Cancer Epidemiology 37 (2013) 593–600

Contents lists available at SciVerse ScienceDirect

Cancer Epidemiology

The International Journal of Cancer Epidemiology, Detection, and Prevention

journal homepage: www.cancerepidemiology.net

# Bayesian area–age–period–cohort model with carcinogenesis age effects in estimating cancer mortality

Zhiheng Xu [a,*], Vicki S. Hertzberg [b]

[a] The Food and Drug Administration, Silver Spring, MD 20993, USA
[b] Emory University, The Rollins School of Public Health, Department of Biostatistics and Bioinformatics, Atlanta, GA 30322, USA

ARTICLE INFO

*Article history:*
Received 23 October 2012
Received in revised form 1 July 2013
Accepted 4 July 2013
Available online 26 July 2013

*Keywords:*
AAPC
Bayesian
Prior
Spatial
Temporal
Nonidentifiability

ABSTRACT

*Objective*: Area–age–period–cohort (AAPC) model has been widely used in studying the spatial and temporal pattern of disease incidence and mortality rates. However, lack of biological plausibility and ease of interpretability on temporal components especially for age effects are generally the weakness of AAPC models. We develop a Bayesian AAPC model where carcinogenesis age effect is incorporated to explain age effects from the underlying disease process. An autoregressive prior structure and an arbitrary linear constraint are used to solve the nonidentifiability issues. *Methods*: Two multistage carcinogenesis models are employed to derive the hazard functions to substitute the age effects in the AAPC models. The Iowa county-wide lung cancer mortality data are used for the model fitting and Deviance Information Criteria (DIC) is used for model comparison. *Results*: Our study shows that conventional AAPC model (DIC = 19,231.30), AAPC model with Armitage–Doll age effect (DIC = 19,233.00) and with two-stage clonal expansion (TSCE) age effect (DIC = 19,234.70) achieved the similar DIC values which indicated consistent model fitting among three models. The spatial pattern shows that the high spatial effects are clustered in the south of Iowa and also in largely populated areas. The lung cancer mortality rate is continuously declining by birth cohorts while increasing by the calendar period until 2000–2004. The age effects show an increasing pattern over time which can be easily explained by Armitage–Doll carcinogenesis model since we assume a log-linear relationship between age and hazard function. *Conclusions*: Our finding suggests that the proposed Bayesian AAPC model can be used to replace the conventional AAPC model without affecting model performance while providing a more biological sound approach from the underlining disease process.

Published by Elsevier Ltd.

## 1. Introduction

Disease mapping is an important topic in epidemiology to study the space and time variation for the risk of disease. Many general or more heavily parameterized Bayesian models have been proposed to study the spatio-temporal mappings of disease rates [1,2]. With the development of open-source software (such as WinBUGS or R), the research in fully Bayesian disease mapping is expanding. For example, Bernardinelli et al. [3] extends the Besag et al. [4] BYM model to allow area specific random effects for temporal trends. Knorr-Held [5] tests different spatial–temporal interactions in the model. Ugarte et al. [6] evaluate the performance of different spatio-temporal Bayesian models in disease mapping. Wakefield [7] studies the spatial regression with count data.

Disease incidence and mortality data may vary considerably among different geographical regions. Areas with a small population could result in an extreme observation of incidence and mortality due to the small population at risk. Therefore, to consider the high sampling variability in small areas when estimating disease incidence and mortality rates across each region, we usually add a weight matrix to the set of model parameters to smooth variation among neighboring areas and improve the estimation in the small regions by borrowing strength from their adjacent regions [8]. In an analysis of lung cancer rates in Tuscany, Lagazio et al. [9] proposed a full spatio-temporal Bayesian model which include main effects of area, age, period and cohort as well as area–period and area–cohort interactions. Gaussian first-order and second-order random walk priors (RW1, RW2) were given to age, period and cohort effects in Lagazio's full model. To better explain the biological meaning of age effect in determining cancer mortality rate, multistage carcinogenesis models will be considered in AAPC models to replace the main age effects. It is based on the assumption of fundamental role of age in determining the cancer incidence rates and subsidiary roles of period and cohort in

* Corresponding author at: 10903 New Hampshire Avenue, Silver Spring, MD 20993-0002, USA. Tel.: +1 301 796 6094; fax: +1 301 847 8123.
*E-mail address:* zhiheng.xu@fda.hhs.gov (Z. Xu).

1877-7821/$ – see front matter. Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.canep.2013.07.002

modulating the age effect [10]. Besides modeling cancer incidence rates, many researchers have used carcinogenesis models to study cancer mortality rates. In their ground-breaking paper, Armitage and Doll [11] assumed that mortality gave a good indication of incidence and treated the data as if they referred to age specific incidence rates. Fisher and Holloman [12] and Nordling [13] also found that the cancer death rate increased proportionally with the sixth power of the age from analyzing cancer statistics data from the U.S. and several European countries. Hazelton et al. [14] and Holford and Levy [15] applied multistage carcinogenesis model in studying lung cancer mortality rates in the U.S.

The AAPC model provides a general framework to jointly study the evolution in time and the spatial pattern of the risk of disease [9]. The interaction terms over area can reduce the identifiability burden in the standard AAPC model [16,17]. Gaussian RW1 and RW2 structures on the age, period and cohort effects can improve model estimation and prediction of future mortality rates [18,19].

The identifiability problem is well-known for this type of model since three temporal effects (age, period and cohort) are linearly dependent and two spatial effects are simultaneously included in the analysis. Many approaches have been studied but none of them is adequately solve this identifiability problem entirely [17,20]. We adapt an autoregressive prior structure and an arbitrary linear constraint for the temporal effects in order to control the identifiability issues in model fitting. In addition, the choice of carcinogenesis age effects could help us fitness the inherent non-identifiability problem associated with APC approach [10,21]. In this study, we develop a new Bayesian AAPC model where multistage carcinogenesis models are introduced into the AAPC model to incorporate more biological meaning of the age effects in studying the spatio-temporal pattern of cancer mortality rates. The prior means of age effects in the AAPC model are replaced by the log transformation of hazard functions derived from the Armitage–Doll multistage carcinogenesis model and the TSCE model. The proposed AAPC model is also compared with the conventional AAPC model where age effects are assigned as RW1 or RW2 priors in fitting cancer mortality data. Model selection procedures (DIC) are implemented to compare the performance of several alternative models.

## 2. Method

### 2.1. Data source

The Surveillance, Epidemiology and End Results (SEER) database at National Cancer Institute (NCI) provides cancer incidence and mortality information on the U.S. population. SEER collects data on cancer cases from various locations and sources throughout the U.S. since 1973 with a limited amount of registries and continues to expand to include even more areas and demographics today. Seventeen SEER registry sites are listed which covers approximately 26% U.S. population. Iowa county-wide lung cancer mortality data from SEER are used to evaluate the extended AAPC models proposed in this paper.

### 2.2. Model development

A full area–age–period–cohort (AAPC) model is used to study the spatio-temporal pattern of disease risk [9]. The model incorporates the main effect of area, age, period and cohort, and interaction terms such as the area–cohort and area–period interactions. The model is as follows:

$$\log(\lambda_{iap}) = \nu_i + \mu_i + \theta_a + \gamma_p + \delta_c + \varphi_{ip} + \varphi_{ic},$$

where $\lambda_{iap}$ is the relative risk for the $a$th age group and the $p$th calendar period in the $i$th area, $\nu_i$ is the unstructured spatial term for the spatial heterogeneity effects, $\mu_i$ is the structured spatial term to incorporate spatial clustering effect, $\theta_a$, $\gamma_p$ and $\delta_c$ are the age, period, and cohort main effects, $\varphi_{ip}$ is the space–period interaction and $\varphi_{ic}$ is the space–cohort interaction. The space–age interaction is not considered in this model since we assume homogeneous age effects across different regions. Although smoking status is generally considered as an important factor in predicting lung cancer mortality, we did not include it in this model due to lack of county level prevalence rates of smoking from SEER database. Other national health databases such as Behavioral Risk Factor Surveillance System (BRFSS), either do not provide information on tobacco use at the county level, or cannot be directly used because of different year intervals for age categories and less accurate trend data in prevalence from previous years [22].

To emphasize the focus of this paper, we only discuss the carcinogenesis priors for age effects in Section 2. The prior choices for spatial effects and temporal effects (period and cohort) are listed in the appendix.

Conditional autoregressive priors including Gaussian first-order and second-order random walk priors (RW1, RW2) are commonly used to all three temporal effects. However, lack of biological plausibility and ease of interpretability on temporal components especially for age effects are generally the weakness of using such priors [10,23]. To explain the age effects from the underline disease process, we integrate multistage carcinogenesis into an AAPC model by specification of biologically meaningful priors for age effects. A noninformative prior is assigned to the age effect $\theta_a$ at age group $a$ as

$$\theta_a \sim N(\bar{\theta}_a, \tau),$$

where the mean of normal prior $\bar{\theta}_a$ is replaced by the hazard function derived from carcinogenesis model as $\bar{\theta}_a = \log(h(t_a))$. To obtain the hazard function $h(t_a)$ at age $t_a$, we investigate two widely used carcinogenesis models – Armitage–Doll multistage carcinogenesis model and Moolgavkar's two-stage clonal expansion (TSCE) model.

In the Armitage–Doll multistage carcinogenesis model, the normal stem cells are assumed to undergone multi-stage transformations in developing into cancer cells. Armitage and Doll fixed the probabilities of the occurrence of each stage transition and concluded that the hazard function or the occurrence of first $s-1$ changes will be proportional to the age with a power of $s-1$ [11], i.e.,

$$h(t_a) \propto t_a^{s-1},$$

and the logarithm transformation of hazard function can be written as

$$\log(h(t_a)) = c + (s-1)\log(t_a).$$

Hyperpriors are assigned to Armitage–Doll carcinogenesis parameters $c$ and $s$ as $c \sim N(0, \tau)$ and $s \sim N(6, \tau)$. The precision term $\tau$ in the above normal priors is defined as $\tau \sim Gamma(0.001, 0.001)$.

In the Armitage–Doll multistage carcinogenesis model, the stage transition probability is determined and fixed. However, cancer formation is a stochastic process and cancer growth is uncontrollable and exponential. Therefore, Moolgavkar and his colleagues proposed a two-stage clonal expansion carcinogenesis model where the carcinogenesis process is treated as a birth-death type stochastic process [24]. A normal cell can divide into two normal cells, or die or differentiate, or initiate to intermediate cell. The clonal expansion is occurred at the stage when one

intermediate cell expands to multiple intermediate cells and each intermediate cell has the same mutation rate to become a malignant cell. With the assumption of constant rates, Moolgavkar has obtained a closed form solution for hazard function which is a nonlinear function of age [24–26], i.e.,

$$h(t_a) = r\,pq\,\frac{e^{-qt_a} - e^{-pt_a}}{qe^{-pt_a} - pe^{-qt_a}},$$

where $p$ and $q$ are the roots of quadratic equations with the rate of initiation, $\mu_1$, the rates of division $\alpha$, and death $\beta$, of initial cells, and the rate of malignant conversion, $\mu_2$. Based on the literature review [24–26], we estimate that $\mu_1$and $\mu_2$ are at the magnitude of $10^{-8}$ and much smaller than the rates of division $\alpha$, and death $\beta$. Therefore, we can calculate the approximate intervals for parameters $p$, $q$, and $r$ and their noninformative priors are specified as $r \sim Uniform(0, 10^{-5})$, $p \sim Uniform(-0.2, 0)$, $q \sim Uniform(0, 10^{-5})$.

Markov chain Monte Carlo (MCMC) methods have been used to sample from the posterior density. Due to the non-identifiability issues in the model parameters, certain constraints (see Appendix) are added in the model to improve numerical stability and mixing [27]. Area, age, period and cohort effects are each adjusted by subtracting their respective means [27]. Sensitivity analysis has been implemented to examine the robustness of the posterior inference and relative risk estimates to the prior specifications [7]. WinBUGS software is used to derive the posterior distribution about model parameters from 10,000 iterations after a burn-in of 1000 iterations. The convergence plots of model parameters are provided along with convergence diagnostics. The posterior estimates of main effects and interactions in the AAPC model are summarized by the mean and 95% highest posterior density (HPD) derived from posterior samples. Statistical software R and WinBUGS are used in this study.

## 3. Results

Lung cancer mortality records in the state of Iowa are retrieved from SEER for this study. The death counts are aggregated by county, age group, and period. The state of Iowa has 99 counties and the county adjacent matrix is obtained using the Geographic Information Software – ArcGIS. Since the Armitage–Doll model fit the cancer data well for the age between 25 and 74, we drop the observations with age less than 24 or greater than 74. The total number of age group is 10. To calculate the mortality rate, we retrieve population data from the U.S. Census website per age group in each county of Iowa in the 1980s, 1990s, and 2000s. Five consecutive calendar years were grouped into one period group and six period groups were formed (1980–1984, 1985–1989, 1990–1994, 1995–1999, 2000–2004, 2005–2008). Table 1 describes age and period specific rates (×100,000) and number of cases. In Table 1, at different period, we observe the lung cancer mortality rate consistently increases with the age group which is supported by the carcinogenesis model. In most age groups, we also notice their period effect is also going upward until 2000–2004 where the decreasing trend is shown. To further illustrate the geographic variability in lung cancer mortality rates, we display Iowa county-wide lung cancer mortality rate in three periods (1980–1984, 1990–1994, 2005–2008) for three age groups (45–49, 55–59, 70–74) in Fig. 1. It is difficult to visually detect the spatial pattern in affecting lung cancer mortality rates. Therefore, we apply the AAPC model to study the spatial and temporal patterns in determining the lung cancer mortality rate in Iowa.

To compare different models in fitting lung cancer mortality in Iowa, we use the deviance information criterion (DIC) which is the posterior average of the deviance plus a measure of complexity. Table 2 displays the DIC among different models. The models listed in Table 2 include area–age (AA), area–age–period (AAP), area–age–cohort (AAC), area–age–period and area–period interaction (AAP + AP), area–age–cohort and area–cohort interaction (AAC + AC), area–age–period–cohort (AAPC), area–age–period–cohort and area–period interaction (AAPC + AP), area–age–period–cohort and area–cohort interaction (AAPC + AC), area–age–period–cohort and two interactions (area–period and area–cohort) (AAPC + AP + AC). The smaller the DIC value, the better the model fits the data.

Compared with the DIC values derived from the conventional Bayesian AAPC models where all temporal effects are taking autoregressive priors, we do not see much difference on the improvement of DIC values when the multistage carcinogenesis models are incorporated into AAPC model. However, the age effect in the Bayesian extended AAPC model has a more sound biological meaning since we replace it with the hazard function from the carcinogenesis model. The temporal evolution of age effects derived from the carcinogenesis model clearly demonstrates the association between age and mortality rate. With the introduction of the carcinogenesis model into AAPC model, we reduce the complexity of any possible linear or nonlinear age effects in determining the mortality rate, except for those derived from the Armitage–Doll model and the TSCE model. With similar model fitting criteria (DIC), the extended AAPC model outperforms the conventional AAPC model due to its strong biological meaning of age effects.

The convergence of model estimates for the Bayesian AAPC model was satisfactory through visually checking convergence plots. Figs. 2 and 3 show the convergence plot for posterior estimates from Armitage–Doll and TSCE models. In the sensitivity analysis, three different prior choices for the variance components for the unstructured spatial effects were used and the posterior estimates for the parameter of number of stage in Armitage–Doll carcinogenesis model were compared in Table 3. The posterior means were approximately close.

The model we chose to first fit Iowa lung cancer mortality data is AAPC using Armitage–Doll carcinogenesis model due to its low DIC value and good convergence. We further investigate other model fitting criteria, such as Akaike information criterion (AIC)

**Table 1**
Age–period specific mortality rates (×100, 000) and number of deaths. Lung cancer, Iowa, 1980–2008.

| | Period | | | | | |
|---|---|---|---|---|---|---|
| Age | 1980–1984 | 1985–1989 | 1990–1994 | 1995–1999 | 2000–2004 | 2005–2008 |
| 25–29 | 0.50 (6) | 0.70 (8) | 0.52 (5) | 0.55 (5) | 0.69 (6) | 0.67 (5) |
| 30–34 | 1.02 (11) | 1.28 (14) | 1.75 (19) | 1.78 (17) | 1.31 (12) | 1.73 (12) |
| 35–39 | 4.22 (36) | 3.91 (39) | 5.30 (58) | 5.51 (61) | 4.65 (47) | 2.27 (17) |
| 40–44 | 16.50 (120) | 12.73 (103) | 9.74 (98) | 12.60 (140) | 14.24 (161) | 8.83 (73) |
| 45–49 | 37.39 (249) | 35.25 (241) | 34.08 (271) | 26.87 (266) | 33.00 (362) | 22.81 (206) |
| 50–54 | 77.30 (533) | 83.27 (524) | 73.61 (497) | 62.65 (494) | 58.00 (562) | 45.03 (386) |
| 55–59 | 137.41 (976) | 156.62 (1008) | 137.11 (848) | 137.03 (911) | 116.97 (888) | 80.02 (594) |
| 60–64 | 203.54 (1363) | 227.81 (1482) | 232.45 (1446) | 223.90 (1348) | 212.21 (1310) | 150.32 (847) |
| 65–69 | 260.07 (1556) | 295.36 (1797) | 335.60 (2009) | 316.06 (1755) | 319.67 (1708) | 233.29 (1049) |
| 70–74 | 300.95 (1496) | 351.17 (1801) | 372.57 (1969) | 417.04 (2168) | 406.16 (2029) | 303.45 (1154) |

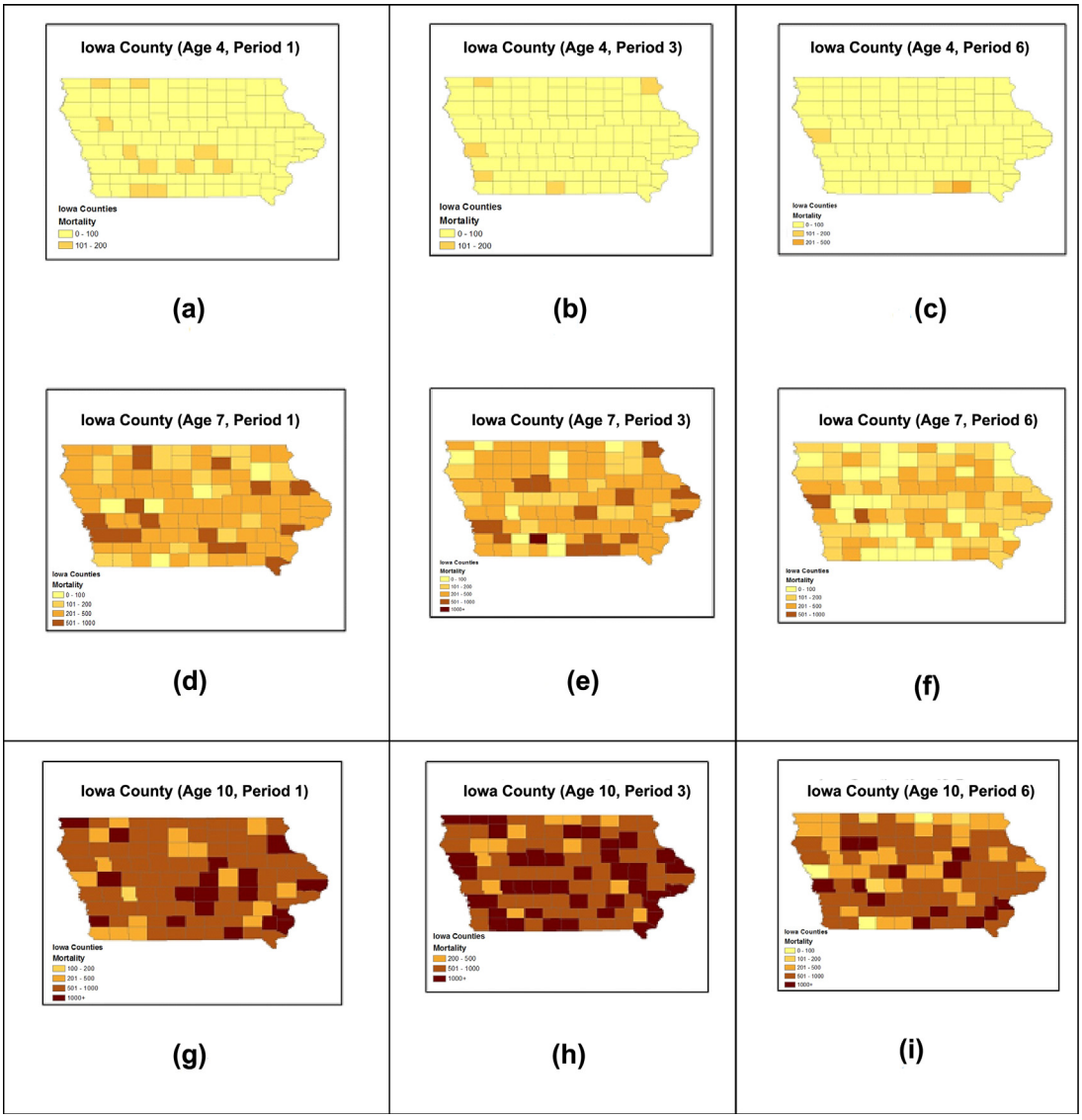


**Fig. 1.** Lung cancer mortality rate in Iowa from 1980 to 2008 for three age groups during three periods. (a) age: 45–49 ($a = 4$), period: 1980–1984 ($p = 1$); (b) age: 45–49 ($a = 4$), period: 1990–1994 ($p = 3$); (c) age: 45–49 ($a = 4$), period: 2005–2008 ($p = 6$); (d) age: 55–59 ($a = 7$), period: 1980–1984 ($p = 1$); (e) age: 55–59 ($a = 7$), period: 1990–1994 ($p = 3$); (f) age: 55–59 ($a = 7$), period: 2005–2008 ($p = 6$); (g) age: 70–74 ($a = 10$), period: 1980–1984 ($p = 1$); (h) age: 70–74 ($a = 10$), period: 1990–1994 ($p = 3$); (i) age: 70–74 ($a = 10$), period: 2005–2008 ($p = 6$).

and the Bayesian information criterion (BIC) and conclude the choice of our final model with AIC = 19,310 and BIC = 19,808. The posterior estimates for age, period and cohort effects are listed in Table 4.

Age, period and cohort main effects and their 95% confidence intervals are displayed in Fig. 4. To improve numeric stability and mixing in the MCMC samplings, we implement model constraints to center the age, period and cohort effects to their means. The age

**Table 2**
Model comparisons using DIC.

| Model | Age effect | | |
|---|---|---|---|
| | ICAR (intrinsic conditional autoregression) prior | Carcinogenesis model priors | |
| | | Armitage–Doll | TSCE |
| AA | 20,966.60 | 20,973.90 | 2096.20 |
| AAP | 19,499.80 | 19,506.30 | 19,501.30 |
| AAC | 19,813.90 | 19,886.80 | 19,842.40 |
| AAP + AP | 19,592.40 | 19,606.6 | 19,575.70 |
| AAC + AC | 19,814.40 | 19,838.20 | 19,823.00 |
| AAPC | 19,231.30 | 19,233.00 | 19,234.70 |
| AAPC + AP | 19,021.30 | 19,230.70 | 19,240.70 |
| AAPC + AC | 18,985.00 | 18,996.30 | 19,235.80 |
| AAPC + AP + AC | 19,227.00 | 19,228.30 | 19,223.40 |

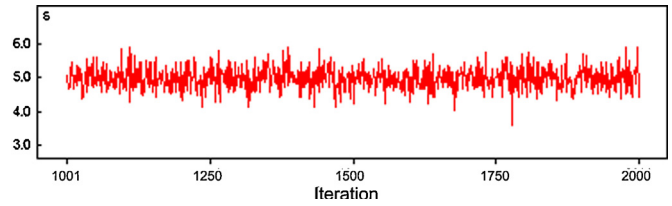


**Fig. 2.** Convergence plots for posterior estimate of Armitage–Doll model.

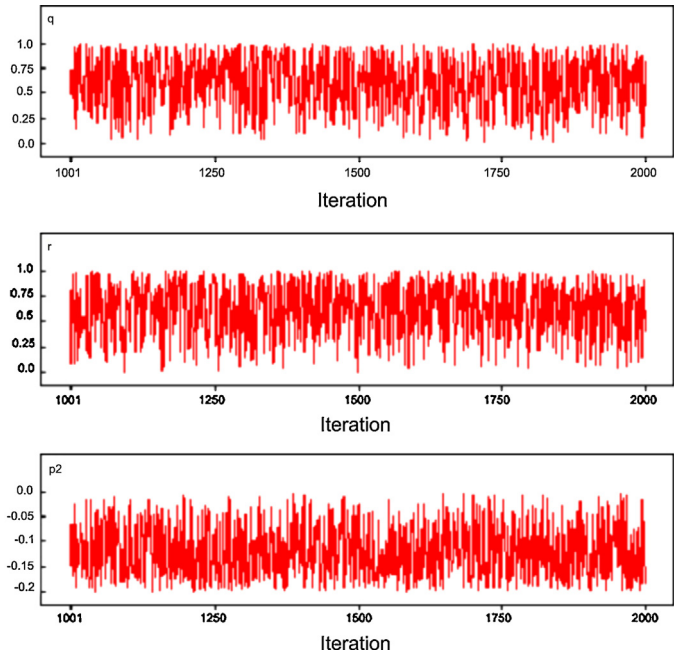


**Fig. 3.** Convergence plots for posterior estimates of TSCE model parameters.

effects show an increasing pattern over time, which means older age leads to higher cancer mortality rate than younger age does as we controlled for other covariates. The age pattern can be easily explained by Armitage–Doll carcinogenesis model since we assume a log-linear relationship between age and hazard function. The change point for the period effects is in period 2000–2004. The period effects are increasing before the year 2000 but sharply decreasing after the year 2000. The anti-smoking campaign has been introduced in the U.S. in the 1990s and since then people's behaviors on smoking have significantly changed which explain the decreasing trend in period effects since 2000. The lung cancer mortality rate is continuously declining by birth cohorts. The main area effects are displayed in Fig. 5 which shows a higher lung cancer mortality rate in the south of Iowa as compared to that in the north of Iowa. In southern Iowa, there are higher rates of radon gas and a higher rate of smoking blue-collar workers, and manufacturing jobs where coal mining previously occurred. The scatter plot of main spatial effects vs. Iowa county population from 2000 to 2004 is displayed in Fig. 6. It supports the conclusion of higher spatial effects in largely populated areas, such as Polk county where Iowa capital city Des Moines is located. Compared to the main area and cohort effects in the AAPC model, the coefficients of area–cohort interactions are much smaller and can be ignored.

To compare the temporal trends between carcinogenesis priors and conventional priors assigned to age effects, we display the posterior estimate for age, period and cohort effects when the age effects are assigned with ICAR (intrinsic conditional autoregression) priors in Table 5 and carcinogenesis priors in Table 4. The carcinogenesis age effects tend to shrink more toward the mean as compared to ICAR age effects. When ICAR priors were assigned to age effects, the trend of posterior period and cohort effects in Table 5 were totally different from Table 4. We observed the decreasing period trend but increasing cohort effects which are hard to explain. For example, it is difficult to explain why the birth cohort effect at 1978–1982 is significantly higher than those at 1908–1912. On the other hand, the decreasing birth cohort trends in Table 4 make more sense since the development of the lung cancer awareness and anti-smoking campaign in last decades have significant changed people's view and behavior in smoking. The comparison between Tables 4 and 5 demonstrates that the incorporation of carcinogenesis age effects can improve the interpretation of temporal trends derived from Bayesian extended AAPC model.

**Table 3**
Sensitivity analysis.

| Prior* | $s$** | |
|---|---|---|
| | Posterior mean | Posterior SD |
| *Gamma*(1, 0.01) | 4.659 | 0.3259 |
| *Gamma*(0.01, 0.01) | 4.672 | 0.3258 |
| *Gamma*(1, 0.02) | 4.658 | 0.3259 |

* This prior is assigned to the variance component for the unstructured spatial effects $v_i$.

** $s$ is the parameter of number of stages in Armitage–Doll Carcinogenesis model.

**Table 4**
Posterior estimates for age, period and cohort effects in the Bayesian extended AAPC model with Armitage–Doll priors for age effects.

| Age | | Period | | Cohort | |
|---|---|---|---|---|---|
| 25–29 ($a=1$) | −3.021 | 1980–1984 ($p=1$) | −0.330 | 1908–1912 ($c=1$) | 0.969 |
| 30–34 ($a=2$) | −2.289 | 1985–1989 ($p=2$) | −0.125 | 1913–1917 ($c=2$) | 0.907 |
| 35–39 ($a=3$) | −1.375 | 1990–1994 ($p=3$) | 0.014 | 1918–1922 ($c=3$) | 0.840 |
| 40–44 ($a=4$) | −0.529 | 1995–1999 ($p=4$) | 0.157 | 1923–1927 ($c=4$) | 0.792 |
| 45–49 ($a=5$) | 0.211 | 2000–2004 ($p=5$) | 0.308 | 1928–1932 ($c=5$) | 0.641 |
| 50–54 ($a=6$) | 0.757 | 2005–2008 ($p=6$) | −0.024 | 1933–1937 ($c=6$) | 0.473 |
| 55–59 ($a=7$) | 1.209 | | | 1938–1942 ($c=7$) | 0.270 |
| 60–64 ($a=8$) | 1.522 | | | 1943–1947 ($c=8$) | −0.021 |
| 65–69 ($a=9$) | 1.714 | | | 1948–1952 ($c=9$) | −0.285 |
| 70–74 ($a=10$) | 1.801 | | | 1953–1957 ($c=10$) | −0.394 |
| | | | | 1958–1962 ($c=11$) | −0.518 |
| | | | | 1963–1967 ($c=12$) | −0.694 |
| | | | | 1968–1972 ($c=13$) | −0.865 |
| | | | | 1973–1977 ($c=14$) | −0.996 |
| | | | | 1978–1982 ($c=15$) | −1.12 |

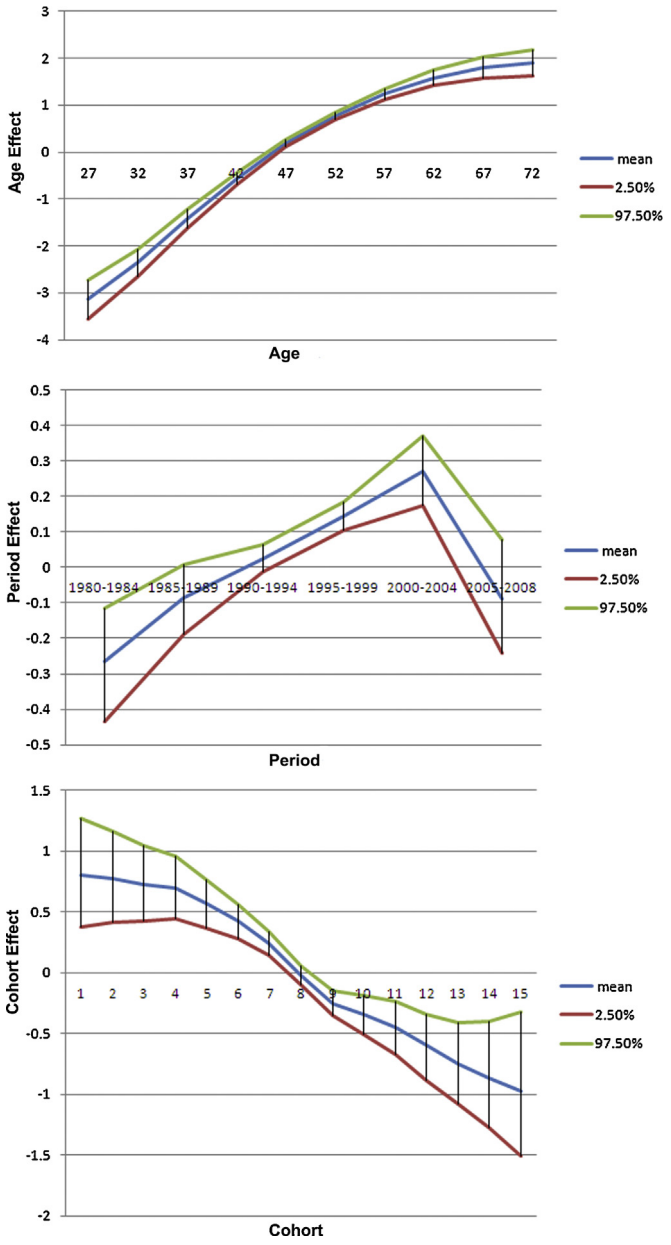


**Fig. 4.** Age, period, and cohort main effects in AAPC model.

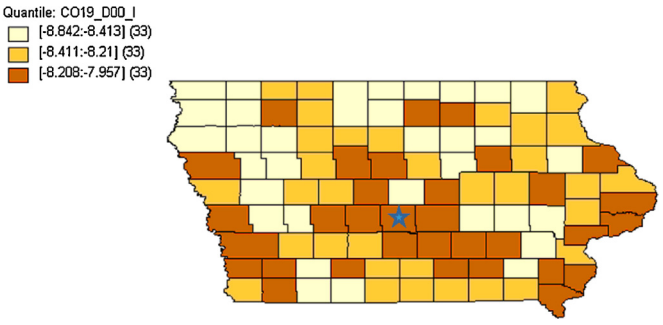


**Fig. 5.** Area main effects in AAPC model.

To demonstrate how the Bayesian extended AAPC model fit the data, we aggregate the mortality rates across all Iowa counties and compare the predicted rates with the observed rates for different age and period groups in Fig. 7a and b. It shows that the predicted rates agree well with the observed mortality rates.

## 4. Discussion

It is very important to determine the trends of disease risk both temporally and spatially [9]. However, it is difficult to explain some of those temporal effects (age, period, and cohort) [28] due to lack of biological meanings. Carcinogenesis models of a typical underlying disease process describe how normal cells are transformed into cancer cells and age is a deterministic factor in the model. Therefore, we propose a new extended AAPC model by incorporating carcinogenesis model into our study to improve our prior knowledge of age effects in determining disease trends. Both Armitage–Doll and TSCE carcinogenesis models are considered in this study. The lung cancer mortality study shows the extended AAPC model with area–cohort interaction and Armitage–Doll age effects can be used to estimate lung cancer risk while we control the age effect from the underline disease process. The convergence of model parameters is guaranteed as well. The extended AAPC model can be used in studying spatial–temporal pattern of cancer mortality with strong biological prior beliefs in the age effects.

Non-identifiability is the common challenge in fitting APC and AAPC models. We have added model constraints in our extended AAPC models in considering the identifiability issues. However, further works are still needed in the extended AAPC models in this area. A different sampling technique which uses multivariate Metropolis steps [9,18] would be a better approach to handle efficiently the identifiability problems. DIC has been widely used in

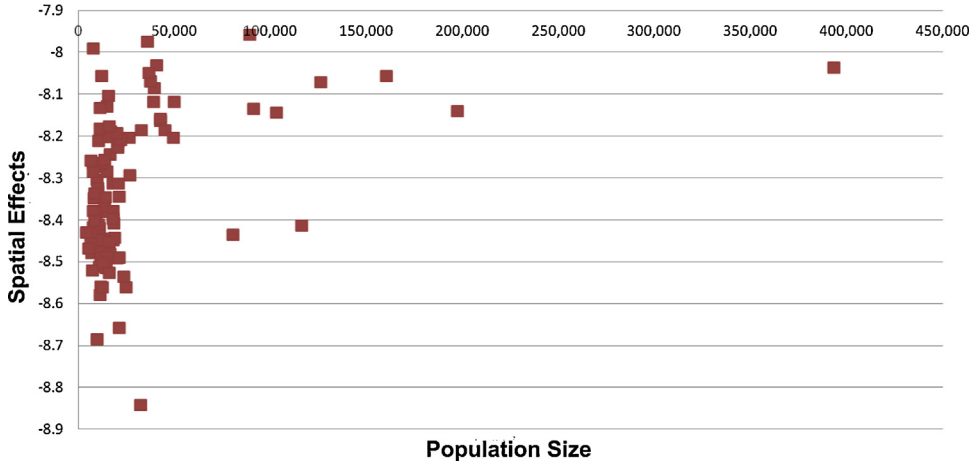


**Fig. 6.** The scatter plot of main spatial effects in lung cancer mortality vs. Iowa county populations from 2000 to 2004.

**Table 5**
Posterior estimates for age, period and cohort effects in the Bayesian extended AAPC model with ICAR priors for age effect.

| Age | | Period | | Cohort | |
|---|---|---|---|---|---|
| 25–29 ($a=1$) | −5.81 | 1980–1984 ($p=1$) | 1.11 | 1908–1912 ($c=1$) | −3.195 |
| 30–34 ($a=2$) | −4.339 | 1985–1989 ($p=2$) | 0.723 | 1913–1917 ($c=2$) | −2.658 |
| 35–39 ($a=3$) | −2.841 | 1990–1994 ($p=3$) | 0.274 | 1918–1922 ($c=3$) | −2.144 |
| 40–44 ($a=4$) | −1.406 | 1995–1999 ($p=4$) | −0.177 | 1923–1927 ($c=4$) | −1.6 |
| 45–49 ($a=5$) | −0.078 | 2000–2004 ($p=5$) | −0.612 | 1928–1932 ($c=5$) | −1.167 |
| 50–54 ($a=6$) | 1.071 | 2005–2008 ($p=6$) | −1.319 | 1933–1937 ($c=6$) | −0.753 |
| 55–59 ($a=7$) | 2.114 | | | 1938–1942 ($c=7$) | −0.363 |
| 60–64 ($a=8$) | 3.014 | | | 1943–1947 ($c=8$) | −0.055 |
| 65–69 ($a=9$) | 3.799 | | | 1948–1952 ($c=9$) | 0.233 |
| 70–74 ($a=10$) | 4.477 | | | 1953–1957 ($c=10$) | 0.783 |
| | | | | 1958–1962 ($c=11$) | 1.27 |
| | | | | 1963–1967 ($c=12$) | 1.602 |
| | | | | 1968–1972 ($c=13$) | 1.926 |
| | | | | 1973–1977 ($c=14$) | 2.676 |
| | | | | 1978–1982 ($c=15$) | 3.445 |

model selection for Bayesian statistics. However, some statisticians were cautious about the performance of DIC (see the discussion in Spiegelhalter et al. [29]). As a result, alternative model comparison methods such as the posterior expectation of Akaike information criterion (EAIC) and the Bayesian information criterion (EBIC) could be considered to demonstrate the agreement with DIC in the future study. Choosing different priors for temporal effects instead of autoregressive Gaussian distribution can also be considered in the Bayesian model. More complicated forms for spatial priors can be added in the future study. For example, Waller [2] suggested to include the distance between county $i$ and $j$ in the formula of computing the weights. Furthermore, covariates such as smoking status, social economic status of the counties might be included in the model.

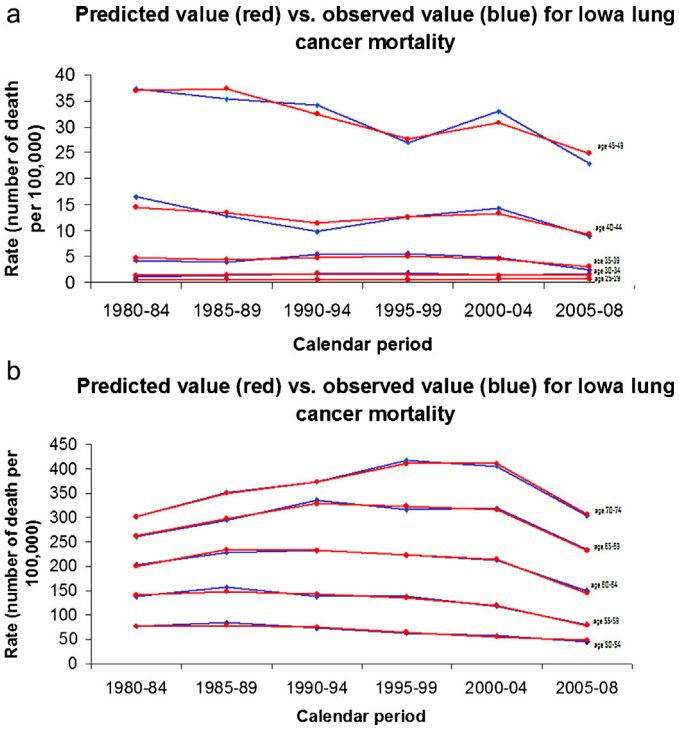


**Fig. 7.** (a) Predicted value (red) vs. observed value (blue) for Iowa lung cancer mortality for age 25–49. (b) Predicted value (red) vs. observed value (blue) for Iowa lung cancer mortality for age 50–74. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

## Conflict of interest statement

None declared.

## Appendix A

### A.1. Priors for spatial effects

The prior for the unstructured spatial effects $\nu_i$ is taken as $\nu_i \sim N(0, \tau)$ where $\tau$ is a hyperprior which is defined as gamma distribution as $\tau \sim Gamma(0.001, 0.001)$. Independence is assumed to all spatial unstructured effects. The intrinsic Gaussian conditional autoregressive priors are considered for the structured spatial effects. Congdon [2006a] showed the joint prior for the structured spatial effect $\mu = (\mu_1, ..., \mu_n)$ can be taken as

$$P(\mu_1, ..., \mu_n) \propto \exp\left[-0.5\kappa^{-1}\sum_{i\sim j} c_{ij}(\mu_i - \mu_j)^2\right],$$

where $c_{ij}$ is the contiguity matrix defined as

$$c_{ij} = \begin{cases} 1, & \text{if area } i \text{ and } j \text{ are first-order neighbors,} \\ 0, & \text{otherwise.} \end{cases}$$

In WinBUGS, we use the distribution car.normal to assign ICAR priors to the joint spatial structured effects $\mu$.

### A.2. Priors for period and cohort effects

Noninformative normal priors with mean at 0 and large variance are used to define the first two period or cohort effects, i.e., $\beta_1, \beta_2 \sim N(0, 100000\frac{1}{\tau}), \gamma_1, \gamma_2 \sim N(0, 100000\frac{1}{\tau})$. For the remaining period or cohort effects, we specify the conditional distribution of period or cohort effect, giving its previous period or cohort effects following a normal distribution with mean at a linear extrapolation from its two immediate predecessors as

$$\beta_j|\beta_1, ..., \beta_{j-1} \sim N(2\beta_{j-1} - \beta_{j-2}, \tau),$$

$$\gamma_k|\gamma_1, ..., \gamma_{k-1} \sim N(2\gamma_{k-1} - \gamma_{k-2}, \tau),$$

$$j = 3, ..., P, k = 3, ..., C.$$

The precision term $\tau$ is assigned as a noninformative Gamma hyperprior at

$\tau \sim Gamma(0.001, 0.001)$.

### A.3. Nonidentifiability constraints

We add several model constraints to control the nonidentifiability problems in APC models. Adding those constraints will improve the numerical stability and mixing in MCMC sampling. The first constraint we use is to centralize all three temporal effects such as

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = 0.$$

Additional constraint we employ is to remove the linear trend in age effects.

Holford [30] suggested to partition age effects into two components (linear slope and residuals) and fitted residuals into APC model. The age effects can be represented as

$$\alpha = \left[ i - \frac{I+1}{2} \right] \alpha^{slo\,ps} + \alpha_i^{res},$$

where $\alpha^{slops}$ is the underlining slope for the age effect and $\alpha_i^{res}$ is the residual or curvature effect for age group $i$. The slope $\alpha^{slops}$ can be estimated through the conventional approach to solve the ordinary least square (OLS) estimate to a linear model ($Y = X\beta + \varepsilon$), where $\beta = (X'X)^{-1}X'Y$. Therefore, we can obtain OLS estimate for $\alpha^{slops}$ as below,

$$\alpha^{slo\,ps} = \left( \sum_{i=1}^{I} \left[ i - \frac{I+1}{2} \right]^2 \right)^{-1} \sum_{i=1}^{I} \left( i - \frac{I+1}{2} \right) \alpha_i,$$

and equivalently,

$$\alpha^{slo\,ps} = \frac{\sum_{i=1}^{I} \left( i - \frac{I+1}{2} \right) \alpha_i}{\frac{I(I+1)(I-1)}{12}}.$$

The age residual terms can be calculated as

$$\alpha_i^{res} = \alpha_i - \left( i - \frac{I+1}{2} \right) \alpha^{slo\,ps},$$

and plugged into the APC model to replace the age effects.

## References


[1] Carlin BP, Lious T. Bayesian methods for data analysis. Boca Raton, FL: Chapman & Hall/CRC, 2009.

[2] Waller L, Carlin BP, Xia H, Gelfand AE. Hierarchical spatio-temporal mapping of disease rates. J Am Stat Assoc 1997;92:607–17.

[3] Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, Songini M. Bayesian analysis of space–time variation in disease risk. Stat Med 1995;14(21–22):2433–43.

[4] Besag J, York J, Mollie A. Bayesian image restoration with two applications in spatial statistics. Ann Inst Stat Math 1991;43:1–59.

[5] Knorr-Held L. Bayesian modelling of inseparable space–time variation in disease risk. Stat Med 2000;19(17–18):2555–67.

[6] Ugarte MD, Goicoa T, Ibanez B, Militino F. Evaluating the performance of spatio-temporal Bayesian models in disease mapping. Environmetrics 2009;20:647–65.

[7] Wakefield J. Disease mapping and spatial regression with count data. Biostatistics 2007;8(2):158–83.

[8] Buenconsejo J, Fish D, Childs JE, Holford TR. A Bayesian hierachical model for the estimation of two incomplete surveillance data sets. Stat Med 2008;27:3269–85.

[9] Lagazio C, Biggeri A, Dreassi E. Age–period–cohort models for disease mapping. Environmetrics 2003;14:475–90.

[10] Jeon J, Luebeck EG, Moolgavkar SH. Age effects and temporal trends in adenocarcinoma of the esophagus and gastric cardia (United States). Cancer Causes Control 2006;17(7):971–81.

[11] Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer 1954;8:1–12.

[12] Fisher JC, Hollomon JH. A hypothesis for the origin of cancer foci. Cancer 1951;4:916–8.

[13] Nordling CO. A new theory on the cancer-inducing mechanism. Br J Cancer 1953;7:68–72.

[14] Hazelton WD, Clements MS, Moolgavkar S. Multistage carcinogenesis and lung cancer mortality in three cohorts. Cancer Epidemiol Biomarkers Prev 2005;14:1171–81.

[15] Holford TR, Levy DT. Comparing the adequency of carcinogenesis models in estimating U.S. population rates for lung cancer mortality. Risk Anal 2012;32:S179–89.

[16] Clayton D, Schifflers E. Models for temporal variation in cancer rates, II: age–period–cohort models. Stat Med 1987;6(4):469–81.

[17] Clayton D, Schifflers E. Models for temporal variation in cancer rates, I: age–period and age–cohort models. Stat Med 1987;6(4):449–67.

[18] Knorr-Held L, Rainer E. Projects of lung cancer mortality in west Germany: a case study in Bayesian prediction. Biostatistics 2001;2:109–29.

[19] Schmid V, Held L. Bayesian extrapolation of space–time trends in cancer registry data. Biometrics 2004;60(4):1034–42.

[20] Holford TR. Understanding the effects of age, period, and cohort on incidence and mortality rates. Annu Rev Public Health 1991;12:425–57.

[21] Luebeck EG, Moolgavkar SH. Multistage carcinogenesis and the incidence of colorectal cancer. Proc Natl Acad Sci U S A 2002;99(23):15095–100.

[22] Iowa Department of Public Health. Behavioral risk factor surveillance system [cited 2013 June 29]; 2013, Available from: http://www.idph.state.ia.us/brfss/.

[23] LaVecchia C, Negri E, Levi F, Decarli A, Boyle P. Cancer mortality in Europe: effects of age, cohort of birth and period of death. Eur J Cancer 1998;34:118–41.

[24] Moolgavkar SH, Luebeck G. Two-event model for carcinogenesis: biological, mathematical, and statistical considerations. Risk Anal 1990;10(2):323–41.

[25] Moolgavkar SH, Luebeck EG. Multistage carcinogenesis: population-based model for colon cancer. J Natl Cancer Inst 1992;84(8):610–8.

[26] Moolgavkar SH, Luebeck EG. Multistage carcinogenesis and the incidence of human cancer. Genes Chromosomes Cancer 2003;38(4):302–6.

[27] Bray I. Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. J R Stat Soc Ser C-Appl Stat 2002;51:151–64.

[28] Richardson DB. A simple approach for fitting linear relative rate models in SAS. Am J Epidemiol 2008;168(11):1333–8.

[29] Spiegelhalter D, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc B 2002;64:583–639.

[30] Holford TR, Zhang ZX, Mckay LA. Estimating age, period and cohort effects using the multistage model for cancer. Stat Med 1994;13(1):23–41.