

# Supplemental Information for *Effect and Pathways Modifiers in a Bayesian Pathways Analysis of the National Human Exposure Assessment Survey for Arsenic in EPA Region 5*

## Definitions of Input Variables and Fitted Bayesian Pathways Models

This document specifies the variables and results from all Bayesian Pathways Models fitted to the NHEXAS data in the paper *Effect and Pathways Modifiers in a Bayesian Pathways Analysis of the National Human Exposure Assessment Survey for Arsenic in EPA Region 5*. It also lists the full-conditional distributions of the model parameters used by our MCMC algorithm.

### 1 Media and Pathway Modifier Variables

See Tables 1 and 2.

### 2 Full Conditional Distributions of the Hierarchical Bayesian Sub-population Model

#### 2.1 Posterior distribution

Letting  $\theta = (\beta, \tau, \omega)$ , the posterior distribution is

$$\begin{aligned} [\theta, \mathbf{X} | \mathbf{Y}] &\propto [\mathbf{Y} | \mathbf{X}, \theta] [\mathbf{X} | \theta] [\theta] \\ &= [\mathbf{Y} | \mathbf{X}, \omega] [\mathbf{X} | \beta, \tau] [\beta] [\omega] [\tau], \end{aligned}$$

assuming independence of the three components within  $\theta$ . The distribution is not available in closed form and so we use a Gibbs sampler Markov chain Monte Carlo (MCMC), based on an augmented (including censored and missing data) joint distribution to sample from this posterior distribution.

Variable	Type	Definition and Units
UR	M	natural log of arsenic concentration in Urine ( $\mu g/\ell$ )
FD	M	natural log of arsenic intake in Food ( $\mu g/day$ )
BV	M	natural log of arsenic intake in Beverage ( $\mu g/day$ )
IA	M	natural log of arsenic concentration in Indoor Air ( $ng/m^3$ )
OA	M	natural log of arsenic concentration in Outdoor Air ( $ng/m^3$ )
PA	M	natural log of arsenic concentration in Personal Air ( $ng/m^3$ )
SL	M	natural log of arsenic concentration in Soil ( $ng/g$ )
SD	M	natural log of arsenic loading in Surface (Sill) Dust ( $\mu g/cm^2$ )
WF	m	natural log of arsenic concentration in Tap Water, Flushed ( $\mu g/\ell$ )
logcrtst	C	natural log of creatinine concentration ( $\mu g/\ell$ )
<i>mlogcrtn</i>	C	negative of the natural log of creatinine concentration ( $\mu g/\ell$ )
<i>logcreatinine</i>	C	the natural log of creatinine concentration ( $\mu g/\ell$ )
<i>male</i>	C, E	binary (0/1) variable indicating the participant is male
<i>female</i>	E	binary (0/1) variable indicating the participant is female
<i>child4to11/Age1</i>	C, E	binary (0/1) variable indicating the participant is between 4 and 11 years old
<i>age</i>	C, E	participant's age (years)
<i>AgeAtoB</i>	C, E	binary (0/1) variable indicating the participant's age is between <i>A</i> and <i>B</i> years old
<i>Age2</i>	C,E	binary (0/1) variable indicating the participant's age is between 12 and 19 years
<i>hhold</i>	E, P	household size
<i>log hhold</i>	E, P	natural log of household size
<i>minushhold</i>	E, P	negative of household size
<i>hrs.home</i>	E	average number of hours participant spends inside home/day
<i>glass</i>	E	average number of glasses of water participant drinks/day
<i>tap.c</i>	P	binary (0/1) variable indicating participant cooks with tapwater
<i>fruit</i>	P	binary (0/1) variable indicating participant eats fruit at least 3 days/three-months
<i>tap.d/tapwater</i>	E	binary (0/1) variable indicating participant drinks tapwater
<i>c.air</i>	E	binary (0/1) variable indicating participant's home has central air
<i>no.c.air</i>	E	binary (0/1) variable indicating participant's home does <i>not</i> have central air
<i>gasequip</i>	P	binary (0/1) variable indicating participant used gas equipment in the past week

Table 1: For each variable used in the Bayesian hierarchical analysis, the variable symbol, its use—depending on the fitted model (M = media, C = component of the creatinine adjustment, E = effect modifier, and P = pathway modifier), and its definition, including units.

Variable	Type	Definition and Units
<i>tobacco/tobacco1</i>	P	binary (0/1) variable indicating the participant is a current smoker
<i>tobacco2</i>	P	binary (0/1) variable indicating the participant is a former smoker
<i>workshop</i>	P	average number of minutes/day in an enclosed workshop
<i>work20</i>	P	binary (0/1) variable indicating participant works $\geq 20$ min./day in a workshop
<i>fish</i>	P	binary (0/1) variable indicating participant eats fish
<i>new.fish</i>	P	binary (0/1) variable indicating participant eats fish more than once/month
<i>no.fish</i>	P	binary (0/1) variable indicating participant eats fish less than once/month
<i>logBMI</i>	C	Natural log of Body Mass Index ( $kg/m^2$ )
<i>race3_2</i>	C, E	a binary(0/1) variable indicating whether the participant is African-American
<i>race3_3</i>	C, E	binary(0/1) variable indicating the participant's race is <i>not</i> Caucasian/African-American
<i>no.wellwater</i>	P	binary(0/1) variable indicating the source of running water is well water
<i>wellwater</i>	P	binary(0/1) variable indicating the source of running water is <i>not</i> well water

Table 2: For each variable used in the Bayesian hierarchical analysis, the variable symbol, its use depending on the fitted model (M = media, C = component of the creatinine adjustment, E = effect modifier, and P = pathway modifier), and its definition, including units.

## 2.2 Full conditional distributions

1. For each  $i$  and  $j$  we set/draw augmented data,  $Y_{ij}^*$ , as follows:

$$Y_{ij}^* \begin{cases} = Y_{ij}, & Z_{ij} = 0; \\ \sim \text{Truncated } N(X_{ij}, 1/\omega_j) \text{ on } (-\infty, M_{ij}], & Z_{ij} = 1; \\ \sim N(X_{ij}, 1/\omega_j), & Z_{ij} = 2. \end{cases}$$

2. For the process variable  $\mathbf{X}$ , we sample each process variable  $\mathbf{X}_j$ ,  $j = 1 \dots, J$ , conditional on the other process variables  $\{\mathbf{X}_k : k \neq j\}$  and the other parameters in the model. We have that

$$\begin{aligned} & [\mathbf{X}_j | \{\mathbf{X}_k : k \neq j\}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Y}^*, \boldsymbol{\omega}] \\ & \propto [\mathbf{Y}_j^* | \mathbf{X}_j, \boldsymbol{\omega}] [\mathbf{X}_j | \{\mathbf{X}_k : k \neq j\}, \boldsymbol{\beta}_j, \boldsymbol{\tau}_j] \prod_{l \in ch_j} [\mathbf{X}_l | \{\mathbf{X}_k : k \neq l\}, \boldsymbol{\beta}_l, \boldsymbol{\tau}_l], \end{aligned}$$

where  $ch_j$  denotes the columns of the process variable  $\mathbf{X}$  that depend on  $\mathbf{X}_j$  in their definition of the conditional mean  $\mu_{ij}$  (i.e., those  $l$  for which  $s_{jk}^X = j$  for some  $k$  or  $s_{lm}^{EX} = j$  for some  $m$ ).

The above expression is equal to

$$\prod_{i=1}^I \left( [Y_{ij}^* | X_{ij}, \omega_j] [X_{ij} | \{X_{ik} : k \neq j\}, \beta_j, \tau_j] \prod_{l \in ch_j} [X_{il} | \{X_{ik} : k \neq l\}, \beta_l, \tau_l] \right),$$

from which we can see that we can sample each  $X_{ij}$ ,  $i = 1, \dots, I$ , independently. Now

$$[Y_{ij}^* | X_{ij}, \omega_j] \propto \exp\left(\frac{\omega_j}{2}(Y_{ij}^* - X_{ij})^2\right),$$

$$[X_{ij} | \{X_{ik} : k \neq j\}, \beta_j, \tau_j] \propto \exp\left(\frac{\tau_j}{2}(X_{ij} - \mu_{ij})^2\right),$$

and for each  $l \in ch_j$

$$[X_{il} | \{X_{ik} : k \neq l\}, \beta_l] \propto \exp\left(\frac{\tau_l}{2}(X_{il} - \delta_{il}(j) - \epsilon_{il}(j))^2\right). \quad (1)$$

In (1) we define

$$\delta_{ij}(o) = \alpha_j + \sum_{\{k:s_{jk}^X \neq o\}} \beta_{jk}^X X_{is_{jk}^X} + \sum_l \beta_{jl}^P P_{is_{jl}^P} + \sum_m \beta_{jm}^E E_{is_{jm}^E} + \sum_{\{n:s_{jn}^{EX} \neq o\}} \beta_{jn}^{EX} E_{is_{jn}^E} X_{is_{jn}^{EX}}$$

to be the conditional mean of  $X_{ij}$  minus the regression terms that depend on  $X_{io}$ , and

$$\begin{aligned} \epsilon_{ij}(o) &= \sum_{\{k:s_{jk}^X = o\}} \beta_{jk}^X X_{is_{jk}^X} + \sum_{\{m:s_{jm}^{EX} = o\}} \beta_{jm}^{EX} E_{is_{jm}^E} X_{is_{jm}^{EX}} \\ &= X_{io} \left[ \sum_{\{k:s_{jk}^X = o\}} \beta_{jk}^X + \sum_{\{m:s_{jm}^{EX} = o\}} \beta_{jm}^{EX} E_{is_{jm}^E} \right], \end{aligned}$$

to be the regression terms that depend on  $X_{io}$ . Always,  $\delta_{ij}(o) + \epsilon_{ij}(o) = \mu_{ij}$ .

Rearranging the terms of  $X_{ij}$ , the distribution of  $X_{ij}$  conditional on all other terms is  $N(q/p, 1/p)$

where

$$q = \omega_j Y_{ij}^* + \tau_j \mu_{ij} + \sum_{l \in ch_j} \tau_l (X_{il} - \delta_{il}(j)) \left( \sum_{\{k:s_{jk}^X = n\}} \beta_{jk}^X + \sum_{\{m:s_{jm}^{EX} = n\}} \beta_{jm}^{EX} E_{is_{jm}^E} \right)$$

and

$$p = \omega_j + \tau_j + \sum_{l \in ch_j} \tau_l \left( \sum_{\{k:s_{jk}^X = n\}} \beta_{jk}^X + \sum_{\{m:s_{jm}^{EX} = n\}} \beta_{jm}^{EX} E_{is_{jm}^E} \right)^2.$$

3. For the regression parameters  $\beta$ , we sample each  $\beta_j$  conditional on the other parameters in the model. Letting

$$\mathbf{A}_j = \begin{bmatrix} 1 & \{X_{1s^X_{jk}}\} & \{P_{1s^P_{jk}}\} & \{E_{1s^E_{jk}}\} & \{E_{1s^E_{jk}} X_{1s^{EX}_{jk}}\} \\ 1 & \{X_{2s^X_{jk}}\} & \{P_{2s^P_{jk}}\} & \{E_{2s^E_{jk}}\} & \{E_{2s^E_{jk}} X_{2s^{EX}_{jk}}\} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \{X_{Is^X_{jk}}\} & \{P_{Is^P_{jk}}\} & \{E_{Is^E_{jk}}\} & \{E_{Is^E_{jk}} X_{Is^{EX}_{jk}}\} \end{bmatrix},$$

the distribution of  $\beta_j$  conditional on the other parameters is

$$\text{MVN}_{r_j}(\mathbf{R}_j^{-1} \mathbf{q}_j, \mathbf{R}_j^{-1}),$$

where  $\mathbf{R}_j = \tau_j \mathbf{A}_j^T \mathbf{A}_j + \Sigma_j^{-1}$  and  $\mathbf{q}_j = \tau_j \mathbf{A}_j^T \mathbf{X}_j + \Sigma_j^{-1} \mathbf{m}_j$ .

4. For each  $j$ , the conditional distribution of the process precision  $\tau_j$ , given the other parameters, is

$$\text{Ga}\left(c_j + I/2, d_j + \sum_{i=1}^N (X_{ij} - \mu_{ij})^2/2\right),$$

where  $\mu_{ij}$  was defined above in Equation (1) of the paper.