Bayesian Disease Mapping for Public Health utilizing Relative Risks and Standard Mortality Ratios

Valentina Arputhasamy, John Molitor

Disease mapping is an approach used in spatial epidemiology to examine the geographical distribution of disease and to identify areas of **excessive disease counts** compared to what one would expect given the demographic characteristics of the region in question (Lawson and Lee 2017; Elliott and Wartenberg 2004). Such approaches have a long history dating back to the 1800s when scientists utilized maps of disease rates to characterize the spread and to predict possible outbreaks of diseases such as cholera (Walter 2001). By employing these methods, one can obtain more robust estimates of disease risk compared to traditional approaches by considering the spatial distribution of disease and by providing measures of uncertainty regarding results obtained.

One approach to examining excess disease counts as compared to what one would expect given a reference population is the calculation of Standardized Incidence Rates (SIRs) (Becher and Winkler 2017) and Relative Risks (RRs) (Tenny and Hoffman 2024). An SIR is a ratio that estimates the observed occurrence of an event in a population relative to the expected occurrence of the event in a larger comparison population. For example, if 20% of individuals in a large reference population have a disease, then in a region with 20 individuals, one would expect 20*0.2 = 4 diseased individuals. If 6 individuals have the disease in the region then the SIR is 50% higher than expected. Such computations can be adjusted for demographic variable such as age and sex, so that expected counts take these factors into account.

A relative risk is like an SIR except the computation is based on a statistical model, such as a Poisson-based count model, and consists of a modeled value as opposed to an empirical point estimate. Such modeling has advantages since it can estimate measures of uncertainty, such as confidence limits and exceedance probabilities (e.g. the probability that the RR is great than one) and allow for spatial “smoothing”, where data from nearby regions can inform estimates of parameters in the region in question. Such modeling is extremely useful in small areas where data is sparse, and SIRs can become “noisy” and unreliable.

HCAI and CCDF datasets consists of data collected at the individual level and includes a variety of demographic and clinical variables, but variables that we are primarily interested in are age, sex, patient zip code, and primary diagnosis. Before beginning the analysis, the data was subset by health condition and modified so that age and sex were categorized into four and two groups, respectively. The categories for age are: 00-19, 20-44, 45-64, and 65+, while the categories for sex are: female (F), male (M), and unknown (U). Although race/ethnicity data was available, it was excluded from the analysis due to small sample sizes within geographic areas, which introduced high variability and uncertainty in the model estimates.

The goal of the analysis is to obtain estimates for the relative risks and standardized incidence ratios of ER visits for each primary diagnosis using a spatial regression model with the R-INLA package (Rue, Martino, and Chopin 2009). Since our outcome (number of ER visits) is a count, we analyze our data using a Poisson regression.

If we let Y_i represent the observed counts of ED visits in zip code i then our model can be specified as follows:

$$Y_i \sim Poisson(E_i\theta_i), i, \dots, n, \tag{1}$$

such that E_i is the expected number of ER visits in zip code i and θ is the relative risk in zip code i. The logarithm of θ_i can be generally expressed as:

$$\log(\theta_i) = \beta_0 + \mu_i + \nu_i, \tag{2}$$

where β_0 is the intercept where and represents baseline log risk, and μ_i and ν_i represent structured and independent spatial effects, respectively. A conditionally auto-regressive (CAR) (Besag 1974) distribution is used to model μ_i which is a structured spatial effect. With this, the distribution can be specified as: $\mu_i|\mu_{-i} \sim N\left(\bar{\mu}_{\delta_i}, \frac{\sigma^2_{\mu_i}}{n_{\delta_i}}\right)$.
On the other hand, $\nu_i \sim N(0, \sigma^2_\nu)$ is the unstructured spatial effect, often known as the error term. Finally, the relative risk, θ_i is used to determine if the risk of ER visit is higher $\theta > 1$ (or lower) in zip code i when compared to the average risk in the reference population. Note that the probability $p = \Pr(\theta > 1)$ denotes the exceedance probability mentioned above and is an important measure of uncertainty. In our study, we have defined two reference groups, which yield two separate sets of relative risks and SIRs. These reference groups are further explained in the following subsections.

Reference Group 1. The “Healthy Places Index" (HPI), is a metric developed by the Public Health Alliance of Southern California (PHASC) to measure the healthiness of various neighborhoods and communities (https://www.healthyplacesindex.org/). The HPI defines a healthy community as one that provides residents with access to quality education, good jobs, safe housing, clean air and water, healthcare, and strong social support using 25 indicators across these areas. In our study, the HPI scores provided in the PHASC's public database were used to define the reference group in our first set of relative risk and SIR calculations. To begin, the HPI scores for every zip code in California obtained from the PHASC database were merged with the HCAI dataset by zip code. Next, the merged dataset was filtered using a pre-selected HPI cutoff value to identify the zip codes with “heathy” communities, defined as the top 25% of the HPI. The quantiles selected to define a “healthy” community have been frequently used in epidemiological studies and was selected after extensive consultation with the AIRE collaborative, community members, and agency partners. The higher the HPI score, the healthier the community. We then divided the total number of observed ER visits in each age/sex category by the total number of individuals belonging to that age/sex category in the population, which was estimated using the American Community Survey (ACS) census data. This calculation provided the expected ER visit rate for each age/sex combination in a healthy population. Using these rates, the expected number of cases was calculated for each zip/age/sex combination. The total number of observed ER visits in each zip code was then divided by total number of expected ER visits in the same area.

In other words, the SIR in the i^{th} zip code was calculated as follows:

$$\mathrm{SIR_i} = \frac{\text{Observed ED Visits}_i}{\text{Expected ED Vists}_\mathrm{i}}, \tag{3}$$

where $\text{Oberved ER Visits}_\mathrm{i}$ is the total number of observed ER visits in zip code i and $\text{Expected ER Visits}_\mathrm{i}$ is the expected number of ER visits in zip code i, calculated based on the assumption that the zip code has the same average rate of ER visits as zip codes with an HPI score above the cutoff. To calculate the relative risks, we supplied R with the expected counts calculated for each zip code, and the neighborhood matrix needed to define the spatial random effects. The "inla" function within the R-INLA package was then used to compute the posterior estimates of the relative risks for each zip code.

Reference Group 2 The reference group in our second set of relative risk and SIR calculations was determined using the expected number of ER visits in as the statewide population, after adjusting for age and sex. The expected rate was calculated by dividing the total number of observed ER visits in each age/sex category by the total number of individuals belonging to that age/sex category in the population, as estimated using the ACS census data. With this expected rate, we then calculated expected number of cases for each zip/age/sex combination. Subsequently, we divided the total number of observed ER visits in each zip code by the expected number of ER visits in the same zip code, yielding the SIR for the i^{th} zip code. Mathematically, the formula for SIR_i remains the same as described in (Becher and Winkler 2017), but with $\text{Expected ER Visits}_\mathrm{i}$ representing the expected number of ER visits in zip code i, calculated under the assumption that the zip code has the same average rate of ER visits as zip codes across the rest of California. The relative risks in this version were calculated using the same methods, with the sole variation being the calculation of expected counts for each zip code, which were determined using the reference group defined in this section.

Conclusion: It is important to recognize that the SIRs are effectively point estimates of the relative risks, but they do not benefit from advanced modeling constructs such as spatial smoothing and measures of uncertainty. These techniques have been appropriately applied in the estimation of the relative risks, allowing for a nuanced understanding of risk that accounts for spatial variation and uncertainty. We hope this clarification emphasizes the distinction between the SIRs and relative risks and the benefits of using relative risks in spatial modeling.

References:

Becher, Heiko, and Volker Winkler. 2017. "Estimating the Standardized Incidence Ratio (SIR) with Incomplete Follow-up Data." *BMC Medical Research Methodology* 17 (April):55. https://doi.org/10.1186/s12874-017-0335-3.

Besag, Julian. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems (with Discussion)." *J1 Roy1 Statist1 Soc1 B*.

Elliott, Paul, and Daniel Wartenberg. 2004. "Spatial Epidemiology: Current Approaches and Future Challenges." *Environmental Health Perspectives* 112 (9): 998–1006. https://doi.org/10.1289/ehp.6735.

Lawson, Andrew, and Duncan Lee. 2017. "Bayesian Disease Mapping for Public Health." In *Handbook of Statistics*. https://doi.org/10.1016/bs.host.2017.05.001.

Rue, H., S. Martino, and N. Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *J R Stat Soc Ser B Stat Methodol* 71. https://doi.org/10.1111/j.1467-9868.2008.00700.x.

Tenny, Steven, and Mary R. Hoffman. 2024. "Relative Risk." In *StatPearls*. Treasure Island (FL): StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK430824/.

Walter, S. D. 2001. "Disease Mapping: A Historical Perspective." In *Spatial Epidemiology: Methods and Applications*, edited by Paul Elliott, Jon Wakefield, Nicola Best, and David Briggs, 0. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198515326.003.0012.