Estimating lung cancer mortality rates in U.S. counties using Bayesian hierarchical Poisson regression models

Melissa Jay
Advisor: Dr. Jacob Oleson

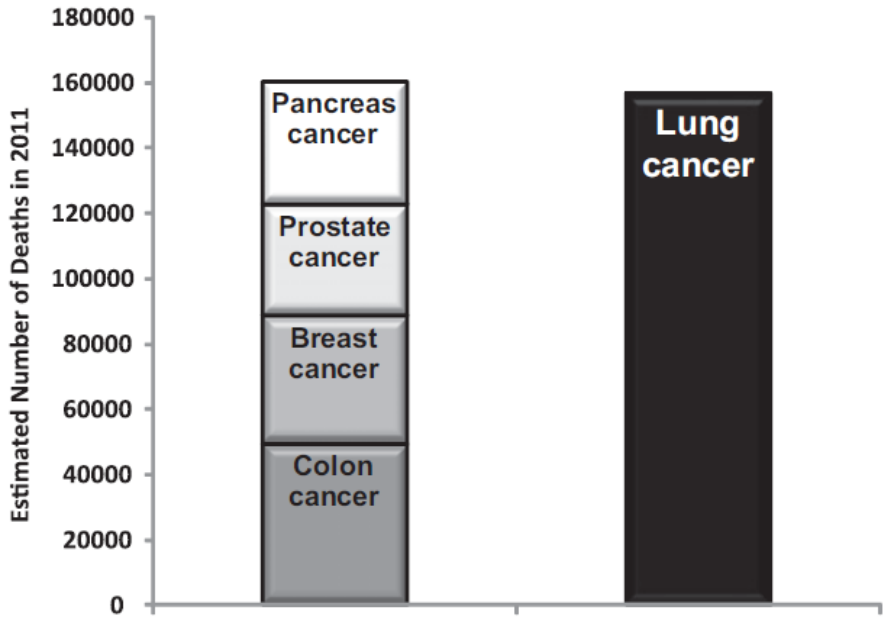
May 3, 2019

Outline

Lung cancer in the United States

- In the U.S., lung cancer is the leading cause of cancer death
- The five year survival rate is just 15.6% based on data from 2007
- Lung cancer incidence and mortality have been decreasing steadily for men but only recently started to decrease for women

- Known risk factors include:
  - Cigarette smoking
  - Secondhand smoke
  - Radon
  - Air pollution
  - Asbestos
  - Diesel exhaust
  - History of lung disease



Source: Dela Cruz et al. (2011)

O'Connor et al. (2018)

- O'Connor et al. studied the relationship between median household income and cancer mortality rate at the county level, while assessing potential mediators
- Concluded that food insecurity, low-quality care, smoking, and physical inactivity had the largest mediating effects

- Methods:
  - Outcome variable was overall cancer mortality rate estimate for 2014
  - Used series of linear regressions
  - Did not account for spatial correlation
  - Took an ad hoc approach to adjust for correlation in model error terms

Mokdad et al. (2017)

- Created annual small area mortality rate estimates for 29 cancers in the United States for the years 1980-2014
- Used spatio-temporal Bayesian hierarchical models to obtain estimates
- Included seven covariates related to race/ ethnicity, median household income, high school graduation rate, and population density at the county level
- Made these data publicly available on the Institute for Health Metrics and Evaluation (IHME) website

Project goals

1. Determine which county-level covariates explain lung cancer mortality rates
2. Estimate lung cancer mortality rates for counties in the U.S.
3. Identify geographic clusters with particularly high lung cancer mortality rates

Data collection

- Collected county-level data from IHME, EPA, CDC, SEER, USDA, and other agencies on all counties for the years 2005-2014
- Lung cancer mortality rates used were age-adjusted rate estimates from Mokdad et al. (2017)
- Variables included in final model:
  - **Behavioral:** prevalence of daily smokers, high alcohol consumption, any alcohol consumption, and obesity
  - **Environmental:** average daily fine particulate matter (PM2.5), radon zone, and diesel emissions
  - **Socioeconomic:** percent unemployed
  - **Geographic:** recoded rural-urban continuum code (RUCC)
  - **Demographic:** proportion male
- Final data set includes 3,108 U.S. counties

Model set-up and notation

$$
\begin{aligned}
Y_i &\sim \text{Poisson}(\theta_i) \\
\log(\theta_i) &= \log(n_i) + \boldsymbol{x_i^{*\prime}}\boldsymbol{\beta} + \gamma_i + \epsilon_i \\
R_i &= \left(\frac{\theta_i}{n_i}\right) * 100000
\end{aligned}
$$

- Y_i = age-adjusted lung cancer death count for county i
- n_i = population of county i
- $\boldsymbol{x_i^{*\prime}}$ = vector of centered and scaled covariates for county i
- R_i = lung cancer mortality rate for county i
- γ_i = spatial random effect for county i
- ϵ_i = overdispersion term for county i

Model set-up and notation (continued)

Priors:

- $\beta_j \sim \text{Normal}(0, 1000)$ for $j = 1 \ldots, J$
- $\epsilon_i \sim \text{Normal}(0, \sigma_Y^2)$
- $\boldsymbol{\gamma} \sim \text{MVN}\left(\mathbf{0}, \sigma_\gamma^2(\boldsymbol{I} - \boldsymbol{W})^{-1}\right)$
- $\sigma_Y^2 \sim \text{Inverse Gamma}(0.01, 0.01)$
- $\sigma_\gamma^2 \sim \text{Inverse Gamma}(0.01, 0.01)$

Implementation:

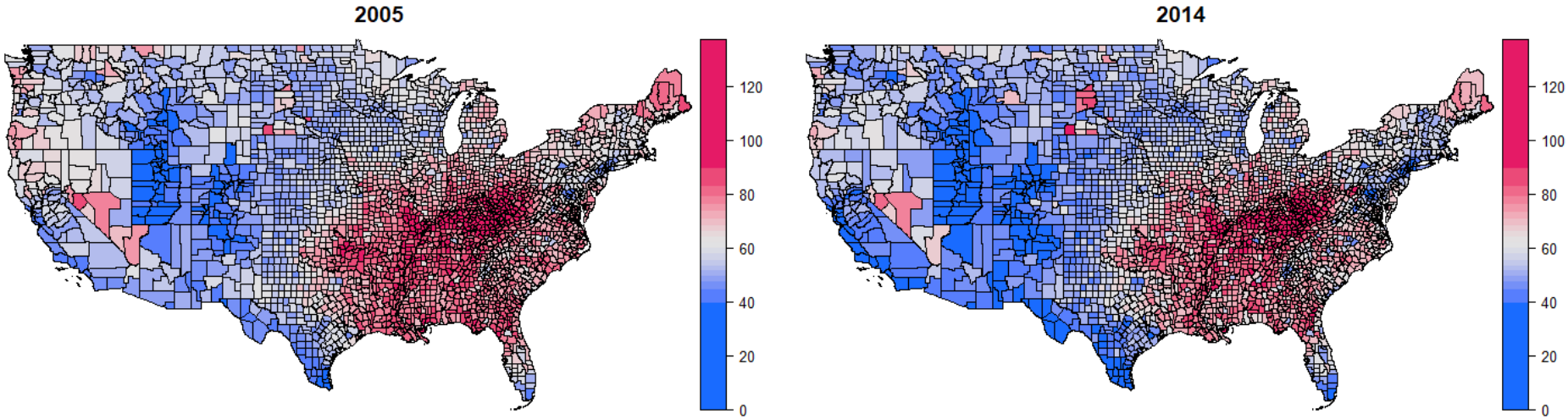
- Used OpenBUGS to implement and run all models
- Ran models for three individual years: 2005, 2010, and 2014

Covariate effects - 2014

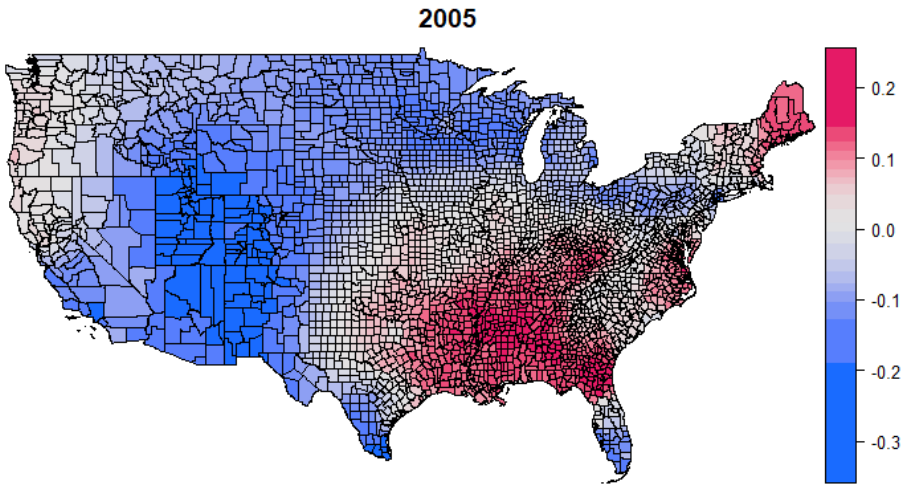
Multiplicative effects of a one standard deviation increase in covariate value on lung cancer mortality rates for counties with the same population size and same spatial random effect

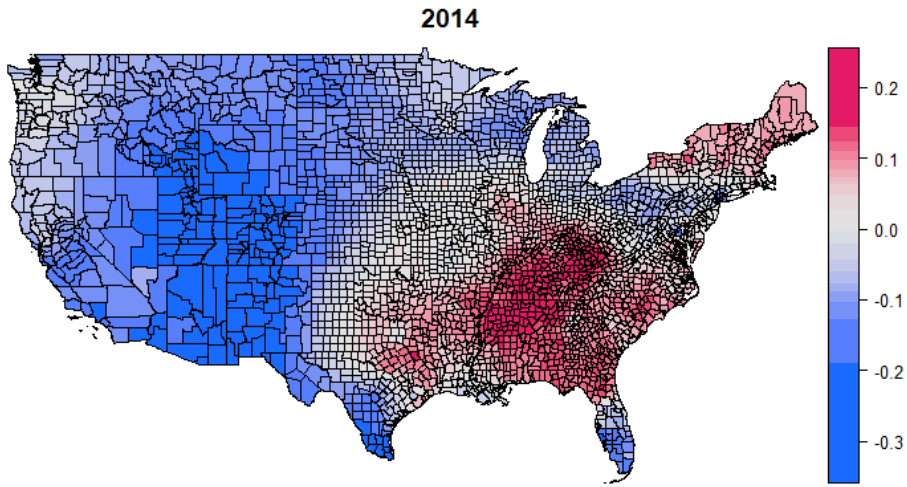
Variable	**Estimate**	**Credible interval**
Daily smokers (%)	1.192	(1.177, 1.207)
Unemployed (%)	1.035	(1.023, 1.048)
Suburban (vs. metropolitan)	0.971	(0.953, 0.990)
Any alcohol consumption (%)	1.026	(1.009, 1.045)
Rural (vs. metropolitan)	0.978	(0.951, 1.004)
Obesity (%)	1.017	(1.004, 1.030)
PM2.5	1.003	(0.987, 1.022)
Diesel emissions	1.003	(0.997, 1.009)
Radon zone	1.002	(0.992, 1.013)
Male (%)	1.002	(0.992, 1.013)
Heavy alcohol consumption	1.001	(0.989, 1.013)

Estimated lung cancer mortality rates



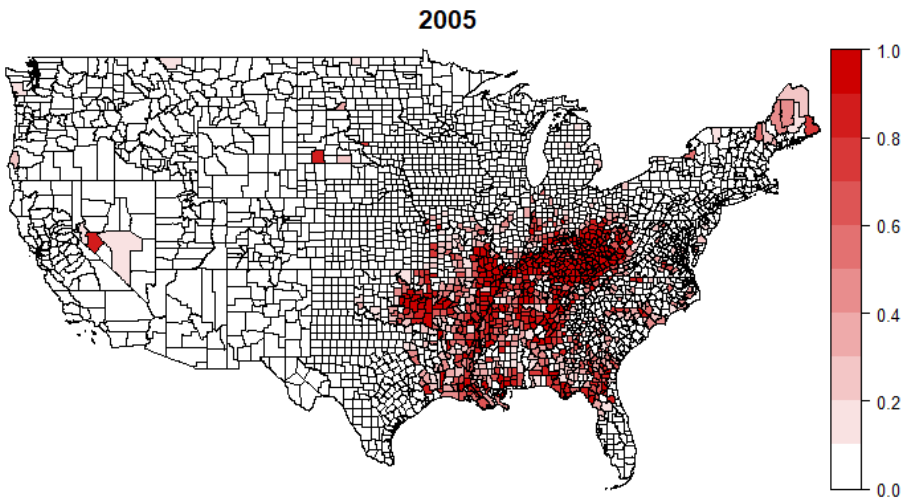
Unexplained spatial variation in mortality rates

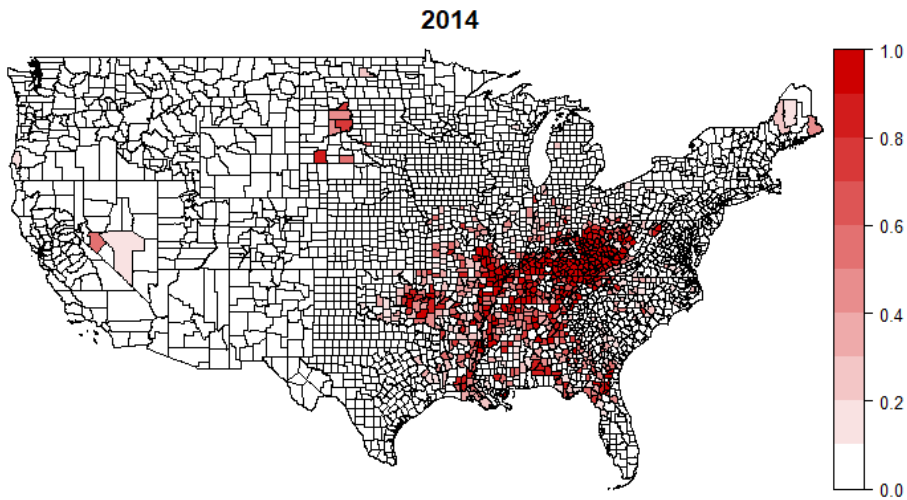




Clustered lung cancer mortality

Probability that estimated lung cancer mortality rate exceeds 80 per 100,000





Conclusions and limitations

- Smoking, unemployment, rurality, alcohol consumption, and obesity partially explained lung cancer mortality rates in the 2005, 2010, and 2014 models
- We identified geographic clusters with particularly high lung cancer mortality rates that could benefit from public health interventions
    - These clusters include counties in KY, WV, TN, MO, LA, and other parts of the southeastern United States
- There still exists spatial variation that is not explained by model covariates
- Limitations:
    - Models do not fully adjust for confounding between radon and smoking
    - Our outcome variable could be subject to measurement error since it is an estimate rather than a raw count
    - Marginal models were run for each of the years of interest, which means temporal correlation was not captured

Future work

- Run model with autoregressive time component on Argon
- Develop composite measure of socioeconomic and other variables to alleviate confounding between radon levels and smoking prevalence
- Incorporate healthcare access variables into models
- Environmental Health Sciences Research Center pilot grant
  - Create cancer risk estimates with an emphasis on environmental covariates and covariates relevant to rural counties

Acknowledgements

- Dr. Jacob Oleson
- Dr. Mary Charlton
- The department of biostatistics

Thank you for listening! Questions?

References

1. Dela Cruz CS, Tanoue LT, Matthay RA. Lung cancer: epidemiology, etiology, and prevention. *Clin Chest Med.* 2011;32(4):605-644. doi:10.1016/j.ccm.2011.09.001.
2. Field RW, Steck DJ, Smith BJ, et al. Residential radon gas exposure and lung cancer: The Iowa radon lung cancer study. *Am J Epidemiol.* 2000;151(11):1091-1102. doi:10.1016/j.ccm.2011.09.001. 10.1093/oxfordjournals.aje.a010153.
3. Mokdad AH, Dwyer-Lindgren L, Fitzmaurice C, et al. Trends and patterns of disparities in cancer mortality among US counties, 1980-2014. *JAMA*. 2017;317(4):388-406. doi:10.1001/jama.2016.20324.
4. O'Conner JF, Sedghi T, Dhopapkar M, Kange MJ, Gross CP. Factors Associated with cancer disparities among low-, medium-, and high-income US counties. *JAMA Netw Open.* 2018;1(6):e183146. doi:10.1001/jamanetworkopen.2018.3146.
5. Puskin JS. Smoking as a confounder in ecologic correlations of cancer mortality rates with average county radon levels. *Health Phys.* 2003;84(4):526-532.

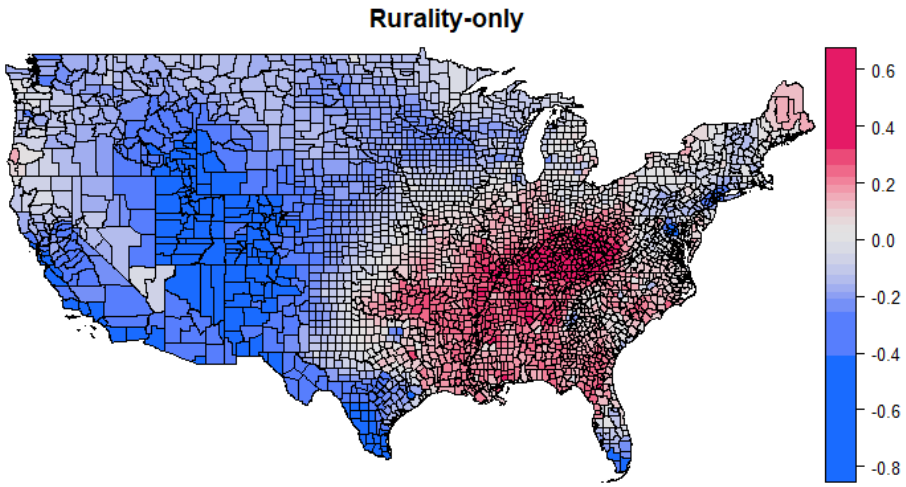
Counties with the highest lung cancer mortality rate estimates - 2014

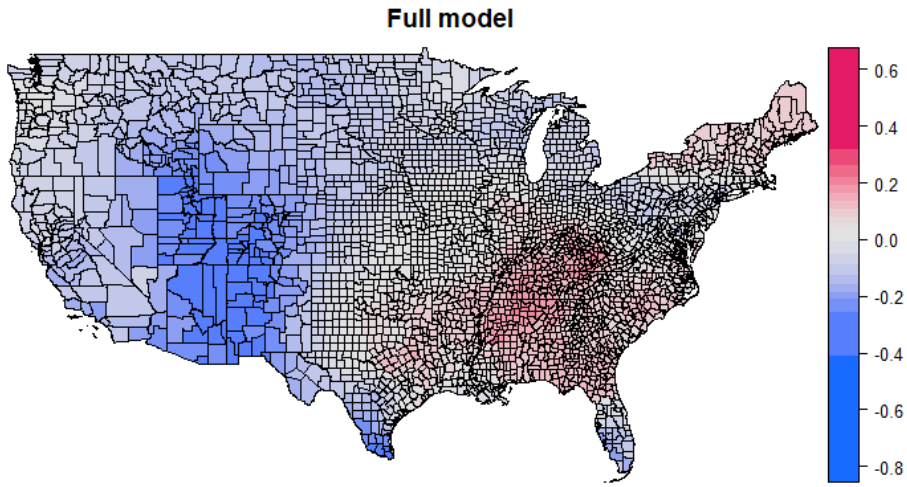
State	County	Estimated mortality rate
KY	Clay	137.29
KY	Leslie	129.46
KY	Knox	126.33
KY	McCreary	125.42
KY	Lee	124.16
KY	Perry	123.65
KY	Jackson	121.14
KY	Breathitt	121.11
KY	Estill	119.44
KY	Elliott	116.26

Counties with the lowest lung cancer mortality rate estimates - 2014

State	County	Estimated mortality rate
UT	Utah	22.54
UT	Wasatch	24.39
UT	Summit	25.38
UT	Davis	25.39
UT	Cache	26.01
NM	Los Alamos	26.23
UT	Morgan	26.60
TX	Kenedy	27.94
CO	Eagle	28.77
UT	Salt Lake	29.24

Rurality-only model vs. full model





- Full model explains more of the spatial variation than the rurality-only model
- Lower DIC for full model compared to rurality-only model

Confounding between smoking and radon

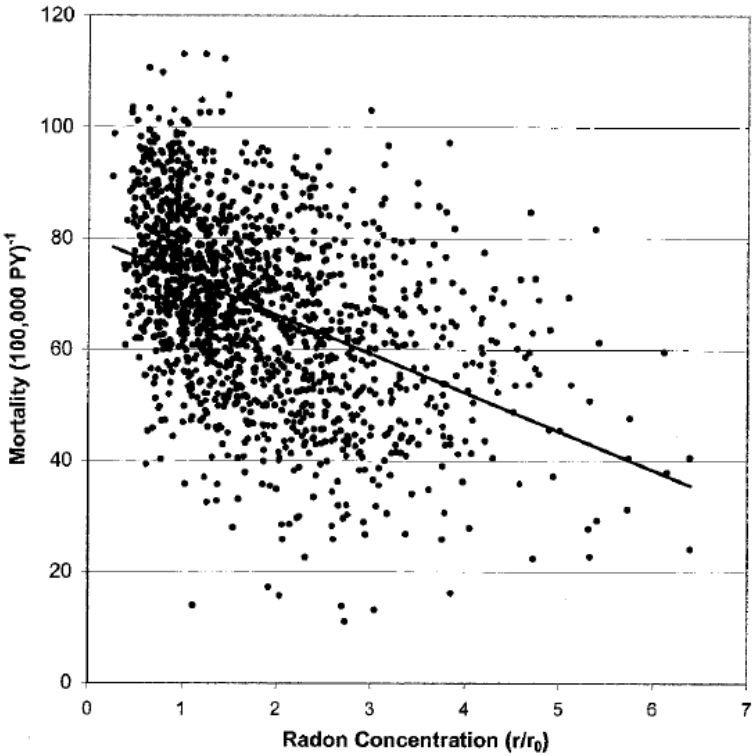


Fig. 1. Lung cancer mortality (1970–1994) for white males vs. measured average radon concentration. Each plotted symbol represents data on one county. The line represents the result of an unweighted linear regression through the points.

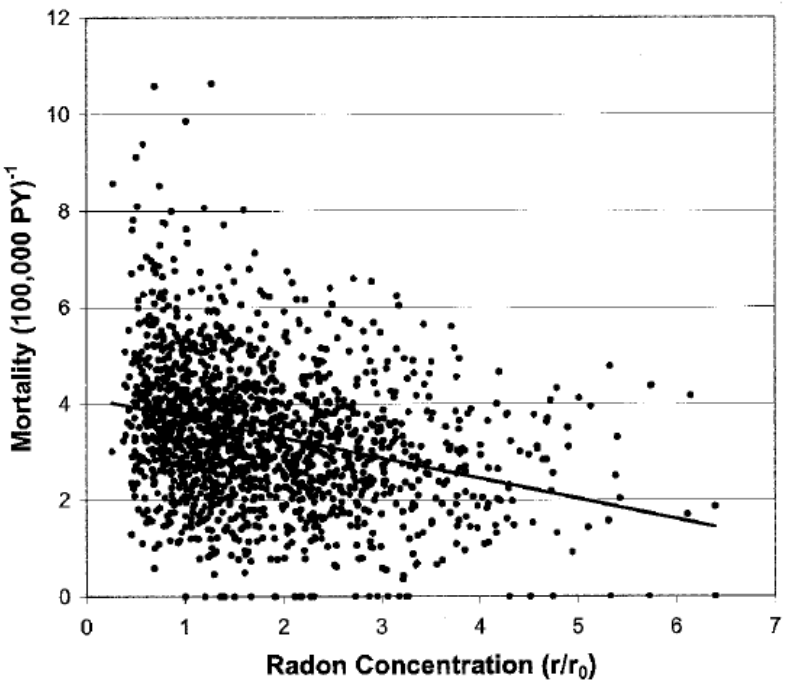


Fig. 2. Oral cancer mortality (1970–1994) for white males vs. measured average radon concentration. Each plotted symbol represents data on one county. The line represents the result of an unweighted linear regression through the points.

Source: Puskin (2003)

Challenges with multi-year spatial data

- Changing county borders and/or county names
  - **Solution:**
    - Use 2010 county definitions and corresponding map for all years
    - Conduct analysis on contiguous U.S. only
    - Use average of neighboring county covariate values when value isn't available
- Annual lung cancer mortality data from SEER suppresses county mortality counts under ten
  - **Solution:**
    - Back transform age-adjusted rate estimates from the IHME to obtain age-adjusted counts
    - Do not make inference on variables used to initially create those estimates