# Bayesian models for elevated disease risk due to exposure to uranium mine and mill waste on the Navajo Nation

<table>
<thead>
<tr>
<th>Journal:</th>
<th><em>Journal of the Royal Statistical Society</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>JRSS-OA-SC-Dec-11-0218</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Original Article - Series C</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>01-Dec-2011</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Stark, Glenn; University of New Mexico, Department of Mathematics and Statistics  
Huerta, Gabriel; Indiana University, Department of Statistics  
Henio-Adeky, Sarah; Southwest Research and Information Center,  
Nez, Teddy; Southwest Research and Information Center,  
Ramone, Sandy; Southwest Research and Information Center,  
Shuey, Chris; Southwest Research and Information Center,  
Cajero, Miranda; University of New Mexico Health Sciences Center, Community Environmental Health Program, College of Pharmacy  
DeGroat, Jeremy; University of New Mexico Health Sciences Center, Community Environmental Health Program, College of Pharmacy  
Downs, Mallery; University of New Mexico Health Sciences Center, Community Environmental Health Program, College of Pharmacy  
Lewis, Johnyee; University of New Mexico Health Sciences Center, College of Pharmacy, Community Environmental Health Program |
| Keywords: | abandoned uranium mines, conditionally specified logistic regression, hypertension, kidney disease, Native American |
Bayesian models for elevated disease risk due to exposure to uranium mine and mill waste on the Navajo Nation

Glenn A. Stark†
University of New Mexico, Albuquerque, USA
Gabriel Huerta
Indiana University, Bloomington, USA
Sarah Henio-Adeky, Teddy Nez, Sandy Ramone and Chris Shuey
Southwest Research and Information Center, Albuquerque, USA
Miranda Cajero, Jeremy DeGroat, Mallery Downs, and Johnnye L. Lewis
University of New Mexico Health Sciences Center, Albuquerque, USA

Summary. Many residents of the Navajo Nation live near to abandoned, unremediated uranium mines, mills, and waste piles. In response to concerns that exposure may increase health risks, and as part of the first comprehensive community health study of the effects of this waste, the Dine Network for Environmental Health (DiNEH) surveyed residents, collecting health, demographic, and behavioral data. Surrogate measures for exposure to waste differentiate between 1) exposures to uranium miners and their families during the active mining period and 2) continuing exposures to all community members, arising from the environmental legacy of uranium mining. Bayesian model averaging is applied to logistic regression and to conditionally specified logistic regression models of the prevalence of kidney disease, diabetes, and hypertension. In addition to accepted risks for these diseases, active-mining exposures are associated with elevated rates of kidney disease, while exposures to legacy waste are associated with elevated rates of hypertension.

Keywords: abandoned uranium mines, conditionally specified logistic regression, hypertension, kidney disease, Native American

1. Introduction and Background

The Navajo Nation is the largest Native-American governed territory in the United States. It covers 27,000 square miles in Arizona, Utah, and New Mexico, and it is home to more than 180,000 people (US Census, 2000). There are 520 abandoned uranium mines and 5 abandoned uranium processing mills on the Navajo Nation, as shown by the map in Figure 1 (USEPA, 2007, 2008a). Within the areas occupied by these abandoned mines and mills, the United States Environmental Protection Agency (USEPA) has identified 1100 features that include portals, prospects, rim strips, open pits, vertical shafts,

†Address for correspondence: Glenn A. Stark, Department of Mathematics and Statistics, MSC01 1115, 1 University of New Mexico, Albuquerque, New Mexico 87131-0001. E-mail: gstark@unm.edu.
processing facilities, and waste piles, ranging in size from less than 1 to more than 100 acres. This is the environmental legacy of a uranium mining boom on the Navajo Nation which began in 1948 when the United States Atomic Energy Commission announced that it would purchase all uranium ore mined in the United States at a guaranteed price. The United States government remained the sole purchaser of uranium until 1971. As the demand for uranium declined and mines ceased operation, hundreds of mines were abandoned and left open with no warnings about the dangers they posed as potential sources of air, soil, and groundwater contamination. The uranium mining boom ended when the last uranium mine on the Navajo Nation ceased operation in 1986 (Brugge and Goble, 2002).

Uranium mine and mill wastes contain elevated concentrations of not only uranium but many other heavy metals and radioactive constituents, as well as industrial chemicals including strong acids or bases and organic solvents introduced during milling operations (USEPA, 2008b). Uranium itself is both a radioactive element and a heavy metal with well-documented toxicity to the kidney (Leggett, 1989). The other components in the wastes present additional health risks either individually or acting in concert (ATSDR, 2011; USEPA, 2007). Environmental releases of these toxic and radioactive contaminants have been documented at mining and milling sites on and near the Navajo Nation (USEPA, 2007, 2008a).

There were many pathways by which residents may have been exposed to contamination during the active mining period as well as pathways that continue for exposures to the unremediated wastes. The term “active-mining related exposures” describes exposures experienced by miners, mill-workers, or their family members including occupational exposures to uranium yellowcake ore and dust, washing clothes of family members who were exposed workers, or secondary exposures of family members while living in a mining camp. The testimony to the United States Congress of a second-generation Navajo uranium miner described his experience living in a uranium mining camp:

I drank uranium-contaminated water from those mines. We washed our clothes in uranium-contaminated water. I watched children going into the mines and playing on the waste piles. We made our coffee with the uranium-contaminated water. In all likelihood I’ve continued to drink uranium-contaminated water through the years (Harrison, 2007).

While there are no longer active mining camps, many residents of the Navajo Nation continue to live in close proximity to abandoned uranium mines or other waste sources that have remained unremediated for decades. The term “environmental legacy exposures” describes continuing exposures to these residents which include inhalation of windblown dust, ingestion and bathing with contaminated groundwater, and consumption of potentially contaminated animals and plants. Because these sites are unmarked and unfenced, residents may also walk or ride horses across them, herd livestock, or collect plant material from these sites. Children growing up near abandoned uranium mines sometimes play in mines and on mine waste piles. In the past, open pit mines collected rainwater, which was sometimes used for drinking, watering livestock, or swimming. The foundations, floors, and walls of some homes were built out of mill tailings and chunks of uranium ore, which remain as current exposure sources (USEPA, 2008a). Active-mining and environmental legacy exposures are likely to have different routes of exposure, doses,
frequencies, and durations, suggesting the possibility that different health risks would be associated with each.

Prior to 2004, no comprehensive epidemiological study had assessed the population health impact of uranium mine and mill waste contamination in this Native American community with a high existing prevalence of kidney disease, diabetes, and hypertension. The possibility that the existing health status and local land use patterns might increase sensitivity, exposure, or both remained a major concern in the communities. Outside of the Navajo Nation, several previous studies have suggested sub-clinical nephrotoxic effects of chronic exposures to low-level, environmental concentrations of uranium and other metals in drinking water (Kurttio et al., 2002; Limson-Zamora et al., 1998; Mao et al., 1995; Wagner et al., 2010).

The Diné Network for Environmental Health (DiNEH) Project was developed in response to concerns raised by Navajo community leaders and health providers about whether chronic environmental exposures to uranium and other metals contained in the abandoned, unremediated waste sites may be contributing to the elevated kidney disease rates observed in their communities. The DiNEH team designed a collaborative research approach with the communities and providers to assess the health risks of these community-level exposures within the context of other known risk factors for kidney disease and related conditions. The primary tool of the initial phase of this assessment, conducted from 2004 to 2010, was a water and land-use survey developed with the communities that generated data on health, demographics, and exposure-related behaviors.

This paper presents the initial results of this first large-scale study to examine the health impact of the environmental legacy of 50 years of uranium mining on this population. Section 2 describes the data and proposes surrogate measures for exposure to uranium mine and mill waste. Section 3 analyzes logistic regression models for kidney disease, diabetes, and hypertension. Section 4 discusses a multivariate conditionally specified logistic regression model. Results are used to develop maps illustrating the spatial distribution of risks associated with environmental legacy exposures, and the rate of increase in risk associated with increases in the number of exposure activities engaged in within the waste areas — a potential surrogate for dose from current and historic exposures. Section 5 compares the analysis of Section 4 with the Bayesian t-link model proposed by O’Brien and Dunson (2004). Section 6 discusses the relevance of current findings and proposes directions for the analysis of ongoing studies.

2. Data

The DiNEH Project, as a community-based participatory research collaborative, reflects researchers working closely with community partners in all phases of the research to ensure the resulting project is both scientifically valid, community-appropriate, and conducted in a manner that is respectful of culture and tradition in this Native American population. All results are communicated to participating and affected communities, clinical-care providers, and policy-makers to inform decisions and reduce risk (deLemos et al., 2007). This collaborative process resulted in several outcomes that increase the team’s confidence in the data: 1) Questions were worded to maximize willing and accurate responses; 2) Interviews were conducted by community members with extensive training in survey administration and data quality, as well as in Navajo language and
culture. Regular and repeated reviews with research staff and Navajo cultural and language experts were held to ensure consistent translation of concepts, allowing interviews to be conducted in English, Navajo, or both languages as suited the participant while still ensuring consistency in the results; and 3) Frequent and informative communication with community members on both the progress and results of the research maintained an atmosphere of trust in the community. The project worked with the following 20 of the 110 Navajo Nation chapters (local governance units), all located in the Eastern Agency: Baca/Prewitt, Becenti, Casamero Lake, Coyote Canyon, Crownpoint, Iyanbito, Lake Valley, Littlewater, Mariano Lake, Nahodishgish, Ojo Encino, Pinedale, Pueblo Pintado, Smith Lake, Standing Rock, Thoreau, Torreon/Star Lake, White Horse Lake, and White Rock. Chapters varied on their history of mining with the number of mines ranging from 0 to more than 30. A commitment was made to the chapters to survey equal numbers of participants in each chapter, based upon the chapter in which a participant voted, with some adjustment due to varied populations and response rates among chapters. The geographical distribution of the participants was reviewed throughout data collection to ensure respondents were representative of mining and non-mining areas within the chapters.

The health, water, and land-use survey conducted by the DiNEH Project was used to collect data from 1,304 participants who were 18 years of age or older and who volunteered to be interviewed about their demographics, water use, and medical histories, as well as their possible exposures to uranium mining and milling waste in both community and occupational settings. Survey participants were recruited at chapter meetings, through radio announcements, at public events, and by word of mouth. The project began with survey information (reported here) and followed with environmental sampling, medical record reviews and biological monitoring still in progress. Although analysis of these later phases continues, early efforts to assess the validity of the data reported here substantiate their validity based on comparisons of summary data for disease prevalence with published Indian Health Services (IHS) prevalence, and initial comparison of self-reported disease with medical record diagnoses.

2.1. Health data

The current analysis focuses on the primary research question, the contribution of uranium exposure to elevated prevalence of kidney disease, and also looks at two related health outcomes, diabetes and hypertension, both of which are also found at elevated prevalence in the population and can be related physiologically. Preliminary analyses of our survey data suggested that the prevalences of kidney disease, diabetes, and hypertension were elevated among participants living in the proximity of abandoned uranium mines and mills (deLemos, 2008; Shuey et al., 2010). Modeling presented here further explores those relationships.

The health data that we analyze in this paper are self-reported. Survey participants were asked if they “have now or ever had kidney disease,” “diabetes,” and “high blood pressure,” as well as a range of other health endpoints. Follow-up probes asked participants if they were currently on dialysis or if they had ever had a kidney transplant. Participants were also asked if they had family histories of kidney disease, diabetes, or hypertension. Additional survey data that are used as covariates in models of disease...
prevalence in this paper include gender, age, and body mass index (BMI).

The BMI measurements analyzed in this paper were calculated from the self-reported height and weight of each participant. We have compared these measurements with BMI measurements based upon technician-measured data for 195 survey participants who volunteered to provide blood and urine in a later phase of the study, between 6 months and 5 years after their initial interview. The correlation between self-reported and technician-measured BMI for these 195 respondents was 0.82. As a point of comparison, McAdams, Van Dam, and Hu (2007) conducted a survey of 10,639 participants to evaluate the validity of BMI based on self-reported data. McAdams et al. (2007) reported correlations ranging from 0.95 for whites to 0.90 for Mexican Americans between self-reported and technician-measured BMI values collected concurrently and concluded that the “accuracy of self-reported BMI is sufficient for epidemiological studies using disease biomarkers” (page 188). Given the time lapse in our comparison, we feel our results are consistent with these conclusions.

2.2. Geospatial data

In this analysis, we used two sources of geospatial data: the locations of the homes of the survey participants and the locations of the abandoned uranium mines and mills within the study area. When a survey was administered in a participant’s home, the location of the home was measured using a hand-held Global Positioning System (GPS) instrument. When the participant was interviewed elsewhere, a GPS measurement was taken later, or the geographic coordinates were identified using the McKinley County or Cibola County Rural Addressing System. The locations of the abandoned uranium mines and mills in the study area were obtained from the U.S. Army Corps of Engineers documentation compiled for the USEPA (USEPA, 2007).

Uranium mining and milling on the Navajo Nation concentrated in four areas: Cameron-Tuba City, Arizona in the southwest; Monument Valley, Arizona and Utah in the north; Red Valley, Arizona-Shiprock, New Mexico in the north-central; and the Navajo portion of the Grants Mineral Belt of New Mexico in the southeast. The DiNEH study area is in and near the Grants Mineral Belt of New Mexico and includes 20 chapters with a range of 0 to more than 30 mines per chapter. Ninety-eight abandoned uranium mines and two abandoned uranium mill processing facilities are located within or within four miles of the boundaries of this study area.

2.3. Exposure surrogates

Because we are interested in life-span exposures to the mixture of contaminants associated with these abandoned sites, current environmental and biomonitoring data are insufficient for exposure quantification. In the absence of direct exposure data, we propose to determine the relative contribution of exposure to environmental toxicants associated with abandoned uranium mine waste by defining surrogate measures for exposure. Survey participants were asked a battery of questions about the occupational and environmental conditions in which they lived and ways in which they might have come into contact with uranium throughout their life. Exposures occurring during the period of active mining are likely to differ substantially from environmental legacy exposures. Legacy exposures are likely to be of a longer duration but of lower concentrations of
contaminants. In addition, the routes of exposure may differ between the two exposure classes. We use answers reported by each survey participant to define a surrogate measure for active-mining related exposures and we combine these responses with a measure of the distance from a participants home to the 100 mine and mill waste sites in the study area to define a surrogate measure for environmental legacy exposures.

We define the active-mining related exposure surrogate ($M$) to be the total number of five possible active mining-related exposures reported by each participant. When asked if they had experienced certain exposures to uranium mine or mill waste: 

\begin{itemize}
  \item [i)] 22\% of survey participants reported that they had washed or handled the clothes of a uranium worker;
  \item [ii)] 10\% had worked in a uranium mine;
  \item [iii)] 4\% had lived in a mining camp;
  \item [iv)] 2\% had worked in a uranium mill; and
  \item [v)] 2\% had worked on a uranium mine or mill reclamation or had hauled uranium ore or tailings in a pickup truck.
\end{itemize}

The distribution of $M$ is shown in the first row of Table 1.

Environmental exposure surrogates are frequently modeled with distance-based functions that have large values for residents who live near to a focus of environmental exposure and that have increasingly smaller values for residents who live farther away. One function that measures this relationship is the inverse-distance measure

$$g(i, m) = \frac{1}{\text{Dist}(i, m)}$$

where Dist($i, m$) is the distance from a focus of environmental exposure, $m$, and the home of the $i^{th}$ survey participant (Waller and Poquette, 1999).

Environmental legacy exposures are exposures to any resident living near to uranium mines, mills, or waste piles that may have continued even after activity ceased and those mines or mills were abandoned. The distances between a respondent’s current home and mine and mill waste sites are useful to measure lifetime exposures in our analysis because the Navajo population is relatively stable with respect to residence. Participants have lived in their current residence for an average of 32 years. One-third of the study sample reported only one residence for their lifetime. In addition, a distance-based exposure surrogate may be calculated for all areas within the study area. This allows us to communicate findings about risk back to communities. We found that many participants live nearer to abandoned uranium mine or mill waste sites than they believed. We found 374 survey participants living within two miles of an abandoned uranium mine or mill, but only 210 (56\%) of these participants were aware that they lived this close to a uranium mine or mill. Participants may simply be unaware how far they actually live from abandoned uranium mines and mills. In addition, many abandoned uranium mine features within the study area are relatively small, unmarked, and/or unfenced, and have weathered considerably in the decades after they have been abandoned, making them unrecognizable as abandoned uranium mines. Although proximity alone in our preliminary analyses was a significant predictor of the diseases of interest here, proximity by itself is an incomplete description for environmental exposure because it fails to account for patterns of behavior that may increase risk of exposure to uranium mine or mill waste. We propose an environmental legacy exposure surrogate that is a generalization of the inverse distance measure and that is appropriate to our analyses because it not only accounts for the many abandoned mines and mills that are possible sources of exposure within the study area, but it also incorporates self-reported behavioral data.
This surrogate therefore presumes that the risk of exposure to the environmental legacy of uranium mining is not only elevated for residents living near to foci of environmental exposure, but also for residents who more often had contact with uranium mine and mill waste over the course of their lives.

We define the environmental legacy exposure surrogate ($L$) to be the mean weighted inverse-distance measure of the $i^{th}$ participant’s home to the 100 abandoned uranium mines and mills in and near the study area:

$$L_i = \frac{1}{100} \sum_{m=1}^{100} \frac{w_i}{\text{Dist}(i, m)}$$

where Dist($i, m$) is the distance between the home of the $i^{th}$ participant and the location of the $m^{th}$ abandoned uranium mine or mill and the weight $w_i$ is the total number of six possible environmental legacy exposures reported by each participant: $i$) 17% of survey participants responded that they had used materials from an abandoned uranium mine or mill; ii) 13% had herded livestock next to a uranium mine, mill or waste dump; iii) 13% had drunk or contacted uranium mine water or waste spills; iv) 13% had played on a uranium tailings pile or waste dump; v) 12% had played outdoors near a uranium mine, mill or waste dump; and vi) 2% had sheltered livestock in an abandoned uranium mine. The distribution of $w$ is shown in the second row of Table 1.

3. Univariate models for kidney disease, diabetes, and hypertension

In this section, we examine logistic regression models for kidney disease, diabetes, and hypertension. The focus of this analysis is to identify relevant sets of predictors for each of these diseases. This serves two purposes. First, we will examine models in which the covariates for the model for each disease may include the other two diseases. This will motivate the structure of a multivariate, conditionally specified logistic regression model for these three diseases, which we will analyze in Section 4. Second, we will identify a set of predictors for each disease that is small enough to allow us to exhaustively examine the entire model space for the conditionally specified logistic regression model in Section 4.

We define the following notation:

(a) $Y = (Y_1, \ldots, Y_N)$ where $Y_i$ is a binary response variable that is 1 if the $i^{th}$ participant reported being diagnosed and 0 otherwise. There are $N = 1304$ survey participants. The upper case $Y$ designates the random variable and the lower case $y = (y_1, \ldots, y_N)$ designates our data.

(b) $X_i = [X_{i0}, \ldots, X_{ip}]$ is a row vector that allows for a model intercept and that contains the values of the model covariates for the $i^{th}$ participant.

(c) $\beta = [\beta_0, \ldots, \beta_p]^T$ is the column vector of model coefficients.

Where describing different models, we attach the superscripts $K$ for kidney disease, $D$ for diabetes, and $H$ for hypertension to our model notation. For example, $X_i^K = [X_{i0}^K, \ldots, X_{ip^K}]$ designates the values of the covariates in the model for kidney disease.

The number of covariates for each model may differ, so we add a superscript to the number of covariates of each model. For example, $p^K$ is the number of covariates in...
3.1. Methods of univariate analysis

To identify good subsets of covariates to model our data, we follow the Bayesian model averaging approach described in Kass and Raftery (1995), in Hoeting et al. (1999), and in Viallefont et al. (2001). This approach is based on the estimation of Bayes factors, which are useful to exploratory model development because they are a method for comparing competing models by evaluating the evidence in favor of each model. To implement this method, each model is defined by the covariates that it includes, or by the number of coefficients that are nonzero. Here \( p \) is the maximum number of covariates that any model may have and so \( 2^p \) is the total number of possible models. We assign the prior probability \( P(\beta_j \neq 0) = 0.5 \) for \( j = 1, \ldots, p \), and so the prior probability for each model \( m = 1, \ldots, 2^p \) is \( P(M = m) = (0.5)^p \). The conditional probability of model \( m \) given \( y \) is

\[
P(M = m | y) = \frac{P(y | M = m)P(M = m)}{\sum_{i=1}^{2^p} P(y | M = i)P(M = i)}.
\]

We choose an arbitrary and fixed model and denote that model by \( M = 1 \). Because \( P(M = m) \) is a uniform prior, we can write Equation (2) as

\[
P(M = m | y) = \frac{P(y | M = m)}{\sum_{i=1}^{2^p} P(y | M = i)} = \frac{B_{m1}}{\sum_{i=1}^{2^p} B_{i1}}
\]

where \( B_{i1} \) is the Bayes factor \( B_{i1} = P(y | M = i) / P(y | M = 1) \).

Many methods to estimate \( P(y | M = i) \) are discussed in Kass and Raftery (1995). The analyses presented in this paper are based upon two variations of the Laplace approximation of \( \log(P(y | M = i)) \). Both are asymptotic results that are appropriate for problems with large sample sizes, and both are straightforward to compute, allowing for an exhaustive comparison of many possible models. Raftery (1995) describes in detail an estimate based upon the Schwartz criterion, also known as the Bayesian information criterion (BIC),

\[
\log(P(y | M = i)) \approx -\frac{1}{2} p_i \log(n) + \log(P(y | \hat{\beta}_i, M = i))
\]

where \( p_i \) is the number of parameters in model \( i \) and where \( \hat{\beta}_i \) is the MLE of the coefficients in model \( i \). Draper (1995) advocates a variant,

\[
\log(P(y | M = i)) \approx \frac{1}{2} p_i \log(2\pi) - \frac{1}{2} p_i \log(n) + \log(P(y | \hat{\beta}_i, M = i))
\]

observing that the inclusion of the \( \frac{1}{2} p_i \log(2\pi) \) term may improve estimates of Bayes factors when they are used to compare models with unequal numbers of parameters. We have compared the analysis based upon both Equations (4) and (5) and have found that they lead to similar results for the analyses discussed in this paper. We therefore only
Bayes factors are used to address problems of model selection as well as problems of model averaging. Model selection is a method for selecting a single model from a set of candidates, based on the data. A Bayesian approach to model selection is to select the model with the greatest posterior probability $P(y|M = i)$. There are other criteria for selecting a single “best” model from a collection of logistic regression models. For example, the Akaike information criteria (AIC) is commonly used for this purpose. Nonetheless, regardless of the criterion used for model selection, the common strategy of estimating a quantity of interest by first selecting a single model and then estimating the quantity of interest based upon that model ignores the problem that the model structure itself is uncertain. Model averaging is a Bayesian method for estimating a quantity of interest by first estimating that quantity for a range of candidate models and then averaging those estimates, weighted by the posterior probability of each model. As in Kass and Raftery (1995), if $\Delta$ is some quantity of interest in our models, the Bayesian model average is

$$E[\Delta|y] = \sum_{i=1}^{2^p} E[\Delta|y, M = i]P(M = i|y). \tag{6}$$

If the quantity of interest is an indicator that the $j^{th}$ model coefficient $\beta_j$ is nonzero, Equation (6) becomes

$$P(\beta_j \neq 0|y) = \sum_{i \in A_j} P(M = i|y) \tag{7}$$

where $A_j$ is the set of all models that contain $\beta_j$.

To use the DiNEH Project data to model the prevalence of kidney disease, let the probability that the $i^{th}$ participant reports kidney disease be specified by a binomial family with a logit link

$$\logit \left( P(Y^K_i = 1|X^K_i) \right) = \log \left( \frac{P(Y^K_i = 1|X^K_i)}{P(Y^K_i = 0|X^K_i)} \right) = X^K_i \beta^K.$$

The elements of $X^K_i$ allow for an intercept and for potential risk factors for kidney disease. These risk factors include $y^D_i$ and $y^H_i$, which are indicators that the $i^{th}$ participant reported a diagnosis of diabetes or hypertension. These risk factors also include gender, age, BMI, a family history of kidney disease, $M$, and $L$. $M$ is the active-mining related exposure surrogate and $L$ is the environmental legacy exposure surrogate. It is reasonable to examine $y^D_i$ and $y^H_i$ as possible covariates because diabetes and hypertension are accepted risk factors for kidney disease (CDC, 2007).

### 3.2. Results of univariate analysis

The first row of Table 2 reports the posterior probability that each model coefficient is nonzero for the kidney disease model, computed with Equation (7). There is strong evidence to include the active-mining related exposure surrogate $M$ in a logistic regression model for kidney disease; the coefficient $\beta^K_M$ is nonzero with posterior probability 0.97. The coefficient for a diagnosis of diabetes is also nonzero with posterior probability 1.00.
The coefficients for hypertension and a family history of kidney disease are nonzero in this model with posterior probabilities 0.26 and 0.45, respectively. This is not compelling evidence to exclude these covariates entirely, and we note that both are accepted risk factors for kidney disease. However, there is no evidence to include the other risk factors in a model for kidney disease; the coefficients for age and gender are both nonzero with posterior probability 0.02, the coefficient for BMI is nonzero with posterior probability 0.05 and the coefficient for the environmental legacy exposure surrogate is nonzero with posterior probability 0.07.

We use a similar approach to analyze which covariates are important to explain diabetes. The covariates that we considered for models for diabetes include a diagnosis of kidney disease, a diagnosis of hypertension, gender, age, BMI, a family history of diabetes, and the exposure surrogates $L$ and $M$. The second row in Table 2 reports the posterior probability that each coefficient in this model is nonzero. There is strong evidence to include the following covariates in a logistic regression model for diabetes: a diagnosis of kidney disease, a diagnosis of hypertension, age, BMI, and a family history of diabetes. Each of these covariates is nonzero with posterior probability 1.00. There is no evidence to include the covariates gender, $M$, or $L$ in this model.

Finally, we analyze models for the prevalence of hypertension. The covariates that we considered for models for hypertension were a diagnosis of kidney disease, a diagnosis of diabetes, gender, age, BMI, a family history of hypertension, and the exposure surrogates. The third row in Table 2 reports the posterior probability that each coefficient in this model is nonzero. The coefficients for a diagnosis of diabetes, age, BMI, and a family history of hypertension as predictors for hypertension are nonzero with posterior probability 1.00, indicating strong evidence to include these coefficients in the model. There is no evidence for including the active-mining related exposure surrogate $M$ in this model. The coefficients for a diagnosis of kidney disease, for the gender of the participant, and for the environmental legacy exposure surrogate $L$ are nonzero with posterior probabilities 0.61, 0.53, and 0.39, respectively. This is not conclusive evidence either to include or to exclude these three covariates in the univariate model. In Section 4, we use Bayesian model averaging to examine the evidence for these variables in the conditionally specified logistic regression model.

The results summarized in Table 2 are used to inform the conditionally specified logistic regression model in the next section. The most important findings are: 1) in each logistic regression model, it is plausible that the coefficients for the other disease endpoints are nonzero, 2) there is strong evidence for including the active-mining related exposure surrogate in a logistic regression model for kidney disease and 3) it is plausible that the parameter for the environmental legacy exposure surrogate is nonzero in a logistic regression model for hypertension.

4. A conditionally specified logistic regression model for kidney disease, diabetes and hypertension

A conditionally specified logistic regression model (Liu, 1994) is an approach, applicable to the DiNEH Project data, for modeling a multivariate binary response vector with covariates. In a conditionally specified logistic regression model, each element of the binary response vector is a covariate for logistic regressions on the other binary response
variables, but each of the conditionally specified logistic regressions may also include additional and different sets of covariates. In this section, we define the framework for conditionally specified logistic regression models for these data, we specify the conditions which guarantee that the conditionally specified models define a proper probability distribution, and we identify the best models of this class that describe our data.

4.1. Model definition

We use a notation that is parallel to the one described in Section 3. \( \mathbf{Y}_i = [Y^K_i, Y^D_i, Y^H_i] \) is a multivariate binary response vector that indicates if the \( i \)th participant has reported a diagnosis of kidney disease, diabetes, or hypertension, respectively. We use the lower case \( y_i = [y^K_i, y^D_i, y^H_i] \) to designate our observed data. We will model the response vector \( \mathbf{Y}_i \) with three covariate vectors \( \mathbf{X}^K_i, \mathbf{X}^D_i, \) and \( \mathbf{X}^H_i \) through three parameter vectors, \( \mathbf{\beta}^K, \mathbf{\beta}^D, \) and \( \mathbf{\beta}^H \). We use the analyses of Section 3 to inform the structure of our covariate matrices, but unlike the logistic regression models of Section 3, we do not include any of the elements of \( \mathbf{Y}_i \) in any of the covariate vectors \( \mathbf{X}^K_i, \mathbf{X}^D_i, \) or \( \mathbf{X}^H_i \). Instead we will use an additional parameter \( \mathbf{\alpha} = (\alpha_{KD}, \alpha_{DK}, \alpha_{KH}, \alpha_{HK}, \alpha_{DH}, \alpha_{HD}) \) to specify the conditional relationships among \( Y^K_i, Y^D_i, \) and \( Y^H_i \).

We define the joint probability distribution \( P(\mathbf{Y}_i = y_i) \) through three conditionally specified probability distributions. First, we specify the probability that the \( i \)th participant reports kidney disease by a binomial family with a logit link, conditional on that participant’s self-reported diagnoses of diabetes \( (Y^D_i) \) and hypertension \( (Y^H_i) \) as follows.

\[
\logit \left( P(Y^K_i = 1 | Y^D_i = y^D_i, Y^H_i = y^H_i, \mathbf{X}^K_i, \mathbf{\beta}^K, \mathbf{\alpha}) \right) = X^K_i \mathbf{\beta}^K + \alpha_{KD} y^D_i + \alpha_{KH} y^H_i \quad (8)
\]

Similarly, we specify the probability that the \( i \)th participant reports diabetes conditional on that participant’s self-reported diagnoses of kidney disease and hypertension as follows.

\[
\logit \left( P(Y^D_i = 1 | Y^K_i = y^K_i, Y^H_i = y^H_i, \mathbf{X}^D_i, \mathbf{\beta}^D, \mathbf{\alpha}) \right) = X^D_i \mathbf{\beta}^D + \alpha_{DK} y^K_i + \alpha_{DH} y^H_i \quad (9)
\]

Finally, we specify the probability that the \( i \)th participant reports hypertension conditional on that participant’s self-reported diagnoses of kidney disease and diabetes as follows.

\[
\logit \left( P(Y^H_i = 1 | Y^K_i = y^K_i, Y^D_i = y^D_i, \mathbf{X}^H_i, \mathbf{\beta}^H, \mathbf{\alpha}) \right) = X^H_i \mathbf{\beta}^H + \alpha_{KH} y^K_i + \alpha_{HD} y^D_i \quad (10)
\]

The ordering of the diseases in the specification is not important. However, a conditionally specified probability distribution does not necessarily define a proper joint probability distribution. Liu (1994) proves a necessary and sufficient condition for the conditionally specified logistic regression models to define a proper joint probability distribution. In our case, (8), (9), and (10) define a proper joint probability distribution when we impose the following restriction:

\[ \alpha_{KD} = \alpha_{DK}, \quad \alpha_{KH} = \alpha_{HK} \quad \text{and} \quad \alpha_{DH} = \alpha_{HD}. \]

Hence, \( \mathbf{\alpha} \) has only three distinct elements.
Liu (1994) also gives the joint probability distribution that follows after this restriction. For simplicity of notation, write the data as $y = (y_1, \ldots, y_N)$ and the model parameters as

$$\theta = (\beta^K_1, \ldots, \beta^K_p, \beta^D_1, \ldots, \beta^D_p, \beta^H_1, \ldots, \beta^H_p, \alpha_{KD}, \alpha_{KH}, \alpha_{DH})$$

The joint distribution of $Y_i$ can be written as

$$P(Y_i = y_i | \theta) = C_i^{-1} \exp \left\{ X^K_i \beta^K y^K_i + X^D_i \beta^D y^D_i + X^H_i \beta^H y^H_i + \alpha_{KD} y^K_i y^D_i + \alpha_{KH} y^K_i y^H_i + \alpha_{DH} y^D_i y^H_i \right\}$$

(11)

The normalizing constant $C_i$ is

$$C_i = \sum_{y^K=0}^1 \sum_{y^D=0}^1 \sum_{y^H=0}^1 \exp \left\{ X^K_i \beta^K y^K_i + X^D_i \beta^D y^D_i + X^H_i \beta^H y^H_i + \alpha_{KD} y^K_i y^D_i + \alpha_{KH} y^K_i y^H_i + \alpha_{DH} y^D_i y^H_i \right\}$$

(12)

and the likelihood function of the parameters given the data is

$$\mathcal{L}(\theta | y) = \prod_{i=1}^N P(Y_i = y_i | \theta)$$

(13)

where $P(Y_i = y_i | \theta)$ is given in Equation (11).

We follow the model averaging approach discussed in Section 3 to identify good subsets of covariates for this conditionally specified logistic regression model and to estimate the posterior expectation of each model parameter, taken over all possible model subsets. In the conditionally specified logistic regression model, there are $2^{18}$ subsets of the covariates that we introduced in Chapter 3. To make it feasible to examine all possible model subsets, we reduced the size of the model space in three ways. First, we eliminated coefficients that were nonzero with posterior probabilities less than 0.10 in the analysis in Section 3. Second, because family history is an accepted risk factor for these diseases, we assigned a prior probability of 1 that the coefficients for a family history of each disease were nonzero. Finally, because there was strong evidence for including these coefficients in the univariate analysis, we forced age and BMI into the models for diabetes and hypertension by assigning a prior probability of 1 that these coefficients were nonzero. This reduced the number of possible models to 128.

4.2. Model averaging results

Table 3 shows the covariates that were included in the conditionally specified logistic regression model. The first column of Table 3 shows the prior probabilities that each of the model parameters is nonzero, the second column shows the posterior probabilities that each of the model parameters is nonzero, and the third column shows the model average estimate of each parameter, calculated by using the MLE of the coefficients for each model as an estimate of the quantity $E[\Delta | y, M = i]$ in Equation (6).
Table 3 shows evidence for including $M$ in the conditionally specified logistic regression model as a predictor for kidney disease and for including $L$ in the conditionally specified logistic regression model as a predictor for hypertension. The evidence in favor of $M$ is strong. From a prior assumption $P(\beta_M^K \neq 0) = 0.5$ that $\beta_M^K$ is nonzero, we found that this parameter is nonzero with a posterior probability of 0.95. The evidence in favor of $L$ in the model is weaker. Using the Schwartz criterion estimate as in Equation (4) we calculated that $\beta_L^H$ is nonzero with a posterior probability of 0.59. Using the alternate approximation in Equation (5) we calculated that $\beta_L^H$ is nonzero with a posterior probability of 0.75. Both estimates of this posterior probability show an increase over our prior assumption of 0.50.

The third column in Table 3 gives an approximation of the parameters for the conditionally specified logistic regression model that is calculated as an ensemble over 128 different model structures. Taken as an average over all of the model structures, we estimate $E(\beta_M^K | y) = 0.43$. In other words, the active-mining related exposure surrogate is associated with elevated prevalence of kidney disease in this population. This finding corresponds to an odds ratio of approximately 1.54 for a participant who reported one active-mining related exposure versus a participant who reported none. In addition, we estimate $E(\beta_L^H | y) = 1.04$, which is evidence that the environmental legacy exposure surrogate is associated with elevated prevalence of hypertension in this model. In the following section, we will examine further the model with the highest posterior probability, and we will develop risk maps for the impact of this covariate based upon that model.

### 4.3. Analysis of the best model

We will now examine the “best model” in the sense that we will choose the model with the highest posterior probability. This model is defined by Table 4. We examine the model defined in Table 4 for two reasons. First, we wish to compare the posterior distributions of the parameters for a conditionally specified logistic regression model with an alternative model structure, the t-link model discussed in Section 5. Second, because this single model has a posterior probability of 0.38, it is representative of those models which include the coefficient $\beta_L^H$. Not only is the posterior probability of this model higher than that of any other model that contains $\beta_L^H$, but also the posterior means of the parameters in the model defined in Table 4 are similar to the posterior expectation of the parameters, taken over all models where the coefficient $\beta_L^H$ is nonzero.

In order to examine this model further, we follow a Bayesian approach to estimate the posterior distribution of all the parameters in the model defined by Table 4. The model parameters are assigned independent Normal priors, so $\pi(\theta) \sim N(0, \sigma_\theta^2 I)$ where the prior variance $\sigma_\theta^2$ is large. The results presented in this paper are based upon the $\sigma_\theta = 100$. An analysis to prior sensitivity showed that results were almost identical for $\sigma_\theta = 10$, $\sigma_\theta = 1000$, and $\sigma_\theta = 10000$. The posterior distribution of the model parameters given the data is $f(\theta | y) \propto \mathcal{L}(\theta | y) \pi(\theta)$ where $\mathcal{L}(\theta | y)$ is defined in Equation (13).

We estimated this posterior distribution by using random walk Metropolis-Hastings steps to construct a Markov chain of simulated samples from $f(\theta | y)$, which we denote $\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)}\}$ where $\theta^{(t)} = \{\theta_1^{(t)}, \ldots, \theta_p^{(t)}\}$ (Givens and Hoeting, 2005). Specifically, to iteratively simulate each $\theta_j^{(t)}$, we took the following steps:
We initialized the parameter chains at an arbitrary value and set all the tuning parameters equally. We generated 25,000 samples, which were discarded as a burn-in period to allow for convergence. We then set the tuning parameters $\sigma^2$ proportional to the standard deviations of the posterior distributions that were generated during the burn-in period and generated an additional 300,000 samples. Because these samples were highly correlated, 95% of the remaining samples were “thinned” (discarded), leaving 15,000 samples of the posterior distribution for each parameter to be analyzed.

The posterior means and standard deviations of model coefficients of the “best model” are shown in Table 4. The posterior distributions of $\beta^H_M$ and $\beta^H_L$ in this model are plotted with a solid line in Figure 2. The posterior probabilities for $\beta^H_M$ and $\beta^H_L$ being greater than zero in this model are 0.9997 and 0.9982, respectively. The impact of the environmental legacy exposure surrogate $L$ can be expressed as an odds ratio $\exp\{L(s,w)\beta^H_L\}$ for each grid point. In this expression, the term $L(s,w)$ designates the values of the environmental legacy exposure surrogate for some resident living at grid point $s$ who experienced $w$ environmental legacy exposures as in Equation (1). Figure 3 plots the odds ratio $\exp\{L(s,w)\beta^H_L\}$ for this model for a resident with one of the environmental legacy exposures versus a resident with none (top) and a resident with two environmental legacy exposures versus a resident with none (bottom). This figure shows 0.5 posterior quantile of these odds ratios. These odds ratios are greater than 1 everywhere within the study area. This is also the case, even at the 0.025 quantile of the posterior distribution.

5. A multivariate t-link model for kidney disease, diabetes, and hypertension

In Section 4, we examined a conditionally specified logistic regression model for multivariate binary data. We now compare the modeling results of the conditionally specified logistic regression model for the DiNEH Project data with the t-link model proposed by O’Brien and Dunson (2004). The t-link is a modification of the well-known multivariate binary data. We now compare the modeling results of the conditionally specified logistic regression model for multi-
Recall that $Y_i = [Y_i^K, Y_i^D, Y_i^H]$ is a multivariate binary response vector that indicates if the $i^{th}$ participant has reported a diagnosis of kidney disease, diabetes, or hypertension, respectively. The three covariate vectors have the same structure as the conditionally specified logistic regression model shown in Table 4. $X^K_i$ allows for an intercept and includes an indicator for a family history of kidney disease and the active mining related exposure surrogate $M$. $X^D_i$ allows for an intercept and includes age, BMI, and an indicator for a family history of diabetes. $X^H_i$ allows for an intercept and includes age, BMI, an indicator for a family history of hypertension, and the environmental legacy exposure surrogate $L$. The three parameter vectors associated with these covariate vectors are $\beta^K$, $\beta^D$, and $\beta^H$.

We model the multivariate binary response vector through a latent multivariate variable, $t_i = (t^K_i, t^D_i, t^H_i)^T$ such that

$$Y^j_i = 1(t^j_i > 0)$$

where the subscript $i$ identifies the participant $i = 1, \ldots, N$ and the superscript $j \in \{K, D, H\}$ identifies the disease. To model the mean of this latent variable, we let $X_i$ be a block diagonal matrix

$$X_i = \begin{bmatrix} X^K_i & 0 & 0 \\ 0 & X^D_i & 0 \\ 0 & 0 & X^H_i \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta^K \\ \beta^D \\ \beta^H \end{bmatrix}$$

be the vector of all model coefficients, so that

$$X_i \beta = \begin{bmatrix} X^K_i \beta^K \\ X^D_i \beta^D \\ X^H_i \beta^H \end{bmatrix}.$$  

We define the priors level of the multivariate latent variable to be

$$t_i \sim N(X_i \beta, \sigma^2 \phi_i^{-1} R)$$

$$\phi_i \sim \text{Gamma}(\nu/2, \nu/2)$$

where the matrix $R$ expresses the latent association between the observed binary variables. $R$ is a parameter to be estimated, subject to the restrictions that it is symmetric and that its diagonal elements are 1. We assigned independent normal priors to the unique elements of the matrix $R$ and to the elements of the vectors $\beta^K$, $\beta^D$, and $\beta^H$. We assigned the constant values $\nu = 7.3$ and $\sigma^2 = 2.39$, which are the restrictions required for the marginal distribution of $t_i = (t^K_i, t^D_i, t^H_i)^T$ to closely approximate a logistic distribution (O’Brien and Dunson, 2004).

We estimated the posterior distributions of the model parameters by using the Metropolis-Hastings algorithm described in detail in O’Brien and Dunson (2004) and implemented by Mwallili (2005), generating a Markov chain of 120,000 samples. We discarded the initial 20,000 samples as a burn-in period to allow for convergence and thinned the chain by discarding 90% of the remaining to reduce correlation of the posterior samples, leaving 10,000 samples for analysis.
For this model, the mean and standard deviation of the posterior distribution of $\beta^K_M$ are $E(\beta^K_M|y) = 0.37$ and $SD(\beta^K_M|y) = 0.09$. The mean and standard deviation of the posterior distribution of $\beta^H_L$ are $E(\beta^H_L|y) = 1.33$ and $SD(\beta^H_L|y) = 0.56$. The posterior distributions of $\beta^K_M$ and $\beta^H_L$ are shown by the dashed line in Figure 2. Although the t-link model defines the relationship between the three disease variables differently from a conditionally specified logistic regression model, the posterior distribution of these coefficients are more conservative than but similar to the posterior distributions of $\beta^K_M$ and $\beta^H_L$ in the conditionally specified logistic regression model in Section 4. Both model structures lead us to conclude that the active mining related exposure surrogate that we have proposed is associated with elevated rates of kidney disease in our data. In addition, our proposed environmental legacy exposure surrogate is associated with elevated rates of hypertension, regardless of the model structure we used for our analysis.

6. Discussion

In this paper we analyzed survey data of 1,304 residents of the Eastern Agency of the Navajo Nation to model the prevalence of kidney disease, diabetes, and hypertension, using covariates derived from self-reported environmental histories and from geospatial data as surrogates for potential exposure to uranium. This is the first comprehensive community health study of its kind, especially among Native American populations, and the first to study the health impact of environmental contamination from abandoned uranium mine and mill waste within the context of an existing elevated prevalence of kidney disease and hypertension and other known risk factors for these diseases within this community.

We proposed surrogate measures for active-mining-related and environmental legacy exposures to uranium mine and mill waste that have allowed us to distinguish between the impacts from these two distinct types of exposures that likely differed in route and dose. Distance from legacy waste and the number of activities leading to exposures to that waste also provide two surrogates for evaluating dose within the legacy exposures. Because many of the exposures of concern are historical or cumulative, reconstruction of quantitative exposure histories is difficult. Valid exposure surrogates provide a mechanism for not only identifying those at risk, but also for prioritizing interventions to reduce risks in a manner that most effectively improves population health.

The construction of these exposure surrogates incorporates self-reported data that counted the number of ways that an individual may have been exposed to uranium mine or mill waste. In addition, the environmental legacy exposure surrogate uses a distance-based function that describes how the impact of exposure to abandoned uranium mine or mill waste increases for individuals who live closer to sources of environmental contamination. We found strong evidence that the active-mining related exposure surrogate is associated with elevated prevalence of kidney disease in this population. We also found strong evidence that the environmental legacy exposure surrogate is associated with elevated prevalence of hypertension in this population in a manner consistent with a dose-response relationship based on the combined distance and activity exposure surrogates. We developed maps that identify regions within the study area which our models suggest have relatively higher risk of disease based upon their proximity to abandoned uranium mines and mills. These risk maps also show how continued environmental legacy
exposures further impact models for elevated risk of disease. These results are useful to
communicate with community members who continue to live in these areas and who
continue to be at risk from exposure to abandoned uranium mine and mill waste.

The sheer number of sites and volume of waste at the legacy sites present a problem
that far exceeds available resources for removal of waste. Therefore, the graphic heat
maps presented here have provided a tool that is being used in efforts to reduce risk
by 1) informing community members living in these areas about activities that increase
risk and the areas that present the greatest risk; 2) working with decision-makers to
help inform risk reduction actions through prioritization of highest risk sites, improved
signage, and additional data collection and outreach; and 3) working with clinicians to
develop rapid screening of patients at greatest risk through identifying the location of
residence and a discrete set of current and historical activities associated with increased
risk.

In this paper we examined the DiNEH Project data using several methods of analysis.
We began with a univariate approach to model kidney disease, diabetes and hypertension
separately. In each of these models it is reasonable to use the other two diseases as co-
variates, and this assumption was supported by our data. This motivated a multivariate
modeling approach, which we implemented with a conditionally specified logistic regres-
sion model. Because these three diseases have different risk factors, we used a Bayesian
approach to identify good models subsets by calculating the posterior probability that
each model parameter was nonzero. This allowed us to specify a conditionally specified
logistic regression model in which each disease had a different set of covariates. This
multivariate approach allowed us to model the relationships between these three diseases
and to propose a different set of covariates for each disease, distinguishing between the
effects of the two exposure surrogates within the model. A conditionally specified logistic regression model is not the only model for multivariate binary data. We compared
the results of the conditionally specified logistic regression model with a Bayesian t-link
model. Although they have different structures, both identified evidence for two different
surrogate measures for exposure to uranium mine and mill waste in a multivariate model
for kidney disease, diabetes, and hypertension.

Other authors have conducted studies of univariate and multivariate binary response
data that point towards possible extensions of this paper. Draper (1995) applied Bayesian
model averaging to the analysis of univariate binary data to account for uncertainties
not only in model covariates but also in the link function. Chib and Greenberg (1998)
introduced a method for calculating Bayes factors for the multivariate probit model.
Sabanés Bové and Held (2011) compared Monte Carlo methods for calculating Bayes
factors for univariate binary data with methods based upon the Laplace approximation.
Finally, Ghosh and Clyde (2011) describe a method for computing estimates of model
probabilities that does not rely upon asymptotic approximations such as the Schwartz
criterion estimate. Future analysis of the DiNEH project data could focus upon address-
ing the uncertainty of underlying model structure as well as the uncertainty of the model
parameters and could also employ Monte Carlo methods for calculating Bayes factors to
support the validity of the Laplace approximations used in this article for this class of
models.

While the use of surrogates for exposure have many useful applications and help us in
understanding risks posed by these waste sites, there remain limitations. Although the
probabilities in the posterior distribution indicate a strong association between disease endpoints and exposure, the precise contaminants and levels of exposure contributing to these associations cannot be determined. The exposures are to mixed wastes generated during mining and milling as previously noted, and the actual individual exposures reflect a cumulative life-span exposure rather than recent exposures readily quantified through biomonitoring. In addition, the disease endpoints are based on self-report, which although we have taken steps to validate through medical record reviews, may be subject to a degree of error or change over time. The work reported here is the first analytical phase of an ongoing effort that is continuing to analyze clinical data from a subset of participants, as well as to assess biomarkers of diseases in blood and urine samples in an effort to identify an early indicator of disease to improve diagnostics. Finally, detailed analysis of water consumption data collected as part of this effort will also be incorporated in exposure modeling in future studies. Through these efforts, we will continue to build on and refine the results reported here in efforts to further understand risks and improve health in these communities.

Acknowledgements

Funding for this work has been provided by the National Institute for Environmental Health Sciences, RO1 ES014565, R25 ES013208, and P30 ES-012072. Glenn Stark was a fellow in the Program of Interdisciplinary Biological and Biomedical Sciences (PiBBs) from 2009-10, funded by a grant from the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health, 5 T32 EB009414. The authors wish to thank the 20 participating Navajo Chapters and the 1304 chapter members who participated in the study. We also acknowledge the following for their work in collecting, validating, and managing the data: Harris Arthur (in memoriam), Eastern Navajo Health Board, Jerry Elwood, Harrison Gorman, Thomas Manning, Alta McCabe, Bernice Norton, Bernadette Pacheco, Tommy Rock, and Bess Seschillie. We also gratefully acknowledge the contributions of Jamie deLemos and Margaret Menache for initial analyses upon which this work has built, and Ed Bedrick and Curtis Miller for their input. We also acknowledge the guidance of the DiNEH Project Community Research Advisory Board: Edward Carlisle, Jay DeGroat, Herbert Enrico, Thomas Manning, Lynnea Smith, and Jean Whitehorse. Work in this project has been conducted under approval and review of the UNM HRRC and the Navajo Nation Human Research Review Board. Finally, we acknowledge the ongoing collaboration from Navajo Nation Division of Health, and Environmental Protection Agency, and US Environmental Protection Agency and Army Corps of Engineers for sharing data and helping in the characterization of exposures.

References

Elevated disease risk due to exposure to uranium waste on the Navajo Nation


Elevated disease risk due to exposure to uranium waste on the Navajo Nation


Fig. 1. Locations of 520 abandoned uranium mines and five abandoned uranium mills on the Navajo Nation. There are 98 discrete abandoned uranium mines and two abandoned uranium mills within four miles of the DiNEH study area.
Fig. 2. Posterior distributions of the coefficients $\beta^K_M$ and $\beta^H_L$ in the conditionally specified logistic regression model (dotted line) and for the t-link model (dashed line).
Fig. 3. Heat maps showing the 0.50 posterior quantile of $\exp\{\beta^H_L\}$ for residents with one environmental legacy exposure (top) and residents with two environmental legacy exposures (bottom).
Table 1. The distributions of $M$ and $w$

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4+</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>0.74</td>
<td>0.16</td>
<td>0.07</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>$w$</td>
<td>0.71</td>
<td>0.12</td>
<td>0.07</td>
<td>0.04</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 2. The posterior probability that each coefficient is non-zero in a logistic regression model. K, D, and H are indicators that the respondent was diagnosed with kidney disease, diabetes, and hypertension, respectively. FH is an indicator for a family history of disease.

<table>
<thead>
<tr>
<th>Model</th>
<th>K</th>
<th>D</th>
<th>H</th>
<th>Gender</th>
<th>Age</th>
<th>BMI</th>
<th>FH</th>
<th>M</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kidney Disease</td>
<td>1.00</td>
<td>0.26</td>
<td>0.03</td>
<td>0.03</td>
<td>0.06</td>
<td>0.45</td>
<td>0.96</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td>1.00</td>
<td>1.00</td>
<td>0.08</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.04</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.61</td>
<td>1.00</td>
<td>0.53</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.09</td>
<td>0.39</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. $P(\beta_i \neq 0)$, the prior probability that each coefficient is nonzero; $P(\beta_i \neq 0|y)$, the posterior probability that each coefficient is nonzero, and $E(\beta_i|y)$, the expected value of each coefficient.

| Kidney Disease | $P(\beta_i \neq 0)$ | $P(\beta_i \neq 0|y)$ | $E(\beta_i|y)$ |
|----------------|---------------------|----------------------|----------------|
| Intercept      | 1.0                 | 1.00                 | -4.32          |
| Family History | 1.0                 | 1.00                 | 0.73           |
| $M$            | 0.5                 | 0.95                 | 0.43           |
| $L$            | 0.5                 | 0.04                 | 0.03           |

| Diabetes       | $P(\beta_i \neq 0)$ | $P(\beta_i \neq 0|y)$ | $E(\beta_i|y)$ |
|----------------|---------------------|----------------------|----------------|
| Intercept      | 1.0                 | 1.00                 | -2.82          |
| Family History | 1.0                 | 1.00                 | 0.88           |
| Age            | 1.0                 | 1.00                 | 0.04           |
| BMI            | 1.0                 | 1.00                 | 0.05           |
| $M$            | 0.5                 | 0.03                 | 0.00           |
| $L$            | 0.5                 | 0.04                 | 0.02           |

| Hypertension   | $P(\beta_i \neq 0)$ | $P(\beta_i \neq 0|y)$ | $E(\beta_i|y)$ |
|----------------|---------------------|----------------------|----------------|
| Intercept      | 1.0                 | 1.00                 | -1.75          |
| Family History | 1.0                 | 1.00                 | 0.78           |
| Age            | 1.0                 | 1.00                 | 0.06           |
| BMI            | 1.0                 | 1.00                 | 0.06           |
| Gender         | 0.5                 | 0.24                 | -0.07          |
| $M$            | 0.5                 | 0.06                 | 0.01           |
| $L$            | 0.5                 | 0.59                 | 1.04           |

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\alpha_{KD}$</th>
<th>$\alpha_{KH}$</th>
<th>$\alpha_{DH}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{KD}$</td>
<td>1.0</td>
<td>1.00</td>
<td>1.46</td>
</tr>
<tr>
<td>$\alpha_{KH}$</td>
<td>1.0</td>
<td>1.00</td>
<td>0.75</td>
</tr>
<tr>
<td>$\alpha_{DH}$</td>
<td>1.0</td>
<td>1.00</td>
<td>1.86</td>
</tr>
</tbody>
</table>
Table 4. Posterior means and standard deviations of the coefficients of the conditionally specified logistic regression model with the highest posterior probability. $M$ is the active-mining related exposure surrogate and $L$ is the environmental legacy exposure surrogate.

<table>
<thead>
<tr>
<th>Kidney Disease</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.37</td>
<td>0.26</td>
</tr>
<tr>
<td>Family History</td>
<td>0.72</td>
<td>0.26</td>
</tr>
<tr>
<td>$M$</td>
<td>0.44</td>
<td>0.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Diabetes</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.83</td>
<td>0.16</td>
</tr>
<tr>
<td>Age</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>BMI</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>Family History</td>
<td>0.89</td>
<td>0.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hypertension</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.83</td>
<td>0.13</td>
</tr>
<tr>
<td>Age</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>BMI</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>Family History</td>
<td>0.78</td>
<td>0.15</td>
</tr>
<tr>
<td>$L$</td>
<td>1.81</td>
<td>0.64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{KD}$</td>
<td>1.48</td>
<td>0.30</td>
</tr>
<tr>
<td>$\alpha_{KH}$</td>
<td>0.75</td>
<td>0.31</td>
</tr>
<tr>
<td>$\alpha_{DH}$</td>
<td>1.88</td>
<td>0.16</td>
</tr>
</tbody>
</table>