STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:4619–4642
Published online 20 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2884

Iowa radon leukaemia study: A hierarchical population risk model for spatially correlated exposure measured with error

Brian J. Smith[1,*,†], Lixun Zhang[2] and R. William Field[3]

[1]*Department of Biostatistics, The University of Iowa, Iowa City, IA, U.S.A.*
[2]*Departments of Biostatistics and Geography, The University of Iowa, Iowa City, IA, U.S.A.*
[3]*Departments of Occupational and Environmental Health and Epidemiology, The University of Iowa, Iowa City, IA, U.S.A.*

SUMMARY

This paper presents a Bayesian model that allows for the joint prediction of county-average radon levels and estimation of the associated leukaemia risk. The methods are motivated by radon data from an epidemiologic study of residential radon in Iowa that include 2726 outdoor and indoor measurements. Prediction of county-average radon is based on a geostatistical model for the radon data which assumes an underlying continuous spatial process. In the radon model, we account for uncertainties due to incomplete spatial coverage, spatial variability, characteristic differences between homes, and detector measurement error. The predicted radon averages are, in turn, included as a covariate in Poisson models for incident cases of acute lymphocytic (ALL), acute myelogenous (AML), chronic lymphocytic (CLL), and chronic myelogenous (CML) leukaemias reported to the Iowa cancer registry from 1973 to 2002. Since radon and leukaemia risk are modelled simultaneously in our approach, the resulting risk estimates accurately reflect uncertainties in the predicted radon exposure covariate. Posterior mean (95 per cent Bayesian credible interval) estimates of the relative risk associated with a 1 pCi/L increase in radon for ALL, AML, CLL, and CML are 0.91 (0.78–1.03), 1.01 (0.92–1.12), 1.06 (0.96–1.16), and 1.12 (0.98–1.27), respectively.

KEY WORDS: Bayesian methods; leukaemia risk; Markov chain Monte Carlo; Poisson regression; residential radon exposure; spatial statistics

*Correspondence to: Brian J. Smith, Department of Biostatistics, The University of Iowa, 200 Hawkins Drive, C22 GH, Iowa City, IA 52242-1009, U.S.A.
†E-mail: brian-j-smith@uiowa.edu

Contract/grant sponsor: National Institute of Environmental Health Sciences; contract/grant numbers: R01 ES05653, P30 ES05605
Contract/grant sponsor: National Cancer Institute, NIH; contract/grant number: R01 CA85942

Received 13 January 2006
Accepted 15 February 2007

1. INTRODUCTION

The American Cancer Society estimates that 35 070 new cases of leukaemia will be diagnosed in the United States in 2006 [1]. In fact, in men less than 40 years of age and women under 20, leukaemia is the most common cause of cancer death. However, over 60 per cent of leukaemias occur in people over 50 years of age. Leukaemia is classified as either myelogenous or lymphocytic, reflecting the cell type of origin, and can be further classified as either acute or chronic. Acute lymphocytic leukaemia (ALL), a rapidly progressing cancer affecting lymphocytes, is the most common form of childhood leukaemia, whereas acute myelogenous leukaemia (AML), a rapidly progressive cancer affecting immature cells of the bone marrow, exhibits the highest incidence of the four forms of leukaemia among adults [2]. In the United States, the occurrence of leukaemia is more common in males than females and also more likely to occur in white individuals of European decent.

A substantial body of leukaemia research supports a causal effect of medically related therapies such as chemotherapy, radiation therapy, growth hormones, and antibiotics. Alternatively, environmental risk factors for leukaemia are understudied. While some chemicals like pesticides, herbicides, solvents, dioxins, butadienes, ethylene oxides, and styrenes as well as infectious agents have been suggested to cause various forms of leukaemia, only benzene and external penetrating ionizing radiation have widespread scientific acceptance as causative agents. In regard to lifestyle-related risks, Korte *et al.* [3] have suggested that tobacco smoking may be responsible for up to three-fifths of AML mortality. The chemical constituents of tobacco that may pose a risk of leukaemia include benzene, polonium-210 (radon-222 decay product), and polycyclic aromatic hydrocarbons.

As mentioned above, ionizing radiation in the form of external penetrating radiation (X-rays and gamma rays) has been implicated as a causative agent in the induction of AML, ALL, and chronic myelogenous leukaemia (CML) since 1944 [4, 5]. In addition, a recent scientific study [6] has noted a possible causal association between radon exposure in mines and chronic lymphocytic leukaemia (CLL). The potential for the development of leukaemia from internal ionizing radiation exposure has received less attention. As pointed out by Kendall and Smith [7], there is significant uncertainty in the dose estimates for the haematopoietic bone marrow for short-lived radon decay products since the initial models were designed to address dose estimates for longer-lived radionuclides. In addition, the significant uncertainty associated with the process of translocation [8] of radon progeny and the half-time for absorption to blood significantly affects the accuracy of dose estimates.

Ecologic radon studies examine associations between disease outcomes and exposures aggregated over geographic regions. The exposures in such studies cannot be observed directly and are generally predicted from a sample of radon measurements within each region. Errors in predictive measures of regional radon exposures have largely been ignored. Misleading risk estimates can result from analyses that ignore covariate measurement error [9–13]. As in previous ecologic radon studies, our primary goal is to investigate the associations between disease and exposure. However, we take a novel approach that more appropriately accounts for important sociodemographic factors and uncertainty in predicted radon exposures.

We present an analysis of the effects of radon on leukaemia risk. Our data include radon measurements collected in a case-control study of residential radon in Iowa and incident leukaemia cases from state cancer registries. The radon measurements were taken at geographic points in space, whereas the leukaemia data are available as counts at the county level. To combine the point-referenced radon and areal leukaemia data, a geostatistical approach is used to model radon as

DOI: 10.1002/sim

a continuous spatial process and to predict county-average concentrations. The predicted averages are, in turn, included as a covariate in Poisson models for incident leukaemia cases. Within a fully Bayesian framework, we develop a hierarchical model to simultaneously predict radon and estimate the associated leukaemia risk. Consequently, the resulting risk estimates accurately reflect the uncertainty in predicting radon, which includes incomplete spatial coverage, spatial variability, characteristic differences between homes, and detector measurement error.

Components of our modelling strategy have been considered by other authors. Richardson *et al.* [14] use Bayesian methods to estimate the effect of radon on leukaemia rates for geopolitical districts in Great Britain. Although their risk model includes spatial and non-spatial components to account for extra-Poisson variability in the rates, radon is treated as a fixed covariate in the analyses. A recent paper by Toti *et al.* [15] explores a covariate measurement error model for radon exposure in the case-control setting. Their error model assumes that radon measurements vary randomly about the group means for 'similar houses', but does not directly account for the spatially correlated nature of radon. Our approach is most similar to that of Zhu *et al.* [16] in which predicted average, ambient ozone levels are linked to the occurrence of asthma-related emergency room visits in Atlanta zip codes. Zhu *et al.* also discuss the computational challenges of implementing their sophisticated Bayesian model and ultimately rely on 'shortcuts' to reduce the complexities of their algorithms. Simplifying assumptions are not uncommon when dealing with complicated Bayesian models, and the authors present simulation results to validate their methods. Unfortunately, the fully Bayesian nature of their model is not realized in the implementation of their algorithms. Such concessions are not made in the algorithms for the Bayesian models presented in this paper.

Our paper is organized as follows. The motivating radon and leukaemia data are described in Section 2. A Bayesian hierarchical model linking residential radon to incident leukaemia cases is developed in Section 3. Spatial correlation, detector measurement error, and systematic mean differences in radon measurements are considered in the model development. In addition, we show how the model can be used to predict county-average radon concentrations. Results from the proposed Bayesian model are provided in Section 4, along with those from a model that ignores the radon prediction error. We conclude with Section 5 in which a discussion of our analytic approach is given.

2. MATERIALS

2.1. Iowa radon study

Field *et al.* [17] conducted an epidemiologic case-control study in Iowa to estimate the effect of residential radon on lung cancer risk. Data were collected over a four year span beginning in 1993. The study enrolled 413 incident lung cancer cases and 614 population-based, disease-free controls. Although risk estimation was the primary aim of the Iowa Study, the detailed environmental data that were collected provide a unique opportunity to characterize the distribution of outdoor and indoor radon. Alpha-track detectors were used to obtain year-long radon measurements at subject homes and outdoor sites [18]. Since lung cancer cases were over-sampled in the Iowa study, only data from the 614 control subjects are included in our current analyses. In particular, we utilize the following: (1) 2590 radon measurements from the disease-free study participant homes; (2) the home floors on which radon detectors were located; and (3) 136 measurements from 109 outdoor

DOI: 10.1002/sim

sites across the state. At least one radon measurement was taken on each floor of the home, resulting in an average of 4.2 measurements per home. The geographic distribution of homes mirrors that of the general Iowa population since control subjects were a population-based sample, whereas the outdoor sites were chosen to be more uniformly distributed across the state.

2.2. SEER leukaemia data

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute collects and publishes cancer incidence and survival data from 14 population-based cancer registries and three supplemental registries. Each cancer registry covers geographic areas as determined by the registry's ability to operate and maintain a high-quality reporting system [19]. The SEER program began in 1973 with registries in Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco-Oakland, and Utah. Seattle-Puget Sound was added in 1974 and Atlanta in 1975. Although additional registries joined the SEER program after 1978, we only used 1973–2002 data from the original nine. In particular, we obtained registry data for the calculation of incident leukaemia and total population counts by county, 5-year age strata, gender, race, and calendar year.

2.3. County area resource file

The Area Resource File (ARF), as made available by the Bureau of Health Professions [20], contains over 6000 variables for each county in the US. More than 50 data source files were used to create the ARF. We extracted the values for variables 'per cent high school diploma or higher', 'unemployment rate', 'urban influence code', and 'per cent below poverty'. Urban influence codes divide US counties into 12 groups based on population and commuting data for metropolitan counties and on adjacency to metro areas for non-metropolitan counties. A value of one represents the most urban counties and a value of 12 represents the most rural ones.

3. METHODS

3.1. Leukaemia risk model

We wish to study the effect of residential radon on leukaemia risk in Iowa, while controlling for potential confounding effects of other covariates. To do so, we first partition the Iowa population into demographic subgroups defined by race and calendar year. Then, we calculate the number of leukaemia cases that would be expected in each subgroup if the age- and gender-specific rates from the nine original SEER registries were applied to the Iowa population. The expected number of cases in the kth county and lth demographic subgroup can be expressed as

$$E_{kl} = \sum_i m_{kli}\, \lambda_i$$

where the summation is over age-gender strata, m_{kli} is the person-years at risk in the Iowa population, and λ_i is the stratum-specific rate from the SEER population. Note that the ratio of observed cases y_{kl} to expected cases E_{kl} is the maximum likelihood estimate for the *standardized morbidity ratio*, i.e. $\widehat{\mathrm{SMR}}_{kl} = y_{kl}/E_{kl}$. For the risk analyses, we assume that the observed cases

DOI: 10.1002/sim

follow a Poisson model of the following form:

$$\begin{aligned} y_{kl} &\sim \text{Poisson}(E_{kl}e^{\psi_{kl}}) \\ \psi_{kl} &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_{kl} + \beta_r \bar{r}'_k + \theta_{kl} + \phi_k \\ \theta_{kl} &\overset{iid}{\sim} \text{N}(0, \sigma_h^2) \\ \boldsymbol{\phi} &\sim \text{N}(\mathbf{0}, \sigma_c^2(D_w - \rho_c C)^{-1}) \end{aligned} \tag{1}$$

where $\mathbf{x}_{kl}$ is a vector of known covariates with corresponding mean parameters $\boldsymbol{\beta}$; $\bar{r}'_k$ is the predicted county-average radon level; and β_r is the effect of radon in the model. The $e^{\psi_{kl}}$ parameter can be interpreted as the true standardized morbidity ratio. Statewide heterogeneity is accounted for with the θ_{kl} parameters, and spatial correlation not explained by the covariates is modelled with the ϕ_k parameters. In the specification above, the $\boldsymbol{\phi}$ vector of spatial parameters follows a conditional autoregressive (CAR) spatial model, where σ_c^2 is the variance parameter, C is a proximity matrix of indicator variables $c_{kk'}$ that equals one if counties k and k' share a border and zero otherwise, and D_w is a diagonal matrix containing the number of neighbours for each county such that $(D_w)_{kk} = c_{k+}$. The ρ_c parameter is included and bounded by the reciprocal of the largest and smallest eigenvalues of $D_w^{-1/2} C D_w^{-1/2}$ to ensure a non-singular covariance matrix in the CAR specification [21].

The $\bar{r}'_k$ are unobservable since they represent integrated averages over geographic regions, within which radon concentrations vary continuously. Thus, we are faced with the challenge of predicting these averages. Ideally, our approach should also account for the level of uncertainty in the predicted radon levels. In the next section, we describe a geostatistical model that provides the predictive distribution for the county averages we seek.

3.2. Distribution for measured radon concentrations

In this section, we introduce the models to be used in characterizing the distribution of residential radon. Our approach is based on experience gained in previous analyses of the Iowa Study data. For a more comprehensive account of the justification and validation of our radon models, see Smith and Field [22].

Although outdoor and indoor radon both arise from uranium deposits in the soil, measurements from the two environments are expected to differ systematically and are modelled accordingly. In the model formulation, we differentiate between measurements taken at outdoor and indoor home sites. Furthermore, we allow for multiple observations at each geographic location. Let $r_{os,ij}$ denote the jth measurement from the ith outdoor site for $i = 1, \ldots, n_{os}$, and $r_{hm,ij}$ the jth measurement from the ith home site for $i = 1, \ldots, n_{hm}$.

We specify the following model for the outdoor measurements:

$$\begin{aligned} \ln r_{os,ij} &= \beta_{os} + z(s_i) + \varepsilon_{os,ij} \\ \varepsilon_{os,ij} &\overset{iid}{\sim} \text{N}(0, \sigma_{os}^2) \end{aligned} \tag{2}$$

where β_{os} is an overall mean parameter, $\varepsilon_{os,ij}$ is an independent error term, and σ_{os}^2 is the error variance. The $z(s_i)$ parameter accounts for spatial correlation among radon concentrations, as described in Section 3.3.

DOI: 10.1002/sim

Differences in building characteristics add to the variability of radon concentrations indoors. Radon primarily enters homes through the floors closest to the ground. Thus, higher home radon concentrations may result when there are cracks or openings in the building foundation. Likewise, the presence of a basement may lead to higher concentrations due to the increased surface area in contact with the ground. Once inside the home, radon dilutes as it rises up to higher floors; thus, setting up a gradient of concentrations with the highest concentrations occurring in basements [23]. The distribution within homes is affected by factors such as the use of forced air furnaces that increase the movement of air between floors. Additionally, outdoor concentrations are generally lower than indoor concentrations where the enclosed nature of homes leads to the accumulation of radon therein. Lower concentrations have also been noted in older homes which tend to be draftier than more recently constructed homes. Therefore, models for indoor radon should allow for systematic differences between floors as well as homes. Consequently, we model home radon measurements as

$$\begin{aligned} \ln r_{hm,ij} &= \boldsymbol{\beta}_{hm}^{\mathrm{T}} \mathbf{x}_{hm,ij} + \gamma_i + z(s_i) + \varepsilon_{hm,ij} \\ \gamma_i &\overset{iid}{\sim} \mathrm{N}(0, \sigma_{bh}^2) \\ \varepsilon_{hm,ij} &\overset{iid}{\sim} \mathrm{N}(0, \sigma_{wh}^2) \end{aligned} \tag{3}$$

where $\mathbf{x}_{hm,ij}$ is a vector of mean covariates with corresponding parameters $\boldsymbol{\beta}_{hm}$; γ_i is an exchangeable random effect for the home, with variance σ_{bh}^2; $z(s_i)$ is the latent spatial parameter; and $\varepsilon_{hm,ij}$ is an independent error term, with variance σ_{wh}^2. Mean differences between homes may be due to unmeasured or unknown housing characteristics. In the absence of such information, differences between homes are captured with the home random effects. We also note that the error variance is a combination of variability due to systematic differences within the home and random detector measurement error.

3.3. Spatial correlation

Outdoor and indoor radon measurements exhibit spatial correlation because both originate from uranium deposited within the Earth's crust. Knowledge of deposited uranium and other soil characteristics that affect surficial radon concentrations is often incomplete. Consequently, it is natural to think of radon measurements as arising from a latent spatial process. Indeed, this is the rationale for our inclusion of the $z(s_i)$ parameters in equations (2) and (3). In this section, we explicitly define the latent spatial process as a stationary, multivariate Gaussian distribution. Let $\mathbf{z}^{\mathrm{T}} = (z(s_1), \ldots, z(s_n))$, for $i = 1, \ldots, n$ and $n = n_{os} + n_{hm}$, denote the vector of latent spatial parameters that correspond to the unique geographic sites at which radon measurements were obtained. Then, their assumed distribution is

$$\mathbf{z} \sim \mathrm{N}(\mu_{\mathbf{z}}(\boldsymbol{\beta}_s), \sigma_s^2 R_{\mathbf{z}}(\rho_s))$$

where $\mu_{\mathbf{z}}(\boldsymbol{\beta}_s)$ and $R_{\mathbf{z}}(\rho_s)$ are the mean vector and correlation matrix, respectively; ρ_s is a spatial correlation parameter; and σ_s^2 is a scalar variance parameter. Covariates that vary as a function of geographic location may be included in the mean vector. Since we do not have such covariates for our analyses, the latent process is assumed to have zero mean. Consequently, statewide mean radon concentrations are accounted for with the β_{os} and $\boldsymbol{\beta}_{hm}$ parameters in the previously specified models for outdoor and home radon measurements.

DOI: 10.1002/sim

For the spatial distribution, we chose to model the correlation between geographic sites s_i and $s_{i'}$ as

$$(R_{\mathbf{z}}(\rho_s))_{ii'} = c_s(s_i - s_{i'}; \rho_s)$$

where $c_s(s_i - s_{i'}; \rho_s)$ is a function of ρ_s and the distance between sites. To model the spatial correlation, we employ an isotropic Gaussian function of the form

$$c_s(s_i - s_{i'}; \rho_s) = \exp\{-\|s_i - s_{i'}\|^2/\rho_s^2\}$$

where $\|s_i - s_{i'}\|$ is calculated as the great circle distance (in miles) between sites [24], and ρ_s controls the rate of decay, as a function of distance. Many other correlation structures exist [25], and any parametric structure can be used in our model implementation. We considered both the Gaussian and exponential correlation functions in preliminary analyses. Leukaemia risk estimates did not differ between the two. We selected the Gaussian function because it allowed for more precise estimation of the radon distribution. Regardless, the specification of a spatial distribution that is a continuous function of the distance between point-referenced measurements makes our radon model geostatistical in nature. An advantage of geostatistical models is the ability to make prediction at unmeasured geographic sites—a property that we will utilize in the following section.

3.4. Predicted county-average radon

The specific aim of our analysis is to assess the leukaemia risk posed by residential radon. Since our measures of disease are county-specific leukaemia rates, we need corresponding county radon exposures in order to completely define our risk model. As a measure of exposure for county k, we use the average predicted radon concentration defined as

$$\bar{r}'_k = \exp\left\{ \int_{B_k} w(s) \ln r'(s)\, \mathrm{d}s \Big/ \int_{B_k} w(s)\, \mathrm{d}s \right\}$$

where B_k denotes the geographic region, or *block*, $r'(s)$ is a predicted radon concentration at site s, and $w(s)$ is a spatial weighting function. The $w(s)$ function allows for weighted averaging of radon concentrations within a county. We make use of this function in our analysis to weight concentrations by population density. In particular, we define $w(s)$ to be the population densities (people per square mile) for Iowa state zip codes, as reported in the 2000 U.S. Census and summarized here in Figure 1. Consequently, predicted radon concentrations in more densely populated regions receive more weight in our calculation of county averages. Predicted concentrations at geographic sites are based on the outdoor and indoor models in equations (2) and (3) and are assumed to be free of random home-to-home variability as well as detector measurement error. Consequently, the predicted county-average home radon concentration can be written as

$$\bar{r}'_k = \exp\left\{ \boldsymbol{\beta}_{hm}^{\mathrm{T}} \mathbf{x}_{hm} + \int_{B_k} w(s) z(s)\, \mathrm{d}s \Big/ \int_{B_k} w(s)\, \mathrm{d}s \right\}$$

for a given set of covariate values $\mathbf{x}_{hm}$. An analogous formula can be written for outdoor concentrations. Unfortunately, the integral cannot be solved exactly due to the irregular shape of geographic regions. Hence, numerical methods must be used to approximate the integral. One approach is to

DOI: 10.1002/sim

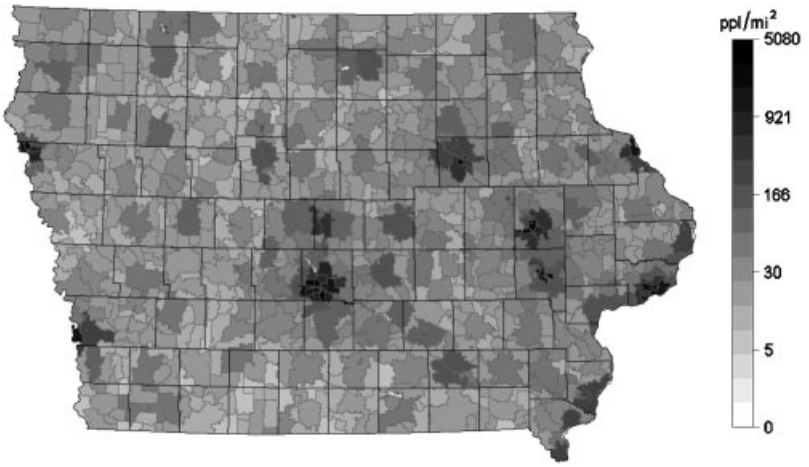

Figure 1. Population density (people per square miles) by Iowa state zip codes.

replace the integration by summation over a new set of sites $\{s_i' : i = 1 \ldots L_k\}$ so that

$$\bar{r}_k' \approx \exp\left\{\boldsymbol{\beta}_{hm}^{\mathrm{T}}\mathbf{x}_{hm} + \sum_{i=1}^{L_k} w(s_i')z(s_i') \Big/ \sum_{i=1}^{L_k} w(s_i')\right\} \tag{4}$$

If the sites are randomly sampled from uniform distributions within blocks, this is Monte Carlo integration. If they are generated from a low-discrepancy sequence, the approximation is referred to as quasi-Monte Carlo integration. Integration error for the quasi-Monte Carlo method is almost always smaller than that for standard Monte Carlo [26]. In any case, the approximation can be made arbitrarily accurate by varying the number of grid sites.

For the numerical integration in our application, we employ a fixed grid of approximately 3000 equally spaced sites $\{s_i'\}$ across Iowa, thus partitioning the state into 18.5 square-mile regions. A computational advantage of this approach is that the grid sites, and distances between them, remain constant throughout the iterative routine employed in the data analyses and need only be specified once—prior to the first iteration. Furthermore, lattice-based grid sites represent a type of low-discrepancy sequence, and thus our integration scheme is quasi-Monte Carlo. In so far as we can consider the Iowa counties to be rectangles, our integration technique also corresponds to the rectangle rule. Another possible approach to the integration is the Riemann approximation, although it has the disadvantage of being more awkward to implement for irregularly shaped blocks.

Denote the latent parameters for the set of grid sites by $\{z(s_i') : i = 1, \ldots, L\}$, and let $\mathbf{z}'$ be the vectorization of these parameters. Then, the joint distribution for the latent spatial parameters at the observed Iowa sites and unobserved grid sites is

$$\begin{pmatrix}\mathbf{z}\\ \mathbf{z}'\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \sigma_s^2 \begin{pmatrix} R_{\mathbf{z}}(\rho_s) & R_{\mathbf{z},\mathbf{z}'}(\rho_s) \\ R_{\mathbf{z},\mathbf{z}'}^{\mathrm{T}}(\rho_s) & R_{\mathbf{z}'}(\rho_s)\end{pmatrix}\right) \tag{5}$$

DOI: 10.1002/sim

Consequently, the predictive distribution for $\mathbf{z}' \mid \mathbf{z}, \sigma_s^2, \rho_s$ is a normal with mean

$$\mu_{\mathbf{z}'}(\boldsymbol{\beta}_s) + R_{\mathbf{z},\mathbf{z}'}^{\mathrm{T}}(\rho_s) R_{\mathbf{z}}^{-1}(\rho_s)\mathbf{z}$$

and variance

$$\sigma_s^2 (R_{\mathbf{z}'}(\rho_s) - R_{\mathbf{z},\mathbf{z}'}^{\mathrm{T}}(\rho_s) R_{\mathbf{z}}^{-1}(\rho_s) R_{\mathbf{z},\mathbf{z}'}(\rho_s))$$

Simulation methods will be used to sample from the multivariate predictive distribution for $\mathbf{z}'$. Predicted county-average radon concentrations are then estimated from these samples according to equation (4).

3.5. Joint posterior distribution

A fully Bayesian approach is taken to obtain the joint posterior distribution of all model parameters. The joint posterior is proportional to the following product of the likelihood functions associated with the models in equations (1)–(3), the distribution for the latent spatial parameters in equation (5), and prior distributions for the model parameters:

$$\left[\prod_k \prod_l f(y_{kl} \mid \boldsymbol{\beta}, \beta_r, \bar{r}'_k, \theta_{kl}, \phi_k)\right] f(\boldsymbol{\theta} \mid \sigma_h^2) f(\boldsymbol{\phi} \mid \sigma_c^2, \rho_c) \pi(\boldsymbol{\beta}, \beta_r, \sigma_h^2, \sigma_c^2, \rho_c)$$

$$\times \left[\prod_i \prod_j f(\ln r_{os,ij} \mid \beta_{os}, \mathbf{z}, \sigma_{os}^2)\right] \pi(\beta_{os}, \sigma_{os}^2)$$

$$\times \left[\prod_i \prod_j f(\ln r_{hm,ij} \mid \boldsymbol{\beta}_{hm}, \sigma_{bh}^2, \mathbf{z}, \sigma_{wh}^2)\right] \pi(\boldsymbol{\beta}_{hm}, \sigma_{bh}^2, \sigma_{wh}^2)$$

$$\times f(\mathbf{z}' \mid \mathbf{z}, \sigma_s^2, \rho_s) f(\mathbf{z} \mid \sigma_s^2, \rho_s) \pi(\sigma_s^2, \rho_s)$$

where $\bar{r}'_k$ is a function of the home radon mean parameters $\boldsymbol{\beta}_{hm}$ and the predicted latent parameters $\mathbf{z}'$, as defined in equation (4). We assume independent prior distributions for the analyses. Normal prior distributions are specified for each of the β mean parameters and inverse-gamma distributions for the σ^2 variance parameters. Uniform prior distributions are used for the ρ_c and ρ_s spatial correlation parameters.

The constant of proportionality for the posterior cannot be derived analytically. Therefore, we use Markov chain Monte Carlo (MCMC) methods to sample from the joint posteriors. Standard software routines, such as those found in WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs), cannot be used to perform MCMC sampling in our case because of the complex nature of the hierarchical model and the size of the radon data set. Instead, we implemented an MCMC algorithm for our model with the R programming language (http://www.r-project.org). Details of our sampling algorithm can be found in the Appendix. A potential advantage of R is that algorithms can often be implemented more quickly, relative to a compiled language. MCMC algorithms for geostatistical models are computationally intensive since the spatial correlation matrix must be updated and inverted at each iteration of the sampler. Our R program relies on the BLAS [27] library for linear algebra computations. The associated run times are highly dependent on the performance of these libraries. More than a five-fold increase in speed was achieved when using optimized ATLAS [28],

DOI: 10.1002/sim

versus reference, versions of BLAS. Software programs for fitting the proposed Bayesian model are available from the first author (BJS).

4. RESULTS

Five separate analyses were performed. The Bayesian hierarchical model of Section 3 was applied to each of the four leukaemia types (ALL, AML, CLL, and CML) separately as well as to all four combined. The same set of radon measurements from the Iowa Study were used throughout the analyses. In the following sections, the results of these analyses are presented. Details of the prior distributions are given in Section 4.1. Estimates for the predictive distribution of radon are provided in Section 4.2, and leukaemia risk estimates are summarized in Section 4.3.

4.1. Prior specifications and convergence diagnostics

Vague $N(0, 0.001)$ prior distributions are specified for each of the β mean parameters and Gamma$(0.001, 0.001)$ for the inverse variance parameters $1/\sigma^2$, where the Gamma(α, β) distribution is parameterized with mean equal to α/β. The CAR covariance matrix specified in the risk model of Section 3.1 is proper if $\rho_c \in (-1.78, 1.00)$. We assume a Uniform prior distribution over this range. Similarly, a Uniform$(0, 100)$ prior distribution is specified for the ρ_s spatial range parameter. The upper bound of 100 corresponds to an upper limit of 173 miles on the distance at which spatial correlation between radon measurements decays to 0.05 and is believed to include all possible values of this quantity. Furthermore, the upper bound proved to be large enough so as to have no discernable impact on posterior inference in our analyses.

Three parallel chains with dispersed starting values were generated for each of the analyses. Fifty-thousand iterations were run for each chain, of which 1000 were discarded as a burn-in sequence and every 5th subsequent iteration retained. Thus, posterior estimates are based on 29 400 autocorrelated samples from the posterior distribution. Convergence of the chains was assessed with the diagnostics of Gelman and Rubin [29] as well as with graphical checks of the output. Bayesian credible intervals are reported as highest probability density intervals (HPDs) computed using the method of Chen and Shao [30]. Convergence diagnostics and posterior summaries were performed with the BOA software [31].

4.2. Predictive distribution for radon

For the analyses, we included indicator variables as covariates in the mean structure of equation (3) to allow for systematic differences in radon concentrations between the basement, first floor, and second or higher floors. The associated model parameters are denoted as β_{hm0}, β_{hm1}, and β_{hm2}. Each of the five Bayesian analyses that we performed included the same set of radon measurements from the Iowa Study. Since the distributions for radon and disease risk are modelled jointly, the resulting predictive distributions for radon, including the associated model parameters, can differ. In our application, most of the information about the predictive distribution is provided by the measured radon concentrations rather than the disease outcomes, hence the radon results from the five analyses are similar. In the interest of brevity, we present in Table I only summaries of the geostatistical parameters from the analysis of all four leukaemia types combined. As expected, mean radon concentrations are lowest in outdoors and highest in basements. Since the log-transformed radon measurements are modelled directly, predicted statewide radon concentrations can be obtained

DOI: 10.1002/sim

Table I. Posterior summaries of the parameters in the geostatistical model for radon, estimated from the combined analysis of all four leukaemia types.

Radon parameter	Mean	SD	95 per cent HPD
β_{os}	−0.255	0.064	(−0.381, −0.128)
σ_{os}	0.292	0.024	(0.248, 0.342)
β_{hm0}	1.564	0.066	(1.431, 1.690)
β_{hm1}	0.936	0.066	(0.800, 1.059)
β_{hm2}	0.853	0.067	(0.718, 0.979)
σ_{bh}	0.707	0.023	(0.662, 0.752)
σ_{wh}	0.270	0.004	(0.262, 0.278)
σ_s	0.242	0.110	(0.160, 0.328)
ρ_s	35.1	13.6	(15.6, 67.4)

by exponentiating the β parameters. In particular, the posterior geometric mean outdoor radon concentration in Iowa is 0.77 pCi/L (95 per cent HPD 0.68–0.88 pCi/L), whereas the first floor geometric mean is 2.55 pCi/L (2.23–2.88 pCi/L). Posterior summaries of the county-specific average, first floor radon concentrations are mapped in Figure 2. The county averages are based on the underlying continuous spatial process assumed in the geostatistical model and exhibit patterns similar to those observed in other radon mapping efforts [22, 32].

Separate variance components are included in the geostatistical model to account for detector measurement error, unexplained differences between homes, and spatial dependencies. The model parameter σ^2_{bh} provides a measure of the between-home variance not explained by covariates in the model. Within-home variance is captured by the σ^2_{wh} parameter and includes both random measurement error and room-to-room variability. Posterior mean estimates indicate that the variance between homes is almost seven times higher than that within homes. Similar estimates are obtained for the within-home variance and the outdoor measurement error variance σ^2_{os}, which may indicate that the former is primarily a function of random detector measurement error. The σ^2_s variance parameter provides a measure of the variability in radon measurements attributable to the underlying spatial process. Its posterior mean falls in-between the means for the between and within-home variances. Correlation in the underlying spatial process is assumed to decay as a Gaussian function of distance and the parameter ρ_s. The posterior mean of 35.1 miles implies correlations of 0.98, 0.73, 0.27, and 0.05 at distances of 5, 20, 40, and 60 miles, respectively; although it should be noted that there is considerable variability in the decay parameter as indicated by the wide 95 per cent HPD of 15.6–67.4.

4.3. Leukaemia risk estimates

Since higher socio-economic status is a possible risk factor for lymphoproliferative disease [33–35], we conducted analyses with univariate Poisson models to examine the effect of the ARF factors described in Section 2.3. Frequentist, likelihood-based methods were used in this variable selection phase of the analyses since it would have been time-prohibitive to use the full Bayesian model. The select ARF factors were all non-significant except for urban influence code. Linear and quadratic effects for this variable were selected based on the results from likelihood ratio testing. Temporal

DOI: 10.1002/sim

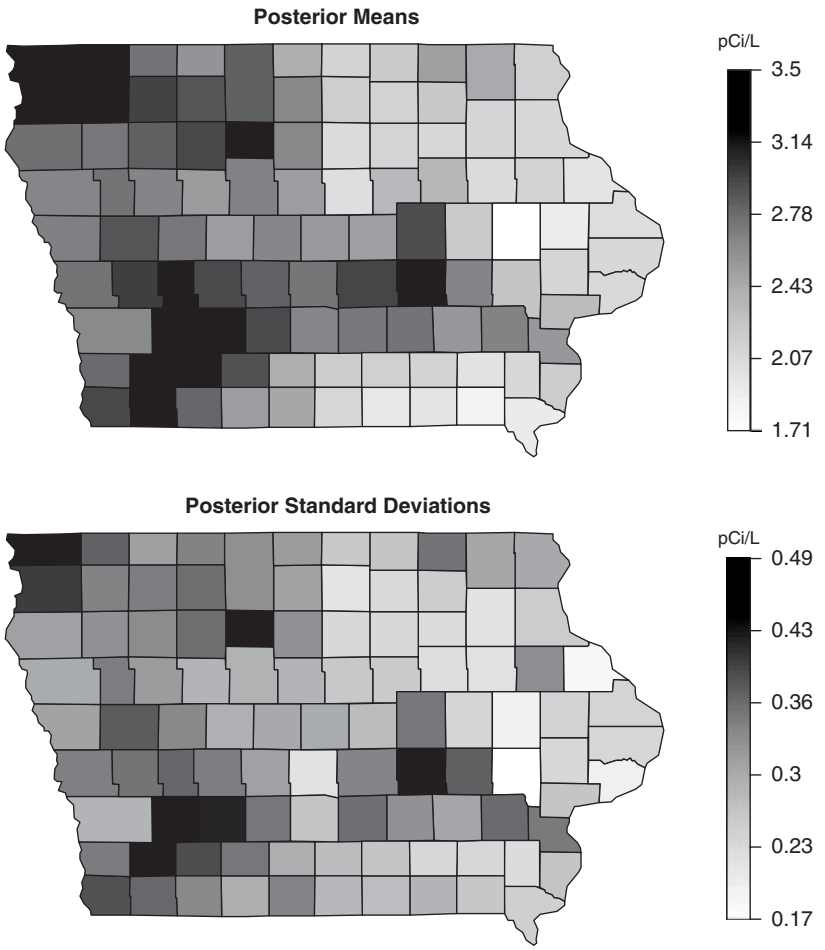

Figure 2. Posterior means and standard deviations for the predicted county-average, first floor radon concentrations, estimated from the combined analysis of all four leukaemia types.

and race effects were also examined in univariate analyses. Three calendar periods—1973–1982, 1983–1992, and 1993–2002—were considered. Since the number of cases for races other than black and white was very small in Iowa (less than 0.63 per cent for all four types of leukaemia), we compared black race to all others (non-black). Both year and race were significant at the 5 per cent level.

For the full Bayesian analyses, we included an indicator variable for black race, an integer variable indexing the three calendar year periods, linear and quadratic effects for urban influence code, and the predicted county-average, first-floor radon concentrations. County-specific disease rates are summarized in Figure 3 with maximum likelihood estimates of the age and gender-adjusted SMRs. Similarly, SMRs for the race categories, calendar periods, and county influence codes are included in Tables II and III, along with posterior summaries of the relative risks from the Bayesian analyses.

DOI: 10.1002/sim

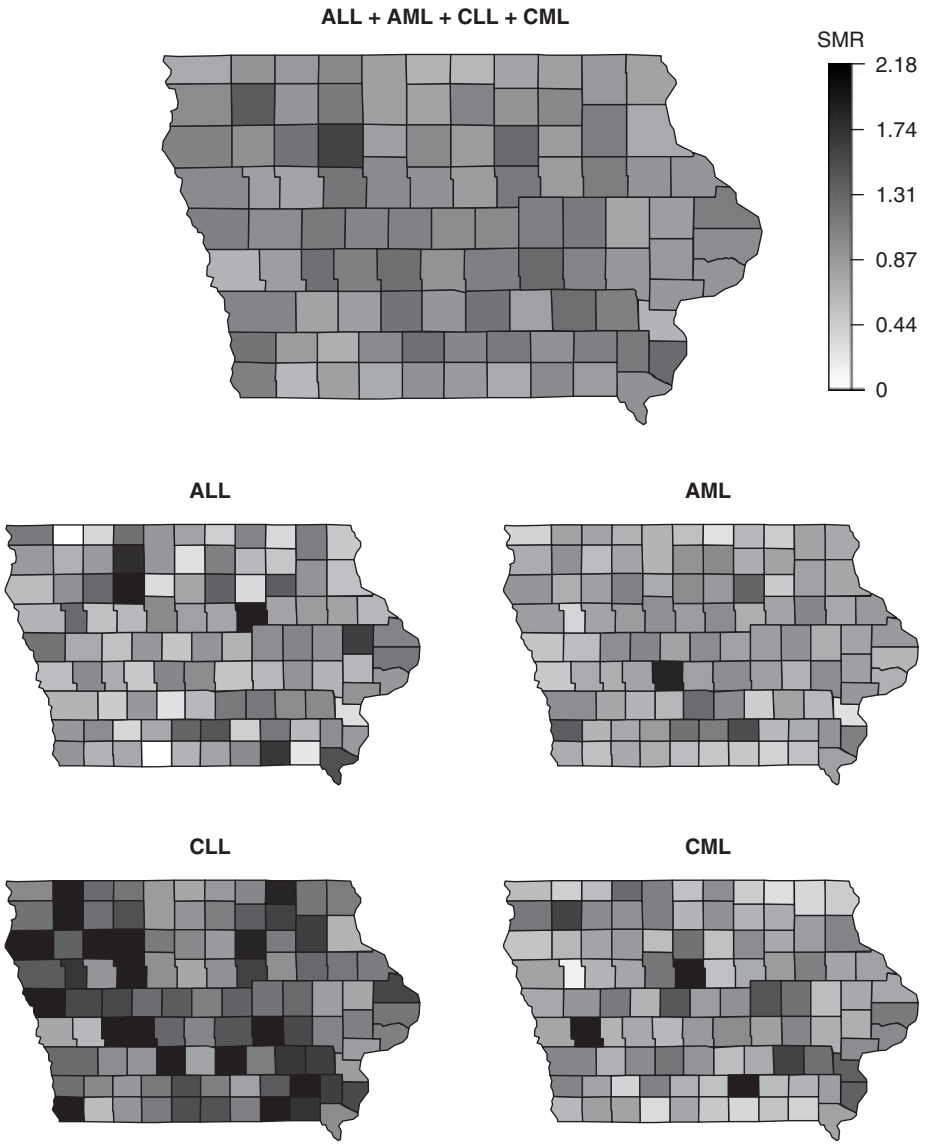

Figure 3. Spatial maps of the Iowa county leukaemia rates, age- and gender-standardized by the disease-specific rates from the nine original SEER registries.

Risk estimates for acute and chronic lymphocytic leukaemia tended to be lower in blacks, while estimates for the myelogenous leukaemia subtypes exhibited wide 95 per cent HPDs and did not provide strong evidence of an increased or decreased risk for blacks. We observed an increasing temporal trend in AML risk and, to a lesser extent, a decreasing trend in CML. Relative risk estimates were fairly homogeneous across levels of the county influence code variable, with the possible exception of increased AML and CML risk in the most rural counties.

DOI: 10.1002/sim

Table II. Summary statistics and posterior mean (95 per cent HPD) relative risk estimates from the Bayesian analyses for acute lymphocytic leukaemia (ALL) and acute myelogenous leukaemia (AML).

		ALL				AML			
Risk covariate	Values	y	$\widehat{\text{SMR}}^*$	Relative risk		y	$\widehat{\text{SMR}}^*$	Relative risk	
Race	Non-black	1189	1.07	1.0	†	3124	1.02	1.0	†
	Black	19	0.80	0.74	(0.42, 1.09)	30	0.96	0.92	(0.61, 1.27)
Year	1973–1982	408	1.03	1.0	†	931	0.96	1.0	†
	1983–1992	375	1.01	1.06	(0.98, 1.13)	929	0.91	1.12	(1.07, 1.17)
	1993–2002	425	1.15	1.12	(0.97, 1.28)	1294	1.18	1.25	(1.14, 1.36)
Urban influence code	2 (Urban)	609	1.07	1.0	†	1435	1.05	1.0	†
	5	81	0.98	0.96	(0.84, 1.07)	227	0.99	1.01	(0.94, 1.09)
	6	192	1.08	0.96	(0.83, 1.09)	543	1.03	1.02	(0.93, 1.11)
	7	27	0.90	0.97	(0.84, 1.10)	99	1.02	1.03	(0.94, 1.13)
	8	146	1.24	0.99	(0.86, 1.13)	356	1.02	1.05	(0.96, 1.15)
	9	90	0.99	1.02	(0.87, 1.17)	285	1.00	1.07	(0.97, 1.17)
	10	21	1.01	1.06	(0.87, 1.25)	74	1.03	1.09	(0.98, 1.22)
	11	35	0.94	1.11	(0.86, 1.37)	103	0.92	1.12	(0.97, 1.28)
	12 (Rural)	7	0.64	1.18	(0.82, 1.56)	32	0.81	1.16	(0.95, 1.37)
Radon	+1 pCi/L	‡	‡	0.91	(0.78, 1.03)	‡	‡	1.01	(0.92, 1.12)

*Standardized morbidity ratios are computed as the number of observed leukaemia cases in Iowa (y) divided by the expected number based on age- and gender-specific rates from the nine original SEER registries.

†Reference category for posterior relative risk estimates.

‡Only relative risk estimates are reported since radon is treated as a continuous covariate in the analyses.

Our model allows for the prediction of average radon concentrations outdoors, on each floor of the home, or any combination thereof. Field *et al.* [36] compared dosimetric models for radon exposure and observed the highest lung cancer risk estimates for ambient radon concentrations linked to individual subject-level mobility, followed by radon concentrations on the first floors of individual homes. Since mobility data are unavailable for our analyses, we relate leukaemia risk to first floor radon concentrations. In particular, an indicator variable for the first floor is included in the prediction model (4). Also note that radon appears as a linear effect in the risk model (1) which implies that the relative risk is constant at a given unit increase in radon.

Tables II and III include posterior summaries for a 1 pCi/L increase in the predicted county radon averages. In comparison to other measurement units for radon, 1 pCi/L is equivalent to 37 Bq/m^3 and an annual exposure of 0.16 Working Level Months (WLM) at an assumed average of 19 h/day spent indoors and an equilibrium ratio of radon decay products to radon of 40 per cent. The posterior means presented in the tables show a negative relationship between radon and ALL risk, and increasingly positive relationships for AML, CLL, and CML. Bayesian analyses further allow one to estimate the probability that risk is above or below a given value. Suppose, for example, that exposures are classified as important protective or risk factors if their associated relative risks are less than 0.95 or greater than 1.05, respectively. The probability of such occurrences can easily be estimated from the results of our analyses. In this case, there is a 0.759 probability that the ALL relative risk for a 1 pCi/L increase in radon is less than 0.95, whereas the probabilities that the

DOI: 10.1002/sim

Table III. Summary statistics and posterior mean (95 per cent HPD) relative risk estimates from the Bayesian analyses for chronic lymphocytic leukaemia (CLL) and chronic myelogenous leukaemia (CML).

		CLL				CML			
Risk covariate	Values	y	$\widehat{\text{SMR}}^*$	Relative risk		y	$\widehat{\text{SMR}}^*$	Relative risk	
Race	Non-black	5393	1.37	1.0	†	1679	1.07	1.0	†
	Black	29	0.89	0.67	(0.43, 0.92)	20	1.29	1.22	(0.70, 1.76)
Year	1973–1982	1601	1.30	1.0	†	563	1.14	1.0	†
	1983–1992	1967	1.49	1.01	(0.97, 1.05)	547	1.04	0.95	(0.90, 1.01)
	1993–2002	1854	1.30	1.02	(0.94, 1.10)	589	1.03	0.90	(0.80, 1.02)
Urban influence code	2 (Urban)	2182	1.29	1.0	†	746	1.06	1.0	†
	5	408	1.37	0.95	(0.88, 1.02)	153	1.29	0.94	(0.85, 1.04)
	6	1011	1.43	0.94	(0.87, 1.01)	287	1.05	0.95	(0.84, 1.06)
	7	193	1.46	0.94	(0.86, 1.02)	57	1.13	0.97	(0.86, 1.09)
	8	639	1.39	0.93	(0.86, 1.02)	203	1.13	1.01	(0.89, 1.13)
	9	565	1.47	0.94	(0.85, 1.02)	149	1.01	1.06	(0.93, 1.20)
	10	153	1.54	0.94	(0.86, 1.04)	34	0.91	1.13	(0.97, 1.31)
	11	193	1.28	0.96	(0.85, 1.07)	56	0.97	1.23	(1.01, 1.47)
	12 (Rural)	78	1.43	0.97	(0.83, 1.12)	14	0.69	1.36	(1.02, 1.71)
Radon	+1 pCi/L	‡	‡	1.06	(0.96, 1.16)	‡	‡	1.12	(0.98, 1.27)

*Standardized morbidity ratios are computed as the number of observed leukaemia cases in Iowa (y) divided by the expected number based on age- and gender-specific rates from the nine original SEER registries.

†Reference category for posterior relative risk estimates.

‡Only relative risk estimates are reported since radon is treated as a continuous covariate in the analyses.

corresponding relative risks exceed 1.05 for AML, CLL, and CML are 0.215, 0.568, and 0.849. The risk estimates presented thus far represent summary statistics based on samples drawn from the joint posterior distribution. Those samples were also used to construct posterior density plots for the leukaemia relative risks associated with a 1 pCi/L increase in radon, as shown in Figure 4.

Summaries of the parameters that were included in the risk model to account for extra-Poisson variability are given in Table IV. As specified in equation (1), statewide heterogeneity is represented by the $\boldsymbol{\theta}$ parameters which are assumed to be random draws from a normal distribution with zero mean and a standard deviation of σ_h. In order to provide insight into the σ_h parameter, consider the estimated posterior mean from the ALL analysis of 0.063 at which the $\theta_i \overset{iid}{\sim} \text{N}(0, 0.063^2)$. The 5th and 95th quantiles of this Normal distribution are ± 0.12 and would correspond to relative risks of 0.89 and 1.13 in the Poisson model. Thus, the relative risks associated with the heterogeneity parameters are of magnitudes similar to those for the demographic covariates. Spatial variability in the incidence rates that is not accounted for by the $\boldsymbol{\beta}$ parameters in equation (1) is characterized by the σ_c and ρ_c parameters in the CAR specification. We note that the values for these two must be interpreted with care since

$$\phi_i \mid \phi_j, j \neq i \sim \text{N}\left(\rho_c \sum_j c_{ij}\phi_j / c_{i+}, \sigma_c^2 / c_{i+}\right)$$

DOI: 10.1002/sim

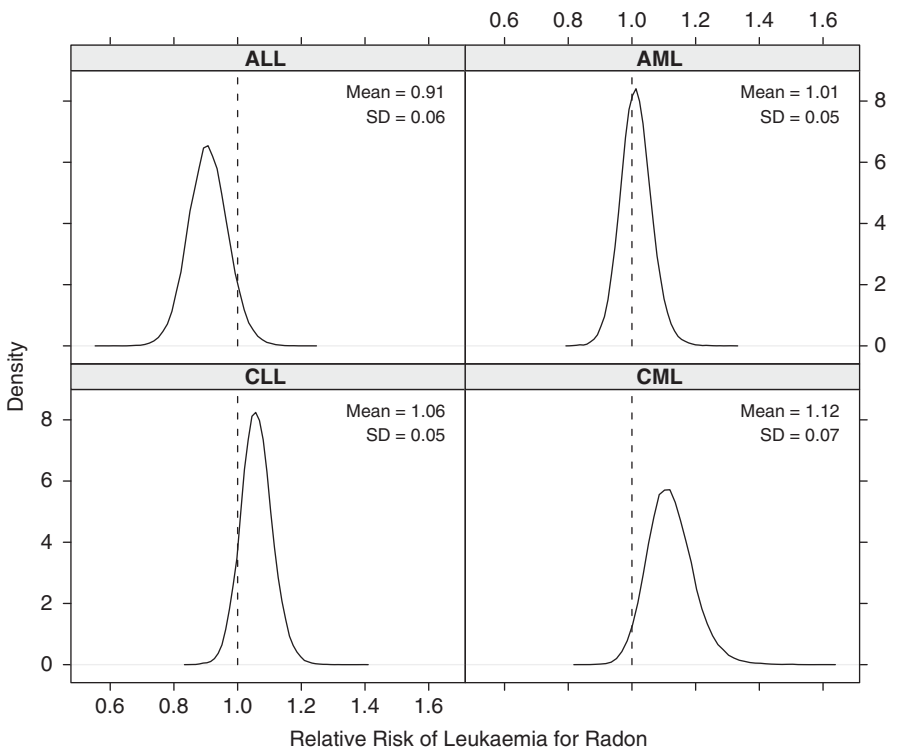

Figure 4. Posterior distribution of the leukaemia risk associated with a 1 pCi/L increase in county-average, first floor radon concentrations.

Table IV. Posterior estimates for the variance parameters in the Poisson risk model for incidence leukaemia cases.

	ALL		AML		CLL		CML	
Risk parameter	Mean	95 per cent HPD	Mean	95 per cent HPD	Mean	95 per cent HPD	Mean	95 per cent HPD
σ_h	0.063	(0.017, 0.123)	0.065	(0.019, 0.123)	0.125	(0.073, 0.175)	0.071	(0.017, 0.145)
σ_c	0.094	(0.015, 0.224)	0.103	(0.015, 0.226)	0.117	(0.017, 0.249)	0.092	(0.014, 0.219)
ρ_c	−0.390	(−1.651, 0.930)	−0.243	(−1.543, 1.000)	−0.090	(−1.488, 1.000)	−0.248	(−1.546, 1.000)

Specifically, σ_c by itself should not be interpreted as a measure of the strength of spatial association because $\rho_c = 0$ implies independence. To examine the dependencies among counties implied by the extra-Poisson variance parameters, we studied the posterior distribution of Moran's I spatial correlation coefficient [37] applied to the $\boldsymbol{\phi}$ model parameters. In all of the analyses, a consistently weak residual spatial association was observed, with CLL rates exhibiting the largest values for Moran's I (posterior mean = 0.02; 95 per cent HPD −0.25–0.35).

DOI: 10.1002/sim

4.4. Model comparisons

Our fully Bayesian analysis represents an ambitious attempt to model simultaneously residential radon and leukaemia risk in Iowa. We prefer such a modelling approach because it provides (1) an estimate of the joint posterior distribution of all model parameters and (2) leukaemia risk estimates that account for prediction error in the radon exposure covariate. There are, however, other reasonable approaches that can be taken.

To complete the analysis, we offer a comparison of our proposed model to three alternatives. We begin by noting that our joint modelling approach allows county leukaemia rates to inform about the spatial distribution of radon and hence the predicted radon exposure covariate in the leukaemia risk model. Since this has the potential to distort the exposure–disease relationship, we consider a second 'composite' model in which leukaemia rates are not allowed to inform about the radon distribution. The composite model has the same functional form as that given in Section 3 but differs from the original model with regards to the MCMC implementation. In particular, the MCMC algorithm is modified (as described in the Appendix) to produce a predictive distribution for radon that is independent of the county leukaemia counts. Alternatively, we consider a third model in which radon exposure is treated as a fixed covariate in the leukaemia risk analysis. We do this by first analysing only the radon data to obtain the predictive distribution for county-average radon exposures. Then, we take the posterior mean county radon averages to be fixed covariate values in subsequent Poisson regression analyses of the leukaemia data. Our fourth model in the comparison is the same as the third, except that population density weights are replaced with uniform weights in the averaging of county radon exposures.

Model comparisons are carried out with respect to the deviance information criterion (DIC), defined as

$$\mathrm{DIC} = E_{\vartheta}[D(\mathbf{y} \mid \vartheta)] + p_D$$

where ϑ represents the parameters in the Poisson risk model (1), $D(\mathbf{y} \mid \vartheta)$ is the associated deviance function, and p_D is a measure of the effective number of model parameters [38]. Smaller values of the deviance function are indicative of models that provide better fits to the data. The effective number of parameters is included in the DIC formula as a penalty term since the deviance function necessarily decreases as the number of parameters increases. Consequently, comparisons based on the DIC aim to strike a balance between model goodness-of-fit and parsimony. Preference is given to models with smaller DIC values.

Summarized in Table V are DIC results from analyses based on our proposed exposure-risk model (Model 1) as well as the three alternative models described previously (Models 2–4). Among the first three models which utilize county radon averages weighted by population density, our joint model provides the smallest DIC values.

In general, Model 3 may be preferred to Model 2 because the difference in DIC values for the ALL and CML analyses are most notably in favour of the former. Thus, Model 3 may be the preferred choice after Model 1. Recall that the third model differs from our joint approach in that county averages are computed separately and included as a fixed covariate in Poisson regression models. The risk estimates from this simplified approach are very close to those given in Tables II and III for the joint model. Specifically, for a 1 pCi/L increase in radon, Model 3 yields posterior mean (95 per cent HPD) relative risks estimates of 0.90 (0.78, 1.03), 1.01 (0.92, 1.11), 1.05 (0.96, 1.14), and 1.12 (0.99, 1.25) for ALL, AML, CLL, and CML, respectively. The similar estimates from the separate analysis of radon and leukaemia data suggest that the

DOI: 10.1002/sim

Table V. Comparison of deviance information criterion (DIC) for leukaemia risk models with different predictors of county-average radon.

Model	Radon covariate	Radon-risk modelling	Population density weights	DIC			
				ALL	AML	CLL	CML
1	Random	Joint	Yes	1145.4	1539.0	1765.7	1307.3
2	Random	Composite	Yes	1146.5	1539.8	1768.0	1309.7
3	Fixed	Separate	Yes	1145.7	1539.9	1768.3	1308.1
4	Fixed	Separate	No	1144.8	1540.0	1767.9	1308.2

Note: Smaller DIC values indicate preferable models.

exposure-disease relationship is not distorted in the joint model. From a computing perspective, Model 3 has the advantage of requiring shorter run times because the same radon estimates are used in each of the leukaemia analyses. Hence, the estimates need only be computed once. However, the MCMC algorithms for Model 3 are only marginally less complex than the algorithm for the joint model. There is almost complete overlap in the code used to fit the two models. Therefore, the programming burden is essentially the same.

Models 3 and 4 are very similar with respect to both the DIC values in Table V and the risk estimates (not shown). Thus, the use of population density in calculating county radon averages appears to have a negligible effect on the results. Indeed, the correlation between posterior mean county radon averages based on population density weights *versus* uniform weights is 0.99. The high correlation is not surprising given that the population density does not vary much within Iowa counties, which tend to be either predominantly urban or predominantly rural.

5. DISCUSSION

In the present study, a hierarchical Bayesian model is used to examine the relationship between aggregate (county) radon levels and incident leukaemia cases in Iowa. Application of our model to all four leukaemia types combined yields a relative risk estimate of 1.04 (95 per cent HPD 0.98–1.10) for a 1 pCi/L increase in the mean county-average radon. Previous studies of radon-exposed miners have provided somewhat equivocal findings concerning the risk posed by radon exposure. Roscoe [39] also reported statistically non-significant elevations for leukaemia (SMR = 1.6; 95 per cent CI 0.8–2.7) for a retrospective cohort study of 3238 uranium miners in the Colorado Plateau of the United States. A comprehensive pooled analyses by Darby *et al.* [40] of 11 retrospective cohort studies of radon-exposed underground miners noted a significant increased relative risk of leukaemia (RR = 1.93; 95 per cent CI 1.19–2.95) for cohorts of miners in their first 10 years of employment. While the overall risk of leukaemia was also positive, it was not significant (RR = 1.16; 95 per cent CI 0.90–1.47). A recently published retrospective cohort study of 23 043 Czech uranium miners by Řeřicha *et al.* [6] reported a significant association between radon exposure and leukaemia (p-value = 0.014). When the high categorical radon exposure of 110 WLM was compared to the referent low radon exposure category of 3 WLM, a RR of 1.75 (95 per cent CI 1.10–2.78) was found for the four types of leukaemia combined. Laurier *et al.* [41] have previously reviewed 19 ecologic studies published between 1987 and 2000 that examined the association between radon exposure and leukaemia. Overall, ecologic studies using various

DOI: 10.1002/sim

surrogate measures of residential radon exposures, such as average county radon concentrations based on short-term radon measurements, yielded fairly consistent results suggesting an association between radon exposure and the risk of leukaemia at the aggregate level.

The marginally negative association (RR = 0.91; 95 per cent HPD 0.78–1.03) reported in Table II between ALL and county radon level has been found elsewhere. Collman *et al.* [42] conducted an ecologic study using cancer mortality data from North Carolina, USA. The RR for male ALL was less than 1.0, although the relationship was non-significant. Using Poisson regression, Viel [34] also reported a negative association between radon and ALL in an ecologic study which accounted for such covariates as socio-economic status (estimated by the percentage of workmen in the employed population), a linear geographical gradient (expressed in terms of latitude and longitude for each area), and indoor gamma rays. In another study that examined radon exposure and childhood ALL at the postcode sector level in the UK counties of Devon and Cornwall, Thorne *et al.* [43] reported a non-significant (p-value = 0.28) reduced risk for those exposed to higher radon levels ($\geqslant 100\,\mathrm{Bq/m^3}$) as compared to lower levels ($<100\,\mathrm{Bq/m^3}$). Haque and Kirik [44], employing linear regression on data from the UK, also found a positive, but non-significant, correlation between radon concentration and ALL.

The null relative risk of 1.01 (95 per cent HPD 0.92–1.12) between AML and radon level reported in this study has been noted by previous authors. Toti *et al.* [15] used a Bayesian hierarchical model to analyse data from a case-control study and did not find an association between adult myeloid leukaemia and indoor radon concentration (OR $>185\,\mathrm{Bq/m^3}$ *versus* $<80\,\mathrm{Bq/m^3} = 1.0$, 95 per cent Cr.I. 0.2–2.9). Also using a case-control study design, Steinbuch *et al.* [45] reported that indoor residential radon exposure was not found to be associated with risk of AML among children. The inverse association between radon level and AML risk among those <2 years was attributed to chance. Forastiere *et al.* [46] used conditional logistic regression in their case-control study to estimate odds ratios for radon, gamma radiation, municipality, and dwelling characteristics. No significant association between adult myeloid leukaemia and exposure to indoor radon was detected. Other ecologic analyses support a positive relationship between radon and AML [34, 44, 47].

In the current study, we report evidence of a positive association between county radon levels and both CLL (RR = 1.06; 95 per cent HPD 0.96–1.16) and CML (1.12; 0.98–1.27). Studies investigating a possible association between radon exposure and the subsequent development of either CLL or CML are less frequent. In ecologic analyses, Alexander *et al.* [48] reported non-significant positive findings for both CLL and CML using census data, while Haque and Kirk [44] found a significant positive association between radon and both leukaemia types. In their collaborative analysis of 11 cohort studies of miners, Darby *et al.* [40] reported a positive non-significant relationship between radon exposure and CLL. In the recent retrospective cohort study of Czech uranium miners by Řeřicha *et al.* [6], the incidence of CLL was positively associated with radon exposure (p-value = 0.016). When the authors compared a high radon exposure (110 WLM) to a low exposure (3 WLM), they found a relative risk of 1.98 (CI 1.10–3.59) for CLL. CML exhibited a similar, but non-significant, association. To our knowledge, the paper by Řeřicha *et al.* is the first analytic epidemiologic study to find a statistically significant increased risk of CLL for any type of radiation exposure.

While the findings of ecologic analyses should be limited to hypothesis generating and cannot be used to determine risk, our hierarchical Bayesian modelling approach offers several advantages over previous ecologic studies. The first advantage lies in our model for the radon data that allows for prediction of county-average radon levels while accounting for spatial dependencies, systematic differences between homes, and detector measurement error. Secondly, the Bayesian analysis

DOI: 10.1002/sim

provides the joint distribution of all model parameters. Since the parameters are modelled simultaneously, relative risk estimates accurately reflect uncertainties in the predicted radon exposure covariate. The third advantage is that the resulting joint distribution allows probability statements to be made about the model parameters. For instance, parameters fall within the reported 95 per cent HPD intervals with probability equal to 0.95. Likewise, statements can be made about the probability that a relative risk ranges above or below a given value, as illustrated in Section 4.3. Finally, since age- and gender-specific rates are modelled in the Poisson regression, we are able to consider leukaemia in all age groups instead of focusing on only children or only adults. Alternatively, the effects of race, calendar period, and urban/rural regions are controlled for with additional covariates in the risk model.

The study has several other strengths. For example, Iowa has the highest mean residential radon concentration in the United States and provides a wide distribution of county radon concentrations for use in the analyses. In addition, the National Cancer Institute, Surveillance Epidemiology and End Results registry provides highly accurate information on the incidence of leukaemia subtypes occurring historically in Iowa's 99 counties.

In conclusion, we present a Bayesian hierarchical risk model to characterize the effects of a radon exposure covariate that cannot be measured directly. Although the model is applied to population-level incidence rates, our approach can be generalized to individual-level risk models for which spatial dependencies and measurement error are considerations in the prediction of an exposure covariate. The risk estimates presented in this study, performed in a state with high radon levels, as well as the recent findings from Řeřicha *et al.* suggest that an analytic study examining the risk posed by prolonged exposure to residential radon progeny and the subsequent development of leukaemia may be warranted.

APPENDIX A: THE MCMC ALGORITHM

Markov chain Monte Carlo sampling is an iterative numerical method for generating correlated draws from the joint posterior distribution of model parameters [49]. The general idea is to draw samples from the posterior full conditional distributions of the parameters at each MCMC iteration. Existing random number generators can be used to obtain draws from full conditionals that are standard distributions. Such is the case for the variance parameters σ^2 and the mean outdoor radon parameter β_{os} in our model. Although the full conditionals for the correlation parameters ρ_c and ρ_s are of non-standard forms, we are able to utilize shrinkage slice sampling [50] to easily and efficiently sample these. For the remaining parameters, sampling is carried out with Metropolis–Hastings algorithms.

Sampling of the mean home radon parameters $\boldsymbol{\beta}_{hm}$ and the latent spatial parameters $(\mathbf{z}, \mathbf{z}')$ poses a unique challenge because each set appears in multiple levels of the model. Recall that the predicted county-average radon level $\bar{r}'_k$ is defined in (4) as a function of $\boldsymbol{\beta}_{hm}$ and $\mathbf{z}'$. Hence, the home mean parameters appear in both the radon measurement model (3) as well as the risk model (1). Likewise, the latent spatial parameters appear in both the spatial model (5) and risk model (1).

Before giving the details of our Metropolis–Hastings implementations, we first describe the algorithm in general. Consider observed data y and a partitioning of the model parameters ϑ into d subvectors such that $\vartheta = (\vartheta_1, \ldots, \vartheta_d)$. Values for the subvectors are to be sampled sequentially at each iteration t of the MCMC sampler. Various numerical sampling methods are available. The Metropolis–Hastings algorithm is one such method and involves the generation of new *candidate*

DOI: 10.1002/sim

points from a *proposal* distribution according to the following algorithm:

(a) Draw a candidate point ϑ_i^* from the proposal distribution $q_i(\vartheta_i, \vartheta^{t-1})$ where $\vartheta^{t-1}=(\vartheta_1^t, \ldots, \vartheta_{i-1}^t, \vartheta_i^{t-1}, \ldots, \vartheta_d^{t-1})$ is the vector of current values for all subvector parameters.
(b) Compute the ratio

$$r = \frac{p(\vartheta_i^* \mid \vartheta_{(i)}^{t-1}, y) q_i(\vartheta_i^{t-1}, \vartheta_i^*)}{p(\vartheta_i^{t-1} \mid \vartheta_{(i)}^{t-1}, y) q_i(\vartheta_i^*, \vartheta_i^{t-1})}$$

where p is the full conditional distribution of ϑ_i, and $\vartheta_{(i)}^{t-1} = (\vartheta_1^t, \ldots, \vartheta_{i-1}^t, \vartheta_{i+1}^{t-1}, \ldots, \vartheta_d^{t-1})$ is the vector of current values for all parameters other than ϑ_i.
(c) Set

$$\vartheta_i^t = \begin{cases} \vartheta_i^* & \text{with probability } \min(r, 1) \\ \vartheta_i^{t-1} & \text{otherwise} \end{cases}$$

Any distribution from which samples are readily obtainable may serve as the proposal distribution. However, some choices will work better than others, depending on the given problem.

We now turn to the specific use of Metropolis–Hastings in our MCMC algorithm. The posterior full conditional for $\boldsymbol{\beta}_{hm}$ can be written as the product of the risk model likelihood and a function $p(\boldsymbol{\beta}_{hm} \mid \sigma_{bh}^2, \mathbf{z}, \sigma_{wh}^2, \mathbf{r}_{hm})$ which has a known multivariate normal distribution. We use the latter as our proposal distribution to generate a candidate point $\boldsymbol{\beta}_{hm}^*$ in the Metropolis–Hastings step. Consequently, the associated acceptance probability simplifies to

$$r = \frac{\prod_k \prod_l f(y_{kl} \mid \boldsymbol{\beta}, \beta_r, \bar{r}_k'^*, \theta_{kl}, \phi_k)}{\prod_k \prod_l f(y_{kl} \mid \boldsymbol{\beta}, \beta_r, \bar{r}_k', \theta_{kl}, \phi_k)}$$

where

$$\bar{r}_k'^* = \exp\left\{ \boldsymbol{\beta}_{hm}^{*\mathrm{T}} \mathbf{x}_{hm} + \sum_{i=1}^{L_k} w(s_i') z(s_i') \Big/ \sum_{i=1}^{L_k} w(s_i') \right\}$$

and can be calculated easily at each MCMC iteration. The same approach can be applied to generate candidate points $(\mathbf{z}^*, \mathbf{z}'^*)$ for the latent spatial parameters. In particular, the associated full conditional can be factored into a product of the risk likelihood and a known multivariate normal distribution that we use as the proposal. The resulting acceptance probability r has the same form as that given above. However, in the sampler for the spatial parameters, the predicted radon average is evaluated at the candidate point as follows:

$$\bar{r}_k'^* = \exp\left\{ \boldsymbol{\beta}_{hm}^{\mathrm{T}} \mathbf{x}_{hm} + \sum_{i=1}^{L_k} w(s_i') z^*(s_i') \Big/ \sum_{i=1}^{L_k} w(s_i') \right\}$$

Note that county leukaemia counts may inform about the radon distribution since the $\boldsymbol{\beta}_{hm}$ and $\mathbf{z}'$ parameters appear in the risk likelihood. To prevent this effect of the leukaemia data, one could set the acceptance probability to $r = 1$ in the aforementioned Metropolis–Hastings steps for the

DOI: 10.1002/sim

radon parameters, although the MCMC results would be something other than samples from the joint posterior distribution of all model parameters. Proposal distributions for Metropolis–Hastings sampling of the remaining parameters in the risk model mean structure are constructed using the methods of Gamerman [51].

ACKNOWLEDGEMENTS

This publication was made possible in part by grant numbers R01 ES05653 and P30 ES05605 from the National Institute of Environmental Health Sciences and grant number R01 CA85942 from the National Cancer Institute, NIH. The authors are particularly grateful to Dr. Charles Lynch, Department of Epidemiology, University of Iowa for providing this study with information on the incidence of leukaemia.

REFERENCES

1. Jemal A, Siegel R, Ward E, Murray E, Xu J, Smigal C, Thun MJ. Cancer statistics. *CA*: *A Cancer Journal for Clinicians* 2006; **56**:106–130.
2. Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ. Cancer statistics. *CA*: *A Cancer Journal for Clinicians* 2005; **55**:10–30.
3. Korte JE, Hertz-Picciotto I, Schulz MR, Ball LM, Duell EJ. The contribution of benzene to smoking-induced leukemia. *Environmental Health Perspectives* 2000; **108**(4):333–339.
4. March HC. Leukemia in radiologists. *Radiology* 1944; **43**:275–278.
5. NRC. *Assessment of the Scientific Information for the Radiation Exposure Screening and Education Program*. Board on Radiation Effects Research, Division on Earth and Life Studies, National Research Council, National Academy of Science. The National Academies Press: Washington, DC, 2005.
6. Řeřicha V, Kulich M, Řeřicha R, Shore DL, Sandler DP. Incidence of leukemia, lymphoma, and multiple myeloma in Czech uranium miners: a case-cohort study. *Environmental Health Perspectives* 2006; **114**(6):818–822.
7. Kendall GM, Smith TJ. Doses to organs and tissues from radon and its decay products. *Journal of Radiological Protection* 2002; **22**:389–406.
8. Oberdorster G, Oberdorster E, Oberdorster J. Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. *Environmental Health Perspectives* 2005; **113**(7):823–839.
9. Fung KY, Krewski D. On measurement error adjustment methods in Poisson regression. *Environmetrics* 1999; **10**(2):213–224.
10. Heid M, Küchenhoff H, Wellmann J, Gerken M, Kreienbrok L, Wichmann HE. On the potential of measurement error to induce differential bias on odds ratio estimates: an example from radon epidemiology. *Statistics in Medicine* 2002; **21**(21):3261–3278.
11. Lubin JH, Boice JD, Samet JM. Errors in exposure assessment, statistical power and the interpretation of residential radon studies. *Radiation Research* 1995; **144**:329–341.
12. Reeves GK, Cox DR, Darby SC, Whitley E. Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine* 1998; **17**(19):2157–2177.
13. Stidley CA, Samet JM. Assessment of ecologic regression in the study of lung cancer and indoor radon. *American Journal of Epidemiology* 1994; **139**(3):312–322.
14. Richardson S, Monfort C, Green M, Draper G, Muirhead C. Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. *Statistics in Medicine* 1995; **14**(21–22):2487–2501.
15. Toti S, Biggeri A, Forastiere F. Adult myeloid leukaemia and radon exposure: a Bayesian model for a case-control study with error in covariates. *Statistics in Medicine* 2005; **24**(12):1849–1864.
16. Zhu L, Carlin BP, Gelfand AE. Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmetrics* 2003; **14**(5):537–557.
17. Field RW, Steck DJ, Smith BJ, Brus CP, Fisher EL, Neuberger JS, Platz CE, Robinson RA, Woolson RF, Lynch CF. Residential radon gas exposure and lung cancer: the Iowa radon lung cancer study. *American Journal of Epidemiology* 2000; **151**:1091–1102.
18. Field RW, Lynch CF, Steck DJ, Fisher EL. Dosimetry quality assurance: Iowa residential radon lung cancer study. *Radiation Protection Dosimetry* 1998; **78**:295–303.

DOI: 10.1002/sim

19. NCI. *Surveillance*, *Epidemiology*, *and End Results* (*SEER*) *Program*, *SEER*Stat Database: Incidence—SEER 9 Regs Public-use* (*1973–2002*). National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, April 2005. URL: http://www.seer.cancer.gov. Based on the November 2004 submission.
20. Bureau of Health Professions. Area Resource File (ARF), February 2004.
21. Banerjee S, Carlin BP, Gelfan AE. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC: New York, Boca Raton, FL, 2004.
22. Smith BJ, Field RW. Effects of housing factors and surficial uranium on spatial prediction of residential radon in Iowa. *Environmetrics* 2006.
23. Fisher EL, Field RW, Smith BJ, Lynch CF, Daniel J. Spatial variation of residential radon concentrations: the Iowa radon lung cancer study. *Health Physics* 1998; **75**(5):506–513.
24. Banerjee S. Essential geodesics for the spatial statistician. *Technical Report 009*, Division of Biostatistics, University of Minnesota, 2003.
25. Cressie NAC. *Statistics for Spatial Data* (revised edn). Wiley-Interscience: New York, 1993.
26. Morokoff WJ, Caflisch RE. Quasi-Monte Carlo integration. *Computational Physics* 1995; **122**:218–230.
27. Lawson CL, Hanson RJ, Kincaid DR, Krogh FT. Basic linear algebra subprograms for Fortran usage. *ACM Transactions on Mathematical Software* 1979; **5**(3):308–323. URL: http://doi.acm.org/10.1145/355841.355847
28. Whaley RC, Petitet A, Dongarra JJ. Automated empirical optimization of software and the ATLAS project. *Parallel Computing* 2001; **27**(1–2):3–35.
29. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**:457–511.
30. Chen MH, Shao QM. Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* 1999; **8**(1):69–92.
31. Smith BJ. Bayesian Output Analysis Program (BOA), version 1.1.5. http://www.public-health.uiowa.edu/boa, 23 March 2005.
32. Steck DJ, Field RW, Lynch CF. Exposure to atmospheric radon. *Environmental Health Perspectives* 1999; **107**:123–127.
33. Cook-Mozaffari PJ, Darby SC, Doll R, Forman D, Hermon C, Pike MC, Vincent T. Geographical variation in mortality from leukaemia and other cancers in England and Wales in relation to proximity to nuclear installations. *The British Journal of Cancer* 1989; **59**:476–485.
34. Viel JF. Radon exposure and leukemia in adulthood. *International Journal of Epidemiology* 1993; **22**(4):627–631.
35. Wolff SP. Leukemia risks and radon. *Nature* 1991; **352**(6333):288.
36. Field RW, Smith BJ, Steck DJ, Lynch CF. Residential radon exposure and lung cancer: variation in risk estimates. *Journal of Exposure Analysis and Environmental Epidemiology* 2002; **12**(3):197–203.
37. Moran PAP. Notes on continuous stochastic phenomena. *Biometrika* 1950; **37**(1/2):17–23.
38. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with Discussion). *Journal of the Royal Statistical Society*, *Series B* 2002; **64**(4):583–639.
39. Roscoe RJ. An update of mortality from all causes among white uranium miners from the Colorado plateau study group. *American Journal of Industrial Medicine* 1997; **31**(2):211–222.
40. Darby SC, Whitley E, Howe GR, Hutchings SJ, Kusiak RA *et al.* Radon cancers other than lung cancer in underground miners: a collaborative analysis of 11 studies. *Journal of the National Cancer Institute* 1995; **87**:378–384.
41. Laurier D, Valenty M, Tirmarche M. Radon exposure and the risk of leukemia: a review of epidemiological studies. *Health Physics* 2001; **81**(3):272–288.
42. Collman GW, Loomis DP, Sandler DP. Radon-222 concentration in groundwater and cancer mortality in North Carolina. *International Archives of Occupational and Environmental Health* (*Historical Archive*) 1988; **61**(1–2):13–18.
43. Thorne R, Foreman NK, Mott MG. Radon in Devon and Cornwall and paediatric malignancies. *European Journal of Cancer* 1996; **32**(2):282–285.
44. Haque AKMM, Kirk AE. Environmental radon and cancer risk. *Radiation Protection Dosimetry* 1992; **45**(1): 639–642.
45. Steinbuch M, Weinberg CR, Buckley JD, Robison LL, Sandler DP. Indoor residential radon exposure and risk of childhood acute myeloid leukaemia. *The British Journal of Cancer* 1999; **81**(5):900–906.
46. Forastiere F, Sperati A, Cherubini G, Miceli M, Biggeri A, Axelson O. Adult myeloid leukaemia, geology, and domestic exposure to radon and gamma radiation: a case control study in central Italy. *Occupational and Environmental Medicine* 1998; **55**(2):106–110.

DOI: 10.1002/sim

47. Evrard AS, Hemon D, Billon S, Laurier D, Jougla E, Tirmarche M, Clavel J. Ecological association between indoor radon concentration and childhood leukaemia incidence in France, 1990–1998. *European Journal of Cancer Prevention* 2005; **14**(2):147–157.
48. Alexander FE, McKinney PA, Cartwright RB. Radon and leukaemia (letter). *Lancet* 1990; **335**:1008–1012.
49. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**:97–109.
50. Neal RM. Slice sampling. *Annals of Statistics* 2003; **31**:705–767.
51. Gamerman D. Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* 1997; **7**:57–68.

DOI: 10.1002/sim