Joint Modeling of Disease Pairs

Gergely TÓTH[1], Sándor KABOS[1,*] and György SURJÁN[2]

[1] ELTE Eötvös Loránd University, Faculty of Social Sciences, Departement of Statistics, H-1117 Pázmány sétány 1/A. Budapest, Hungary.
[2] GYEMSZI National Institute for Quality- and Organizational Development in Healthcare and Medicines, H-1054 Hold utca 1. Budapest, Hungary.
E-mails: toth.gergo@gmail.com; kabos@tatk.elte.hu; surjan.gyorgy@gyemszi.hu

* Author to whom correspondence should be addressed; Tel.: +3613722500 ext. 6828

Received: 10 January 2012 /Accepted: 29 February 2012 / Published online: 10 March 2012

Abstract
Aim: Exploring the spatial patterns in joint distribution of incidences of two diseases. *Material and method:* A Poisson-Binomial regression model was used in analysing hospitalisation counts data in Hungary, 2008. Model parameters were estimated by MCMC implemented in WinBUGS. *Results:* Spatial patterns of laryngeal and hypopharyngeal cancer differ significantly from that of their ratio. *Conclusion:* The Poisson-Binomial model proposed here might help clarify us the different spatial dependencies of the sum and the ratio of incidences of two diseases.

Keywords: Epidemiologic Studies; Regression Analysis; Poisson Distribution; Binomial Distribution.

Introduction

Lots of relevant statistical analyses are available for malignant neoplasm of hypopharynx and malignant neoplasm of larynx, but most of them are based on clinical data. Brugere et al. stated that locations of cancer of larynx, pharynx, and mouth differ significantly according to the consumption of alcohol and brown tobacco [1]. Statistical term 'significant' is used here in reference to their data as a sample (3500 inpatients of hospital of Institute Curie, Paris examined from 1975 to 1982), but there is no clear definition of the population from where this sample were taken. Further investigations are needed when adapting these results in an analysis of public health data.

Several papers (e.g. Nakai [2]) demonstrated evidences that peptic ulcer is a multifactorial disease with both bacterial (Heliobacter pylori) and psychosomatic causes. Davidovic et al. pointed out that the age-risk dependencies of gastric and duodenal ulcer are different from each other [3].

Hospitalisation counts classified by these two diagnoses are commonly assumed to be conditionally independent Poisson random variables in epidemiological modelling [4]. Speaking in regression framework: responses are independent conditioned on all the explanatory variables are controlled. The real situation is that only a small part of relevant explanatory variables are known in epidemiologic studies. Statisticians remove the effects of known factors (most commonly the age and gender) when calculating standardised incidence ratio (SIR), but there is no commonly-acknowledged statistical method testing for association of SIR values of two diseases. The common association tests cannot be used correctly because of unknown distribution of SIR values.

One possibility to overcome this problem is the shared component model for the joint spatial analysis of several diseases proposed by Knorr-Held and Best [5]. The Poisson-Binomial model we will use here decomposes the joint distribution of counts into marginal and conditional distributions. Our motivating ideas are similar to that of the shared component model but our method is restricted to two variables only.

Material and Method

Hospitalisation Incidence Data

The analyses are based on GYEMSZI data of hospital admissions (incl. clinical units) classified by admission diagnosis (using its ICD-10 codes), age category, gender and place of residence of patients. This database is a Hungarian national data repository that stores records of all hospital cases based on a defined standard minimal basic data set. The data collection started in 1993; the current database stores data consistently since 2004. Patient identifier is replaced by a pseudocode that does not identify the real person but enables to couple the records belonging to the same patient.

Diseases treated here are the malignant neoplasm of hypopharynx (ICD-10 code: C13) and the malignant neoplasm of larynx (ICD-10 code: C32). The mean annual hospitalisation incidence rate in Hungary is about 25 inpatient cases per 100,000 citizens/year for C13 and 60 inpatient cases per 100,000 citizens/year for C32.

Results of an analysis of gastric ulcer (ICD-10 code: K25) and duodenal ulcer (ICD-10 code: K26) data are demonstrated in Appendix 3. The mean annual hospitalisation incidence rate in Hungary is about 180 inpatient cases per 100,000 citizens/year for K25 and 150 inpatient cases per 100,000 citizens/year for K26.

The Poisson-Binomial Regression Model

The conditional Poisson-Binomial regression method is specified in Appendix 1. The well-known technique of decomposing a joint distribution into its marginal and conditional distributions stems from Barndorff-Nielsen [6]. Many statistical textbooks (e.g. Agresti [7] Ch. 3.1.) treat decomposing a joint distribution into its marginal and conditional distributions the case we use here: if $Y^{(1)}$ and $Y^{(2)}$ are two independent Poisson random variables and $Z=Y^{(1)} + Y^{(2)}$ then Z is also Poisson and conditional distribution $Y^{(1)}|Z$ is Binomial of order Z Our idea was to apply this decomposition in the Poisson regression framework and to estimate parameters of Z and $Y^{(1)}|Z$ independently. An example of artificial data is shown in Appendix 2. to demonstrate how to describe different spatial patterns of Z and $Y^{(1)}|Z$ by fitting our model. Computer implementation of this model is easy using the built-in implementation of Knorr-Held and Best method in WinBUGS [8].

Results

Analysis of C13 and C32 Incidence Data

Raw incidence rates were transformed into the standardised ones. The table of hospitalisation incidence numbers classified by age and gender (Table 1.) divided by the population numbers cell by cell form the table of age-gender specific hospitalisation incidence rates (Table 2.).

Table 1. Hospitalisation incidences of C13 and C32 patients by age and gender in Hungary, 2008

C13	MALE	FEMALE	Total
0-34	5	3	8
35-64	1754	260	2014
65-xx	440	79	519
Total	2199	342	2541

C32	MALE	FEMALE	Total
0-34	11	5	16
35-64	3379	643	4022
65-xx	1653	270	1923
Total	5043	918	5961

Standardised Incidence Ratio (SIR) has been calculated for each of the 174 sub-regions: the actual incidence number divided by the expected incidence number. The expectation has been calculated under the assumption of risk homogeneity (the risk of being hospitalised depends only on age and gender), so it is the sum for all age-gender categories the population number of this age-gender category multiplied by the age-gender specific rate of Table 2.

Table 2. Hospitalisation incidence rates (/100 000 persons) of C13 and C32 patients by age and gender in Hungary, 2008

C13	MALE	FEMALE
0-34	0.22	0.14
35-64	90.63	12.38
65-xx	74.23	7.66

C32	MALE	FEMALE
0-34	0.49	0.23
35-64	174.60	30.62
65-xx	278.86	26.18

The following set of maps of SIR values of sub-regions can be conceived as a visual test for the assumption of risk homogeneity.

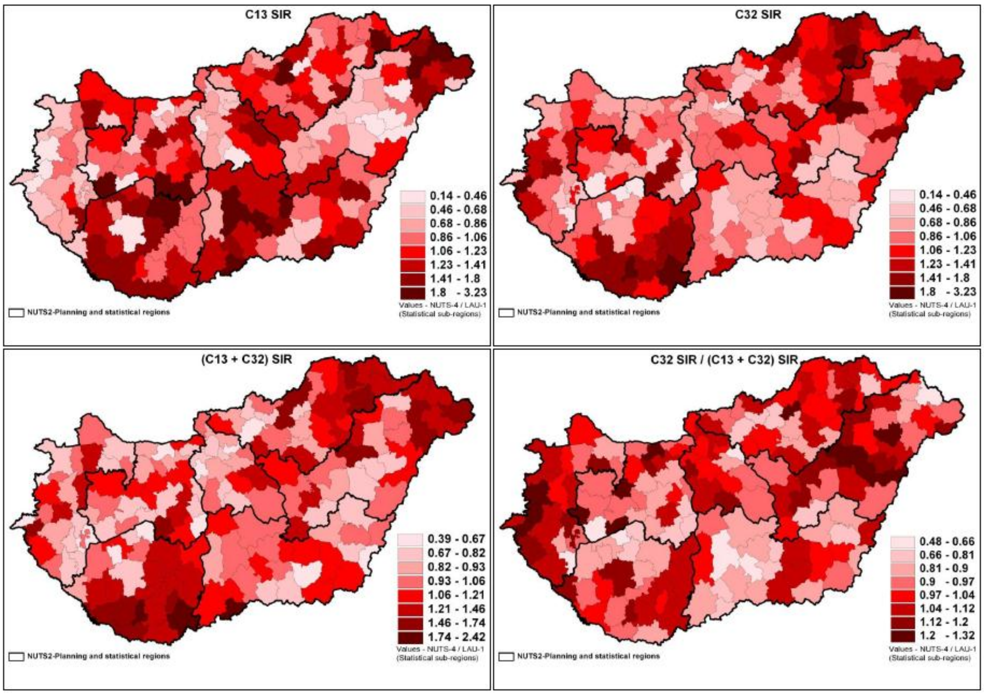
C13 SIR
0.14 - 0.46
0.46 - 0.68
0.68 - 0.86
0.86 - 1.06
1.06 - 1.23
1.23 - 1.41
1.41 - 1.8
1.8 - 3.23
Values - NUTS-4 / LAU-1 (Statistical sub-regions)
NUTS2-Planning and statistical regions
C32 SIR
0.14 - 0.46
0.46 - 0.68
0.68 - 0.86
0.86 - 1.06
1.06 - 1.23
1.23 - 1.41
1.41 - 1.8
1.8 - 3.23
Values - NUTS-4 / LAU-1 (Statistical sub-regions)
NUTS2-Planning and statistical regions
(C13 + C32) SIR
0.39 - 0.67
0.67 - 0.82
0.82 - 0.93
0.93 - 1.06
1.06 - 1.21
1.21 - 1.46
1.46 - 1.74
1.74 - 2.42
Values - NUTS-4 / LAU-1 (Statistical sub-regions)
NUTS2-Planning and statistical regions
C32 SIR / (C13 + C32) SIR
0.48 - 0.66
0.66 - 0.81
0.81 - 0.9
0.9 - 0.97
0.97 - 1.04
1.04 - 1.12
1.12 - 1.2
1.2 - 1.32
Values - NUTS-4 / LAU-1 (Statistical sub-regions)
NUTS2-Planning and statistical regions

Figure 1. SIR values of the 174 sub-regions of Hungary, **(a)** C13 SIR, above left **(b)** C32 SIR, above right **(c)** (C13+C32) SIR, down left **(d)** C32/ (C13+C32) SIR, down right

SIR values appearing on Figure 1 seem to have spatial autocorrelation and cross-correlation between the two diseases. Moran statistics might be used in testing for risk homogeneity, but it is obvious, that the hypothesis of homogeneity of SIR values must be rejected. Moreover, one can easily reveal that maps **a**, **b** and **c** are similar to each other, while map **d** is different. These impressions can be turned into regular statistical tests using Poisson-Binomial model as follows.

Table 3 consists of parameter estimation of Poisson-Binomial model realised in WinBUGS. Notations of Appendix 1 can be summarised using the multilevel generalised linear model terminology. There are two outcome variables: Z and Y. The first one, Z consists of the sum of incidence numbers of (C13 + C32) by age-category, gender and sub-region and assumed to be conditionally independent realisations of a Poisson random variable with random parameter mu. The second one, Y consists of the incidence numbers of C32 by the same categories and assumed to be Binomial of order Z and with random parameter p. Explanatory variables $X^{(\mu)}$ and $X^{(P)}$ are the indicator variable of Dél-Dunántúl, South-West Region of Hungary (that is they are 1 for sub-regions included in this South-West Region and they are 0 for all other parts of country). Linear predictor on the first level is the natural logarithm of parameter mu and modelled as the logarithm of age-gender standardised expected incidence number (an offset variable) plus an intercept parameter plus a regression parameter **b** multiplied by explanatory variable $X^{(\mu)}$ plus an error term. Linear predictor on the second level is the logit of parameter p and modelled as the sum of an intercept parameter and a regression parameter **beta** multiplied by explanatory variable $X^{(\mu)}$ plus another error term.

Table 3. Parameter estimation in Poisson-Binomial model

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
a	-0.2447	0.0732	0.00513	-0.3875	-0.2453	-0.0944	751	2250
b	0.8454	0.1480	0.01031	0.5581	0.8474	1.1330	751	2250
alpha	0.2408	0.0610	0.00426	0.1158	0.2426	0.3559	751	2250
beta	0.0786	0.1255	0.00874	-0.1611	0.0774	0.3232	751	2250
deviance	1818.0	25.59	0.81890	1770.0	1817.0	1870.0	751	2250

One can infer from Figure 1 that the SIR values of disease (C13+C32) seem to be higher in the South-West Region than elsewhere. This visual impression is confirmed by Table 3. because coefficient **b** proved to be positive and statistically significant (its 0.95-level confidence interval is [0.558, 1.133] and the 0 value is outside of this interval). On the other hand the hypothesis **beta=0** must be accepted (because its 0.95-level confidence interval [-0.1611, 0.3232] contains the 0 value) and this fact is interpreted that ratios C32 SIR/(C13 + C32) SIR in the South-West Region do not differ significantly from that of other parts of country.

Appendix 3 treats with another pair of diseases: gastric ulcer (K25) and duodenal ulcer (K26). The specification of Poisson-Binomial model was extended by time factor as data refer to five years from 2004 to 2008. An explanatory variable "Social indicator of the economic underdevelopment status" was included in the model and its effect proved to be significant.

Discussion

Calculation of standardised incidence ratio (SIR) is a widely-used and efficient tool to compare risks between populations of different age-gender distribution and our maps follow this presentation technique. Results of statistical tests treated above were all consistent with the visual impressions of SIR maps. It is important to note that his is not always the case. It is enough to refer to the simple fact that C13 SIR/(C13 + C32) SIR + C32 SIR/(C13 + C32) SIR $\neq$ 1 to demonstrate the unavailability of linear models in analysing ratio of SIR values. This is because diseases C13 and C32 have different age-gender specific risks (as shown in Table 3.) so standardisation for C13, C32 and (C13+C32) means three completely different procedures.

Indicator explanatories were chosen for the sake of simplicity. South-West region was assigned arbitrarily: if $X^{(\mu)}$ and $X^{(P)}$ were the indicator variable of West Region (Nyugat-Dunántúl), then **b** would not be significant but **beta** would be significant. A continuous explanatory variable was treated in Appendix 3, but many other possible explanatory variables could have been considered here. Our main aim was to demonstrate how the Poisson-Binomial model can be used in an epidemiological research.

The medical relevance of our approach is twofold. On one hand it can be helpful to discover common etiological factors for different diseases. On the other hand it might reveal the level of uncertainty of ICD coding: the codes of closely related diseases are possibly mixed up.

Conclusion

Usual procedures based on SIR mapping are limited to investigate only one disease but cannot be used correctly when analysing the joint distribution of a disease pair. A multilevel Poisson-Binomial model is introduced here to perform significance tests for ratio of incidences of two diseases.

Conflict of Interest

The authors declare that they have no conflict of interest.

Authors' Contributions

KS designed and coded the statistical analyses, TG helped design the statistical analyses and he compiled the maps and SGy was counsellor in modelling epidemiological data and in interpretation results. All authors read and approved the final manuscript.

Acknowledgements

This paper is an extended version of a poster presentation held by the authors in Applied Biostatistics in Life Sciences, Joint Meeting of ISCB National Groups and IPA Cross-border Co-operation Program, Szeged, Hungary on 4th November, 2011.

This research was partially supported by TÁMOP 4.2.1/B-09/1/KMR-2010-003.

References

1. Brugere J, Guenel P, Leclerc A, Rodriguez J. Differential Effects of Tobacco and Alcohol in Cancer of the Larynx, Pharynx, and Mouth. Cancer 1986;57:391-395.
2. Nakai Y, Fukunaga M. Peptic Ulcer. Japan Medical Association Journal 2003;46(2):61-65.
3. Davidovic M, Svorcan P, Milanovic P, Antovic A, Milosevic D. Specifics of Helicobacter pylori Infection/NSAID Effects in the Elderly. Romanian Journal of Gastroenterology 2005;14(3):253-258.
4. Waller LA, Gotway CA. Applied Spatial Statistics for Public Health Data. New York: John Wiley; 2004.
5. Knorr-Held L, Best NG. A shared component model for detecting joint and selective clustering of two diseases. J R Statist Soc A 2001;164(1):73-85.
6. Barndorff-Nielsen 0. Information and Exponential Families in Statistical Theory. New York: John Wiley; 1978.
7. Agresti A. Categorical Data Analysis. New York: John Wiley; 1990.
8. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 2000;10:325-337.