Analysis of Age-specific rates and age-standardized rates using Bayesian hierarchical model

PHP2530

Zhaoxiang Ding

September 11, 2024

1. Introduction

In the field of spatial epidemiology, disease mapping is a valuable tool as it can used assess patterns in incidence or mortality of a given disease(Elliott and Wartenberg, 2004). Understanding the distribution of diseases is crucial for public health professionals to allocate resources effectively. Disease mapping, a valuable tool in spatial epidemiology, allows researchers to analyze disease patterns(Elliott and Wartenberg, 2004). When the interest lies into the field of comparing different disease that belong to the same family in a limited population region, rates derived from raw data will be subject to high variability(Jay et al., 2021). This is especially true for disease that might be in low-prevalence under some region and age group. Under these situations, Bayesian hierarchical modeling offers a promising solution for analyzing such data. (Jay et al., 2021).

Age-specific rates and age-standardized rates are valuable tools for disease mapping that can be used to compare the health of different populations. Unlike crude rates, which are calculated by dividing the number of cases by the population at risk and multiplying by a certain number (typically 100,000 for cancer rates), age-specific rates and age-standardized rates offer a more accurate comparison of disease risk, especially when the disease is strongly influenced by the age distribution(Klein, 2001). Age-specific rates are calculated by determining a crude rate for each age group separately, while age-standardized rates are obtained by calculating a weighted average of the age-specific rates based on the population proportions of a standard population(Buescher, 2008; dos Santos Silva, 1999). However, the estimated rates may be unreliable for diseases with low prevalence, where the number of incidence cases or deaths in the population is low, leading to a significant number of zero counts in the datasets. In general, when the number of incidences in a specific age group population is less than 20, the corresponding age-specific rate is considered unreliable, as even a small change in the number of incidences can result in a substantial change in the rates. When calculating age-standardized rates, these unreliable age-specific rates are not taken into account.

In the presence of excessive zero counts, common models like the Poisson regression model may fail as they may predict fewer zeros than observed in the dataset (Aregay et al., 2018). Hurdle models and zero-inflated models, such as the Poisson hurdle model and the

zero-inflated Poisson (ZIP) model, are two popular approaches to appropriately analyze count data that contains excessive zeros (Arab, 2015). Both models are mixture models. The Poisson hurdle model is a mixture of a point mass at zero and a zero-truncated Poisson distribution. The model includes 2 stages. First, the model uses a Bernoulli distribution to determine whether the count is zero or not. Next, the non-zero counts are further modeled through a zero-truncated Poisson distribution. The model assumes that every individual in the study population is "at risk" of the outcome, the difference is only in how high the risk is. The ZIP model, on the other hand, distinguishes between those who are at risk and those who are not (Arab, 2015). In this project, we will use the Poisson hurdle model as the assumption is more aligned with the characteristics of diseases.

Several studies have using Poisson hurdle model to estimate zero-excess disease counts data. Jay et al. (2021) using a Poisson hurdle model with multilevel regression to model the age specific and age standardized rates for low-prevalence diseases in small areas. They analyze the county level age adjusted mortality rates for liver cancer and colorectal cancer in Midwest region of the United states. Their models includes 2 multilevel regression model include random effects of county level and year to separately estimated the probability of whether the count is zero or not and the rate parameter for the zero-truncated Poisson model.

Based on Jay et al. (2021) work, we propose a modified version of Bayesian hierarchical hurdle model frameworks which does not cluster on region, but cluster on cancer type. The number of incidence of receiving any type cancer can be related to other type of cancer. Such relation may come from some unobserved con-founders, such as a higher participation in cancer screening or unhealthy behaviours (Mokdad et al., 2018), as long as second cancer, which are defined as a new cancer that's unrelated to any previous cancer diagnosis in the patient (Society, 2002). The proposed model can predict the expected incidence counts, as long as corresponding age-specific and age standardised rate.

The remainder of this article is organized as follows. In 2, we describe the data will use to analyze. In 3 we describe the model and the prior we will use. In 4 we will show the posterior likelihood of our model, and in 5 we will present the results of our model.

2. Data Structure and Notation

The data we will use to analyze is the Cancer registration statistics 2013-2017, England. The data sets come from National Cancer Registration and Analysis Service within Public Health England; Office for National Statistics (Caul and Broggio, 2019). The data set include cancer diagnoses and age-standardised incidence rates for all types of cancer by age, sex and region including breast, prostate, lung and colorectal cancer. Age structure is set with 20 groups from 0, 1-4 years old to 90+ years old . The cancer type in the data is coded using the tenth version of International Statistical Classification of Diseases and Related Health Problems (ICD-10). Under this coding system, same type of cancer will be assigned with different code based on their minor difference(e.g. Kidney cancer will be assigned with code C64, C65, C66 and C68). We merged the data with similar cancer feature and separate the cancer into several types based on Table 2 from World Health Organization. The age specific rate in the data set is calculated by:

$$ASR_k = (r_k/p_k) \times 100000$$

Table 1: Cancer type included in our work and corresponding ICD-10 Code

Cancer	**ICD-10 code**
Oesophagus	C15
Stomach	C16
Colorectal	C18 to C20, C21.8
Lung	C33 to C34
Breast	C50
Kidney and Urinary Tract	C64 to C66, C68
Bladder	C67
Non-Hodgkin lymphoma	C82 to C85
Myeloma	C90
Leukaemia	C91 to C95

Where ASR_k is the age-specific rate for age group k, r_k is the registrations in age group k and p_k is the population in age group k. k = 0, 1-4, 5-9, ... , 85-89, and 90 and over. The Age-standardised rates is computed from direct standardisation: age- and sex-specific rates in each group in the populations to be compared are multiplied by the corresponding number of people in a 'standard' population, World or (here) European Standard Population, and then summed to give an overall rate per 100,000 population. The formula is given by:

$$I(ASR/E) = \{\sum_k ASR_K P_k\}_k / \sum P_k$$

where P_k is the European standard population in age group k.

The dataset shares the same zero-excess problems as we described above. As for young age group (e.g. age group of below 0, or 0-5 years old) are less likely to get cancer, the incidence counts of these age groups are zero. We apply our model on this dataset to see whether the model can overcome the obstacles and successfully predict the incidence rate.

3. Methodology

3.1 Model

The following model and prior come from Jay et al. (2021) work except stated otherwise.

Let $Y_{i,j,k}$ denotes the raw incidence count for cancer type i (modification made in this work)during year j for age group k, and assume there are I cancer types, J years and K groups in the dataset. The possion hrudle model can be written as:

$$P(Y_{i,j,k} = y_{i,j,k}|\pi_{i,j,k}, \theta_{i,j,k}, n_{i,j,k}) = \begin{cases} 1 - \pi_{i,j,k} & y_{i,j,k} = 0 \\ \pi_{i,j,k} * \frac{\theta_{i,j,k}^{y_{i,j,k}} exp(-\theta_{i,j,k})}{y_{i,j,k}!(1-exp(-\theta_{i,j,k}))} & y_{i,j,k} > 0 \end{cases} \quad (1)$$

Where $0 < \pi_{i,j,k} < 1$ and $\theta_{i,j,k} > 0$. We allow each cancer type and year to have its own $\pi_{i,j,k}$, the probability of at least one death. The reason is that the probability may varied a lot based on the different type of cancer and age. $\theta_{i,j,k}$ also varies among cancer type, year and age group and serve as the mean counts from a Poisson distribution. $\pi_{i,j,k}$ can be derived based on a logit transformation of a regression equation includes a fixed effect for

each age group, a fixed effect for the interaction between age group and log of the population size, a random effect for cancer type, and a random effect for time:

$$logit(\pi_{i,j,k}) = \mathbf{x_k^T}\boldsymbol{\alpha_1} + log(n_j, k) * \mathbf{x_k^T}\boldsymbol{\alpha_2} + \gamma_{1,j} + \delta_{1,j} \tag{2}$$

Here, $\mathbf{x_k}$ is a K $\times$ 1 vector of age group indicators, $n_{j,k}$ is the population size for age group k during year j(modification made in this work), $\gamma_{1,i}$ is a spatial random effect term for cancer type i, and $\delta_{1,j}$ is a temporal random effect term for year j. The parameter vectors, $\boldsymbol{\alpha_1}$ and $\boldsymbol{\alpha_2}$, represent the coefficients for the fixed age effects and the interaction terms between age group and log of the population size, respectively. They are parametrized using a reference age group (k^*) as the intercept and an additional effect for each of the remaining age groups. For example, if $k^* = 3$ and $k = 2$, then $\mathbf{x_2} = (0, 1, 1, \ldots, 0)$.

In the second stage of the hurdle model, $\theta_{i,j,k}$ is estimated with a log-linear regression model includes a set of coefficients ($\boldsymbol{\beta}$) for the age group fixed effects, a spatial random effect ($\gamma_{2,i}$), a temporal random effect ($\delta_{2,i}$), and an uncorrelated heterogeneity term ($\sigma_{i,j}$):

$$log(\theta_{i,j,k}) = log(n_{i,j,k}) + \mathbf{x_k^T}\boldsymbol{\beta} + \gamma_{2,i} + \delta_{2,j} + \sigma_{i,j} \tag{3}$$

The coefficients of the fixed effects ($\boldsymbol{\beta}$) are once again parametrized by using the reference age group

3.2 Priors

We assign noninformative or weakly prior so that the posterior is largely shaped by the data. We assign independent diffuse normal priors to the coefficients of the fixed effects. Thus, $\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2} \sim Normal(0, 10 * I)$ and $\boldsymbol{\beta} \sim Normal(0, 10 * I)$, where the normal distributions are parametrized in terms of the precision. And the prior for random effects are as follow:

$$\boldsymbol{\gamma_1}|\tau_{\gamma_1} \sim MVN(0, \tau_{\gamma_1}\mathbf{C})$$
$$\boldsymbol{\gamma_2}|\tau_{\gamma_2} \sim MVN(0, \tau_{\gamma_2}\mathbf{C})$$

where τ_{γ_1} and τ_{γ_2} are overall precision terms and $\mathbf{C}$ is the cancer type covariance matrix follows inverse Wishart distribution, a commonly used distribution that have been used to model the relationship between different deceases in previous researches(Jahan et al., 2020).

$$C \sim IW(V, n)$$

Where V is a fixed symmetric positive definite matrix of size $n \times n$. This can also be written as:

$$\tau_C \sim W(\Gamma, n)$$

Where $\tau_C = C^{-1}$ is the precision matrix for kth region, which is a Wishart prior with degrees of freedom n set equal to the number of cancer types considered in the model and the scale matrix Γ is specified as an identity matrix so that the priors are minimally informative(Jahan et al., 2020). weakly informative prior distributions are specified on the overall precision terms, τ_{γ_1} and τ_{γ_2}. We utilize a Half-Cauchy prior on the overall SDs, so that $1/\sqrt{\tau_{\gamma_1}} \sim Half - Cauchy(10)$ and $1/\sqrt{\tau_{\gamma_2}} \sim Half - Cauchy(10)$, where 10 represents

the scale parameter. Setting the scale parameter to 10 ensures that there is a probability of 0.76 that the overall SDs are below 25. This distribution aligns with our prior beliefs while allowing the data to primarily drive the posterior inference.

Auto-regressive priors are used to capture the correlation between each year's counts and the prior year's counts and allow for temporal smoothing for temporal random effects in equation 2 and 3:

$$\delta_{1,1} \sim Normal(0, \tau_{\sigma_1}(1-\rho_1^2))$$

$$\delta_{1,j} \sim Normal(\delta_{1,j-1} * \rho_1, \tau_{\sigma_1}) \; for \; j = 2, ..., J$$
$$\delta_{2,1} \sim Normal(0, \tau_{\sigma_1}(1-\rho_1^2))$$

$$\delta_{2,j} \sim Normal(\delta_{2,j-1} * \rho_1, \tau_{\sigma_2}) \; for \; j = 2, ..., J$$

Where we set $\tau_{\sigma_1} = \tau_{\sigma_2} = 100$ and $\rho_1 = \rho_2 = 0.4$ based on the best value identified in (Jay et al., 2021).

Finally, the prior distribution for the uncorrelated heterogeneity terms is $\epsilon_{i,j}$ $Normal(0, \tau_\epsilon)$. Once again, we use a weakly informative Half-Cauchy(10) prior on the SD of these random effects.

4. Computation

4.1 Likelihood

The Bayesian hierarchical modeling framework allows us to assume the $Y_{i,j,k}$ are conditionally independent within the likelihood level of the model(Arab, 2015). Therefor, The full likelihood of our model can be expressed as:

$$\begin{aligned}
&\mathcal{L}(\alpha_{1,k}, \alpha_{2,k}, \beta_k, \gamma_{1,i}, \gamma_{2,i}, \delta_{1,j}, \delta_{2,j}, \epsilon_{i,j}, \tau_{\gamma_1}, \tau_{\gamma_2}, \tau_\epsilon, C | y_{i,j,k}, n_{i,j,k}, X, \tau_{\delta_1}, \tau_{\delta_2}) = \\
&\prod_{i,j,k} f(y_{i,j,k}, \alpha_{1,k}, \alpha_{2,k}, \beta_k, \gamma_{1,i}, \gamma_{2,i}, \delta_{1,j}, \delta_{2,j}, \epsilon_{i,j}, \tau_{\gamma_1}, \tau_{\gamma_2}, \tau_\epsilon, C, n_{i,j,k}, X, \tau_{\delta_1}, \tau_{\delta_2}, \Gamma, \nu) = \\
&\prod_{i,j,k} f(y_{i,j,k} | \alpha_{1,k}, \alpha_{2,k}, \beta_k, n_{i,j,k}, \gamma_{1,i}, \gamma_{2,i}, \delta_{1,j}, \delta_{2,j}, \epsilon_{i,j}) f(\alpha_{1,k}) f(\alpha_{2,k}) f(\beta_k) f(\gamma_{1,i} | \tau_{\gamma_1}, C) \\
&f(\gamma_{2,i} | \tau_{\gamma_2}, C) f(\delta_{1,j}) f(\delta_{2,j}) f(\epsilon_{i,j} | \tau_\epsilon) f(\tau_\epsilon) f(\tau_{\gamma_1}) f(\tau_{\gamma_2}) f(C) \sim \\
&P(Y_{i,j,k} = y_{i,j,k}) * N(\alpha_{1,k} | 0, 1) * N(\alpha_{2,k} | 0, 1) * N(\beta_k | 0, 1) * N(\gamma_{1,i} | 0, \tau_{\gamma_1} * C[i,i]) * \\
&N(\gamma_{2,i} | 0, \tau_{\gamma_1} * C[i,i]) * N(f(\delta_{1,j}) | f(\delta_{1,j-1}), 100) * N(f(\delta_{2,j}) | f(\delta_{2,j-1}), 100) * N(\epsilon_{i,j} | 0, \tau_\epsilon) * \\
&Cauchy(1/\sqrt{\tau_{\gamma_1}} | 0, 10) * Cauchy(1/\sqrt{\tau_{\gamma_2}} | 0, 10) * Cauchy(1/\sqrt{\tau_\epsilon} | 0, 10) * IW(C | \Gamma^{-1}, M)
\end{aligned} \tag{4}$$

Where Γ^{-1} is the inverse of an identical matrix and M is the number of cancer types. Recall $1/\sqrt{\tau_{\gamma_1}}$, $1/\sqrt{\tau_{\gamma_2}}$ and $1/\sqrt{\tau_\epsilon}$ follow Half-Cauchy distribution instead of Cauchy distribution. But since $1/\sqrt{\tau_{\gamma_1}}$, $1/\sqrt{\tau_{\gamma_2}}$ and $1/\sqrt{\tau_\epsilon}$ are always bigger than 0 in our cases, the only difference between Half-Cauchy PDF and Cauchy PDF is constant and can be absorbed.

The above likelihood is tedious and difficult to compute. However, recall that both $\pi_{i,j,k}$ and $\theta_{i,j,k}$ are combinations of other parameters. Therefor, the likelihood can be written as:

$$\begin{aligned}&\mathcal{L}_1(\alpha_{1,k}, \alpha_{2,k}, \beta_k, \gamma_{1,i}, \gamma_{2,i}, \delta_{1,j}, \delta_{2,j}, \epsilon_{i,j}, \tau_{\gamma_1}, \tau_{\gamma_2}, \tau_\epsilon, C | y_{i,j,k}, n_{i,j,k}, X, \tau_{\delta_1}, \tau_{\delta_2}) = \\ &\mathcal{L}_2(\pi_{i,j,k}, \theta_{i,j,k}, \tau_{\gamma_1}, \tau_{\gamma_2}, \tau_\epsilon, C | y_{i,j,k}, n_{i,j,k}, X, \tau_{\delta_1}, \tau_{\delta_2})\end{aligned} \tag{5}$$

And,

$$\mathcal{L}_2(\pi_{i,j,k}, \theta_{i,j,k}, \tau_{\gamma_1}, \tau_{\gamma_2}, \tau_\epsilon, C|.) = \prod_{i,j,k} P(y_{i,j,k}|\pi_{i,j,k}, \theta_{i,j,k}) f(\pi_{i,j,k}, \theta_{i,j,k}) f(\tau_{\gamma_1}, \tau_{\gamma_2}, \tau_\epsilon, C) \tag{6}$$

Where,

$$\begin{aligned} f(\pi_{i,j,k}, \theta_{i,j,k}) &= f(\pi_{i,j,k}) f(\theta_{i,j,k}) \sim Constant \\ f(\tau_{\gamma_1}, \tau_{\gamma_2}, \tau_\epsilon, C) &= f(\tau_{\gamma_1}) f(\tau_{\gamma_2}) f(\tau_\epsilon) f(C) \\ &\sim Cauchy(1/\sqrt{\tau_{\gamma_1}}|0, 10) * Cauchy(1/\sqrt{\tau_{\gamma_2}}|0, 10) * \\ &\quad Cauchy(1/\sqrt{\tau_\epsilon}|0, 10) * IW(C|\Gamma^{-1}, M) \end{aligned} \tag{7}$$

Recall $P(y_{i,j,k}|\pi_{i,j,k}, \theta_{i,j,k})$ is a 2 stage model, and when $y_{i,j,k} = 0$, $y_{i,j,k}$ is only depend on $\pi_{i,j,k}$. Therefor, the likelihood can also be written as:
When $y_{i,j,k} = 0$,

$$\mathcal{L}_2 \sim \prod_{i,j,k} (1 - \pi_{i,j,k}) * N(\gamma_{1,i}|0, \tau_{\gamma_1} * C[i,i]) * Cauchy(1/\sqrt{\tau_{\gamma_1}}|0, 10) * IW(C|\Gamma^{-1}, M) \tag{8}$$

When $y_{i,j,k} > 0$,

$$\begin{aligned}\mathcal{L}_2 \sim \prod_{i,j,k} &P(y_{i,j,k}) * N(\gamma_{1,i}|0, \tau_{\gamma_1} * C[i,i]) * Cauchy(1/\sqrt{\tau_{\gamma_1}}|0, 10) * IW(C|\Gamma^{-1}, M) * \\ &N(\gamma_{2,i}|0, \tau_{\gamma_2} * C[i,i]) * Cauchy(1/\sqrt{\tau_{\gamma_2}}|0, 10) * N(\epsilon_{i,j}|0, \tau_\epsilon) * Cauchy(1/\sqrt{\tau_\epsilon}|0, 10)\end{aligned} \tag{9}$$

4.2 Algorithm

Even though $\mathcal{L}_2$ is much more simpler and easy to compute compare to $\mathcal{L}_1$, it's still not conjugated and thus can not use Gibbs sampler to draw posterior. Instead, we use Stan which will use No-U-Turn sampler (NUTS) to draw posterior. We sampled 2 chains with 1000 iterations and 500 iterations for warm-up. Computation is ran on R and Rstan are used to link R with Stan. Both the code of R script and Stan code are provided in Section 5.2. Reference age group in $\mathbf{X}$ is set as 40-45 as this age group has the average incidence counts in most cancer types in our dataset.

4.3 Age-specific and age-standardized rate

In this section, we derive the formula used to compute age specific and standardized rate from our model's outputs. We will calculate the expected number of incidence in each year, age group and cancer type, dived it by the population and times 100,000 to get the corresponding age specific rate. And base on the equation of age standardized rate given above to compute the corresponding age standardized rate.

We first derived the expectation of $Y_{i,j,k}$. By definition $E[Y_{i,j,k}] = \sum_{y_{i,j,k}=0}^{\infty} y_{i,j,k} \times P(Y_{i,j,k} = y_{i,j,k})$, which can also be written as:

$$\begin{aligned}
E[Y_{i,j,k}] &= \sum_{y_{i,j,k}=0}^{\infty} y_{i,j,k} * (1-\pi_{i,j,k})^{I(y_{i,j,k}=0)} * \left[\pi_{i,j,k} * \frac{\theta_{i,j,k}^{y_{i,j,k}} exp(-\theta_{i,j,k})}{y_{i,j,k}!(1-exp(-\theta_{i,j,k})}\right]^{I(y_{i,j,k}>0)} \\
&= 0 + \sum_{y_{i,j,k}=1}^{\infty} y_{i,j,k} * \pi_{i,j,k} * \frac{\theta_{i,j,k}^{y_{i,j,k}} exp(-\theta_{i,j,k})}{y_{i,j,k}!(1-exp(-\theta_{i,j,k})} \\
&= \sum_{y_{i,j,k}=0}^{\infty} y_{i,j,k} * \pi_{i,j,k} * \frac{\theta_{i,j,k}^{y_{i,j,k}} exp(-\theta_{i,j,k})}{y_{i,j,k}!(1-exp(-\theta_{i,j,k})} \\
&= \frac{\pi_{i,j,k}}{1-exp(-\theta_{i,j,k})} * \sum_{y_{i,j,k}=0}^{\infty} y_{i,j,k} * \frac{\theta_{i,j,k}^{y_{i,j,k}} exp(-\theta_{i,j,k})}{y_{i,j,k}!} \\
as\ E[Poisson(\theta_{i,j,k})] &= \frac{\theta_{i,j,k}^{y_{i,j,k}} exp(-\theta_{i,j,k})}{y_{i,j,k}!} = \theta_{i,j,k} \\
E[Y_{i,j,k}] &= \frac{\pi_{i,j,k}}{1-exp(-\theta_{i,j,k})} * \theta_{i,j,k}
\end{aligned} \tag{10}$$

The age-specific rate then can be computed by:

$$R_{i,j,k} = \left(\frac{\pi_{i,j,k}}{1-exp(-\theta_{i,j,k})} * \theta_{i,j,k}]/n_{i,j,k}\right) * 100000$$

Finally, age standardized rate can be computed by:

$$I(ASR/E) = \{\sum_{k} R_{i,j,k} P_k\}_k / \sum P_k$$

5. Data Analysis

In this section, we use our Bayesian model to predict the age-specific rates and age-adjusted rate in England from year 2013 to 2017 based on the datasets mentioned in 2.

5.1 Model's results

After 1000 iterations, all the parameters have converged (shown in Figure 1 , further details of the convergence regarding other parameters can be seen in Section 5.2).

To evaluate our model's prediction, we compared the models predicted expected incidence counts with the observations. Result(Figure 2) shows that overall our model can estimate the data but tend to underestimate in most cases. It also worth to notice that the model's prediction sometimes deviate from the true value (observations) a lot, indicating that the prior we set may still be too informative or because the model's structure need to be further refined.

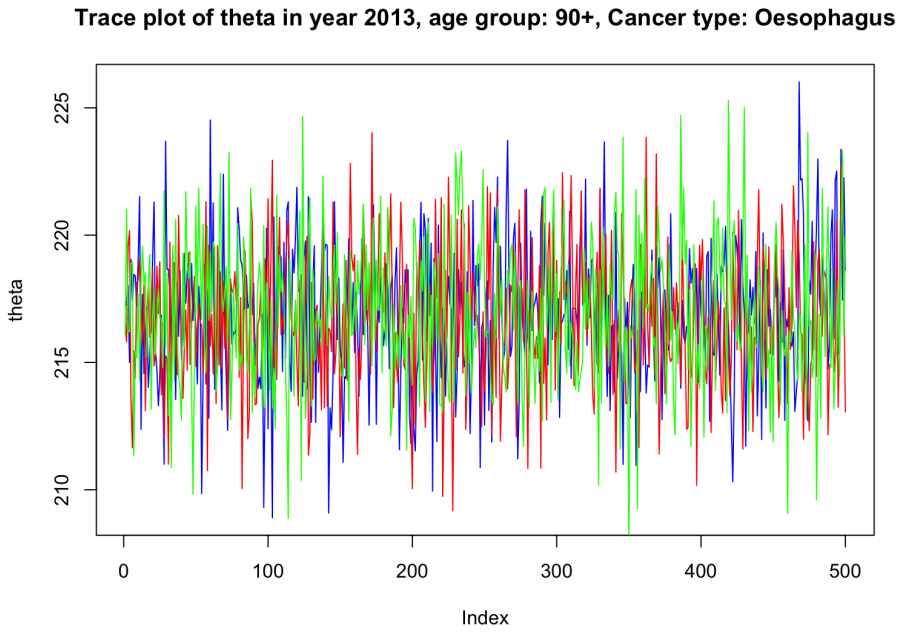

Figure 1: Trace plot for theta

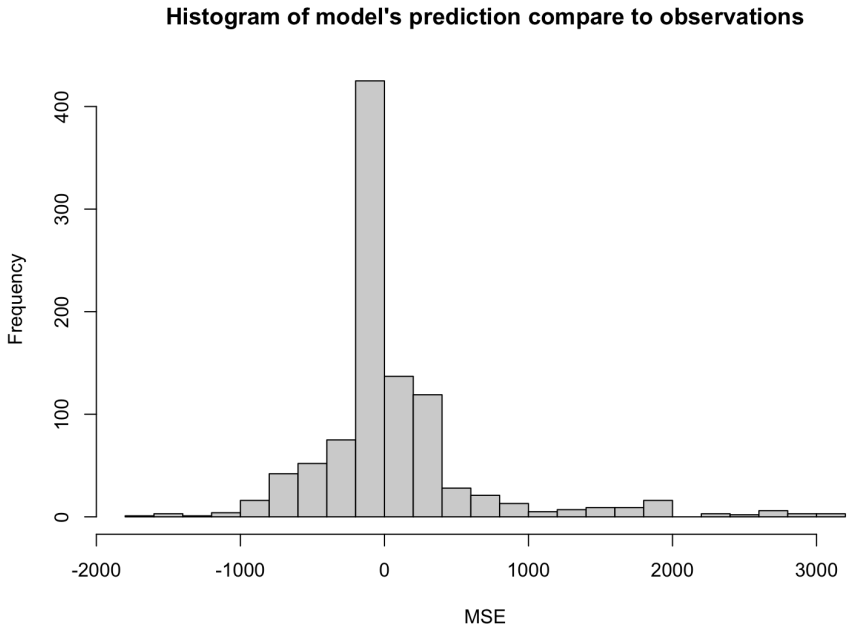

Figure 2: Model's prediction v.s. Observations

The predicted age-standardized rate and the age-standardized rate using direct standardization method are shown in Table 5.2. The office of statistics did not provide the age-standardized rate for Kidney and Urinary Tract cancer, only the population's average age-standardized rate. However, our model's results fall between the the interval of observed male rate and female rate, indicating that our model can capture the underlying cancer pattern at the population level. Our model's results vary more across years compared to the direct standardization method. In comparison to all other predicted results, the model appears to underestimate the rate for lung cancer. Over the 5 years, the predicted age-standardized rate for lung cancer is consistently lower than the lowest direct age-standardized rate or close to it. No overestimates were observed. This finding aligns with our previous results, suggesting that the model may underestimate the incidence counts. The reason for this discrepancy is unknown and requires further investigation, possibly due to incorrect priors or other factors. In our analysis, the model's results do not differ significantly from the observations and direct standardization method, indicating that both methods yield accurate results. The impact of our model in situations where the direct standardization method may fail (e.g., low incidence counts) needs to be explored further.

5.2 Conclusion

In this work, we modified an existing Bayesian hierarchical model, change the prior corresponds to the random effect of region into explaining the random effect of different cancer types in order to predict age-standardized rate. New prior were added so that we can correctly modeled the covariance of different cancer types' incidence rate. We then applied our model on a dataset which researchers use the same dataset to derive age-standardized rate through direct standardization method. The difference between 2 methods exists but are minor. This indicate that overall, our model is precise, but minor modification can still be made to achieve better precision. Further development of the model can include adding new structure that enable randomness among genders.

Table 2: Model predicted age standardised rate and direct age standardised rate

	2013		2014	
	Predicted	Observed(male, female)	Predicted	Observed(male, female)
Oesophagus	13.11	22.2, 8.9	12.37	21.9, 8.8
Stomach	10.31	17.3, 7.3	9.2	16.0, 6.7
Colorectal	65.4	86.8, 56.2	61.71	84.5, 56.4
Lung	69.59	92.5, 64.4	66.27	91.6, 65.2
Breast	80.6	1.3, 169.8	80.31	1.5, 173.4
Kidney and Urinary Tract	19.58		20.44	
Bladder	16.06	30.3, 8.9	14.59	29.1, 8.2
Non-Hodgkin's lymphoma	22.17	27.6, 19.9	22.29	28.2, 19.4
Myeloma	8.52	12.4, 7.6	7.72	11.7, 7.7
Leukaemia	15.76	21.5, 11.8	15.82	21.5, 11.9
	2015		2016	
	Predicted	Observed(male, female)	Predicted	Observed(male, female)
Oesophagus	12.05	22.8, 8.7	12.23	22.6, 8.3
Stomach	9.17	16.0, 7.0	8.81	15.4, 6.4
Colorectal	61.44	84.6, 56.8	60.2	84.4, 55.4
Lung	65.56	89.4, 65.6	66.5	89.8, 65.5
Breast	78.51	1.4, 170.2	75.2	1.3, 167.9
Kidney and Urinary Tract	20.17		20.24	
Bladder	14.25	28.1, 8.3	14.53	27.3, 8.2
Non-Hodgkin's lymphoma	21.9	27.5, 19.6	22.19	28.0, 19.8
Myeloma	7.62	11.8, 7.2	7.74	11.7, 7.4
Leukaemia	16.2	21.4, 12.5	15.52	20.8, 11.7
	2017			
	Predicted	Observed(male, female)		
Oesophagus	10.07	22.2, 8.1		
Stomach	5.66	14.4, 6.2		
Colorectal	24.3	81.9, 54.9		
Lung	49.86	86.9, 67.0		
Breast	15.67	1.3, 166.7		
Kidney and Urinary Tract	7.08			
Bladder	8.46	27.6, 8.2		
Non-Hodgkin's lymphoma	7.71	27.6, 19.5		
Myeloma	4.35	12.4, 7.5		
Leukaemia	7.6	21.1, 12.2		

Appendix A.

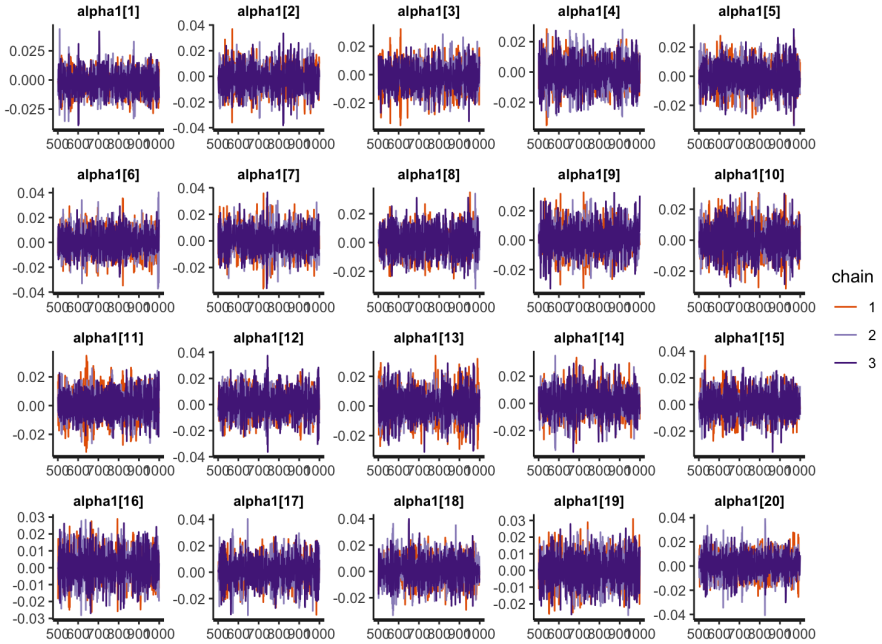

Figure 3: Trace plot for alpha 1

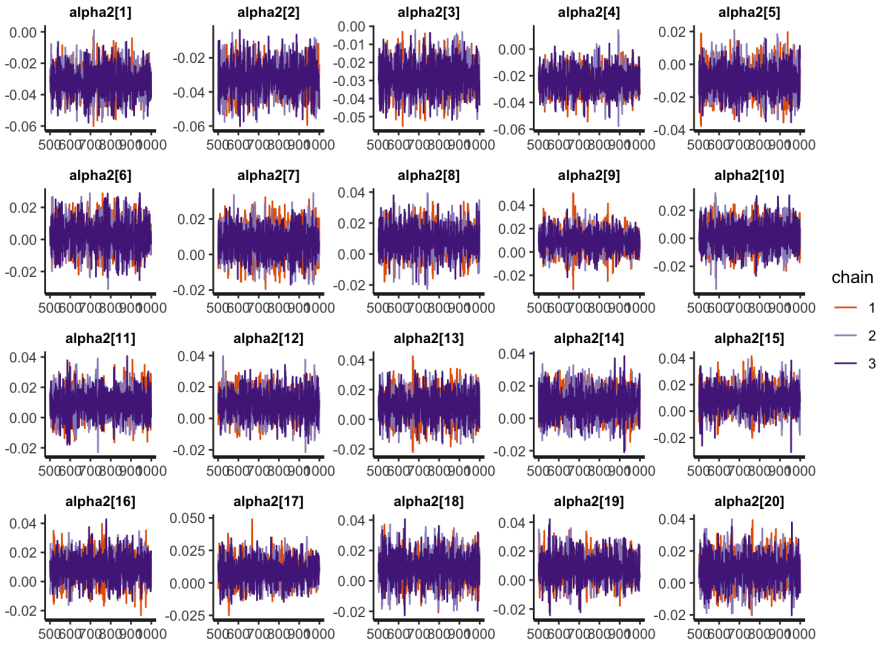

Figure 4: Trace plot for alpha 2

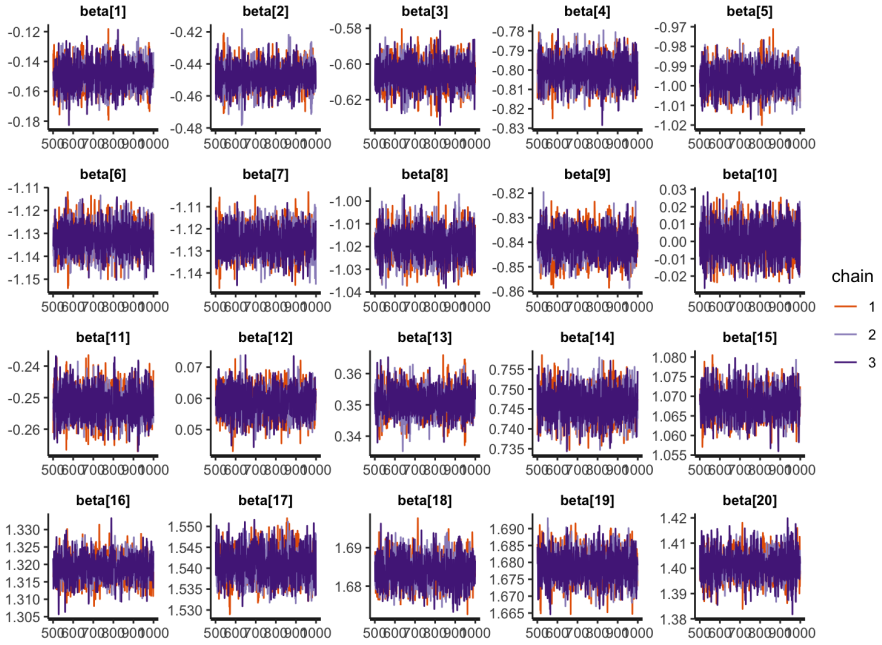

Figure 5: Trace plot for beta

Appendix B.

Code availability: https://github.com/ZhaoxiangD/2530final

References

Ali Arab. Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *International journal of environmental research and public health*, 12(9):10536–10548, 2015.

Mehreteab Aregay, Andrew B Lawson, Christel Faes, Russell S Kirby, Rachel Carroll, and Kevin Watjou. Zero-inflated multiscale models for aggregated small area health data. *Environmetrics*, 29(1):e2477, 2018.

PA Buescher. Age-adjusted death rates. statistical primer no. 13. *State Center for Health Statistics, North Carolina Department of Health and Human Services*, 2008.

Sarah Caul and John Broggio. Cancer registration statistics, england 2017. *Off. Natl. Stat*, pages 1–16, 2019.

Isabel dos Santos Silva. *Cancer epidemiology: principles and methods*. IARC, 1999.

Paul Elliott and Daniel Wartenberg. Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives*, 112(9):998–1006, June 2004. ISSN 1552-9924. doi: 10.1289/ehp.6735. URL `http://dx.doi.org/10.1289/ehp.6735`.

Farzana Jahan, Earl W. Duncan, Susana M. Cramb, Peter D. Baade, and Kerrie L. Mengersen. Multivariate bayesian meta-analysis: joint modelling of multiple cancer types using summary statistics. *International Journal of Health Geographics*, 19(1), October 2020. ISSN 1476-072X. doi: 10.1186/s12942-020-00234-0. URL `http://dx.doi.org/10.1186/s12942-020-00234-0`.

Melissa Jay, Jacob Oleson, Mary Charlton, and Ali Arab. A bayesian approach for estimating ageâadjusted rates for lowâprevalence diseases over space and time. *Statistics in Medicine*, 40(12):2922–2938, March 2021. ISSN 1097-0258. doi: 10.1002/sim.8948. URL `http://dx.doi.org/10.1002/sim.8948`.

Richard J Klein. *Age adjustment using the 2000 projected US population*. Number 20. Department of Health & Human Services, Centers for Disease Control and . . . , 2001.

Ali H Mokdad, Katherine Ballestros, Michelle Echko, Scott Glenn, Helen E Olsen, Erin Mullany, Alex Lee, Abdur Rahman Khan, Alireza Ahmadi, Alize J Ferrari, et al. The state of us health, 1990-2016: burden of diseases, injuries, and risk factors among us states. *Jama*, 319(14):1444–1472, 2018.

American Cancer Society. *Cancer prevention and early detection*. American Cancer Society, 2002.