Supplemental Materials: *A causal inference framework for cancer cluster investigations using publicly available data*

Rachel C. Nethery[1], Yue Yang[1], Anna J. Brown[2], Francesca Dominici[1]
[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health
[2]Department of Statistics, University of Chicago

1 Identifiability of the cSIR

Assume that the ignorability and causal consistency assumptions hold, as stated in the main manuscript. Consider the numerator of the SIR, $E\left[Y(T=1)|T=1\right]$. Note that

$$E\left[Y(T=1)|T=1\right] = E_{\boldsymbol{X}}\left[E\left[Y(T=1)|T=1,\boldsymbol{X}\right]\right] = E_{\boldsymbol{X}}\left[E\left[Y|T=1,\boldsymbol{X}\right]\right]$$

where the last equality holds by causal consistency. Similarly, for the denominator, $E\left[Y(T=0)|T=1\right] = E_{\boldsymbol{X}}\left[E\left[Y(T=0)|T=1,\boldsymbol{X}\right]\right]$. Now, we invoke the ignorability assumption, which states that T is independent of $Y(T=0)$ conditional on $\boldsymbol{X}$, so that
$E\left[Y(T=0)|T=1,\boldsymbol{X}\right] = E\left[Y(T=0)|T=0,\boldsymbol{X}\right]$. Thus, we have

$$E\left[Y(T=0)|T=1\right] = E_{\boldsymbol{X}}\left[E\left[Y(T=0)|T=0,\boldsymbol{X}\right]\right] = E_{\boldsymbol{X}}\left[E\left[Y|T=0,\boldsymbol{X}\right]\right]$$

by applying causal consistency as above. Thus, we see that both the numerator and denominator of the cSIR are identifiable and can be estimated with observed data.

2 Supplemental information for the simulations

Here we provide supporting information for the simulation studies presented in the main manuscript. Table 1 shows the parameter values used to generate the simulated data under each of the four simulation scenarios. Table 2 provides additional details of the simulation results for simulations 3 and 4 that were omitted from the main manuscript. Specifically, it shows the performance of the methods when the value of the true SIR used to generate the table is varied. In general, these results are consistent with the ones shown in the main manuscript, so we do not provide more commentary on those results here.

Figure 1 shows the rate of coverage of the null SIR value, 1, for the 95% credible/confidence interval for each method in simulations 3 and 4 as the true SIR value is increased from 1.1 to 2. The coverage of the null gives a sense of the power of our method to detect exposure effects. For cSIR, coverage of the null value decreases as the true SIR increases, as we would expect. We begin to see reasonable power to detect non-null SIRs when the true SIR is above 1.5, which means we are able to detect relatively small exposure effects. CDC's coverage of the null value is erratic and does not reflect trends in the true SIR, i.e., coverage of the null does not consistently decrease as the true SIR value increases. In the main manuscript, we discussed how PR's instability results in highly conservative coverage, and we see that this result holds not only for the true SIR but also for the null value.

Table 1: Parameter values used to simulate data. The values are chosen to reflect the anticipated magnitude and direction of the relationship between each variable and the exposure/outcome.

		Sim 1	Sim 2	Sim 3	Sim 4
Exposure Model	Intercept (γ_0)	-1.15	0	-1.15	0
	MoneyFood (γ_{11})	0	0.0009	0	0.0009
	MoneySmoke (γ_{12})	0	0.015	0	0.015
	P65+ (γ_{13})	0	0.003	0	0.003
	PMale (γ_{14})	0	-0.001	0	-0.001
	PWhite (γ_{15})	0	-0.01	0	-0.01
	Unemploy (γ_{16})	0	0.004	0	0.004
	Commute (γ_{17})	0	0.002	0	0.002
	Income (γ_{18})	0	-0.01	0	-0.01
Outcome Model	Intercept (α_0)	-5.99	-5	-5	-5
	Exposure (α_1)	0	0	$\{\log(1.1), \log(1.2),$ ..., $\log(2)\}$	$\{\log(1.1), \log(1.2),$ ..., $\log(2)\}$
	MoneyFood (α_{21})	0	0.007	0.007	0.007
	MoneySmoke (α_{22})	0	0.015	0.015	0.015
	P65+ (α_{23})	0	0.03	0.03	0.03
	PMale (α_{24})	0	-0.001	-0.001	-0.001
	PWhite (α_{25})	0	-0.02	-0.02	-0.02
	Unemploy (α_{26})	0	0.004	0.004	0.004
	Commute (α_{27})	0	0.002	0.002	0.002
	Income (α_{28})	0	-0.005	-0.005	-0.005

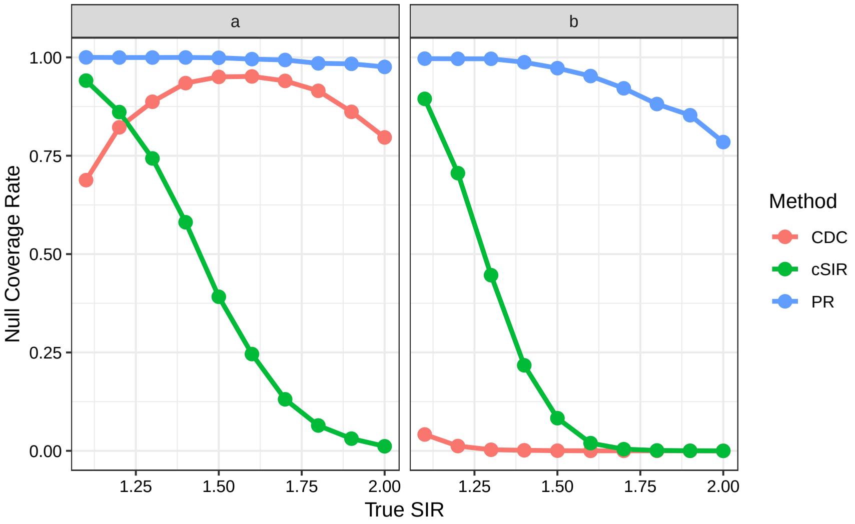

Figure 1: Trends in the rate of coverage of the null SIR as the true SIR increases in (a) simulation 3 and (b) simulation 4. This reflects the power of each method to detect exposure effects.

Table 2: Detailed results for simulations 3 and 4 comparing the proposed cSIR method with the standard cancer cluster SIR estimation method (CDC) and a similar Poisson regression approach (PR). Shown are the bias in the point estimate, the coverage rate of the true SIR for 95% confidence/credible intervals, and the width of the 95% confidence/credible intervals.

	True SIR	Method	Bias	Coverage True SIR	CI Width
Simulation 3	1.1	CDC	-0.33	0.43	0.54
		PR	-0.06	1.00	12.89
		cSIR	-0.03	0.95	0.73
	1.5	CDC	-0.54	0.09	0.58
		PR	1.23	1.00	49.08
		cSIR	-0.05	0.94	0.87
	2	CDC	-0.84	0.00	0.60
		PR	0.73	1.00	49.08
		cSIR	-0.07	0.93	1.02
Simulation 4	1.1	CDC	0.56	0.13	0.80
		PR	0.02	1.00	14.29
		CSIR	-0.01	0.95	0.53
	1.5	CDC	0.51	0.26	0.83
		PR	3.76	0.99	112.96
		CSIR	-0.01	0.94	0.62
	2	CDC	0.35	0.58	0.84
		PR	3.26	1.00	112.96
		CSIR	-0.02	0.94	0.72

3 Prediction model fitting and validation

We build the prediction models on a training set of kidney and bladder cancer incidence data from New York (NY). The training set includes census block group (CBG) cancer incidences for all counties in NY besides Broome county (where Endicott is located) and six counties selected to be withheld as a test set. The six counties in the test set were specifically chosen to reflect a wide range of of demographic features, so that we can assess whether our model performs better in areas with particular characteristics. Specifically, we selected two highly populous counties in the New York City area (Westchester and Orange Counties), two moderately populous counties representing smaller cities (Saratoga and Oswego Counties), and two sparsely populated counties (Allegany and Livingston Counties).

To the training data, we fit our multinomial regression (MR) prediction model as described in detail in the main text, i.e.

$$\log(\pi_j) = \boldsymbol{Z}_j'\boldsymbol{\beta} + \log(P_j) - \log(\sum_{l\in\psi(j)} e^{\boldsymbol{Z}_l'\boldsymbol{\beta}} P_l)$$

The following 11 CBG-level predictors ($\boldsymbol{Z}$) are used: percent of the population age 65+, percent of the population male, percent of the population white, rural indicator, percent of the adult population unemployed, average commute time, median household income, total dollars spent on smoking products as a portion of per capita income, percent of total dollars spent on food that was spent on food outside the home, percent of the population that reports exercising at least 2 times per week, and percent of the population working in the agriculture, mining, construction, or manufacturing industries. Independent $N(0, 1)$ prior distributions are placed on each component of $\boldsymbol{\beta}$. Although this prior choice may initially seem restrictive, realistically we would not expect a one-unit increase in any of the predictor variables to increase/decrease the incidence rate by more than a factor of $exp(2) = 7.4$, thus it seems reasonable to place low probability on $\boldsymbol{\beta}$ values larger than 2. Very flat (non-informative) prior distributions often cause stability problems in generalized linear models. Moreover, the large size of our training data ($12,427$ CBGs) ensures that the posterior distributions primarily reflect information in the data rather than the prior. 200,000 posterior samples of all parameters were obtained with the Markov Chain Monte Carlo sampler following 10,000 burn-in samples. Exponentiated $\boldsymbol{\beta}$ point estimates (posterior means) and 95% credible intervals are provided in Table 3. Traceplots of the posterior samples for all 11 elements of $\boldsymbol{\beta}$, shown in Figure 2 for the kidney cancer model and Figure 3 for the bladder cancer model, demonstrate good convergence. Metropolis sampler acceptance rates were 0.29 and 0.30 for the models.

We test the predictive accuracy of our model, and compare it to a more ad-hoc prediction approach, which we call a rescaled Poisson regression (PR). Recall that the purpose of our prediction model is to predict incidence at the CBG level, conditional on the observed county level incidence. For our MR model, we perform prediction for all CBGs in a county by collecting posterior predictive samples from a multinomial distribution with the observed county level incidence as the 'number of trials' parameter and the model-predicted probabilities for the CBGs in the county as the event probability parameters. The competing rescaled PR method instead fits a standard frequentist Poisson regression to the CBG incidences, obtains CBG level predictions from the model, and then rescales the predictions for all the CBGs within a county so that they sum to the observed county-level incidence. We compare the mean square error (MSE) of the predictions from these models as well as the proportion of observed cancer incidences falling in the 95% confidence/credible intervals for their predictions (Coverage). We compute these model fit metrics for the CBGs in the training set of NY data, the CBGs in the test set of the NY data, and separately for each of the individual

Table 3: Exponentiated point estimates and 95% credible intervals (CI) for each of the predictors in the kidney cancer and bladder cancer incidence prediction models.

	Kidney		Bladder	
Variable	Estimate	95% CI	Estimate	95% CI
MoneyFood	0.9543	0.9243, 0.9848	0.9805	0.9549, 1.0066
MoneySmoke	0.9762	0.8784, 1.0819	1.0177	0.9253, 1.1185
Rural	1.0970	1.0224, 1.1779	1.0775	1.0214, 1.1364
P65Plus	1.0278	1.0247, 1.0309	1.0379	1.0354, 1.0404
PMale	0.9927	0.9865, 0.9990	1.0044	0.9992, 1.0094
PWhite	1.0045	1.0035, 1.0055	1.0113	1.0104, 1.0122
Unemploy	0.9970	0.9934, 1.0006	0.9947	0.9915, 0.9978
Commute	1.0060	1.0029, 1.0091	1.0036	1.0010, 1.0062
Income	1.0011	0.9999, 1.0023	1.0028	1.0019, 1.0038
Industry	1.0009	0.9980, 1.0039	1.0001	0.9977, 1.0025
Exercise	0.9994	0.9983, 1.0004	0.9974	0.9961, 0.9985

counties in the test set. Moreover, we have also obtained (not publicly available) census tract level kidney and bladder cancer incidence data for the entire state of Idaho from the Cancer Data Registry of Idaho. Although census tracts have larger populations than CBGs, we can adjust for this through the population size offset term in the models. Thus, we also test the predictive power of the models on the Idaho cancer data to give a sense of their external validity.

The results are shown in Table 4 for kidney cancer and Table 5 for bladder cancer. While the MSEs are similar for the two models, our MR model always provides better coverage than the PR model. Our MR model's substantial improvements in uncertainty estimation for the predictions is critical, since these uncertainties are passed into the estimation stage model and reflected in the final effect estimates. For both models, we see similar predictive accuracy in the training and test sets. Based on the county-specific results from the test set, there are no clear trends in the prediction performance across county population sizes. Finally, the predictive performance for the census tract incidences in Idaho differs considerably between the kidney and bladder cancer models. We observe a notable increase in the MSE for the kidney cancer model in Idaho, while the bladder cancer MSE for Idaho is on par with the MSEs from the NY data. However, for both types of cancer, our MR model demonstrates strong coverage in the Idaho data. Because the uncertainties from the MR prediction models are reliable and are accounted for in the cSIR estimation models, we are not concerned about the modest decrease in predictive accuracy for the Idaho kidney cancer data.

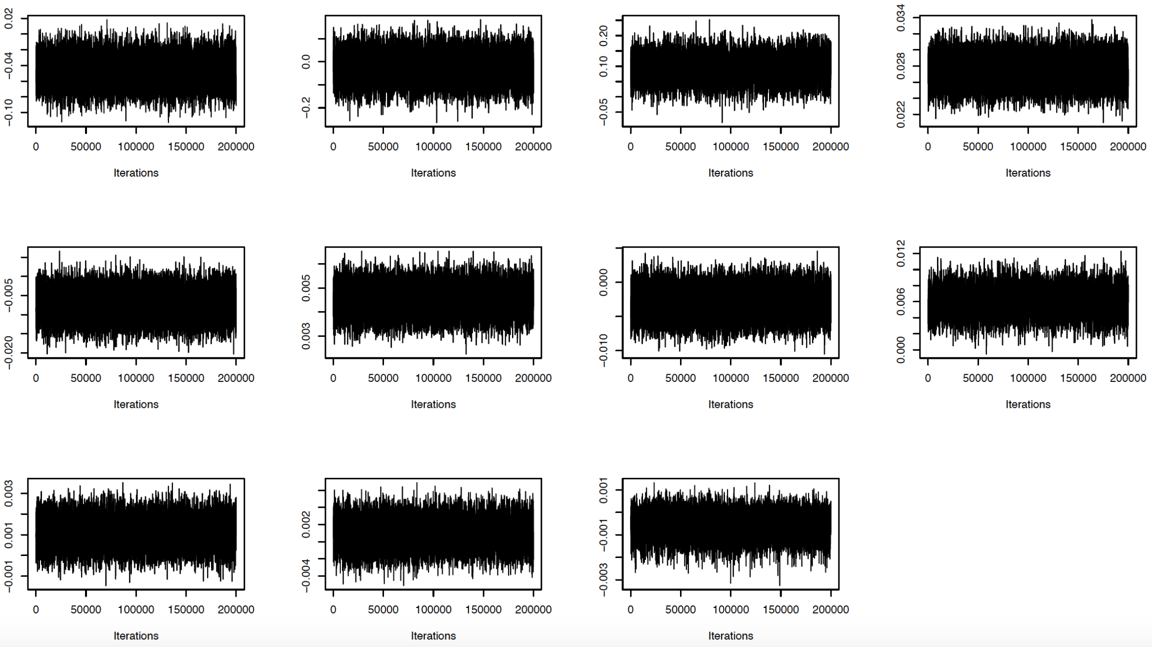

Figure 2: Traceplots for the kidney cancer prediction model.

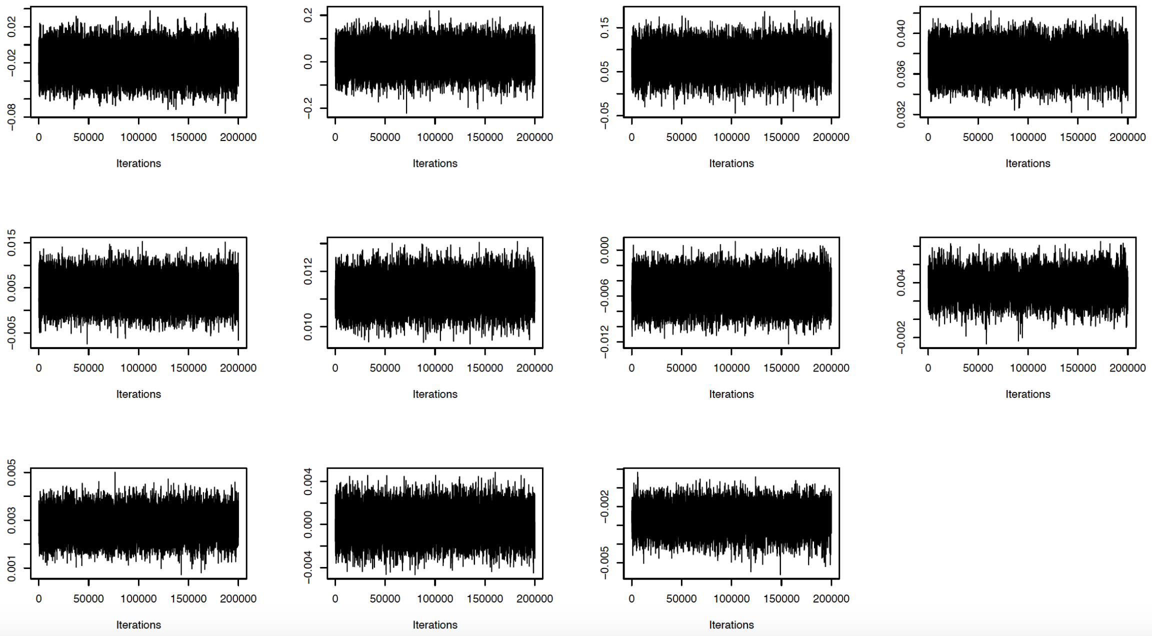

Figure 3: Traceplots for the bladder cancer prediction model.

Table 4: Kidney cancer incidence prediction model fit comparing our proposed Bayesian multinomial regression model (MR) and a frequentist Poisson regression model (PR). Model fit criteria shown are (1) mean square error (MSE) of predictions and (2) proportion of incidences falling in the 95% confidence/credible interval for their predictions (Coverage). These metrics are evaluated separately for CBGs in the training set from NY, CBGs in the test set from NY, and CBGs in the individual counties in the test set, and for census tract level cancer incidence data from the state of Idaho.

	Method	MSE	Coverage
Train	MR	1.32	0.58
Train	PR	1.32	0.04
Test	MR	1.54	0.62
Test	PR	1.56	0.05
Allegany	MR	0.69	0.65
Allegany	PR	0.69	0.07
Livingston	MR	1.08	0.72
Livingston	PR	1.08	0.02
Orange	MR	1.53	0.58
Orange	PR	1.52	0.07
Oswego	MR	2.26	0.62
Oswego	PR	2.28	0.03
Saratoga	MR	1.49	0.75
Saratoga	PR	1.49	0.01
Westchester	MR	1.56	0.60
Westchester	PR	1.59	0.05
Idaho	MR	4.12	0.82
Idaho	PR	4.11	0.09

Table 5: Bladder cancer incidence prediction model fit comparing our proposed Bayesian multinomial regression model (MR) and a frequentist Poisson regression model (PR). Model fit criteria shown are (1) mean square error (MSE) of predictions and (2) proportion of observed incidences falling in the 95% confidence/credible interval for their predictions (Coverage). These metrics are evaluated separately for CBGs in the training set from NY, CBGs in the test set from NY, and CBGs in the individual counties in the test set, and for census tract level cancer incidence data from the state of Idaho.

	Method	MSE	Coverage
Training set	MR	2.18	0.68
Training set	PR	2.18	0.03
Test set	MR	2.37	0.73
Test set	PR	2.38	0.04
Allegany	MR	1.76	0.85
Allegany	PR	1.77	0.00
Livingston	MR	2.35	0.77
Livingston	PR	2.38	0.05
Orange	MR	2.46	0.67
Orange	PR	2.48	0.03
Oswego	MR	2.50	0.81
Oswego	PR	2.49	0.04
Saratoga	MR	3.28	0.79
Saratoga	PR	3.28	0.03
Westchester	MR	2.17	0.73
Westchester	PR	2.19	0.05
Idaho	MR	2.55	0.74
Idaho	PR	2.52	0.06

4 cSIR estimation model fitting

Recall that the cSIR estimation model is fit to the matched data and has the following form:

$$\log(E\left[\tilde{Y}_i^{(b)}\right]) = \alpha_0 + T_i\alpha_1 + \boldsymbol{X}_i'\boldsymbol{\alpha}_2 + \log(P_i)$$

where $\tilde{Y}_i^{(b)}$ is the observed cancer incidence for the exposed units and NY controls and the b^{th} posterior predictive sample of Y_i from the prediction model for non-NY control units. For the analysis presented in the main manuscript (A-1), we placed independent $N(0, 0.5)$ prior distributions on α_0, α_1, and the components of $\boldsymbol{\alpha}_2$. Here we also provide the results from a sensitivity analysis (A-2) with $N(0, 1)$ priors.

We have 200,000 $\tilde{Y}_i^{(b)}$ posterior samples from the prediction model, and we iterate through these to collect 200,000 posterior samples of the estimation model parameters, 150,000 of which are discarded as burn-in and 50,000 are retained for estimation. We show here the results for the A-1 and A-2 models fit to the 5:1 propensity score (logistic regression with linear terms) matched data, the results from the other matched datasets are similar. The exponentiated point estimates (posterior means) and 95% credible intervals for the coefficients in the models are shown in Table 6 and Table 7 for A-1 and A-2 respectively. The differences in the results of A-1 and A-2 are minor, indicating little sensitivity to the variance of the prior distributions. The A-1 traceplots, shown in Figure 4 for kidney cancer and Figure 5 for bladder cancer, demonstrate good convergence. The A-1 Metropolis sampler acceptances rates are 0.26 and 0.30 for the models. The A-2 traceplots in Figure 6 (kidney) and Figure 7 (bladder) also show convergence, and the Metropolis acceptance rates are 0.35 and 0.39.

Table 6: Exponentiated point estimates and 95% credible intervals for each of the variables in the cSIR estimation models (A-1).

	Kidney		Bladder	
Variable	Estimate	95% CI	Estimate	95% CI
Intercept	1.277	0.225, 4.100	1.254	0.229, 4.020
MoneyFood	0.936	0.718, 1.201	0.847	0.654, 1.065
MoneySmoke	0.916	0.202, 2.674	1.306	0.264, 4.028
P65Plus	1.044	0.955, 1.132	1.033	0.967, 1.102
PMale	0.943	0.787, 1.132	1.034	0.877, 1.221
PWhite	0.980	0.917, 1.041	0.984	0.942, 1.046
Unemploy	1.030	0.953, 1.111	1.025	0.960, 1.101
Commute	1.023	0.926, 1.126	1.004	0.928, 1.086
Income	1.010	0.942, 1.084	1.041	0.983, 1.104
Industry	0.977	0.931, 1.017	0.973	0.940, 1.007
Exercise	1.013	0.942, 1.081	0.981	0.917, 1.038
Exposure	0.748	0.299, 1.497	1.571	0.886, 2.680

We also perform a sensitivity analysis in which we omit MoneyFood, MoneySmoke, and Exercise from the matching/adjustment set. The remaining variables represent the subset that can be easily obtained for small areas from the census or other publicly available sources. We use 5:1 propensity score matching from a logistic regression model with linear terms. We fit the cSIR estimation model using the same specifications as above (A-1 priors). For both the kidney and bladder cancer models, we observe only minor changes in the results using this subset of the confounders. The

Table 7: Exponentiated point estimates and 95% credible intervals for each of the variables in the cSIR estimation models (A-2).

	Kidney		Bladder	
Variable	Estimate	95% CI	Estimate	95% CI
Intercept	1.645	0.128, 7.101	1.587	0.121, 7.068
MoneyFood	0.915	0.685, 1.166	0.862	0.680, 1.094
MoneySmoke	1.006	0.108, 3.876	1.439	0.196, 5.072
P65Plus	1.044	0.961, 1.127	1.035	0.958, 1.113
PMale	0.961	0.799, 1.166	1.021	0.871, 1.192
PWhite	0.983	0.932, 1.038	0.987	0.938, 1.037
Unemploy	1.028	0.955, 1.096	1.019	0.949, 1.091
Commute	1.012	0.918, 1.105	1.000	0.920, 1.082
Income	1.009	0.940, 1.082	1.033	0.978, 1.092
Industry	0.982	0.942, 1.021	0.973	0.936, 1.010
Exercise	1.013	0.955, 1.076	0.980	0.929, 1.031
Exposure	0.699	0.220, 1.579	1.643	0.903, 2.726

kidney and bladder cancer cSIR estimates and 95% credible intervals are 0.75 (0.27, 1.57) and 1.52 (0.81, 2.64), respectively.

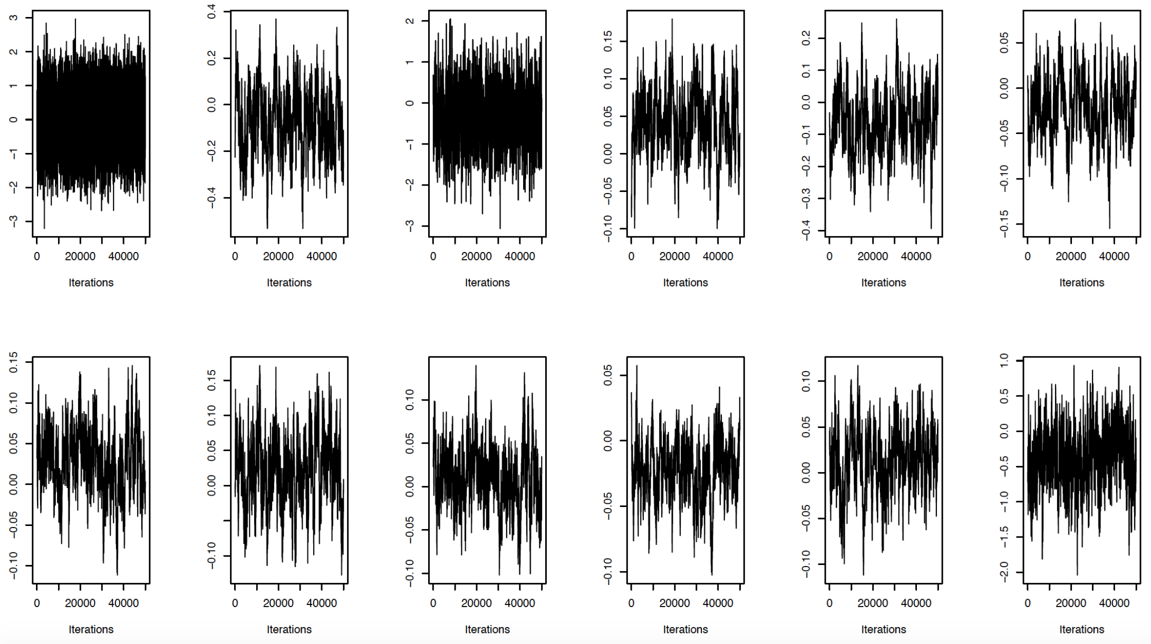

Figure 4: Traceplots for the kidney cancer cSIR estimation model (A-1).

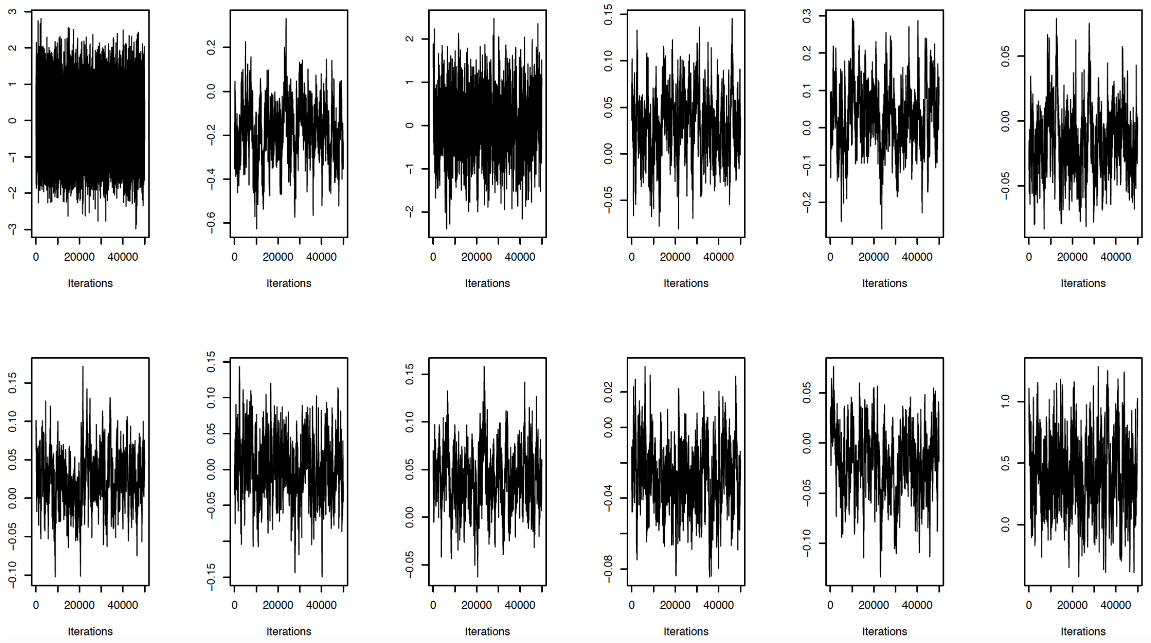

Figure 5: Traceplots for the bladder cancer cSIR estimation model (A-1).

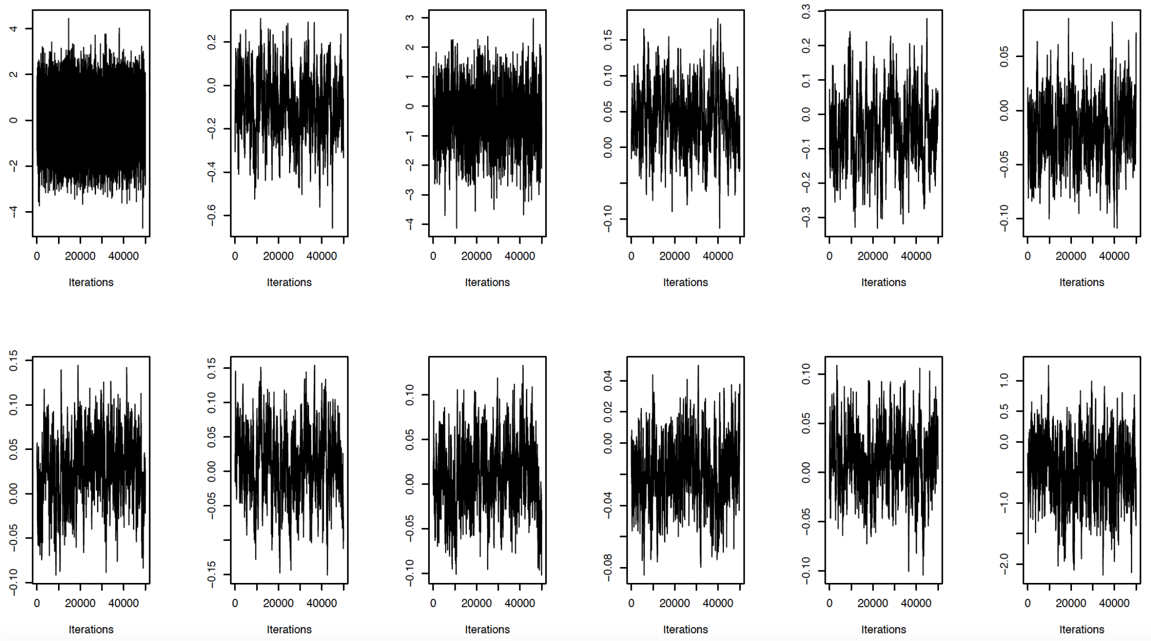

Figure 6: Traceplots for the kidney cancer cSIR estimation model (A-2).

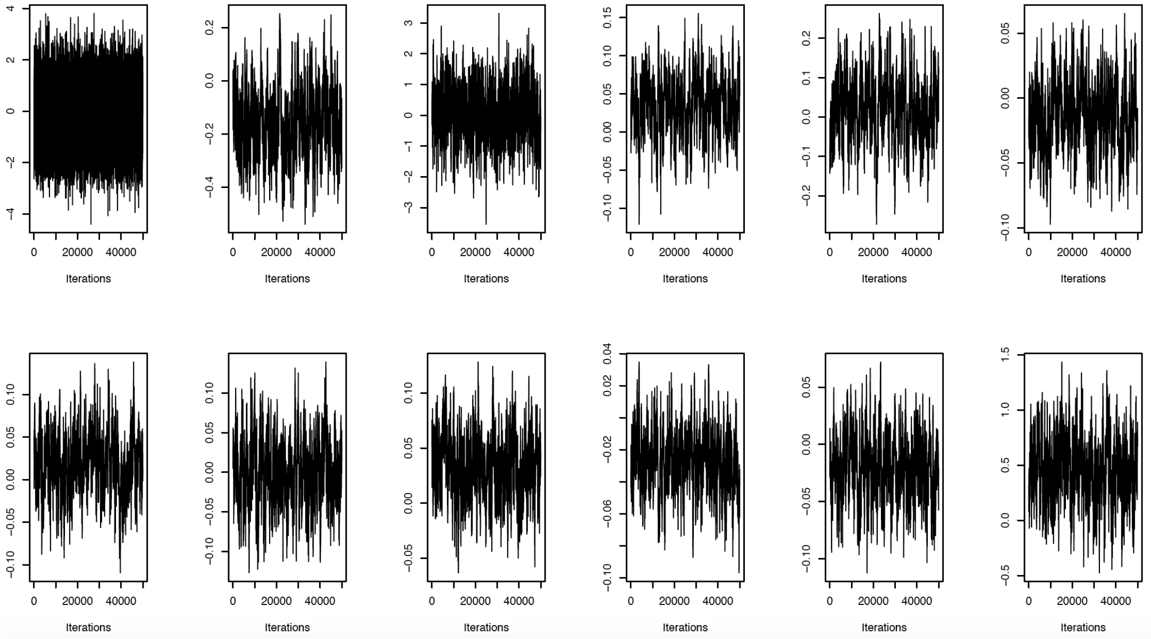

Figure 7: Traceplots for the bladder cancer cSIR estimation model (A-2).

5 Prediction Model Likelihood Details

Assume for simplicity that we have 3 counties, each containing 3 census block groups (CBGs), for a total of 9 CBGs in the data. Letting the CBGs be indexed by j, we assume that the cancer incidences in the CBGs follow $Y_j \sim Poisson(\lambda_j)$ for $j = 1, \ldots, 9$. Let CBGs $j = \{1,2,3\}$ be nested within county 1, CBGs $j = \{4,5,6\}$ be nested within county 2, and CBGs $j = \{7,8,9\}$ be nested within county 3.

Now we want the distribution of the CBG incidences conditional on the county incidence, i.e.,

$$Y_1, Y_2, Y_3 | \sum_{j=1}^{3} Y_j = B$$
$$Y_4, Y_5, Y_6 | \sum_{j=4}^{6} Y_j = C$$
$$Y_7, Y_8, Y_9 | \sum_{j=7}^{9} Y_j = D$$

We will first show that these conditional distributions are multinomial distributed. Since we know $\sum_{j=1}^{3} Y_j \sim Poisson(\sum_{j=1}^{3} \lambda_j)$, we have that

$$\begin{aligned} f\left(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3 \middle| \textstyle\sum_{j=1}^{3} Y_j = B\right) &= \frac{f\left(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3 \cap \sum_{j=1}^{3} Y_j = B\right)}{f\left(\sum_{j=1}^{3} Y_j = B\right)} \\ &= \frac{\prod_{j=1}^{3} \frac{e^{-\lambda_j} \lambda_j^{y_j}}{y_j!}}{\frac{e^{-\sum_{j=1}^{3} \lambda_j} \left(\sum_{j=1}^{3} \lambda_j\right)^B}{B!}} \\ &= \left(\frac{B!}{y_1!\, y_2!\, y_3!}\right) \frac{\prod_{j=1}^{3} \lambda_j^{y_j}}{\left(\sum_{j=1}^{3} \lambda_j\right)^B} \\ &= \left(\frac{B!}{y_1!\, y_2!\, y_3!}\right) \prod_{j=1}^{3} \frac{\lambda_j^{y_j}}{\left(\sum_{j=1}^{3} \lambda_j\right)^{y_j}} \\ &= B! \prod_{j=1}^{3} \frac{1}{y_j!} \left(\frac{\lambda_j}{\left(\sum_{j=1}^{3} \lambda_j\right)}\right)^{y_j} \end{aligned}$$

Letting $\pi_j = \frac{\lambda_j}{\left(\sum_{j=1}^{3} \lambda_j\right)}$ we see that $\sum_{j=1}^{3} \pi_j = 1$ and thus $f\left(Y_1, Y_2, Y_3 \middle| \sum_{j=1}^{3} Y_j = B\right)$ is a multinomial distribution with parameters $\{\pi_1, \pi_2, \pi_3\}$. Note that we can also write the distribution in the following way:

$$f\left(Y_1, Y_2, Y_3 \middle| \textstyle\sum_{j=1}^{3} Y_j = B\right) = B! \prod_{j=1}^{3} \frac{1}{y_j!} \pi_j^{y_j} = \prod_{j=1}^{3} \frac{(B!)^{1/3}}{y_j!} \pi_j^{y_j}$$

Now define $K_1 = K_2 = K_3 = B$, and $\psi(1) = \psi(2) = \psi(3) = \{1,2,3\}$, i.e., the $\psi(j)$ are sets containing the indices of all the CBGs in same county as CBG j. Using these, we can write

$$f\left(Y_1, Y_2, Y_3 \middle| \sum_{j=1}^{3} Y_j = B\right) = \prod_{j=1}^{3} \frac{(K_j!)^{1/||\psi(j)||}}{y_j!} \pi_j^{y_j}$$

The same results hold for $f(Y_4, Y_5, Y_6 | \sum_{j=4}^{6} Y_j = C)$ and $f(Y_7, Y_8, Y_9 | \sum_{j=7}^{9} Y_j = D)$.

It is trivial to show that $f\left(Y_1, Y_2, Y_3 \middle| \sum_{j=1}^{3} Y_j = B\right)$, $f(Y_4, Y_5, Y_6 | \sum_{j=4}^{6} Y_j = C)$ and $f(Y_7, Y_8, Y_9 | \sum_{j=7}^{9} Y_j = D)$ are independent when the individual Y_j are independent. Thus, the likelihood of all the data can be written as:

$$f\left(Y_1, Y_2, Y_3 \middle| \sum_{j=1}^{3} Y_j\right) \times f\left(Y_4, Y_5, Y_6 \middle| \sum_{j=4}^{6} Y_j\right) \times f\left(Y_7, Y_8, Y_9 \middle| \sum_{j=7}^{9} Y_j\right) = \prod_{j=1}^{9} \frac{(K_j!)^{1/||\psi(j)||}}{y_j!} \pi_j^{y_j}$$

This corresponds to the form of the likelihood appearing in the main manuscript on page 13. We have chosen to use the K_j and $\psi(j)$ notation for convenience in the scenario with differing numbers of CBGs within each county.