perspective*; "Software Pioneers: Contributions to Software Engineering" Manfred Broy, Ernst Denert eds; Springer, 2002.

[8] Ourusoff N., Using Jackson Structured Programming (JSP) and Jackson Workbench to Teach Program Design, "Informing Science" June 2003.

[9] Płodzień J., Stemposz E., *Analiza i projektowanie systemów informatycznych*, Polsko-Japońska Wyższa Szkoła Technik Komputerowych 2003.

Robert Milewski
Anna Justyna Milewska
Department of Statistics and Medical Informatics,
Medical University of Bialystok

Jacek Jamiołkowski
Department of Public Health,
Medical University of Bialystok

Jan Czerniecki
Department of Biology and Pathology
of Human Reproduction, Institute of Animal
Reproduction and Food Research of Polish
Academy of Sciences in Olsztyn

Jan Domitrz
Sławomir Wołczyński
Department of Reproduction and
Gynecological Endocrinology,
Medical University of Bialystok

THE STATISTICAL MODULE FOR THE SYSTEM OF ELECTRONIC REGISTRATION OF INFORMATION ABOUT PATIENTS TREATED FOR INFERTILITY USING THE IVF ICSI/ET METHOD

Abstract: Infertility treatment using IVF methods requires to the collection, storage and analysis of large quantities of various types of data. Created at the University Hospital in Bialystok, system of electronic registration of information about patients treated for infertility using the IVF ICSI/ET method, turned out to be useful in the process of data collection and storage of information about treated couples. However, it does not satisfy the condition relating to the need to analyze the data collected. For this reason, system developers have taken the trouble of improving it with a statistical module that fulfills hopes connected with it. This module consists of two main parts which generally may be called: descriptive statistics and neural network. The first part of the module refers to the designation and presentation of descriptive statistics. They are based on a number of key features of the treatment process, as well as the juxtaposing the designated statistics, broken down into groups defined by the grouping variables. The second part concerns the neural network to predict the efficacy of the treatment. The network which has been used here provides nearly 90% probability treatment failure and can be used for the prediction of negative cases.

Introduction

The infertility treatment is a process that requires the collection, and above all, the analysis of large quantities of specific data. The concept of a hospital information system is well known in medical computer sciences [1], but existing popular software on the medical application market is not designed to collect specialized data. The majority of such applications collect

ISBN 978–83–7431–266–0 ISSN 0860-150X

the most common data, mainly concerning personal information, course of treatment, findings, procedures performed, diagnosis, therapy, medical education and hospital administration [2]. In many cases, they cover the demand of medical units, mainly due to administrative needs and requirements imposed by the National Health Fund [3]. However, due to ever-changing medical procedures, one should not count on the market to find a dedicated medical software application that will be ready to collect data essential to our field. For this reason, specialized systems have been created to handle data related to the narrow scope of medical activities, most often a specific clinic or department.

The system of registration information about patients treated with the IVF ICSI/ET method

One of the specialized applications related to a narrow scope of medical activities is an electronic system for recording information on patients treated for infertility using the IVF ICSI/ET method implemented in the Department of Reproduction and Gynecological Endocrinology of the University Hospital in Bialystok [4]. One of the main tasks of the clinic's activities is to treat couples affected by infertility problem. A number of complex procedures are carried out here, among which many procedures are associated with the process of in vitro fertilization (IVF) combined with intracytoplasmic sperm injection (ICSI), and then with the transfer of obtained embryo (ET) to the uterus [5]. This process involves the storage of vast amounts of information which if recorded on paper is difficult for statistical analysis. Such analysis is essential in maintaining a high efficacy of treatment.

The created application is the result of a collaboration between staff engaged in the treatment of infertility and those involved in programming. It is the result of many years of detailed consultation and cooperation allowing the whole group to fully understand the issue, both from the medical aspect as well as from the development environment in which the application has been made.

The application has been made in programming environment Delphi 2007 [6, 7], while the database in Microsoft Access [8] from Microsoft Office 2003 packet.

The application created in the first stage of the project, and described in [4], has allowed for collecting and managing large amounts of detailed data on infertility treatment. However, it was not equipped with automatic

statistical tools which is necessary to control and maintain an appropriate level of treatment efficacy. Therefore it was necessary to carry out the second phase of the project which should complete the system by the statistical module.

The statistical module of the system for registering information about patients treated with the IVF ICSI/ET method

This module consists of two main parts which generally can be called: descriptive statistics and neural network.

The first part of the module refers to the designation and presentation of descriptive statistics based on a number of key features of the treatment process. It also includes the juxtaposing the designated statistics, broken down into groups defined by the grouping variables. Presented in graphic form, the information concerns the following characteristics – the dependent variables (Fig. 1):

- the number of cumulus oophorus,
- the number of correct cumulus oophorus,
- the number of atretic cumulus oophorus,
- the number of luteinizing cumulus oophorus,
- the number of MII oocytes,
- the number of MI oocytes,
- the number of GV oocytes,
- the number of atretic oocytes,
- the number of 1PN, 2PN and 3PN embryos,
- the number of 2-blastomere embryos,
- the number of degenerating embryos,
- the number of embryos (class A, B, C, D) in the second day of culture,
- the number of embryos (class A, B, C, D) in the third day of culture,
- the effectiveness of treatment (getting pregnant).

Among the grouping variables which allow for the compilation of statistics into two or more groups were (Fig. 1):

- presence (or absence) of individual causes of infertility (ovulation disorders, fallopian factor, endometriosis, PCOS, male factor, idiopathic cause, the other cause),
- the selected type of treatment protocol [4],
- the number of the cycle of treatment for a given pair,
- the results of sperm analysis,
- the year of the treatment.

Fig. 1. Dependent variables and grouping variables included in the statistical module of the system

Presented in Fig. 1 additional options, allow the selection of the presented: measure of location and measure of dispersion. It is possible to choose the arithmetic mean and standard deviation, typically used for normal distributions (or at least symmetrical), and medians and quartiles, usually chosen when the distribution is not normal (and certainly when it is not symmetrical). There is also the possibility of exclusion from the presentation of extreme values, and usage of a logarithmic scale instead of a standard scale.

The graph showing the arithmetic mean and standard deviation for the number of embryos of class B in the second day of culture, presented in two groups, designated due to the presence of polycystic ovary syndrome as a cause of infertility, is shown in Fig. 2.

The largest marks (green rectangles) show the average level in each group and the smaller (red rectangles) point the standard deviation, set aside for both sides than the average. Because the option "extreme values" is selected, there are also the smallest rectangles, which indicate the minimum and maximum values in each group. The graph allows to make a visual assessment of whether the occurrence of PCOS affects the number of embryos class B in the second day of culture. Of course, there is also the option of an enhanced program to automatically perform the appropriate statistical tests that would give the answer whether there are statistically significant differences in the level of the dependent feature between the groups designated because of the grouping features. However, in the current version, the system is limited only to the graphic presentation of results.

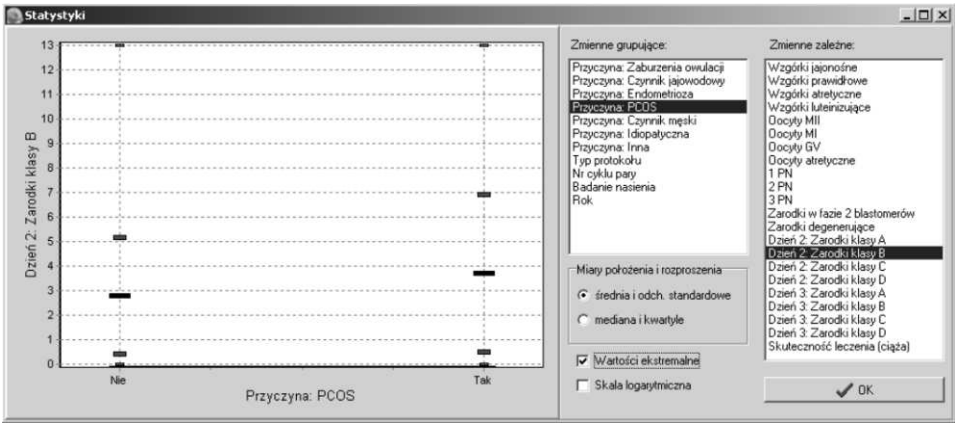


Fig. 2. The arithmetic mean and standard deviation for the number of embryos of class B in the second day of culture

Analogous graphs, except that for measurement: median and quartiles, are shown in Fig. 3.

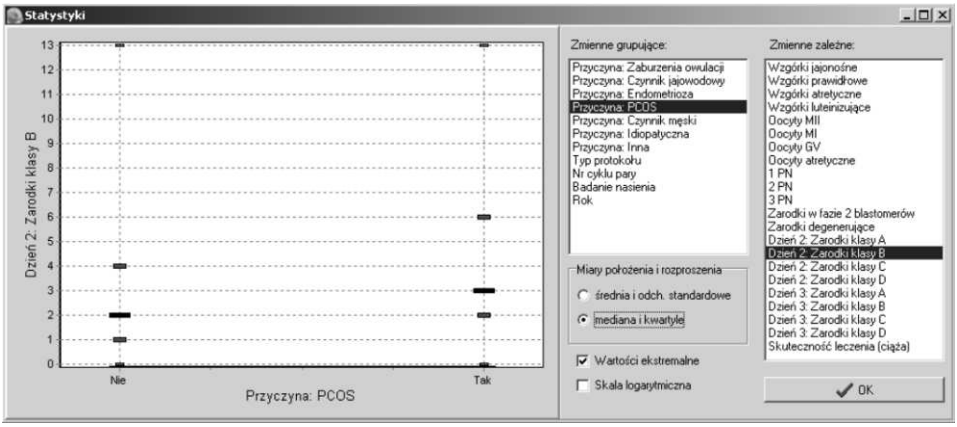


Fig. 3. Median and quartiles for the number of embryos class B in the second day of culture

A similar situation, but after excluding extreme values – minimum and maximum (unselecting the option) is shown in Fig. 4. Of course the scale of the presented graphs is adapted to the smallest and largest values of both analyzed groups.

Most of the presented characteristics are shown by charts, presenting their location and dispersion measures, possibly including the extremes (the

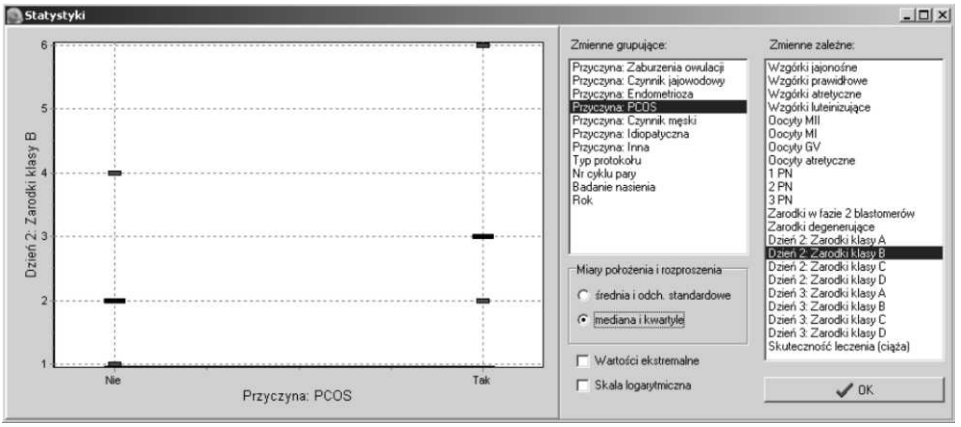


Fig. 4. Median and quartiles without extreme values for the number of embryos class B in the second day of culture

equivalent of the classical box-whiskers plot). However, there is one (but probably the most important) feature, which is presented using a bar chart. This feature means the effectiveness of medical treatment, that is the percentage of pregnancy (Fig. 5).

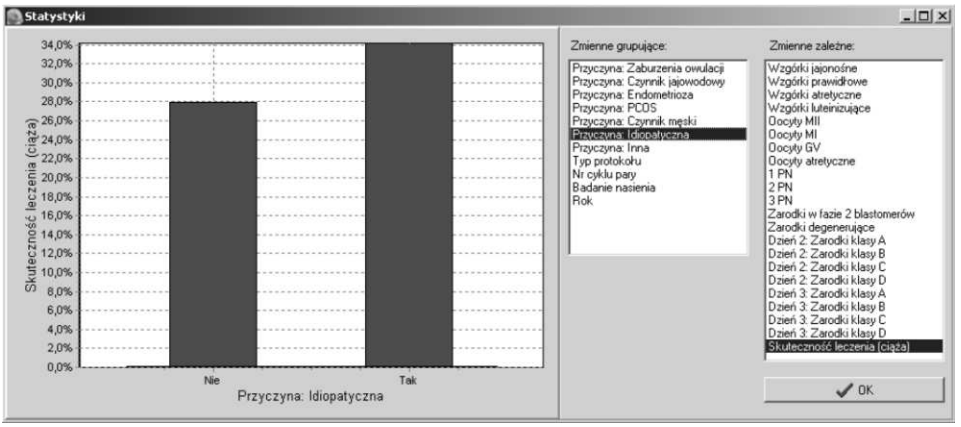


Fig. 5. Efficacy of treatment in groups designated due to the occurrence of idiopathic infertility

The graph shows a clear difference in efficacy of treatment between the two groups. In the group of patients with an idiopathic infertility, the treatment efficacy is higher (about 34%), while in the second group, it reaches 28%.

The neural network to predict efficacy of treatment

The second part of the statistical module concerns the neural network to predict efficacy of treatment. In the paper [9], authors described the trained neural network which with a probability of nearly 90% predicts failure of treatment using the IVF ICSI/ET method and can be used for prediction of negative cases. This neural network was implemented to the system for collecting information about the treatment. The outcome of the calculation is presented graphically. Colored rectangle shows the probability of the success of treatment for the couple, based on the prediction made by the trained neural network. The cut-off level is marked, which is the border between the positive and negative prediction.

Fig. 6 shows a case in which the rate of prediction is strongly shifted to the right, which theoretically predicts the success of the treatment. However, the trained network is not suitable for accurate prediction of positive cases, but only the negative. In this case the optimism of the treated couple can increase only the fact, that the rate did not hit the left edge of the rectangle, which would very likely conclude that the treatment in this cycle will fail.

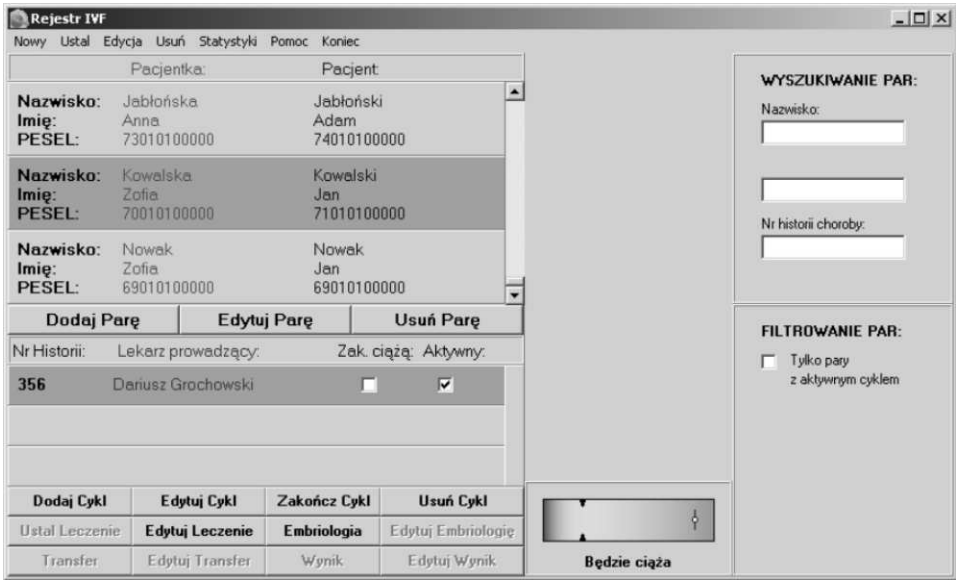


Fig. 6. The neural network forecast on the success of infertility treatment

A situation in which the prognosis of the network is not optimistic, is presented for another couple (Fig. 7).

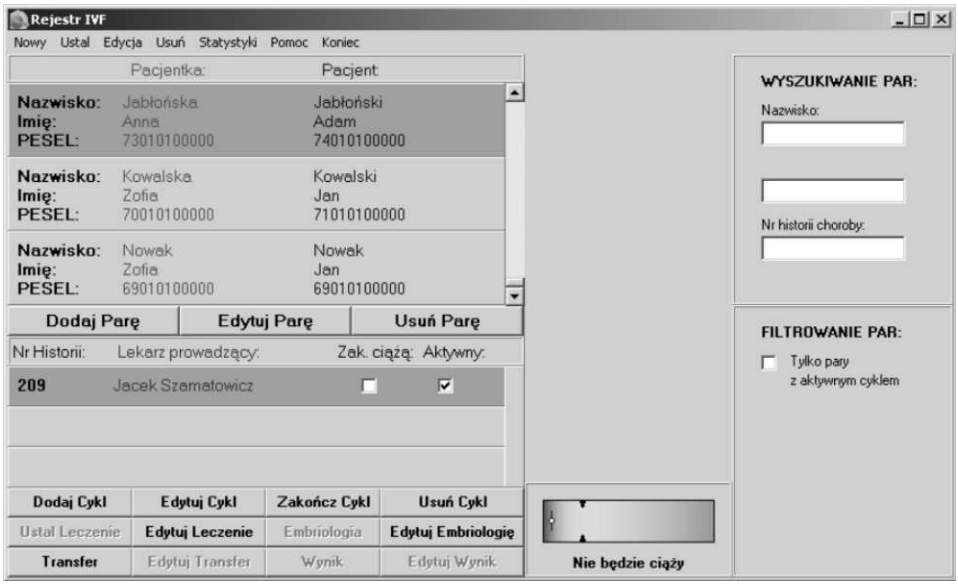


Fig. 7. The negative prognosis of the neural network, concerning the success of infertility treatment

Conclusions

The system of collecting information about treated patients after reaching the stage of stability and testing by the staff not involved in the process of programming, has been implemented for use in the Department of Reproduction and Gynecological Endocrinology in the Medical University of Bialystok. Until now data for more than a thousand pairs being treated for infertility using the IVF ICSI/ET method has been collected. However, the statistical module of the system is just being introduced in that clinic.

One of the first observed effects of applying the described above system was the improvement of the accuracy and reliability of input data [4]. This was mainly caused by forcing the need to fill some key fields of applications, which was a response to the frequently encountered lack of data. It is hoped that the dissemination of the statistical module will increase the awareness of physicians about the effectiveness of their choices made, which should result in a further increase of efficacy of the treatment. Both, graphically descriptive statistics and implemented neural network play an important role here to recognize the negative cases.

Tasks for the future

The created system with the statistical module has been implemented for use, but it means the completion of the second (not final) stage of the planned work.

The third stage of the application development will be to adapt it to network working, with the possibility of multi-user access data. Currently, the database is built on a single, selected computer. After adjusting it to network working, many users from different locations might also add, edit, and analyze patient data. This step involves not only ensuring a stable multi-user access to the data, but also the security of information collected.

After completion of the third stage of work, the application will be a major and comprehensive system, that could be adopted as a standard for the collection of information in units dealing with infertility treatment using IVF methods.

REFERENCES

[1] Piętka E. Zintegrowany system informacyjny w pracy szpitala. Wydawnictwo Naukowe PWN, Warszawa 2004.

[2] Kącki E., Kulikowski J. L., Nowakowski A., Waniewski E. Systemy komputerowe i teleinformatyczne w służbie zdrowia. Biocybernetyka i Inżynieria Biomedyczna, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2003.

[3] Trąbka W. Szpitalne systemy informatyczne. Uniwersyteckie Wydawnictwo Medyczne Vesalius, Kraków 1999.

[4] Milewski R., Jamiołkowski J., Milewska A. J., Domitrz J., Wołczyński S. 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), 2009.

[5] Radwan J. (pod red.) Niepłodność i rozród wspomagany. Wydawnictwo Termedia, Poznań 2005.

[6] Boduch A. Delphi 2005. Kompendium programisty. Wydawnictwo Helion 2005.

[7] Wybrańczyk M. Delphi 7 i bazy danych. Wydawnictwo Helion 2003.

[8] Feddema H. Microsoft Access. Podręcznik administratora. Wydawnictwo Helion 2006.

[9] Milewski R., Jamiołkowski J., Milewska A. J., Domitrz J., Szamatowicz J., Wołczyński S. Prognozowanie skuteczności procedury IVF ICSI/ET – wśród pacjentek Kliniki Rozrodczości i Endokrynologii Ginekologicznej – z wykorzystaniem sieci neuronowych. Ginekologia Polska, 80 (12), 2009.

Piotr Ziniewicz
Paweł Malinowski
Stanisław Zenon Mnich
Department of Statistics and Medical Informatics
Medical University of Bialystok

CLINICAL DEPARTMENT INFORMATION SYSTEM DEVELOPMENT

Abstract: Advanced informatics systems should reflect the diversity of its users and their needs. An important part of the development process is to study the expectations of its users. This, in result, enables the creation of many different models reflecting specific aspects of such a system.

Introduction

Dynamic development of medical informatics and the growing users requirements, caused informatics systems to evolve into more and more complex applications. Such systems consist of multiple cooperating components located on different machines and communicating with each other using different methods. Programs designed to perform individual tasks become a relic of the past and are slowly replaced by more complex and expanded systems for organized work management. JeNaK management system (from Polish Jednostka Naukowo-Kliniczna – "Clinical and Scientific Unit") is based on the current needs of doctors, researchers, teachers and administrative workers employed at the Medical University of Białystok. Considering the huge problem complexity, the key to the success of the task is to properly build a data structure, logic and application interface. This structure should open the possibility to easily support its expansion in order to meet the changing needs of its users. The whole system should include many possible action scenarios, taking into account the individual needs of each of its user. The aim is to define the structure of the system, meeting the requirements stated above, and its implementation. In this paper, primary collection entity participating in modeled information system will be defi-

ISBN 978–83–7431–266–0 ISSN 0860-150X

ned together with their relational dependencies. The next step will be to define a set of parameters for each entity with their type and data range. Firebird database engine and Embarcadero RAD Studio 2010 was used to implement the system.

Information systems

Last 30 years have brought great changes in healthcare information technology support. The introduction of computers and information systems in healthcare resulted from the need to save on very large and constantly growing healthcare costs. Additional reasons were the increasing amount of information that describes the patient's health state, the need of archiving and granting authorized access to that data, or its transfer between medical units. Now, medical data not only "goes after the patient", but also undergoes the process of wider scientific analysis, for inventing new, more effective and cheaper therapies. Today on the software market there are many hospital information systems that manage and circulate information (van Bemmel, Musel 1997). There are even standards for communication between different systems, such as HL7 (Benson 2010) or DICOM (Pianykh 2008). Many of them contributed for the progress in medicine, patient service at the same time reducing medical costs.

Most of the modern hospital information systems are focused on medical and organizational information circulation, essential for the hospital operation as a therapeutic unit. However, there are no systems that supports potential clinical nature of such "units", including training and research carried out within. The JeNaK system tackles these problems. One of this system's assumptions is large versatility in creating a system handling practically any clinical unit including research projects, without the need for greater modification.

The informatics system is a computer-assisted information system. This information system is defined as a combination of procedures for collecting, processing and transmitting information to support and improve the process of management, decision-making and control. It is a necessary component of any organization or company, including healthcare units.

Construction of the information system

By analyzing the structure of the exemplary information system, one can extract four basic elements (Trąbka/1999, p. 19): hardware, software, data, and users. Hardware is a collection of all specialized devices used for

system construction. Continuous technological progress makes it cheaper and may face increasingly higher user requirements. The essential part is a desktop computer. Usually it is assumed that it has specified parameters, in order to ensure smooth operations, increasing user productivity and comfort. Computers are usually connected together with network devices to exchange information. Today, there are virtually no new computer systems that function on a single computer. Another important part of the information system is software. Software is a combination of programs which supervise work of computers, manage databases and allow the interaction of the system with the users, perform their commands and protect security and confidentiality. Hardware without the end user software is useless. Users are a central element of the information system. The system itself is created to meet their needs. One can distinguish two groups of system users: professionals and end-users. Professionals create and manage quality software and define hardware requirements. End-users are a team of people who work with the finished system. Information system supports data processing to provide ability to enter and process information for potential users.

The process of creating an information system

Creating and managing the rich information system is a fairly difficult task. This is not something new. Already at the turn of the 1960s and 70s, 20th-century programmers first observed fiasco of the major programming projects (Sommerville 2003, p. 71). Software has been delivered late, was unreliable, inefficient and often exceeded the estimated costs of its creation (Brooks 2000, s. 40). Failed projects was a result of incomplete engineering mechanisms and software modeling. Software development quite significantly differs from other manufacturing processes, because the product – software – is "virtual". When producing physical object one can observe the process of its construction and clearly compare with accepted assumptions. Software cannot be touched or seen. Furthermore, programmers are unable to test its functionality for a considerable length of time for the process of its creation is long. Secondly, there are no standards of software creation. In many other areas of engineering, the manufacturing process is checked and tested, and also well known and understood. It is impossible to determine if a concrete process of software creation will lead to the occurrence of specific errors.

Now software is created using objects and classes. These concepts have direct connection with the reality. Biological equivalent class is a species of

animals, and individual objects can be thought of as individuals of these species. Note that some species share with others certain properties (may have a common ancestor). In addition to defined measurable attributes (such as growth or weight), class can have operations (term "behavior" can be used). Classes have property inheritance, so they may use certain operations defined in other classes. Term "entity" instead of "object" is used when turning to database category.

One of the most important achievements in the field of software engineering was the universal modeling language (UML-Alhir 2007). UML Specification 2.2 provides definitions of 14 models that describe many aspects of the system that is created. Each of these models is presented using an appropriate diagram. These models may be divided into two groups in terms of their content and relevance:

1. Structural group of models, these reflect a static system structure using objects, entities, attributes, relationships, operation. These are:
   1.1. class diagram – describes the structure of the system by presenting classes, their attributes and associations,
   1.2. component diagram – describes how the system is divided into components and shows dependencies between them,
   1.3. profiles diagram – describes the so-called metamodel of the system that is a description closer to natural language,
   1.4. structure diagram – describes the internal structure of classes and operations within this structure,
   1.5. implementation diagram – model equipment that is used in the implementation,
   1.6. objects diagram – a complete or partial view of the structure of the system in a specific time,
   1.7. packages diagram – describes the breakdown of the system into logical parts and their relationship to each other,
2. Behavioral group of models – these reflect the dynamic structure of the system, showing the influence of objects on each other, interactions and internal states. In this group one can specify:
   2.1. activities diagrams – show running system and activities of the user step-by-step
   2.2. use cases diagrams – show full or partial functionality of system
   2.3. states diagrams – show the internal state of the object and its changes
   2.4. interaction diagrams – these are rich diagrams that describe the system's behavior. Due to precision, they are created only for critical elements of the system. One of the interaction diagrams is

the communication diagram which shows the interaction between classes as an orderly string message. Sequence diagram shows the interaction between objects as an orderly string message, in addition, it specifies the lifetime of these objects. Diagram of interaction is the general view of the interaction of objects and classes. Each of the nodes in this diagram is a diagram of interaction itself. Time dependency diagram is a specific type of the interaction diagram in which attention is focused on time dependencies.

During the design of JeNaK part of these diagrams were created to make project management more efficient. However, their volume significantly exceeds the volume of this work. Therefore, only the most interesting elements of the system are presented here with their descriptions.

JeNaK functionality

The creation of use case diagrams was the first step during the design of the system. These diagrams show potential users (actors) and system functions (use cases) that can be used by them. Linking actors with functions executed directly by them is marked as a straight lines. Complex functionality of the system can be then split into a series of simpler steps, creating a grid of links between them. There are two types of links between use cases:

1. inclusion marked by «include» – execution of one function always includes an execution linked function,
2. extension marked by «extend» – execution of one function may include an execution linked function.

The example of using these two kinds of calls is shown in Figure 1. Diagram of figure. 1 reflected the patient's visit to a doctor. During the visit it is necessary to conduct an interview with the patient (anamnesis). However, this not always involves a prescription writing. The same combination of operations must be offered by the created system.

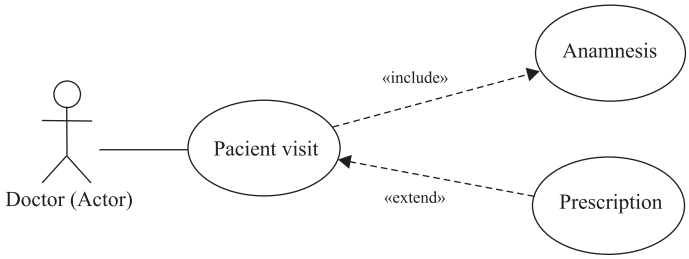


Fig. 1. An exemplary use case diagram

Additionally, scenarios are created for each of the functions that are called by the user. They consist of 5 parts:

1. Participating actors – lists of actors that may call specific functionality,
2. Basic events plot – they describe basic, the simplest course of actions using the sequence listing of interactions between an actor and a system,
3. Alternative events plot – describes all possible aberrations from the basic plot together with their consequences,
4. Time dependency – provides information about how often action is called and how it is time-consuming,
5. Results obtained by actors after finishing use case – gives final effects arising from the execution of a function.

The process of creating a diagram started with conducting surveys concerning duties performed by employees of individual clinical and scientific units. Data was collected from personnel. Below are few examples of many survey questions:

1. What is your role in the unit? (position and brief description),
2. Please list all duties performed in your work,
3. What documents and information you work with and which ones should you archive?,
4. Which steps related to your work you wish to automate? (in the context of the above questions),
5. How does students knowledge verification system look like in the context of your subjects?,
6. How do you collect material for research?,
7. How does staying patient records system look like?

As a result of information analysis gathered from surveys, four types of workers were distinguished based on duties they performed. These types are not disjoint, which means that one employee may perform several function types.

1. Technicians – engaged in administrative and organizational activities and caring for the clinic's equipment. They care about efficient allocation of students into groups and classes, and coordinate this with teachers. An example of a technician is a secretary or a laboratory technician,
2. Scientists – people who conduct research. Often they are interest in unusual cases of patients or rare diseases or complications. Scientists have their acquits in form of scientific publications and gain degrees and awards. Data on the acquits is regularly transmitted to appropriate college structures,

3. Medicals – they deal with patients therapy process, from the patient reception at the ward, forwarded by treatment and commissioning medical check-up, until the end of his/her stay. They deal with both registration and execution, make medical records, register patients observation and perform parts of research. An additional function is to find the unusual cases of illness and inform scientists about them,
4. Teachers – are involved in the didactic process and everything related with it. They are responsible for preparation of student lists and class scheduling, along with reservation of rooms, in which they are assisted by a technician. Teachers also manage "electronic students log" and electronic presence list in accordance with the nature of the imposed teaching obligations. They are also responsible for the preparation of teaching materials to support the learning process and control its effects.

Figure 2 presents a use case diagram of functions related to the administrative job of a secretary employee. Due to the transparency of the diagram it was limited only to the administrative part of the secretarial work. Technicians' duties also spread to the teachers' and scientists' areas. The most common work of a secretary technician is mail management. As it may be seen on the diagram it was split to the three functions: Incoming mail registration, Outgoing mail disposition and Application and requests management. The third one is worth paying more attention. Any application requests funds and results with assets change, therefore it was distinguished as a separate function of mail management. Application for purchase demands to check the funds state and eventually to accept an asset that was purchased. Such function is related to the "Surveillance state funds" and "Keeping the books counting" functions. Despite the fact that all requests and applications that were realized, it automatically changes the state of the funds. There is an additional function named: "Correcting the funds state". This function is a plain "backup" in the case if the real state of funds is different from that counted by system one. All functions of the "Socio-occupational cases" are related to the "Outgoing mail disposition" function due the fact that their files have to be eventually send to the Human Resources Department. "Application for purchase creation" use case scenario is presented below.

1. Actors:
   1.1. Technician,
2. Default event string:
   2.1. System displays "External mail management" window
   2.2. Actor selects "Application and request management"

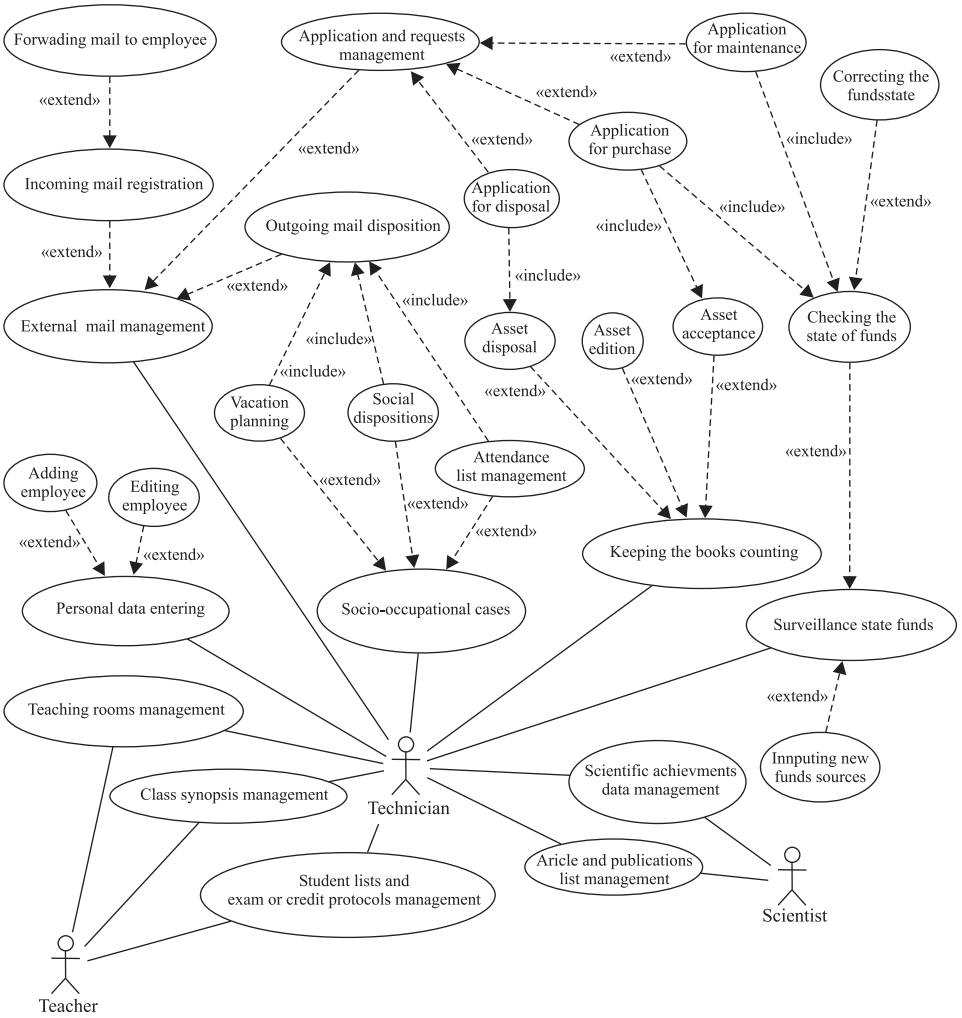


Fig. 2. Secretary use case diagram

2.3. System displays “Application and request management” window
2.4. Actor selects “Application for purchase”
2.5. System displays “Application for purchase” form
2.6. Actor fills subject of purchase description
2.7. Actor acknowledge the value of a subject and fill it to the form
2.8. Actor selects “Add funds source” function
2.9. System displays “Funds state” window
2.10. Actor selects account and closes window
2.11. System displays “Enter amount” dialog window

2.12. Actor enters maximum cash amount to be taken from selected account
2.13. If not all funds sources selected go to step 2.7
2.14. Actor fills reason of purchase
2.15. Actor selects OK option
2.16. System checks if value of subject is less or equal to the funds source amount.
2.17. Actor prints out a paper version of application and send it to the acceptation of head.

3. Alternative event string:
   3.1. Value of subject is greater than funds source amount,
      3.1.1. System displays a warning and returns to the „Application for purchase" form,
4. Time dependencies:
   4.1. Frequency of usage: normal: ~ 5 times a year, pile up season: ~ 5 times
   4.2. Expected pile up: one time a year at the end of year,
   4.3. Typical realization time: $\sim$ 2–5 min,
   4.4. Maximal realization time: unspecified,
5. Values obtained by actors after using:
   5.1. Prepared electronic and paper versions of application,
   5.2. Mail database record,
   5.3. Funds database record (funds reservation).

Overview of JeNaK internals: sample state diagrams

After a range of activities assisted by JeNaK system, it is necessary to define stored data. This is done through the extraction of the set of objects (material or not) participating in the use case diagram. Examples of objects include employee, patient, student (as physical objects) or diet, thesis, score (as intangible objects), etc. All this data is saved in the database as entities. Part of this database, which is related to didactic process was presented in (Ziniewicz, P.; Milewski, R.; Malinowski, P.; Mnich, S. Z. 2010).

Use case diagram focuses on the presentation of the operations supported by the system JeNaK. Functionality of the system is completely separated from the database content. However, experience shows that such separation is quite illusory. Theoretically, every system process stored data according to user's wishes. For the sake of common sense and internal integrity the system attempts to monitor user's actions and warns before,

or eventually denies unauthorized operations execution. Most modern software systems are event-driven, which means that they continuously wait for the occurrence of some external or internal event such as a mouse click, a button press, a time tick, or an arrival of a data packet. After recognizing the event, such systems react by performing an appropriate computation that may include manipulating the hardware or generating "soft" events that trigger other internal software components. Once the event handling is complete, the system goes back to waiting for the next event. The response to an event generally depends on both the type of the event and on the internal state of the system and can include a change of the state leading to state transition. The pattern of events, states, and state transitions among these states can be abstracted and represented as a finite state machine (FSM).

The concept of an FSM is important in event-driven programming because it makes the event handling explicitly dependent on both the event-type and on the state of the system. When used correctly, a state machine can drastically cut down the number of execution paths through the code, simplify the conditions tested at each branching point, and simplify the switching between different modes of execution. Conversely, using event-driven programming without an underlying FSM model can lead programmers to produce error prone, difficult to extend and excessively complex application code.

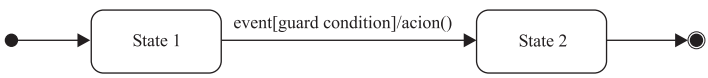


Fig. 3. Sample state diagram

The UML state diagrams are directed graphs in which nodes denote states and connectors denote state transitions. Figure 3 shows sample UML, where one can distinguish:

1. entry point, denoted by solid circle
2. exit point, denoted by circle with solid circle inside
3. internal state, denoted by rounded rectangle labeled by state name
4. state transition from one to another, denoted by arrow, and specially labeled

The arrow label contains information about transition. Label has format event[guard condition]/action(). As mentioned before, event can be generated (triggered) by hardware, software or user. There are also events that occur in specific time (denoted by when(time/date)), or after a specific period of time (denoted by after(duration)). When the event is triggered,

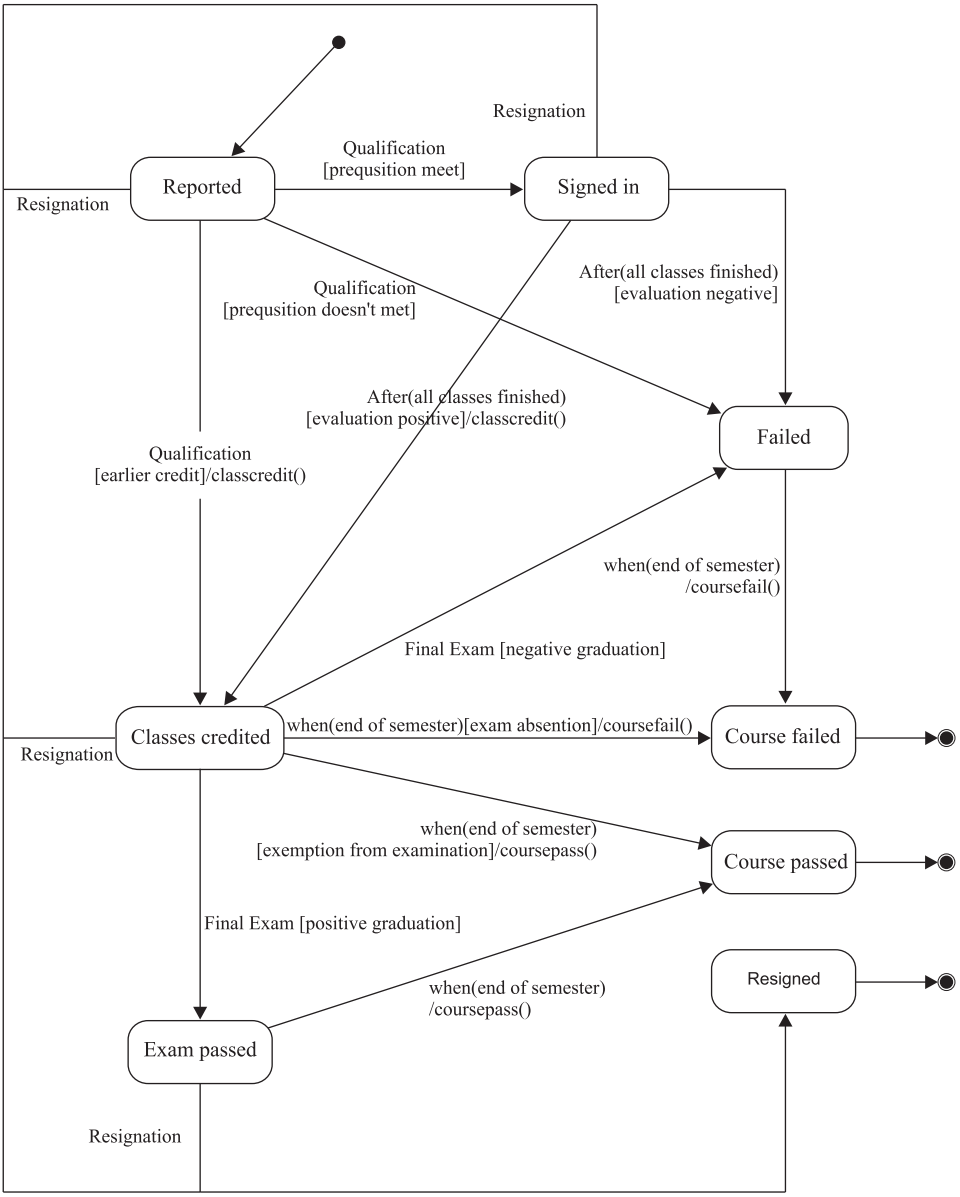


Fig. 4. Student state diagram

guard condition is checked, and if is met, transition to another state occurs, and action() is performed. Latest UML standard (Sommerville 2003) defines many other features in addition, but they will not be used in this work.

State diagram describes the internal state of the selected object in the system, as well as its changes. To create such a diagram it is necessary to

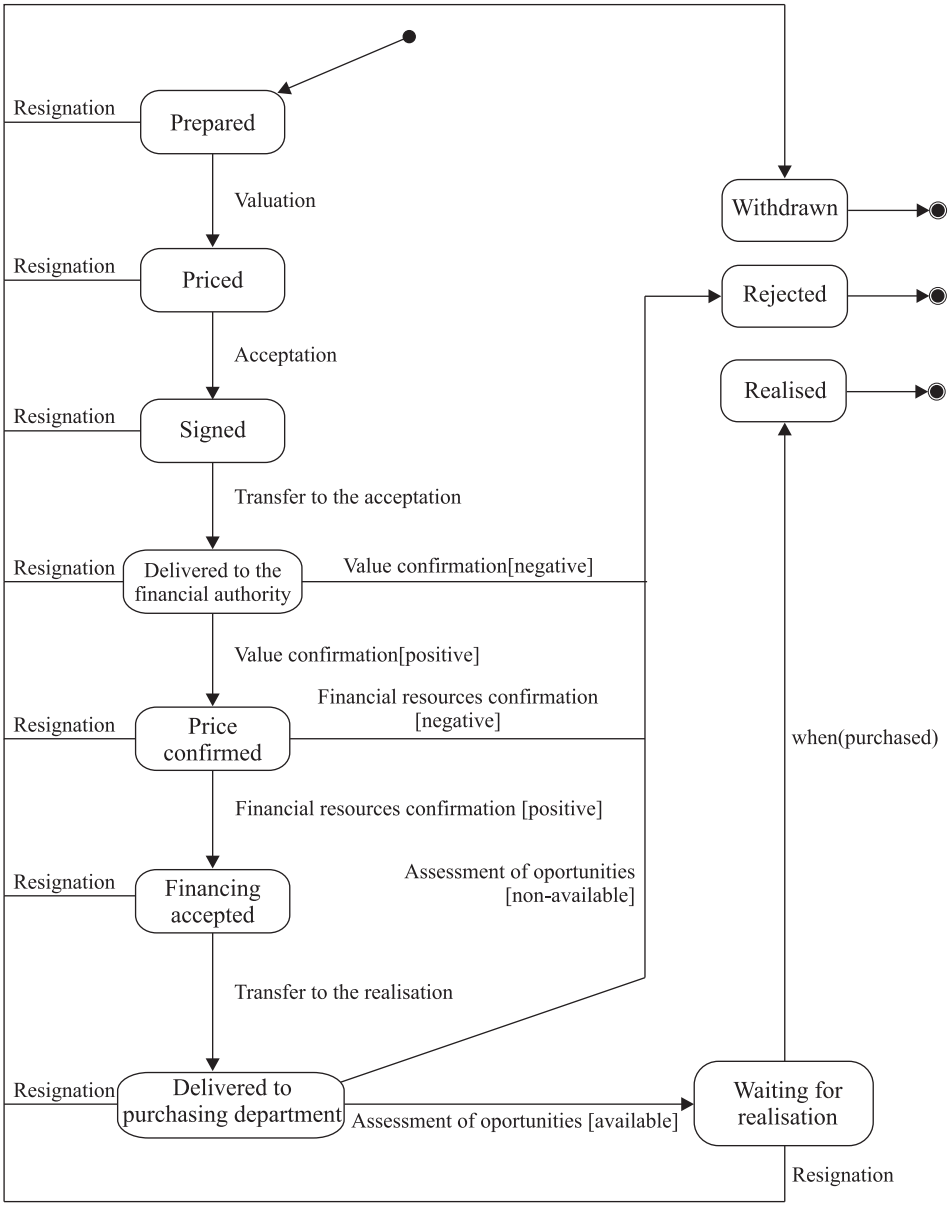


Fig. 5. Application for purchase

track most operations supported by the system, and examine the data on which they operate. Typically, such diagrams are created for data that must be saved for a considerable time, often going through many changes during processing. Usually, for such object, end user will distinguish properties like

status, type, etc. especially where a large number of states and transitions creates fairly dense network of connections.

Inside JeNaK system there are many objects that change their states. Very good example of states diagram can be presented on the base of "student" object, which is presented in Figure 4. To obtain a credit student has to report to the department first. At this stage it is checked if he meets requirements or not. If the student meets requirements, he/she is treated as a legal member of a course and moves to the "Signed in" state. When all classes are completed he/she can change the state to one of the two: "Classes credited" or "Failed". Which of these two will depend on the Evaluation results. At this stage student the waits for an exam and then changes his state to "Course passed" or "Course Failed" accordingly. Worth of attention is the fact that states "Failed" and "Course Failed" are distinguished. The second one is forced by the time and definitive and the first one is more like "continuous". It sometimes happens that the student completed the course before. In such case the during qualification event we have condition: "earlier credit". At the "Classes credited" we have an "exemption from examination" time condition also to omit standard path of the student. At any stage of the path the student can also resign. In such case he/she is treated differently to the one who failed the course because it was his/her conscious decision.

One of the most common documents in secretarial work is an application for purchase. Such document has many stages of preparation which are presented on Figure 5. It has to be prepared, priced, accepted by the head of the department, confirmed by the financial department, etc. At any stage it can be withdrawn by a source user (employee). Subject of the purchase can be priced by the user or by an expert. It does not matter because the application for purchase even when priced has to be confirmed by an expert – once during "Value confirmation" event, and the second time when it is delivered to the purchasing department. It is worth noticing that between "Preparation" and "Waiting for realization" stages the time lapse may be significant. Meanwhile some subjects such as computer equipment may be withdrawn from the market. Prices may also change significantly during this time. "Assessment of opportunities" event has to deal with such cases. At the "Delivered to purchasing department" stage, the expert examines the possibility of purchase and decides if the application will be realized or rejected due to the lack of purchase possibility. Another point of interest are that exit points from "Withdrawn" and "Rejected" which are distinguished. In the first case we deal with a conscious decision of the user, in the second one the user is independent.

Conclusions

Typical hospital information systems have been designed to manage medical and financial aspects of the clinic. They store data related to hospitalized patient, drugs disposal, healthcare insurance organizations accounting, and internal accountancy. Presented work shows a project of a system that manages clinic and scientific information. Unlike typical SSI, focus was put here on scientific, educational and administrative aspects of a single unit.

This system does not replace the whole SSI system, but cooperates with it using standards DICOM and HL7, extending its functionality to include management of research work and finances of an individual unit. It may also cooperate with other systems operating in the administrative units. This extension mechanism and system flexibility enables easy customization of the nature of the research work carried out in a given period, however this requires further research.

REFERENCES

[1] van Bemmel J. H. i Mused M. A. 1997. *Handbook of Medical Informatics*. Berlin, Germany: Springer-Verlag.

[2] Pianykh O. S. 2008. Digital Imaging and Communications in Medicine A Practical Introduction and Survival Guide. Berlin, Germany: Springer-Verlag.

[3] Benson T. 2010. Principles of Health Interoperability HL7 and SNOMED. London: Springer-Verlag.

[4] Sommerville I. 2003. *Inżynieria oprogramowania*. Warszawa: Wydawnictwa Naukowo-Techniczne.

[5] Brooks F. P. 2000. *Mityczny osobomiesiąc*. Warszawa: Wydawnictwa Naukowo-Techniczne.

[6] Alhir S. S. 2007. *Guide To Applying The UML*. New York: Springer-Verlag.

[7] Trąbka W. 1999. *Szpitalne Systemy Informatyczne*. Kraków: Uniwersyteckie Wydawnictwo Medyczne „Vesalius".

[8] Ziniewicz P.; Milewski R.; Malinowski P.; Mnich S. Z. 2010. Informatyczny system zarządzania jednostką naukowo-kliniczną Uniwersytetu Medycznego w: *Współczesne wyzwania strukturalne i menadżerskie w ochronie zdrowia*. Praca zbiorowa pod red. Romana Lewandowskiego i Ryszarda Walkowiaka. Olsztyn: Olsztyńska Wyższa szkoła Informatyki i Zarządzania im. prof. T. Kotarbińskiego.