Empirical Bayes Estimators under Nonparametric Priors for Disease Mapping of HIV/AIDS

Apinya Surawit, Chukiat Viwatwongkasem*, Pichitpong Soontornpipit, Prasong Kitidamrongsuk
Department of Biostatistics, Faculty of Public Health,
Mahidol University
Bangkok, Thailand
chukiat.viw@mahidol.ac.th

Piangchan Rojanavipart
Faculty of Public Health,
Bangkok, Thailand
piangchan.roj@mahidol.ac.th

Sutthi Jareinpituk
Department of Epidemiology, Faculty of Public Health,
Mahidol University
Bangkok, Thailand
sutthi.jar@mahidol.ac.th

Abstract—Empirical Bayes (EB) method is a statistical technique in which the prior distribution by Bayesian rule is estimated from the data. This paper focuses on generation of the prior distribution from previously informative data to improve the efficiency of estimation on the standardized morbidity ratio (SMR) of newly diagnostic HIV/AIDS infected cases. Data source of newly diagnosed HIV/AIDS infection is provided by the NAP (National AIDS program) project, collected by the National Health Security Office (NHSo) in Thailand 2008 to 2013. After constructing and testing empirical prior distribution of previous SMR data during 2010-2012, we found that the distribution of prior followed a normal distribution at approximately 0.1 p-value. The further results of EB estimation indicated that in Thailand 2013, the mean of SMREB overall the country was 0.85 that decreases slightly in HIV infection, compared with the past five years from 2008 to 2012 as the national standard reference. The maximum SMR of 3.26 found in Nong Bua Lamphu, and the minimum value of 0.14 in Roi Et. Empirical Bayes estimates with normal prior perform well with smaller variance, leading to the narrow width of the credible interval (CI).

Keywords—standardized morbidity ratio; disease mapping; nonparametric prior; empirical Bayes; HIV/AIDS

I. INTRODUCTION

Empirical Bayes (EB) method is a statistical technique that the prior information in Bayes' rule is estimated from the data. The EB method is directly related to the Bayes’ theorem with difference in the sense that: for the fully Bayesian approach, the prior is assumed to be fixed distribution before any data are observed, whereas the prior for the EB approach is an estimated distribution from the observed data [1]. In a typical EB approach, the conditional probability density given the unknown parameters is assumed in order to create the likelihood function coping with the observed data. The unknown parameters are treated as a random variable but they are not enough information about the distribution (prior). The idea of EB method is a generation of an empirical prior distribution obtained from the previous knowledge data. Hence, the outcomes in terms of the posterior distribution, the posterior mean, and the posterior variance are further investigated and obtained. Note that the construction of the prior from the past knowledge data is an essential part of the EB method. The posterior distribution as an outcome is described by combining the likelihood of the observed data with the prior of the previously informative data. While Markov Chain Monte Carlo (MCMC) method is also the effective tool in finding the posterior outcomes.

The motivational example is on the disease mapping about HIV/AIDS data. The HIV/AIDS distribution contains difficult pattern to understand information of disease occurred because the illness has been occurred in various sex and age groups. In addition, HIV/AIDS morbidity is one among of the public health surveillance issues that continue to spread in many areas. According to a report by the World Health Organization (WHO) in 2016, there were 36.7 million people living with HIV/AIDS worldwide, and more than 1 million people have died of AIDS. During the past 10 years the incidence of HIV infections increased to 5 million cases [2]. Recently, the incidence of HIV infection in Thailand has begun to decline slowly. However, geographically spatial distribution of HIV/AIDS infection looks doubtful because of a shortage of information.

Typically, a measure of the spatial heterogeneity of a disease distribution used in epidemiology and public health is the standardized morbidity ratio (SMR). The SMR as a relative risk among areas is commonly used in disease mapping since the SMR itself has an advantage of controlling the confounding variables such as gender and age. Since a small event size or a rare disease situation might have an effect on the calculation of SMR, the smoothed risks of SMR are recommended to adjust [3, 4]. Indeed, EB method being a statistical approach can provide the smoothed estimates.
The main objective of this paper is to construct the prior distribution from the underlying informative data to improve the efficient estimation on the SMR of HIV/AIDS infection, and to demonstrate how the EB method can lead to a smoothed map with a few of extreme estimates [5,6]. The use of EB method in disease mapping shows the visual representations of complex data that help the researchers to test the hypotheses about the various factors of disease, to identify the high-risk areas for public health surveillance and to aid in policymaking to allocate resources and to prevent disease.

II. STATISTICAL METHODS

A. Data Sources and Characteristics

The registration data set of newly diagnosed HIV/AIDS infection in the NAP (National AIDS program) were collected by the National Health Security Office (NHSO) Thailand, 2008 to 2013. The scope of this study identified the SMR of newly diagnosed HIV/AIDS infected on the whole Thailand in 2013, compared with the national standard reference during a past 5 years of age and sex adjusted infection rates in 2008 to 2012. Normally, the analysis of spatial data with respect to time is aggregated in certain time windows like 5 or 10 years [7].

B. Standardized Morbidity Ratio

Let \( \lambda_i = SMR_i = O_i / E_i \) be considered as the standardized morbidity ratio (SMR, ) for province \( i (i=1,...,n) \). Given that \( O_i \) is the observed number of newly diagnosed HIV/AIDS infection cases and \( E_i \) is the expected number of HIV/AIDS infection cases where \( E_i \) is calculated by the adjustment of age and sex confounders within the indirect method of standardization in Thailand during period 2008 to 2012.

C. Standardized Morbidity Ratio

Let a random sample of \( n \) observations of SMR data \( \lambda = (\lambda_1,...,\lambda_n) \) be the conditionally normal distribution with density \( f(\lambda_i | \theta) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left( -\frac{(\lambda_i - \theta)^2}{2\sigma^2} \right) \) where \( \theta \) refers to a variously unknown mean and \( \sigma^2 \) is a known variance. In other words, we have

\[
\lambda_i | \theta \sim N(\theta, \sigma^2), \quad E(\lambda_i | \theta) = \theta, \quad Var(\lambda_i | \theta) = \sigma^2
\]  

(1)

Suppose we take heterogeneity prior \( \theta \) where \( \theta = (\theta_1,...,\theta_n) \) having any density \( p(\theta) = p(\mu, \tau^2) \) with unknown mean \( \mu \) and known variance \( \tau^2 \) as

\[
\theta \sim p(\mu, \tau^2), \quad \mu = E(\theta), \quad \tau^2 = Var(\theta)
\]  

(2)

The marginal likelihood of \( \lambda = (\lambda_1,...,\lambda_n) \) can be obtained as

\[
f(\lambda) = \prod_{i=1}^{n} \left[ \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left( -\frac{(\lambda_i - \theta)^2}{2\sigma^2} \right) p(\theta) \right]
\]

(3)

In the beginning step, if the prior \( p(\theta) \) is proved to be a normal distribution \( p(\theta) = N(\mu, \tau^2) \) with mean \( \mu \) and variance \( \tau^2 \), by plugging the normal prior into Eq. (3), then marginal likelihood of \( \lambda \) is obtained as

\[
f(\lambda) = \frac{1}{\left[ 2\pi(\theta^2 + \tau^2) \right]^{n/2}} \exp \left( -\frac{1}{2(\theta^2 + \tau^2)} \sum_{i=1}^{n} (\lambda_i - \mu)^2 \right)
\]  

(4)

Next the posterior distribution of \( \theta \) given \( \lambda \) follows

\[
p(\theta | \lambda) = \frac{N(\lambda | \theta) p(\theta)}{f(\lambda)}
\]  

(5)

We see that the figures of marginal likelihood, posterior distribution, posterior mean, and posterior variance, all depend upon the distribution of prior \( p(\theta) \). Regularly, Markov Chain Monte Carlo (MCMC) method is used to find those posterior outcomes. From the initial step, the posterior distribution of \( \theta \) given \( \lambda \) under the normal prior is also a normal distribution:

\[
p(\theta | \lambda) = N(\sigma^2 \mu + \tau^2 \lambda, \sigma^2 \tau^2)
\]  

(6)

Writing \( B = \sigma^2 / (\sigma^2 + \tau^2) \), the posterior mean is \( E(\theta | \lambda) = B \mu + (1 - B) \lambda \), and the posterior variance is \( Var(\theta | \lambda) = B \tau^2 = (1 - B) \sigma^2 \). In the next section, the unknown prior distribution will be estimated by generating the empirically observed data of SMR for HIV/AIDS infection.

D. Empirical Prior Distribution

The prior data of 231 SMRs in the past three years from 2010-2012 provided the mean of 0.95, the standard deviation (SD) of 0.53, the minimum of 0.12, and the maximum of 3.45. To construct the empirical prior, we start to form a frequency distribution, a cumulative distribution, and a histogram of these SMRs to describe the shape of data and then we use the above mean and variance to represent the population mean and the population variance in testing the goodness-of-fit of the observational data to a specific distribution.

Table 1 shows the relative frequency f(SMR) as the empirical probability density function (PDF) of SMR, and the cumulative relative frequency F(SMR) as the empirically cumulative distribution function (CDF) of SMR, respectively.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Frequency (SMR)</th>
<th>F(SMR)</th>
<th>f(SMR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00-0.34</td>
<td>21</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>0.35-0.70</td>
<td>60</td>
<td>0.35</td>
<td>0.26</td>
</tr>
<tr>
<td>0.71-1.05</td>
<td>73</td>
<td>0.67</td>
<td>0.32</td>
</tr>
<tr>
<td>1.06-1.40</td>
<td>38</td>
<td>0.83</td>
<td>0.16</td>
</tr>
<tr>
<td>1.41-1.75</td>
<td>18</td>
<td>0.91</td>
<td>0.08</td>
</tr>
<tr>
<td>1.76-2.10</td>
<td>13</td>
<td>0.97</td>
<td>0.06</td>
</tr>
<tr>
<td>2.11-2.45</td>
<td>5</td>
<td>0.99</td>
<td>0.02</td>
</tr>
<tr>
<td>&gt;2.45</td>
<td>3</td>
<td>1.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Fig. 1 shows a graphical shape of data sets. It is possible to have either a slightly positive skewed distribution or a symmetric normal distribution.

\[\mu_{\lambda,\sigma} = \mu + \lambda \sigma \]  

\[\tau = \frac{\sigma}{\lambda} \]

\[p(\theta \mid \lambda, \mu, \sigma^2) = N(\mu, \tau^2) \]

\[\text{Empirical Bayes Estimation} \]

F. Empirical Bayes Estimation

Since the empirical prior \( p(\theta) \) is accepted to be a normal distribution \( \theta \sim N(\mu, \tau^2) \) with \( \mu \) its mean and \( \tau^2 \) its variance, both unknown. We replace \( p(\theta) = N(\mu, \tau^2) \) in the marginal distribution and the posterior distribution (using Bayes’ rule) can be achieved as

\[p(\theta \mid \lambda, \mu, \sigma^2) = N(B \mu + (1 - B) \lambda, (1 - B)\sigma^2) \]  

The associated posterior mean, which leads to empirical Bayes estimates for the \( SMR \) is

\[SMR_{_{EB}} = \hat{\theta}_{_{EB}} = E(\theta \mid \mu, \tau^2, \lambda) = (1 - \hat{B})\lambda + \hat{B} \mu \]

Also, the posterior variance of the \( SMR \) is

\[\text{Var}(\theta \mid \mu, \tau^2, \lambda) = (1 - \hat{B})\sigma^2 \]

Where \( \hat{B} = \sigma^2 / (\sigma^2 + \tau^2) \), \( \lambda = \lambda = (1/n) \sum_{i=1}^{n} \lambda_i \), \( \tau^2 = (S^2 - \sigma^2) \), and \( S^2 = \sum_{i=1}^{n} (\lambda_i - \bar{\lambda})^2 / (n - 1) \).

### III. Results

In this study, analysis of disease mapping by using the SMR of newly diagnosed HIV/AIDS infected cases overall in Thailand 2013, the basic unit is the province that composes of \( n = 77 \). The number of newly diagnosed HIV/AIDS cases in 2013 was 18,081 while there was the total population at risk with blood test of 623,265. The mean of \( SMR_{EB} \) for overall the country in 2013 was 0.85 that decreases slowly in HIV infection, compared with the past five years from 2008 to 2012 as the national standard reference. There were fifteen provinces of Thailand having \( SMR_{EB} \) more than 1 in which these provinces had a higher risk.

Eight provinces with highest order \( SMR_{EB} \) were Nong Bua Lam Phu (\( SMR_{EB} \): 3.26 (95% CI: 3.14, 3.37)), Chumphon (\( SMR_{EB} \): 2.43 (95% CI: 2.31, 2.54)), Samut Prakan (\( SMR_{EB} \): 2.26 (95% CI: 2.14, 2.38)), Udon Thani (\( SMR_{EB} \): 1.95 (95% CI: 1.83, 2.06)), Pathumthani (\( SMR_{EB} \): 1.58 (95% CI: 1.47, 1.70)), Krabi (\( SMR_{EB} \): 1.46 (95% CI: 1.35, 1.58)), Singburi (\( SMR_{EB} \): 1.45 (95% CI: 1.33, 1.56)), and Mukdahan (\( SMR_{EB} \): 1.41 (95% CI: 1.30, 1.53)). Empirical Bayes estimates with normal prior perform well with the smaller variances, leading to the narrow width of the credible interval (CI) and the results are presented in Table II.

### TABLE II. PROVINCES WITH SMR-Empirical Bayes (EB) ESTIMATES

<table>
<thead>
<tr>
<th>Province name</th>
<th>SMR(95%CI)</th>
<th>SMR_{EB}(95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Nong Bua Lam Phu</td>
<td>3.29 (2.25, 4.32)</td>
<td>3.26 (3.14, 3.37)</td>
</tr>
<tr>
<td>2  Chumphon</td>
<td>2.45 (2.41, 3.48)</td>
<td>2.43 (2.31, 2.54)</td>
</tr>
<tr>
<td>3  Samut Prakan</td>
<td>2.28 (1.24, 3.31)</td>
<td>2.26 (2.14, 2.38)</td>
</tr>
<tr>
<td>4  Udon Thani</td>
<td>1.96 (0.93, 2.99)</td>
<td>1.95 (1.83, 2.06)</td>
</tr>
<tr>
<td>5  Pathumthani</td>
<td>1.59 (0.56, 2.63)</td>
<td>1.58 (1.47, 1.70)</td>
</tr>
<tr>
<td>6  Krabi</td>
<td>1.47 (0.44, 2.51)</td>
<td>1.46 (1.35, 1.58)</td>
</tr>
<tr>
<td>7  Singburi</td>
<td>1.45 (0.42, 2.49)</td>
<td>1.45 (1.33, 1.56)</td>
</tr>
<tr>
<td>8  Mukdahan</td>
<td>1.42 (0.39, 2.46)</td>
<td>1.41 (1.30, 1.53)</td>
</tr>
<tr>
<td>9  Prachinburi</td>
<td>1.39 (0.35, 2.42)</td>
<td>1.38 (0.26, 1.50)</td>
</tr>
<tr>
<td>10 Angthong</td>
<td>1.37 (0.34, 2.41)</td>
<td>1.36 (1.25, 1.48)</td>
</tr>
</tbody>
</table>
Disease mapping of HIV/AIDS infection in the year 2013 identified the top five provinces similar to the work of Viwatwongkasem et al. [8], which having the highest risk are Nong Bua Lamphu, Chumphon, Samut Prakan, Udorn Thani, and Pathumthani. The empirical Bayes estimates with normal prior are robust and more interesting. Consider, the credible interval (CI) of SME_EB, the length of the estimator are shown to be shorter than the traditional SMR all through the 77 provinces. In the small area data which less population of Nong Bua Lam Phu, Mukdahan, Chumphon and Samut Songkram, the SME_EB and CI are still shown the shorter range of CI than in the SMR traditional.

Thus, estimation of relative risk using empirical Bayes approach are visual representation of heterogeneity explained in the geographical data of disease mapping. In Thailand, the SMR (more than 1) of HIV infection were Central region of Thailand (13%), Southern (9.09%), Northern (3.89%) and Northeast (3.89%). Each region is increased slowly of HIV infection, compared with the past five years from 2008 to 2012 as the national standard reference. Therefore, the result helps to generate hypotheses about the various factors of the spread of HIV infection and associated with risk behaviors in each region. It is the public health surveillance and to prevent the spread of the disease. We recommend that further work be addition of covariate information.

In our study, using classification of SMR based on the histogram of empirical prior which we divided into the number of 5 components. The heavily elevated risk or the 5th component contains 8 (10.39%) provinces. The medium elevated risk or the 4th component has 11 (14.29%) provinces. The slightly increased risk or the 3rd component is 19 (24.68%) provinces. Finally, the slightly reduced risk or the 2nd component is 29 (37.66%) provinces. We show the quartile map based on the SMR as well as the one based on Eq. (9), which can be seen in Fig. 2.

The report of Bureau of Epidemiology [9] in the year 2013 identified provinces with the highest prevalence of HIV/AIDS were Phuket (73.92 per 100,000 mid-year population), Trat and Rayong. The result of our of the highest risk HIV infection does not relate to the report of the Bureau of Epidemiology because of the data from the different sources, the different time period, and the different target populations. This study the registration data set of newly diagnosed HIV infection out of the population at risk scrutinized with blood test from the NAP program supported by the NHSO during 2008 to 2012. Whereas, the numerator data of report of Bureau of Epidemiology are data of the accumulation (or prevalence) of HIV/AIDS individuals from the past to the present and the denominator data are the mid-year population. However, the researcher is waiting for analysis of current data of HIV infection from the NHSO after that the data will be compared with data report of the country more and more.

![Fig. 2. Disease map of the base on SMREB using a Normal prior of HIV/AIDS infections case in Thailand for the years 2013.](image)

### IV. DISCUSSION

### ACKNOWLEDGMENT

This paper was partially supported by the Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok, Thailand.

### REFERENCES


