Wiley StatsRef:
Statistics Reference Online

Quantifying Worker Exposures Using Bayesian Statistical Methods in Industrial Hygiene

By *Caroline P. Groth*[1, 2], *Sudipto Banerjee*[3], *Gurumurthy Ramachandran*[4], *Patricia A. Stewart*[5], *Mark R. Stenzel*[6], *and Mohammed A. Virji*[2]

Keywords: *industrial hygiene, Bayesian methods, workers, exposures, occupational health*

Abstract: Workplaces are dynamic and complex locations in which workers can be exposed to a variety of harmful chemicals. One of the industrial hygienists' primary responsibilities is the evaluation of workers' exposures to such chemicals. To do so, appropriate statistical modeling strategies are needed to accurately describe these exposures. Bayesian methods provide a useful framework to quantify and characterize workers' exposures in the complex work environment. This article presents several Bayesian modeling strategies to improve the understanding of workers' exposures while accounting for measurements below the analytic limit of detection. These methods allow industrial hygienists to characterize exposures by summarizing measured exposures, thereby investigating relationships between exposures to multiple chemicals, statistically identifying and describing exposure determinants, and estimating exposures when measurements are unavailable.

1 Introduction

Industrial hygienists (IHs) are responsible for anticipating, recognizing, evaluating, and controlling hazards in the workplace. This endeavor requires a good understanding of the sources of hazards, exposures from these sources, and the path between the two. Due to the dynamic nature and interactions of the workplace, workforce, and work practices, occupational exposures are highly variable, and their evaluation is very complex. In performance of their responsibilities[1], IHs, after collecting exposure measurements, must summarize those measurements to compare with occupational standards or recommendations.

[1]WVU School of Public Health, West Virginia University, One Medical Center Drive, Morgantown, WV, USA
[2]National Institute for Occupational Safety and Health, Respiratory Health Division, Morgantown, WV, USA
[3]UCLA Fielding School of Public Health, University of California – Los Angeles, Los Angeles, CA, USA
[4]Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
[5]Stewart Exposure Assessments, LLC, Arlington, VA, USA
[6]Exposure Assessment Applications, LLC, Arlington, VA, USA

DOI: 10.1002/9781118445112.stat08376

To perform the exposure assessment, IHs might also need to identify determinants of exposure, that is, factors affecting exposure concentrations[2, 3] that allow for evaluation of primary sources of exposure for purposes of exposure reduction. IHs may collect quantitative measurements or use deterministic or mathematical models to estimate exposure levels. Deterministic and mathematical modeling can estimate exposures quickly for less cost and only requires a limited number of measurements to validate the model[3, 4]. Imputing missing exposures from relationships between one or more measured chemicals and a chemical not measured can also reduce the number of measurements that need to be collected.

Particular challenges in assessing exposures, however, are small sample size, censored data (data below the limit of detection (LOD); the LOD is a threshold below which the particular analytical method cannot detect with precision; *see* **Detection Limits**), autocorrelation (*see* **Autocorrelation Function**), and large exposure variability, all of which can result in large uncertainties in characterizing exposures. Bayesian methods offer approaches to simultaneously incorporate multiple sources of information such as professional judgment and modeling results with measured exposure data and integrate over the censored part of the distribution to improve accuracy and reduce uncertainty. In addition, these methods offer the desired summary metrics such as mean, standard deviation, and various quantiles such as the 95th percentile and, most importantly, quantify uncertainty around these metrics to truly appreciate the possible range of the metrics of interest.

This article first describes Bayesian terms and concepts and presents a basic statistical structure. It then discusses the various challenges and presents modifications of that statistical structure to reflect the analysis of measurement data. Examples are provided in the supplementary materials to demonstrate how Bayesian methods were used to improve exposure estimation in various settings.

2 Statistical Methods

2.1 Bayesian Methods

In the supplementary materials, we provide a brief discussion of Bayesian statistical principles including definitions of the Bayesian terms used here (*see* **Bayesian Inference**; **Bayes Linear Analysis**; **Bayesian Robustness**; **Conjugate Families of Distributions**; **Decision Theory**; **Distributional Inference**; **Bayesian Model Selection**; **Hierarchical Models – Theory**; **Jeffreys' Prior Distribution**; **Prior Distributions**; **Bayesian Analysis and Markov Chain Monte Carlo Simulation**; **Credible Intervals**; **Bayesian Inference of Markov Processes**; **Posterior Distribution: Introduction**; **Bayesian Methods**; **Prior Distribution Elicitation**).

2.2 Modeling Single Chemical Exposures

2.2.1 Modeling when all measurements are above the limit of detection

If we are interested in modeling exposures of a single group of related measurements, for example, an exposure group, the Bayesian model requires only an intercept in the regression framework (*see* **Regression, Bayesian**). We model the natural log of the chemical measurements (*see* **Lognormal Distribution – Basic**) to maintain the model requirement of normality based on the general observation that environmental and occupational measurements are log normally distributed.

Let Y_i be the natural log of the measurement i, where $i = 1, \ldots, n$. Then, assuming that all measurements are above the LOD and are independent and identically distributed (iid), we can write $Y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ to

Wiley StatsRef: Statistics Reference Online, © 2014 – 2022 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08376

denote that the natural log of the measurements is approximately normally distributed with mean μ and variance σ^2. This represents our likelihood (*see* **Likelihood**) and can be rewritten in terms of a product of the following form: $\prod_{i=1}^{n} N(Y_i \mid \mu, \sigma^2)$.

Our Bayesian specification is completed by introducing priors (*see* **Prior Distributions**), often specified to be weakly informative to allow the data to drive the inference. For example, we specify the prior for μ as $\mu \sim N(a, b)$ with mean a and variance b; larger values of b correspond to weaker priors. The prior on the variance, σ^2, is set to be an inverse gamma distribution such that the variance is positive (*see* **Normal-Gamma (Prior) Density**). Using the notation of Gelman *et al.*[5] with parameters c and d, this prior is written as $\text{IG}(\sigma^2 \mid c, d)$. These yield the joint distribution of the parameters and the data, assuming that all measurements are above LOD, which is proportional to the posterior distribution,

$$p(\mu, \sigma^2 \mid \{Y_1, Y_2, \ldots, Y_n\}) \propto N(\mu \mid a, b) \times \text{IG}(\sigma^2 \mid c, d) \times \prod_{i=1}^{n} N(Y_i \mid \mu, \sigma^2) \quad (1)$$

We conduct Bayesian inference by sampling from the above distribution; this model can be developed easily using Markov Chain Monte Carlo (MCMC) or other direct algorithms available in many software languages (see supplements for details). The posterior samples for $\{\mu, \sigma^2\}$ are directly plugged into the formulas for the geometric mean (GM; e^{μ}) and the geometric standard deviation (GSD; e^{σ}) to immediately produce the corresponding posterior samples for GM and GSD. From the GM, GSD, and the number of observations (n) we can calculate the posterior estimates of the arithmetic mean (AM) and the 95th percentile.

2.2.2 Modeling when some measurements are below the limit of detection

Oftentimes, a measurement result is below the analytic LOD. In such cases, we do not know the actual value, other than it falls between the LOD and 0 (known as left censored; *see* **Censoring, Left**). To include censored data in our single chemical Bayesian analysis, we introduce a censored set called $C_Y = \{\forall i : Y_i \leq \text{LOD}_i\}$ designating "for all" ($\forall$) Y_i below the particular LOD_i for that observation. A specific LOD is associated with each measurement and may vary from one measurement to another. We also introduce an observed set $O_Y = \{\forall i : Y_i > \text{LOD}_i\}$. Then, we modify (1) to accommodate censoring as

$$p(\mu, \sigma^2 \mid \{Y_1, Y_2, \ldots, Y_n\}) \propto N(\mu \mid a, b) \times \text{IG}(\sigma^2 \mid c, d) \times \prod_{i \in O_Y} N(Y_i \mid \mu, \sigma^2) \times \prod_{i \in C_Y} \Phi\left(\frac{\text{LOD}(Y_i) - \mu}{\sigma}\right) \quad (2)$$

where the products over i are taken over the sets O_Y and C_Y according to the corresponding measurement being observed or censored, respectively, and $\Phi(Q)$ denotes the standard normal cumulative density function of Q. This framework allows for multiple values (from applying MCMC methods, see the supplementary materials and **Markov Chain Monte Carlo Methods**; **Markov Chain Monte Carlo (MCMC)**; **Markov Chain Monte Carlo Methods**; **Markov Chain Monte Carlo Algorithms**; **Bayesian Analysis and Markov Chain Monte Carlo Simulation**; **Markov Chain Monte Carlo, Introduction**; **Markov Chain Monte Carlo, Convergence and Mixing in**) to serve as replacements for each censored measurement, quantifying uncertainty in the censored measurement. This uncertainty is incorporated into the estimates of the AM, GM, GSD, and the 95th percentile.

2.2.3 Accounting for within-worker variability/repeated measures

Among the sources of exposure variability for a worker are the between- and within-worker variability (*see* **Repeated Measures**). If many measurements are available on a set of workers and there is a sufficient

DOI: 10.1002/9781118445112.stat08376

number of measured workers (and information to know which measurement belongs to which worker), then we should account for within-subject variability. Not accounting for this variability can result in a misleadingly low variance, which can result in a lower value of the mean and 95 percentile than is true. If a worker performs the same job assignment every day for the same amount of time, the underlying concept of modeling these two sources of variability may be appropriate, even with limited data. However, when workers perform multiple job assignments over days with different exposure levels, calculating within-worker variability from limited data, that is, that do not capture all job assignments, will result in an inappropriately low estimate of within-worker variability.

To evaluate within-subject variability, we can simply add a random effect (*see* **Fixed and Random Effects**) to the model framework above. Let Y_{ij} be the ith observation on subject j, and γ_j be the random effect with 0 mean and variance τ^2 (defined as between-subject variability). Then, the joint (posterior) distribution becomes (is proportional to) the following:

$$N(\gamma_j \mid 0, \tau^2) \times N(\mu \mid a, b) \times \mathrm{IG}(\sigma^2 \mid c, d) \times \mathrm{IG}(\tau^2 \mid c, d) \\ \times \prod_{i \in O_Y} N(Y_{ij} \mid \mu + \gamma_j, \sigma^2) \times \prod_{i \in C_Y} \Phi\left(\frac{\mathrm{LOD}(Y_i) - (\mu + \gamma_j)}{\sigma}\right) \tag{3}$$

If limited information (i.e., few measurements and/or high censoring) is available for each subject, or if few subjects are available in a group, variances may greatly inflate, leading to high GSDs, resulting in inappropriately high AMs, and 95th percentiles. Identifying informative priors (e.g., restricting the GSD to particular ranges) may help to restrict variances. See the supplementary materials for an example.

2.3 Modeling Multiple Chemical Exposures/Mixtures

The previous models have assumed that we are interested in only one chemical exposure. However, often chemicals are components of mixtures to which workers can be exposed. Within a chemical mixture, the concentrations of the chemicals may be correlated. Moreover, the chemicals may be characterized by different amounts of censoring. It may be possible to use one chemical in the mixture to estimate exposures to a second chemical that may have more censored measurements.

Groth *et al*.[6, 7] extended the framework in (2) to model one or more chemical exposures simultaneously using other chemical exposures in the model as predictors. For a single predictor, let Y_i be the natural log of the chemical being estimated, and X_i be the natural log of a measurement of another chemical. Generally, the more highly observed chemical (i.e., the one with the lowest percentage of censoring) should be used as the predictor because it increases the amount of information provided to the model compared to using a more highly censored chemical. Define the observed sets (O_Y and O_X) and censored sets (C_Y and C_X) as earlier for Y and X. Also, let μ_X and σ_X^2 be the mean and variance of X, respectively, while each Y_i can be modeled conditional on X_i with mean $\beta_0 + \beta_1 X_i$ and common variance $\sigma_{Y|X}^2$. Then, assuming weakly informative priors on the regression coefficients (e.g., $N(a, b)$), means, and variances (IG(c,d)), we arrive at the following distribution:

$$N(\mu_X \mid a, b) \times \mathrm{IG}(\sigma_X^2 \mid c, d) \times N(\boldsymbol{\beta} \mid a\mathbf{1}, b\mathbf{I}) \times \mathrm{IG}(\sigma_{Y|X}^2 \mid c, d) \times \prod_{i \in O_X} N(X_i \mid \mu, \sigma_X^2) \\ \times \prod_{i \in O_Y} N(Y_i \mid \beta_0 + \beta_1 X_i, \sigma_{Y|X}^2) \times \prod_{i \in C_X} \Phi\left(\frac{\mathrm{LOD}_i(X) - \mu_X}{\sigma_X}\right) \times \prod_{i \in C_Y} \Phi\left(\frac{\mathrm{LOD}_i(Y) - (\beta_0 + \beta_1 X_i)}{\sigma_{Y|X}}\right) \tag{4}$$

Groth *et al*.[7] further extended this framework to include multiple chemical predictors. For more information, see the supplementary materials (and *see* **Multivariate Analysis, Bayesian: Overview with Examples**).

Wiley StatsRef: Statistics Reference Online, © 2014–2022 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08376

2.4 Identifying and Describing Exposure Determinants

Another important area of exposure evaluation is the identification of exposure determinants. Determinants can be used to identify groups of comparably exposed workers, such that the distribution of the group's measurements represents the exposure distribution of all members of the group, thus obviating the need for taking measurements on all workers. In addition, using determinants to identify groups of measurements that are expected to be more comparable than all available measurements in a dataset is likely to result in more accurate summary statistics because the statistics use measurements that are likely more representative (as determined by the determinants) of the group than an entire body of measurements.

Modeling a single exposure determinant or multiple exposure determinants typically proceeds from (2) or (3), but with the appropriate measurements identified with their determinants. Thus, we expand the term by adding regression coefficients, $\beta_0, \beta_1, \ldots, \beta_k$, a coefficient associated with each determinant. Customarily, wide normally distributed priors are typically placed on the regression coefficients to allow them to be primarily estimated from the data. For information on extending this to evaluate the impact of multiple determinants, see the supplementary materials.

2.5 Estimating Missing Worker Exposures

The effect on the within-worker variability of not having taken measurements on days with unique exposures is described in Section 2.2.3. In some cases, however, measurements for one chemical may be collected that generally reflect the actual exposure variability, but such measurements are not available for all of the same exposure scenarios for another chemical (missing data, *see* **Missing Data and Imputation**). In this case, Bayesian statistics can predict the estimates of the second chemical using the relationship (e.g., the correlation) between the two chemicals on days when both were measured and then apply that relationship to the days of missing data. An example of this approach is described in the supplementary materials.

2.6 Prior Information

Most of the examples described in this article used weakly informative prior information. However, for many exposure scenarios it may be beneficial to add stronger prior information to the model to help inform exposures. Additional information could be drawn from a variety of sources including from expert judgment, other group exposures/relationships in the study population, or even from other studies.

Incorporating expert judgment (*see* **Expert Judgment**) as prior information has been shown to positively impact exposure estimation in industrial hygiene. Ramachandran *et al*.[8] demonstrated that IHs without an intimate familiarity with the workplace can still provide useful judgments that can be accommodated as prior information. Specifically, even these judgments from IHs who were not intimately familiar with the workplace were shown to lead to strong estimates of exposure, even in the presence of sparse exposure data. Banerjee *et al*.[9] and Vadali *et al*.[10] further described how priors formed from expert judgments can be used alongside exposure data to help inform decisions and determine accurate epidemiologic exposure classification. As described in Ramachandran[11] and Zhang *et al*.[12], expert information and historical plant information can also be incorporated with physical models (such as a two-zone model) information to develop more reasonable exposure estimates than not taking this approach.

DOI: 10.1002/9781118445112.stat08376

One common concern with Bayesian methods is that if the priors are too informative and also incorrect, the priors may lead to inaccurate conclusions. However, researchers have identified methods to help mitigate this concern. For an example of this, see the supplementary materials.

2.7 Exposure–Health Outcome Relationships

The purpose of exposure assessment may be for intervention of some kind (e.g., engineering controls, protective equipment, and training), but exposure assessments are also used in epidemiologic studies. In many epidemiological studies, job exposure matrices (JEMs, matrices of exposures for workers in different scenarios; *see* **Job-Exposure Matrices**) are used to assign exposure estimates derived from measurements defined by a set of exposure determinants. All study subjects that have the same determinants are then assigned a single exposure estimate, allowing the investigation of the exposure–disease relationships (*see* **Statistics in Environmental Health Risk Assessment**). JEMs may be entirely based on professional judgment of a group of experts or on sparce historical monitoring data available. Bayesian methods hold tremendous promise for estimating exposures in these situations by facilitating data from multiple sources such as professional judgment, modeling, and limited monitoring to be integrated to maximize the utility of all available information[11].

These methods also allow for development of exposure estimates of various exposure parameters with credible intervals, which indicate the estimation uncertainty. Historically, most epidemiologic research has not accounted for uncertainty intervals (or their equivalent, confidence intervals) because of the inability to account for these intervals using standard statistical modeling strategies. However, with the advancement of Bayesian statistics, it has become easier to incorporate the exposure models with estimates of uncertainty and the disease risk models. For information on potential benefits, see the supplementary materials.

2.8 Other Extensions

These models can also be extended to incorporate spatial information, time-series data, and hierarchical structures. See the supplementary materials for more details.

3 Conclusion

In this article, we have described the different ways IHs can use Bayesian modeling strategies that account for measurements below the LOD to estimate exposures among workers. Bayesian methods are flexible frameworks, allowing researchers to quantify exposures in the complex situations often present in work environments. We have presented several potential frameworks in which IHs can build more complex models to account for various needs in exposure assessment. These methods allow IHs to compare exposure distributions between various groups of workers based on exposure determinants, investigate the relationships between various exposures, statistically identify and describe determinants of exposure, and estimate worker exposures when measurements are not available. With the help of these Bayesian methods, IHs can describe and characterize the dynamic workers' exposure in greater detail than before and provide additional information about uncertainty to epidemiologic studies for further investigation of exposures–disease outcomes. The hope is that these new data-based research methods will lead to development of new safety measures and provide protection for workers in the future.

Wiley StatsRef: Statistics Reference Online, © 2014–2022 John Wiley & Sons, Ltd.

DOI: 10.1002/9781118445112.stat08376

Acknowledgement

Sudipto Banerjee would like to acknowledge funding from NIH 1R01ES030210 and NIH R01ES027027-01A. Patricia Stewart and Mark Stenzel acknowledge funding from the National Institute of Environmental Health Sciences contract HHSN273201600003I. The findings and conclusions in this article are those of the author(s) and do not necessarily represent the official position of the National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention.

Supplementary Materials

Additional supporting information may be found online in the Supporting Information section in the HTML rendition of this article. Quantifying Worker Exposures Using Bayesian Statistical Methods in Industrial Hygiene

Related Articles

Bayesian Statistics in Quantitative Risk Assessment; **Occupational Health and Medicine**; **Occupational Epidemiology, Statistics in healthy worker effect+**; **Personal Exposure Monitoring**; **Personal Exposure Monitoring**; **Exposure Assessment**; **Exposure Rating**; **Exposure Effect**; **Synergy of Exposure Effects**; **Job-Exposure Matrices**; **Sampling in Industrial Standards**

References

[1] Stewart, P. and Stenzel, M. (2000) Exposure assessment in the occupational setting. *Appl. Occup. Environ. Hyg.*, **15** (5), 435–444.
[2] Virji, M.A. and Armstrong, T. (2009) Statistical models of exposure determinants, in *Mathematical Models for Estimating Occupational Exposure to Chemicals*, 2nd edn (eds C.B. Keil, C.E. Simmons, and T.R. Anthony), American Industrial Hygiene Association Press, Fairfax, VA, pp. 71–80.
[3] Ramachandran, G. (2008) Toward better exposure assessment strategies–the new niosh initiative. *Ann. Occup. Hyg.*, **52** (5), 297–301.
[4] Arnold, S.F., Stenzel, M., Drolet, D., and Ramachandran, G. (2016) Using checklists and algorithms to improve qualitative exposure judgment accuracy. *J. Occup. Environ. Hyg.*, **13** (3), 159–168.
[5] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2013) *Bayesian Data Analysis*, vol. 2, Taylor & Francis, Boca Raton, FL.
[6] Groth, C., Banerjee, S., Ramachandran, G. *et al*. (2017) Bivariate left-censored Bayesian model for predicting exposure: preliminary analysis of worker exposure during the *Deepwater Horizon* oil spill. *Ann. Work Expo. Health*, **61** (1), 76–86.
[7] Groth, C.P., Banerjee, S., Ramachandran, G. *et al*. (2018) Multivariate left-censored Bayesian modeling for predicting exposure using multiple chemical predictors. *Environmetrics*, **29** (4), e2505.
[8] Ramachandran, G., Banerjee, S., and Vincent, J.H. (2003) Expert judgment and occupational hygiene: application to aerosol speciation in the nickel primary production industry. *Ann. Occup. Hyg.*, **47** (6), 461–475.
[9] Banerjee, S., Ramachandran, G., Vadali, M., and Sahmel, J. (2014) Bayesian hierarchical framework for occupational hygiene decision making. *Ann. Occup. Hyg.*, **58** (9), 1079–1093.
[10] Vadali, M., Ramachandran, G., and Mulhausen, J. (2009) Exposure modeling in occupational hygiene decision making. *J. Occup. Environ. Hyg.*, **6** (6), 353–362.
[11] Ramachandran, G. (2001) Retrospective exposure assessment using Bayesian methods. *Ann. Occup. Hyg.*, **45** (8), 651–667.
[12] Zhang, Y., Banerjee, S., Yang, R. *et al*. (2009) Bayesian modeling of exposure and airflow using two-zone models. *Ann. Occup. Hyg.*, **53** (4), 409–424.

DOI: 10.1002/9781118445112.stat08376