Regression Models in Analysis of the Laryngeal Cancer Incidence Trends in Females in Podkarpackie Voivodship

Jan Gawełko, Marek Cierpiał-Wolan
University of Rzeszów, Poland

Sebastian Wójcik
Statistical Office in Rzeszów, Poland
Pedagogical University of Cracow, Poland

Abstract
Laryngeal cancer is the most common cancer of the head and neck. The main predisposing factor is exposure to cigarette smoke, alcohol, occupational factors and HPV infections. An analysis of the incidence of laryngeal cancer in Poland shows that since the beginning of the 1990s the incidence in females has been increasing, while in Podkarpackie Voivodship there is very slight increase. The aim of the work is to explore model-based assessment of dynamics of cancer incidence and to analyze the causes of changes in the incidence of laryngeal cancer in females in Podkarpackie Voivodship in the years 1990–2012, including lifestyle, socio-economic situation, and making comparisons with the incidence trends in the country as a whole. For this purpose, a retrospective analysis of cases of laryngeal cancer in Podkarpackie Voivodship in the years 1990–2012 has been performed. Data have been obtained from the publication of the Department of Epidemiology of Podkarpackie Center of Oncology in Rzeszow and the Centre of the Maria Skłodowska-Curie Institute of Oncology in Warsaw. Dynamics of cancer incidences derived from raw data is misleading. Therefore, to analyze the dynamics of the phenomenon three regression models have been used to remove random disturbance: ARX(1), SVR and Poisson regression model. The models have been compared based on standard statistics. In Podkarpackie the model-based absolute number of cases per year in females increased slightly between 1990–2012. In Poland, the model-based absolute number of cases in females increased over 30% in the last two decades in the years 1990–2012. The percentage share of cases of laryngeal cancer in females among all malignant cancers decreased by 0.5% to 0.4% in the years 1990–2012. The average age of the incidence for females increased from 55.6 years in 1989 to 65.4 in 2010. Dynamics analysis based on raw data solely may produce misleading results in opposition to a model based approach. A model-based approach seems to be relevant especially for the ill-behaved time series for such a number of cancer incidences.

Keywords: cancers in females, laryngeal cancer, epidemiology, regression models, forecasting methods

JEL: I12, I15, C53

Introduction
Head and neck tumors account for about 5% of all malignant tumors registered in Poland (approximately 7% in males and over 1% in females). In 2012, they were the sixth most common group of malignant neoplasms in males and the 12th in females in Poland. Squamous cell carcinomas represent over 90% (Kawecki and Nawrocki 2013). Laryngeal cancer is the most common cancer diagnosed in the head and neck area. The risk of developing this type of cancer increases with age, in Poland reaching the highest incidence in females in the age group 55–64 years, and in males at the age of 60–64 (Wojciechowska, Diłkowska, and Zatoński 2012). In Podkarpackie this age is significantly higher and the peak incidence is recorded in females aged 65–69, and in males aged 70+ (Grądalska-Lampart et al. 2014).
The main causative factor of squamous cell carcinomas of the head and neck is exposure to carcinogens contained in tobacco smoke, alcohol consumption, exposure to occupational agents and inadequate oral hygiene as well as mechanical irritation of mucous membranes. In the case of laryngeal cancer, simultaneous exposure to tobacco smoke and alcohol is particularly risky (carcinogenic synergistic effect). The aim of the study was to present the analysis of the dynamics and causes of changes in the incidence of laryngeal cancer in females in Podkarpackie in the years 1990–2012 in comparison with data on the country level, including lifestyle of the residents and socio-economic situation in the region, and to produce forecasts up to the year 2020 with the emphasis on applying regression models.

1 Material and methods

A retrospective analysis of the incidence of laryngeal cancer in the years 1990–2012 has been carried out for the area of Podkarpackie Voivodship (until 1999 the area of former voivodships: krośnieńskie, przemyskie, rzeszowskie and tarnobrzeskie). In order to calculate the crude rate and the percentage (structure indicators) for laryngeal cancer (C32 according to ICD 10 classification) standard statistical methods based on demographic data of the Statistical Office in Rzeszów, Poland were used, whereas similar data for Poland was compiled and the structure of morbidity was compared based on the data published by the Department of Epidemiology of the Centre of the Maria Skłodowska-Curie Institute of Oncology in Warsaw. Since the incidence of cancer in the analyzed period is often characterized by considerably high variability and randomness, the calculation of the dynamics of the year-on-year incidence for the series in which there are random changes is associated with the fact that the determined dynamics are not very reliable and often do not reflect what is happening in the analyzed period. In order to eliminate or reduce random disturbance distorting the information coming from the dynamics indicators an econometric model was used. The regression model was used to smooth the time series, which allowed for the removal of random disturbance. Such an approach was used to calculate the dynamics of the incidence of tongue and tonsils cancers in the United States for the population aged 20–44, among other things (Shiboski, Schmidt, and Jordan 2005), while developing a linear trend model for logarithmic crude incidence rates.

The available series of the years 1964–2012 was analyzed and the 1990–2012 period was selected to build the model. The lack of data from the years 1997–1998 was imputed by linear interpolation. An analysis of the series showed that for the trend models there is a very clear correlation between residuals of the first order (the higher orders are irrelevant). Therefore, an autoregressive component was included in the model after testing different options. Three different approaches were used: the linear ARX(1) model, Poisson regression and SVR model. All of the models were based on a calendar variable and lagged dependent variable as the explanatory variables. In the first step the smoothed value of the number of cases of cancers broken down to males and females was obtained using the ARX model (autoregressive with exogenous input) of the first order with the drift of the time variable as the exogenous variable. The model took the form

\[ y_t = \alpha + \beta y_{t-1} + \gamma t + \varepsilon_t, \]

where:
- \( y_t \) — number of cases in the \( t \) period,
- \( \varepsilon_t \) — random component,
- \( \alpha, \beta, \gamma \) — model parameters estimated by least squares method.

Another tool used was the Support Vector Machine regression (\( \varepsilon \)-SVR). This is a method based on support vectors. The desired approximating function has the form:

\[ f(x) = \langle w, x \rangle + b \quad \text{for} \quad w \in X, \; b \in R, \]

where:
- \( x = (t, y_{t-1}) \in R^2 \) — input data vector,
- \( \langle \cdot, \cdot \rangle \) — denotes scalar product,
- \( w = (w_1, w_2) \) — weight vector,
- \( b \) — intercept.
In the separable case, this problem can be formally written as the optimization problem in the following form:

$$\text{minimization } \frac{1}{2} \| w \|^2, \quad \text{where } \| w \|^2 = \langle w, w \rangle,$$

with conditions

$$\left\{ \begin{array}{l}
y_i - \langle w, x_i \rangle - b \leq \varepsilon \\
\langle w, x_i \rangle + b - y_i \leq \varepsilon. 
\end{array} \right.$$

In the non-separable case, the optimization problem can be extended so that it continues to minimize $\| w \|$ but also prediction errors $\xi_i$, which we accept. The optimization problem takes the form

$$\text{minimization } \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*),$$

with conditions

$$\left\{ \begin{array}{l}
y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\
\langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^*, \\
\xi_i, \xi_i^* \geq 0
\end{array} \right.$$

where $C > 0$ is a selectable constant. It reflects a compromise between a large margin and the admission of $\xi_i$ errors. The smaller the $C$, the smaller the impact of the second component in (5), and the position of the classification plane will prefer a larger margin between classes, even at the expense of the existence of large $\xi_i$ errors. The higher the $C$, the more important the second component in (5), and the minimization will focus on eliminating $\xi_i$ errors at the expense of lower margins. The size of the error is described by the so-called $\varepsilon$-insensitive cost function of the form

$$|\xi|_\varepsilon := \begin{cases} 0, & \text{for } |\xi| \leq \varepsilon \\
|\xi| - \varepsilon, & \text{otherwise}. \end{cases}$$

Let us note that the number of cases is number data. Their character suggests choosing Poisson regression as an alternative. Poisson regression is a generalized linear model, in which the logarithmic function is the binding function, and the modelled variable is in the form of Poisson distribution. In our case, the model takes the form

$$\ln E(y_t|y_{t-1}) = \alpha + \beta y_{t-1} + \gamma t.$$

Model parameters are estimated using maximum likelihood method by maximizing the following likelihood function

$$L(\alpha, \beta, \gamma|y_{t-1}, t) = \prod_{t=2} e^{-e^{(\alpha+\beta y_{t-1}+\gamma t)}} \cdot e^{y_t}(\alpha+\beta y_{t-1}+\gamma t) \cdot y_t!.$$

The dynamics of the incidence of cancers $d_{r,r+1}$ for the period of $r$ to $r+1$ was calculated for the analyzed period in accordance with the formula

$$d_{r,r+1} = \frac{\hat{y}_{r+1}}{\hat{y}_r} \cdot 100,$$

where $\hat{y}_r$ is number of cases in the year of the regression model.

The calculation of the dynamics using the smoothed series reduces the impact of a single year-on-year leap on the result.

### 2 Results

Table 1 presents results of estimation. Models were compared using the Akaike information criterion, coefficient of determination and MAPE. The analysis indicates a better quality of SVR results for all the analyzed time series.
Parameters are comparable only up to a sign, not their values: Poisson regression is an exponential model, SVR is estimated on the standardized variables thus constant is zero. The figures present original series and model-based smoothed series for comparison of performance of models. In the years 1990–2012, 3934 cases of laryngeal cancer were registered in Podkarpackie, including 3632 for males and 302 for females. The incidence for females amounted to 7.7% of all laryngeal cancers. The absolute number of cases per year for females decreased from 15 cases in 1990 to 12 cases in 2012 (fig. 1). The crude rates of the incidence decreased from 1.3 in 1990 to 1.1 in 2012 per 100 thousand population. The value of the standardized rates for laryngeal cancer for females increased from 0.5 to 0.7 in the years 1999–2012. In the years 1990–2012, the percentage share of laryngeal cancer among all cancers decreased from 0.7% to 0.3% for females (Gawełko 2016; Grądalska-Lampart et al. 2014). At the country level, data for the years 1990–2012 vary. The absolute number of cases of laryngeal cancer was 56448 in total, with 50082 cases for males and 6366 (fig. 3) for females. Thus, the cases for females amounted to 11.3% of all cases of laryngeal cancer. In Poland, in the last two decades, the number of cases among females increased by almost 50%—from 209 in 1990 to 308 in 2012 (Didkowska and Wojciechowska 2015; Didkowska et al. 2007; Didkowska et al. 2003; Didkowska, Wojciechowska, and Zatoński 2011, 2013; Wojciechowska et al. 2005, 2006; Wojciechowska, Didkowska, and Zatoński 2004, 2008, 2012, 2014; Wojciechowska et al. 2016; Zatoński and Tyczyński 1993, 1994, 1995, 1996, 1997, 1998, 1999).

In Poland, the crude rate of the incidence in females increased from 1.1 in 1990 to 1.5 in 2012. The value of the standardized rates for laryngeal cancer for females increased from 0.8 to 0.9 during 1990–2012. In Poland, the percentage share of the incidence of laryngeal cancer among all cancers in females decreased from 0.5% to 0.4% in the years 1990–2012. Based on the above material an attempt was made to assess the dynamics of laryngeal cancer in females in the years 1990–2012. The figures present original series and model-based smoothed series for comparison of performance of models.

| Tab. 1. Statistics for regression models of the numbers of cases of laryngeal cancer in the years 1990–2012 |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
|                | ARX(1) | Poisson | SVR | ARX(1) | Poisson | SVR |
| Statistics of |         |         |     |         |         |     |
| model          |         |         |     |         |         |     |
| $R^2$          | 0.31    | 0.29    | 0.45 | 0.93    | 0.87    | 0.95 |
| AIC            | 259.92  | 261.06  | 249.11 | 460.54  | 490.14  | 440.60 |
| MAPE           | 33.9%   | 34.4%   | 28.7% | 9.5%    | 14.9%   | 7.5% |
| Parameters     |         |         |     |         |         |     |
| ($p$-values)   |         |         |     |         |         |     |
| Constant       | 3.680   | 4.2657  | 0    | 27.98   | 1.5110  | 0   |
| ($0.00237$)    | ($< 0.001$) | ($< 0.001$) | ($< 0.001$) | ($< 0.001$) | ($< 0.001$) | ($< 0.001$) |
| Calendar       | 0.110   | 0.0079  | 3.969 | 2.15    | 0.0139  | 5.118 |
| variable       | ($0.00510$) | ($< 0.001$) | ($0.00423$) | ($< 0.001$) | ($< 0.001$) | ($< 0.001$) |
| $y_{t-1}$      | 0.202   | 0.0038  | 3.666 | 0.61    | 0.0238  | 2.888 |
| ($0.16374$)    | ($0.0923$) | ($< 0.001$) | ($< 0.001$) | ($< 0.001$) | ($< 0.001$) | ($< 0.001$) |

Fig. 1. The incidence of laryngeal cancer (C32) in females in Podkarpackie, absolute numbers

Parameters are comparable only up to a sign, not their values: Poisson regression is an exponential model, SVR is estimated on the standardized variables thus constant is zero. The figures present original series and model-based smoothed series for comparison of performance of models. In the years 1990–2012, 3934 cases of laryngeal cancer were registered in Podkarpackie, including 3632 for males and 302 for females. The incidence for females amounted to 7.7% of all laryngeal cancers. The absolute number of cases per year for females decreased from 15 cases in 1990 to 12 cases in 2012 (fig. 1). The crude rates of the incidence decreased from 1.3 in 1990 to 1.1 in 2012 per 100 thousand population. The value of the standardized rates for laryngeal cancer for females increased from 0.5 to 0.7 in the years 1999–2012. In the years 1990–2012, the percentage share of laryngeal cancer among all cancers decreased from 0.7% to 0.3% for females (Gawełko 2016; Grądalska-Lampart et al. 2014). At the country level, data for the years 1990–2012 vary. The absolute number of cases of laryngeal cancer was 56 448 in total, with 50 082 cases for males and 6 366 (fig. 3) for females. Thus, the cases for females amounted to 11.3% of all cases of laryngeal cancer. In Poland, in the last two decades, the number of cases among females increased by almost 50%—from 209 in 1990 to 308 in 2012 (Didkowska and Wojciechowska 2015; Didkowska et al. 2007; Didkowska et al. 2003; Didkowska, Wojciechowska, and Zatoński 2011, 2013; Wojciechowska et al. 2005, 2006; Wojciechowska, Didkowska, and Zatoński 2004, 2008, 2012, 2014; Wojciechowska et al. 2016; Zatoński and Tyczyński 1993, 1994, 1995, 1996, 1997, 1998, 1999). In Poland, the crude rate of the incidence in females increased from 1.1 in 1990 to 1.5 in 2012. The value of the standardized rates for laryngeal cancer for females increased from 0.8 to 0.9 during 1990–2012. In Poland, the percentage share of the incidence of laryngeal cancer among all cancers in females decreased from 0.5% to 0.4% in the years 1990–2012. Based on the above material an attempt was made to assess the dynamics of laryngeal cancer in females in the years 1990–2012. The figures present original series and model-based smoothed series for comparison of performance of models.
The year-to-year dynamics for the period 1990–2012 with real data amounted to 98.39, while the dynamics calculated from the smoothed data ranged from 100.21 to 100.58, depending on the model. This shows that the observed decrease is the result of random swings and once they are removed or reduced, we can observe a slight upward trend of incidence for this period. For instance, year-to-year dynamics for 1991–2012 is 102.048 which could be interpreted as an increase what is inconsistent with previous conclusion from dynamics for 1990–2012 with real data. The opposite situation is observed in Poland. The dynamics calculated with real data amounted to 102.19, and with smoothed data it ranged from 99.98 to 100.07. In the analyzed period, considering signs of parameters and their p-values, we conclude there was a slight increase.
To summarize, we can see that the use of the model significantly reduced the volatility of the dynamics of the cancer incidence. Basically, the SVR model smoothed the time series the most, as evidenced by the values of the dynamics.

3 Discussion

The differences between geographical distribution of mortality and cancer incidence in Poland have been studied since the 1960s. These studies were aimed at differences between individual areas of intensity of smoking and alcohol consumption, exposure to occupational agents, types of diet and reproductive behavior, among other things (Zatoński and Pukalla 1993). In this regard, Podkarpackie, which is a part of the Eastern Poland macro-region, is characterized by a number of individual and specific features, which have a significant impact on problems discussed in this paper. In order to find the causes of this distinctiveness, known risk factors were analyzed against the background of social and moral transformations.

In terms of socio-economic development, Podkarpackie (as other voivodships of the so-called “eastern wall”), is in the group of voivodships with lower transport accessibility, market responsiveness, level of industrial development, transformation of the economy and the labor market situation (Długosz 2008). Also, the phenomenon of unemployment and severely perceived discrimination against certain categories of workers—e.g., by gender (females), is connected with the inefficient functioning of the market (Kaszuba and Sokół 2005; Minkiewicz 2001). Similarly, GDP per capita is one of the lowest. In 2013, in Podkarpackie, it was 30.6 thousand and was slightly higher than in Lubelskie Voivodship—PLN 30.4 thousand, and slightly lower than in Świętokrzyskie Voivodship—PLN 31.4 thousand, with GDP for Poland at the level of PLN 43.0 thousand, and the highest one in Mazowieckie Voivodship—PLN 69.0 thousand. Average monthly gross wages and salaries amounted to PLN 3 412 in 2014, which placed the region 15th in Poland, although the difference between Podkarpackie and the average for Świętokrzyskie, Lubuskie and Kujawsko-Pomorskie voivodships did not exceed PLN 30. The registered unemployment rate with regard to voivodships was 11.6% in Podkarpackie Voivodship in August 2016, which placed the region 14th in Poland. At the same time, unemployment for the country was 8.4% in total, 10.2% in Lubelskie Voivodship and 11% in Świętokrzyskie Voivodship. Residents of Podkarpackie live longer than people in other voivodships. In 2012, an increase in the life expectancy of 1.2 years for males was registered compared to 2008, and of 0.8 years for females. Also, the rate of the average life expectancy in Podkarpackie is the highest for males and the second highest for females in the country. In 2014, these numbers for Poland amounted to 73.8 years for males and 81.6 for females (Dmochowska 2015; Rutkowska 2015). Epidemiological data on mortality and cancer rates for Podkarpackie Voivodship are clearly different from data for other regions of the country. The number of deaths from laryngeal cancer (C32) in females was 2 per 147 in the country in 2011, and together with Lubelskie Voivodship (also 2 deaths per 147) it occupied the last place in the list of deaths due to laryngeal cancer in different voivodships in Poland (Didkowska, Wojciechowska, and Zatoński 2013).

The frequency of laryngeal, pharyngeal and esophageal cancer is 20 times higher in the population of smokers than in non-smokers. In the countries of Central and Southern Europe the percentage of male smokers is higher than female smokers, but the frequency of smoking in females is very high. The reason for this can be socio-cultural changes and an increase in the acceptance of smoking by females, emancipation of females and the promotion of smoking (often by tobacco companies) as an integral element of “a western lifestyle,” as well as increased availability of tobacco products in packages addressed to females (Kaleta 2013).

In Eastern Europe, the percentage of female smokers has been decreasing very slowly, while in Western Europe, USA, Canada and Australia there was a significant reduction in the proportion of smoking females and mortality due to smoking related diseases (Kowalewska 2013). Research

proves that females are less likely to smoke than males, which also applies to females at a young age (high school girls, students), also at childbearing age laryngeal cancer is very rare. Females are also less dependent on nicotine than males, but it is significantly more difficult for them to quit the addiction (Klimberg and Żarnowska 2006; Kobiorska-Nowak and Mierzwa 2005; Kowalewska 2013; Rzeźnicki et al. 2007). A worrying phenomenon is, however, a high level of passive exposure to tobacco smoke (Szcześćh et al. 2014). In Poland, the documented change in the percentage of smokers may be the result of the implementation of preventive programmes on tobacco control. 51% of males and 25% of females smoked daily in Poland in the years 1990–1994, and 30.9% of males and 17.9% of females in 2011 (Dmochowska 2015).³

The results of the 2005 National Multicenter Health Survey programme (WOBASZ) showed a great regional diversity among female smokers in Poland. In 10 voivodships there are more than 25% of females regularly smoking cigarettes. The highest numbers are in Warmińsko-Mazurskie, Lubuskie and Dolnośląskie voivodships—32%–34%, and in Lubelskie and Świętokrzyskie voivodships—23% and 22%. The habit of smoking is the least frequent in Podkarpackie Voivodship—10% and Małopolskie Voivodship—12%. Females who never smoked represent the highest percentage in Podkarpackie Voivodship—80%, and the lowest in Warmińsko-Mazurskie Voivodship—58% (Polakowska et al. 2005). In this context, it is worth mentioning that from 1999–2012 the national average incidence of laryngeal cancer in females increased from 1.2/100 000 to 1.5/100 000, and half of the voivodships reported an increase in the incidence of laryngeal cancer in females. These voivodships included: Warmińsko-Mazurskie Voivodship (increase from 0.4/100 000 to 2.7/100 000), Lubuskie Voivodship (increase from 0.4/100 000 to 2.3/100 000).

Alcohol, similarly to smoking, is a factor increasing the risk of developing oral cavity, laryngeal and pharyngeal cancers. It is estimated that approximately 40% of patients with head and neck cancer are alcoholics. In the USA, France and Italy alcohol and/or tobacco are responsible for about 75% of these tumors. In Nordic countries that percentage amounted to 35% for males and to 27% for females. Alcohol increases the risk of cancer approximately 2–3 times, and abusing it along with tobacco increases that risk approximately 15 times (Castellsague et al. 2004; Franceschi et al. 1990; Jurkiewicz, Dzaman, and Rapiejko 2006; Majszyk, Bruzgielewicz, and Osuch-Wójcikiewicz 2014; Mashberg et al. 1993). Alcohol consumed by females aged 20–39 in the amount of 3.9 g per day and by females aged over 40 in the amount of 4.9 g per day is a risk for developing laryngeal cancer, according to WHO.

A longitudinal study on dissemination of drinking and alcohol abuse among females in Poland showed that in the years 1971 to 1991, the highest number of female drinkers were in the age group of 21–34 years, and the lowest number in the age group of over 65 years. It has also been shown that the problem affected mostly females living in cohabitation, employed and unemployed, unmarried, and the least: widows, females running a household, pensioners and annuitants (Bury, Godlewski, and Wojtyś 2000). Females’ social roles can be a risk factor, but these roles are imposed on them by age, gender, membership of a particular social group or ethnic group. Females overcoming alcohol addiction are much more likely to be compulsive smokers than non-smoking females at the same age (82% and 34% respectively) (Kobiety i alkohol 1997; Anderson and Baumberg 2007; Bernacka et al. 2012; Bury, Godlewski, and Wojtyś 2000; Józefowicz-Korczyńska et al. 2014; Jurkiewicz, Dzaman, and Rapiejko 2006).

Conclusions

In the years 1990–2012, in Poland, the incidence of laryngeal cancer in females shows an increase both in absolute numbers and crude rates, standardized rates and dynamics. In the same period, the incidence of laryngeal cancer in females from Podkarpackie shows only slight increase in terms of absolute numbers and crude rates, standardized rates and dynamics. The model-based approach produces results more robust to random disturbance, especially in the case of time series with low

---

incidence of cancer. In the years 1989–2010, in Podkarpackie, the average age of the incidence in both females and males increased substantially from the mid fifth to the mid sixth decade of life. Industrialization, which is lower than in other regions, and constant reduction of emission of gases and dust in Podkarpackie may be related to the laryngeal cancer incidence trends. The percentage of female smokers and alcohol abusers, which was recorded in Podkarpackie Voivodship, and which is the lowest in Poland, may indicate a correlation between moral factors and laryngeal cancer incidence trends in this region.

References


