MEDWIN PUBLISHERS
Committed to create value for Researchers

Epidemiology international journal

# Using BRFSS Data to Estimate County-Level Colorectal Cancer Screening Prevalence in Missouri

**Jiang Du[1], Dongchu Sun[1,2]*, Chester Lee Schmaltz[1] and Jeannette Jackson Thompson[1]**

[1]University of Missouri, Columbia, USA

[2]East China Normal University, China

**Research Article**
Volume 2 Issue 1
**Received Date:** March 23, 2018
**Published Date:** May 03, 2018

***Corresponding author** Dongchu Sun, University of Missouri, Columbia, MO 65210, USA and East China Normal University, Shanghai 200062, China, Tel: 1-573-882-7675; Email: sund@missouri.edu

## Abstract

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and the third leading cause of cancer death in both men and women in the US. Colorectal cancer screening (CRCS) serves an important role in the early detection of colorectal cancer and reduces the mortality rate. It is recommended that people aged 50 and older should take CRCS regularly. However, not all people follow the guideline. The state-level CRCS prevalence can be estimated from the Behavioral Risk Factor Surveillance System (BRFSS). Efforts to advocate CRCS are often conducted locally, often at the level of county or county equivalent. Knowing county-level CRCS prevalence can be important to making relevant policies. However, BRFSS does not provide county-level CRCS prevalence estimates. We examined the possibilities of using BRFSS data for county-level estimates with small area estimation (SAE) techniques. Demographic information from both BRFSS and U.S. Census population file were used in our models. In addition, county attributes related to education levels and house incomes were used to improve the estimates. A random spatial effect was also added to capture other county attributes not included in the model. We took the 2012 Missouri BRFSS (MO-BRFSS) data as an example to get county-level CRCS prevalence estimates. To evaluate the results, estimates from 2011 Missouri County Level Study (MO-CLS), which is a BRFSS-like survey but collected hundreds of responses for each county in Missouri, was used as “true” values. The evaluation results indicated the inclusion of county attributes improved the estimates significantly, but not the random spatial effect. The estimates from MO-BRFSS showed similar patterns as those from MO-CLS but less accurate.

**Keywords:** Small area estimation; Bayesian hierarchical models; Survey estimates; BRFSS

## Introduction

### Background about Colorectal Cancer Screening

Colorectal cancer (CRC) is the third most commonly diagnosed cancer and the third leading cause of cancer death in both men and women in the U.S. CRC incidence rates increased from 1975 through the mid-1980s [1]. But have since decreased with the exception of a slight, unexplained bump in rates between 1996 and 1998. Declines have accelerated during the past few years. The incidence rates decreased by more than 4% from 2008 to 2010. The large declines over the past decade have largely been attributed to the detection and removal of polyps as a result of increased colorectal cancer screening (CRCS). As for the mortality rates, they have been decreasing along with the incidence rates. From 2001 to 2010, CRC mortality rates decreased by about 3% per year, compared to declines of about 2% per year in the 1990s. The declines in mortality rates have been attributed to improvements in treatment, changing patterns in CRC risk factors and screening. CRCS plays an important role in the early detection of CRC, which can provide patients higher chance of survival. When screening identifies a colorectal tumor in its early stages, the cost of treatment is often much less expensive than if the tumor is detected later in the course of the disease.

CRCS is important but unfortunately, not everyone takes the screening as recommended, which suggest people over 50 years old should take regular screenings. The possible reasons may include the unawareness of the screening, the cost of screening, the nature of the screening procedures (some people feels uncomfortable with them), etc. Knowing the current CRCS prevalence can be important for policymakers making plans to protect people against CRC. Efforts to promote CRCS are generally conducted locally, often at the level of the county or county equivalent. Counties with low CRCS prevalence should be monitored and policies should be made to encourage people to participate CRCS. However, estimates of CRCS prevalence are often available at the state level, not at finer scales like county level.

### Existing Surveys Covered CRCS in Missouri

There are two kinds of regularly conducted surveys in Missouri which contain information about CRCS prevalence. One is the Behavioral Risk Factor Surveillance System (BRFSS) and the other one is Missouri County-Level Study.

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services [2]. Since its creation by the U.S. Centers for Disease Control and Prevention (CDC) in 1984, the BRFSS has been conducted annually in all 50 states as well as the District of Columbia and three U.S. territories. The Health and Behavioral Risk Research Center at the University of Missouri-Columbia collects BRFSS data for Missouri. To adjust the sampling bias and make sure the data collected are representative of the population for the state, raking (iterative proportional fitting) was used to accomplish this goal [3]. The sample weight for a particular sample generated by raking can be interpreted as the inverse probability of a "likely" or "unlikely" the person been selected. For every two years, questions related to CRCS prevalence are asked in Missouri. For interviewers aged 50 and older, they were asked if they never had a sigmoidoscopy or colonoscopy, which are two major types of CRCS. Based on responses to this question, CRCS prevalence for Missouri can be obtained. We denote the Missouri BRFSS data as MO-BRFSS data. In 2012, there were 5,310 adults interviewed by randomly selected household landline telephone numbers. Additionally, 1,403 randomly selected adult cellphone-mostly users participated in the interview. A CRCS prevalence of 66.5% was reported for Missouri.

The Missouri County-level Study, which we denote as MO-CLS, followed standard CDC BRFSS methods and techniques. The questionnaire used for the interviews contains questions from the BRFSS and CDC Adult Tobacco Survey (ATS). Different from MO-BRFSS which only aims state-level estimates, MO-CLS aims at producing accurate county-level estimates by collecting much more data for each county. For example, the 2011 CLS goal was to complete 47,200 landline interviews with Missouri adults ages 18 and older [4]. The goals were as follows:

- 800 interviews in Jackson County and St. Louis County with 400 interviews among African Americans and other races and 400 interviews among whites.
- 800 interviews in St Louis City with 400 interviews among African Americans and 400 interviews among whites and other races.
- 400 completed interviews in the rest of 112 counties.

Additionally, a goal was established to obtain 4,720 interviews with adult cellphone-only users. Data from cell phone interviews were combined with landline data for analysis at the state and regional levels. Between January and December 2011, interviews were completed with 47,261 landline and 4,828 cell phone only users, totaling 52,089 completed interviews. The same raking method was used for MO-CLS to produce weighting variables. Note that the weighting variable in MO-BRFSS is to make survey sample represent the whole Missouri State, while the weighting variable in MO-CLS makes the survey sample represent each county in Missouri, which produce feasible ways to conduct county-level estimates. For interviewers aged 50 and older, they were asked if they never had a sigmoidoscopy or colonoscopy, which is the same as the question asked in MO-BRFSS. At the state level, 2011 MO-CLS reported CRCS prevalence as 66.2%, which is close to 2012 MO-BRFSS result. The MO-CLS is ideal for producing county-level prevalence estimates, but it is conducted less frequently as MO-BRFSS. Currently, MO-CLS were only conducted for years 2003, 2007, 2011 and 2016.

## Aim of the Paper

It is useful to get county-level CRCS prevalence estimates and MO-CLS can directly serve the purpose. However, the discontinuity in conducting years limits us from monitoring changes over time. MO-BRFSS can give us the state-level estimates every two years, but not for the county-level estimates. There are many counties in Missouri have zero or rather small sample sizes in MO-BRFSS data, which prevent us from getting county-level estimates directly. In this paper, we want to examine the possibilities of using MO-BRFSS data to produce county-level estimates. The process of estimating prevalences for counties with no samples falls into the category of small area estimation (SAE) problems. Rao (2015) is a classical text contains various methods for SAE problems [5]. SAE techniques have been widely used in survey context when one tries to estimate the quantity of interest for an area with no sample available. When applying SAE methods, the survey sampling design should be considered to reduce biases in the estimates.

Once county-level estimates were obtained from MO-BRFSS data, we will use MO-CLS to evaluate the results. The county information is no longer publicly available for MO-BRFSS data after 2012. In addition, the 2016 MO-CLS data is not available at the time of writing this paper. Due to these limitations, we decided to use 2011 MO-CLS data and 2012 MO-BRFSS data as two sources to produce county-level CRCS prevalence estimates. The state-level estimate shows little difference between 2011 MO-CLS and 2012 MO-BRFSS. Thus we assume the prevalence stay unchanged from 2011 to 2012. Due to much large sample sizes in 2011 MO-CLS, its estimates are treated as “true” values. Estimates from MO-BRFSS with SAE methods will be compared to those “true” values and results will be evaluated.

## Review of SAE Methods for Surveys

We briefly review some SAE methods used for survey analysis which have already been proposed in the literature. As in a typical survey analysis, population totals or population proportions (prevalences) are often quantities of interest. Our purpose is to use survey samples to estimate population truth. There are two types of approaches used in practice. The first one is the design-based approach and the other one is the model-based approach.

The design-based approach is carried out based on distributions of the individuals that could appear in the survey. Weighting variables can be interpreted as the inverse of the inclusion probability for each individual. In MO-BRFSS or MO-CLS, the weights obtained from raking are normalized so that the sum of the weights equals the true population size. In other words, the weights for one individual means how many people he (she) is representing for in the whole population. Design-consistent estimators can be generated by the weighted average/sum of sample quantities based on weighting variables. The Horvitz–Thompson (HT) estimator is commonly used in practice [6]. Let $Y_{ik}$ be a binary health outcome for individual $k$ in county $i$ ( $i=1,\ldots,I$ and $k=1,\ldots,N_i$ ), where $N_i$ is the population size for county $i$ and usually known. In a survey, a sample of size $n_i$ is drawn from each county $i$ with sample values $y_{ik}$. The true prevalence for each county $i$, denoted as $P_i$, can be calculated as

$$P_i = \frac{1}{N_i}\sum_{k=1}^{N_i} Y_{ik}. \qquad (1)$$

Suppose the calculated inclusion probability for individual $k$ in county $i$ is $\pi_{ik}$ and the weight is $w_{ik}=1/\pi_{ik}$. For normalized weights, $\sum_{k=1}^{N_i} w_{ik} = N_i$.

Then the HT estimator of $P_i$ is given by

$$\hat{p}_i^{HT} = \frac{1}{N_i}\sum_{k=1}^{N_i}\frac{y_{ik}}{\pi_{ik}}, \qquad (2)$$

with variance

$$\widehat{Var}\left(\hat{p}_i^{HT}\right) = \frac{1}{N_i^2}\left[\sum_{k=1}^{N_k}\left(\frac{1-\pi_{ik}}{\pi_{ik}^2}\right)y_{ik}^2 + \sum_{k=1}^{N_k}\sum_{k\neq k'}\left(\frac{\pi_{ik,ik'}-\pi_{ik}\pi_{ik'}}{\pi_{ik}\pi_{ik'}}\right)\frac{y_{ik}y_{ik'}}{\pi_{ik,ik'}}\right],$$

(3)

where
e $\pi_{ik,ik'}$ is the sampling probability for the pair of individuals $ik$ and $ik'$. For a description of constructing $\pi_{ik,ik'}$, see Lumley (2010, Section 7.2) [7]. The HT estimator is a design-unbiased estimator of $P_i$ and often called the direct estimator since it only uses the responses from the area of interest. One limitation for HT method is that the estimator cannot be calculated for area with no samples. Other design-based methods use a model for construction of the estimators. Pfeffermann (2013) [8] contains an overview of recent developed design-based methods which can produce estimates for non-sampled areas [8].

The model-based approach assumes a hypothetical infinite population from which the responses are drawn. The model-based approach is appealing since standard statistical modeling machinery can be applied. However, it is difficult to implement since the sampling mechanism needs to be modeled. One needs to include all variables used in the sampling process, which are often not available. Even if available, the complex survey itself is hard to model. Gelman [9] describes the issues. Despite of the issues, developing model-based approaches has been an active research area [9]. Mercer, et al. [10] compares several existing methods [10]. The simplest model is the binomial regression model which assumes

$$y_i \mid P_i \sim Binomial(n_i, P_i) \quad \text{and} \quad \text{logit}(P_i) = \beta_0 + u_i + v_i \quad (4)$$

where $y_i = \sum_{k=1}^{n_k} y_{ik}$, $\beta_0$ is the overall intercept, $u_i$ is the spatial effect for county $i$ and $v_i$ is random error, which is similar to a BYM mode [11]. However, such model completely ignores the survey design. Many methods has been proposed for SAE methods which account for survey design by either include the variance estimate $\widehat{Var}\left(\hat{p}_i^{HT}\right)$ as part of the model, for example the logit-normal model [10], the arcsine model [12], or directly modeling the sample weighting variable, for example using penalized splines [13], or combining pseudo-likelihood with the effective sample size calculated by weighting variable [14], or just regressing the response variable on a Gaussian process function of the weighting variable [15]. There are still many other approaches that we have not mentioned.

Unfortunately, the weighting variable is not available in MO-BRFSS data in terms of producing county-level estimates. Thus the model-based approaches mentioned above with weights being part of the model cannot be applied. Similar to our situation, Cadwell, et al. [16] used Bayesian multilevel models to estimate diabetes prevalences for 3,141 counties in the US using BRFSS data [16]. Estimating MO county-level CRCS prevalences is essentially the same problem. Survey respondents were classified into $J$ categories based on age, race and sex. Because of low prevalences of diabetes (5.5% national wide), they assume the number of sample people who have diabetes for county $i$ and category $j$, namely $y_{ij}$ follows a Poisson distribution

$$y_{ij} \mid P_{ij} \sim Poisson\left(n_{ij}P_{ij}\right); \quad i = 1,\dots,I \quad \text{and} \quad j = 1,\dots,J \quad (5)$$

Where $I$ is the number of counties, $n_{ij}$ is the sample size and $P_{ij}$ is the prevalence for category $j$ of county $i$. They further model $P_{ij}$ through

$$\log\left(P_{ij}\right) = \beta_{ij}, \qquad (6)$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{iJ})'$ follows a multivariate normal distribution with mean $\mu_{s[i]}$ and covariance matrix $\boldsymbol{\Sigma}$. Here $s[i]$ denote the state where county $i$ belongs to. They assumed independent normal priors on $\mu_{s[i]}$s and Inverse-Wishart prior on $\boldsymbol{\Sigma}$. From the posterior samples, the posterior predictive distributions of county-level prevalence $P_i$ was obtained with the use of true population sizes from Census data. The mean of the posterior predictive distribution was used as the estimate of $P_i$.

## Outline of our approach

We used Cadwell, et al. (2010) [16] as our main reference with modifications to address our problem [16]. Firstly, we assumed binomial distribution for $y_{ij}$ since the CRCS prevalence is around 60% in our case, where Poisson approximation may not work well. Secondly, we avoided modeling the dependence between different categories of people and assigning an Inverse-Wishart prior on the covariance matrix. Compared to their data, we only have 115 counties with relatively much smaller sample size. Thus the covariances between different categories will be hard to estimate. For simplicity, we assumed independence among categories. Lastly, we added county-level covariates: the percentage of people below high school education level, the percentage of people below 9th-grade education level, the percentage of people above bachelor education level and the median house income. The inclusion of these variables improved our estimates.

## Methods

### Data

Missouri is comprised of 114 counties and the City of St. Louis. For the 2012 MO-BRFSS data, we only selected people over 50 years old as 50 is the recommended age to start CRCS, which left us with 4,605 samples. After removing unknown counties and unknown responses about CRCS, we had a total of 3,807 samples for analysis. In MO-BRFSS study, the sample size from each county was too small to make conclusive estimates by county. Therefore, those areas were clustered into seven MO-BRFSS regions shown in Figure 1. Table 1 contains the sample size for each region. These seven regions are the smallest area units where survey results are reported. Figure 2 shows the sample sizes for each county. Although 3,807 observations seem quite a lot at the state level, it is not enough to get estimates for every county since most samples were from a highly populated urban areas, in which case we have highly unbalanced distribution of survey samples across all counties. For example, there were 37 counties in MO have zero sample (white color) and only 15 counties had more than 50 samples, where 50 is the minimum sample size that CDC will report the prevalence for a Chronic Disease indicator. Therefore, getting accurate county-level CRCS prevalence estimates is quite challenging.

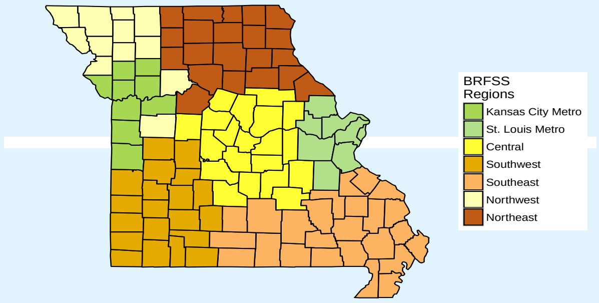


Figure 1: 2012 MO-BRFSS regions.

| Region | Sample Size |
|---|---|
| Kansas City Metro | 767 |
| St. Louis Metro | 1030 |
| Central | 493 |
| Southwest | 485 |
| Southeast | 433 |
| Northwest | 355 |
| Northeast | 244 |

Table 1: 2012 MO-BRFSS regions in MO with sample sizes.

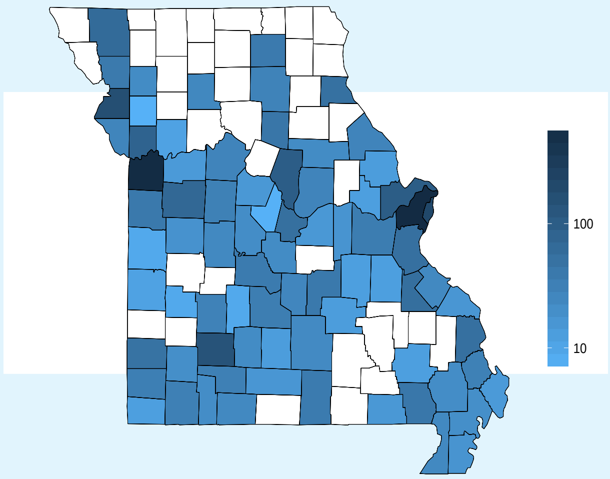


Figure 2: 2012 MO-BRFSS sample sizes counties in Missouri.

For 2011 MO-CLS data, more samples were obtained for each county. After cleaning the data, we have 35,590 samples for people over 50 years old with valid county and demographic information. The sample sizes for all counties ranged from 209 to 706, with a median 300. Weighting variables were also created for county-level estimates in this survey. The availability of the weighting variable allows us to get direct county-level estimates using the HT estimator. With large sample size in MO-CLS, we may assume the direct estimates are accurate.

The U.S. Census Bureau publishes population estimates by demographic characteristics for all counties. We used 2012 Census county projections to obtain the estimates for the population size for different demographic categories in each county in Missouri, which were used for cross-classify 2012 MO-BRFSS data later on. Here we ignore the uncertainty of the population sizes. These population size information will be used to adjust our estimates to match the population distribution for each county.

The county attribute variables can be obtained using SEER*Stat software [17]. It provides a convenient, intuitive mechanism for the analysis of Surveillance, Epidemiology, and End Results Program (SEER) and other cancer-related databases. The SEER*Stat calculated the county attribute variables for 2010-2014 based on the Census American Community Survey (ACS) 5-year files. We used the five years 2010 to 2014 as the year 2012 is the middle year of the span, which we would expect less bias. Note that year 2012 is the one we want to study.

### Prevalence Estimator

Respondents in MO-BRFSS were classified into different categories based on their county at diagnosis, age (50–64, 65–74, 75+), gender (male, female) and race (white, non-white). For each county $i$, $i=1,\dots,I$ where $I=115$ in Missouri, the three age groups, two genders and two races yield twelve different categories. Let $J=12$ be the total number of categories and we use the following notations for each category $j$, $j=1,\dots,J$, in county $i$:

- $n_{ij}$ : the MO-BRFSS sample size, which is the number of respondents;
- $y_{ij}$ : the number of respondents who have had CRCS out of all $n_{ij}$ respondents;
- $N_{ij}$ : the true population size based on 2010 Census data;
- $Y_{ij}$ : the true population total who have had CRCS out of $N_{ij}$ people;
- $P_{ij}$ : the true proportion of people who have had CRCS.

In the variables above, $n_{ij}$ and $y_{ij}$ are known through our MO-BRFSS data. If $n_{ij}=0$, we set $y_{ij}=0$. $N_{ij}$ is also known from 2010 Census data. The other two quantities $Y_{ij}$ and $P_{ij}$ are unobserved. For county $i$, the unobserved prevalence is:

$$P_i = \frac{\sum_{j=1}^{J} Y_{ij}}{\sum_{j=1}^{J} N_{ij}}, \qquad (7)$$

which is the quantity we are interested in.

We define $Z_{ij}=Y_{ij}-y_{ij}$ to be the total number of people in 2012 who have had CRCS in county $i$ category $j$ but not included in the survey. Then (7) is equivalent to

$$P_i = \frac{\sum_{j=1}^{J} (Z_{ij}+Y_{ij})}{\sum_{j=1}^{J} N_{ij}}. \qquad (8)$$

### Bayesian Binomial Regression

We used a Bayesian binomial regression framework to estimate $P_{ij}$. To be specific, we assume $y_{ij}$ follows a Binomial distribution with number of trials $n_{ij}$ and success probability $P_{ij}$:

$$y_{ij} \sim \text{Binomial}\left(n_{ij}, P_{ij}\right). \qquad (9)$$

A logit transformation $v_{ij}=\log\left(P_{ij}/\left(1-P_{ij}\right)\right)$ was used in the second level regression. With all available covariates, $v_{ij}$ can be modeled as:

$$v_{ij}=\mu+\alpha_{r(i)}+\beta_j+\phi_1 x_{1i}+\phi_2 x_{2i}+\phi_3 x_{3i}+\phi_4 x_{4i}+u_i+\epsilon_{ij}, \qquad (10)$$

where

- $\mu$ is the overall intercept;
- $\alpha_{r(i)}$ is the regional effect of region $r(i)\in\{1,2,\dots,7\}$ (recall that we have seven BRFSS regions for 2012 MO-BRFSS);
- $\beta_j$ is effect for the $j$ th demographic category;
- $x_{1i}$ is the percentage of people below high school education level, with coefficient $\phi_1$;
- $x_{2i}$ is the percentage of people below 9th-grade education level, with coefficient $\phi_2$;
- $x_{3i}$ is the percentage of people above bachelor education level, with coefficient $\phi_3$;
- $x_{4i}$ is the median house income, with coefficient $\phi_4$;
- $u_i$ is a random spatial effect which accounts for extra county level variability not included in our model if exist;
- $\epsilon_{ijkl}$ is the over dispersion term accounting for extra variability not included in our model.

For $\alpha_{r(i)}$ s and $\beta_j$ s, sum-to-zero constraints are added for identifiability issues.

It is helpful to write our model in vector notation. We define $\mathbf{y} = (\mathbf{y}_1', \dots, \mathbf{y}_I')'$ where $\mathbf{y}_i=(y_{i1},\dots,y_{iJ})'$. The vectors $\mathbf{n},\mathbf{N},\mathbf{P}$ and $\mathbf{v}$ are defined in the same way as $\mathbf{y}$. The likelihood function (9) can then be rewritten as:

$$\mathbf{y} \sim \text{Binomial}(\mathbf{n},\mathbf{P}). \qquad (11)$$

For the regression (10), we define a $IJ\times 7$ design matrix $\mathbf{X}_\alpha$ with each row being an indicator where only the $r(i)$th element is one and all remaining elements are zeros. We collect $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_7)'$. Then $\mathbf{X}_\alpha\boldsymbol{\alpha}$ gives a vector to indicate the region each element in $\mathbf{v}$ belongs to. By the same methods, we create a $IJ \times J$ design matrix $\mathbf{X}_\beta = \mathbf{I}_J \otimes \mathbf{1}_I$ with vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ where $\mathbf{I}_J$ is an identity matrix of size $J$, $\mathbf{1}_I$ is a vector with all ones of size $I$ and $\otimes$ is the Kronecker product. For the county attributes, let $\phi = (\phi_1, \phi_2, \phi_3, \phi_4)'$ and $\mathbf{x}_i = (x_{1i}, x_{2i}, x_{3i}, x_{4i})'$, then $\phi_1 x_{1i} + \phi_2 x_{2i} + \phi_3 x_{3i} + \phi_4 x_{4i} = \mathbf{x_i}'\phi$. The design matrix for $\phi$ can be written as $\mathbf{X}_C = \mathbf{1}_J \otimes (\mathbf{x}_1, \ldots, \mathbf{x}_I)'$. For the random spatial effect $u_i$, let $\mathbf{u} = (u_1, \ldots, u_I)'$. The design matrix for $\mathbf{u}$ is $\mathbf{X}_u = \mathbf{1}_J \otimes \mathbf{I}_I$. Under all these definitions, model (10) can be rewritten as

$$\mathbf{v} = \mu \mathbf{1}_{IJ} + \mathbf{X}_\alpha \boldsymbol{\alpha} + \mathbf{X}_\beta \boldsymbol{\beta} + \mathbf{X}_C \phi + \mathbf{X}_u \mathbf{u} + \boldsymbol{\epsilon} \qquad (12)$$

where $\boldsymbol{\epsilon}$ a size $IJ$ vector collects all $\epsilon_{ij}$ in the same way as $\mathbf{v}$.

## Prior Distributions on the Regression Parameters

Flat priors are used for parameters $\mu$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\phi$. We assume independence among all these parameters. Normal distributions with large variances are used as their prior distributions:

$$\mu \sim N(0,\gamma), \quad \boldsymbol{\alpha} \sim N_7(\mathbf{0},\gamma\mathbf{I}), \quad \boldsymbol{\beta} \sim N_J(\mathbf{0},\gamma\mathbf{I}), \quad \phi \sim N_4(\mathbf{0},\gamma\mathbf{I}), \qquad (13)$$

where $N_m(\cdot,\cdot)$ indicate a multivariate normal distribution of dimension $m$, $\mathbf{0}$ is a vector of zeros and $\mathbf{I}$ is an identity matrix with dimensions corresponding to the multivariate normal distribution. Here we use a large variance $\gamma$ for the purpose of non-informative priors.

For the over-dispersion parameter $\boldsymbol{\epsilon}$, we assume

$$\boldsymbol{\epsilon} \sim N_{IJ}(\mathbf{0}, \delta_0 \mathbf{I}), \qquad (14)$$

where $\delta_0$ is the variance parameter for $\boldsymbol{\epsilon}$.

We used a CAR model for the spatial effect $\mathbf{u} = (u_1, \ldots, u_I)'$ in our analysis. We have a brief review here. For each $u_i$ defined at county $i$, the CAR model assumes it has at least one neighboring county. We use $i' \sim i$ to denote that county $i'$ is adjacent to $i$. Based on the locations for all the counties, let $\mathbf{C} = (c_{i'i})_{I \times I}$ be an adjacency matrix describes the neighboring structure of the counties in MO. The element $c_{i'i} = 1$ if county $i'$ adjacent to county $i$ and $c_{i'i} = 0$ otherwise. By convention, a county will not be the neighbor of itself so we define $c_{ii} = 0$. Then the CAR model specifies the conditional densities of $u_i$ given all the other variables to be:

$$\left(u_i \mid u_{i'}, i' \neq i\right) \sim N\left(\rho \sum_{i' \sim i} u_{i'}, \delta\right), \qquad (15)$$

where $\delta > 0$ specifies the conditional variance, $\sum_{i' \sim i} u_{i'}$ is the sum of neighboring variables around county $i$, and $\rho$ specifies the strength of the relationship between county $i$ and its neighbors. It has been shown that (15) is equivalent to the multivariate normal distribution [18].

$$[\mathbf{u} \mid \delta, \rho] \sim N_I(\mathbf{0}, \delta(\mathbf{I} - \rho\mathbf{C})^{-1}). \qquad (16)$$

The matrix $\delta(\mathbf{I} - \rho\mathbf{C})^{-1}$ need to be positive definite, which requires $\rho$ in the range:

$$\rho \in \left(\frac{1}{\min \, v_i}, \frac{1}{\max \, v_i}\right), \qquad (17)$$

where the $v_i$ s are eigen values of the adjacency matrix $\mathbf{C}$. For counties in Missouri, the range for $\rho$ is $(\rho_{\min}, \rho_{\max})$, where $\rho_{\min} = -0.347$ and $\rho_{\max} = 0.173$.

## Selection of Hyperpriors

Parameters in the priors distribution need to be specified for a full Bayesian analysis. The variance parameter $\delta_0$ in (14) was given an Inverse-Gamma $(a_0, b_0)$ prior distribution with density proportional to

$$[\delta_0 \mid a_0, b_0] \propto \frac{1}{\tau_1^{a_0 - 1}} \exp\left(-\frac{b_0}{\delta_0}\right), \quad \text{for } \delta_0 > 0. \qquad (18)$$

Here we use $[\delta_0 \mid a_0, b_0]$ to denote the probability density function of $\delta_0$ conditional on $a_0$ and $b_0$.

The parameter $\rho$ controls the strength of spatial association and is restricted in SEER (17). Thus we used a uniform prior distribution

$$\rho \sim \text{Unif}\left(\rho_{\min}, \rho_{\max}\right). \qquad (19)$$

To complete our hierarchical model, we still need a prior on the variance parameter $\delta$ in (16). Instead of assigning a prior distribution on $\delta$ independently from $\delta_0$, we connect $\delta_0$ with $\delta$ by the noise-to-signal ratio $\eta = \delta_0 / \delta$. We adapted the idea from Cheng and

Speckman (2012) [19], where they used a scaled Pareto prior on $\eta$ in Bayesian spline settings,

$$\left[\eta | a\right] = \frac{a}{(a+\eta)^2}, \quad \eta > 0. \quad (20)$$

The use of noise-to-signal ratio has the benefit of introducing dependence between unstructured error $\boldsymbol{\epsilon}$ and spatially structured random effect $\mathbf{u}$ through their variance parameter. If we define the proportion of spatial variance as

$$\phi = \frac{\delta}{\delta_0 + \delta}, \quad (21)$$

then the $\phi$ is distributed as

$$[\phi | a] = \frac{a}{[(a-1)\phi+1]^2}. \quad (22)$$

In this sense, the total variation is separated into the structured spatial variation with proportion $\phi$ and the unstructured random noise with proportion $1-\phi$. Figure 3 shows the density plot of $\phi$ with $a = 0.5, 1$ and $2$. When $a = 1, \phi \sim \text{Uniform}(0,1)$; when $a < 1$, more weight is put on the spatial random effect; when $a > 1$, more weight is put on the unstructured random effects. In our application, we choose $a = 1$ so that

$$[\phi | a=1] \sim \text{Unif}(0,1), \quad (23)$$

which treats the spatial variation and unstructured error variation equally likely.

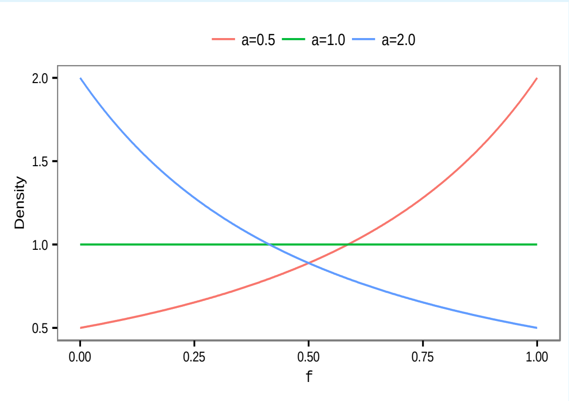


Figure 3: Prior distributions of spatial variance proportion $\phi$ under different values of $a$.

### Estimates of CRCS Prevalence

Recall in (8) $Z_{ij}$ is needed for county-level CRCS prevalence estimates. From our hierarchical model, the posterior predictive distribution of $Z_{ij}$ is

$$Z_{ij} | \mathbf{y}, \mathbf{n}, \mathbf{N} \sim \text{Binomial}(N_{ij} - n_{ij}, P_{ij}) \quad (24)$$

with expected value $E[Z_{ij} | \mathbf{y}, \mathbf{n}, \mathbf{N}] = (N_{ij} - n_{ij})P_{ij}$.

Then the prevalence can be estimated as

$$\hat{P}_i = E(P_i | \mathbf{y}, \mathbf{n}, \mathbf{N}) = E\left( \frac{\sum_{j=1}^{J} (Z_{ij} + Y_{ij})}{\sum_{j=1}^{J} N_{ij}} | \mathbf{y}, \mathbf{n}, \mathbf{N} \right) = \frac{\sum_{j=1}^{J} [\mathrm{E}[\mathrm{Z}_{ij} | \mathbf{y}, \mathbf{n}, \mathbf{N}] + \mathrm{Y}_{ij}]}{\sum_{j=1}^{J} N_{ij}}. \quad (25)$$

Posterior distributions of $P_{ij}$ can be obtained from our Bayesian hierarchical models, thus posterior predictive distribution of $\hat{P}_i$ can also be obtained.

## Results

We fit the model described in previous section with 2012 MO-BRFSS data. We treat the regression model (12) for the linear predictor $\mathbf{v}$ as our full model, which contains all covariates available in our analysis: the demographic information, county attributes and random spatial effects. However, not all covariates may be needed to produce reasonable estimates. We want to know how good the estimates are with the absence of some covariates. Three additional models with fewer covariates were fitted. Table 2 contains the forms of all different models we checked. Model 1 is our full model contains all covariates. In Model 2 we remove the random county effect. This is considered because with the presence of county attributes and the sparsity of our data, the random county effect can be hard to estimate. Model 3 only contains the demographic covariates and the random county effect, which mimics the situation where county attributes are not available. In the end, Model 4 is the simplest model including only the demographic covariates. Note that the regional effects were retained for all models.

| | Covariates | $\mathbf{v}$ |
|---|---|---|
| Model 1 | demographic + county attributes + spatial | $\mu\mathbf{1}_{IJ} + \mathbf{X}_\alpha\boldsymbol{\alpha} + \mathbf{X}_\beta\boldsymbol{\beta} + \mathbf{X}_c\phi + \mathbf{X}_3\mathbf{u} + \epsilon$ |
| Model 2 | demographic + county attributes | $\mu\mathbf{1}_{IJ} + \mathbf{X}_\alpha\boldsymbol{\alpha} + \mathbf{X}_\beta\boldsymbol{\beta} + \mathbf{X}_c\phi + \epsilon$ |
| Model 3 | demographic + spatial | $\mu\mathbf{1}_{IJ} + \mathbf{X}_\alpha\boldsymbol{\alpha} + \mathbf{X}_\beta\boldsymbol{\beta} + \mathbf{X}_3\mathbf{u} + \epsilon$ |
| Model 4 | Demographic | $\mu\mathbf{1}_{IJ} + \mathbf{X}_\alpha\boldsymbol{\alpha} + \mathbf{X}_\beta\boldsymbol{\beta} + \epsilon$ |

Table 2: Models with different covariates.

## Computation

A Markov Chain Monte Carlo (MCMC) algorithm was used to generate the posterior distributions of parameters in our model. A Gibbs sampling algorithm was implemented in R and C++ with the full conditional densities derived in Appendix. The full conditional distribution for $\boldsymbol{\psi} = c(\mu, \boldsymbol{\alpha}', \boldsymbol{\beta}', \phi')'$ is multivariate normal distributions. The full conditional distribution for $\delta_0$ is inverse-gamma distributions. These random variables can be sampled directly with available software. For all the other parameters, Adaptive Rejection Metropolis Sampling (ARMS) introduced in Gilks, Best, and Tan (1995) [20] was used to get the posterior samples via a C++ code modified from the C code by Gilks [21]. In our application, the following hyper parameter values were used:

$$a_0 = b_0 = a = 1, \quad \gamma = 10^4. \qquad (26)$$

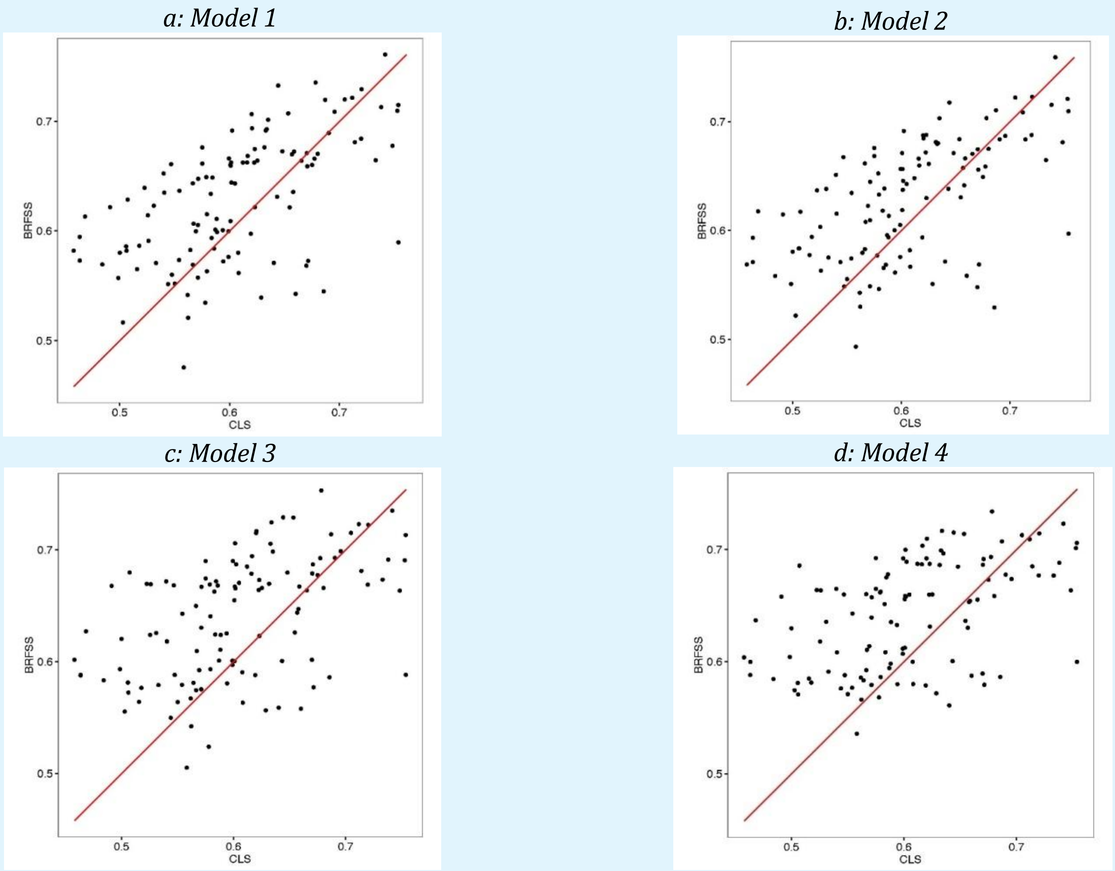
We used 20,000 posterior samples for all models in Table 2 after discarding the first 10,000 ones. Convergences were monitored via trace plots and posterior density plots.

## Model Evaluation

In our application, 20,000 posterior samples of $\hat{P}_i$ were obtained for each county $i$ based on the posterior predictive distribution in (25). The mean $\hat{p}_i$ of the posterior predictive distribution was used as the estimate of $P_i$. We define $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_I)'$ to be the estimates for all counties from MO-BRFSS. The direct HT estimates of $P_i$ s from MO-CLS were also obtained as our "true" values for comparison and we denote them as $\mathbf{p}^{CLS} = (p_1^{CLS}, \ldots, p_I^{CLS})'$. Figure 4 contains the scatter plots of estimates from MO-BRFSS against CLS for Models 1, 2, 3 and 4. The closer are the points from the diagonal line, the better our estimates are.

To evaluate the performance of different models in Table 2, we measure the closeness between the estimates from one of the four models and the estimates from CLS. We used the mean absolute difference (MAD), the Pearson correlation and Spearman correlation as three metrics. Table 3 contains the results for all model.

Smaller values of MAD and larger values of Pearson/Spearman correlation will indicate $\hat{\mathbf{p}}$ is closer to $\mathbf{p}^{CLS}$. Firstly, we notice that the MAD and Pearson Correlations are consistent with each other. A model with a smaller MAD value will have a Pearson larger correlation value. However, the Spearman correlation shows a different trend when comparing Models 1 and 2, or Models 3 and 4. Secondly, by comparing Models 1 and 2 with Models 3 and 4, we notice the presence of county attributes covariates improved our estimates. Thirdly, by comparing Model 1 with Model 2, or Model 3 with Model 4, we notice that the inclusion of the random county effect performs slightly worse in terms of MAD and Pearson correlation. However, the random county effect slightly increases the Spearman correlation. This can be noticed from Figure 4 as the points for Models 1 and 3 spread more evenly along the diagonal line than Models 2 and 4. But the improvement is rather small. In situations where a large amount of counties with zero or small sample sizes, the spatial effects are hard to make a difference. However, it does not hurt much to include random spatial effect when prediction is our main interest. Finally, Models 1 and 2 performs the best based on our criteria and the differences between them is negligible. For illustration, we mapped the CRCS prevalences estimates from Model 1 in Figure 5, which also contains the spatial plots of the estimates from CLS for compression. The white color represents the statewide CRCS prevalence estimated from CLS, which is around 0.66. In general, estimates from Model 1 differed from CLS by 0.05118 (or 5.118%) on average. Accurate county prevalence estimates may not be obtained solely from MO-BRFSS data. However, the spatial variation of CRCS prevalence for all counties can be reflected by MO-BRFSS data if we comparing the two spatial plots in Figure 5.



*a: Model 1*

*b: Model 2*

*c: Model 3*

*d: Model 4*

Figure 4: Scatter plots of estimates from MO-BRFSS versus CLS for Models 1, 2, 3 and 4.

| | MAD | Pearson | Spearman |
|---|---|---|---|
| Model 1 | 0.05118 | 0.582 | 0.606 |
| Model 2 | 0.04949 | 0.591 | 0.603 |
| Model 3 | 0.05894 | 0.487 | 0.514 |
| Model 4 | 0.05887 | 0.491 | 0.507 |

Table 3: Model evaluation.

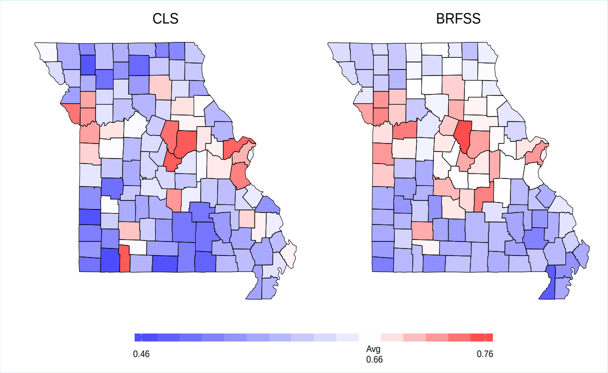


Figure 5: Spatial plots for CRCS prevalence estimated from CLS and MO-BRFSS Model 2.

## Discussion

In this paper, we explored the possibilities of utilizing MO-BRFSS data for county-level CRCS prevalence study. Model-based survey analysis methods were combined with small area techniques to produce county CRCS prevalence in Missouri. Adjustments based on 2010 Census population data were used to correct the bias from MO-BRFSS data, which has no weighting variable for a regular county-level survey analysis. Besides the demographic covariates, we include county attributes to improve our estimates. However, due to zero or small sample sizes for many counties in MO-BRFSS data, our attempt of including random county effect did not bring us any significant improvements in our estimates.

We classified people into 12 categories based on their age groups, gender and race. Ideally, finer classification with more demographic variables can provide more accurate estimates. However, our MO-BRFSS data does not support such fine categories. In general, our model based on MO-BRFSS data can provide similar but less accurate point estimates compared to CLS and spatial variation can be effectively investigated.

The estimates from CLS were treated as “true” values for comparison. However, the variance of the HT estimator was not considered. In addition, we assumed the prevalences of CRCS are the same between the year 2011 and 2012. The state-level prevalence estimates between them are very close. However, we don’t know if the similarity holds at county-level.

We only studied CRCS prevalence in our paper. In general, BRFSS data contain many health-related questions so that similar studies can be conducted. However, zero or small sample sizes for some counties may still be a block when one tries to produce accurate county-level estimates.

## Acknowledgement

Most work is done while the first author is a research assistant at Missouri Cancer Registry. Sun's research was supported by NSF grant SES-1260806, Chinese 111 Project B14019 and Chinese NSF grant 11671146.

## References

1. Atlanta: American Cancer Society (2014) Colorectal Cancer Facts & Figures 2014-2016.

2. The Behavioral Risk Factor Surveillance System https://www.cdc.gov/brfss/about/index.htm. [Accessed 27 March 2018]

3. Missouri BRFSS data reports, [Online]. Available: http://health.mo.gov/data/cls. [Accessed 27 March 2018].

4. (2011) Missouri County-Level Study, [Online]. Available:http://health.mo.gov/data/cls/designmethodology.php. [Accessed 27 March 2018].

5. Rao JNK (2015) Small-Area Estimation. Wiley Online Library.

6. Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association 47(260): 663-685.

7. Lumley T (2011) Complex surveys: A guide to analysis using R. 565(Volume), John Wiley & Sons.

8. Pfeffermann D (2013) New important developments in small area estimation. Statistical Science 28(1): 40-68.

9. Gelman A (2007) Struggles with survey weighting and regression modeling. Statistical Science 22(2): 153-164.

10. Mercer L, Wakefield J, Chen C, Lumley T (2014) A comparison of spatial smoothing methods for small area estimation with sampling weights. Spat Stat 8: 69-85.

11. Besag J, York J, Molli A (1991) Bayesian image restoration, with two applications in spatial statistics," Annals of the institute of statistical mathematics 43(1): 1-20.

12. Raghunathan TE, Xie D, Schenker N, Parsons VL, Davis WW, et al. (2007) Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. Journal of the American Statistical Association 102(478): 474-486.

13. Vandendijck Y, Faes C, Kirby RS, Lawson A, Hens N (2016) Model-based inference for small area estimation with sampling weights. Spatial Statistics 18: 455-473.

14. Chen C, Wakefield J, Lumely T (2014) The use of sampling weights in Bayesian hierarchical models for small area estimation. Spatial and spatio-temporal epidemiology 11: 33-43.

15. Si Y, Pillai NS, Gelman A (2015) Bayesian nonparametric weighted sampling inference. Bayesian Analysis 10(3): 605-625.

16. Cadwell BL, Thompson TJ, Boyle JP, Barker LE (2010) Bayesian small area estimates of diabetes prevalence by US county, 2005. J Data Sci 8: 173-188.

17. SEER*Stat. [Online]. Available: https://seer.cancer.gov/seerstat/. [Accessed 27 March 2018].

18. Sun D, Tsutakawa RK, Speckman PL (1999) Posterior distribution of hierarchical models using CAR (1) distributions. Biometrika, pp. 341-350.

19. Cheng CI, Speckman PL (2012) Bayesian smoothing spline analysis of variance. Computational Statistics & Data Analysis 56: 3945-3958.

20. Gilks WR, Best NG, Tan KKC (1995) Adaptive rejection Metropolis sampling within Gibbs sampling," Applied Statistics, pp: 455-472.

21. "Adaptive Rejection Sampling," [Online]. Available: https://www1.maths.leeds.ac.uk/~wally.gilks/adaptive.rejection/web_page/Welcome.html [Accessed 27 March 2018].