National Cancer Institute

Model-based State Level Estimates for Cancer Related Knowledge Variables using HINTS Data

Benmei Liu, Ph.D.

Division of Cancer Control and Population Science

January 9, 2014

Rockville, MD

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

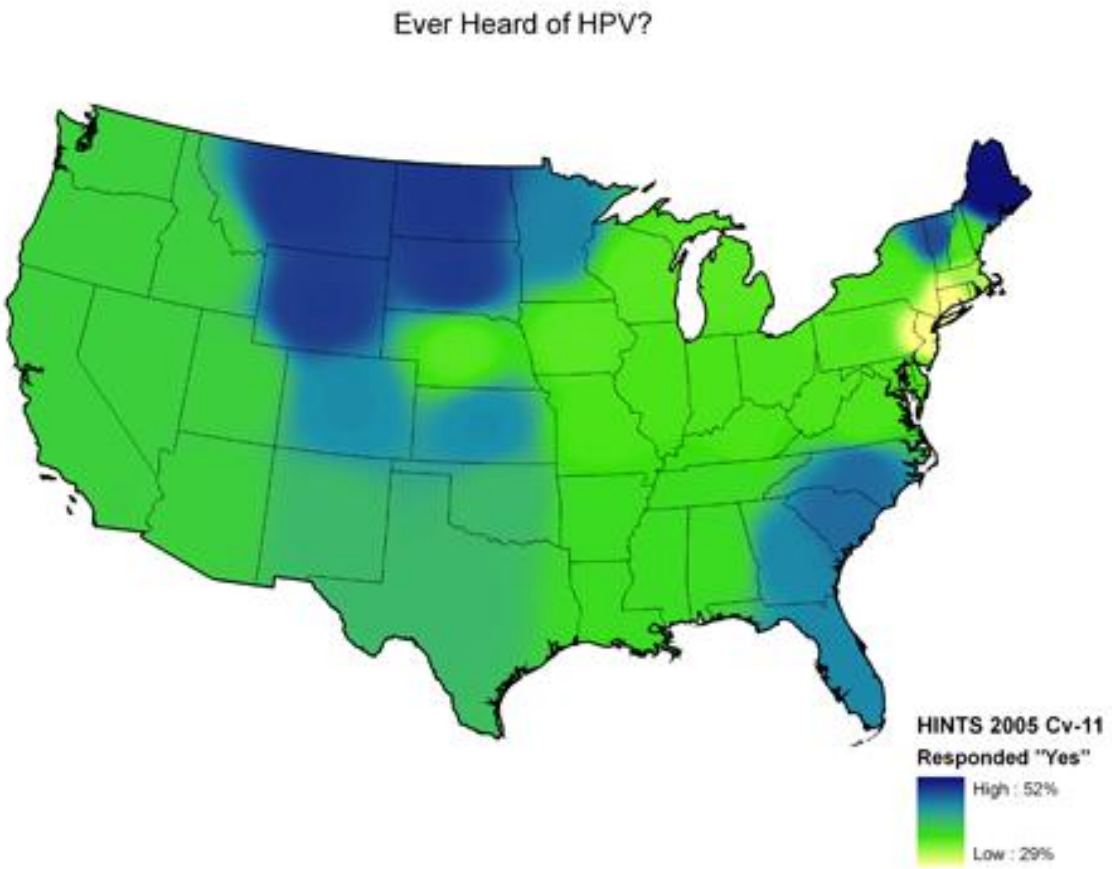
National Institutes of Health

Outline

- Motivation of the project
- Research goal
- Proposed SAE model and implementation
- Evaluation of the model-based estimates
- Model-based estimates on US maps
- Discussions

Motivation of the Project

GIS maps using HINTS data could provide a visual representation of possible geographic relationships in HINTS cancer-related knowledge variables (http://hints.cancer.gov/allMaps.aspx)



Motivation of the Project (cont'd)

- Due to instability in some state values from relatively small sample sizes, the GIS maps that have been developed cannot provide specific state-level estimates of HINTS variables.
- People are highly interested in getting estimates for each state for the variables of interest.
- Small area estimation techniques need to be considered

Model-based SAE Techniques

- Borrowing strength from relevant sources (Census/ Administrative information, related surveys)
- Methods of combining Information
  - Choose good small area models
  - Use good statistical methodology

Ref: Rao (2003); Jiang & Lahiri (2006)

Application of SAE Techniques in Estimating Proportions

- Estimate cancer risk factors & screening behaviors for states and counties by combining data from Behavior Risk factor Surveillance System (BRFSS) and National Health Interview Survey (NHIS) (http://sae.cancer.gov/)
- Estimate poverty rates for states, counties, and school districts in the Census Bureau's Small Area Income and Poverty (SAIPE) program (http://www.census.gov/did/www/saipe/)
- Estimate substance rates for states with data from the National Survey on Drug Use and Health (NSDUH) (http://www.samhsa.gov/data/NSDUH/2k11State/NSDUHsae2011/index.aspx)
- Estimate proportions at the lowest level of literacy for states and counties with data from the National Assessment of Adult Literacy (NAAL) (http://nces.ed.gov/naal/estimates/overview.aspx)

Research Goal of the Project

- Estimate state level proportions of people who answered 'YES' to the following cancer-related knowledge questions using HINTS:
  - Does smoking increase your chance of cancer a lot? (CK13 in 2003)
  - Does Lung cancer will cause most deaths? (CK15 in 2003)
  - Have you ever heard of a sigmoidoscopy or colonoscopy? (CC15 in 2003)
  - Have you ever heard of a stool blood test? (CC4 in 2003)
  - At what age people supposed to start having sigmoidoscopy or colonoscopy? Proportion of people whose answer is 50. (CC24 in 2003)
  - Have you ever heard about HPV? (CV11 in 2005)
  - Have you ever looked for cancer information from any sources? (HC09 in 2003, CA08 in 2005 and HC08 in 2008)
  - Have you ever looked for information about health or medical topics from any source? (HC01 in 2008)

Notations

y_{ik} : a binary response for unit *k* in state i;

s_i : the set of sampled units in state i;

n_i : the sample size in state i;

w_{ik} : the sampling weight for unit *k* in state i;

$\mathbf{x}_i$: the vector of auxiliary variables;

$$k = 1, \ldots, N_i; \quad i = 1, \ldots, m$$

- Parameters of interest are the population proportions:

$$P_i = \sum_{k=1}^{N_i} y_{ik} / N_i$$

Direct Estimates of P_i and Associated Variances

- Direct estimates (design-unbiased):

$$p_{iw} = \frac{\sum_{k \in s_i} w_{ik} y_{ik}}{\sum_{k \in s_i} w_{ik}}, i = 1, \ldots, m$$

- Variances of the direct estimates:

- $VAR(p_{iw}) = \frac{P_i(1-P_i)}{n_i} * DEFF_i, i = 1, \ldots, m.$

➢ Problem of p_{iw}: imprecise when sample sizes n_i are small

➢ Solution: Model-based approach

One Commonly Used Area Level Model

- The well known Fay-Herriot model (Fay & Herriot 1979):

  - Sampling model: $p_{iw}|P_i \sim N(P_i,\ d_i)$;
  - Linking model: $P_i = x_i'\beta + v_i$; where $v_i \sim N(0, A)$

→ The sampling variance d_i is assumed known

Proposed Small Area Model

- Fay-Herriot model with arcsin transformation:

Let $z_i = \arcsin\left(\sqrt{p_{iw}}\right)$; (Carter & Rolph, 1974 JASA)

- Sampling model: $z_i|\theta_i \sim N\left(\theta_i, \frac{DEFF_i}{4n_i}\right)$;
- Linking model: $\theta_i = x_i'\beta + v_i$; where $v_i \sim N(0, A)$

→ Goal: To estimate $P_i = sin^2(\theta_i)$.

!! Model was chosen based on an extensive simulation study

Estimate the Design Effects

- The Design effect or $Deff$ is the ratio of the actual variance of a sample to the variance of a simple random sample of the same number of elements;
- $Deff$ can be estimated using the typical survey software such as SAS PROC SURVEYMEANS or SUDAAN;
- Design effects estimated at the state level are not stable due to small sample sizes;
- Design effects estimated at the Census regional level are used instead for smoothing purpose.

Model Assumptions

- To estimate $P_i = sin^2(\theta_i)$, we use hierarchical Bayesian method.

- Prior assumptions:

$$\beta \propto 1;\ A \sim unif(0, 100)$$

- Three chains for each model:
  - 10,000 iterations each after 5,000 burn in; thinning to 5000
  - Gibbs sampling algorithm for drawing samples from the joint posterior distribution

Auxiliary Variables

- The pool of auxiliary variables include:
  - 27 state level demographic & socio-economic variables obtained from Census 2000 and other administrative records
- Classical model selection procedures are applied to reduce the number of auxiliary variables for each outcome

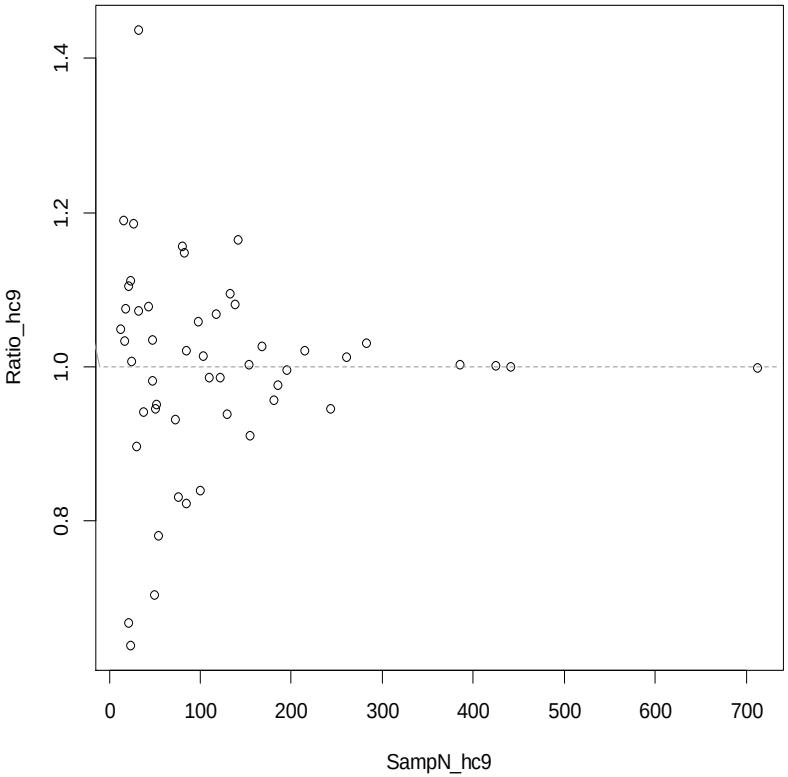
Evaluate the Model-based Estimates

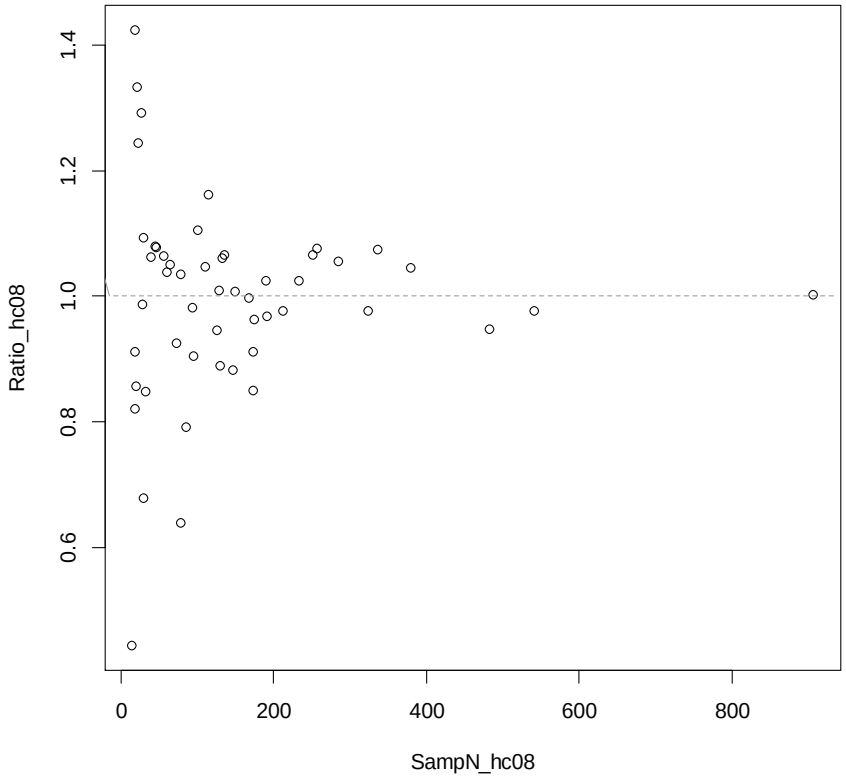
Plot the ratios of the direct estimates over the model-based estimates against the sample size.

- The ratio is expected to converge to 1 as the sample size gets larger

Diagnosis for proportion of people who have ever looked for cancer information from any sources

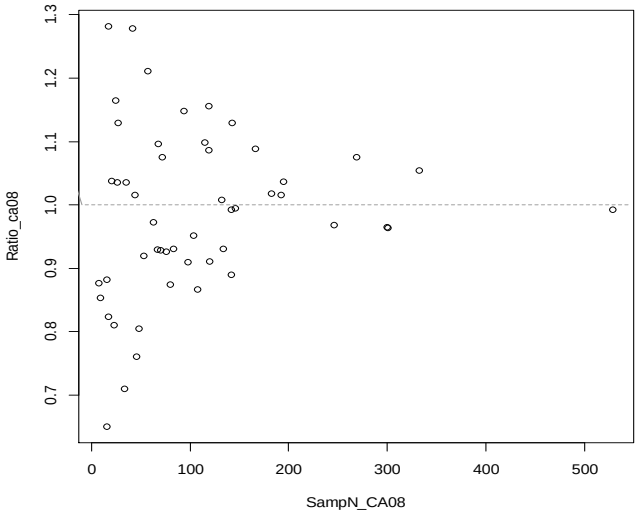
Figure 1: Ratio of the direct estimates over the model-based estimates for HC09 and HC08

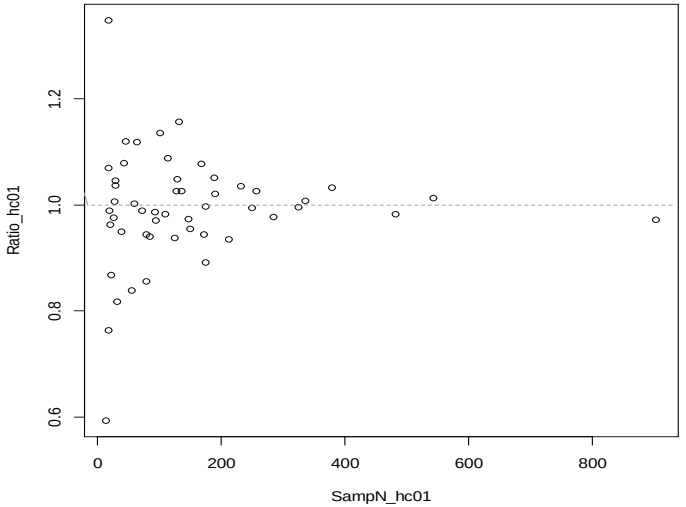


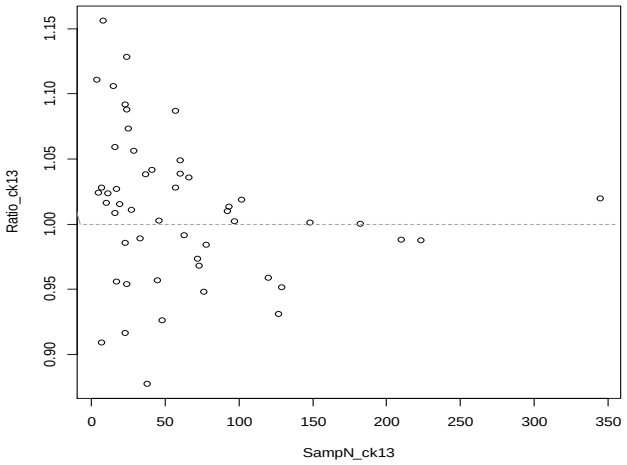


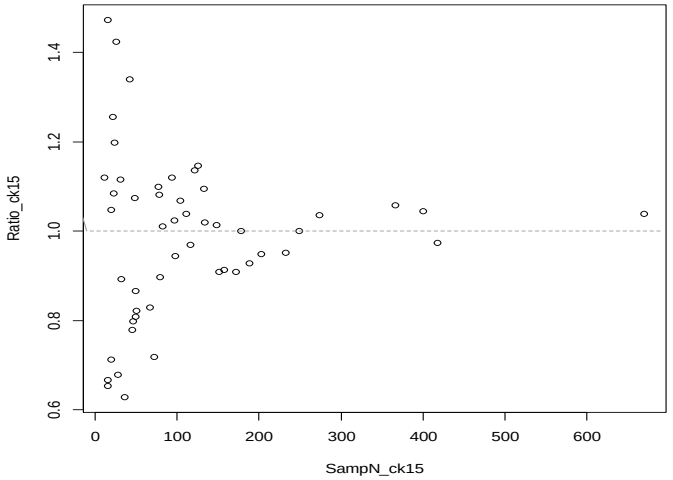
Diagnosis for Other Outcomes

Figure 2: Ratio of the direct estimates over the model-based estimates for CA08, HC01, CK13 and CK15



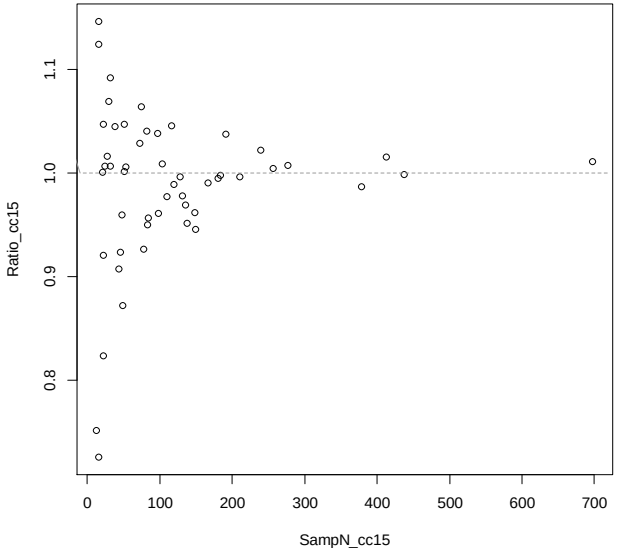


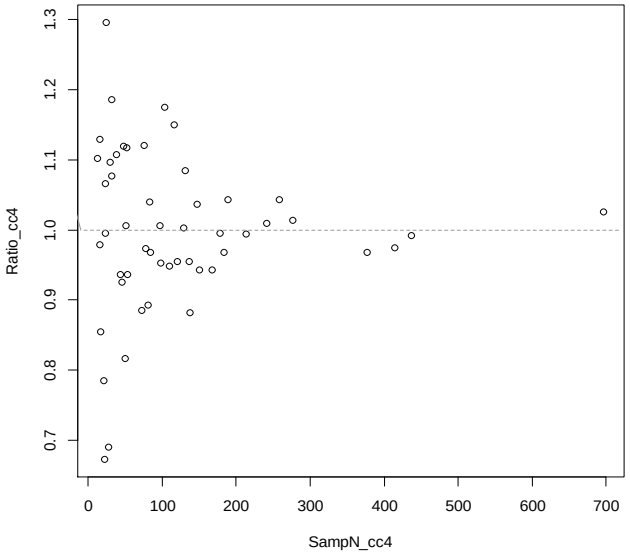


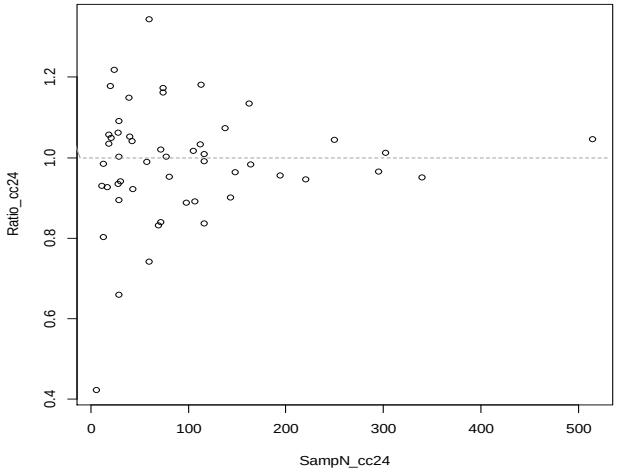


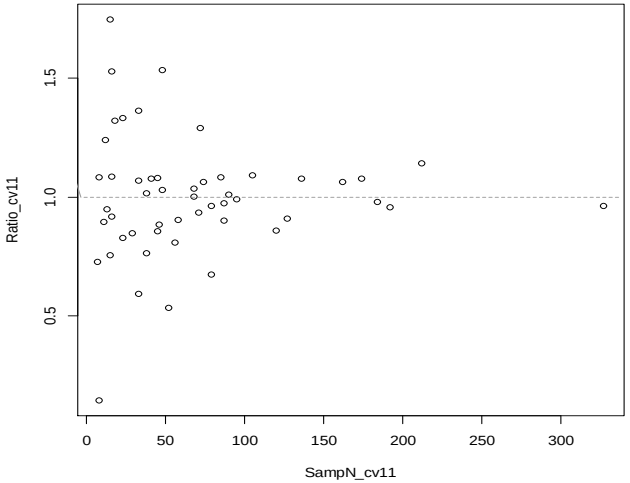
Diagnosis for Other Outcomes (Cont'd)

Figure 3: Ratio of the direct estimates over the model-based estimates for CC15, CC4, CC24 and CV11





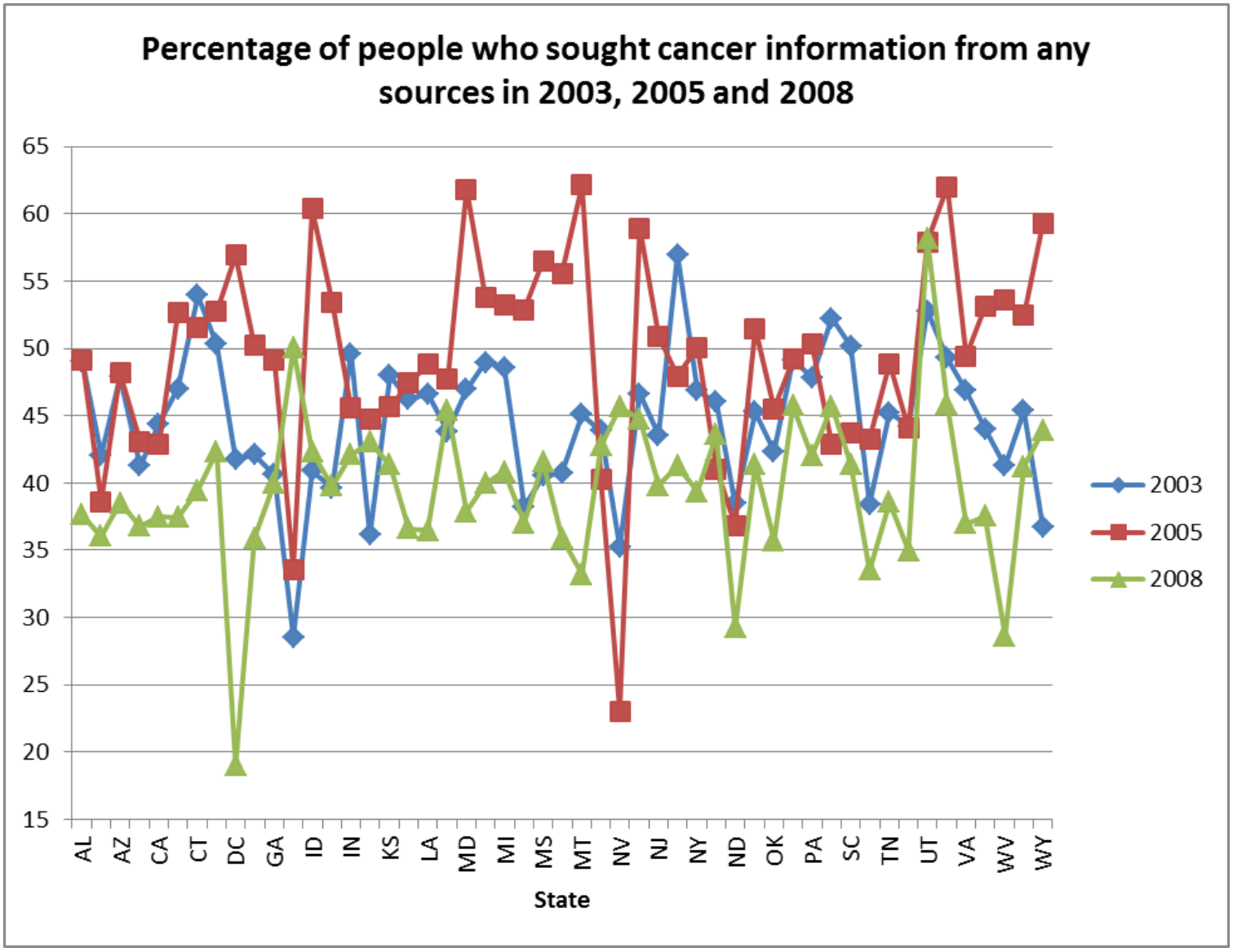




Further Model Diagnosis to Assess the Goodness of Fit of the Proposed Model

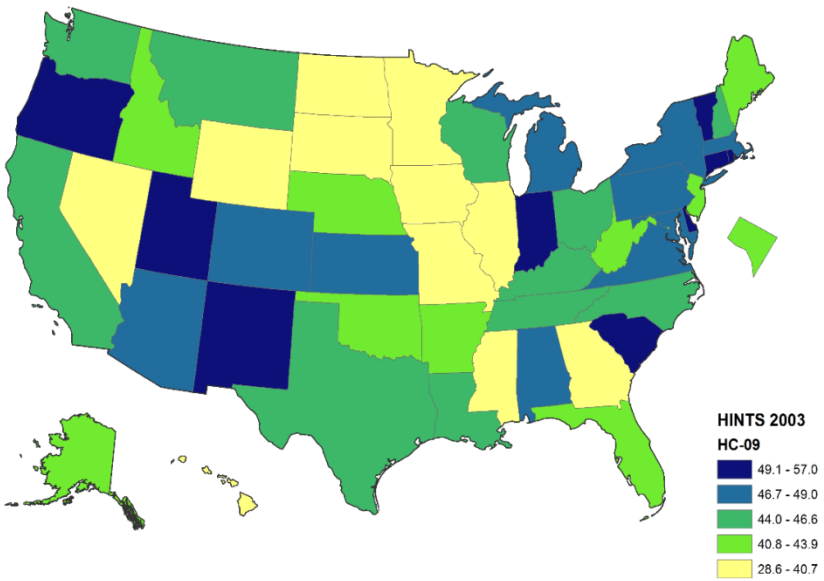
- Check the overall fit of the proposed model using method of posterior predictive *p*-value (Gelman & Meng 1996)
- Assess model fit at individual state level by computing two measures (Rao 2003, Sec 10.2):
  - State level measure providing information on the degree of consistent overestimation or under estimation of the observed value
  - State level measure which is similar to a cross-validation standardized residual but uses the full predictive density

Model-based Estimates for Cancer Information Seeking – Scatter Plots

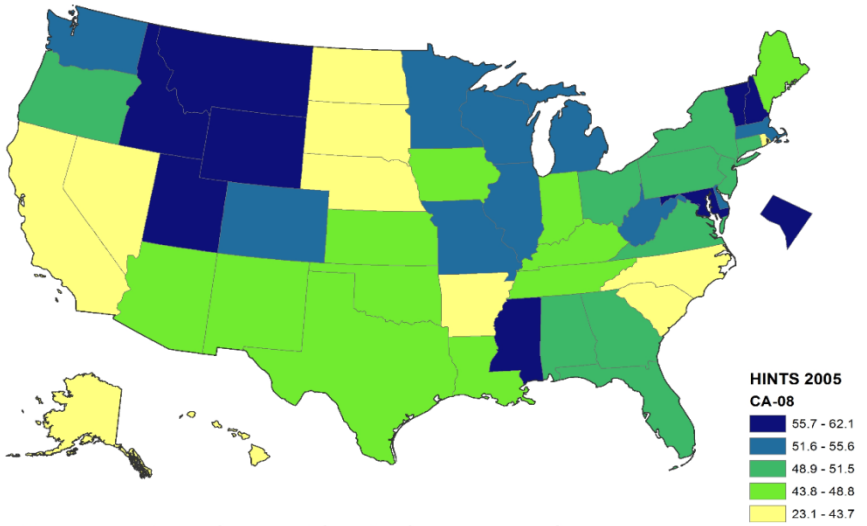


Model-based Estimates for Cancer Information Seeking - Maps

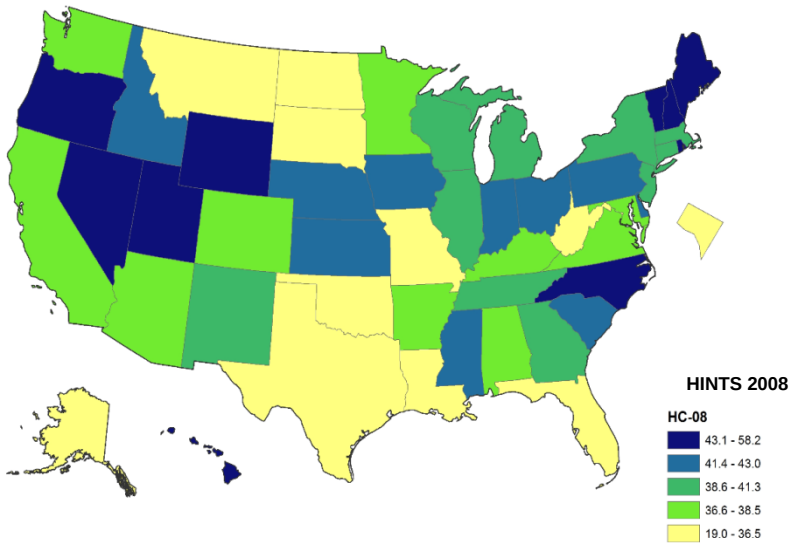
Have you ever looked for cancer information from any sources?



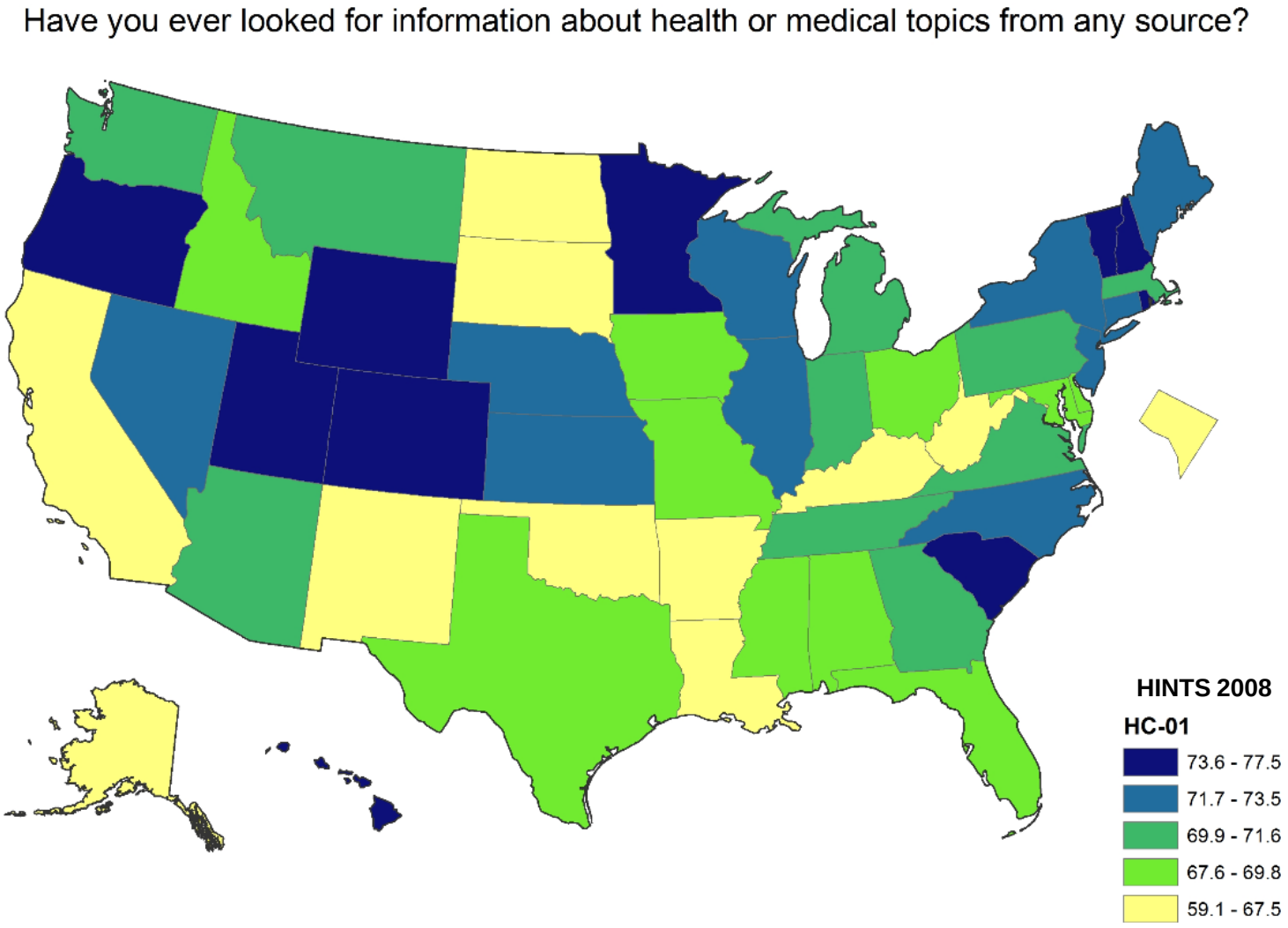
Have you ever looked for cancer information from any sources?



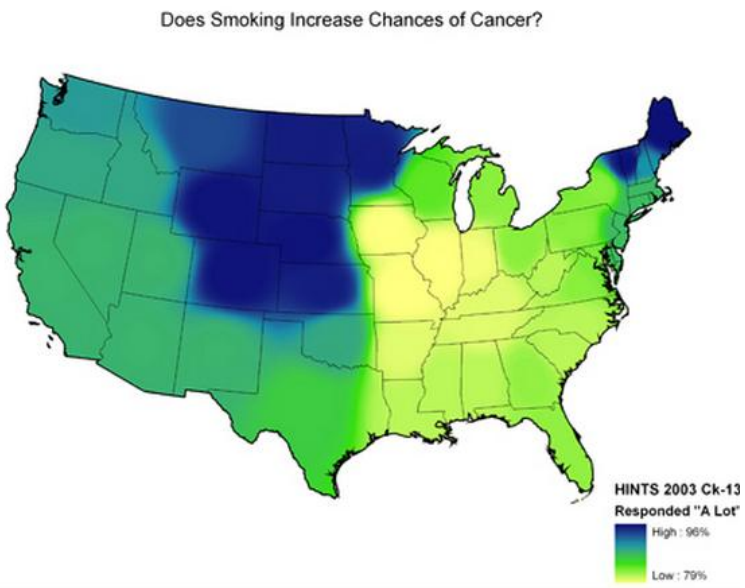
Have you ever looked for cancer information from any sources?

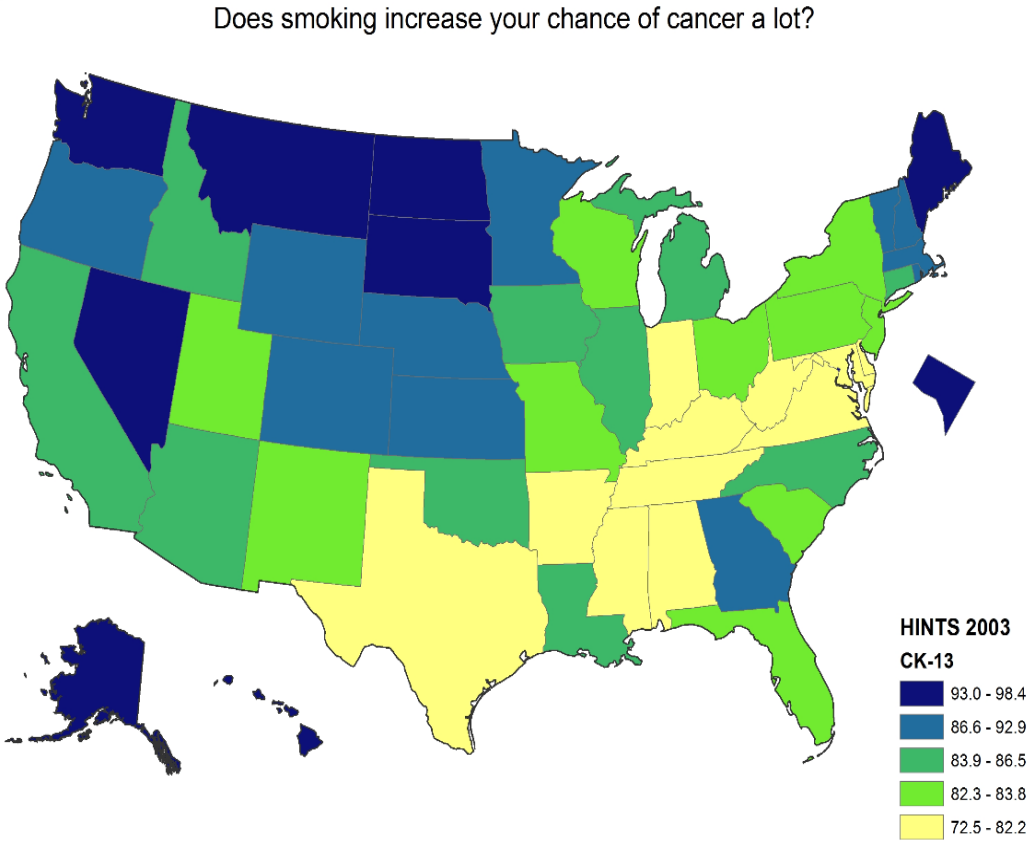


Model-based Estimates for Health Information Seeking - Maps

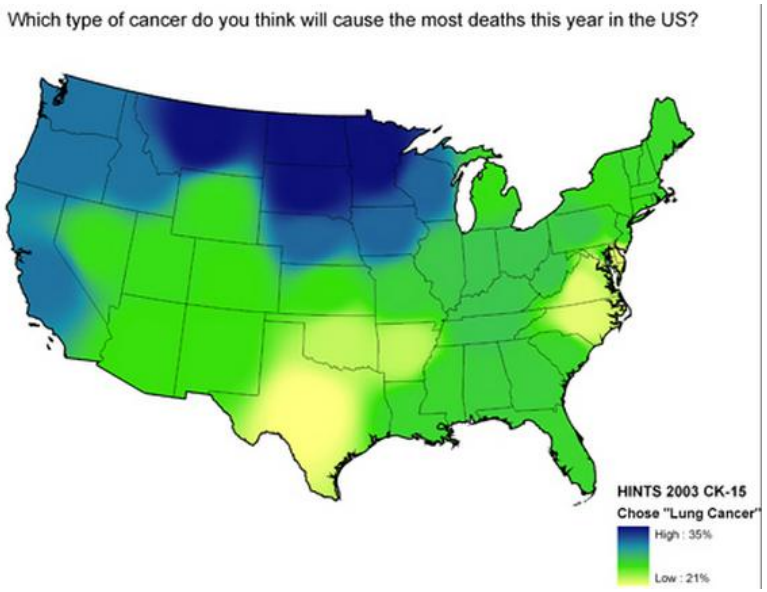


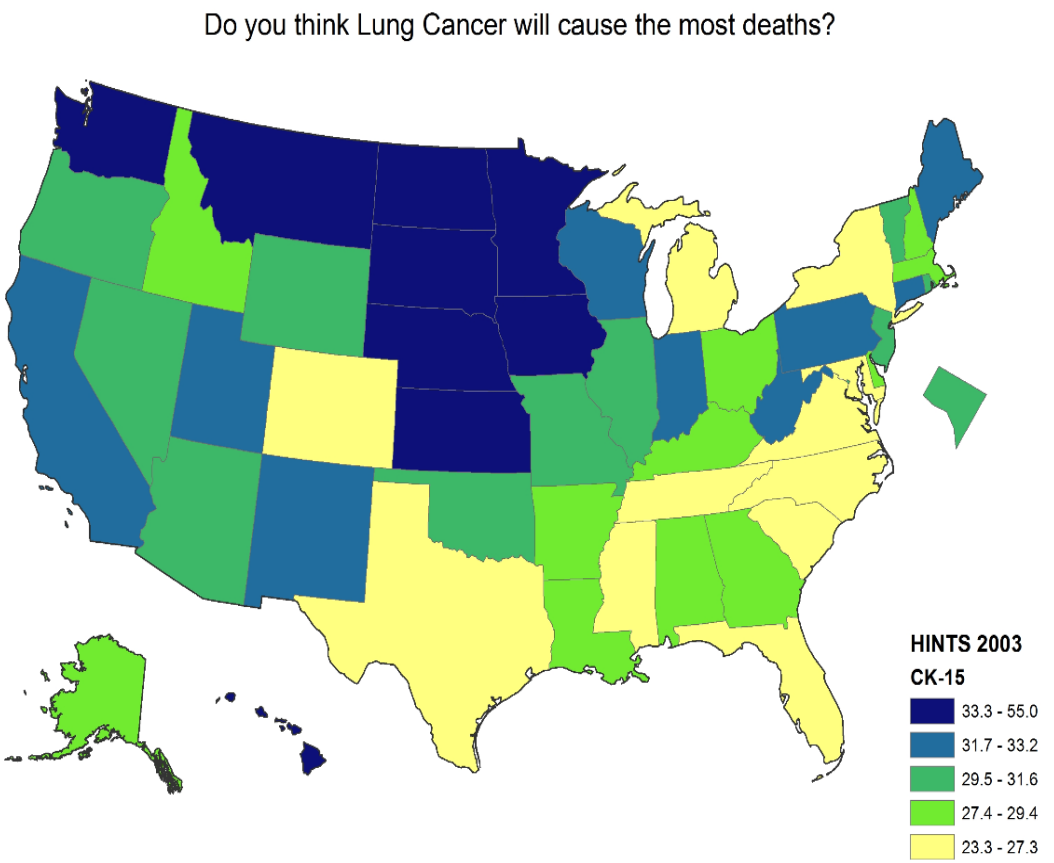
Smoking Increase Chance of Cancer: Compare the Smoothed Direct Estimates With Model-Based Estimates



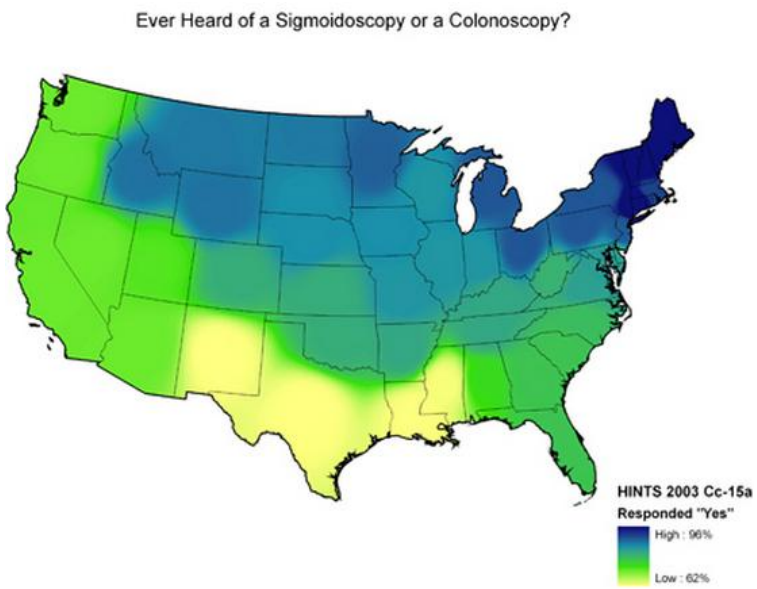


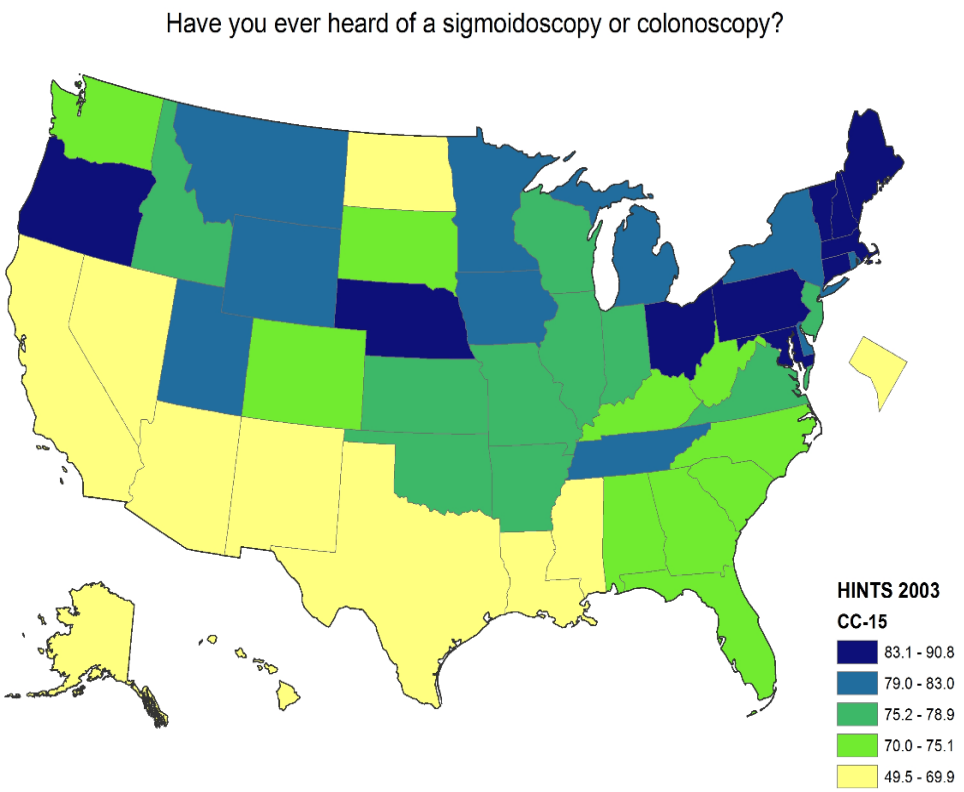
Lung Cancer Cause Most Deaths: Compare the Smoothed Direct Estimates With Model-Based Estimates



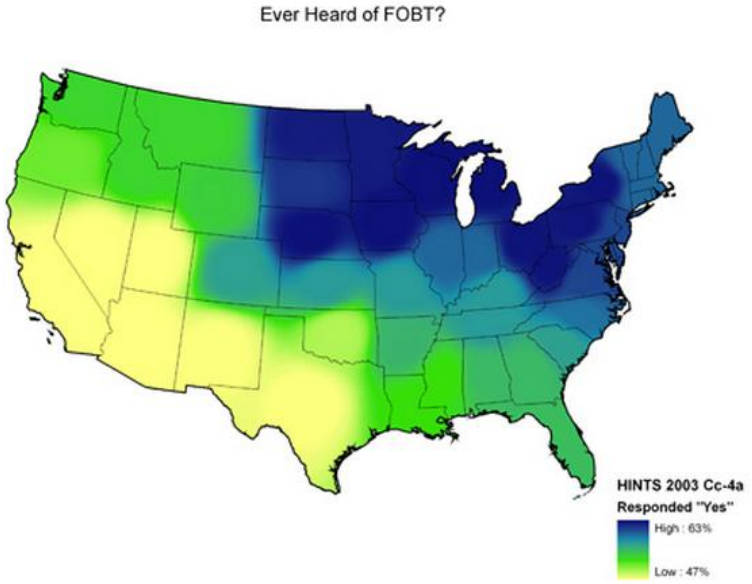


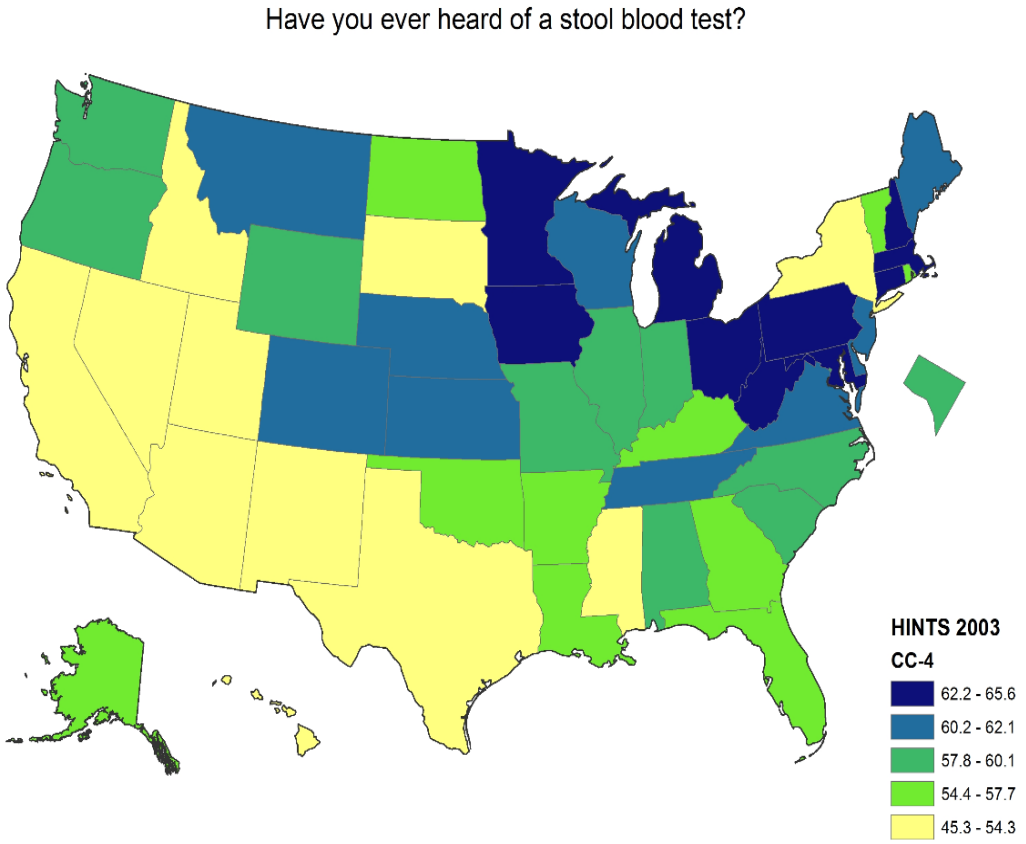
Colon Cancer Screening: Compare the Smoothed Direct Estimates With Model-Based Estimates



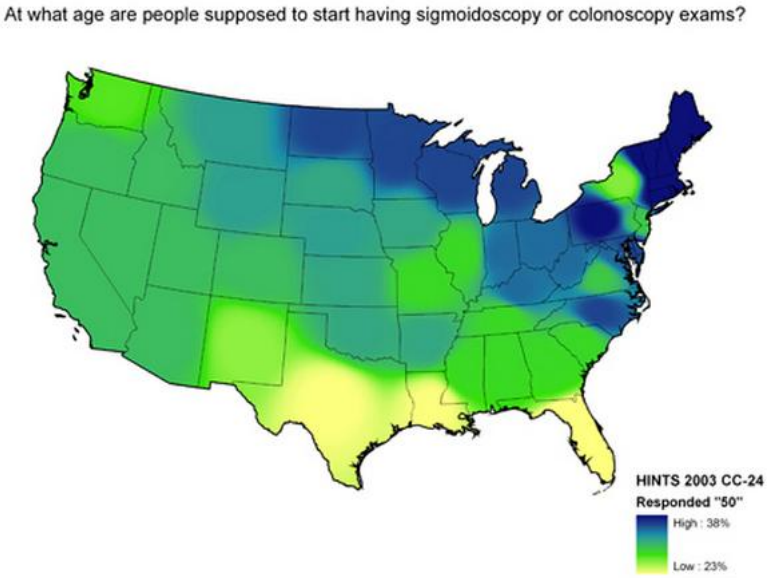


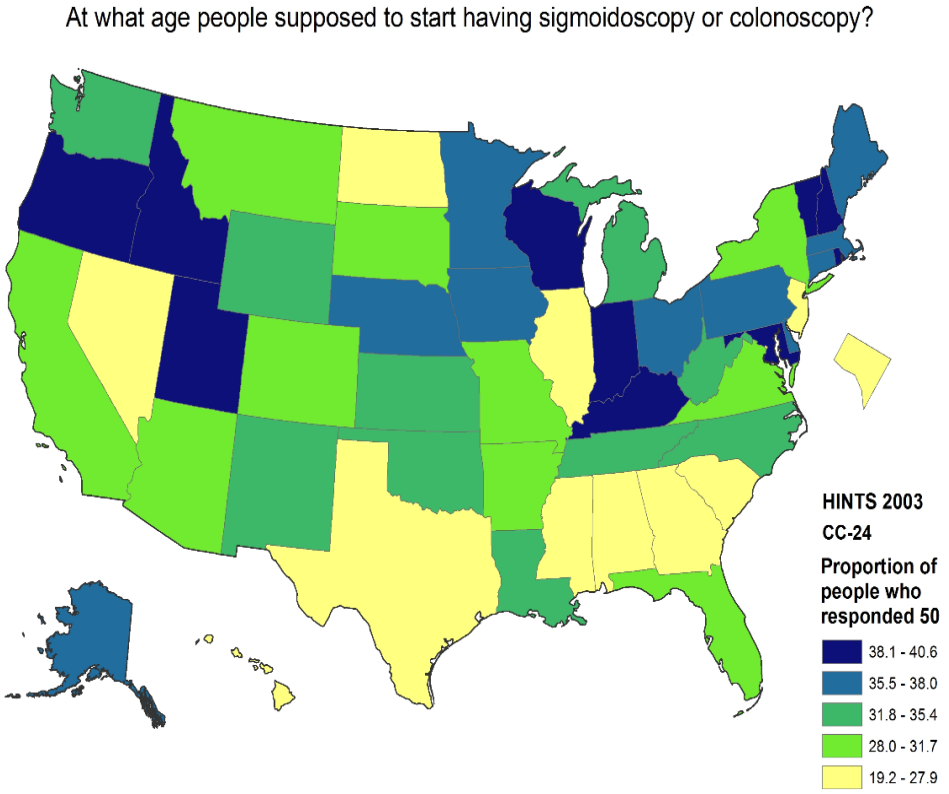
Colon Cancer Screening: Compare the Smoothed Direct Estimates With Model-Based Estimates



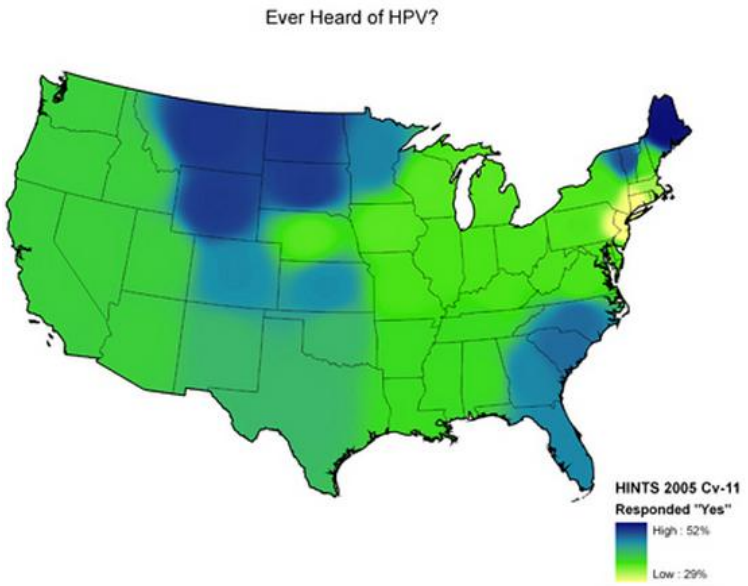


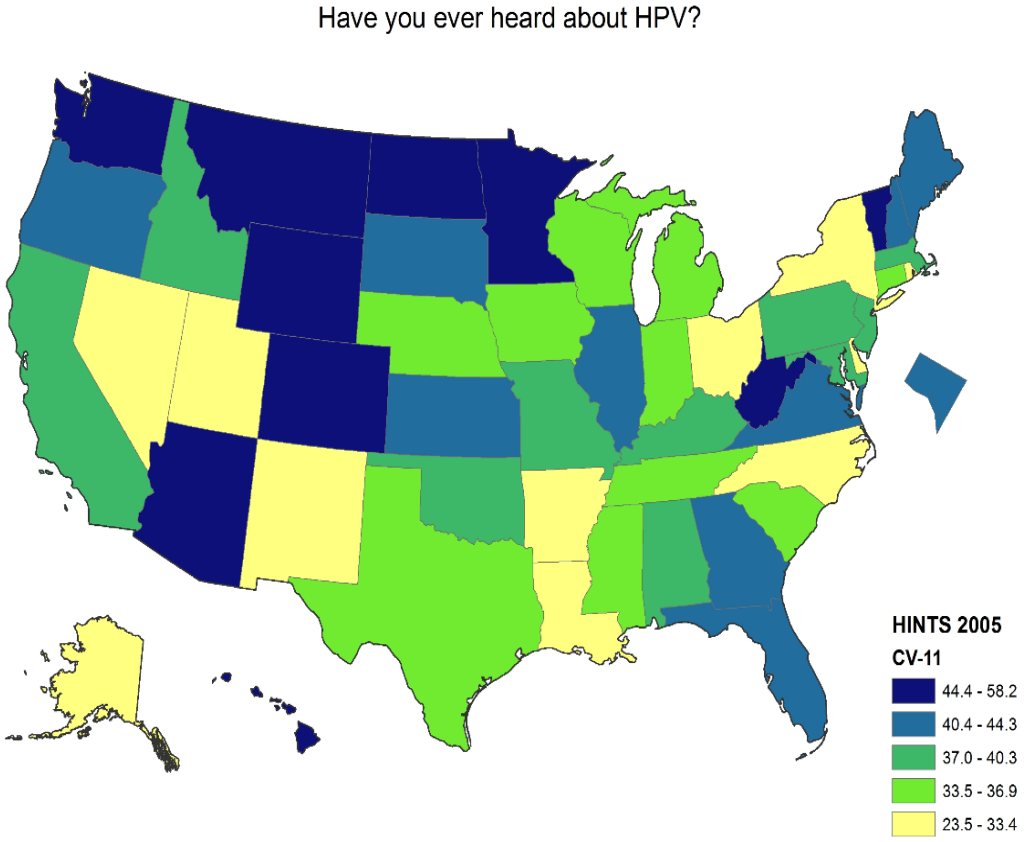
Start Age for Colon Cancer Screening: Compare the Smoothed Direct Estimates With Model-Based Estimates





Ever Heard about HPV: Compare the Smoothed Direct Estimates With Model-Based Estimates





Discussions

- The evaluations confirm that the proposed model works for all the outcomes
- The model-based estimates are expected to be better than the direct estimates on average if the model used are appropriate
- For states with little HINTS sample, the estimates increasingly depend on using the demographic information to produce estimates for states with “similar” profiles from across the state in terms their covariates

References

- Carter, G.M., and Rolph, J.E. (1974), “Empirical Bayes methods applied to estimating fire alarm probabilities,” *Journal of the American Statistical Association*, 69, 880-885.
- Citro, C., and Kalton, G. (Eds.). (2000). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*. Washington, DC: National Academy Press.
- Kish (1965). *Survey Sampling*, New York: John Wiley and Sons.
- Gelman, A., and Meng, S-L. (1996). Model checking and Model-Improvement, in W.R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall, pp. 145-161.
- Maples, J., and Bell, W.R. (2005). Evaluation of school district poverty estimates: Predictive models using IRS income tax data*. Proceedings of the American Statistical Association.*
- Rao, J.N.K. (2003). *Small area estimation*. New York: John Wiley and Sons.
- Robert, C.P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

Any Questions?

Thank you!

Contact info:

Benmei Liu

liub2@mail.nih.gov