Exploring Neighborhood Inequality in Female Breast Cancer Incidence in Tehran using Bayesian spatial models and Spatial Scan Statistic

Running title: breast cancer incidence mapping with Bayesian spatial methods

Erfan Ayubi¹, Mohammad Ali Mansournia¹, Ali Motlagh², Alireza Mosavi-Jarrahí³, Ali Hosseini⁴, Kamran Yazdani¹*

1- Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
2- Department of Radiotherapy, Shahid Baheshti University of Medical Sciences, Tehran Iran
3- Department of Health and Community Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
4- Department of Geography and Urban Planning, University of Tehran, Tehran, Iran

**Funding**: This work derived from Erfan Ayubi’s PhD thesis and was supported by Tehran University of Medical Sciences

* **Corresponding author**: Kamran Yazdani, (MD, PhD)

Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

E-Mail: kyazdani48@yahoo.com
Exploring Neighborhood Inequality in Female Breast Cancer Incidence in Tehran using Bayesian spatial models and Spatial Scan Statistic

Abstract:

Objectives. To explore the spatial pattern of female breast cancer (BC) incidence at neighborhood level in Tehran.

Methods: Present study included registered incident cases of female BC during March 2008 to March 2011. The raw standardized incidence ratio (SIRs) of BC for each neighborhoods were estimated by comparing observed cases relative to expected cases. The estimated raw SIRs were smoothed by Besag, York and Mollie (BYM) spatial model and the spatial empirical Bayes method. The purely spatial scan statistic was used to identify spatial clusters.

Results: There were 4,175 incident BC cases in the study area from 2008 to 2011 and of them 3,080 successfully geocoded to neighborhood location. The higher than expected rates of BC were found in neighborhoods located in northern and central of Tehran, whereas those with lower rates were appeared in southern areas. The most likely cluster of higher than expected incidence BC involved neighborhoods in district of 3 and 6 with observed to expected ratio of 3.92 (p<0.001) whereas, the most likely cluster of lower than expected involved neighborhoods in districts of 17, 18 and 19 with observed to expected ratio of 0.05 (p-value<0.001).

Conclusions: Neighborhood inequality in incidence of BC exists in Tehran. The findings can be a basis for resources allocation and preventive strategies for at risk areas.

Key words: Breast Neoplasms, Spatial Analysis, Geographical disparity, Tehran
Introduction

Statistics showed that incidence rate of breast cancer (BC) is 24 per 100,000 in women in Iran [1]. The number of newly annually BC cases is projected from 5,000 in 2000 to 15,000 in 2030 [2]. BC have poor prognosis in Iran so that, it is the 3rd leading cause of death from cancers, accounting for 16% of cancer deaths [3]. It found that majority of BC patients are diagnosed at advanced stages in Iran [3]. This is probably the reason why BC have a poor prognosis.

The incidence and mortality of BC have been attributed to many individual level risk factors [4-9]. Regardless of these risk factors, it found that incidence and mortality of BC can be associated with place and area based risk factors [9, 10]. While many studies in other countries consider spatial pattern of BC at census tract or zip code level [9, 11, 12], spatial pattern in incidence of BC have been studied only at province level in Iran [13-15]. For example, overall incidence of BC in population living in Tehran province was 31.5 per 100,000 which it was more than other provinces [16]. Therefore, studies should focus on identifying spatial pattern of BC incidence in finer geographic scales to understand health needs and health care allocation [9-11].

When studying spatial pattern of disease in finer geographic scale, however, there are some challenges that must be considered. The estimated rates and observed associations can be along with degree of bias due to spatial autocorrelation, population size heterogeneity and small area effect [17]. Two methods of empirical spatial bayes smoothing and Besag, York and Mollie (BYM) spatial model are used to offset these challenges with considering spatial autocorrelation and spatial heterogeneity among geographic units [11, 18, 19].
With considering these issues in mind, our objectives in this study were (1) to estimate the smoothed standardized incidence ratio (SIR), (2) to identify the clusters with higher or lower than expected incidence of female BC in Tehran.

Materials and methods

Study area

A retrospective study design was used in Tehran, Capital of Iran. This city has 22 districts. The geographical unit of the study was 374 neighborhoods in Tehran city.

Data sources

The information of incident female BC of Tehran during March 2008 to March 2011 were obtained from Iran’s cancer registry of ministry of health. Street of residence of patients was geocoded to neighborhood location. Population of women aged 15 and over for each neighborhoods was obtained from national census 2006 and 2011.

Statistical analysis

Raw standardized incidence ratio (SIR)

The number of the observed events in each neighborhood follow a Poisson distribution

\[ O_i \sim \text{Poisson}(E_i \theta_i) \]  

Where \( O_i, E_i \) and \( \theta_i \) are the observed, number, the expected number and relative risk for neighborhood \( i \) respectively. Expected events is calculated as follows:
Where \( n_i \) is number of women aged 15 and over in neighborhood \( i \).

The standardized incidence ratio (SIR) can be calculated by the observed to expected ratio.

The raw SIRs per neighborhoods were expected to be dispersed due to extra-Poisson variability or over-dispersion. To offset this challenge, the raw SIRs can be smoothed by Besag, York and Mollie (BYM) spatial model and the spatial empirical Bayes (SEB) method.

BYM model

Overdispersion or extra-Poisson variability is a challenges when Poisson model to be applied for the count data in the spatial analysis. Ignoring the extra-Poisson variation in the data set in the analysis can lead to incorrect results. Overdispersion occurs in the presence of spatial autocorrelation in the residual values. The concept of spatial autocorrelation declare that due to spatial components the local estimates of disease risk for the neighbor areas are assumed to be correlated. Effect of overdispersion due to spatial autocorrelation on the results is strong in the small area problem [18]. This problem exists in the data set when the population size or number of the events of areas is low.

In the spatial analysis framework, the model often would not be fully specified in presence of unmeasured spatial and non-spatial factors. To offset this the aforementioned challenge, Hierarchical Models such the Besag-York-Mollie (BYM) model is introduced [18].
In the BYM model, unmeasured spatial factors are controlled using suitable random effects.

With considering equation 1:

\[ \log(\theta_i) = \alpha + u_i + v_i \]  

where $\alpha$ is a log-relative risk baseline, $v_i$ and $u_i$ indicates random effects components regarding to spatial and non-spatial factors.

Spatial autocorrelation across neighborhoods ($v_i$) induced by the Conditional Autoregressive (CAR) model. The CAR represent risk factors with spatial structures so that specific risk estimate of given area will tend to shrunk toward a local mean. CAR model in the BYM model is as follow:

\[ [v_i, v_j, i \neq j, \tau_v^2] \sim N(\bar{v}_i, \tau_v^2) \]

Where

\[ \tau_v^2 = \frac{\tau_v^2}{\sum w_{ij}} \quad \text{and} \quad \bar{v}_i = \frac{1}{\sum w_{ij}} \sum_j v_j w_{ij} \]

If area $i$ and $j$ are neighbor of each other, the weight is equal 1 and otherwise is 0.

The random effect of $u_i$ represent risk factors with non-spatially structure so that that specific risk estimate of given area will tend to shrunk toward a global mean of the study area. This component in BYM model is as follow:

\[ u_i \sim N(0, \tau_u^2) \]

Two parameters of $\tau_v = \frac{1}{\sigma_v^2}$ and $\tau_u = \frac{1}{\sigma_u^2}$ are two precision parameters of two aforementioned random effects. The proper distribution for $\tau_v$ and $\tau_u$ is the Gamma distribution $G(a, b)$ with...
expected value $\frac{a}{b}$ and variance $\frac{a}{b^2}$. In this study based on previous studies to select suitable Gamma distribution [17, 18, 20], we used $a_\psi = 0.5$ and $b_\psi = 0.005$ for spatially structured random effect and $a_u = 0.5$ and $b_u = 0.005$ for non-spatially structured random effect.

We implement Markov Chain Monte Carlo (MCMC) simulation for estimating the model parameters in the BYM model. The Gibbs sampler as a specific MCMC was used to produce random samples through parameter space. The convergence of the model was evaluated by Brooks–Gelman–Rubin (BGR) statistics [18]. This statistic evaluates MCMC convergence by comparing the within-chain variance to the between-chain variance and the values close to 1 indicate the degree of convergence [18]. We run MCMC model with 100,000 iterations and the first 5000 iterations ignored as burn-in. Iterations started from over-dispersed initial values on four parallel chains. OpenBUGS version 3.2.3 is used to implement the BYM model.

Spatial Empirical Bayes (SEB) methods

Another available method for correcting bias in raw estimates of SIR is Spatial Empirical Bayes (SEB) methods. SEB method has been proposed that the rate of neighborhoods in areas without clear spatial patterns and those in areas with obvious spatial patterns are shrunk toward the global mean and local mean of the study area respectively [21]. In this method, posterior of $\theta_i$ does depend on the data $O_i$ and $E_i$ from the other regions $j \neq i$. In other words, parameters of prior distribution are not fixed and will estimated empirically and based on all available data. Smoothing raw SIRs with empirical bayes methods was done using 2nd order queen weights in GeoDa.

Detection and identification of BC clusters
Neighborhood variation in incidence of BC (regardless of staging), early and late stage at diagnosis were determined by the purely spatial scan statistic, discrete Poisson model, using SaTScan software (v9.4.2). The analysis requires number of cases, population counts and the geographical coordinates (longitude and latitude) for each locations. The standard purely spatial scan statistic imposes a circular window (spatial cluster) on the map and it move across the study area to compare the number of disease cases in a geographic area (θin) with disease cases outside that area (θout). Since the results of this analysis can be sensitive to model parameters, particularly window size, the maximum spatial cluster size is defined using Gini coefficient [22]. It argued that Gini coefficient is more intuitive and systematic way to identify the best collection and non-overlapping of clusters [22].

The number of cases in each location is Poisson-distributed for cancer incidence, so we applied the exponential model based spatial scan statistic using SaTScan software, a Poisson model is a typical model. The likelihood ratio statistic (LRS) of Poisson distribution (under test hypothesis; Ho: θin =θout; Ha: θin ≠ θout) for a specific window is proportional to 1:

\[
\frac{c^c}{E[c]^c} \left( \frac{C - c}{C - E[c]} \right)^{c-c}
\]

Where C is the total number of BC cases, c is the observed number of BC cases within window and E[c] is crude expected number of cases within the window under the null hypothesis, C-E[c] is expected number of cases outside the window.

Statistical significance of detected clusters is evaluated using randomization testing or Monte Carlo Hypothesis testing because the exact distribution of LRS is unknown. Under null hypothesis a large number of random dataset are generated and the LRS value for each of
random dataset is computed. The Monte Carlo p-value of a window is computed as \( \frac{R_{\text{best}} + 1}{R + 1} \), where \( R_{\text{best}} \) is the number of random dataset which its LRS is higher than LRS under real dataset and \( R \) is total number of random dataset. A window would be statistical significance at \( \alpha = 0.05 \) when its LRS is higher than approximately 95% of LRS of random dataset. The windows with highest statistical significance likelihood ratio would be defined as most likely, secondary and tertiary cluster respectively. P-value of < 0.05 using 999 permutations was considered as statistically significant of Moran’s Index and spatial clusters. Sufficient statistical power was warranted by with 999 replication in Monte Carlo simulation. All cartographic manipulations and displays were done in ArcGIS 10.3.

**Results**

There were 4,175 incident BC cases in the study area from 2008 to 2011 and of them 3,080 successfully geocoded to neighborhood location. The number of BC cases ranged from 0 to 86 across neighborhoods in Tehran. The most female BC was found in northern part of Tehran (Fig 1). According to Moran’s Index, the null hypothesis of zero spatial autocorrelation is rejected for all the studied neighborhood based characteristics (Moran’s Index: 0.08, p-value<0.05).

**Spatial distribution of BC**

The results of three methods of raw SIR, BYM model and SEB method indicate a neighborhood inequality in incidence of female BC in Tehran. The neighborhoods with higher than expected incidence of BC were found in neighborhoods which located in districts of 1, 2, 3, 5, 6 and 7, where in northern and central part of Tehran. In other hand, neighborhoods with lower than
expected rates were located in districts of 15, 15, 17, 18, 19, 20, 21 and 22, where in southern and southwest of Tehran (Fig 2 to 4).

Fig 2 display the estimated raw SIR of female BC in Tehran during 2008-2011. The median (IQR) of female BC based on raw SIR was 0.52 (1.33). The estimated raw SIR ranged from 0 to 14.84. In 82 neighborhoods the value of raw SIR was 0 because no BC cases found in these neighborhoods and in addition, 37% of the neighborhoods had SIR higher than 1. The results of smoothed SIR using SEB method illustrated in Fig 3. The median (IQR) of female BC based on SEB method was 0.50 (1.13). As expected there was degree of shrinkage in the estimated SIR so that in 2 neighborhoods the value of SIR was 0 and range of SIR has been narrowed. The median (IQR) of female BC based on BYM model was 0.60 (1.14), none neighborhood had SIR of 0 and 30% of the neighborhoods had SIR higher than 1 (Fig 4).

Spatial clusters of BC

Table 1 describe characteristics of most likely clusters of BC. Fig 5 display geographic pattern of most likely clusters of BC. There was a statistical dispersion in detected clusters of female BC incidence (Gini index=0.47). The clusters with higher than expected incidence of BC were appears in northern, northern east and near city centers in the study area. Lower than expected incidence clusters of BC were appears in southern part of the study area. The most likely cluster of higher than expected incidence BC was located in areas near to center of Tehran, including neighborhoods in districts of 3 and 6 with observed to expected ratio of 3.92 (p<0.001), implying that the incidence of BC was 3.92 times higher within this cluster compared to the rest of the study area. The most likely cluster of lower than expected incidence BC included the
neighborhoods in districts of 17, 18 and 19 with observed to expected ratio of 0.05 (p-value<0.001), implying that the incidence of BC was 0.95 lower within this cluster compared to the rest of the study area.
Discussion

A neighborhood inequality in female BC incidence was found in Tehran during 2008 to 2011. The most likely clusters of higher than expected BC were located in central, northern and northeast of Tehran whereas, the most likely clusters of lower than expected incidence were located in southern part of Tehran.

Spatial analysis at finer scales can provide useful information about actually at risk areas. Our results showed that the smoothed rates of BC incidence are variably distributed within a specific district, therefore performing spatial study at the districts level fail to identify these inequality within district. For example, on average, the neighborhoods in district 19 had lower rates of BC, however, there are a few neighborhoods in this district with high rates of BC. In return, the neighborhoods in districts 1, 2 and 3 had higher rates of BC, but there are a few neighborhoods with low rates within these districts.

While spatial analysis of cancer measures at geographic resolutions such as census tract and ZIP codes has been well studied in developed countries [11, 12, 23, 24], it has not received well attention in developing countries. In Iran, spatial analysis of cancer measures using GIS and SaTScan mainly conducted at provinces or counties [13, 25, 26] so that, evidences about spatial pattern of cancer measures in finer resolutions such as neighborhood level are limited. In one study by Rohani et al. it has been identified that population living in the districts of 1, 2, 3 and 6 had highest age specific rate (ASRs) for BC incidence [27].

The results of smoothed rates and spatial analysis with SaTScan spatial scan statistic showed that incidence of BC is health problem in areas near center and northern parts of Tehran, which, corresponding with wealthy areas and higher degrees of educational attainment and more
expenditure on health care activities \[28\]. It has been identified that women living in wealthy areas spent more expenditures on health care activities such as screening tools and BC would be more diagnosed \[29\]. In addition to this factor, they have better access to cancer treatment facilities and adjuvant therapies and likely have better survival \[30\].

Several methodological issues about spatial analysis with SaTScan should be mentioned. It argued that the hierarchical approach (SaTScan Default) for selecting maximum size of clusters may lead to unnecessarily large and less informative clusters \[22\]. In current study, the maximum size for spatial cluster was based on Gini coefficient. It suggested that the Gini coefficient provides a greater degree of information about non-overlapping clusters with avoiding overly large clusters with relatively small relative risk (RR) and smaller clusters with higher RR \[22\]. SaTScan scan statistics allows a better understanding of spatial patterns with adjustment for covariates. Previous published studies have demonstrated that adjustment for area based characteristics such as census tract poverty and individual characteristics of patients including age, race/ethnicity or stage at diagnosis can change observed pattern of clusters \[12, 24\].

Our analysis has some advantages and limitations. Main advantage of the present study is that to our knowledge, this is the first study to explore the spatial patterns of female BC at diagnosis at the neighborhood level in Iran. This type of spatial analysis at the neighborhood level can provide useful information to policy makers for allocation resources to truly needy areas. As expected the raw SIRs per neighborhoods to be dispersed due to extra Poisson variability to offset this challenge, we smoothed the raw SIRs by BYM spatial model and the spatial empirical Bayes (SEB) method. Main objective of BYM model is that it can take into account spatial
autocorrelation in an efficient way but the ability of BYM model is limited where geographical units are different in sizes and shapes [31].

One of these limitations of the study is related to how missing data induce bias in our results. Missing in surveillance data such as cancer registries is inevitable and it influenced by many factors such as gender, age, SES and etc. As expected in ecological studies and spatial analysis, the ecological fallacy and the modifiable areal unit problem (MAUP) are potential sources of misleading interpretations. Other problem that not accounted in this study is a phenomenon named edge effect. It means results for neighborhoods near the administrative borders must be interpreted with caution because for example, the socioeconomic indicators of neighborhoods outside of the studied region may affect the characteristics of residents near the borders. Finally the geocoding of the street address can induce degree of misclassification in the results.

In conclusion, female BC incidence differently distributed across neighborhoods in Tehran. Higher than expected spatial clusters of BC incidence were appeared in central and northern parts of Tehran whereas, those with lower than expected incidence were located in southern part of Tehran. The observed neighborhood inequality can be a basis for allocation resources and preventive strategies for truly needy areas.

Conflict of interest: The authors have nothing to disclose.

Acknowledgment: This study was derived from Erfan Ayubi’s PhD thesis in Department of Biostatistics & Epidemiology, School of Public Health, Tehran University of Medical Sciences. We thank staff of cancer office of Iran's Ministry of Health and Medical Education for providing the data.
References

17. Lawson AB, Browne WJ, Rodeiro CLV. Disease mapping with WinBUGS and MLwiN: John Wiley & Sons; 2003
Table 1. High and low risk cluster for female breast cancer (BC) incidence using by spatial scan statistics in Tehran (2008, 11

<table>
<thead>
<tr>
<th></th>
<th>Optimal Gini coefficient</th>
<th>MSC</th>
<th>Clustered detected</th>
<th>Involved District</th>
<th>At risk population</th>
<th>Observed cases (O)</th>
<th>Expected cases (E)</th>
<th>Annual cases per 100000</th>
<th>O/E</th>
<th>RR**</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total BC incidence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n=3080)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Areas with high rates</td>
<td>0.47</td>
<td>0.04</td>
<td>Primary</td>
<td>3, 6</td>
<td>58039</td>
<td>217</td>
<td>55.37</td>
<td>126.4</td>
<td>3.92</td>
<td>4.14</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Secondary</td>
<td>3, 4</td>
<td>29134</td>
<td>145</td>
<td>27.79</td>
<td>165.9</td>
<td>5.22</td>
<td>5.43</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tertiary</td>
<td>4, 8</td>
<td>45449</td>
<td>161</td>
<td>43.36</td>
<td>118.1</td>
<td>3.71</td>
<td>3.86</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Areas with low rates</td>
<td></td>
<td></td>
<td>Primary</td>
<td>17, 18, 19</td>
<td>111902</td>
<td>6</td>
<td>106.75</td>
<td>1.8</td>
<td>0.05</td>
<td>0.05</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Secondary</td>
<td>16, 20</td>
<td>124216</td>
<td>27</td>
<td>118.50</td>
<td>7.2</td>
<td>0.23</td>
<td>0.22</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tertiary</td>
<td>14</td>
<td>113341</td>
<td>23</td>
<td>108.12</td>
<td>6.8</td>
<td>0.21</td>
<td>0.21</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

MSC, maximum size cluster

** Relative Risk is calculated as the observed divided by the expected within the cluster divided by the observed divided by the expected outside the cluster
Figure 1. The number of observed female breast cancer across neighborhoods in Tehran during 2008-2011
Figure 2. The estimated raw standardized incidence ratio (SIR) of female breast cancer incidence in Tehran (2008-2011)
Figure 3. The estimated standardized incidence ratio (SIR) of female breast cancer incidence using spatial empirical bayes method in Tehran (2008-2011)
Figure 4. The estimated standardized incidence ratio (SIR) of female breast cancer incidence using Besag, York and Mollie (BYM) spatial model in Tehran (2008-2011)
Figure 5. Spatial clusters of female BC incidence in Tehran (2008-2011)