Appendix for “Small area estimation of cancer risk factors and screening behaviors in U.S. counties by combing two large national health surveys” by Benmei Liu, Van Parsons, Eric J. Feuer, Qiang Pan, Machelle Town, Trivellore E. Raghunathan, Nathaniel Schenker, Dawei Xie

This appendix contains the details of the small area models and the county-level covariates (Table A1) used for the data periods 2004-2007 and 2008-2010.

**The Small Area Model**

As an extension to the Raghunathan et al [1] model, the following multi-level mixed effect model was developed to include the cell-phone only component into the model for one outcome of interest at a specific data period (e.g., *colorectal endoscopy* screening rates in 2008-2010):

Level 1:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ z_i \end{pmatrix} = \begin{pmatrix} \arcsin\left(\sqrt{p_{y1i}}\right) \\ \arcsin\left(\sqrt{p_{y2i}}\right) \\ \arcsin\left(\sqrt{p_{y3i}}\right) \\ \arcsin\left(\sqrt{p_{zi}}\right) \end{pmatrix} \sim N_4\left[\begin{pmatrix} \theta_{1i} \\ \theta_{2i} \\ \theta_{3i} \\ (1+\delta_i)\theta_{1i} \end{pmatrix}, \begin{pmatrix} 1/(4\tilde{n}_{1i}) & 0 & 0 & 0 \\ 0 & 1/(4\tilde{n}_{2i}) & 0 & 0 \\ 0 & 0 & 1/(4\tilde{n}_{3i}) & 0 \\ 0 & 0 & 0 & 1/(4\tilde{n}_{zi}) \end{pmatrix}\right] \quad (1)$$

Where $p_{y1i}, p_{y2i}$, and $p_{y3i}$ are the NHIS direct estimates of the outcome (i.e., survey weighted proportions) and $y_{1i}$, $y_{2i}$, and $y_{3i}$ are the corresponding estimates after arcsin-square-root transformation for households with landline phone, households with cellphone only, and households without any phone in the NHIS in area $i$ ( $i = 1, \dots, m$) for a specific time period (2008-2010). A working covariance matrix which assumes an independence structure among the estimates is used. Covariate variables used are listed in Table S1. The variable $p_{zi}$ is the BRFSS direct estimate of the outcome and $z_i$ is the corresponding estimate after arcsin-square-root transformation. The county-level direct estimates $p_{y1i}, p_{y2i}, p_{y3i}$ and $p_{zi}$ are ratio estimators. The bias of those estimators is negligible for large samples [2]. Our grouping of multiple years of data into data periods enlarges county-level sample sizes thus help reduce the potential bias of those estimators especially for counties with smaller sample sizes. The parameters $\theta_{1i}, \theta_{2i}$, and $\theta_{3i}$ are the unknown population parameters corresponding to the direct estimates after arcsin-square-root transformation. The parameter $(1+\delta_i)$ measures the proportionate bias in the BRFSS estimates relative to the NHIS estimate (see page 479 of Raghunathan et al [1]). The variables $\tilde{n}_{1i}, \tilde{n}_{2i}, \tilde{n}_{3i}, \tilde{n}_{zi}$ are the effective sample sizes (sample sizes divided by estimated design effects) corresponding to the direct estimates.

Level 2:

$$\omega_i = \boldsymbol{\beta} X_i + \eta_i, \text{and}\, \eta_i \sim N_4(\mathbf{0}, \boldsymbol{\Sigma}), \tag{2}$$

where $\omega_i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \delta_i)'$, $X_i$ is a $p \times 1$ vector of covariates, $\boldsymbol{\beta}$ is a $4 \times p$ matrix of regression coefficients and $\boldsymbol{\Sigma}$ is a $4 \times 4$ covariance matrix.

Both $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are unknown hyperparameters and need to be estimated from the model fitting with the observed data.

*Model implementation and inference*

The ultimate goal was to obtain prevalence estimates (with standard errors) for all areas (counties) for each outcome of interest for each data period. Suppose that $M_{1i}$ and $M_{2i}$ denote the proportions of target population living in households with landline phones and cellphones only for county $i$. The inferential quantity (i.e., the estimand) of interest is the composite proportion:

$$\mu_i = M_{1i} \sin^2 \theta_{1i} + M_{2i} \sin^2 \theta_{2i} + (1 - M_{1i} - M_{2i}) \sin^2 \theta_{3i}\,. \tag{3}$$

Estimation of $\mu_i$ involves estimation of $\theta_{1i}, \theta_{2i}$, $\theta_{3i}$, $M_{1i}$ and $M_{2i}$. Given the complex nature of the model and the relatively large number of parameters to estimate, we use a fully hierarchical

Bayesian approach to estimate $\theta_{1i}$, $\theta_{2i}$, and $\theta_{3i}$. We assume a diffuse proper prior for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ with columns of $\boldsymbol{\beta}$ having independent multivariate normal distributions, $N_p(\mathbf{0},\ 10^4\mathbf{I}_p)$, where $\mathbf{I}_p$ is a $p \times p$ identity matrix. The covariance matrix $\boldsymbol{\Sigma}$ is assumed to follow a Wishart distribution with $d_0 = 4$ degrees of freedom and scale matrix $\mathbf{R}_0$, where $\mathbf{R}_0 = 10^{-4}\mathbf{I}_4$. These prior distributions are relative diffuse, but assure that the posterior distributions will be proper.

The Markov Chain Monte Carlo (MCMC) technique of Gibbs sampling [3] is adopted and implemented using the GAUSS programming software [4]. Ten parallel chains, each of length 10,000, were used in Gibbs sampling. The first 5,000 draws from each sequence were discarded, and then the next 5,000 were included in computing posterior means and variances. Draws were pooled across the 10 parallel sequences, so that a total of 50,000 draws were used to compute each summary. The Gelman-Rubin potential scale reduction factor $\hat{R}$ [5] is used to assess convergence of the MCMC models. For counties without any sample from the NHIS or BRFSS, the final estimates are predicted from the same model through the Gibbs sampling process. Technical details on how the Gibbs sampling works can be seen at the Appendix of Raghunathan et al [1].

A two-step small area modeling approach is developed to estimate $M_{1i}$ and $M_{2i}$ for data periods 2004-2007 and 2008-2010. Step 1 estimates $M_i^* = (1 - M_{1i} - M_{2i})$ using a linear mixed model $\hat{y}_i = x_i'\beta + v_i + e_i$, where $\hat{y}_i$ is the direct estimate of $M_i^*$ obtained from the NHIS after taking the arcsin-square-root transformation of the direct estimates, $x_i$ are a set of covariates selected using principle component analysis, $v_i \sim N(0, \sigma_v^2)$ is the random effect, and $e_i \sim N(0, \sigma_e^2)$ is the error term. Step 2 estimates $M_i^{**} = M_{1i}/(M_{1i} + M_{2i})$ using the same modeling approach as used in step 1. A fully Bayesian approach is used to estimate $M_i^*$ and $M_i^{**}$. Finally, $M_{1i}$ and

$M_{2i}$ come be computed using the results obtained from step 1 and step 2. The final accepted MCMC values for $M_{1i}$ and $M_{2i}$ are combined with those MCMC values for $\theta_{1i}, \theta_{2i}$ and $\theta_{3i}$ to compute the posterior mean, standard deviation, and selected percentiles of $\mu_i$ using formula (3).

References


1. Raghunathan T E, Xie D, Schenker N, Parsons VL, Davis WW, Dodd KW, Feuer EJ. Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association* 2007, 102(478)**,** 474-486.
2. Cochran W.G. Sampling techniques. John Wiley & Sons Inc.
3. Gelfand AE and Smith AMF. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990, 85, 398–409.
4. Aptech Systems. *Gauss: Advanced Mathematical and Statistical Systems*. Version 5, Black Diamond, Washington, 2003.
5. Gelman A. and Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992, 7, 457–472.

Table A1: The county-level covariates

| Variables | Source |
| --- | --- |
| 1. Proportion of persons who are Black 2005-09[1] | USA Census County Stat |
| 2. Proportion of persons who are Hispanic 2005-09 | USA Census County Stat |
| 3. Proportion of persons with high school+ education 2005-09 | USA Census County Stat |
| 4. Proportion of persons with college+ education 2005-09 | USA Census County Stat |
| 5. Property tax per capita, 2002 | USA Census County Stat |
| 6. Local government revenue per capita, 2002 | USA Census County Stat |
| 7. Federal expense per capita 2005-09 | USA Census County Stat |
| 8. Social security beneficiaries, 2005-09 | USA Census County Stat |
| 9. Mean income per capita, 2005-09 | USA Census County Stat |
| 10. Median household income, 2005-09 | USA Census County Stat |
| 11. Proportion of persons under poverty, 2005-09 | USA Census County Stat |
| 12. Proportion of persons living in rural area, 2000 census | USA Census County Stat |
| 13. Unemployment rate, 2005-09 | USA Census County Stat |
| 14. Violence and property crimes, 2005-08 | USA Census County Stat |
| 15. Retail, eating and drinking expense per household, 2007 | USA Census County Stat |
| 16. Household size, 2005-09 | USA Census County Stat |
| 17. Proportion of households with female head, 2005-09 | USA Census County Stat |
| 18. Proportion of households with children under 18, 2005-09 | USA Census County Stat |
| 19. Proportion of households with only one person, 2005-09 | USA Census County Stat |
| 20. Births, 2005-09 | USA Census County Stat |
| 21. Deaths, 2005-09 | USA Census County Stat |
| 22. Population, 2005-09 | USA Census County Stat |
| 23. Persons per square mile, 2010 census | USA Census County Stat |
| 24. Proportion of persons aged 65+ among those aged 18+, 2005-09 | USA Census County Stat |
| 25. Median home value, 2005- 09 | USA Census County Stat |
| 26. Proportion of workers with commute time less than 30 minutes, 2005-09 | USA Census County Stat |
| 27. Buying power index | USA Census County Stat |
| 28. EPA green book nonattainment status, 2004-2006 | BRFSS Supplement file |
| 29. Number of dentists per 100k population in 1998 | BRFSS Supplement file |
| 30. Emergency room visits per 100k population in 2004 | BRFSS Supplement file |
| 31. Limited-service eating places per 100k population in 2005 | BRFSS Supplement file |
| 32. Fitness & recreation sports centers per 100k population in 2005 | BRFSS Supplement file |
| 33. Short term general hospital admissions per 100k population in 2004 | BRFSS Supplement file |
| 34. Short term general hospital beds per 100k population in 2004 | BRFSS Supplement file |
| 35. Short term general hospitals per 100k population in 2004 | BRFSS Supplement file |
| 36. Beer, wine & liquor stores per 100k population | BRFSS Supplement file |
| 37. General practice office based MDs per 100k population | BRFSS Supplement file |

[1] Multiple years are averaged

| | |
|---|---|
| 38. Medical specialist MDs per 100k population | BRFSS Supplement file |