NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

06-13

A Bayesian multivariate analysis of children's exposure to pesticides

N. Cressie, M. Morara, B. Buxton, N. McMillian, W. Strauss and N. Wilson

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Environmentrics 00, 1–19

DOI: 10.1002/env.XXXX

A Bayesian multivariate analysis of children's exposure to pesticides

N. Cressie[a*], M. Morara[b], B. Buxton[b], N. McMillan[b], W. Strauss[b], N. Wilson[c]

Summary: In this article, we present a multivariate Bayesian analysis of the relationships, in preschool children, between environmental pathways of exposure to a non-persistent pesticide, chlorpyrifos (CPF), and its corresponding biomarker in urine, trichloropyridinol (TCP). The analysis uses the three years of data from the Pesticide Exposures of Preschool Children Over Time (PEPCOT) study. Hierarchical Bayesian analysis of pathways of exposure has gained popularity in recent years, where missing and censored data are modeled, and measurement and regression errors are accounted for in a single hierarchical statistical model. Here we consider multivariate pathways, where CPF and its metabolite TCP are modeled jointly in the environmental media. In this article, we analyze each of the three years of the study, focusing on the within-year multivariate nature of the PEPCOT data set. We present the results in a way that allows for an easy comparison of the fitted parameters over time.

Keywords: BHM; biomarker; environmental media; exploratory data analysis; PEPCOT study.

1. INTRODUCTION

George Casella and Noel Cressie were Past President and President, respectively, of the American Statistical Association's Section on Statistics and the Environment (ENVR) in 1998. We believe that George would have appreciated the science, the frequentist exploratory

[a] *National Institute Applied Statistics Research Australia, University of Wollongong, NSW 2522, Australia*
[b] *Battelle Memorial Institute, Columbus, OH 43201, USA*
[c] *Battelle Memorial Institute, Durham, NC 27713, USA*
[*] *Correspondence to: N. Cressie, National Institute Applied Statistics Research Australia, University of Wollongong, NSW 2522, Australia. E-mail: ncressie@uow.edu.au*

This paper has been submitted for consideration for publication in *Environmetrics*

data analysis, and the Bayesian inference that underly the analysis below. He will be missed in so many ways.

Environmental epidemiologic studies aim to characterize relationships between complex and often subtle human exposures to environmental agents and adverse health effects within target populations. Over the past 15 years, there has been significant research in developing biomarkers of exposure in urine and blood, since they can be cost-effective metrics of exposure along exposure pathways involving, for example, air, water, food, soil, and dust.

The study of Pesticide Exposures of Preschool Children Over Time (PEPCOT) sought to estimate the changes in aggregate exposures to targeted pesticides for selected preschool children over a three-year time period (Wilson *et al.*, 2009). The targeted pesticides in the PEPCOT study included pyrethroid and organophosphate pesticides and acid herbicides, which are or have been used in homes, schools, and other settings in which young children might be in contact with them. The PEPCOT study investigated the aggregate exposures of sibling pairs, living in the same household, to the targeted pesticides. Data were collected three times over the study period.

Children can be exposed to environmental pollutants through multiple contamination pathways and multiple routes (inhalation, dietary ingestion, non-dietary ingestion, and dermal absorption). Compared to adults and other children, young children may have increased exposures to environmental pollutants, because of what the children eat and drink, where they spend their time, and what they spend their time doing. Because young children's development changes so fast, relatively small differences in ages can result in relatively large differences in total exposure. Furthermore, the impact of the exposures may be greater on young children, because of their smaller body masses and immature body systems (Perera, 1977; Schettler, 2001; Mendola *et al.*, 2002; Wigle *et al.*, 2007). Very young children learn about their environment by exploring not only the appearance and texture of objects, but also their taste and smell. Thus, non-dietary ingestion may also play an important role

in their exposures. Several questionnaire-based and epidemiologically based studies have implicated pesticide exposures and exposures to other xenobiotics as possible causes of children's health problems (Goldman, 1995; Landrigan *et al.*, 1999, 2004; Birnbaum and Fenton, 2003; Eskenazi *et al.*, 2007; Adams *et al.*, 2009).

In this article, we consider the case of childhood exposure to a non-persistent pesticide, chlorpyrifos (CPF), and its metabolite and corresponding biomarker of exposure in urine, namely 3,5,6-trichloro-2-pyridinol, or trichloropyridinol (TCP). The PEPCOT study is one of very few that measured both pesticide and metabolite concentrations in air, dust, soil, and food, as well as the metabolite in urine in a repeated-measures study. Due to the fact that the urinary TCP metabolite concentration is related to exposure to both CPF and TCP in environmental media, any statistical analyses of the PEPCOT data must consider both chemicals *jointly*.

The extent of one individual's exposure depends on a large number of factors, including physical and chemical properties of the toxic pollutants, environmental properties that govern the fate and transport of the pollutants through different environmental media (e.g., air, water, food, soil, dust), and behavioral, nutritional, and other factors that determine the extent to which an individual comes into contact with the pollutants. Historically, quantitative (statistical) models were often pieced together. For example, fate and transport might be modeled separately from behavior and human activities; then the models would be combined, often without fully accounting for model uncertainties or correlations among factors.

Over the past decade, Bayesian hierarchical models (BHMs) have been growing in popularity for addressing complex quantitative problems, such as those posed by human-exposure studies. One of the major advantages of a BHM for quantitative human-exposure studies, is that it offers the flexibility to combine, in a single model, data from different sources that inform different aspects of the exposure scenario and where there are different

levels of variability. Also, as a practical matter, the BHM can deal coherently with the censoring of data that often occurs when measuring the presence of chemicals in media and in blood or urine.

The type of statistical problem posed by these exposure-biomarker studies has been recently addressed by authors like Clayton *et al.* (2002), McMillan *et al.* (2006), Cressie *et al.* (2007), Santner *et al.* (2008), and Craigmile *et al.* (2009). These articles considered human exposures to toxic metals in the environment, focusing on one metal (arsenic or lead) at a time. While this is a reasonable approach, humans are quite often simultaneously exposed to multiple pollutants, and a model that considers all of them jointly could potentially provide more accurate and precise predictions of exposure and inferences about significant pathways. Morara *et al.* (2010) gave a BHM that accounted for multivariate exposures. In this paper, we use that BHM to analyze bivariate (CPF and TCP) data from the PEPCOT study of preschool children. To our knowledge, it represents the first time that a multivariate statistical analysis has been applied within the context of exposure-biomarker pathways investigations.

Section 2 describes the PEPCOT study and discusses the data used in the multivariate analysis. Section 3 presents exploratory data analyses and associated data summaries of the PEPCOT data. Section 4 presents the multivariate BHM, including data models, process models, and priors. Section 5 presents the results from fitting the BHM, and a discussion of these results is provided in Section 6.

2. PEPCOT STUDY

The PEPCOT study (Wilson *et al.*, 2009) sought to estimate changes in exposure for a small group of preschool-aged children over a three-year period. The changes considered were in aggregate exposures to selected pesticides, and interpersonal variability in these exposures

was assessed for children living in different homes and children living in the same home. The study was conducted from June, 2002 to May, 2007, with field sampling in 2003, 2004, and 2005 in 50 households located within one-hour driving time from Durham, North Carolina. In each of the 50 households, two (or more) children were recruited, such that one child was age three years in the first sampling year, and the other child was a younger sibling.

In the age-range of the children in the study, a small difference in age can make a big difference in the child's stage of development. Thus, we expect that differences in aggregate exposure between siblings will arise because they spend different time in different micro-environments, their activities are different, they ingest different foods (and sometimes non-foods), they have different breathing rates, they have different hand-to-mouth behaviors, and so forth.

The sampling objective was to collect environmental and personal samples once a year for three consecutive years (2003, 2004, 2005) in each of the 50 households. Each household was sampled in the same season (spring, summer, or fall). Sampling in the second and third years for each family was scheduled within two weeks of the date of the first sampling event. During each annual visit, samples were collected over the course of 24 hours. Environmental samples were collected from indoor and outdoor air, indoor-carpeted floor dust, soil, food-preparation surface wipes, and uncarpeted-floor surface wipes. Personal samples collected from each child (by the parents or other adult household members) included duplicate diet samples (liquid and solid food eaten during the 24-hour period), hand-wipe samples, and first-morning-void urine samples. Other supplemental questionnaire and survey information included food and activity diaries, household characteristics, and other ancillary information. The multimedia samples were extracted using Soxhlet, sonication, or accelerated solvent techniques; then they were analyzed by gas chromatography/mass spectrometry in the selected ion monitoring mode. Liquid food was not included in this analysis because most liquid food samples had no discernible levels of the target analytes (including CPF and TCP).

Prior to the analysis, all data were converted to molar concentrations to ensure that the intake of one molecule of either CPF or TCP was considered to produce one molecule of TCP in the urine, and then the data were transformed by taking natural logarithms. The units of measurement for the various exposure and environmental samples are shown in Table 1. Working on the log scale is a standard approach in statistical analysis of environmental and biomarker data, since they usually follow a log-normal distribution. Moreover, working on the log scale often makes the statistical errors additive and homoscedastic.

[Table 1 about here.]

3. EXPLORATORY DATA ANALYSIS (EDA) OF THE PEPCOT DATA

This section reports within-media and between-media summary statistics for the CPF and TCP data in all environmental and biological media. In particular, simple regression models were fitted for all pairs of variables (i.e., chemical concentrations in all media) separately for the CPF data and the TCP data.

During this exploratory data analysis (EDA), for cases where the CPF or TCP level in a sample was found to be below the laboratory method detection limit (MDL), the concentration was set to $\log(\text{MDL}/\sqrt{2})$ (see, for example, Hornung and Reed, 1990). Note that this method may lead to biased estimates in a contamination analysis (Succop *et al.*, 2004; Baccarelli *et al.*, 2005), and therefore we only used it in the EDA. Indeed, one of the main features of the BHM presented in this paper (Section 4) is the ability to impute missing and censored data from the probability-distributional models and the available data.

The summary statistics from our EDA are presented in Table 1 of the Supplemental Material section, and they include the estimated mean on the log scale, $\hat{\mu}$, with associated $(2.5\%, 97.5\%)$ confidence limits (a bold value means it is statistically significantly different from 0 at the 0.05 level), the estimated standard deviation on the log scale, $\hat{\sigma}$, and

sample size n (with the associated number of missing values in parentheses). Statistics are provided for each sampling medium, each CPF and TCP analyte, and each sampling year. The (approximate) sample size, $n = 100$, reflects sampling for two children in each of 50 households. Generally, the data indicate decreasing levels over time of CPF and TCP in the environmental media (i.e., hand wipes, floor dust, indoor air, and outdoor air), but either relatively flat or increasing levels in solid food and urine (the exposure measure). Simple correlation coefficients, $\hat{\rho}$, between CPF and TCP levels (on the log scale) are also presented. Note that the correlations are generally high, especially in media like floor dust, indoor air, and outdoor air. This feature of the data was part of the motivation for conducting a multivariate statistical analysis, since such an approach is designed to take advantage of various correlations and dependencies in the data.

The corresponding histograms are presented in Figures 1-3 of the Supplemental Material section, which show the general range and shape of the univariate-data distributions. In most cases, these histograms indicate reasonably symmetric distributions, although some cases of skewness (e.g., CPF, TCP in outdoor air in all three years) are also evident.

The results of the regressions between pairs of media are shown in Table 2 of the Supplemental Material section, and they include the intercept $\hat{\mu}$, with associated $(2.5\%, 97.5\%)$ confidence limits, slope $\hat{\beta}$, with associated $(2.5\%, 97.5\%)$ confidence limits (a bold value means statistical significance at the 0.05 level), coefficient of determination R^2, and sample size n (with the associated number of missing values in parentheses). Each sub-table represents the regression of the first medium listed as a function of the second medium. The corresponding scatter plots are presented in Figures 4-9 of the Supplemental Material section, which show the general shape of dependence among all pairs of environmental and biological media, with tighter data clouds indicating a stronger correlation (e.g., floor-dust and indoor-air CPF and TCP measurements in all three years). For the purposes of EDA, potentially significant correlations between different media can be judged by examining the

confidence bounds for the slope and highlighting cases where the bounds do not contain the value of zero (shown in bold text in Table 2 of the Supplemental Material section). These cases indicate strong correlations and hence are cases that might be expected to result in important exposure pathways under the BHM analysis (Section 5). Of all the cases highlighted in the regression table (a total of 41 cases), nearly half of them (18 cases) involve correlations among CPF and TCP in hand wipes, floor dust, and indoor air. This could indicate transport of the two analytes between floor dust and indoor air, and from there onto the hands of children and the hand wipes. In addition, correlations with TCP levels in urine, the biomarker of primary interest in this analysis, are seen in 12 cases, namely correlation with TCP or CPF in solid food (3 cases), floor dust (2 cases), indoor air (3 cases), and outdoor air (4 cases). These findings were used to motivate the pathways model shown in the next section (see Figure 1).

Because sampling in the PEPCOT study involved pairs of children within households, we repeated the EDA using a mixed model to account for the within-household correlations. However, the t-statistics calculated to test the significance of the regression parameters were largely unaffected and did not warrant our modeling the within-household correlations in the hierarchical Bayesian model.

4. MULTIVARIATE BAYESIAN HIERARCHICAL MODEL

The objective of this article is to use data from the PEPCOT study to assess the magnitude and statistical importance of various environmental and personal exposure pathways, all the way from the sources of pesticide contamination, CPF, to urinary TCP as a human-exposure biomarker. The assessment is based on regression coefficients relating CPF and TCP levels in different environmental and biological media using the multivariate BHM given by Morara *et al.* (2010).

Bayesian hierarchical modeling offers a coherent way to handle missing data, non-detects, and measurement error *simultaneously*, by separating the data model from the "true process" model (see, for example, Gelman *et al.*, 2003). It makes use of conditional probability distributions, where we write $[A|B]$ to denote the conditional distribution of the variable A given the variable B.

Consider a population of N^I individuals from whom measurements of CPF and TCP are collected in $N^X - 1 = 5$ environmental media (solid food, hand wipe, floor dust, indoor air, outdoor air), and measurements of TCP are taken in urine. In our case, $N^I = 100$ and $N^X = 5 + 1 = 6$.

Let Y_{ijs} and X_{ijs} represent the measured log value and the true log value associated with individual i, medium j, and species $s \in \{1 = \text{CPF}, 2 = \text{TCP}\}$, respectively. Let Z_{ijs} be the logarithm of the MDL. Use S^A to indicate the set of indices (i, j, s) for which there are measurements reported and S^B to indicate the set of indices (i, j, s) for which the measurements are censored and simply reported to be below the MDL. The *data model* expresses the distribution of Data (here log measurements, including those that are left-censored) given the Process (here log of the true CPF and TCP concentrations) and the Parameters:

$$[\text{Data}|\text{Process}, \text{Params}] = \prod_{(i,j,s)\in S^A} N\left(Y_{ijs}, X_{ijs}, \omega_{js}\right) \times \prod_{(i,j,s)\in S^B} \Phi\left(Z_{ijs}, X_{ijs}, \omega_{js}\right) , \qquad (1)$$

where $N(x, m, t)$ and $\Phi(x, m, t) = \int_{-\infty}^{x} N(y, m, t)dy$ denote the normal probability density function and the normal cumulative distribution function, respectively, with mean m and precision (i.e., the reciprocal of the variance) t. Notice that our notation emphasizes the precision parameter rather than the variance parameter.

The log of the true value X defines the Process, and it is modeled using a pathways model involving linear regression with normal errors. The pathways are defined using sets of indices, indicating the conditional dependencies of one medium given the others, which we

call selector sets:

$$S_j \subseteq \{1, \ldots, N^X\} \setminus \{j\}; \quad j = 1, \ldots, N^X. \tag{2}$$

The pathways model used for the PEPCOT data, which was motivated by the EDA results given in Section 3, is displayed in Figure 1. Notice that the selector sets must define an acyclic directed graph (see, for example, Lauritzen, 1996). If we index the media as:

$$\begin{array}{lll} \text{urine} = 1\,; & \text{solid food} = 2\,; & \text{hand wipe} = 3\,; \\ \text{floor dust} = 4\,; & \text{indoor air} = 5\,; & \text{outdoor air} = 6, \end{array}$$

then the selector sets associated with the pathways in Figure 1 are:

$$S_1 = \{2,3,4,5,6\}\,; \quad S_2 = \emptyset\,; \quad S_3 = \{4\}\,; \quad S_4 = \{5\}\,; \quad S_5 = \{6\}\,; \quad S_6 = \emptyset\,.$$

[Figure 1 about here.]

The Process is made up of Biomarker (i.e., log of the true TCP values in urine) and Environment (i.e., log of the true CPF and TCP values in the environmental media). Hence the *process model* can be written as

$$[\text{Biomarker}|\text{Environment}, \text{Params}] \times [\text{Environment}|\text{Params}]$$

We use univariate regressions to model the biomarker in urine, where we only have TCP (the CPF is metabolized in the body), and bivariate regressions for the environmental media, where both CPF and TCP are present.

Since the TCP in the urine comes from both CPF and TCP exposure, we write

$$[\text{Biomarker}|\text{Environment}, \text{Params}] = \prod_{i=1}^{N^I} N\left(X_{i12},\ \mu_{12} + \sum_{k \in S_1} \sum_{s \in \{1,2\}} \beta_{1ks} X_{iks},\ \tau_{122}\right), \tag{3}$$

where the subscript "1" indicates "urine," followed by the subscript "2" or "22" to indicate "TCP." Further, $\mu_{12} \in \mathbb{R}$, $\beta_{1ks} \in \mathbb{R}$, and $\tau_{122} > 0$ represent the intercept, the regression coefficients, and the precision, respectively, of the univariate regression. In the environmental media, we model CPF and TCP jointly as,

$$[\text{Environment}|\text{Params}] = \prod_{i=1}^{N^I} \prod_{j=2}^{N^X} N\left(X_{ij},\ \mu_j + \sum_{k \in S_j} B_{jk} X_{ik},\ \tau_j\right), \tag{4}$$

where $X_{ij} = \begin{bmatrix} X_{ij1} \\ X_{ij2} \end{bmatrix}$, $\mu_j = \begin{bmatrix} \mu_{j1} \\ \mu_{j2} \end{bmatrix}$, $B_{jk} = \begin{bmatrix} \beta_{jk1} & 0 \\ 0 & \beta_{jk2} \end{bmatrix}$, and $\tau_j = \begin{bmatrix} \tau_{j11} & \tau_{j12} \\ \tau_{j21} & \tau_{j22} \end{bmatrix}$

represent the dependent variables, the intercepts, the regression coefficients, and the precision matrices, respectively, of the *bivariate* regression models. Note that the model captures the relationship between the pesticide CPF and the metabolite TCP within each environmental medium j, through the off-diagonal covariance terms in τ_j. Then equations (3) and (4) together define the *process model*. The joint distribution, conditional on the parameters, is obtained by multiplying equations (1), (3), and (4).

The power of Bayesian hierarchical modeling lies in the possibility of using prior information about the parameters in the model. This is called the *parameter model* (or the prior), which we write here as [Params]. For linear regression models with normal errors, the standard choices for prior distributions are gamma/Wishart for the precision, and normal for the regression coefficients (including the intercept). Assuming independence between the parameters, these priors, which are conjugate, result in,

$$\begin{aligned}[\text{Params}] = [\omega, \mu, \beta, \tau] \ &= \ \prod_{js} G(\omega_{js}, s^{\omega}_{js}, r^{\omega}_{js}) \times \prod_{js} N(\mu_{js}, m^{\mu}_{js}, t^{\mu}_{js}) \\ &\quad \times \prod_{js} N(\beta_{js}, m^{\beta}_{js}, t^{\beta}_{js}) \times \prod_{j} W(\tau_j, \nu^{\tau}_j, R^{\tau}_j)\,, \end{aligned} \tag{5}$$

where $G(\omega, s, r)$ denotes the gamma probability density function with shape s and rate (i.e.,

the reciprocal of the scale); and $W(\tau, \nu, R)$ denotes the Wishart probability density function with degrees of freedom ν and rate matrix (i.e., the inverse of the scale matrix) R.

The parameters in the prior distributions are called hyper-parameters and are fixed. They are usually set to values that give non-informative prior distributions or, if possible, to values determined through subjective judgment or previous similar studies (Gelman *et al.*, 2003). In our case, we use extra documentation that was available with the data to provide values for the measurement error. In particular, we set the prior data precision as follows:

$$\omega_{js} = \left(\frac{2}{\ln(1+\epsilon_{js})}\right)^2, \quad j = 1, \ldots, N^X, \quad s = 1, 2, \tag{6}$$

where ϵ_{js} is the relative measurement error associated with medium j and analyte s. This yields a degenerate prior distribution, which is numerically implemented by a tight gamma distribution with mean ω_{js} and a very small variance, as discussed below. Table 2 shows the relative measurement errors ϵ_{js} for the levels of CPF and TCP in the various sampled media. The factor 2 in equation (6) comes from setting twice the standard deviation of the data model, $2/\sqrt{\omega}$, equal to the log measurement error, $\log(1+\epsilon)$.

[Table 2 about here.]

Bayesian models are usually fitted via Markov chain Monte Carlo (MCMC) sampling (see, for example, Robert and Casella, 2004). As expected, we see that the stability of fitting our BHM this way is controlled by both precisions, ω and τ. To achieve appropriate ergodic behavior of the Markov chain, the prior hyper-parameters for τ were chosen based on the assumption that the precision of the process model is unlikely to be greater than the precision of the data model. The opposite assumption, aside from being conceptually hard to justify, can lead to numerical instabilities during MCMC sampling. (Allowing the sampling of values of τ significantly greater than ω can force X to over-fit the process model regardless of the data; in the MCMC, this makes the process-model residuals very small, which in turn pushes

τ to even larger values.) To make it unlikely for τ to take values larger than ω, we set the rate matrix in the Wishart prior distributions for τ_j as follows:

$$R_j^\tau \equiv \begin{bmatrix} r_{j1}^\tau & 0 \\ 0 & r_{j2}^\tau \end{bmatrix} = \begin{bmatrix} N^I/\omega_{j1} & 0 \\ 0 & N^I/\omega_{j2} \end{bmatrix}; \quad j = 1, \ldots, N^X. \tag{7}$$

Table 3 shows the values of the gamma and Wishart hyper-parameters. The shape and rate hyper-parameters, s^ω and r^ω, for the gamma priors were chosen such that the mean $s^\omega/r^\omega = \omega$, and the variance $s^\omega/(r^\omega)^2 \ll \omega$ is very small relatively to the mean. The degrees of freedom of the Wishart distribution, ν^τ, were set equal to the non-informative value of 0, resulting in an improper prior (notice that any small value for the degrees of freedom in the Wishart prior distribution would not significantly affect the posterior distribution, since that value is added to the population size N^I), and the rate matrix was set according to equation (7). Non-informative improper priors were chosen for μ and β.

[Table 3 about here.]

5. RESULTS

Samples from the posterior distribution, which is proportional to the product of (1), (3), (4), and (5), were obtained via MCMC simulation. The MCMC sampler was implemented in C++ using a dedicated C++ object library for MCMC sampling (Morara, 2008).

For each one of the three years in the study, 10^3 samples were obtained by drawing 10^6 samples and keeping one draw every 10^3 draws. This long thinning period was chosen to break, as much as possible, the autocorrelation in the chain. Sources of autocorrelation in the samples are: the high level of missing values and non-detects in some of the media (in particular, floor dust); and the high correlation between CPF and TCP in some of the media. A burn-in of 1 million MCMC iterations was also chosen before any samples were taken. The

MCMC simulation, for a total of 33 million iterations (11 million for each of 3 years), ran in about 8 hours on a PC with a 2.66 GHz Intel® Core™2 Duo CPU.

The marginal posterior parameter estimates obtained from the chains are presented in Tables 4-9. Each estimate is made up of three values: the median and, in parentheses, the 2.5 and 97.5 percentiles. A bold median indicates that the 95% prediction interval does *not* contain 0.

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

[Table 9 about here.]

Each sub-table represents the multivariate (CPF, TCP) regression in the indicated media. Posterior summaries of the intercepts, slopes, and inverse precisions (i.e., variances) of the two CPF, TCP components are shown, together with the correlation between CPF and TCP. As described in the previous section, the urine regression model estimates TCP in urine given CPF and TCP in the environmental media, while the environmental regression models estimate CPF and TCP in one environmental medium given CPF and TCP in the other environmental media, according to the modeled pathways.

One major strength of the multivariate BHM is that it simultaneously accounts for the interdependence and cross-dependence between CPF and TCP in all the environmental media, and for the cross-dependence of TCP in urine on both CPF and TCP in the environmental media. Our multivariate hierarchical approach provides a more parsimonious model (i.e., a simpler model including fewer parameters) than, for example, the series of

regression analyses conducted as part of the EDA described in Section 2). In fact, while the exploratory pairwise regressions indicated 41 possibly significant pathways, the multivariate BHM indicated only 21 possibly important pathways. Generally, the important multivariate BHM pathways were a subset of the pairwise regression pathways, although there were two important BHM pathways (between indoor air and outdoor air) that were not significant in the exploratory regressions.

6. DISCUSSION

From an exposure and environmental-protection perspective, an important objective of the PEPCOT study, and other similar studies, is to sort through the data and try to determine which pathways represent significant transport of pollutants through the part of the environment where ultimately the study participants are exposed to them. If a more simplistic regression approach were used to interpret the data, similar to the results of the EDA shown in Table 3, then the findings would be somewhat mixed and inconclusive. Significant regressions were indicated between the urine biomarker and four of the five environmental media (i.e., all media except hand wipes), although no adjustments in the EDA were made for multiple comparisons. In addition, significant regressions were also indicated for virtually all pairs of environmental media, suggesting that CPF and TCP move relatively freely around the entire household micro-environment. Faced with these (exploratory) findings that *suggest* that everything is correlated with everything else, it becomes more difficult to identify the environmental-protection priorities and determine ways to limit exposures effectively. In contrast, our results from fitting a multivariate BHM indicate simpler and more focused findings.

Only two important pathways of TCP to the urine biomarker are indicated: one pathway comes directly from solid food, and a second pathway comes directly from outdoor air.

Other strong pathways of both CPF and TCP are indicated within the four environmental media, namely hand wipes, floor dust, indoor air, and outdoor air. However, none of the environmental media, other than outdoor air, indicates an important pathway to the urine biomarker. As such, the fitted BHM suggests that exposure-mitigation efforts in micro-environments, like those in the PEPCOT study, should emphasize limiting exposures to CPF and TCP in solid food and outdoor air.

It should be noted that only two of the participating households in the PEPCOT study used CPF, and the measured concentrations of CPF in all environmental media were very low. Indeed, the US Environmental Protection Agency required that CPF be phased out in residential and other settings where children could be exposed, starting in 2000. The CPF in the environment most likely came from the few agricultural uses that were still permitted, and from the more persistent compound TCP from residual amounts in the environment.

Analogous multivariate BHMs for scenarios with greater pesticide use or use of current, less-volatile, and less-persistent pesticides will likely indicate different pathways. Multivariate statistical models and multivariate BHMs offer opportunities to combine disparate data for different pollutants, measured in a variety of environmental media, into a single statistical framework. In turn, this allows the model to simultaneously account for a multitude of inter-correlations in a more efficient and logically consistent way than more traditional approaches that use a series of bivariate analyses. In summary, our approach results in a more parsimonious model containing fewer significant parameters and a simpler interpretation of suggested pathways.

Acknowledgments and Contributions

This research was administered through the American Chemistry Council's (ACC) Long-Range Research Initiative, and it was jointly funded by the EPA's National Center for Environmental Research (NCER) and the ACC. The research on the PEPCOT study was

funded through STAR Grant R829363 to Battelle Memorial Institute.

Noel Cressie was Principal Investigator on the ACC-funded, five-year joint collaboration between The Ohio State University and Battelle in the area of BHM of human exposures. The research in this submission has been built up from published papers that considered univariate exposures, to the current submission on multivariate exposures. He guided the conceptual modeling and then the detailed statistical uncertainty quantification of each component of the hierarchical statistical model. He co-wrote all sections of the paper with coauthors and coordinated all revisions to this submission. Michele Morara contributed to developing the statistical model, developed the mathematical formalism, the MCMC sampling equations, implemented the MCMC sampler in C++, performed the EDA and MCMC analyses including choosing the prior hyper-parameters, produced the results, figures and tables, and co-wrote the abstract, modeling, EDA, results, and discussion sections of the paper. Bruce Buxton served as senior statistician, oversaw the EDA, provided the interpretation of the results, and co-wrote the EDA, results, and discussion sections of the paper. Nancy McMillan developed the initial univariate pathways Bayesian model that was then generalized to the multivariate case, served as senior Bayesian statistician, contributed to developing the multivariate model, and contributed to writing the paper. Warren Strauss provided the PEPCOT dataset, served as PEPCOT-data-analysis expert, performed initial EDA of the PEPCOT data, and co-wrote the introduction and data sections of the paper. Nancy Wilson provided the PEPCOT dataset, served as senior exposure expert, contributed to the interpretation of the results, and co-wrote the introduction and data sections of the paper. Kate Calder, Peter Craigmile, and Tom Santner, who are not coauthors, were involved in the formulation of hierarchical multivariate pathways models during the lifetime of the ACC contract.

The authors would like to thank the editors and referees for their helpful suggestions.

Bibliography

REFERENCES

Adams RD, Lupton D, Good AM, Bateman DN, 2009. UK childhood exposures to pesticides 2004-2007: A TOXBASE toxicovigilance study. *Archives of Disease in Childhood* **94**: 417–420.

Baccarelli A, Pfeiffer R, Consonni D, Pesatori AC, Bonzini M, Patterson DG, Bertazzi PA, Landi MT, 2005. Handling of dioxin measurement data in the presence of non-detectable values: Overview of available methods and their application in the Seveso chloracne study. *Chemosphere* **60**: 898–906.

Birnbaum LS, Fenton SE, 2003. Cancer and developmental exposure to endocrine disrupters. *Environmental Health Perspectives* **111**: 389–394.

Clayton CA, Pellizzari ED, Quackenboss JJ, 2002. National Human Exposure Assessment Survey: Analysis of exposure pathways and routes for arsenic and lead in EPA Region 5. *Journal of Exposure Analysis and Environmental Epidemiology* **12**: 29–43.

Craigmile PF, Calder CA, Li H, Paul R, Cressie N, 2009. Hierarchical model building, fitting, and checking: A behind-the-scenes look at a Bayesian analysis of exposure pathways. *Baysian Analysis* **4**: 1–62.

Cressie N, Buxton BE, Calder CA, Craigmile PF, Dong C, McMillan NJ, Morara M, Santner TJ, Wang K, Young GS, Zhang J, 2007. From sources to biomarkers: A hierarchical Bayesian approach for human exposure modeling. *Journal of Statistical Planning and Inference* **137**: 3361–3379.

Eskenazi B, Marks AR, Bradman A, Harley K, Barr DB, Johnson C, Morga N, Jewell NP, 2007. Organophosphate pesticide exposure and neurodevelopment in young Mexican-American families. *Environmental Health Perspectives* **115**: 792–798.

Gelman A, Carlin JB, Stern HS, Rubin DB, 2003. *Bayesian Data Analysis, Second Edition*. Chapman and Hall/CRC, Boca Raton.

Goldman L, 1995. Case studies of environmental risks to children. *Future Child* **5**: 27–33.

Hornung RW, Reed L, 1990. Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene* **5**: 46–51.

Landrigan PJ, Claudio L, Markowitz SB, Berkowitz GS, Brenner BL, Romero H, Wetmur JG, Matte TD, Gore AC, Godbold JH, Wolff MS, 1999. Pesticides and inner-city children: Exposures, risks, and prevention. *Environmental Health Perspectives* **107**: 431–437.

Landrigan PJ, Kimmel CA, Correa A, Eskenazi B, 2004. Children's health and the environment: Public health issues and challenges for risk assessment. *Environmental Health Perspectives* **112**: 257–265.

Lauritzen SL, 1996. *Graphical Models*. Oxford University Press, New York.

McMillan NJ, Morara M, Young GS, 2006. Hierarchical Bayesian modeling of human exposure pathways and routes. In *ASA Proceedings of the 2006 Joint Statistical Meetings, Section on Statistics and the Environment*, American Statistical Association, Alexandria, VA, pp. 2492-2503.

Mendola P, Selevan SG, Gutter S, Rice D, 2002. Environmental factors associated with a spectrum of neurodevelopmental deficits. *Mental Retardation and Developmental Disabilities Research Reviews* **8**: 188–197.

Morara M, 2008. Object oriented library for Markov chain Monte Carlo simulation. United States Patent No. 7409325, assignee: Battelle Memorial Institute.

Morara M, Buxton B, Strauss W, Wilson N, McMillan N, Cressie N, 2010. Multivariate hierarchical Bayesian analysis of human exposure pathways. Technical Report 839, Department of Statistics, The Ohio State University, Columbus, OH.

Perera FD, 1977. Environment and cancer: Who are susceptible? *Science* **278**: 1068–1073.

Robert CP, Casella G, 2004. *Monte Carlo Statistical Methods, Second Edition*. Springer-Verlag, New York.

Santner TJ, Craigmile PF, Calder CA, Paul R, 2008. Demographic and behavioral modifiers of arsenic exposure pathways: A Bayesian hierarchical analysis of NHEXAS data. *Environmental Science and Technology* **42**: 5607–5614.

Schettler T, 2001. Toxic threats to neurologic development of children. *Environmental Health Perspectives* **109**: 813–816.

Succop PA, Clark S, Chen M, Galke W, 2004. Imputation of data values that are less than a detection limit. *Journal of Occupational and Environmental Hygiene* **1**: 436–441.

Wigle DT, Arbuckle TE, Walker M, Wade MG, Liu S, Krewski D, 2007. Environmental hazards: Evidence for effects on child health. *Journal of Toxicology and Envirvironmental Health, Part B, Critical Reviews* **10**: 3–39.

Wilson NK, Strauss WJ, Iroz-Elardo N, Chuang JC, 2009. Exposures of preschool children to chlorpyrifos, diazinon, pentachlorophenol, and 2,4-diphenoxyacetic acid over 3 years from 2003 to 2005: A longitudinal study. *Journal of Exposure Science and Environmental Epidemiology* **20**: 546–558.

FIGURES

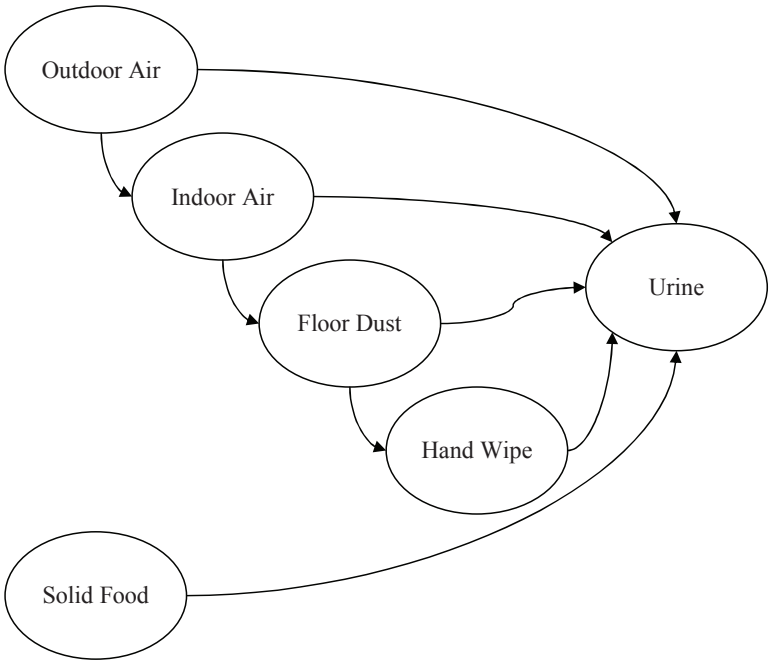

Figure 1. Exposure Pathways for CPF and TCP

TABLES

Table 1. Units of Measurement

Medium	Unit
urine	$nmol/mL$
solid food	$nmol/g$
hand wipe	$nmol/m^2$
floor dust	$nmol/g$
indoor air	$nmol/m^3$
outdoor air	$nmol/m^3$

Table 2. Measurement-Error Estimates

Medium	ϵ	Analyte
urine	11%	CPF, TCP
solid food	18%	CPF, TCP
hand wipe	6%	CPF, TCP
floor dust	9%	CPF, TCP
indoor air neutral	10%	CPF
indoor air acid	18%	TCP
outdoor air neutral	26%	CPF
outdoor air acid	17%	TCP

Table 3. Hyper-Parameters for Gamma and Wishart Prior Distributions

Medium	$\ln(1+\epsilon)$	ω	s^{ω}	r^{ω}	ν^{τ}	r^{τ}
urine	0.10	3.67×10^{2}	10^{18}	2.72×10^{15}	0	2.78×10^{-1}
solid food	0.17	1.46×10^{2}	10^{18}	6.85×10^{15}	0	6.99×10^{-1}
hand wipe	0.06	1.18×10^{3}	10^{18}	8.49×10^{14}	0	8.66×10^{-2}
floor dust	0.09	5.39×10^{2}	10^{18}	1.86×10^{15}	0	1.89×10^{-1}
indoor air CPF	0.10	4.40×10^{2}	10^{18}	2.27×10^{15}	0	2.32×10^{-1}
indoor air TCP	0.17	1.46×10^{2}	10^{18}	6.85×10^{15}	0	6.99×10^{-1}
outdoor air CPF	0.23	7.49×10^{1}	10^{18}	1.34×10^{16}	0	1.36
outdoor air TCP	0.16	1.62×10^{2}	10^{18}	6.16×10^{15}	0	6.29×10^{-1}

Table 4. MCMC Regression Estimates for the Urine pathway

Urine ∼ SolidFood + HandWipe + FloorDust + IndoorAir + OutdoorAir			
	Year 1	Year 2	Year 3
$\mu_{0,TCP}$	1.040 (-0.535, 2.627)	-1.861 (-4.811, 1.195)	-1.523 (-3.941, 0.992)
$\beta_{SolidFood,CPF}$	0.116 (-0.080, 0.298)	0.049 (-0.140, 0.235)	-0.013 (-0.154, 0.130)
$\beta_{SolidFood,TCP}$	**0.635** (0.362, 0.900)	0.285 (-0.035, 0.599)	**0.271** (0.124, 0.416)
$\beta_{HandWipe,CPF}$	0.065 (-0.120, 0.255)	-0.060 (-0.631, 0.587)	0.143 (-0.276, 0.529)
$\beta_{HandWipe,TCP}$	-0.196 (-0.531, 0.119)	-0.241 (-1.326, 0.882)	0.064 (-0.145, 0.273)
$\beta_{FloorDust,CPF}$	0.039 (-0.247, 0.323)	-0.196 (-0.664, 0.284)	-0.131 (-0.430, 0.184)
$\beta_{FloorDust,TCP}$	0.098 (-0.163, 0.350)	0.260 (-0.970, 1.491)	0.165 (-0.004, 0.320)
$\beta_{IndoorAir,CPF}$	0.226 (-0.234, 0.664)	0.210 (-0.229, 0.634)	0.166 (-0.147, 0.484)
$\beta_{IndoorAir,TCP}$	0.036 (-0.379, 0.472)	0.176 (-0.144, 0.498)	-0.429 (-0.861, 0.011)
$\beta_{OutdoorAir,CPF}$	0.026 (-0.200, 0.261)	-0.179 (-0.385, 0.019)	-0.084 (-0.389, 0.230)
$\beta_{OutdoorAir,TCP}$	0.174 (-0.103, 0.443)	**0.296** (0.029, 0.574)	**0.506** (0.039, 0.960)
σ^2_{CPF}	**0.620** (0.439, 0.892)	**0.854** (0.594, 1.215)	**0.427** (0.301, 0.610)

Table 5. MCMC Regression Estimates for the SolidFood pathway

	SolidFood		
	Year 1	Year 2	Year 3
$\mu_{0,CPF}$	**-8.475** (-8.889, -8.133)	**-8.028** (-8.343, -7.740)	**-8.204** (-8.535, -7.923)
$\mu_{0,TCP}$	**-4.662** (-4.875, -4.452)	**-4.801** (-4.948, -4.657)	**-4.719** (-4.943, -4.499)
σ^2_{CPF}	**2.363** (1.551, 3.782)	**1.949** (1.382, 2.858)	**1.809** (1.221, 2.795)
σ^2_{TCP}	**1.060** (0.783, 1.484)	**0.502** (0.379, 0.687)	**1.223** (0.924, 1.662)
$\rho_{CPF,TCP}$	**0.615** (0.434, 0.752)	**0.307** (0.095, 0.495)	**0.354** (0.140, 0.533)

Table 6. MCMC Regression Estimates for the HandWipe pathway

HandWipe $\sim$ FloorDust			
	Year 1	Year 2	Year 3
$\mu_{0,CPF}$	**-0.865** (-1.299, -0.456)	**-0.688** (-0.987, -0.383)	**-0.691** (-1.067, -0.298)
$\mu_{0,TCP}$	**-0.942** (-1.141, -0.756)	**-1.378** (-1.627, -1.177)	**-1.563** (-2.007, -1.205)
$\beta_{FloorDust,CPF}$	**0.679** (0.444, 0.945)	**0.422** (0.274, 0.569)	**0.748** (0.565, 0.979)
$\beta_{FloorDust,TCP}$	**0.376** (0.248, 0.511)	**1.017** (0.789, 1.309)	**0.305** (0.115, 0.542)
σ^2_{CPF}	**1.450** (0.922, 2.417)	**0.897** (0.621, 1.330)	**0.375** (0.224, 0.684)
σ^2_{TCP}	**0.466** (0.306, 0.727)	**0.355** (0.190, 0.714)	**0.948** (0.534, 1.756)
$\rho_{CPF,TCP}$	0.088 (-0.183, 0.348)	**0.727** (0.369, 0.895)	0.067 (-0.322, 0.435)

Table 7. MCMC Regression Estimates for the FloorDust pathway

FloorDust ~ IndoorAir			
	Year 1	Year 2	Year 3
$\mu_{0,CPF}$	**1.789** (0.854, 2.698)	**2.748** (1.918, 3.624)	**2.179** (1.194, 3.188)
$\mu_{0,TCP}$	**2.116** (1.009, 3.252)	**2.571** (1.553, 3.580)	**4.949** (2.834, 7.085)
$\beta_{IndoorAir,CPF}$	**0.781** (0.569, 0.990)	**0.989** (0.804, 1.180)	**1.016** (0.805, 1.232)
$\beta_{IndoorAir,TCP}$	**0.590** (0.382, 0.806)	**0.652** (0.457, 0.847)	**1.164** (0.782, 1.554)
σ^2_{CPF}	**1.109** (0.788, 1.643)	**0.759** (0.524, 1.154)	**1.163** (0.838, 1.705)
σ^2_{TCP}	**1.257** (0.886, 1.854)	**0.786** (0.555, 1.147)	**1.851** (1.246, 2.874)
$\rho_{CPF,TCP}$	**0.389** (0.143, 0.592)	**0.336** (0.080, 0.551)	**0.441** (0.220, 0.623)

Table 8. MCMC Regression Estimates for the IndoorAir pathway

IndoorAir $\sim$ OutdoorAir			
	Year 1	Year 2	Year 3
$\mu_{0,CPF}$	**-2.204** (-2.802, -1.598)	**-2.907** (-4.041, -1.761)	**-3.100** (-4.492, -1.678)
$\mu_{0,CPF}$	**-2.835** (-3.617, -2.037)	**-4.443** (-5.845, -3.028)	**-3.981** (-5.336, -2.537)
$\beta_{OutdoorAir,CPF}$	**0.312** (0.227, 0.397)	**0.206** (0.047, 0.368)	**0.215** (0.029, 0.404)
$\beta_{OutdoorAir,TCP}$	**0.332** (0.222, 0.447)	0.106 (-0.104, 0.318)	**0.237** (0.041, 0.447)
σ^2_{CPF}	**1.424** (1.085, 1.920)	**1.278** (0.968, 1.736)	**1.415** (1.048, 1.965)
σ^2_{TCP}	**1.408** (1.065, 1.908)	**1.040** (0.784, 1.415)	**0.681** (0.510, 0.937)
$\rho_{CPF,TCP}$	**0.913** (0.866, 0.945)	**0.638** (0.499, 0.745)	**0.862** (0.792, 0.909)

Table 9. MCMC Regression Estimates for the OutdoorAir pathway

	OutdoorAir		
	Year 1	Year 2	Year 3
$\mu_{0,CPF}$	**-6.574** (-6.913, -6.234)	**-6.938** (-7.201, -6.690)	**-7.396** (-7.666, -7.165)
$\mu_{0,TCP}$	**-6.959** (-7.297, -6.669)	**-6.616** (-6.806, -6.437)	**-6.888** (-7.052, -6.740)
σ^2_{CPF}	**2.865** (2.118, 3.989)	**1.449** (1.057, 2.094)	**1.225** (0.849, 1.840)
σ^2_{TCP}	**1.876** (1.304, 2.778)	**0.738** (0.530, 1.074)	**0.489** (0.339, 0.727)
$\rho_{CPF,TCP}$	**0.634** (0.479, 0.753)	**0.346** (0.139, 0.530)	**0.806** (0.693, 0.880)