RESEARCH ARTICLE

Comparison of Bayesian Spatial Ecological Regression Models for Investigating the Incidence of Breast Cancer in Iran, 2005-2008

Ahmad Haddad Khoshkar[1], Tohid Jafari Koshki[2], Behzad Mahaki[1]*

Abstract

    <u>Background</u>: Breast cancer is the most prevalent kind of cancer among women in Iran. Regarding the importance of cancer prevention and considerable variation of breast cancer incidence in different parts of the country, it is necessary to recognize regions with high incidence of breast cancer and evaluate the role of potential risk factors by use of advanced statistical models. The present study focussed on incidence of breast cancer in Iran at the province level and also explored the impact of some prominent covariates using Bayesian models. <u>Materials and Methods</u>: All patients diagnosed with breast cancer in Iran from 2005 to 2008 were included in the study. Smoking, fruit and vegetable intake, physical activity, obesity and the Human Development Index (HDI), measured at the province level, were considered as potential modulating factors. Gamma-Poisson, log normal and BYM models were used to estimate the relative risk of breast cancer in this ecological investigation with and without adjustment for the covariates. <u>Results</u>: The unadjusted BYM model had the best fit among applied models. Without adjustment, Isfahan, Yazd, and Tehran had the highest incidences and Sistan-Baluchestan and Chaharmahal-Bakhtiari had the lowest. With the adjusted model, Khorasan-Razavi, Lorestan and Hamedan had the highest and Ardebil and Kohgiluyeh-Boyerahmad the lowest incidences. A significantly direct association was found between breast cancer incidence and HDI. <u>Conclusions</u>: BYM model has better fit, because it contains parameters that allow including effects from neighbors. Since HDI is a significant variable, it is also recommended that HDI should be considered in future investigations. This study showed that Yazd, Isfahan and Tehran provinces feature the highest crude incidences of breast cancer.

Keywords: Disease mapping - breast cancer - gamma poisson model - lognormal model - BYM model

Introduction

    Breast cancer is the most prevalent kind of cancer and the common cause of death madeby cancers among women in western countries. Occurrence of breast cancer is different in various regions and this heterogeneous distribution suggests that its incidence may be affected by environmental factors (Fehm et al., 2013).

    The global burden of the cancer has been growingin developed countries. The reasons of this phenomenon include aging, growth of population, and also some behavioral factors such as smoking which lead to cancer (Bray et al., 2012).

    Cancer is the second cause of death after cardiovascular diseases in developed countries and the third one in developing countries. The number of these deaths is estimated to reach 13.1 million ones annually in 2030. Cancers are the cause of 45 to 50 percent of deaths among 45 to 64 year old women. This high rate, demonstrates significant increase of death caused by cancer in the last century.It has different reasons. Risk factors which have

related to the cancer, increased considerably in the last century include air pollution, smoking, diet changes, insufficient physical activity, obesity, stress and so on. It is possible to control 40 percent of the deaths caused by cancer if the risk factors are recognized and managed properly. So, recognizing these factors, in every region, is necessary to control and prevent cancer (Kerenyi et al., 1990; Akbari et al., 2008).

    Breast cancer is the most prevalent kind of cancer among women. Its standardized incidence rates are 86.4, 27.3, 38.9, 26, 32.6 and 28.25 in developed countries, developing countries and the entire world, Asia, eastern Asia and Iran, respectively.This cancer is the most common type of cancer in all provinces except Ardebil and Kohgilueh-Boyerahmad. It is the third prevalent cancer after esophagus and stomach cancers in Ardebil and is the second common cancer after skin cancer in Kohgilueh-Boyerahmad (Akbari et al., 2008).

    The rate of cancer is predicted to be increased tremendously in future due to life expectancy growth and tendency to follow western lifestyle (RahmatiYamchi et

*Department of Biostatistics, School of Public Health, Isfahan University of Medical Sciences, [1]Isfahan, [2]Sabzevar, Iran *For correspondence: Behzad.mahaki@gmail.com*

al., 2013).

According to the results of many conducted investigations, smoking, fruit and vegetable intake, physical activity, over weightiness and socio-economic status are the influencing and measurable factors which affect the incidence of cancer (Snowdon et al., 1984; Kahn et al., 1998; Yost et al., 2001; Bakhtiari et al., 2004; HolakoueeNaeeni et al., 2006; Kruk, 2007; Shahar et al., 2010; Saki et al., 2011; Sharif Zadeh et al., 2011; Arkani et al., 2012; Abbastabar et al., 2013; RahmatiYamchi et al., 2013; Tehrani et al., 2013; VahedianShahroodi et al., 2013; Kruk, 2014).

A lot of studies have been conducted about breast cancer and its risk factors (Yost et al., 2001; Bakhtiari et al., 2004; Kruk, 2007; 2014; Shahar et al., 2010; Mahaki et al., 2011; Saki et al., 2011; Sharif Zadeh et al., 2011; Abbastabar et al., 2013; RahmatiYamchi et al., 2013; VahedianShahroodi et al., 2013). Most of them have been at individual level and those with ecological approach (Jafari-Koshki et al., 2014) have not evaluated risk factors related to the incidences.

Regarding to ageing process of Iranian population and increase in development of the cancer and its considerable geographical variation in different parts of country, it is necessary to recognize the regions with high incidence and the role of its most significant risk factors to improve prevention process by use of advanced statistical models.

The present study investigates the incidence of breast cancer in Iran at province level and also explores the impact of some covariates including smoking, over-weightiness or obesity, physical activity, fruit and vegetable intakes, and human development index by use of Bayesian models.

Materials and Methods

All registered patients suffering from breast cancer in Iran from 2005 to 2008 were included in this study. These data were extracted from the report by Non-communicable Diseases Management Center of Iranian Ministry of Health and Medical Education which are published annually.

Also the data of four high risks including smoking, fruit and vegetable intake, physical activity and obesitywere extracted from annual reports of Non-communicable Disease Risk Factors Surveillance System of the Ministry of Health. Smoking was considered as the multiplication of the percent of smokers and the mean number of cigarettes smoked daily in each province. Over-weightiness or obesity was the population proportion in each province with body mass index more than 25 (BMI>25). The fruit and vegetable variable was illustrated as the sum of daily consumption of fruit and vegetables in each province. Physical activity was calculated using a combined index called Metabolic Equivalent (MET). Human Development Index is geometric mean of achieved success in every province in three main dimensions of human development including long and healthy life, having access to knowledge and normal life standards and extracted from annual reports of Central Bank. Its range is from zero to one (Akbari et al., 2008). Ecological analysis is an epidemiologically discovering method and investigates the relation between diseases prevalence and risk factors which are measured in groups instead of individuals. The simplest risk estimation method is usingstandardized mortality rate (SMR) by dividing observed (Y_i) by expected frequencies (E_i). This method isresulted from Maximum Likelihood Estimation. However, this estimator has two shortcomings; firstly, SMR does not consider spatial correlation and, secondly, it gives over dispersed estimates.

Ecological models

The following models were used to study the incidence and also to assess association between incidence and the covariates. In these models, $\{Y_i, i=1,\ldots,n\}$ and $\{E_i, i=1,\ldots,n\}$ represent the number of observed and expected breast cancer cases forprovince i respectively. It is assumed that Y_i has Poisson distribution with the rate of $\mu_i=\theta_i E_i$ where θ_i represents relative risk for province i. In the following models, $\sum_{h=1}^{H} \beta_h x_{ih}$ represents linear combination of risk factors $x_1 x_2,\ldots, x_H$ with corresponding coefficient β_h. In all following equations, α is the overall level of relative risk.

-Gamma-poisson regression model

It regards the relation between incidence and risk factors in every province via following log-linear model;

$$\log \theta_i = \alpha + \sum_{h=1}^{H} \beta_h x_{ih}$$

The most important problem of this model is disregarding spatial correlation. Variability cannot be illustrated by ecological variables in many cases. Moreover, in this model, variation becomes more than the expected variability of Poisson model. Since over dispersion happens in this situation, it is added to model. In fact, the rate of over dispersion depends on the score of heterogeneity found among relative risks. In Bayesian structure, it is suggested for this model that θ_i has Gamma distribution. So, the achieved posterior distribution will be Gamma (Mahaki et al., 2011; Jafari-Koshki et al., 2014).

-Lognormal model

Although Gamma prior distribution seems suitable for risk rate mathematically, Gamma-Poisson model has some restrictions. Because adjusting suitability of the independent variable is difficult and involving the spatial correlation between the rates of regions is impossible. Lognormal model is more flexible for relative risk formulation and is defined as followed (Asmarian et al., 2013).

$$\log \theta_i = \alpha + v_i + \sum_{h=1}^{H} \beta_i x_{ih}$$
$$v_i \sim N(0, \tau_v^2)$$

-BYM model

This model considers two sources of changes for justifying the heterogeneity the rate of incidence in every region in addition to independent variables. It also models the rate of incidence through following equation.

$$\log \theta_i = \alpha + u_i + v_i + \sum_{h=1}^{H} \beta_i x_{ih}$$

$$[u_i | u_j, i \neq j, \tau_v^2] \sim N(\bar{u}_1, \tau_1^2)$$

$$\bar{u}_1 = \frac{1}{\sum_j \omega_{ij}} \Sigma u_i \omega_{ij}, \quad \tau_1^2 = \frac{\tau_u^2}{\sum_j \omega_{ij}}$$

$$\omega_{ij} = 1 \ \textit{Ifi and j are adjacent}$$

$$\omega_{ij} = 0 \ \textit{If i andj are not adjacent}$$

Where u_i and v_i represent non-structural and structural heterogeneity respectively. They are also supposed to have normal distribution and conditional auto regressive normal distribution. Regarding conditional auto regressive, the rate of incidence in every region was supposed to be dependent on the incidence rate in every neighbor regions (Lawson et al., 1999; Lawson et al., 2003).

These models were fitted using OpenBUGS version 3.2.1. Convergence was checked by using Brooks-Gelman-Robin plots. Significance tests for parameters were done by the use of Bayesian credible intervals (CrI) which are equivalent to p-value.

Results

The means of risk factors including smoking, over weightiness and obesity, physical activity, fruit and vegetable intake and HDI are, respectively, 124.6365, 43.4193, 32.8371, 2.53425 and 0.7357. Their variances are also, in turn, 950.5683, 32.9568, 47.1048, 0.2785 and 0.0024.

The total number of known breast cancer cases has been 25152 in Iran from 2005 to 2008. The most cases were found in Isfahan (2214 cases) and Khorasan-Razavi (2218 cases). The least incidence was seen in Kohgiluye-Boyerahmad (67 cases).

According to the results shown in Table 1, the role of all risk factors was found to be significant in Poisson-Gamma model. This model disregards the possibility of correlation among provinces. So, this model may be disguising.

Regarding non-structural heterogeneity in lognormal model, the impact of variables was adjusted and risk factors including over weightiness and Human Development Index were significant. This shows that over weightiness and HDI cause to increase the rate of cancer incidence.

Just HDI as a descriptive variable was significant in BYM model. It shows that the risk of breast cancer incidence is more in the regions having higher HDI.

Figure 1A shows the province's relative risks regarding non-structural heterogeneity and without adjusting of risk factors in Gamma-Poisson model. According to this map, Yazd and Tehran province have the highest incidence risk of breast cancer and Sistan-Baluchestan province has the lowest incidence risk. Generally, central provinces have a higher risk of breast cancer incidence. Figure 1B shows the relative risk of provinces according to non-structural heterogeneity without adjusting the impact of risk factors in Lognormal model. These maps show that northwestern and southeastern provinces have a lower incidence risk of breast cancer among which Sistan-Baluchestan and Kohgiluyah-Boyerahmad have the least incidence of breast cancer.

Figure 1C shows the relative risk of the provinces regarding non-structural heterogeneity with adjusting risk factors in Lognormal model. This map illustrates that Khorasan-Razavi and Hamedan provinces have the highest incidence risk of breast cancer and Kohgiluyah-Boyerahmad, Sistan-Baluchestan, and Ardebil provinces have the lowest incidence risk of breast cancer.

Figure 1D shows the province's relative risk in BYM model without adjusting for risk factors and taking structural and non-structural heterogeneity into account. According to this map, the most incidence risk of breast cancer was seen in Isfahan and Tehran provinces and the
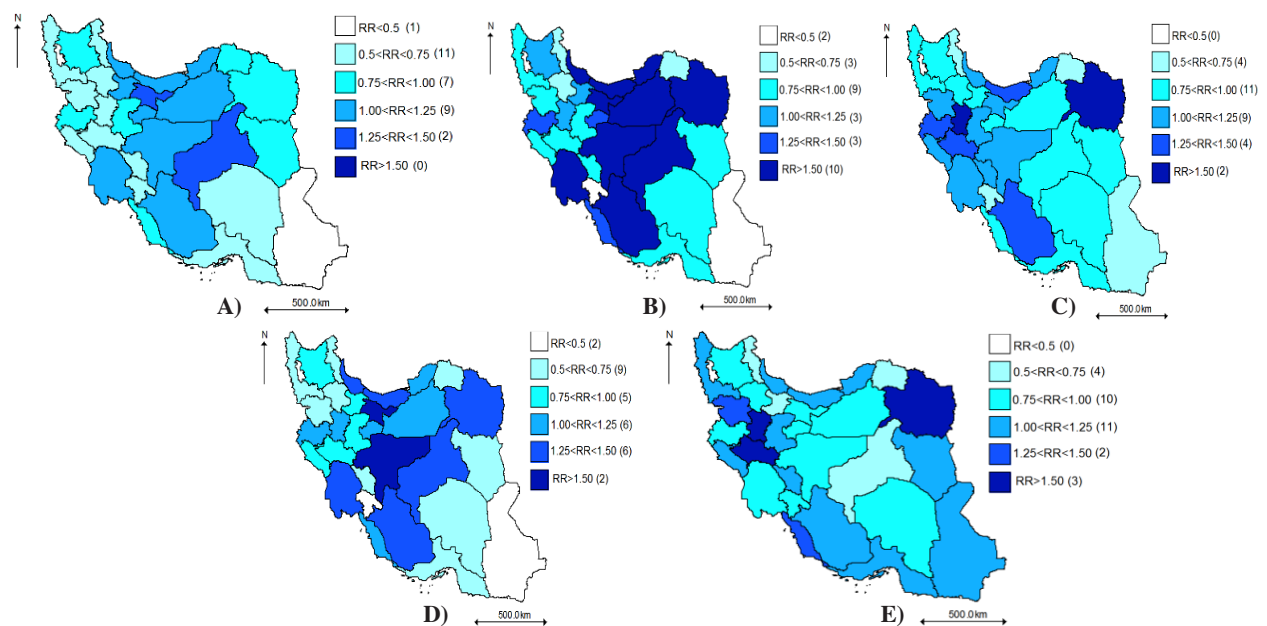

Figure 1. Maps of Incidence Risk of Breast Cancer A) Gamma-Poisson without Adjusting Risk Factor Adjustment B) Lognormal without Adjusting Risk Factor Adjustment C) Lognormal with Adjusting Risk Factor Adjustment D) BYM without Adjusting risk Factor Adjustment E) BYM with Adjusting Risk Factor Adjustment

Table 1. Estimation of Coefficients of Relative Risks Effects for Ecological Regression Models

Risk factors	Poisson-Gamma		Lognormal		BYM	
	coefficient	95%CrI	coefficient	95%CrI	coefficient	95%CrI
Smoking	-0.002	(-0.0025,-0.001)	-0.002	(-0.005,0.002)	-0.002	(-0.006,0.002)
over weightiness	-0.018	(-0.0223,-0.014)	0.022	(0.001,0.042)	0.021	(-0.002,0.043)
physical activity	-0.005	(-0.008,-0.001)	0.004	(-0.110,0.019)	0.004	(-0.011,0.020)
fruit &vegetable intakes	-0.098	(-0.148,-0.049)	-0.022	(-0. 231,0.184)	-0.025	(-0.226,0.181)
HDI	7.426	(6.970,9.700)	6.122	(3.908,8.418)	6.172	(3.794,8.527)
DIC for models with covariates	888.6		298.2		888.6	
DIC for models without covariates	-		298.1		171.4	

minimum incidence was found in Sistan-Baluchestan and Kohgiluyah-Boyerahmad provinces.

Figure 1E shows the relative risk adjusted for risk factors in BYM model that consider structural and non-structural heterogeneity. According to this map, Yazd, Ghazvin, Ardebil and Khorasan-Shomalihad the least incidence risk of breast cancer and Khorasan-Razavi, Lorestan and Hamedan had the maximum of incidence risk.

Table 1 also compares the goodness of fitness of Gamma-Poisson, Lognormal and BYM models with and without considering risk factors by using of DIC. BYM has best goodness of fitness without the presence of risk factors, because it contains structural and non-structural heterogeneity. Gamma-Poisson model without presence of risk factors has the worst fit. Regarding the presence of risk factors, BYM and Lognormalmodels have rather the same fit. By considering risk factors, no change is made in the fit of lognormal model. Gamma-Poisson model has poor fit because it disregards the spatial correlation among the provinces. Therefore, it seems better to use BYM model in ecological analysis in comparison to other ecological regression models.

Discussion

BYM model without adjusting the impact of risk factors showed that central provinces including Isfahan, Yazd and Tehran have the most incidences of breast cancer and Northernprovinces including Fars, Khuzestan and Khorasan-Shomali provinces have the highest incidence after them. North-eastern and south-western provinces have the minimum rates of incidence risk among all provinces. These results are in according to the results reported by Jafari et al. (2014) and Mahaki et al. (2011) studies. However in Jafari et al results were crude and were not adjusted for risk factors. With adjusting risk factors, Khorasan-Razavi, Lorestan and Hamedan have the most incidence of breast cancer and Ardebil, Kohgiluyah-Boyerahmad have the least incidence. In general, south-western provinces have lower incidence of breast cancer. The results show that the incidence of breast cancer is higher in provinces with higher HDI.

The comparison of maps B and D, C and E shows that in spite of finding different amounts of relative risks, ranks of relative risks in all provinces are identical. I.e. provinces with higher relative risk in map B and D have higher relative risk in map C and D. This is in accordance with the result of Clayton and Kaldor study (Clayton et al., 1987).

As shown in Table 1, direct impact of HDI in breast cancer incidence is affirmed. It is compatible with the results of Yost et al study (Yost et al., 2001). Barrios et al found positive relation between HDI and breast cancer incidence as well. In their study, the correlation coefficient between HDI and breast cancer incidence was 0.68 (Barrios et al., 2013).

It could be due to air pollution, more exposure to carcinogens in urban regions or because of diagnosing more cases of cancer in provinces with higher HDI, promotion of western life style and pollution increase accompanied by HDI development. HDI increase can cause to develop diagnosis of people with cancer. Although Abbastabar et al study (Abbastabar et al., 2013) showed the significant relation between breast cancer and fruit and vegetable intake, this significant relation was not found in the present study. This difference could be due to the conducting of the present investigation in ecological (province) level, whereas their study was done in individual level. The results of this study should not be interpreted in individual level because this may be misleading due to a phenomenon known as ecological fallacy. Since the urban lifestyle develops increasingly and direct association has been proved between HDI and incidence rate of breast cancer, it is necessaryto conduct some measures to provide awareness for the people and improve their lifestyle for preventing from growing the incidence of breast cancer.

Air pollution, family history, neonate feeding situation and other covariates were not available at province level. So, we suggest conducting further ecological researches regarding these factors as well.

References

Abbastabar H, SoleimaniDodran M, HamidiFard P, et al (2013). The relationship between breast cancer and the most common non-contagious disease risk factors: an ecologic study. *J Health Syst Res*, **8**, 369-74.

Akbari M, Abachizadeh K, Khayamzadeh M, et al (2008). Iran cancer report. cancer research center shahidbeheshti university of medical sciences Tehran, Qom: Darolfekr.

Arkani M, Safaii A, Karimi K, et al (2012). Association of the insulin gene polymorphism and colorectal cancer. *Koomesh J*, **13**, 172-6.

Asmarian NS, Kavousi A, Salehi M, Mahaki B (2013). Comparison of point poisson kriging and empirical bayesian methods in mapping of gastrointestinal cancer incidence rate in Iran. *J Health Syst Res*, **9**, 277-85.

Bakhtiari A, Hajian K, Faramarzi M (2004). Examine the

relationship between demographic, social and lifestyle in women with breast cancer. *Babol Univers J Med Sci*, **9**, 55-62.

Barrios C, Werutsky G, Martinez-Mesa J (2013). Abstract P1-09-06: breast cancer in the pan-American region: inequities in incidence and mortality rates according to the human development index. *Cancer Research*, 73.

Bray F, Jemal A, Grey N, Ferlay J, Forman D (2012). Global cancer transitions according to the human development index (2008-2030): a population-based study. *Lancet Oncol J*, **13**, 790-801.

Clayton D, Kaldor J (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics J*, **43**, 671-81.

Fehm T, Ruckhäberle E, Lammert E, Zeeb M (2013). Introduction to breast cancer. *Principles Metabolism Health Diseases J*, 379-84.

HolakoueeNaeeni K, Ardalan A, Mahmoudi M, MotovallianA, Yahia Poor Y (2006). Study on risk factors for breast cancer in Mazandaran in 1383. *Public Health Inst Health Res J*, **4**, 27-36.

Jafari-Koshki T, Schmid VJ, Mahaki B (2014). Trends of breast cancer incidence in Iran during 2004-2008: A Bayesian space-time model. *Asian Pac J Cancer Prev*, **15**, 1557-61.

Kahn HS, Tatham LM, Patel AV, Thun MJ, Heath Jr CW (1998). Increased cancer mortality following a history of nonmelanoma skin cancer. *Jama J*, **280**, 910-2.

Kerenyi NA, Pandula E, Feuer G (1990). Why the incidence of cancer is increasing: the role of 'light pollution'. *Medical Hypotheses J*, **33**, 75-8.

Kruk J (2007). Association of lifestyle and other risk factors with breast cancer according to menopausal status: a case-control study in the Region of Western Pomerania (Poland). *Asian Pac J Cancer Prev*, **8**, 513-24.

Kruk J (2014). Lifestyle components and primary breast cancer prevention. *Asian Pac J Cancer Prev*, **15**, 10543-5.

Lawson A, Biggeri A, Bohning D, et al (1999). Disease mapping and risk assessment for public health: John Wiley & Sons.

Lawson AB, Browne WJ, Rodeiro CLV (2003). Disease mapping with WinBUGS andMLwiN: John Wiley & Sons;

Mahaki B, Mehrabi Y, Kavousi A, et al (2011). Multivariate disease mapping of seven prevalent cancers in Iran using a shared component model. *Asian Pac J Cancer Prev*, **12**, 2353-8.

Mahaki B, Mehrabi Y, Kavousi A, Mohammadian M, Khodakarim S (2013). Applying gamma-poisson, lognormal, and BYM Models in comparing relative risk of suicide among provinces of Ilam, Iran. *J Health Syst Res*, **9**, 86-95.

RahmatiYamchi M, Zarghami N, RahbaniNoubar M, Najafipour R, Mobsser M (2013). Correlation between telomerase gene expression and different stages of breast cancer and obesity. **2**, 35-43.

Saki A, Hajianzadeh E, Tehranian N (2011). Evaluation of breast cancer risk factors using analysis of tree models. *OfoghDanesh J*, **17**, 60-7

Shahar S, Salleh RM, Ghazali AR, Koon PB, Mohamud W (2010). Roles of adiposity, lifetime physical activity and serum adiponectin in occurrence of breast cancer among Malaysian women in Klang Valley. *Asian Pac J Cancer Prev*, **11**, 61-6.

Sharif Zadeh G, Hosseni M, Kermani T, Atai M, Akhbari H (2011). Factors associated with breast cancer-a case-control study. *Birjand Univers J Medical Sci*, **18**, 191-9.

Snowdon DA, PHILLIPS RL, Choi W (1984). Diet, obesity, and risk of fatal prostate cancer. *Am J epidemiol*, **120**, 244-50.

Tehrani M, MohseniS (2013). Case finding and clinical study of obese patients. *Iranian J Diabetes Metabolism*, **12**, 393-8.

VahedianShahroodi M, Poorhaji F, Ismaili H, Poorhaji F (2013). The role of physical activity in the prevention of breast cancer. proceedings of the eighth student conference of medical universities in East.1.

Yost K, Perkins C, Cohen R, Morris C, Wright W (2001). Socioeconomic status and breast cancer incidence in California for different race/ethnic groups. *Cancer Causes Control J*, **12**, 703-11.