Iowa Radon Leukemia Study: A Hierarchical Population Risk Model for Spatially Correlated Exposure Measured with Error

Brian J. Smith
Lixun Zhang
William R. Field

Department of Biostatistics
College of Public Health
The University of Iowa

ENAR Spring Meeting
March 13, 2007

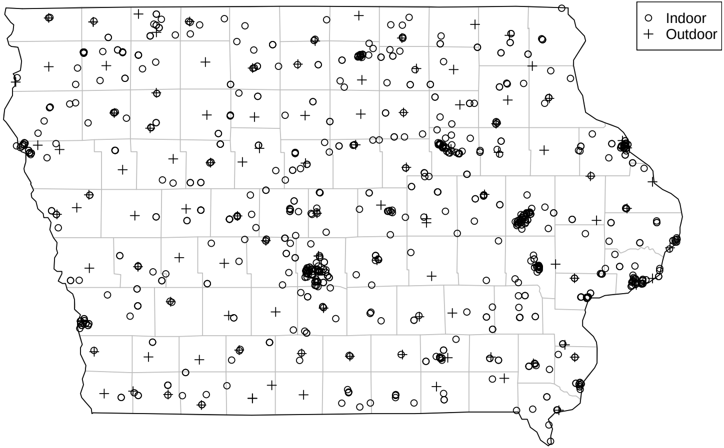
Introduction

- Man-made sources of exposure to ionizing radiation (x-rays and gamma rays) have been implicated as causative agents in several forms of Leukemia.
- Radon is a naturally occurring gas that is the largest contributor to individuals' background exposures to radiation.
- The association between leukemia risk and individuals' exposure to residential radon has not been studied.
- Ecologic radon studies examine associations between disease outcomes and exposures aggregated over geographic regions.
- Exposures in such studies cannot be observed directly and are generally predicted from samples of radon measurements within geographic regions.

Iowa Radon Lung Cancer Study

- Epidemiologic case-control study in Iowa to estimate the effect of residential radon on lung cancer risk.
- 2,590 radon detectors were installed in 614 population-based, control subject homes.
- At least one measurement was taken on each floor of the home, for an average of 4.2 measurements per home.
- Distribution of home sites is similar to that of the general population.
- 136 radon measurements collected from an approximately uniform grid of 109 outdoor sites.

Geographic Radon Monitoring Sites



SEER Cancer Registry

- The NCI Surveillance, Epidemiology, and End Results (SEER) Program collects cancer incidence and survival data.
- Our dataset includes incidence cancers from 1973–2002 in Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco-Oakland, and Utah.
- Leukemia and total population counts by disease type, county, 5-year age strata, gender, race, and calendar year.
- Leukemia cases in Iowa included:
  - 1,189 Acute Lymphocytic Leukemias (ALL)
  - 3,124 Acute Myelogenous Leukemias (AML)
  - 5,393 Chronic Lymphocytic Leukemias (CLL)
  - 1,679 Chronic Myelogenous Leukemias (CML)

Standardized Mortality Ratio

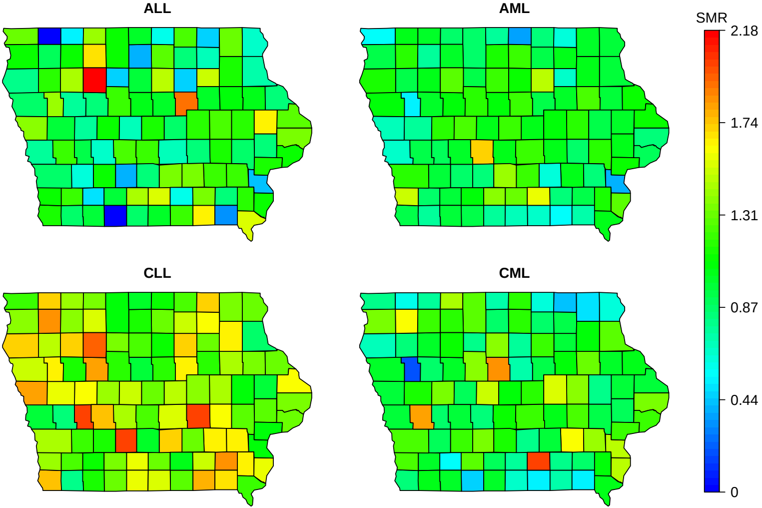
- The expected number of cases in the k^{th} county and l^{th} demographic subgroup is denoted

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

where the summation is over age-gender strata, m_{kli} is the person-years at risk, and λ_i is the stratum-specific rate based on all SEER registry data.

- The observed standardized mortality ratios (SMR) are the number of observed / expected cases and are plotted by Iowa counties on the following slide.

Age and Gender-Standardized Leukemia SMRs



Analysis Goals

- Study the effect of residential radon on leukemia risk in Iowa, while controlling for potential confounding effects of important covariates.
- Relate point-referenced radon and areal leukemia data using a geostatistical model to characterize the continuous spatial distribution of radon.
- Apply a hierarchical Bayesian model to simultaneously predict county-average radon and estimate the associated leukemia risk.
- Resulting risk estimates accurately reflect the uncertainty in predicting radon.

Leukemia Risk Model

- The reported number of cases is assumed to have the following Poisson distribution

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

- $\mathbf{x}_{kl}$ is a vector of known covariates with mean parameters $\boldsymbol{\beta}$.
- β_r is the effect of the predicted county-average radon $\bar{r}'_k$ on leukemia rate.
- θ_{kl} accounts for state-wide heterogeneity and ϕ_k spatial correlation not due to radon.

Outdoor Radon Model

- Let $r_{os,ij}$ denote the j^{th} radon measurement from outdoor site $i = 1, \ldots, n_{os}$.
- We specify the following model for outdoor measurement:

$$\begin{aligned} \ln r_{os,ij} &= \beta_{OS} + z\left(s_i\right) + \epsilon_{os,ij} \\ \epsilon_{os,ij} &\overset{iid}{\sim} N\left(0, \sigma^2_{os}\right) \end{aligned}$$

  where β_{os} is an overall mean parameter, $\epsilon_{os,ij}$ is an exchangeable error term, and σ^2_{os} is the measurement error variance.
- The $z\left(s_i\right)$ parameter accounts for spatial correlation among radon measurements.

Home Radon Model

- Let $r_{hm,ij}$ denote the j^{th} radon measurement from home site $i = 1, \ldots, n_{hm}$.
- We model home radon measurements as

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

where $\mathbf{x}_{hm,ijk}$ and $\boldsymbol{\beta}_{hm}$ are vectors of covariates and corresponding mean parameters; γ_i is a random effect for the home; $z\left(s_i\right)$ is the latent spatial parameter; and $\epsilon_{hm,ij}$ is an exchangeable error term.

Home Radon Model

- In the present analysis, only indicator variables for the floors on which the measurements were taken are included in the covariate vector $\mathbf{x}_{hm,ij}$.
- The error variance σ^2_{wh} is a combination of variability due to systematic differences within the home and random detector measurement error.
- The between-home variance σ^2_{bh} includes unaccounted for differences in home environments.

Latent Spatial Process for Radon

- Let $\mathbf{z}^T = (z(s_1), \ldots, z(s_n))$, for $i = 1, \ldots, n$ and $n = n_{OS} + n_H$.
- The assumed distribution for these latent spatial parameters is

$$\mathbf{z} \sim N\left(\mathbf{0}, \sigma_s^2 R_{\mathbf{z}}(\rho_s)\right)$$

where $R_{\mathbf{z}}(\rho_s)$ is correlation matrix; σ_s^2 is a scalar variance; and ρ_s is a correlation parameters such that

$$\left(R_{\mathbf{z}}(\rho_s)\right)_{ii'} = \exp\left\{-\left\|s_i - s_{i'}\right\|^2 / \rho_s^2\right\}.$$

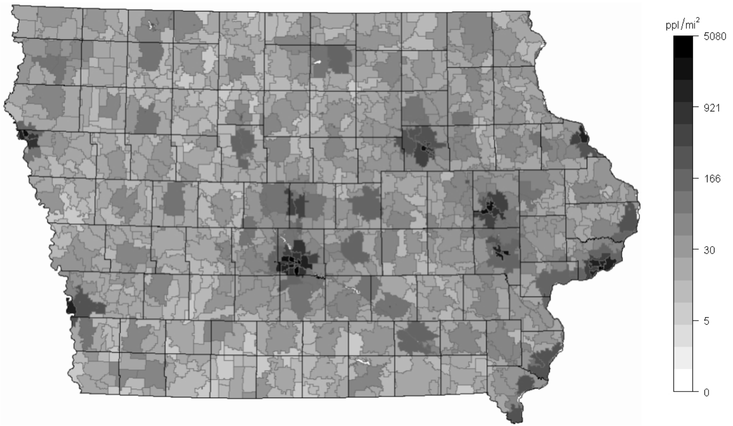
Predicted County-Average Radon

- The county radon concentrations $\bar{r}'_k$ is needed to complete the specification of our risk model.
- For this, we use the average predicted radon concentration defined as

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

- B_k denotes the geographic region.
- $r'(s)$ is a predicted radon concentration at site s.
- $w(s)$ is a function of location that allows for weighted averaging of radon concentrations within a county.
- We set $w(s)$ to be the population densities (people per square mile) for Iowa state zip codes, as reported in the 2000 U.S. Census.

Population Density by Zip Codes



Numerical Integration

- County-average, first-floor radon concentration is the primary radon exposure of interest.
- The previous integral cannot be solved explicitly due to the irregular shape of geographic regions.
- An alternative is to replace the integration by summation over a new set of sites $\{s_i' : i = 1, \ldots, L_k\}$ so that

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

- We employ a fixed grid of approximately 3,000 equally spaced sites $\{s_i'\}$ across Iowa, thus partitioning the state into 18.5 square-mile regions.

Predictive Distribution

- The joint distribution for the latent spatial parameters at observed and unobserved grid sites is

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

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

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

and variance

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

Joint Posterior Distribution

A fully Bayesian approach is taken to obtain the joint posterior distribution of all model parameters.

$$\left[\prod_k \prod_l f\left(y_{kl} \mid \boldsymbol{\beta}, \beta_r, \bar{r}'_k, \theta_{kl}, \phi_k\right)\right] f\left(\boldsymbol{\theta} \mid \sigma_h^2\right) f\left(\boldsymbol{\phi} \mid \sigma_c^2, \rho_c\right) \pi\left(\boldsymbol{\beta}, \beta_r, \sigma_h^2, \sigma_c^2, \rho_c\right)$$
$$\times \left[\prod_i \prod_j f\left(\ln r_{os,ij} \mid \beta_{os}, \mathbf{z}, \sigma_{os}^2\right)\right] \pi\left(\beta_{os}, \sigma_{os}^2\right)$$
$$\times \left[\prod_i \prod_j f\left(\ln r_{hm,ij} \mid \boldsymbol{\beta}_{hm}, \sigma_{bh}^2, \mathbf{z}, \sigma_{wh}^2\right)\right] \pi\left(\boldsymbol{\beta}_{hm}, \sigma_{bh}^2, \sigma_{wh}^2\right)$$
$$\times f\left(\mathbf{z}' \mid \mathbf{z}, \sigma_s^2, \rho_s\right) f\left(\mathbf{z} \mid \sigma_s^2, \rho_s\right) \pi\left(\sigma_s^2, \rho_s\right)$$

where $\bar{r}'_k$ is a function of the home radon mean parameters $\boldsymbol{\beta}_{hm}$ and the predicted latent parameters $\mathbf{z}'$.

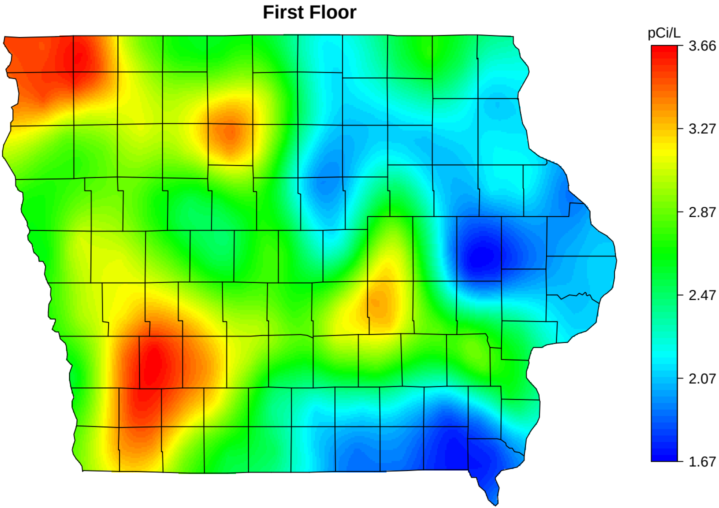
Prior Specification and MCMC Methods

- Vague, independent priors:

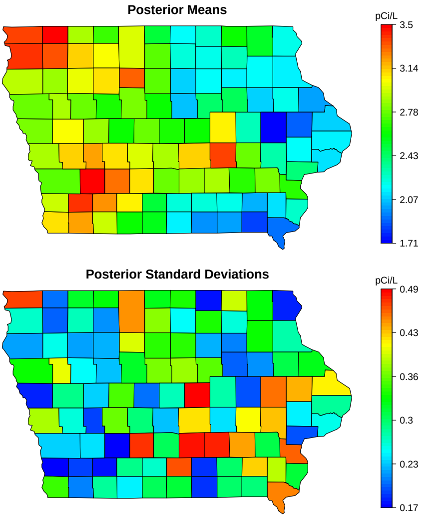
Mean	β	$\sim$	*Normal* $(0, 1000)$
Variance	$1/\sigma^2$	$\sim$	*Gamma* $(0.001, 0.001)$
CAR	ρ_c	$\sim$	*Uniform* $(-1.78, 1.00)$
Spatial Decay	ρ_s	$\sim$	*Uniform* $(0, 100)$

- Markov chain Monte Carlo methods used to sample from the joint posterior.
- Simulated three parallel chains of 50,000 iterations each.
- Analyses performed with a version of R compiled against an ATLAS-tuned BLAS matrix algebra library.

Posterior Mean First Floor Radon



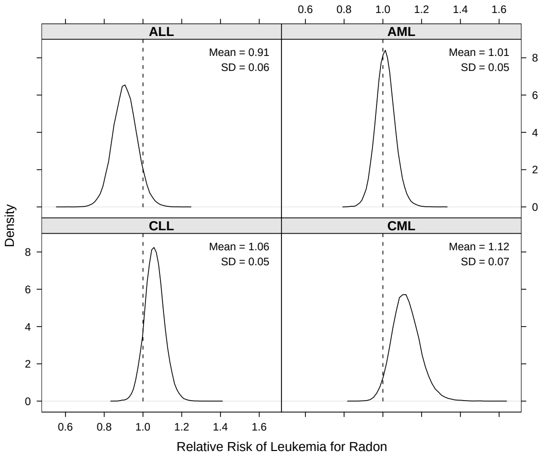
Posterior Mean First Floor County-Average Radon



Posterior Risk Estimates

Covariate		ALL		AML	
		Mean	95% HPD	Mean	95% HPD
Race	Non-black	1.0	-	1.0	-
	Black	0.74	(0.42, 1.09)	0.92	(0.61, 1.27)
Year	1973–1982	1.0	-	1.0	-
	1983–1992	1.06	(0.98, 1.13)	1.12	(1.07, 1.17)
	1993–2002	1.12	(0.97, 1.28)	1.25	(1.14, 1.36)
		ALL		AML	
Race	Non-black	1.0	-	1.0	-
	Black	0.67	(0.43, 0.92)	1.22	(0.70, 1.76)
Year	1973–1982	1.0	-	1.0	-
	1983–1992	1.01	(0.97, 1.05)	0.95	(0.90, 1.01)
	1993–2002	1.02	(0.94, 1.10)	0.90	(0.80, 1.02)

Posterior Distributions of Leukemia Risk for 1 pCi/L Increase in Radon



Discussion

- The Iowa study offers several advantages over previous ecologic radon studies.
- Bayesian approach allows for prediction of county-average radon levels while accounting for spatial dependencies, systematic differences between homes, and detector measurement error.
- Joint distribution of all model parameters.
- Relative risk estimates accurately reflect uncertainties in the predicted radon exposure covariate.
- 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.
- SEER registry provides accurate information on the leukemia incidence in all counties.
- Ecologic findings should be limited to hypothesis generating and not to determine risk.

Discussion

- The Iowa study offers several advantages over previous ecologic radon studies.
- Bayesian approach allows for prediction of county-average radon levels while accounting for spatial dependencies, systematic differences between homes, and detector measurement error.
- Joint distribution of all model parameters.
- Relative risk estimates accurately reflect uncertainties in the predicted radon exposure covariate.
- 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.
- SEER registry provides accurate information on the leukemia incidence in all counties.
- Ecologic findings should be limited to hypothesis generating and not to determine risk.

Discussion

- The Iowa study offers several advantages over previous ecologic radon studies.
- Bayesian approach allows for prediction of county-average radon levels while accounting for spatial dependencies, systematic differences between homes, and detector measurement error.
- Joint distribution of all model parameters.
- Relative risk estimates accurately reflect uncertainties in the predicted radon exposure covariate.
- 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.
- SEER registry provides accurate information on the leukemia incidence in all counties.
- Ecologic findings should be limited to hypothesis generating and not to determine risk.

Discussion

- The Iowa study offers several advantages over previous ecologic radon studies.
- Bayesian approach allows for prediction of county-average radon levels while accounting for spatial dependencies, systematic differences between homes, and detector measurement error.
- Joint distribution of all model parameters.
- Relative risk estimates accurately reflect uncertainties in the predicted radon exposure covariate.
- 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.
- SEER registry provides accurate information on the leukemia incidence in all counties.
- Ecologic findings should be limited to hypothesis generating and not to determine risk.

Discussion

- The Iowa study offers several advantages over previous ecologic radon studies.
- Bayesian approach allows for prediction of county-average radon levels while accounting for spatial dependencies, systematic differences between homes, and detector measurement error.
- Joint distribution of all model parameters.
- Relative risk estimates accurately reflect uncertainties in the predicted radon exposure covariate.
- 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.
- SEER registry provides accurate information on the leukemia incidence in all counties.
- Ecologic findings should be limited to hypothesis generating and not to determine risk.

Discussion

- The Iowa study offers several advantages over previous ecologic radon studies.
- Bayesian approach allows for prediction of county-average radon levels while accounting for spatial dependencies, systematic differences between homes, and detector measurement error.
- Joint distribution of all model parameters.
- Relative risk estimates accurately reflect uncertainties in the predicted radon exposure covariate.
- 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.
- SEER registry provides accurate information on the leukemia incidence in all counties.
- Ecologic findings should be limited to hypothesis generating and not to determine risk.

Discussion

- The Iowa study offers several advantages over previous ecologic radon studies.
- Bayesian approach allows for prediction of county-average radon levels while accounting for spatial dependencies, systematic differences between homes, and detector measurement error.
- Joint distribution of all model parameters.
- Relative risk estimates accurately reflect uncertainties in the predicted radon exposure covariate.
- 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.
- SEER registry provides accurate information on the leukemia incidence in all counties.
- Ecologic findings should be limited to hypothesis generating and not to determine risk.