Using BRFSS Data to Estimate County–level Colorectal Cancer Screening Prevalence in Missouri

J Du, MA[1,2]; CL Schmaltz, PhD[1,3]; J Jackson-Thompson, MSPH, PhD[1,3,4]

[1] *Missouri Cancer Registry and Research Center (MCR-ARC);* [2] *University of Missouri-Columbia (MU), College of Arts & Sciences, Dept. of Statistics;*
[3] *MU School of Medicine, Dept. of Health Management & Informatics;* [4] *MU Informatics Institute, Columbia, Missouri*

Background

In the US, colorectal cancer (CRC) is the **3rd** most common cancer in both men and women.

- Colorectal cancer screening (CRCS) is recommended for people over 50 years of age.
- Behavioral Risk Factor Surveillance System (BRFSS) data in 2012 show that 66.5% of people in Missouri aged 50 and older have had the screening. However, county–level CRCS prevalence cannot be directly obtained from BRFSS due to small or even zero sample sizes.
- Missouri conducted a County–level Study (CLS) in 2011 aimed for accurate county–level estimates. Questions asked in CLS were similar to those in BRFSS. Since much larger sample sizes were obtained for counties in Missouri, CLS could obtain direct estimates for CRCS; however, CLS is not regularly conducted.
- Cadwell, *et al.* (2010)* used Bayesian methods to estimate diabetes prevalence for all US counties. The methods were used by CDC (`https://www.cdc.gov/diabetes/atlas/obesityrisk/County_Methods.html`).

Objective

Use small area estimation techniques to estimate county–level CRCS prevalence in Missouri for people age 50+ with 2012 BRFSS data and compare with results from 2011 CLS.

Data Overview

- Missouri is comprised of 114 counties and the City of St. Louis.
- In 2012 MO–BRFSS, the sample size from each county was too small to make conclusive estimates by county. Therefore, those areas were clustered into seven BRFSS regions (Figure 1). Numbers in parenthesis are the sample sizes for respondents aged 50 or older which are suitable for our CRCS study after removing respondents with unknown county, unknown response, etc.

Kansas City Metro	(767)
St. Louis Metro	(1030)
Central	(493)
Southwest	(485)
Southeast	(433)
Northwest	(355)
Northeast	(244)

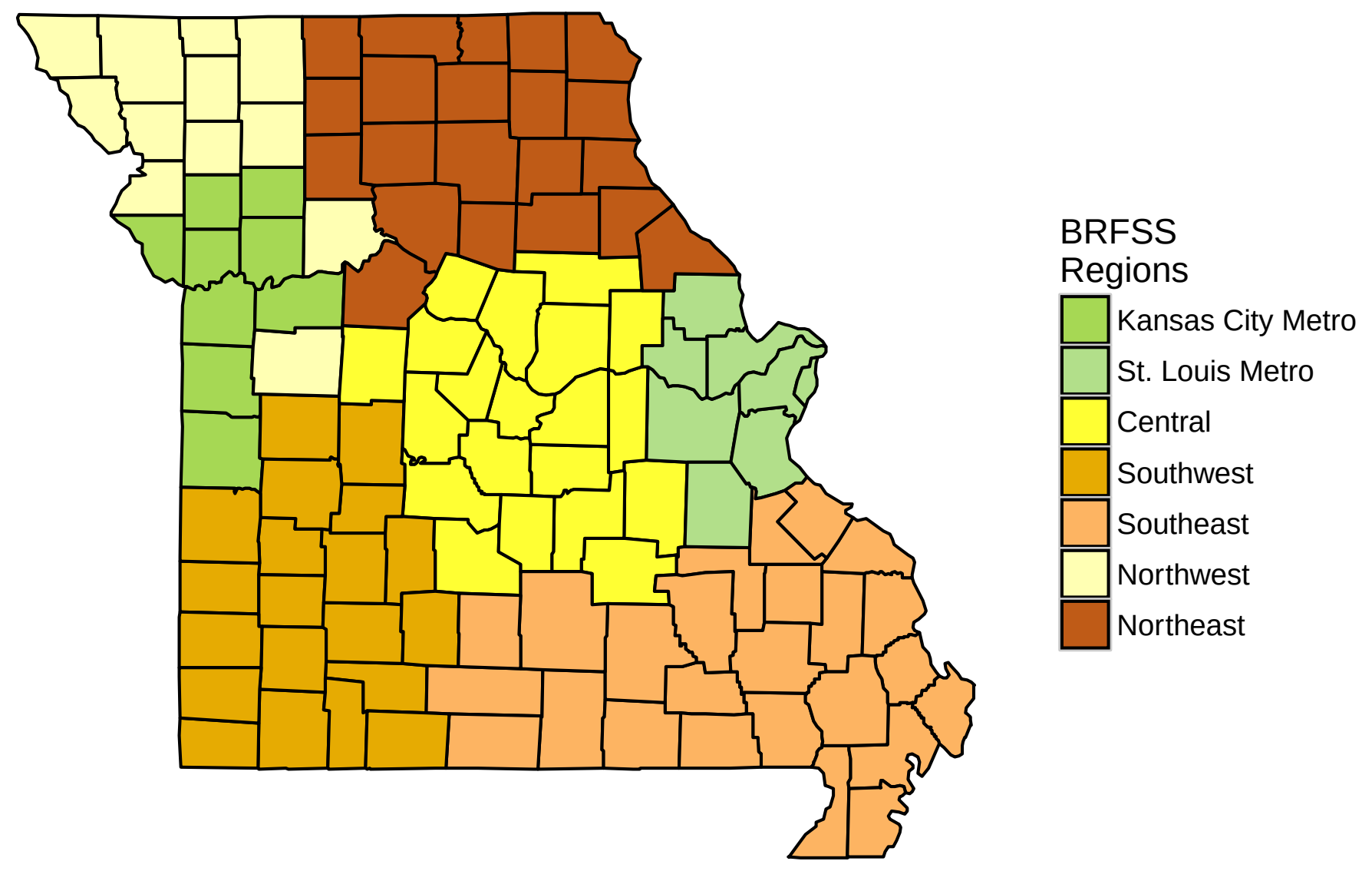

Figure 1: BRFSS regions in MO

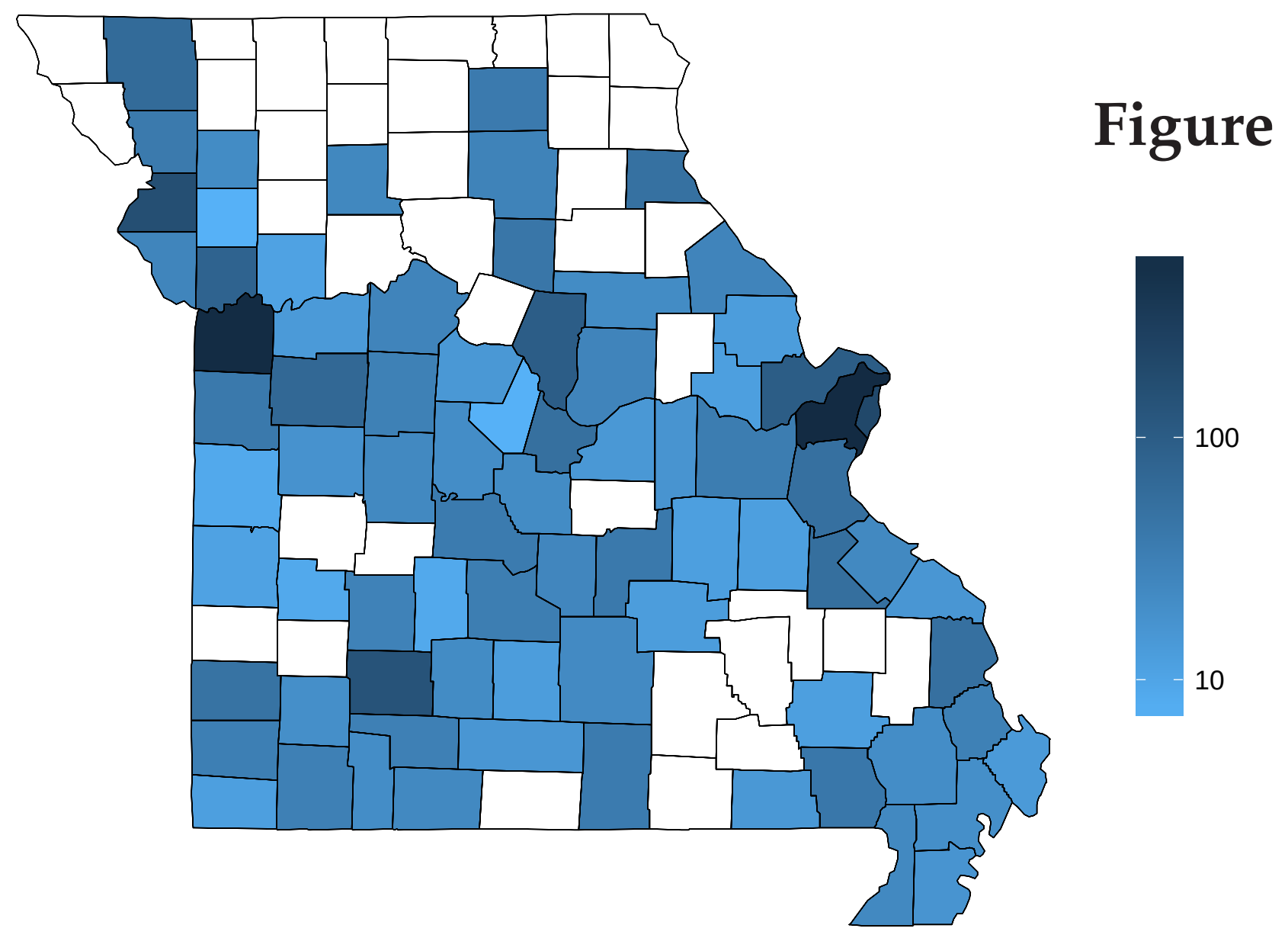

Figure 2 shows the sample size for each county. Thirty-seven counties in MO had zero sample (shown in white); only 15 counties had a sample more than 50.

Figure 2: Sample sizes for counties in MO.

Methods: Notations

Respondents were classified into 12 groups based on age (50–64, 65–74, 75+), gender and race (white, non-white). Let

- n_{ijkl} : sample size (number of respondents) in county $i \in \{1, ..., 115\}$, age group $j \in \{1, 2, 3\}$, gender $k \in \{1, 2\}$ and race $l \in \{1, 2\}$;
- N_{ijkl}: true population size for category $ijkl$, which is obtained from Census data;
- y_{ijkl}: number of respondents who have had CRCS for category $ijkl$;
- Y_{ijkl}: true population total for people who have had CRCS for category $ijkl$;
- p_{ijkl}: proportion of people who have had CRCS for category $ijkl$.

Methods: Models

A Bayesian binomial regression framework to estimate Y_{ijkl}:

$$\begin{aligned} y_{ijkl} &\sim \text{Binomial}(n_{ijkl}, p_{ijkl}) \\ \text{logit}(p_{ijkl}) &= \alpha_{r(i)} + \beta_j + \gamma_k + \theta_l \\ &\quad + \boldsymbol{X}_i \boldsymbol{\psi} + u_i + \epsilon_{ijkl} \end{aligned}$$

where $\epsilon_{ijkl} \sim \text{Normal}(0, \delta_0)$ is the error term and

- $\alpha_{r(i)}$ is the intercept for region $r(i)$ where county i belongs to;
- $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ are the age effects;
- $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$ are the gender effects;
- $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are the race effects;
- $\boldsymbol{\psi}$ contains some county attribute effects like medium income, percentage of people below high school, *etc*; and $\boldsymbol{X}_i$ is i^{th} row of the corresponding design matrix $\boldsymbol{X}$;

The fixed effects above all follow Normal(0,100) priors.
The county effects $\boldsymbol{u} = (u_1, ..., u_{115})$ were modeled with a proper CAR prior

$$\begin{aligned} \boldsymbol{u} &\sim \text{Normal}(\boldsymbol{0}, \delta_1 \boldsymbol{B}^{-1}) \\ \boldsymbol{B} &= \boldsymbol{I} - \rho \boldsymbol{C} \end{aligned}$$

where $\boldsymbol{C}$ is the adjacency matrix to describe the neighborhood structure for counties in MO, ρ measures the spatial correlation strength and $\delta_0 = \delta_1/\eta_1$ measures the spatial variance. Their prior distributions (or densities) are:

$$\begin{aligned} \rho &\sim \text{Unif}(0, \lambda_I^{-1}), \\ [\eta_1] &= \frac{1}{(1+\eta_1)^2}, \quad \eta_1 > 0, \end{aligned}$$

where λ_I is the largest eigenvalue of $\boldsymbol{B}$.

Methods: Estimation

A Markov Chain Monte Carlo algorithm was used to obtain posterior samples for p_{ijkl}. For category $ijkl$, there are $N^-_{ijkl} = N_{ijkl} - n_{ijkl}$ people outside MO–BRFSS data and

$$Y^-_{ijkl} \sim \text{Binomial}(N^-_{ijkl}, p_{ijkl})$$

people who have had CRCS.
For each posterior samples of p_{ijkl} we obtained posterior predictive samples for $Y_{ijkl} = Y^-_{ijkl} + y_{ijkl}$ and

$$p_i = \frac{\sum_{j=1}^3 \sum_{k=1}^2 \sum_{l=1}^2 Y_{ijkl}}{\sum_{j=1}^3 \sum_{k=1}^2 \sum_{l=1}^2 N_{ijkl}}.$$

Therefore, the estimated CRCS prevalence for county i is the mean of the posterior predictive samples of p_i.

Results: Comparison between BRFSS and CLS

The county–level CRCS prevalence estimates from BRFSS generally agree with those from CLS, with an average 5.05%-point difference across all counties in MO. Counties with large sample sizes tend to have more similar estimates to CLS.

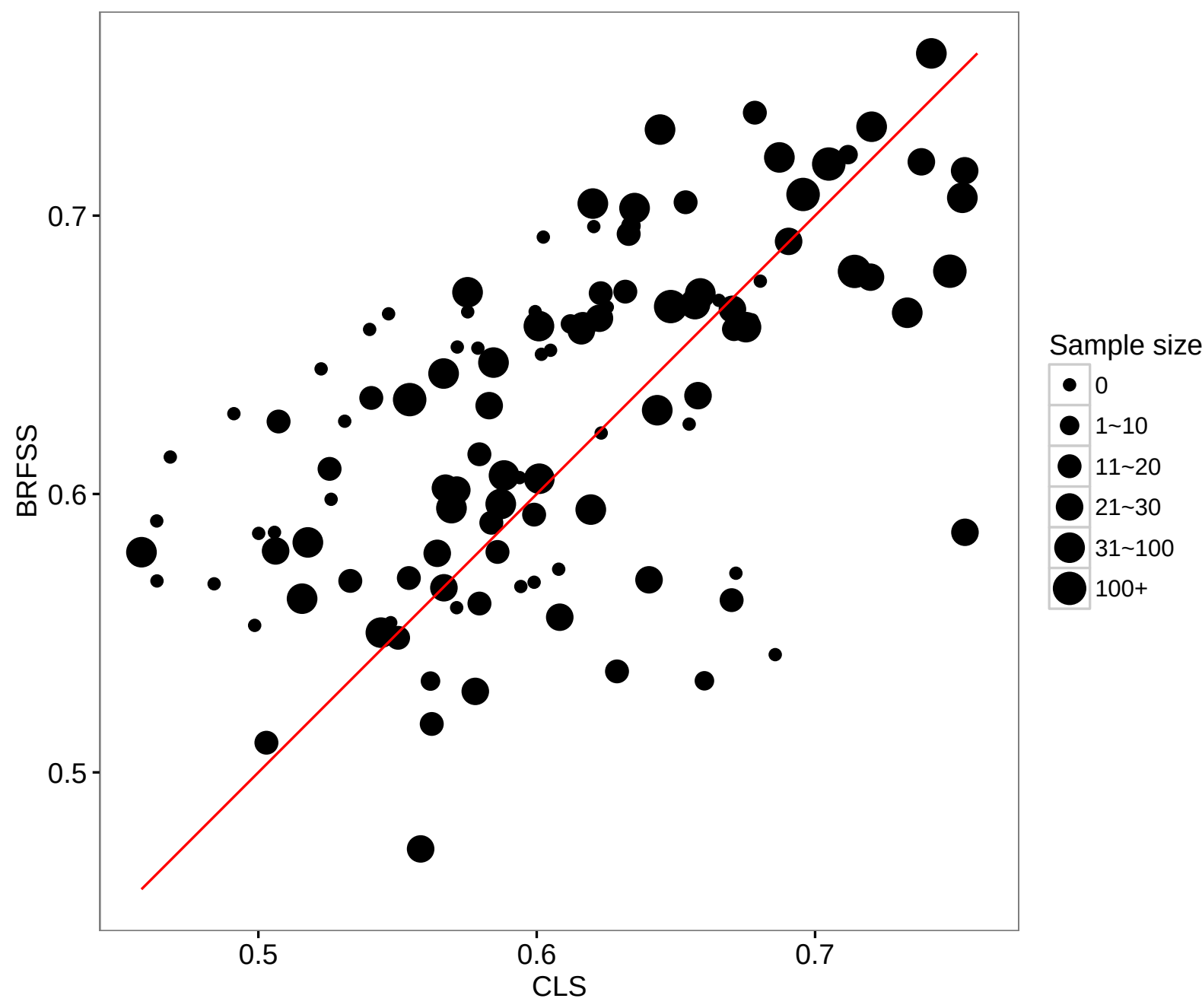

Figure 3: Scatter plot of CRCS prevalence estimates for all counties: CLS vs. BRFSS.

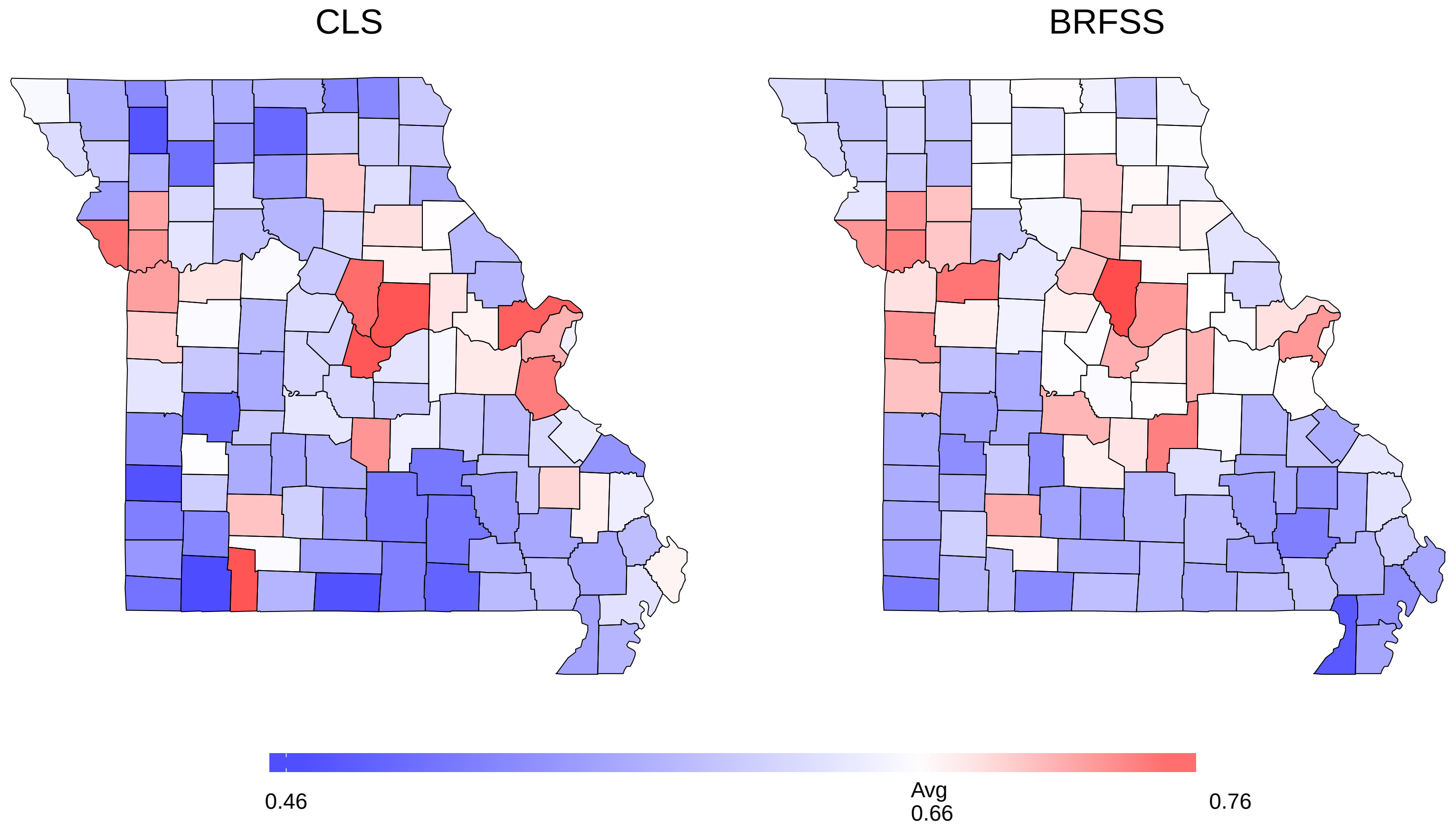

Figure 4: Maps for CRCS prevalence estimates from CLS vs. BRFSS. The white color is the state average CRCS prevalence in Missouri. Our estimates from BRFSS generally have the same spatial variation as those from CLS, but the northern counties are mostly overestimated.

Conclusions

- The differences between BRFSS and CLS for counties with high/low CRCS prevalence are still noticeably large.
- In BRFSS, small or zero sample sizes for counties in Missouri potentially produce biased estimates. It is hard to estimate a whole county's prevalence based only on several or tens of people.
- When BRFSS is the only source to estimate county–level prevalence, our model can still provide reasonable estimates at county level.
- We also used models in Cadwell, *et al.* (2010) to obtain CRCS prevalence estimates. However, due to small sample sizes in Missouri compared to all samples in US, covariances among classes of people were hard to estimate, which added more uncertainty compared to our model.
- We classified people into 12 groups in our analysis. However, when detailed population sizes are available, finer clarification with more demographic variables may help improve the results.
- In our evaluation of the results, we treat CLS (2011) as the true prevalence for comparison. However, the uncertainty from CLS itself was not considered.

The Missouri Cancer Registry and Research Center (MCR-ARC) is supported in part by a cooperative agreement between the Centers for Disease Control and Prevention (CDC) and the Missouri Department of Health and Senior Services (DHSS) (5NU58DP003924-05) and a Surveillance Contract between DHSS and the University of Missouri (MU). We thank DHSS for access to CLS data.

Contact: Jiang Du (jdx66@mail.missouri.edu)

*Cadwell BL, Thompson TJ, Boyle JP, Barker LE (2010). Bayesian small area estimates of diabetes prevalence by U.S. county, 2005. *Journal of Data Science* 8(1): 173-188.