Liet. matem. rink., T. 43, spec. nr., 509-514

Analysis of cervical cancer risk factors

Laura LAPKAUSKAITĖ (MII), Živilė GUDLEVIČIENĖ (OI)
e-mail: laural@ktl.mii.lt

1. Introduction

The aim of this paper is to investigate risk of invasive cervical cancer in relation to the presence and type of Human Papillomavirus (HPV) infection and its epidemiological co-factors in Lithuania. Risk factors were widely studied in case-control studies in Spain, Columbia, Thailand, Brazil, Mexico, USA ([4], [5]). A case-control study of risk factors among Lithuanian women was carried out at the Oncology Institute (OI) of Vilnius University in 2000–2002.

In this paper, to estimate the risk of cervical cancer associated with HPV infection and risk factors, odds ratios (age-adjusted) with 95% confidence intervals are calculated as approximations of relative risks by use of unconditional logistic regression ([1] is a typical reference to categorical data analysis). It reveals the main risk factors, as the presence of HPV 16, lower socio-economic conditions, limited education, younger age of sexual debut, ever smoking, late age of the first menstruation, the increasing number of births, and the increasing interval since last Pap smear. The results are presented in Chapter 4.

Logistic regression has been widely used in medical research recently (see, e.g., [5]). In order to describe a more sophisticated structure, log-linear models are used. They reveal that cervical cancer is directly related to the presence of HPV, age, and the interval since last Pap smear. Since log-linear models are not as popular yet as logistic regression, the definition and some of their examples are presented in Chapter 3.

2. Statistical analysis

2.1. *Description of the data*

The total of 200 patients with cervical carcinoma and 225 control subjects were included in the study and examined with a view to determine the presence and type of HPV. 145 patients and 194 control subjects were interviewed to obtain information with regard to 22 supposed cervical cancer risk factors (age, height, weight, sexual behavior, reproductive history, contraceptive practices, smoking habits, histories of sexually transmitted diseases, cervical cytologic screening histories, and socioeconomic status).

2.2. *Odds ratios*

Logistic regression is widely used for modelling of case probability. In medical research it is usually used to calculate odds ratios and their 95% confidence intervals as a measure of relative risk. Since the occurrence of cancer obviously depends on age, and the case and control groups significantly differ in age, age-adjusted odds ratios are calculated. It simply means that age is included in the list of predictors.

Consider a dichotomous response variable Y with outcomes *event* ($Y = 1$) and *no-nevent* ($Y = 0$) and a dichotomous risk factor variable X that takes values $X = 0$ (if factor X is absent) and $X = 1$ (X is present). The *odds* are a ratio of event and nonevent probabilities. The *odds ratio* (*OR*) is defined as a ratio of the odds for those with the risk factor ($X = 1$), O_1, to the odds for those without the risk factor ($X = 0$), O_0. Thus,

$$OR = O_1/O_0 = \frac{\Pr(Y = 1|X = 1)}{\Pr(Y = 0|X = 1)} : \frac{\Pr(Y = 1|X = 0)}{\Pr(Y = 0|X = 0)}.$$

In applications the odds ratio is usually interpreted as a measure of risk (e.g., how much HPV infection increases the risk of cancer (*event*)).

Logistic regression enables us to describe the impact of several risk factors:

$$\log(\Pr(Y=1|X_1,\ldots,X_k)/\Pr(Y=0|X_1,\ldots,X_k)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

Here β_i are parameters of the model. A better interpretation than β_i itself, has the odds ratio e^{β_i}, which describes a change in the odds for any increase by a unit in the corresponding risk factor X_i.

Results. In the fitted logistic regression model, HPV 16 infection appeared to be the strongest risk factor of cervical cancer. Its odds ratio of *Positive* versus *Negative* (reference) is 86.4 (odds to have cervical cancer are 86 times higher for women infected by HPV16). Table 1 presents age-adjusted odds ratios associated with cervical cancer risk factors, frequencies of persons in case and control groups by categories of risk factors. Other significant risks included lower socio-economic conditions, increasingly limited education, decreasing age at the first intercourse, late age of the first menstruation, the increasing number of live births, ever smoking, and the increasing interval since last Pap smear.

In this study, other risk factors (increasing number of sexual partners, the use of oral contraceptives, number of abortions, etc.) have also been investigated, but the significant association between them and the cervical cancer development was not found.

In order to investigate a more sophisticated structure, log-linear models are used.

2.3. *Log-linear models*

A formal definition of a log-linear model and its interpretations illustrated by examples are presented in this chapter.

Table 1

Risk factors for cervical cancer in Lithuania

Risk factor	Categories	Case subj.	Control subj.	Odds ratio*	95% conf. interval	Pr > χ^2
HPV16	Negative	20	164	1.00		<0.0001
	Positive	180	61	86.40	(34.0; 251.3)	
Socio-economic status	Other	42	86	1.00		<0.0001
	Blue-collar w.	52	17	3.66	(1.59; 8.71)	
	White-collar w.	51	91	0.47	(0.24; 0.92)	
Education	Higher	20	96	1.00		<0.0001
	Professional	68	55	7.32	(3.50; 16.27)	
	High	29	34	7.09	(2.71; 19.87)	
	Uncompl. high	28	9	32.42	(8.13; 151.51)	
Age at first intercourse, y	> 20	53	65	1.00		0.0036
	18-20	74	80	2.88	(1.49; 5.77)	
	< 18	18	49	3.74	(1.42; 10.48)	
Age of first menstruation, y	12-15	112	163	1.00		0.0082
	< 12	2	17	0.18	(0.02; 0.85)	
	> 15	31	13	2.95	(1.22; 7.51)	
No. of live births	1-2	91	110	1.00		0.0317
	0	14	70	0.57	(0.24; 1.29)	
	$\geqslant 3$	40	14	2.24	(1.04; 5.02)	
Smoking status	Never	105	148	1.00		0.0077
	Ever	40	46	2.50	(1.29; 4.98)	
Int. since last Pap smear, y	$\leqslant 1$	23	73	1.00		0.0258
	$\geqslant 2$ or never	122	120	2.11	(1.10; 4.11)	

* age-adjusted

Suppose $\mathcal{L}$ is an index set of finite cardinality of a cross-classified data table, where $n = |\mathcal{L}|$ is the total number of elements in the set $\mathcal{L}$. Consider data $\{Y_l, l \in \mathcal{L}\}$, where Y_l is a count of the lth cell.

The set $\{Y_l, l \in \mathcal{L}\}$ is isomorphic to the vector $(Y_{i(l)},\ l \in \mathcal{L})^T$, where $i(l)$: $\mathcal{L} \to \{1, \ldots, n\}$ is some enumeration of the cells. In the sequel we identify $(Y_{i(l)},\ l \in \mathcal{L})^T$ to $\mathbf{Y} = (Y_l,\ l \in \mathcal{L})^T$.

In fact Y_l can be more general. The definition of a log-linear model does not require for Y_l to be nonnegative or discrete. Statistical inference for Y_l's is most thoroughly developed in three sampling settings: Poisson sampling, multinomial sampling, and product multinomial sampling.

Let us denote:

$$M_l = \mathbf{E}Y_l, \quad m_l = \log(M_l), \quad l \in \mathcal{L}.$$
$$\mathbf{M} = \mathbf{EY}; \quad \mathbf{m} = (m_l,\ l \in \mathcal{L})^T.$$

DEFINITION ([2]). A random vector $\mathbf{Y}$ follows a *log-linear model* if $\mathbf{m} \in \mathcal{S}$ for some known linear subspace $\mathcal{S}$ of $\mathbf{R}^n$.

Several examples to illustrate this rather general concept are presented below.

EXAMPLE 1. Let us analyze a cross-classified table of two factors A and B. The cross-classified table of frequencies, based on N observations, is as follows:

	B	
A	Y_{11}	Y_{12}
	Y_{21}	Y_{22}
	Y_{31}	Y_{32}

Then $\mathbf{Y} = (Y_{11}, Y_{12}, Y_{21}, Y_{22}, Y_{31}, Y_{32})^T$, $\mathcal{L} = \{(i,j)\colon\ i = 1,2,3,\ j = 1,2\}$, $n = 6$.

For independent random factors A and B, the underlying log-linear model (the *main-effects-only model* $[A]\ [B]$) takes the following form:

$$\ln M_{ij} = \alpha + \alpha_i^{(A)} + \alpha_j^{(B)}, \quad \text{where } i = 1,2,\ j = 1.$$

Another form is:

$$\ln \mathbf{EY} = \alpha e + \alpha_1^{(A)} e_1^{(A)} + \alpha_2^{(A)} e_2^{(A)} + \alpha_1^{(B)} e_1^{(B)},$$

which means that for this model with the Poisson sampling ($\sum_{ij} Y_{ij} = N$ is a random variable), $\mathcal{S}$ is a 4-dimensional linear subspace, generated by column vectors $e = (1,1,1,1,1,1)^T$, $e_1^{(A)} = (1,1,0,0,0,0)^T$, $e_2^{(A)} = (0,0,1,1,0,0)^T$, $e_1^{(B)} = (1,0,1,0,1,0)^T$.

EXAMPLE 2. Let us analyze a cross-classified table of three factors A, B, C with the number of states n_A, n_B, and n_C, respectively.

a) The model with arbitrary positive cell means M_{ijk} is called a *saturated model* ($[A\ B\ C]$). It is usually parameterized in the following way:

$$\ln M_{ijk} = \alpha + \alpha_i^{(A)} + \alpha_j^{(B)} + \alpha_k^{(C)} + \alpha_{ij}^{(AB)} + \alpha_{ik}^{(AC)} + \alpha_{jk}^{(BC)} + \alpha_{ijk}^{(ABC)}, \qquad (1)$$

where $i = 1, \ldots, n_A - 1$, $j = 1, \ldots, n_B - 1$, $k = 1, \ldots, n_C - 1$ and the parameter α introduced has a similar interpretation as in ANOVA models, namely, α^A, α^B, and α^C describe so-called main effects, and, e.g., $\alpha^{(AC)}$ describes the effects of log-linear interaction between A and C.

b) The model $[A\ C]\ [B\ C]$ states that A and B are *conditionally independent*, given C:

$$\ln M_{ijk} = \alpha + \alpha_i^{(A)} + \alpha_j^{(B)} + \alpha_k^{(C)} + \alpha_{ik}^{(AC)} + \alpha_{jk}^{(BC)}, \qquad (2)$$

where $i = 1, \ldots, n_A - 1$, $j = 1, \ldots, n_B - 1$, $k = 1, \ldots, n_C - 1$.

Results. The log-linear analysis revealed that cervical cancer is directly related to the presence of HPV, age, and the interval since last Pap smear (Fig. 1). Table 2 presents the main effects and log-linear interactions which are involved in the model. The chi-square

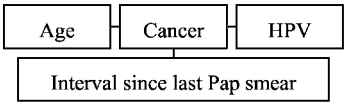

Fig. 1. Cervical cancer risk factors.

Table 2
Maximum likelihood analysis of variance

Source	DF	Chi-Square	Pr > χ^2
HPV	1	20.73	<0.0001
Interval since last Pap smear	1	61.16	<0.0001
Age	4	30.91	0.0039
Cancer	1	9.57	0.0020
Cancer*HPV	1	82.00	<0.0001
Cancer*Age	4	83.06	<0.0001
Cancer*Int. since last Pap smear	1	18.39	<0.0001
Likelihood Ratio	**20**	**19.61**	**0.4824**

test for each effect is a Wald test based on the information matrix from the likelihood calculations. The likelihood ratio statistic compares a specified model with the saturated model (2) and is an appropriate goodness-of-fit test for the model. The obtained model [HPV Cancer] [Age Cancer] [Int Cancer] states that three factors, namely, the presence of HPV, age, and the interval since last Pap smear, are conditionally independent given the presence of cervical cancer (3).

3. Conclusions and discussion

Many epidemiological studies have found a strong association between Human Papillomavirus and cervical neoplasia (e.g., [4]). HPV 16 is also the most important risk factor in cervical carcinogenesis in our investigation (OR = 86.4; 95% CI = (34.0, 251.3)). Other studies reported an elevated risk of cervical cancer associated with nonattendance of school, lower socioeconomic level, and ever smoking ([5], [6]). Our data analysis confirmed the influence (age-adjusted) of a lower socio-economic level and ever smoking, other risk factors being the lower educational level, younger age of the first intercourse, late age of the first menstruation, increased number of live births, and the increasing interval since last Pap smear (Table 1).

Using the log-linear analysis, a direct association between a cervical cancer development and the prevalence of HPV 16 DNA, the interval since last Pap smear, and women's age has been established (Table 2). In addition, those three factors can be considered as independent.

Acknowledgement. The authors would like to thank Marijus Radavičius for his valuable advice.

References

[1] A. Agresti, *Analysis of Ordinal Categorical Data*, John Wiley & Sons, Inc., New York (1984).
[2] T.J. Santer, D.E. Duffy, *The Statistical Analysis of Discrete Data*, Springer-Verlag, Inc., New York (1989).
[3] N.E. Breslow, N.E. Day, Statistical methods in cancer research: the analysis of case-control studies, *IARC Sci. Publ.*, **1**(32), Lyon, IARC (1980).
[4] *Human Papillomaviruses*, IARC, Monographs on the evaluation of carcinogenic risks to humans, **64** (1995).
[5] S. Chichareon, R. Herrero *et al.*, Risk factors for cervical cancer in Thailand: a case-control study, *J. Natl. Cancer Inst.*, **90**(1), 50–57 (1998).
[6] S. Sanjose, F.X. Bosch *et al.*, Socioeconomic differences in cervical cancer: two case-control studies in Colombia and Spain, *Amer. J. Publ. Health*, **86**, 1532–1538 (1996).

Gimdos kaklelio vėžio rizikos faktorių analizė

L. Lapkauskaitė, Ž. Gudlevičienė

Šis taikomasis darbas skirtas gimdos kaklelio vėžio rizikos faktorių nustatymui, naudojant logistinę regresiją bei log-tiesinius modelius. Ligos-kontrolės studija buvo atlikta Vilniaus universiteto onkologijos institute 2000–2002 m. Naudojant log-tiesinius modelius, rasta tiesioginė vėžio priklausomybė nuo infekuotumo žmogaus papilomos virusu, amžiaus ir laiko intervalo nuo paskutinio Pap tepinėlio. Kadangi log-tiesinis modeliavimas nėra taip plačiai žinomas, kaip logistinė regresija, pateiktas trumpas jo įvadas.