



Published in final edited form as:

*Open J Stat.* 2015 February 6; 5: 46–50. doi:10.4236/ojs.2015.51006.

## Statistical Significance of Geographic Heterogeneity Measures In Spatial Epidemiologic Studies

Min Lian<sup>1,2</sup>

Min Lian: mlian@wustl.edu

<sup>1</sup>Department of Medicine, Washington University School of Medicine, St. Louis, Missouri, USA

<sup>2</sup>Alvin J. Siteman Cancer Center, Barnes-Jewish Hospital and Washington University School of Medicine, St. Louis, Missouri, USA

### Abstract

Assessing geographic variations in health events is one of the major tasks in spatial epidemiologic studies. Geographic variation in a health event can be estimated using the neighborhood-level variance that is derived from a generalized mixed linear model or a Bayesian spatial hierarchical model. Two novel heterogeneity measures, including median odds ratio and interquartile odds ratio, have been developed to quantify the magnitude of geographic variations and facilitate the data interpretation. However, the statistical significance of geographic heterogeneity measures was inaccurately estimated in previous epidemiologic studies that reported two-sided 95% confidence intervals based on standard error of the variance or 95% credible intervals with a range from 2.5<sup>th</sup> to 97.5<sup>th</sup> percentiles of the Bayesian posterior distribution. Given the mathematical algorithms of heterogeneity measures, the statistical significance of geographic variation should be evaluated using a one-tailed *P* value. Therefore, previous studies using two-tailed 95% confidence intervals based on a standard error of the variance may have underestimated the geographic variation in events of their interest and those using 95% Bayesian credible intervals may need to re-evaluate the geographic variation of their study outcomes.

### Keywords

Spatial Epidemiology; Heterogeneity; Statistical Significance; 95% Confidence Interval; 95% Credible Interval

### 1. Background

Spatial epidemiology is an important methodology to deal with spatial-correlated issues in epidemiologic studies. One of its core tasks is to determine geographic variations and quantify the magnitude of geographic variations in diseases, health behaviors, or environmental exposures [1]. Some published epidemiologic studies inappropriately estimated the statistical significance of geographic heterogeneity measures of examined events.

---

I declare no conflict of interest.

The generalized linear mixed model and the Bayesian spatial hierarchical model are the most commonly applied to fit the data with a multilevel spatial structure. A geographic variation can be directly quantified as neighborhood-level variance ( $\sigma^2$ ) from parameter estimations of the multilevel model fitting. However, this variance has no meaningful unit and is difficult to interpret. Spatial statisticians and epidemiologists have developed two state-of-the-art heterogeneity measures, the Median Odds Ratio (MOR, Equation 1) [2–4] and the Interquartile Odds Ratio (IqOR, Equation 2) [5], to facilitate the interpretation of geographic heterogeneity of an event.

$$\begin{aligned} \text{MOR} &= \exp(Z_{0.75} \times \sqrt{2 \times \text{VAR}}) \\ &= \exp(0.9539 \times \sqrt{\text{VAR}}), \end{aligned} \quad (1)$$

Where  $\text{VAR}$  is the neighborhood-level variance, while  $Z_{0.75}$  is the  $Z$  value of the Gaussian distribution at the 75<sup>th</sup> percentile (0.6745).

$$\begin{aligned} \text{IqOR} &= \exp\left((Z_{0.875} - Z_{0.125}) \times \sqrt{\text{VAR}}\right) \\ &= \exp(2.3007 \times \sqrt{\text{VAR}}), \end{aligned} \quad (2)$$

Where  $Z_{0.875}$  and  $Z_{0.125}$  are the  $Z$  values of the Gaussian distribution at the 87.5<sup>th</sup> and 12.5<sup>th</sup> percentiles (1.1504,  $-1.1504$ ), respectively.

Both MOR and IqOR are derived from the variance and are always greater than or equal to one. Larger values of MOR and IqOR denote greater geographic variations in the event of interest. The MOR reflects the average difference of risk when comparing two subjects who have the same individual characteristics and are selected randomly from two different neighborhoods. The IqOR represents the average difference of risk when comparing the first quartile of study subjects residing in neighborhoods with the highest risk to the fourth quartile of study subjects residing in neighborhoods with the lowest risk [3, 5]. Similarly, the Median Rate Ratio (MRR) and the Interquartile Rate Ratio (IqRR) can be estimated in a prospective study, and the Median Hazards Ratio (MHR) and the Interquartile Hazard Ratio (IqHR) [6] are for time-to-event studies. To facilitate the explanation, the MOR and IqOR are applied in the following discussions.

## 2. Issues in determining the statistical significance of geographic heterogeneity measures

Geographic variations can be qualitatively assessed by using neighborhood-level variance estimation derived from a generalized linear mixed model. The modeling conducted by a commonly used statistical analysis package, such as the SAS, also gives a  $Z$  value and a corresponding  $P$  value based on an approximately normal distribution of the estimated parameter. With the standard error of the variance from the multilevel model fitting, a 95% CI is able to be computed mathematically. However, one cannot perform a generalized linear mixed analysis to estimate the statistical significance and 95% CIs of the MOR and IqOR because both MOR and IqOR are derived from the variance and do not have their own standard errors.

Alternatively, a Bayesian spatial hierarchical model with a Markov Chain Monte Carlo (MCMC) simulation has been used to estimate geographic heterogeneities. In this setting, the 95% Bayesian credible interval (CrI), defined by the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles of Bayesian posterior distribution of the geographic heterogeneity measure, has been commonly reported.

In the estimation of a fixed effect of an exposure, its statistical significance can be identified if the 95% confidence/credible interval of its regression coefficient does not cross zero. However, this empirical method conflicts with the nature of geographic heterogeneity measures. Two unreasonable results are usually reported in the studies in which the 95% CI or CrI of geographic heterogeneity measures were used to determine their statistical significance. The 95% CI of the variance could cross zero based on an approximately normal distribution ( $X \pm 1.96 \times SE$ ). This is unreasonable because the variance should always be greater than or equal to 0. In addition, the 2.5<sup>th</sup> percentile of the Bayesian posterior distribution of the variance is always greater than 0 and consequently the MOR and IqOR are always greater than one. This leads to the overestimation of geographic disparities.

### 3. Example and Solution

#### 3.1 Example

A simulation analysis was performed to illustrate the issues relevant to the statistical significance of spatial heterogeneity measures. It is assumed that a population of colorectal cancer (CRC) survivors come randomly from 100 neighborhoods, each with 5–20 patients, and that the probability of smoking for each patient is 0.2~0.5. A random simulation generated a dataset that included 1245 patients and 420 smokers. Multilevel logistic regression is applied to quantify small-area geographic variation in smoking behavior among these CRC patients (Equation 3).

$$y_{ij} \sim \text{Binomial}(p_{ij})$$

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 N_i + \beta_2 X_{ij} + u_i \quad (3)$$

Where  $p_{ij}$  is the probability of smoking for patient  $j$  who resides in neighborhood  $i$ ;  $\beta_0$  is the intercept;  $\beta_1$  and  $\beta_2$  are the fixed coefficients of neighborhood- and individual-level covariates, respectively;  $N_i$  is characteristics of neighborhood  $i$ ; and  $X_{ij}$  is a vector of individual-level covariates;  $u_i$  is the random effect between neighborhoods with a normal assumption:  $u_i \sim (0, \sigma^2)$ .

To simplify the explanation, an empty model without neighborhood- and individual-level covariates was fit to estimate the overall geographic heterogeneity of smoking among these CRC patients using the Bayesian hierarchical approach with a MCMC simulation in WinBUGS (Version 1.4.3, MRC, UK). After 50,000 iterations for the convergence, additional 50,000 iterations were run to obtain the posterior estimates of three spatial heterogeneity measures. Because the dataset was simulated randomly, the geographic variation in smoking was expected to be small.

Table 1 shows the Bayesian parameter estimates of three heterogeneity measures. Based on an approximately normal assumption, the 95% CIs of three geographic measures were computed as  $\mu \pm 1.96 \times \sigma$ . Alternatively, the 95% CrIs of three geographic measures were expressed as the range from their 2.5<sup>th</sup> to their 97.5<sup>th</sup> percentiles. However, the inconsistent results were observed when comparing the 95% CIs of the variance, MOR and IqOR to their 95% CrIs. The 95% CI of the variance crossed zero and the 95% CIs of both MOR and IqOR crossed 1, suggesting no significant geographic variation in smoking behavior among CRC survivors. In contrast, the 95% CrI of the variance was more than zero and the 95% CrIs of the MOR and IqOR were greater than one, suggesting a significant geographic variation in smoking behavior.

### 3.2 Solution

Table 2 shows that, the variance is a non-negative measure, and MOR and IqOR are never less than one. The null hypothesis of the statistical test should be that the variance equals to zero and both MOR and IqOR equal to one, that is, there is no significant geographic variation in the event of interest. Meanwhile, the alternative hypothesis of the statistical test should be that the variance is greater than zero, and both MOR and IqOR are greater than one. Therefore, the statistical test is theoretically one-tailed, rather than two-tailed. The critical value for the significance level at 0.05 is 1.645 instead of 1.960. The statistical significance should be denoted directly using one-tailed (right-tailed)  $P$  value. One may not report the 95% CI or the interval between the 2.5<sup>th</sup> and the 97.5<sup>th</sup> percentiles of Bayesian posterior distribution (95% CrI) of geographic heterogeneity measures to avoid the misinterpretation of geographic variations. In fact, a one-tailed  $P$  value for the variation/heterogeneity estimation has been given from a generalized linear mixed model fitting using common statistical analysis packages, such as the SAS. For the heterogeneity estimation from a Bayesian hierarchical model, one should compute the corresponding statistics, based on the prior distribution of the variance, to obtain their one-tailed  $P$  value to determine its statistical significance. In the simulated example, since the  $Z$  value for the variance is:  $(0.007-0)/0.009=0.778$ , the geographic variation in smoking among CRC survivors is not statistically significant using 1.645 as the cutoff for the significance level at 0.05.

## 4. Discussions

The purpose of this study is to point out an inappropriate method that was used to determine the statistical significance of geographic heterogeneity measures. The simulated data suggested that empirically reporting of the 95% CI/CrI of geographic heterogeneity measures may lead to misunderstanding of the statistical significance of geographic variations of an event.

According to the nature of geographic heterogeneity measures, the statistical inference should be one-tailed (right-tailed). It is inappropriate to report a two-tailed 95% CI/CrI of a heterogeneity measure in spatial epidemiologic studies. It could mislead one in understanding the statistical significances of heterogeneity measures. In the studies using standard errors to obtain two-tailed  $P$  values or 95% CIs, geographic variations in the events may be underestimated because a two-tailed test is more conservative than a one-tailed test. In contrast, the studies using the interval between the 2.5<sup>th</sup> and the 97.5<sup>th</sup> percentiles of a

Bayesian posterior distribution to obtain a 95% CrI may overestimate the statistical significance of geographic variation of the event because a Bayesian 95% CrI never crosses zero for the variance and one for both MOR and IqOR. The issue of statistical significance of geographic heterogeneity measures, which was discussed in this paper, is also extendible to a general multilevel study aiming to investigate the variation(s) in one or multiple event(s) of interest across a non-spatial higher level, such as healthcare providers or medical service facilities.

## Acknowledgments

This work is supported partly by a Career Development Award (K07 CA178331) and a Research Award (R21 CA169807) funded by the National Cancer Institute, and a Research Award (R01 AA021492) funded by the National Institute of Alcohol Abuse and Alcoholism, both at the National Institutes of Health. I also thank the service provided by the Health Behavior, Communication and Outreach Core, which is supported by the National Cancer Institute Cancer Center Support grant (P30 CA091842) awarded to the Alvin J. Siteman Cancer Center at Barnes-Jewish Hospital and Washington University School of Medicine.

## Abbreviations and Acronyms

<b>VAR</b>	variance
<b>MOR</b>	median odds ratio
<b>MRR</b>	median rate ratio
<b>MHR</b>	median hazard ratio
<b>IqOR</b>	interquartile odds ratio
<b>IqRR</b>	interquartile rate ratio
<b>IqHR</b>	interquartile hazard ratio

## References

1. Elliott P, Wartenberg D. Spatial epidemiology: current approaches and future challenges. *Environmental health perspectives*. 2004; 112:998–1006. [PubMed: 15198920]
2. Larsen K, Petersen JH, Budtz-Jorgensen E, Endahl L. Interpreting parameters in the logistic regression model with random effects. *Biometrics*. 2000; 56:909–914. [PubMed: 10985236]
3. Larsen K, Merlo J. Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *American journal of epidemiology*. 2005; 161:81–88. [PubMed: 15615918]
4. Merlo J, Chaix B, Ohlsson H, Beckman A, Johnell K, Hjerpe P, Rastam L, Larsen K. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of epidemiology and community health*. 2006; 60:290–297. [PubMed: 16537344]
5. Chaix B, Merlo J, Subramanian SV, Lynch J, Chauvin P. Comparison of a spatial perspective with the multilevel analytical approach in neighborhood studies: the case of mental and behavioral disorders due to psychoactive substance use in Malmo, Sweden, 2001. *American journal of epidemiology*. 2005; 162:171–182. [PubMed: 15972939]
6. Chaix B, Rosvall M, Merlo J. Assessment of the magnitude of geographical variations and socioeconomic contextual effects on ischaemic heart disease mortality: a multilevel survival analysis of a large Swedish cohort. *Journal of epidemiology and community health*. 2007; 61:349–355. [PubMed: 17372297]

**Table 1**

Three Bayesian estimates of three spatial heterogeneity measures in the simulated Example.

Measure	Mean ( $\mu$ )	SD ( $\sigma$ )	2.50%	Median	97.50%	95% Confidence Interval	95% Credible Interval
VAR *	0.007	0.009	0.001	0.004	0.033	0.007 (-0.011, 0.025)	0.004 (0.001, 0.033)
MOR †	1.074	0.045	1.022	1.061	1.189	1.074 (0.986, 1.162)	1.061 (1.022, 1.189)
IqOR ‡	1.190	0.125	1.054	1.155	1.517	1.190 (0.945, 1.435)	1.155 (1.054, 1.517)

\* VAR, variance;

† MOR, median odds ratio;

‡ IqOR, interquartile odds ratio.

**Table 2**

Three spatial heterogeneity measures and their statistical hypotheses.

Measure	Range	Null Hypothesis ( $H_0$ )	Alternative Hypothesis ( $H_1$ )
VAR *	0	VAR = 0	VAR > 0
MOR †	1	MOR = 1	MOR > 1
IqOR ‡	1	IqOR = 1	IqOR > 1

\* VAR, variance;

† MOR, median odds ratio;

‡ IqOR, interquartile odds ratio.