ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

*Ewa Nowakowska-Zajdel*[*], *Małgorzata Muc-Wierzgoń*[**]
*Grażyna Trzpiot*[***], *Alicja Ganczarek*[****]

# CLASSIFICATION OF PATIENTS WITH RESPECT TO SOME GROUP OF FACTORS

**Abstract.** In this paper a classification of examined patients was carried out based on results of multivariate analysis using classification trees. The aim of the analysis was to identify characteristic factors describing groups of patients suffering from colorectal cancer with different stage of disease. Clinical data from medical documentation of the patients with colon cancer were analyzed. Qualitative variables such as sex, clinical stage, histopathology type of cancer and malignancy, weight class, glucose level class and coexistence with other illnesses were used in the analysis.

**Key words**: Classification trees, severity, type and histopathology malignancy, body mass index, glucose level.

## I. INTRODUCTION

Colorectal cancer ranks the second place in regard to occurrence among men and women and is the third cause of death amongst oncological patients in Poland. Each year new cases of colon cancer account for 11 000 while total number of patients newly diagnosed with neoplasm is 110 000 (Nowacki (2006)). Reasons for the disease are still unknown. The most important are genetic predispositions as up 10–15% cases are familial. Environmental factors as poor diet, insufficient physical activity and metabolic disorders (obesity, diabetes) play important role as risk factors for colon cancer (Chang C.K., Ulrich C.M. (2003), Wadden T.A., Brownell K.D., Foster G.D. (2002)). Nowadays many epidemiological research have stated relationships between age, BMI (Body Mass Index),

[*] Ph.D., Department of Internal Diseases, Medical University of Silesia, Katowice.
[**] Professor, Department of Internal Diseases, Medical University of Silesia, Katowice.
[***] Professor, Department of Statistics, The Karol Adamiecki University of Economics, Katowice.
[****] Ph.D., Department of Statistics, The Karol Adamiecki University of Economics, Katowice.

fasting level of glucose and other growth factors as TNF alpha, leptin, insulin like growth factors, adiponectin and others. It is interested to know what medical data concerning colorectal cancer could play important role in division on the separate groups.

The aim of this paper was to identify characteristic factors describing groups of patients suffering from colorectal cancer with different stage of disease.

## II. METHODOLOGY

A lot of the empirical results show that the economic and sociological variables (Gatnar, Walesiak (2004)) do not have normal distribution and are very often described by nominal values, while distributions of variables often have incommunicable values and outlier observations. Thus, it is often the case that we cannot apply the classical methods to the classification of the empirical variables. In this paper, we applied the nonparametric method – the classification trees – to the classification. This method is based on the recursive partitioning of the m-dimension space $X^m$ into homogenous subsets concerning dependent variable y. When dependent variable y is nominal, equation (1) is a classification tree (Breiman, Friedman, Olsen and Stone (1984), Gatnar (1998), Gatnar, Walesiak (2004), Misztal (2007):

$$y = \sum_{k=1}^{K} a_k I\{\mathbf{x}_i \in R_k\} \tag{1}$$

where:

$\mathbf{x}_i$ - multivariate variable, element of $X^m$,

$R_k - k = 1, \ldots, K$, are disjoint regions in the m-dimensional feature space, segment of $X^m$,

$a_k$ – the parameters,

$I(q)$ – an indicator function:: $I(q) = \begin{cases} 1 & \text{if } q \text{ is true} \\ 0 & \text{if } q \text{ is false} \end{cases}$.

## III. EMPIRICAL ANALYSIS

In this part of the paper we analyzed the data from medical documentation of out-patient clinic of Cancer Ward of 4th Hospital in Bytom and Department of Internal Diseases in Bytom, of Medical University of Silesia from 2000 to 2007. We had information on 316 patients about:

– sex (Male, Female) ,
– Body Mass Index (BMI) (Normal weight, Abnormal weight),
– glucose level (Normal glucose, Abnormal glucose),
– coexistence illnesses (cardiovascular diseases, diabetes, other malignant neoplasms),
– tumor location (colon, rectum),
– histopathologic type (mucous adenocarcinoma, adenocarcinoma, adenocarcinoma with necrosis),
– clinical stage of diseases (I, II, III, IV),
– histopathology malignancy (G1, G2, G3).

We analyzed model (1) where the tumor location is the dependent variable y. In model (1) we took into consideration only the qualitative variables which we recognized as factors describing groups of patients suffering from colorectal cancer with different stage of diseases. We used the C&RT (Classification and Regression Trees) recursive partitioning method proposed by Breiman et al. (1984) and available in the STATISTICA PL package. To stop the recursive partitioning, we used three pruning methods: cost-complexity pruning, one Standard Error (1SE) rule and FACT-Fast Algorithm for Classification Trees. The choice results of our classifications are presented in table 1 and in figures 1–3.

Table 1. Results of classifications of tumor location

| Fig. | Recursive partitioning method | Pruning | Classification's error | Cross-validation | Standard deviations |
|---|---|---|---|---|---|
| Fig.1. | C&RT | 1SE rule | 0.45 | 0.44 | 0.03 |
| Fig.2. | C&RT | FACT | 0.43 | 0.43 | 0.03 |
| Fig.3. | C&RT | 1SE rule | 0.53 | 0.45 | 0.02 |

Unfortunately, we obtained significant errors of classification (table 1). They probably resulted from a very high volatility of the factors. However, when we analyzed the trees (figure 1–3), we were able to find a few important dependences.

For example, based on the first tree (figure 1), we can say, that the most important factor in the tumor location is weight, the second important factor is sex and the third most important factor is histopathology type. The patient whose tumor is located in colon in many cases has mucous adenocarcinoma. If that was not mucous adenocarcinoma, the majority of cases that were diagnosed concerned women suffering from abnormal weight (figure 1) or patients (women and men) with abnormal glucose (figure 2).

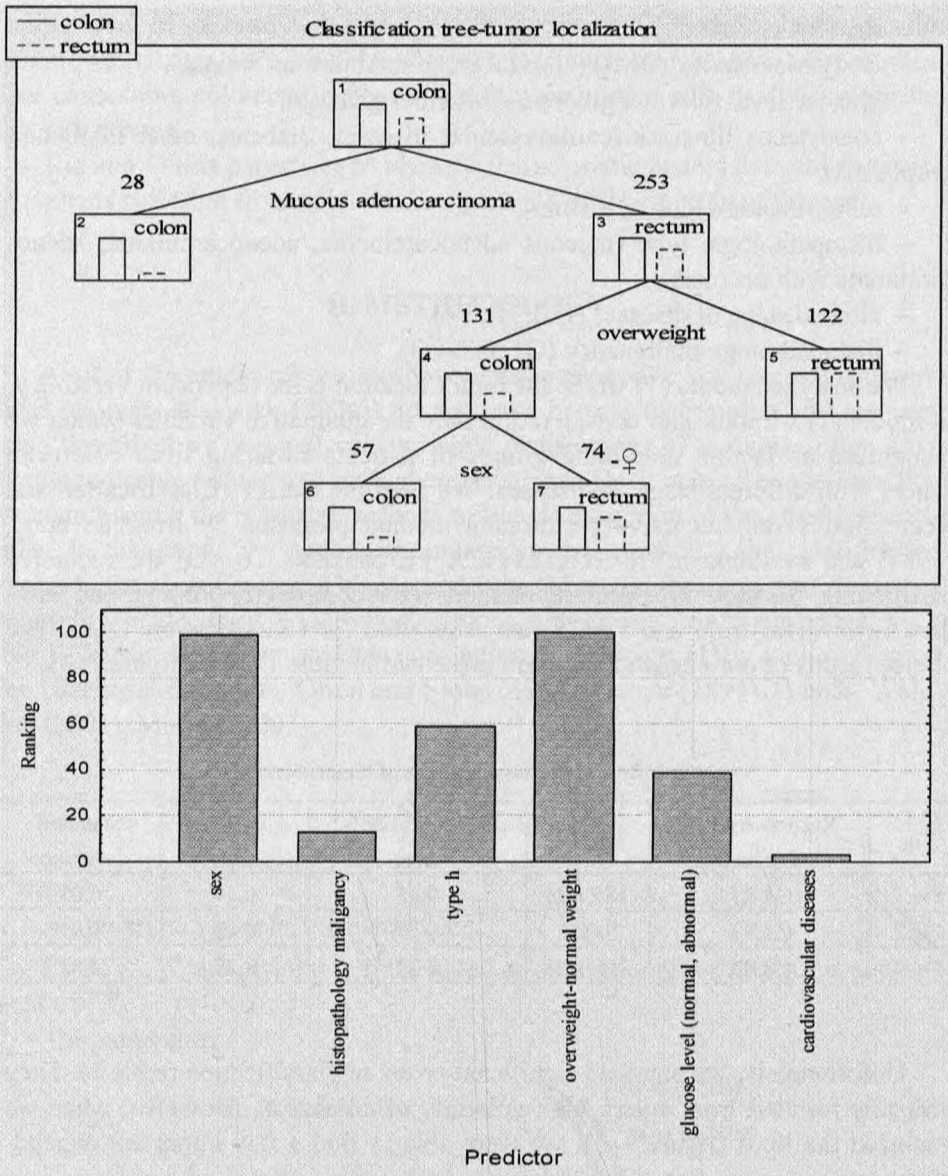


Figure 1. Classification tree and the ranking of predictors

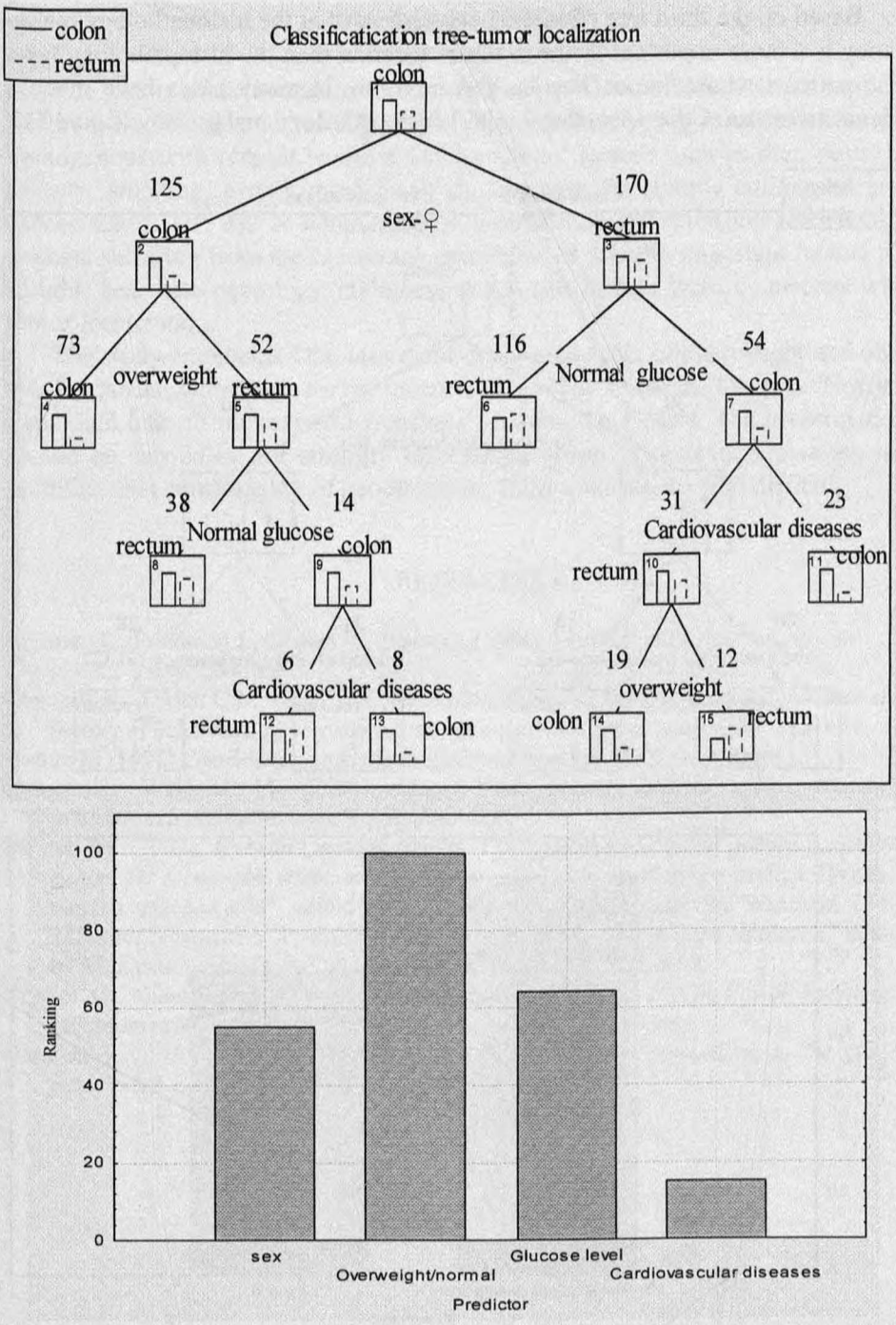


Figure 2. Classification tree and the ranking of predictors

Based on the third tree (figure 3), we can say that the histopathology malignancy is a more important factor in tumor location than the histopathology type. The patients whose tumor was located in colon, in many cases have mucous adenocarcinoma. Otherwise, they had G3 histopathology malignancy (figure 3).

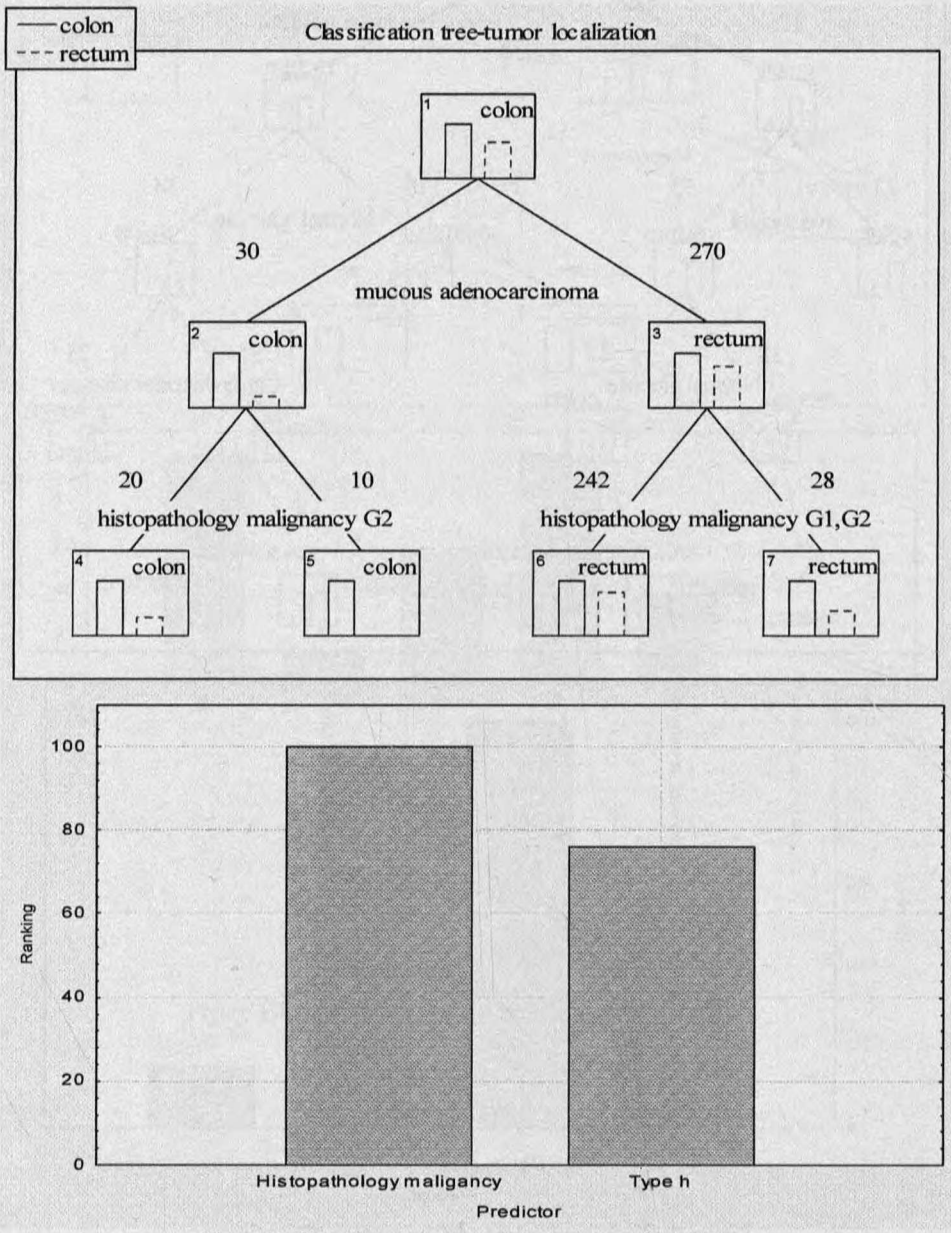


Figure 3. Classification tree and the ranking of predictors

## IV. CONCLUSION

The results of the research are not satisfactory. One point is, that the medical factors were volatile and heterogeneous. The examined group of patients was not homogenous with respect to other environmental factors such as diet, physical activity, smoking, genetic predisposition, non specific chronic colitis and pre-cancer states. The use of nonparametric method (classification tree) to classify patients suffering from the colorectal cancer lets us identify important factors as: weight, sex, histopathology malignancy. All the factors were connected with tumor localization.

The study emphasizes the important explanatory role of overweight and obesity for cancer, which was earlier mentioned in many research. The classification tree could help to illustrate relationships between the factors. The investigation should be carried on for strongly represented group of patients. However, the multifactorial conditioning of neoplasmatic disease makes the trial difficult.

## REFERENCES

Breiman L., Friedman J., Olshen R., Stone C. (1984), *Classification and Regression Trees*, CRC Press, London.

Chang C.K., Ulrich C.M. (2003), *Hyperinsulinaemia and hyperglycaemia: possible risk factors of colorectal cancer among diabetic patients.* Diabetologia, 46, 595–607.

Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, PWN, Warszawa.

Gatnar E., Walesiak M. (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław.

Misztal M. (2007), *Wybrane metody analizy i prognozowania czasu pobytu na OIOM pacjentów z choroba wieńcowa,* „Taksonomia 14. Klasyfikacja i analiza danych – teoria i zastosowania", edited by K. Jajuga and M. Walesiak, AE Wrocław, 1169, 288-296. Nowacki M.P. (2006), *Rak jelita grubego*, „Onkologia kliniczna" edited by M. Krzakowski, Borgis Wydawnictwo Medyczne, Warszawa.

Trzpiot G., Ganczarek A. *The classification of risk on the Polish Power Exchange,* „Ekonometria", edited by J. Dziechciarz, AE Wrocław, in press.

Wadden T.A., Brownell K.D., Foster G.D. (2002), *Obesity: responding to the global epidemic.* J Consult Clin Psychol, 70, 510–525.

*Ewa Nowakowska-Zajdel, Małgorzata Muc-Wierzgoń*
*Grażyna Trzpiot, Alicja Ganczarek*

## KLASYFIKACJA PACJENTÓW ZE WZGLĘDU NA WYBRANĄ GRUPĘ CZYNNIKÓW BADANYCH

Bazując na wynikach analiz metod statystyki wielowymiarowej przeprowadzono klasyfikację grupy badanych pacjentów ze względu na grupę badanych cech.

Celem analizy jest próba wyodrębnienia charakterystycznych grup czynników wśród pacjentów chorujących na raka jelita grubego w różnym stopniu zaawansowania klinicznego.

Analizie poddano wybrane dane epidemiologiczne pochodzące z dokumentacji medycznej chorych z ustalonym rozpoznaniem – rak jelita grubego. Do analizy wykorzystano zmienne jakościowe: płeć, stopień zaawansowania klinicznego choroby, typ i złośliwość histopatologiczną, podział na osoby z wagą prawidłową, nadwagą i otyłością, podział ze względu na stężenie glukozy na czczo w surowicy krwi oraz współistnienie występowania innych chorób.