The use of the basket analysis in a research of the process of hospitalization in the gynecological ward

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

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

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

Abstract. The progress of science and technology allows to create increasingly complex and detailed databases. It leads to the development of modern data analysis methods. Information collected in medical facilities is characterized by great diversity. In this paper we present a description and application of one of the data mining methods, the basket analysis. It will be used on data describing the process of hospitalization on the gynecological ward. A way of searching for association rules the using basket analysis will be presented. This opens great opportunities for the interpretation obtained results.

Introduction

Many companies, government organizations, research centers and medical facilities create extensive databases. Information gathered for many years is a rich source of knowledge. Skilful mining and analyzing allows to improve the operation of facilities, optimize processes and detect irregularities. Basic statistical methods allow to detect some relationships, but may prove that the most interesting and invisible at first glance observations can get away from the researchers note. Meaningful analysis of databases is made possible by data mining techniques, one of them is the basket analysis, also known as the association rules. This method, in contrast to traditional instruments, can detect links between factors, which are rare, for example, relate to an unusual disease entity, which is accompanied by a number of recurring symptoms. A major advantage of this method is the detection of co-existence of many characteristics of an object. Basket analysis indicates even very complex implications. Such observations are undetectable using traditional statistics, especially if the database is large.

ISBN 978–83–7431–296–7 ISSN 0860-150X

Basket analysis – method

Basket analysis is used to find association rules. The term can be understood as all the implications describing collected categorical data. This technique allows for searching rules of the kind: If X then likely Y. At the same time both body and head may incorporate several factors. Appropriate medical database would allow to build the following sentence, for example: if a patient from ward A smokes cigarettes and suffer from obesity, it is probably that the treatment fails.

In basket analysis fast processing of huge data sets is made possible by using a priori algorithm [1] and its subsequent modifications [2]. Basic concepts, associated with it, relate to sales data:

- $I = \{i_1, i_2, \ldots, i_m\}$ – *items* is a set of binary factors.
- Any subset $X \subset I$ is called *itemset*, in particular subset with k elements: *k-itemset*.
- $D = \{(id_1, T_1), (id_2, T_2), \ldots, (id_n, T_n)\}$ transactions database, where:
  - for any j $id_j \subset TID$ is a unique transaction identifier,
  - for any j $T_j \subset I$ is a set of purchased goods.
- $s(X)$ is the number of transactions containing a set X.
- Association rule is called the implication "if X then Y", where both body and head can mean a single category, but also a list of categories (codes):

$$X \Rightarrow Y,$$

  where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$.
- Let n – number of transactions, $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$, then indicators describing the rules:
  - support

$$support(X \Rightarrow Y) = \frac{s(X \cup Y)}{n},$$

  - confidence

$$confidence(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)},$$

  - correlation

$$correlation(X \Rightarrow Y) = \frac{s(X \cup Y)}{\sqrt{s(X) \cdot s(Y)}}.$$

- Determined are also two constants: min_sup and min_conf, which mean the minimum support and minimum confidence, which characterize searched association rules. Appropriate deployment of these limits causes find only the relevant implications.

Construction of association rules is done in two steps:

I. Finding all frequent sets $X \subset I$, it means $\frac{s(X)}{n} > min\_sup$.

II. Generating association rules based on frequent sets $X \subset I$, for example by dividing each of these two subsets such that $A \cup B = X$, $A \cap B = \emptyset$ and $support(A \Rightarrow B) > min\_conf$.

Any computer program would not be able to check in real time all the rules. Using the a priori algorithm optimizes the search process and enables the identification of the most important relationships. Using basket analysis the researcher selects thresholds min_sup and min_conf. At this stage, he can decide whether rules apply to typical events, or whether they have a low support and describe anomalies.

Basket Analysis in Medicine

Not without reason, association rules is a tool that was used in trade. This method is ideal for analyzing shopping baskets [3], it indicates preferences and habits of customers of supermarkets, but also online stores. Familiarity with lists of products most often purchased together makes it possible for the seller to arrange them in the store (on the website) to a client looking for them so he will not miss the promotion addressed to him. The results of basket analysis sometimes confirm associations which are visible at first glance and are obvious to the retailers. However the main purpose of this method is to identify the hidden rules. One may wonder what benefit the discussed tool may bring in medicine. Literature indicates the possibility of using basket analysis in medicine mainly as follows:

- Analysis seemingly exploited of data, where the traditional tools were used [4].
- Exploring the relationship between medical concepts, in order to predict future discoveries (new relations between concepts) [5].
- Diagnostic decision support [6].

Medical data that describe the many features are a great facility that can be analyzed using association rules. It works for a large number of multidimensional variables. Description of each patient can have multiple characteristics, they include parameters describing the condition of the patient, laboratory data, as well as genetic. Collected data are extremely diverse: numeric, ordinal, nominal, as well as images. Organizing, building a multi-way tables for such data or describing by basic statistics can be cumbersome

and sometimes practically impossible. Medical data in many cases can be analyzed using this method. Some of them, first of all nominal variables, can be instantly used to build the association rules. Others may be subjected to transformation, to become the clear implication component. There are various methods and suggestions for standardization of the collected data [7]. It is important that finally obtained results were clear.

Analysis of the basket allows the search of the complex relationships between the characteristics included in the database containing the medical data. Meaningful analysis can bring answers to many questions. What are the differences between patients from large and small cities? What factors affect maintenance of normal weight? On which ward often comes to unusual situations and anomalies? It is possible to find answers to specific questions by introducing certain initial conditions, for example, get only the implications for women living in rural areas or people who have some disease. The advantage of basket analysis is that it also indicates the association rules that are not obvious. It is possible that detected regularity will be a far deviate from the generally accepted practices.

Suppose that there is a database containing information about the membership of the group: people who are overweight, underweight, and proper body weight, consumed products and other dietary habits (number of meals, time of last meal, etc.). Building a multi-way table is almost impossible. It would be too large and unreadable, because most cells would be empty. Attempt to create tables for each product and habit, which often is done, is possible, but also has drawbacks. This form makes the information dispersed and allows to see only double associations. Implications describing the relationship between three or more factors will not be detected.

Before proceeding, a database to the analysis should be prepared so that the gathered codes are components of association rules. A minimum level of support and confidence should also be established. Additionally, it can specify what factors we would like to include in the implications, for example, which food products, patient age, gender, disease entities, etc. In order to avoid unnecessary complexity in the rules, the maximum number of codes in body and head can be specified.

Two examples presented below illustrate how to search and generate association rules. The first one will refer to the traditional area, where basket analysis is used. While the second one will illustrate the mechanism of this tool in medical data. Because of the complexity of the method the presented database is very simplified. It aims to illustrate the possibility of basket analysis for the diversity with which we deal in medical data.

Example 1

Let the database contains information about the contents of 6 shopping baskets. Let us introduce codes for each product:

B – bread
M – milk
F – fruits
C – breakfast cereals
H – ham

[Tab. 1] contains 6 transactions.

Tab. 1. Contents of 6 shopping baskets

TID					
1	B				H
2	B	M	F	C	
3	B	M	F	C	H
4	B	M		C	H
5	B	M	F		
6	B	M		C	H

First, the data are browsed in terms of level of support. Calculated are relative frequencies of each code, then each pair, triple, etc. For further analysis selected are frequent sets that means, those which support is higher than the threshold value. Let min_sup = 60%. In the case of this example results are shown in [Tab. 2]:

Tab. 2. Frequent sets, if *min_sup* = 60%

Support	Frequent sets
100%	B
83%	BM, M
67%	C, H, BC, BH, MC, BMC

While the sets O, CO, MO, PW, CMP, CMW, CMPW, which have support level of 50%, are not frequent sets.

In the second step the level of confidence for all pairs of codes selected in the first phase is taken into account. Conditional probability that the observation containing body also contains head is calculated. Let min_conf = 75%. Partial results table will have the following form [Tab. 3]:

Tab. 3. Support and confidence of selected association rules

Rule	Support	Confidence
M ⇒ B	83%	100%
B ⇒ M	83%	83%
C ⇒ BM	67%	100%
BM ⇒ C	67%	67%
BC ⇒ M	67%	80%
CM ⇒ B	67%	100%
...		

The collected results show some regularity. Particularly interesting are those with high confidence. All persons who have chosen milk also bought bread and they accounted for 83% of respondents. It is worth noting that association rules are not commutative. Consider the following two implications: CBM and BMC. The level of support proves that 67% of people had in the basket at least these three products: B, M and C. The first of these rules means that among the people buying breakfast cereals all bought milk and bread. The second says that among the customers who chose the two most popular products 67% bought also flakes.

Example 2

The following table [Tab. 4] contains information about 10 men. They were divided because of the value of BMI, half of them are obese (A), others are characterized by optimal weight (B). Each of them responded to questions about: cycling, cigarette smoking and cardiovascular diseases. In the following table [Tab. 4], "1" means that the person responded affirmatively to the question, while "0" – negative.

Note that writing the data in convention from Example 1 will receive two times more columns. Each cell contains "0" also carries valuable information. In this case the number of characteristics will be doubled. Let us introduce the following additional codes:

C – rides a bike
D – does not ride a bike
E – smokes cigarettes
F – does not smoke cigarettes
G – there are cardiovascular diseases
H – there are no cardiovascular disease

The above [Tab. 4] would have therefore form [Tab. 5]:

Tab. 4. The database characterizing 10 men

TID	Group	Bike	Cigarettes	Diseases
1	A	0	1	1
2	B	0	0	1
3	B	1	0	0
4	A	0	1	1
5	B	1	1	0
6	A	1	1	0
7	A	0	0	1
8	A	0	1	1
9	B	1	0	0
10	B	1	1	1

Tab. 5. Transcoded database

TID								
1	A			D	E		G	
2		B		D		F	G	
3		B	C			F		H
4	A			D	E		G	
5		B	C		E			H
6	A		C		E			H
7	A			D		F	G	
8	A			D	E		G	
9		B	C			F		H
10		B	C		E		G	

Let $min\_sup = 40\%$ and $min\_conf = 80\%$, then:

Tab. 6. Frequent sets, if *min_sup* = 40%

Support	Frequent sets
60%	E, G
50%	A, B, C, D, DG
40%	F, H, AD, AE, AG, BC, CH, DGA, EG, HC

There can be constructed a lot of frequent of association rules. Because of the large number of implications in the below [Tab. 7] there is just a few of them.

Tab. 7. Association rules and indicators characterizing them

Rule	Support	Confidence	Correlation
D ⇒ G	50%	100%	91%
DG ⇒ A	40%	80%	80%
H ⇒ C	40%	100%	89%
C ⇒ H	40%	80%	89%

The obtained results should be understood as follows. The first implication says that among those not riding the bike all suffer from cardiovascular disease. Additionally, we can see that these two characteristics co-occur very often (correlation (D ⇒ G) = 91%). Another association occurs in 40% of cases and indicates that 80% of those suffering from cardiovascular disease and not riding a bicycle are obese. The last two implications illustrate strong relationship between active pastime and good health.

It is worth noting that these examples are far-reaching simplifications. They illustrate how many dependencies can be found even in the case of a small number of objects and attributes. Databases in which we can apply basket analysis are much more complex.

Use of basket analysis to analyze the process of hospitalization in gynecological ward

The process of hospitalization of patients is very diverse. Even if the same diagnoses are observed, there are significant differences in the course of treatment, medication use or length of staying in the hospital. Here may be helpful the above-mentioned statistical analysis based on data mining methods. These tools allow for finding hidden dependencies in databases and present them in the form of association rules. One of these methods are, for example, artificial neural networks, which can be used to predict the outcome of treatment [8].

This paper presents the use of basket analysis on data from hospital cards from the Department of Gynecology. The created database contains information on more than eight thousand processes of hospitalization, described among others with characteristics: primary and secondary diagnosis, date of admission and discharge, patient age and place of residence. The purpose of the performed analysis was to find associations between the parameters describing the hospitalization of patients on the gynecological ward.

Tab. 8. Codes of patient age

Code for age	Age
wiek 1	up to 19 years
wiek 2	from 20 to 29 years
wiek 3	from 30 to 39 years
wiek 4	from 40 to 49 years
wiek 5	from 50 to 59 years
wiek 6	from 60 to 69 years
wiek 7	70 years and more

Tab. 9. Codes of duration of hospitalization

Code for duration of hospitalization	Duration of hospitalization
dł. hosp.1	1 day
dł. hosp.2	2 days
dł. hosp.3	3 days
dł. hosp.4	4 days
dł. hosp.5	5 days
dł. hosp.6	6 days
dł. hosp.7	7 days
dł. hosp. pow.7	from 8 to 14 days
dł. hosp. pow.14	from 15 to 21 days
dł. hosp. pow.21	more than 21 days

Because of the nature of the analysis data has been transcoded. Age is showed in [Tab. 8], while the duration of treatment in [Tab. 9]. Disease entities are coded according to the international statistical classification of diseases and health problems ICD-10 [9]. Place of residence has been referred to the appropriate NFZ code specifying the patient.

The use of basket analysis allowed to obtain the typical patterns of conduct of hospitalization [Tab. 10]. It is a large group of associations with a single-piece body and head, for which the confidence coefficient is very high and exceeds 80–85%. Associations in these cases generally determine the percentages in the various subgroups.

For example, if patients are under 20 years old, in 95% cases they come from the Podlaskie Province (NFZ 10). We interpret it in the following way: "the youngest patients" which are still schoolgirls rarely change residence

Tab. 10. Association rules describing typical patterns of hospitalization

Min. wsparcie = 1,0%, Min. zaufanie = 80,0%, Min. korelacja = 15,0%
Maks. liczność poprzednika = 1, Maks. liczność następnika = 1

	Poprzednik	==>	Następnik	Wsparcie%	Zaufanie(%)	Korelacja(%)
1	**wiek == 1,**	**==>**	**nfz == 10,**	**2,0**	**94,8**	**16,8**
2	wiek == 4,	==>	nfz == 10,	15,7	83,7	44,6
3	wiek == 5,	==>	nfz == 10,	9,9	93,8	37,5
4	wiek == 6,	==>	nfz == 10,	3,9	95,3	23,6
5	**wiek == 7,**	**==>**	**nfz == 10,**	**2,9**	**95,7**	**20,5**
6	**nfz == 1,**	**==>**	**dł. hosp. == 1,**	**1,8**	**95,0**	**17,1**
7	**nfz == 1,**	**==>**	**kod 1 == Z31**	**1,7**	**88,1**	**22,4**
8	nfz == 3,	==>	dł. hosp. == 1,	2,7	90,6	20,6
9	**kod 1 == N92**	**==>**	**nfz == 10,**	**2,3**	**98,0**	**18,7**
10	nfz == 5,	==>	dł. hosp. == 1,	1,7	90,6	16,3
11	**kod 1 == Z31**	**==>**	**dł. hosp. == 1,**	**29,4**	**100,0**	**70,9**
12	nfz == 6,	==>	dł. hosp. == 1,	1,8	89,3	16,6
13	nfz == 6,	==>	kod 1 == Z31	1,6	80,5	21,0
14	nfz == 7,	==>	dł. hosp. == 1,	7,2	81,0	31,5
15	nfz == 9,	==>	dł. hosp. == 1,	1,5	91,5	15,6

(another district NFZ). Similarly, 96% of the oldest age group are women from Podlaskie Province, which is caused by the fact that elderly women often are cured in a hospital close to home because of feeling of safety and closeness of the family.

We can make a preliminary assessment of the causes of migration of patients from other Polish regions to the analyzed hospital. For example, if the patient comes from the Lower Silesia Province (NFZ 1), in 95% of the cases she is hospitalized 1 day and 88% of patients from this province were performed IVF ICSI/ET procedure (diagnosis code Z31). We see that the analyzed ward offers a one-day medical procedure during which patients come from a distant Lower Silesia Province. Furthermore, we find out that performing IVF procedures on the ward is a cause of migration of patients [10].

Associations with a high level of confidence referring to the main cause of hospitalization ("kod 1") indicate as mainly resident Podlaskie Province. Thus, 98% of patients with a diagnosis of "extensive, frequent and irregular menstruation" (N92), hospitalized on the ward lived in Podlaskie Province. Because most of cases is "a sudden situation" in which patients do not choose away hospitals and go to the nearest one [12]. Here we also observe that if the patient has surgery performed ICSI IVF / ET in 100% of the cases the patient is hospitalized 1 day. This observation confirms the fact that the IVF procedure requires a one-day hospitalization.

Association rules network [Fig. 1] shows the most common heads (with the greatest relative support): NFZ 10 (place of residence in Podlaskie Province), 1 day hospitalization, diagnosis IVF ICSI / ET.

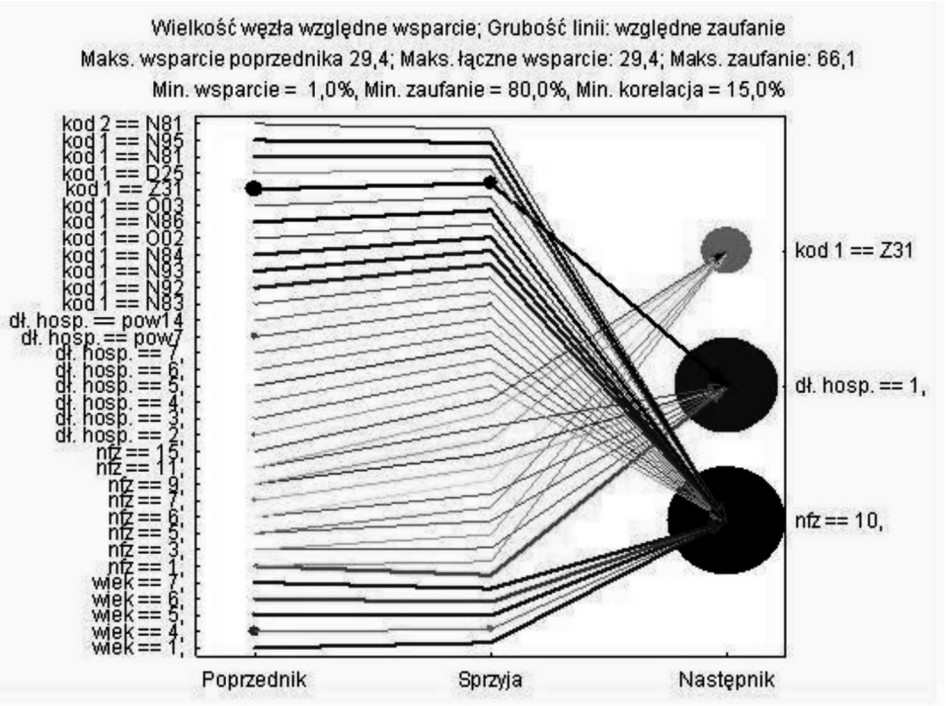

Fig. 1. Association Rules Network, typical patterns of hospitalization

A detailed analysis of the migration patients to this ward can be done apart from the database of those cases, where the resident is assigned to the NFZ 10. 118 association rules were obtained with a confidence above 80%, selected associations are showed in [Tab. 11]. These results support the assessment of migration obtained in [Tab. 10]. We see that the main purpose of hospitalization in the analyzed ward is the treatment of infertility by IVF ICSI/ET. For example, if the patient is assigned to NFZ 3, then in 82% of the cases are performed IVF ICSI/ET procedure, or in 91% of the cases patient is hospitalized for 1 day. Procedure for in vitro fertilization requires hospitalization for one day what we observed with the association rule: if kod 1 = Z31, then in 100% of the cases "dł.hosp = 1". Multicomponent associations also apply to the treatment of infertility. For example, if the patient is 30–39 years old, comes from the NFZ 6 and is hospitalized one day, in 100% of the cases the procedure IVF ICSI/ET (Z31) is performed.

Also popular sets made for patients from outside the Podlaskie Province once again show that the main cause of migration is the treatment of infertility. We see in [Tab. 12] that among patients coming from other provinces 84% were hospitalized 1 day, and 72% had done IVF treatment.

Tab. 11. Association rules for migration of patients from other provinces to the analyzed ward

Min. wsparcie = 1,0%, Min. zaufanie = 80,0%, Min. korelacja = 15,0%
Maks. liczność poprzednika = 10, Maks. liczność następnika = 10
Warunek pomijania: v3=10

	Poprzednik	==>	Następnik	**Wsparcie%**	Zaufanie(%)	Korelacja(%)
1	wiek == 2,	==>	dł. hosp. == 1,	23,7	81,8	48,0
2	wiek == 3,	==>	dł. hosp. == 1,	52,7	89,7	75,0
3	wiek == 3,	==>	kod 1 == Z31	47,4	80,7	72,7
4	wiek == 3,	==>	dł. hosp. == 1,, kod 1 == Z31	47,4	80,7	72,7
5	nfz == 1,	==>	dł. hosp. == 1,	5,3	95,0	24,5
6	nfz == 1,	==>	kod 1 == Z31	4,9	88,1	24,5
7	nfz == 1,	==>	dł. hosp. == 1,, kod 1 == Z31	4,9	88,1	24,5
8	nfz == 2,	==>	dł. hosp. == 1,	2,8	87,9	17,1
9	**nfz == 3,**	**==>**	**dł. hosp. == 1,**	**8,1**	**90,6**	**29,5**
10	**nfz == 3,**	**==>**	**kod 1 == Z31**	**7,4**	**82,4**	**28,9**
11	nfz == 3,	==>	dł. hosp. == 1,, kod 1 == Z31	7,4	82,4	28,9
12	**kod 1 == Z31**	**==>**	**dł. hosp. == 1,**	**72,4**	**100,0**	**92,8**
13	nfz == 5,	==>	kod 1 == Z31	4,6	83,0	23,0
14	nfz == 5,	==>	dł. hosp. == 1,, kod 1 == Z31	4,6	83,0	23,0
15	**wiek == 3,, nfz == 6,, dł. hosp. == 1,**	**==>**	**kod 1 == Z31**	**3,2**	**90,0**	**19,8**

Tab. 12. Popular sets relating to migration of patients

Obliczono częstości zestawów elementów
Min. wsparcie = 1,0%, Min. zaufanie = 80,0%, Min. korelacja = 15,0%
Maks. liczność poprzednika = 10, Maks. liczność następnika = 10
Warunek pomijania: v3=10

	Popularne zestawy	Liczność	Wsparcie%
20	**dł. hosp. == 1,**	2399,000	84,05746
27	**kod 1 == Z31**	2067,000	72,42467
106	**dł. hosp. == 1,, kod 1 == Z31**	2067,000	72,42467
2	**wiek == 3,**	1677,000	58,75964

Association rules network [Fig. 2] shows the main reason for the migration of women as a treatment for infertility. Heads are one-day hospitalization and diagnosis Z31.

We see that the cause of migration is dominated by the treatment of infertility [11]. It is worth asking whether there are other reasons for the choice of this hospital by patients from other provinces. If in the analysis of migrating patients infertility treatment is skipped, we get 17 interesting associations [Tab. 13].

For example, if the patients are 50–59 years and signed up to the treatment of typically gynecological problems – leiomyoma of uterus (D25) in 82% of the cases they come from the Warmia-Masuria Province (NFZ 14). Also in the group of long hospitalizations (15–21 days) 67% of the patients are from the Warmia-Masuria Province. In [Tab. 13] we can observe the coexistence of disease entities. If the second diagnosis is N83 (noninflammatory disorders of ovary), then in 82% of the cases D25 (leiomyoma of uterus) is the main cause of treatment.

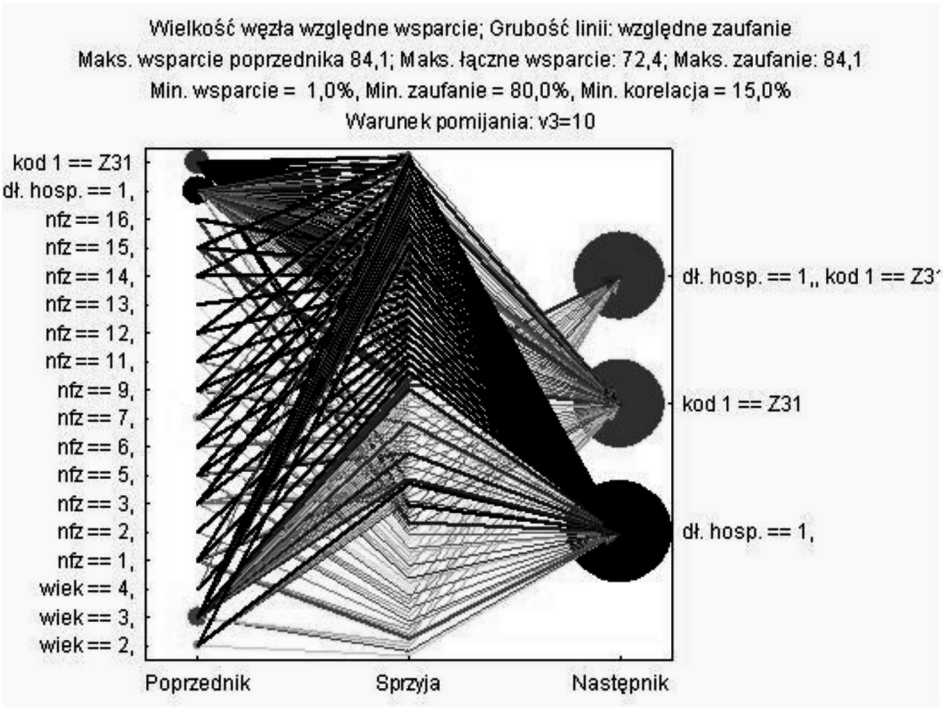

Fig. 2. Association rules network relating to migration of patients

Tab. 13. Association rules for migration of patients from other provinces to the analyzed ward (only gynecological reasons)

Min. wsparcie = 3,0%, Min. zaufanie = 60,0%, Min. korelacja = 15,0%
Maks. liczność poprzednika = 10, Maks. liczność następnika = 10
Uwzględniaj: 889:1004;4994:5162

	Poprzednik	==>	Następnik	Wsparcie%	Zaufanie(%)	Korelacja(%)
1	wiek == 5,	==>	nfz == 14,	14,4	75,9	47,0
2	**dł. hosp. == pow14**	==>	**nfz == 14,**	**3,5**	**66,7**	**21,7**
3	kod 1 == N83	==>	wiek == 2,	7,7	62,9	50,6
4	kod 1 == N93	==>	dł. hosp. == 1,	4,9	70,0	36,6
5	kod 1 == N81	==>	nfz == 14,	3,2	64,3	20,3
6	**kod 2 == N83**	==>	**kod 1 == D25**	**4,9**	**82,4**	**43,1**
7	kod 2 == N81	==>	nfz == 14,	3,5	66,7	21,7
8	wiek == 4,, dł. hosp. == pow7	==>	nfz == 14,	5,3	60,0	25,3
9	wiek == 4,, dł. hosp. == pow7	==>	kod 1 == D25	5,6	64,0	40,6
10	wiek == 5,, dł. hosp. == 1,	==>	nfz == 14,	4,2	70,6	24,5
11	wiek == 5,, dł. hosp. == pow7	==>	nfz == 14,	4,2	70,6	24,5
12	**wiek == 5,, kod 1 == D25**	==>	**nfz == 14,**	**3,2**	**81,8**	**22,9**

In the analysis of gynecological migration we note that there are mainly associations of 14th NFZ. It is the area of the Warmia-Masuria Province, directly adjacent to the Podlaskie Province. The migration of patients is justified by the choice of "better", more prestigious clinical hospital. In

addition, for many years it was the only one in the north-east macroregion III° referral level hospital. More complicated cases that require specialized treatment were directed.

Every hospitalization is associated with specific costs. Shorter treatment is usually more profitable for the hospital. Long hospitalizations are associated with complications and often generate high costs. Therefore it is important to analyze these processes of hospitalization, which last longer than the average duration of treatment in a particular disease entity. In this case all the processes of hospitalization lasting more than 7 days were selected for analysis and a group of more than 1300 items was obtained.

In [Tab. 14] we see that the largest group among the above week hospitalizations are patients with primary diagnosis of D25 (leiomyoma of uterus) – support of 19%. In this group 81% of cases were treated from 8 to 14 days and 19% longer. In the group of long hospitalizations among patients with a diagnosis of N83 (noninflammatory disorders of ovary), 87% stayed on the ward 8–14 days. We obtained association rule: if the first diagnosis is D25 and N83 is the second diagnosis, in 88% of cases duration of treatment is 8–14 days. We have received an interesting information here about co-existing diagnosis: leiomyoma of uterus and noninflammatory disorders of ovaries, which become important, if the patient is treated for a long time.

Tab. 14. Association rules for a long hospitalizations

	Podsumowanie reguł asocjacji Min. wsparcie = 1,0%, Min. zaufanie = 80,0%, Min. korelacja = 15,0% Maks. liczność poprzednika = 10, Maks. liczność następnika = 10 Uwzględniaj: 7112:8428					
	Poprzednik	==>	Następnik	Wsparcie%	Zaufanie(%)	Korelacja(%)
1	wiek == 1,,	==>	dł. hosp. == pow7	1,9	96,2	15,4
2	wiek == 2,,	==>	dł. hosp. == pow7	8,3	82,6	29,8
3	wiek == 5,,	==>	dł. hosp. == pow7	15,3	80,5	40,0
4	**kod 1 == N83**	**==>**	**dł. hosp. == pow7**	**7,4**	**87,5**	**29,1**
5	**kod 1 == D25**	**==>**	**dł. hosp. == pow7**	**18,9**	**81,4**	**44,7**
6	kod 2 == N83	==>	dł. hosp. == pow7	4,7	88,6	23,2
7	wiek == 4,, kod 2 == N83	==>	kod 1 == D25	2,2	93,5	29,8
8	wiek == 4,, kod 2 == N83	==>	dł. hosp. == pow7, kod 1 == D25	2,0	83,9	29,6
9	wiek == 3,, kod 1 == N83	==>	dł. hosp. == pow7	2,2	96,7	16,6
10	wiek == 3,, kod 1 == D25	==>	dł. hosp. == pow7	2,6	85,0	16,9
11	wiek == 4,, kod 1 == D25	==>	dł. hosp. == pow7	11,3	81,9	34,7
12	wiek == 4,, kod 2 == N83	==>	dł. hosp. == pow7	2,1	90,3	15,8
13	wiek == 5,, kod 2 == N83	==>	kod 1 == D25	1,4	82,6	22,6
14	**kod 1 == D25, kod 2 == N83**	**==>**	**dł. hosp. == pow7**	**3,5**	**88,5**	**20,0**
15	wiek == 5,, dł. hosp. == pow7, kod 2 == N83	==>	kod 1 == D25	1,2	80,0	20,5

In [Tab. 15] characteristics for place of residence was also included. We see that 87% of 8–14 days hospitalizations and 89% of hospitalizations lasting 15–21 days concerned Podlaskie Province residents. An analysis of age shows that the older the patient the greater is the percentage of long hospitalizations. Women to up 19 years represent 2% of long hospitalizations,

Tab. 15. The association rules for a long hospitalizations (with the division into NFZ districts)

Podsumowanie reguł asocjacji
Min. wsparcie = 1,0%, Min. zaufanie = 80,0%, Min. korelacja = 15,0%
Maks. liczność poprzednika = 10, Maks. liczność następnika = 10
Uwzględniaj: 7112:8428

	Poprzednik		Następnik	Wsparcie%	Zaufanie(%)	Korelacja(%)
1	wiek == 1,	==>	dl. hosp. == pow7	1,9	96,2	15,4
2	wiek == 2,	==>	dl. hosp. == pow7	8,3	82,6	29,8
3	wiek == 5,	==>	dl. hosp. == pow7	15,3	80,5	40,0
4	dl. hosp. == pow14	==>	nfz == 10,	13,9	87,1	37,0
5	dl. hosp. == pow21	==>	nfz == 10,	6,2	89,0	24,9

women aged 20–29 years – 8%, and patients aged 50–59 years – 15% of all long stays on the ward. Confidence is decreasing, it means that percentage of hospitalizations exceeding 14 days is increasing.

Conclusions

Basket Analysis is a great tool for the medical data mining. This method allows for deeper and more detailed exploration of the collected information than traditional statistics. It may be used for each type of data. Association rules make it possible for a thorough insight into among others case records, diet, habits and customs of patients and medical procedures. It can be applied to monitor the process of hospitalization. It also allows detection of factors that influence the healing process, including those that have not been suspected. Sometimes the basket analysis answers not formulated questions and indicates hidden patterns.

REFERENCES

[1] Agrawal R., Imielinski T., Swami A., Mining association rules between sets of items in large databases, Conference on Management of Data, pp. 207–216, Washington, 1993.

[2] Agrawal R., Ramakrishnan S., Fast Algorithms for Mining Association Rules, IBM Research Raport RJ9839, IBM Almaden Research Center, 1994.

[3] Śniegocka-Jusiewicz M., Analiza koszykowa w badaniach marketingowych, PTE Toruń Working Papers, 17, 2008.

[4] Tadeusiewicz R., Data Mining jako szansa na relatywnie tanie dokonywania odkryć naukowych poprzez przekopywanie pozornie całkowicie wyeksploatowanych danych empirycznych, In: Wątroba J. (red.), Statystyka i Data Mining w badaniach naukowych, StatSoft, pp. 5–30, Kraków 2006.

[5] Hristovski D., Stare J., Peterlin B., Dzeroski S., Supporting discovery in medicine by association rule mining in Medline and UMLS, Medinfo, 10, pp. 1344–1348, 2001.

[6] Miziołek A., Statystyka w medycynie, StatSoft Polska, 2011.
[7] Ordonez C., Santana C. A., de Braal L., Discovering interesting association rules in medical data, In: D. Gunopulos and R. Rastogi, ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery 2000, Dallas, Texas, pp. 78–85, 2000.
[8] Milewski R., Jamiołkowski J., Milewska A. J., et al., Prognozowanie skuteczności procedury IVF ICSI/ET – wśród pacjentek Kliniki Rozrodczości i Endokrynologii Ginekologicznej – z wykorzystaniem sieci neuronowych, Ginekologia Polska, 80 (12), pp. 900–906, 2009.
[9] Bartkowski S. [kom. nauk.], Międzynarodowa statystyczna klasyfikacja chorób i problemów zdrowotnych, Rewizja dziesiąta, T. 1, Vesalius, Kraków 2006.
[10] Radwan J., Wołczyński S., Niepłodność i rozród wspomagany, Termedia, Poznań, 2011.
[11] Milewska A. J., Milewski R., Wołczyński S., Analiza zjawiska migracji pacjentów na Podlasie na przykładzie Kliniki Ginekologii, Polityka Zdrowotna, 7, pp. 71–76, 2008.
[12] Munro M. G., Najnowsze wiadomości dotyczące ostrego krwawienia z dróg rodnych u kobiet niebędących w ciąży, Ginekologia po dyplomie, T. 10, 3 (55), pp. 20–26, 2008.