Original article
UDC: 616-006:115.07:616-036.22

Complex analysis of lung cancer and stomach cancer risk factors

Vladimir LEZHNIN
Eugene POLZIK
Vladimir KAZANTSEV

SCIENTIFIC AND PRACTICAL CENTER FOR MEDICAL, SOCIAL AND ECONOMIC PROBLEMS IN PUBLIC HEALTH, YEKATERINBURG, RUSSIA

BACKGROUND: *We conducted a multifactorial cancer epidemiology case-control study to assess effects of a complex of factors on high cancer morbidity of population of an industrial town in the central part of Russia. The study was carried out in the town of Tula, one of the oldest industrial centers of Russia with a high occupational and industrial load on the population and a high cancer incidence rate exceeding the Russian average rates by 38.4% in men and 89.9% in women.*

METHODS: *We examined 200 lung cancer cases, 206 stomach cancer cases, and 213 controls, all residents of Tula. On each of the 619 study participants we collected personal data, which allowed characterization of cases and controls by a complex of 29 features reflecting well-known lung and stomach cancer risk factors. Mathematical processing was performed using the technique of multi-factorial analysis based on mathematical pattern recognition methods. All tasks were solved in the package of applied programs of pattern recognition KVAZAR.*

RESULTS: *The results of the multifactorial analysis of epidemiological data collected showed that the contribution of smoking in the development of stomach cancer in Tula was 7%, alcohol abuse 4%, occupational exposure to carcinogens12%, a complex of individual biological factors 28%, environmental factors 22%, radiation factors 5%, nutrition habits 10%, and other factors 12%. The contribution of smoking in lung cancer mortality was 27%, alcohol abuse 4%, occupational exposure to carcinogens 14%, a complex of individual biological factors 32%, environmental factors 9%, radiation factors 7%, and other factors 7%.*

CONCLUSION: *The results of the epidemiological study carried out on the basis of methods of a multifactorial analysis in the city of Tula, Russia, showed the effectiveness of such an approach to solving complicated epidemiological tasks. The suggested method allows one to assess a complex effect of different factors on oncologic morbidity in an industrial town, and the prediction of effectiveness of various cancer preventive measures facilitates the choice of the most effective complex of preventive actions on the certain territory and at low cost.*

KEY WORDS: *Lung Neoplasms; Stomach Neoplasms; Risk Factors; Causality; Russia*

INTRODUCTION

In most countries cancer incidence and mortality rates are constantly increasing, and according to available predictions this tendency will persist in the nearest future. At the Congress of the International Union Against Cancer in Rio de Janeiro in 1998, a progressive increase in the number of incident cancer cases in the world was noted: 6 million cancer cases in 1975, 7.6 million in 1985, 8.1 million in 1990, and the expected rate in 2010 is 13 million incident cancer cases. The fact that this increase is taking place on the background of doubtless achievements of modern oncology in diagnostics and treatment of tumors provides reason enough to primary and secondary cancer prevention to be the main means of changing this tendency.

Even though the etiology and pathogenesis of malignant neoplasms remain unclear, the available scientific information about multiple carcinogenic risk factors is quite sufficient for successful

Address correspondence to:
Dr. Vladimir L. Lezhnin, MD, Ph.D., Scientific and Practical Center for Medical, Social and Economic Problems in Public Health, 8b Karl Liebknecht St., 620075 Yekaterinburg, Russia, E-mail: veronica@etel.ru

The manuscript was received: 22. 05. 2003

Provisionally accepted: 23. 07. 2003

Accepted for publication: 19. 11. 2003

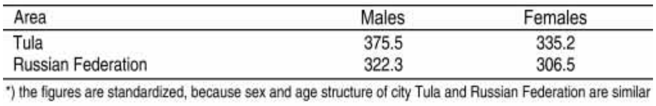
implementation of preventive measures. However, the situation developing at a certain territory is always very peculiar and thus general information concerning the influence of various risk factors in cancer incidence rate is insufficient for developing effective cancer preventive measures. In order to develop a local cancer prevention program, decision-makers need to have the answers to the following questions:

- What factors of carcinogenic risk really affect the population of the region for which the cancer prevention program is being developed?
- Which of the known risk factors have the strongest effect on cancer morbidity of population?
- What is the potential effectiveness of the planned cancer prevention measures?

MATERIALS AND METHODS

The city of Tula (with 520 000 inhabitants) is one of the Russian towns with high cancer incidence rates exceeding the average rates in the Russian Federation by 38.4 % in males and 89.9 % in females (Table 1).

Table 1. Cancer incidence rate for Tula population (per 100,000; as compared to the average rates in the Russian Federation)



Area	Males	Females
Tula	375.5	335.2
Russian Federation	322.3	306.5

*) the figures are standardized, because sex and age structure of city Tula and Russian Federation are similar

Differences between cancer incidence rates in Tula and the average rates for the country as a whole makes it possible to assume that there are strong carcinogenic risk factors in the observed region. It should be noted that Tula is one of the largest industrial centers of Russia. It hosts a large number of ferrous metallurgical, mechanical engineering, and chemical factories accounting for a high intensity of industrial estate and environment contamination.

In the structure of cancer pathology in Tula lung and stomach cancer rank the highest - 52.3 % and 42.9 %, respectively (in the Russian Federation - 42.23 % and 32.7% per 10 000, respectively). Thus, the lung cancer incidence rates in the town exceed the Russian average by 28% in males and 1.5 % in females; stomach cancer incidence rates by 26 % and 40 %, respectively.

The results of this analysis served as a basis for special cancer epidemiology studies carried out to establish causes of cancer incidence in Tula and develop cancer prevention measures.

The following principles served as a basis for them:

1. According to available data, the most prevalent forms of cancer, such as that of lung and stomach, are developing due to the effects of a combination of. environmental, occupational, social, domestic, biomedical, and other factors. The quantitative assessment of the contribution of each factor in the development of lung and stomach cancer is therefore a rather complicated epidemiological task. We think that it is the multifactorial nature of lung and stomach cancer causes that calls for the application of multifactorial techniques and makes traditional monofactorial epidemiological methods nonapplicable for the purpose of solving such tasks.

2. The study should be conducted on the individual level since this approach to the forming of epidemiological material provides an opportunity to obtain the most correct data adequate to the task in question.

The mathematical methods of pattern recognition (PR) may be useful for solving the tasks of multifactorial analysis. Since they are still not often used in epidemiological studies, let us describe them in more details.

By a pattern or class in PR is meant a set of all objects (phenomena, processes, occurrences), which are similar to each other in some fixed respect, for instance, the multitude of lung cancer cases. To recognize a pattern means to indicate the number or name of the pattern to which a given object belongs. Recognizing the object is carried out with the help of a decision rule (a discriminant function), which can be worked out on the object classification computer learning stage preceding that of recognition. The training sample is a set of objects, which provides patterns for training, i.e. a juncture of some subsets of patterns under consideration. The examination (checking) sample is a set of objects used to check the quality of training.

The feature is a description of a property of the object. Features may be both quantitative and qualitative. Solution of pattern recognition problems involves *n*-dimensional vectors, which simulate real objects whereby each component of the simulating vector represents the value of the corresponding feature. Geometrically, the pattern is identified as a multidimensional domain where each point (vector) corresponds to a specific realization of this pattern.

The task for the discriminant analysis may be set as follows. Let $\boldsymbol{k}$ classes of objects $\boldsymbol{X_1}$, ..., $\boldsymbol{X_k}$ be known which in the given set $\boldsymbol{X \subset R^n}$ are represented by their finite non-intersecting subsets $\boldsymbol{X_1}$, ..., $\boldsymbol{X_k}$ of n-dimensional vectors: $\boldsymbol{X_i \cap X_j = \varnothing\ \forall i \neq j}$. The task is to formulate a rule that would help us to classify reliably the vectors as those contained in the training set and newly provided ones. The quality criterion of the decision rule is the percentage of correctly recognized vectors of the examination sample including the objects whose classification is known but they were not included in the training. The examination sample includes 10% to 20% of all of available observations. The worked-out decision rule that provides a high percentage of the correct pattern recognition at the "examination" (80% to 100%) indicates a good differentiation of classes within the pattern chosen.

The construction of the decision rule can be preceded by the analysis of information and the search for the most informative subsystems of features. The choice of the informative subsystem of features consists of selecting a subset of features from a given descriptive system that would ensure a simple, reliable and economic division of the given set of vectors in the corresponding space.

Similar to statistical methods, the quality of the problem solution strongly depends on the sample size. The sample is considered representative if the number of vectors in the training sample of each class is 5 to 10 times higher than the number of features under study, but the final conclusion about the quality of training can be drawn only after the examination procedure. Thus, the percentage of the correctly classified vectors (or, on the contrary, the percentage of misclassification) in the examination sample is the final information criterion of the chosen system of features and representation of the training samples.

Although the interval range of quality indicators in class separation is rarely used in the discriminant analysis based on determinant approaches, i.e. applying no assumptions about the statistical properties of patterns, there still exists the possibility to calculate them. As the distribution of probabilities of misclassification abides by the binomial law in case of random choice of vectors for the examination, we can use a classical scheme of Bernoulli's independent statistical tests to find the confidence interval of the true level of the classifier's mistake. The algorithm of building the interval is described in works on mathematical statistics. For example, Fleiss (2) used this algorithm to assess the conclusions, which were based on one proportion. Duda and Hart (3) provided charts to estimate confidence intervals of the probability of misclassification built according to Bernoulli's scheme. It is possible to estimate the size of the examination sample using this algorithm.

The epidemiological studies held in Tula consisted of several parts:

1. Forming case and control groups. Out of all the cancer cases during the years 2000 and 2001 among Tula citizens, 200 lung cancer cases and 206 stomach cancer cases were chosen. They all had morphological (either histological or cytological) verification of the diagnosis. Cases without such verification were not included into the group.

The control cohort of 213 Tula residents was sampled randomly, yet with account for the following criteria:

a) Its sex and age structure corresponded to that of the adult population of Tula;

b) It included residents of all districts of the town in the proportion close to the territorial distribution;

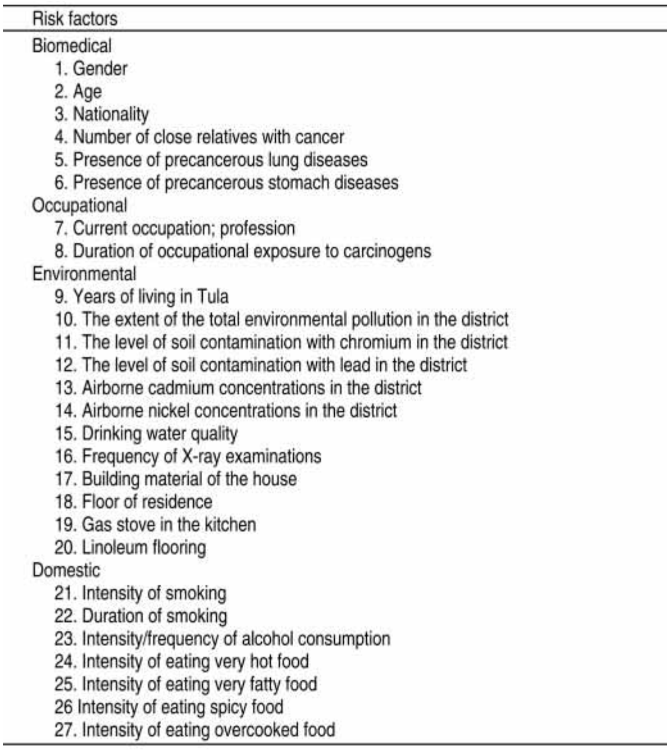
c) The occupational structure of the control cohort was proportional to that of the city as a whole.

Thus, when selecting the control cohort an attempt was made to bring it in accordance with the population structure of Tula.

2. Choice of factors (features) to be studied and forming the databases for the discriminant analysis.

Each of 619 people included in the study was characterized by a complex of 27 features reflecting well-known risk factors of lung and stomach cancer (Table 2).

Table 2. Risk factors under study



Risk factors
Biomedical
1. Gender
2. Age
3. Nationality
4. Number of close relatives with cancer
5. Presence of precancerous lung diseases
6. Presence of precancerous stomach diseases
Occupational
7. Current occupation; profession
8. Duration of occupational exposure to carcinogens
Environmental
9. Years of living in Tula
10. The extent of the total environmental pollution in the district
11. The level of soil contamination with chromium in the district
12. The level of soil contamination with lead in the district
13. Airborne cadmium concentrations in the district
14. Airborne nickel concentrations in the district
15. Drinking water quality
16. Frequency of X-ray examinations
17. Building material of the house
18. Floor of residence
19. Gas stove in the kitchen
20. Linoleum flooring
Domestic
21. Intensity of smoking
22. Duration of smoking
23. Intensity/frequency of alcohol consumption
24. Intensity of eating very hot food
25. Intensity of eating very fatty food
26 Intensity of eating spicy food
27. Intensity of eating overcooked food

Judging by this it may be noted that the number of cases and controls is representative because the number of cases in each group 5 to 10 times exceeds the number of studied features.

At forming the complex of features the task of complete covering of all known and suspected lung and stomach cancer risk factors was set.

All the 27 features might be related to the following groups:

- Biomedical
- Occupational
- Environmental, and
- Domestic

Information from the sphere of biomedical features was obtained from the databases of cancer clinic of Tula. Data concerning the occupational peculiarities of each case (control) - person was obtained by the help of a questionnaire, and the source of information about the environmental contamination by carcinogenic substances was the monitoring data provided by the center of sanitary and epidemiological surveillance of the city of Tula.

A number of features from environmental (ecological) block, characterizing the degree of environmental contamination was also obtained from the monitoring results of environment protection agencies of the city. From the spectrum of the environmental pollutants, 5 factors reflecting the level of substance concentration (NN 10-14 in Table 2) were chosen, because three of them (NN 11, 13, 14) are the IARC recognized carcinogens and N 12 (lead) a "possible carcinogen", and feature n 10 reflects the degree of joint environmental contamination by the toxicants singled out by city ecological service.

The data concerning drinking water quality were also obtained from the center of sanitary - epidemiological surveillance of city of Tula.

The information about the frequency of X-ray examinations, availability of gas stove in the kitchen and linoleum flooring at home, building material of the house and the floor of residence were obtained with the help of a questionnaire. Information concerning the domestic risk factors was also obtained by a questionnaire.

The study used 3 types of factors: quantitative (e.g. age, duration of occupational exposure to carcinogens); qualitative serial (e.g. factor "intensity of smoking" was encoded as follows: 0-"doesn't smoke"; 1-"smokes less that 10 cigarettes per day"; 2- "smokes 10-20 cigarettes per day"; 3- "smokes more that 20 cigarettes per day") and qualitative nominal (e.g. males/females; availability of gas stove/absence of gas stove). Statistical methods give no chance to analyze mathematically the complexes containing various features, but the applied determinist methods of pattern recognition give such an opportunity.

Getting back to the tasks formulated above, one can suggest the following scheme to solve them applying the pattern recognition methodology:

1. Evaluating the sufficiency of the selected complex of factors and its sub-complexes for a reliable description of differences between the objects of the allocated classes;

2. Determining the character (direction) of effect of each factor what can be treated as an increase or a decrease in the probability of a disease under the effect of the factor within the framework of this study;

3. Using mathematical models (decision rules) to forecast the effectiveness of various risk factor controlling effects.

All tasks were solved in the KVAZAR package of applied pattern recognition programs (4). Mathematical processing was done for lung and stomach cancer cases separately. So, the task of a discriminant analysis was solved twice. The volume of processed data in the first case was 413 21-dimensional vectors (200 vectors for the cases and 213 - for the controls) and in the second case - 419 26-dimensional vectors (206 - for the cases and 213 - for the controls).

RESULTS AND DISCUSSION

Lung cancer

According to the above described methodology, in the course of the mathematical data processing it was necessary to decide if the initially selected complex of 21 factors was sufficient for a reliable description of differences between the patterns of lung cancer cases and people without carcinogenic diseases. Three algorithms of the discriminant analysis were applied to solving this task: the algorithm based on the method of potential functions (5), and also the algorithms for construction of homogeneous discriminating committees with majority logic (6,7) and those with seniority logic (8,9).

The best results of recognizing the examination samples were obtained with the algorithm of potential functions: 97.3 % of correct answers in the "no cancer" class and 84.2 % - in the "lung cancer" class. The mean percentage of the correct recognition was 90.7 % with the confidence interval of 81%-96 %. We noticed that a large number of reliable decision rules was obtained by using three different algorithms. This fact indicates that the relationship between the complex of factors under study and the lung cancer morbidity was not found by chance.

Successive solution of discrimination tasks led to the following conclusions:

1. The number of both case and control groups was sufficient enough for obtaining reliable results;

2. The initial complex of features (factors) embraced most of lung cancer risk factor distinctive for Tula.

The highest pattern recognition results were achieved when using the sub-sets of 14-23 most informative factors. The analysis of informativeness of each factor interpreted as the strength of their effects on the development of lung cancer showed that smoking, occupational exposure to carcinogens, chronic nonspecific lung illnesses, and individual factors (age, sex) contributed to lung cancer morbidity in Tula the most. On the one hand, this result was well expected; however, within the framework of this study and using the systemic analysis we gave a quantitative assessment of their role, which was significantly different from that noted in the mono-factorial type of studies. Besides, the fact that our study results did not contradict those of other acknowledged cancer epidemiology studies increased our confidence in their reliability. This circumstance, in its turn, raised our confidence in those results of our study that were initially not so obvious.

For instance, according to the obtained estimates, the influence of the current industrial pollution of environmental media (ambient air, soil, and drinking water) in Tula on cancer morbidity can be considered moderate since we noted no significant link between the lung cancer incidence rate and the fact of people residing in districts with higher levels of environmental contamination by carcinogens (nickel, chromium, cadmium, etc.). We also found indirect evidence of effects of radiation from natural nuclides on the development of lung cancer, which was another nontrivial finding in this study.

If we take the effect of all 21 features on lung cancer development in the population of Tula for 100%, then, based on our findings, the contribution of smoking is 27%, alcohol abuse 4%, occupational exposure to carcinogens 14%, an aggregate of individual biological factors 32%, environmental factors 9%, radiation factors (X-ray examination and radon) 7%, others 7% (Figure 1).

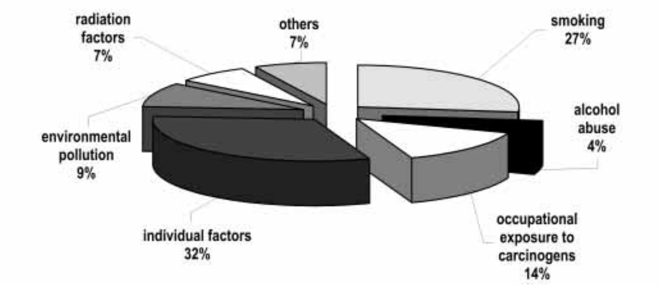


Figure 1. Percent contribution of different risk factors in lung cancer morbidity of the population in Tula

Stomach Cancer

The risk factors of stomach cancer were analyzed similar to those of lung cancer. The best results of pattern recognition 90 % of correct answers during the examination procedure were obtained using three different algorithms and the sub-sets of 16-18 most informative factors. This finding indicated that the initial complex contained the basic factors inducing stomach cancer in the population of Tula.

This study showed that if we take the effect of all 27 analyzed factors on the development of stomach cancer for 100 %, then the contribution of smoking is 7%, alcohol abuse 4%, occupational exposure to carcinogens12%, an aggregate of individual biological factors 28%, environmental factors 22%, radiation factors 5%, nutrition habits 10%, and others 12% (Figure 2).

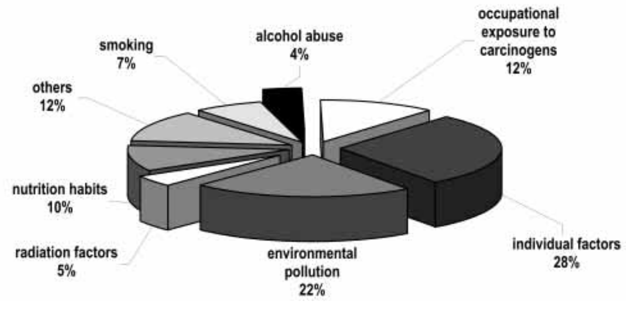


Figure 2. Percent contribution of different risk factors in stomach cancer morbidity of the population in Tula

The important finding of this study was the estimates of possible effectiveness of preventive measures developed to correct different "controlled" factors of carcinogenic risk. Thus, the best results of actions against cancer can be achieved through a complex of preventive measures. Table 3 provides the results of these prognoses.

Table 3. Prognosis of effectiveness of some measures for prevention of lung and stomach cancer in Tula

Lung cancer	Measures	Stomach cancer
10%	Treatment of chronic lung/stomach diseases	20%
12%	Cessation of smoking	3%
0.5%	Cessation of alcohol abuse	1.5%
11%	Elimination of occupational exposure to carcinogens	19%
	Reduction to the allowable level in environmental pollution with:	
1.5%	• lead	-
3%	• nickel	-
-	• cadmium	6%
11%	• total pollution	-
3.5%	Improvement of drinking water quality	5%
1%	Decrease in radon exposure	-
-	Cessation of eating excessively hot food	6%

Based on mathematical modeling, we searched for the optimum "preventive scenario". The model showed that simultaneous implementation of such measures as prevention and timely treatment of precancerous stomach diseases, elimination of occupational exposure to carcinogens in the town factories, reduction in ambient air pollution with cadmium to the minimal level, and cessation of smoking and alcohol abuse could reduce the stomach cancer incidence rate by 45 %. If such risk factors as occupational exposure to carcinogens, smoking, chronic lung diseases, frequency of X-ray examinations, drinking water quality and ambient air pollution were, in their turn, standardized, it could be possible to prevent 65 % of lung cancer incident cases.

CONCLUSION

The results of the epidemiological study carried out on the basis of methods of a multifactor analysis in the city of Tula showed the effectiveness of such an approach to solving complicated epidemiological tasks.

The described method enables one to evaluate the effectiveness of proposed cancer preventive measures and develop the most effective complex of preventive actions providing for the highest efficiency on the certain territory at low cost.

REFERENCES

1. Trapeznikov NN, Axel EM. Statistics of malignant neoplasms in Russia and NIS countries. Moscow; 2001 (in Russian)
2. Fleiss JL. Statistical Methods for Rates and Proportions. New York: John Wiley & Sons.
3. Duda OR, Hart PE. Pattern Classification and Scene Analysis. A Wiley-interscience publication. New York: John Wiley & Sons; 1973.

4. Kazantsev VS. The KVAZAR Package for Pattern Recognition and its Applications. International Journal of Software Engineering and Knowledge Engineering 1993;3(4) (in Russian)

5. Arkadyev AG, Braverman EM. Training computers to classify objects. Moscow: Nauka; 1971. p. 192.

6. Mazurov VD. The method of committees in optimization and classification tasks. Moscow: Nauka; 1990.

7. Ablow CM, Kaylor DJ. A committee solution of the pattern recognition problem. IEEE Trans Inform Theory 1965;11(3):453-5.

8. Beletsky NG. Application of committees for a multiple classification. In: Collection of Scientific Works "Numerical analysis for solving tasks of linear and convex programming". Sverdlovsk: Ural Scientific Center of the Academy of Sciences of the USSR; 1983. p.156-62.

9. Osborne WL. The seniority logic - a logic for committee machine. IEEE Trans Comput 1977; 26(12):1302-6.