

## Shared Uncertainty in Measurement Error Problems, with Application to Nevada Test Site Fallout Data

Yehua Li,<sup>1</sup> Annamaria Guolo,<sup>2</sup> F. Owen Hoffman,<sup>3</sup> and Raymond J. Carroll<sup>4,\*</sup>

<sup>1</sup>Department of Statistics, University of Georgia, Athens, Georgia 30602, U.S.A.

<sup>2</sup>Department of Statistics, University of Padova, Via Cesare Battisti, 241, 35121, Padova, Italy

<sup>3</sup>SENES Oak Ridge, Center for Risk Analysis, 102 Donner Drive, Oak Ridge, Tennessee 37830, U.S.A.

<sup>4</sup>Department of Statistics, Texas A& M University, 3143 TAMU, Texas A& M University,  
College Station, Texas 77843-3143, U.S.A.

*email:* carroll@stat.tamu.edu

**SUMMARY.** In radiation epidemiology, it is often necessary to use mathematical models in the absence of direct measurements of individual doses. When complex models are used as surrogates for direct measurements to estimate individual doses that occurred almost 50 years ago, dose estimates will be associated with considerable error, this error being a mixture of (a) classical measurement error due to individual data such as diet histories and (b) Berkson measurement error associated with various aspects of the dosimetry system. In the Nevada Test Site (NTS) Thyroid Disease Study, the Berkson measurement errors are correlated within strata. This article concerns the development of statistical methods for inference about risk of radiation dose on thyroid disease, methods that account for the complex error structure inherent in the problem. Bayesian methods using Markov chain Monte Carlo and Monte-Carlo expectation–maximization methods are described, with both sharing a key Metropolis–Hastings step. Regression calibration is also considered, but we show that regression calibration does not use the correlation structure of the Berkson errors. Our methods are applied to the NTS Study, where we find a strong dose–response relationship between dose and thyroiditis. We conclude that full consideration of mixtures of Berkson and classical uncertainties in reconstructed individual doses are important for quantifying the dose response and its credibility/confidence interval. Using regression calibration and expectation values for individual doses can lead to a substantial underestimation of the excess relative risk per gray and its 95% confidence intervals.

**KEY WORDS:** Bayes; Berkson error; Classic error; Correlation; Latent variables; MCMC; Measurement error; Monte Carlo EM; Objective Bayes; Radiation epidemiology; Regression calibration; Thyroid disease.

### 1. Introduction

There are many studies relating radiation exposure to disease. We focus here on data from the Nevada Test Site (NTS) Thyroid Disease Study. In the Nevada study, 2491 individuals who were exposed to radiation as children were examined for thyroid disease. The primary radiation exposure to the thyroid glands of these children came from the ingestion of milk and vegetables contaminated with radioactive isotopes of iodine. The idea of the study was to relate various thyroid disease outcomes to radiation exposure to the thyroid. The original version of this study was described by Stevens et al. (1992), Kerber et al. (1993), and Simon et al. (1995). Recently, the dosimetry for the study was redone (Simon et al., 2006), and the study results were reported (Lyon et al., 2006).

The estimation of individual radiation dose that occurred 50 years in the past is well known to be subject to large uncertainties, especially when mathematical models are employed due to the absence of direct measurements of the concentrations of radioactivity in foods or within the thyroid gland of individuals. There are many references on this subject, a good introduction to which is given by Ron and Hoffman (1999) and

the many statistical papers in that volume. Various statistical papers (Reeves et al., 1998; Schafer et al., 2001; Mallick, Hoffman, and Carroll, 2002; Stram and Kopecky, 2003; Lubin et al., 2004; Schafer and Gilbert, 2006) describe measurement error properties and analysis in this context. What is typical in these studies is to build a large dosimetry model that attempts to convert the known data, e.g., about the above-ground nuclear tests, to radiation actually absorbed into the thyroid. Dosimetry calculations for individual subjects were based on age at exposure, gender, residence history, whether as a child the individual was breast-fed, and a diet questionnaire filled out by the parent focusing on milk consumption and vegetables. The data were then input into a complex model and for each individual, the point estimate of thyroid dose (the arithmetic mean of a lognormal distribution of dose estimates) and an associated error term (the geometric standard deviation) for the measurement error were reported.

Generally, the authors engaged in dose reconstruction using mathematical models conclude that radiation doses are estimated with a combination of Berkson measurement error and the classical type of measurement error. The type of model,

pioneered by Reeves et al. (1998), in the log scale of dose says that true log dose  $X$  is related to observed or calculated log dose  $W$  by a latent intermediate  $L$  via

$$X = L + U_b, \quad (1)$$

$$W = L + U_c, \quad (2)$$

where  $U_b$  is the Berkson uncertainty and  $U_c$  is the classical uncertainty. If there is no classical uncertainty, then  $W = L$  and we get the pure Berkson model  $X = W + U_b$ . If there is no Berkson uncertainty, then  $X = L$  and we get the classical additive error model  $W = X + U_c$ .

The starting point of this article is the following. For reasons that we describe in more detail in Section 2, the Berkson measurement errors produced as estimates by dose reconstruction models are now generally thought to be correlated across individuals. The origin of correlated Berkson measurement errors for the NTS Study are described in Section 7.2 of Mallick et al. (2002). In a very different context, with a different dosimetry system, Stram and Kopecky (2003) also describe correlated Berkson errors.

This article is organized as follows. Section 2 describes the NTS Study data in more detail, and gives a description of the modeling of mixtures of classical and correlated Berkson errors.

The only paper of which we are aware that considers the analysis of data subject to a combination of classical measurement error and correlated Berkson errors is Mallick et al. (2002). Their analysis, given in a brief concluding paragraph, was purely Bayesian, but it gave no details as to how to implement the Markov chain Monte Carlo (MCMC) methods, sensitivity of the results to prior specification, etc. Implementation is nontrivial, because the correlation of the Berkson errors means that the latent true log doses  $X$  are correlated and reasonably high dimensional, e.g., as a worst case, of dimension the same as the sample size, namely  $>2000$ . In Section 3, we describe methods for implementing Bayesian MCMC calculations in this context.

In Section 4, we describe frequentist methods for this problem, a topic apparently not previously considered in detail. Stram and Kopecky (2003) suggest a form of Monte Carlo maximum likelihood, but they do not implement the idea and they conclude that it is not even clear whether the method will work. We first show that the standard measurement error technique, regression calibration (Carroll et al., 2006, Chapter 4) is easily implemented if the Berkson errors are independent, but in fact regression calibration fails to account for the correlated case, see Section 4.1.2. This led us to implement a Monte Carlo expectation–maximization algorithm (MCEM; McCulloch, 1997), with the same Metropolis step that we used for the Bayesian analysis.

Section 5 revisits the NTS Study, focusing on the new finding of Lyon et al. (2006) of a significant relationship between dose and thyroiditis and thyroid neoplasm. As background, thyroiditis is an inflammation of the thyroid gland, while a thyroid neoplasm is essentially a tumor, albeit possibly benign. We also discuss the robustness and identifiability of the model in Section 5. In Section 6, we discuss the general effects of the shared Berkson error in two different models and give simulation results to illustrate such effects. Concluding

remarks are given in Section 7. All technical details are available in Web Appendix A.

## 2. The Nevada Test Site Data

### 2.1 Data Structure and Dose-Response Model

The NTS Study for thyroiditis consisted of 2491 individuals, with roughly one-half being female. There were 123 cases of thyroiditis, 103 being women, reflecting their much greater risk even in the absence of excess radiation. There were 20 cases of thyroid neoplasms.

As described just below, we split the data into strata. For person  $i$  in stratum  $s$ , we have binary responses  $Y_{is}$  for thyroiditis, observed covariates  $Z_{is}$ , and unobserved true log dose  $X_{is}$ . The regression model is

$$\text{pr}(Y_{is} = 1 | Z_{is}, X_{is}) = H \left[ Z_{is}^T \beta + \log\{1 + \theta \exp(X_{is})\} \right], \quad (3)$$

where  $H(x) = 1/\{1 + \exp(-x)\}$  is the logistic distribution function. In Lyon et al. (2006),  $Z_{is}$  consists of a vector of ones plus gender. The parameter  $\theta$  is the excess relative risk per gray (ERR/Gy).

As in Mallick et al. (2002), define  $\sigma_{is,B}^2$  and  $\sigma_{is,C}^2$  to be the Berkson and classical variances, where the total uncertainty for an individual is their sum  $\sigma_{is,B}^2 + \sigma_{is,C}^2 = \sigma_{is,\text{total}}^2$ . In the NTS Study, the total variance/uncertainty due to the dose-estimation algorithms is known but the individual components are not. This total variance/uncertainty is derived in Simon et al. (2006, Sections 7 and 8), see also Simon et al. (1995). As Simon et al. (2006) state: “. . . uncertainty was propagated by a combination of Monte Carlo and analytic error-propagation techniques, and the overall distribution of uncertainty for each individual was assumed to be log normal.”

We thus set  $\sigma_{is,C}^2 = \gamma_C \sigma_{is,\text{total}}^2$ , where  $\gamma_C$  is independent of the person and uniformly distributed on the interval from  $a_c = 0.3$  to  $b_c = 0.5$ . This is a slightly different range than that employed by Mallick et al. (2002), and reflects a smaller amount of classical uncertainty than used by them. In addition, the classical measurement errors  $U_{is,C}$  are assumed to be independent normal random variables with mean 0.0 and variances  $\sigma_{is,C}^2$ .

The extent of the correlation of the Berkson errors is not known, and we thus relied on Dr. Hoffman, who is an expert in dose reconstruction of individuals exposed from fallout from the NTS, to estimate the correlation within each stratum by a range. Within a single stratum, we let the Berkson errors  $\{U_{is,B}\}_i$  be normal random variables with means zero, variances  $\sigma_{is,B}^2$ , and a common correlation  $\rho_s$ , whose range is specified. Further comments on stratification and setting ranges for the correlations are provided in Section 2.2.

The dosimetry model was the same as equations (1) and (2), but applied separately to each stratum, so that  $L_{is} = \text{Normal}(\mu_s, \sigma_s^2)$ , log true dose  $X_{is} = L_{is} + U_{is,B}$  and observed log dose  $W_{is} = L_{is} + U_{is,C}$ . The Berkson errors in stratum  $s$  were as described above, to have common correlation  $\rho_s$  on the interval  $[c_s, d_s]$ .

Finally, we note that the classical error is assumed to be independent across subjects, and Berkson error can be correlated within each stratum.

2.2 Stratification and Parameter Specifications

Individuals in the NTS Study were classified into 40 strata depending on (a) the type of exposure; (b) the location of the exposure; and (c) the source of milk that the individual drank. The type of exposure means either the nuclear test called Shot Harry, the nuclear test called Shot Smokey, and all others. The locations were either Washington County Utah, Lincoln County Nevada, Graham County Arizona, other places in Utah, other places in the United States, or all other places. Four of the strata were empty, and one stratum had only three people and was absorbed into a neighboring stratum.

The algorithm that Dr. Hoffman used for setting ranges for the correlation of Berkson uncertainties involved detailed issues of the location of children, the main cause of their radiation exposure, and how they received their milk.

For example, residents of Washington County in Utah experienced similar radiation deposition patterns from similar shots. For most, the dominant source of exposure was fallout from Shot Harry in mid-May of 1953. We thus anticipated higher correlations in uncertainty for those who were exposed in Washington County compared to those from other counties in Utah, who received their doses more from a variety of sources.

In Washington County, some residents received their milk supply from cows in the backyard, while others received their milk supply by local commercial operations. We anticipated more variability between individual cows producing milk than for larger dairies and milk sheds. Thus the degree of “shared” uncertainty will be less for consumers of backyard cows’ milk than for those consuming store-purchased milk.

Thus, Dr. Hoffman set ranges of 50–70% for the shared Berkson correlation of Washington County residents who received their milk from local commercial sources, 40–60% for local backyard milk sources, and roughly 20–40% otherwise. For those from other counties in Utah, who would have lower shared Berkson uncertainty, the ranges are 20–40%, 20–30%, and 20–30%.

In terms of the classical uncertainty, classical uncertainties occur in the location-specific deposition measurements, in the interview process used to obtain information about farming practices from individual milk producers and to obtain dietary, milk source, and residence histories from individual members of the epidemiological cohort. Classical uncertainties also exist in the measured data published in the literature that were used to develop probability distributions for environmental, food chain transfer, and dose conversion coefficients used in the pathway and dose models, as described by Simon et al. (2006) and by Schafer and Gilbert (2006). We selected a range of 30–50% of the uncertainty being classical as a reasonable statement of the importance of these direct measurements on the assigned doses.

3. MCMC Methodology

In this section we outline our methods. Algebraic details of most of the complete conditional distributions and the methods of sampling are given in Web Appendix A.

3.1 Complete Data Likelihood

Let  $n_s$  be the size of the  $s$ th stratum. Let  $\Sigma_s(\rho_s, \gamma_C)$  be the  $n_s \times n_s$  covariance matrix that has

common correlation  $\rho_s$  and variances  $\sigma_{i_s,B}^2 = (1 - \gamma_C)\sigma_{i_s,total}^2$ . Let  $\mathcal{X}_s = (X_{1s}, \dots, X_{n_s s})^T$ ,  $\mathcal{L}_s = (L_{1s}, \dots, L_{n_s s})^T$ , and  $\pi(\beta)\pi(\theta)\prod_s \{\pi(\mu_s)\pi(\sigma_s^2)\}$  be the prior distributions. Then the complete likelihood takes the form

$$\begin{aligned} &\pi(\beta)\pi(\theta)I(a_c \leq \gamma_C \leq b_c) \\ &\times \prod_s \{\pi(\mu_s)\pi(\sigma_s^2)\} \times \prod_s I(c_s \leq \rho_s \leq d_s) \\ &\times \prod_{i,s}^s (H[Z_{is}^T\beta + \log\{1 + \theta \exp(X_{is})\}])^{Y_{is}} \\ &\times (1 - H[Z_{is}^T\beta + \log\{1 + \theta \exp(X_{is})\}])^{1-Y_{is}} \\ &\times \prod_s (\sigma_s^2)^{-n_s/2} \exp\left\{-\frac{1}{2}(\sigma_s^2)^{-1} \sum_{i=1}^{n_s} (L_{is} - \mu_s)^2\right\} \\ &\times \gamma_C^{-N/2} \prod_s \exp\left\{-\frac{1}{2}(\gamma_C\sigma_{i_s,total}^2)^{-1} \sum_{i=1}^{n_s} (W_{is} - L_{is})^2\right\} \\ &\times \prod_s |\Sigma_s(\rho_s, \gamma_C)|^{-1/2} \exp\left\{-(1/2)(\mathcal{X}_s - \mathcal{L}_s)^T \right. \\ &\quad \left. \times \Sigma_s^{-1}(\rho_s, \gamma_C)(\mathcal{X}_s - \mathcal{L}_s)\right\} \end{aligned} \tag{4}$$

3.2 Block Sampling of True Doses

Because of the correlation of the Berkson errors, the major difficulty with implementation lies with sampling the blocks of latent variables within a stratum  $\mathcal{X}_s = (X_{1s}, \dots, X_{n_s s})^T$ . It is possible to do this one element at a time with Metropolis steps, but we found the resulting sampler to be more than an order of magnitude slower, and to mix more poorly for estimation of the crucial excess relative risk per Gy parameter  $\theta$ , than the following block sampler.

For the block sampler, the complete conditional distribution for  $\mathcal{X}_s$  is proportional to

$$\begin{aligned} &\prod_i (H[Z_{is}^T\beta + \log\{1 + \theta \exp(X_{is})\}])^{Y_{is}} \\ &\times (1 - H[Z_{is}^T\beta + \log\{1 + \theta \exp(X_{is})\}])^{1-Y_{is}} \\ &\times |\Sigma_s(\rho_s, \gamma_C)|^{-1/2} \exp\left\{-(1/2)(\mathcal{X}_s - \mathcal{L}_s)^T \right. \\ &\quad \left. \times \Sigma_s^{-1}(\rho_s, \gamma_C)(\mathcal{X}_s - \mathcal{L}_s)\right\}. \end{aligned}$$

Denote the current value of  $\mathcal{X}_s$  as  $\mathcal{X}_{s,curr}$ . The sampling plan is to generate a new value  $\mathcal{X}_{s,new} = \text{Normal}\{\mathcal{L}_s, \Sigma_s(\rho_s, \gamma_C)\}$ , and update  $\mathcal{X}_s$  to the new value  $\mathcal{X}_{s,new}$  with acceptance probability

$$\alpha = \min\left\{1, \frac{\mathcal{P}_{\mathcal{X}_s}(\mathcal{X}_{s,new})}{\mathcal{P}_{\mathcal{X}_s}(\mathcal{X}_{s,curr})}\right\}, \tag{5}$$

where

$$\begin{aligned} \mathcal{P}_{\mathcal{X}_s}(x) &= \prod_i (H[Z_{is}^T\beta + \log\{1 + \theta \exp(x_i)\}])^{Y_{is}} \\ &\times (1 - H[Z_{is}^T\beta + \log\{1 + \theta \exp(x_i)\}])^{1-Y_{is}}. \end{aligned}$$

In this step, for generation of the candidate random variables, computation of  $\Sigma_s^{-1}(\rho_s, \gamma_C)$  can be done exactly, without having to do a massive inverse, see Web Appendix A.

### 3.3 Prior for $\theta$

In our Bayesian analysis, we used two priors for the excess relative risk per Gy,  $\theta$ . The first was a truncated normal distribution, roughly centered at a plausible point estimate and with reasonably large variances, both depending on the problem. A similar choice was made by Mallick et al. (2002).

The second prior for  $\theta$  was based upon a version of Jeffreys prior, as follows. Recall that Jeffreys prior is  $\pi_J(\theta) \propto |I(\theta)|^{1/2}$ , where  $I(\theta)$  is the Fisher's information with respect to  $\theta$ , treating other parameters as constants. If we ignore the strata, and treat  $Z$  and  $X$  as independent, then the log likelihood for  $\theta$  is

$$\begin{aligned} \ell(\theta | Y, X) &= Y[Z^T \beta + \log\{1 + \theta \exp(X)\}] \\ &\quad - \log(1 + \exp[Z^T \beta + \log\{1 + \theta \exp(X)\}]). \end{aligned}$$

One can easily derive that

$$\begin{aligned} I(\theta) &= -E(\partial^2 \ell / \partial \theta^2) \\ &= E(p(\theta)\{1 - p(\theta)\}[\exp(X)/\{1 + \theta \exp(X)\}]^2), \end{aligned}$$

where  $p(\theta) = H[Z^T \beta + \log\{1 + \theta \exp(X)\}]$ . Now make a rare-event approximation, so that  $p(\theta) \approx \exp(Z^T \beta) \times \{1 + \theta \exp(X)\}$  and  $1 - p(\theta) \approx 1$ . Under such approximations, we have

$$I(\theta) \propto E[\exp(2X)/\{1 + \theta \exp(X)\}].$$

Under the Berkson model,  $X = \text{Normal}(\mu_X, \sigma_X^2)$ , where  $\mu_X = \mu_W$  and  $\sigma_X^2 = \text{var}(W) + \text{var}(U_B)$ . We can estimate  $\mu_X$  and  $\sigma_X^2$  using the method of moments. Denoting these estimates as  $\hat{\mu}_X$  and  $\hat{\sigma}_X^2$ , the resulting approximation becomes

$$\pi_J(\theta) \propto \left\{ \frac{1}{\sqrt{\pi}} \sum_{j=1}^J \omega_j \frac{\exp(2\hat{\mu}_X + 2\sqrt{2}\hat{\sigma}_X x_j)}{1 + \theta \exp(\hat{\mu}_X + \sqrt{2}\hat{\sigma}_X x_j)} \right\}^{1/2},$$

where  $x_j$  and  $\omega_j$  are the abscissas and weights for Gauss-Hermite quadrature. The resulting prior is improper: we truncated it to make it proper and used a random walk type of Metropolis scheme to update  $\theta$ .

## 4. Regression Calibration and the EM Algorithm

In this section we discuss two frequentist approaches to the problem. Our purposes in doing so are two: (a) because such methods are intrinsically interesting; and (b) to provide a check on the Bayesian results.

### 4.1 Two-Stage Regression Calibration

4.1.1 *Unshared/independent Berkson uncertainties.* In regression calibration (Carroll et al., 2006), the idea is to replace the true doses,  $\exp(X)$ , by their expectation given all the observed/calculated doses.

In the cases of uncorrelated mixed Berkson-classical uncertainties, and hence for either pure Berkson or pure classical uncertainty, following Reeves et al. (1998) and Mallick et al. (2002), it is readily shown that if we assume that the  $L_{is}$  are independent Normal  $(\mu_s, \sigma_s^2)$ , then given the calculated doses  $W_{is}$ ,  $[X_{is} | W_{is}] = \text{Normal}\{(1 - \lambda_{is})\mu_s +$

$\lambda_{is}W_{is}, \sigma_{is,B}^2 + (1 - \lambda_{is})\sigma_s^2\}$ , where  $\lambda_{is} = \sigma_s^2/(\sigma_s^2 + \sigma_{is,C}^2)$ . If we define  $\nu_{is} = \exp\{(1 - \lambda_{is})(\mu_s + \sigma_s^2/2) + \sigma_{is,B}^2/2\}$ , then

$$E\{\exp(X_{is}) | W_{is}\} = \nu_{is} \exp(\lambda_{is}W_{is}). \quad (6)$$

Given estimates of  $\mu_s$  and  $\sigma_s^2$ , the regression calibration algorithm simply replaces  $\exp(X_{is})$  in (3) by (6). To estimate  $\mu_s$  and  $\sigma_s$ , we compute their maximum likelihood estimated based upon the calculated dose data  $W_{is}$ , recognizing that the latter have mean  $\mu_s$  and variance  $\sigma_s^2 + \sigma_{is,C}^2$ . The bootstrap can then be used to find approximate confidence intervals for the excess relative risk per Gy,  $\theta$ , approximate in the sense that regression calibration is approximate.

Lyon et al. (2006) used regression calibration assuming all uncertainty was Berkson and unshared (independent).

4.1.2 *Shared/correlated Berkson uncertainties.* Because the Berkson uncertainties and the classical uncertainties are independent, this means that  $X_{is}$  is independent of all the calculated doses given  $L_{is}$ , as well as all the other latent intermediate variables. Hence we see that for regression calibration, in the shared/correlated Berkson case,

$$\begin{aligned} E\{\exp(X_{is}) | W_s\} &= E[E\{\exp(X_{is}) | W_s, L_{is}\} | W_s] \\ &= E[E\{\exp(X_{is}) | L_{is}\} | W_s] \\ &= E[E\{\exp(X_{is}) | L_{is}\} | W_{is}] \\ &= E\{\exp(X_{is}) | W_{is}\}, \end{aligned}$$

where  $W_s = (W_{1s}, \dots, W_{n_s s})^T$ . This means that regression calibration does not distinguish between the case that the Berkson uncertainties are independent and that they are correlated.

### 4.2 EM Algorithm

Another way to work out this problem in the frequentist paradigm is to use the MCEM algorithm (McCulloch, 1997). Parenthetically, we attempted to implement a direct Monte-Carlo maximization, but given the dimensionality of the problem with shared uncertainties, not too surprisingly this approach failed.

Rather than attempt to estimate all the correlation parameters as well as the percentage  $\gamma_C$ , as a partial check to see whether Bayesian and frequentist approaches come to roughly the same conclusion, we made the following approximations. First, because the possible intervals for the correlation parameters are generally of length 0.20 in each stratum, we fixed the correlations at the interval midpoints. Also, because the interval for  $\gamma_C$  was from 0.30 to 0.50, we reran our analysis for  $\gamma_C = 0.30, 0.40$ , and 0.50, although we report only 0.40: the other analyses were what would be expected with varying amounts of classical measurement error.

4.2.1 *MCEM for point estimation.* Denote  $\delta = \log(\theta)$  and  $\Theta = (\Theta_1, \Theta_2)$ , where  $\Theta_1 = (\beta, \delta)^T$ ,  $\Theta_2 = \{(\mu_s, \sigma_s^2), s = 1, \dots, S\}^T$ , and let  $\ell_{\text{comp}}(\Theta | \mathcal{Y}, \mathcal{X}, \mathcal{L}, \mathcal{W}, \mathcal{Z})$  be the complete data log likelihood given as

$$\begin{aligned}
 \ell_{\text{comp}}(\Theta | \mathcal{Y}, \mathcal{X}, \mathcal{L}, \mathcal{W}, \mathcal{Z}) &= \ell_1(\Theta | \mathcal{Y}, \mathcal{X}, \mathcal{L}, \mathcal{Z}) + \ell_2(\mathcal{W} | \mathcal{L}), \\
 \ell_1(\Theta | \mathcal{Y}, \mathcal{X}, \mathcal{L}, \mathcal{Z}) &\propto \sum_{s,i} Y_{is} [Z_{is}^T \beta + \log\{1 + \exp(\delta + X_{is})\}] \\
 &\quad - \sum_{s,i} \log(1 + \exp[Z_{is} \beta + \log\{1 + \exp(\delta + X_{is})\}]) \\
 &\quad - \sum_s \left\{ (n_s/2) \log(\sigma_s^2) + \sum_i (L_{is} - \mu_s)^2 / (2\sigma_s^2) \right\} \\
 &\quad - (1/2) \sum_s (\mathcal{X}_s - \mathcal{L}_s)^T \Sigma_s^{-1} (\rho_s, \gamma_C) (\mathcal{X}_s - \mathcal{L}_s), \\
 \ell_2(\mathcal{W} | \mathcal{L}) &\propto - \sum_{s,i} (W_{is} - L_{is})^2 / (2\gamma_C \sigma_{is,\text{total}}^2). \tag{7}
 \end{aligned}$$

Notice that  $\ell_1$  is the likelihood when all missing random effects  $\mathcal{X}$  and  $\mathcal{L}$  are available, while  $\ell_2$  is the likelihood for the surrogate  $\mathcal{W}$ .

The algorithm is as the following.

1. Choose a starting point  $\Theta_0$ . Regression calibration with unshared/uncorrelated Berkson uncertainties is what we used.
2. **E-Step:** Let  $Q(\Theta | \Theta_{\text{curr}}) = E\{\ell_{\text{comp}}(\Theta | \mathcal{Y}, \mathcal{X}, \mathcal{L}, \mathcal{W}, \mathcal{Z}) | \mathcal{Y}, \mathcal{W}, \mathcal{Z}, \Theta_{\text{curr}}\}$ , where  $\Theta_{\text{curr}}$  is the current value of  $\Theta$ . To evaluate this conditional expectation, we used Monte Carlo methods, see Section 4.2.3, to generate  $B$  random samples from conditional distribution  $(\mathcal{X}, \mathcal{L} | \mathcal{Y}, \mathcal{W}, \mathcal{Z}, \Theta_{\text{curr}})$ , and approximate  $Q(\Theta | \Theta_{\text{curr}})$  by

$$\hat{Q}(\Theta | \Theta_{\text{curr}}) = B^{-1} \sum_b \ell_{\text{comp}}(\Theta | \mathcal{Y}, \mathcal{X}^{(b)}, \mathcal{L}^{(b)}, \mathcal{W}, \mathcal{Z}).$$

To do this generation, we generated samples of  $\mathcal{X}_s$  and  $\mathcal{L}_s$  using exactly equation (A.1) in Web Appendix A, the block sampler of Section 3.2 and equation (5).

3. **M-Step:** Maximize  $\hat{Q}(\Theta | \Theta_{\text{curr}})$  with respect to  $\Theta$ .
4. Repeat steps 2 and 3 until the value of  $\Theta$  converges.

The  $M$ -step is rather easy. In the  $M$ -step, the function  $\hat{Q}$  can be factored into two parts, one depending on the parameters  $\Theta_1 = (\beta, \delta)^T$  and the other depending on  $\Theta_{2,s} = (\mu_s, \sigma_s^2)^T$ . We can maximize the two parts separately to get the point estimators. Maximizing  $\hat{Q}$  with respect to  $\mu_s$  and  $\sigma_s^2$  leads to

$$\hat{\mu}_s = B^{-1} \sum_b \sum_i L_{is}^{(b)} / n_s; \quad \hat{\sigma}_s^2 = B^{-1} \sum_b \sum_i (L_{is}^{(b)} - \hat{\mu}_s)^2 / n_s,$$

that is, the sample mean and the sample variance within the  $s$ th stratum.

Maximization of  $\hat{Q}$  with respect to  $\Theta_1 = (\beta, \delta)^T$  does not have a closed form solution. We used the method of scoring, see Web Appendix A for details.

4.2.2 *Estimating the observed information for MCEM.* The likelihood for observed data is

$$f_{\text{obs}}(\mathcal{Y}, \mathcal{W}, \mathcal{Z} | \Theta) = \int f_1(\mathcal{Y}, \mathcal{X}, \mathcal{L}, \mathcal{Z} | \Theta) f_2(\mathcal{W} | \mathcal{L}) d\mathcal{X} d\mathcal{L},$$

where  $f_1$  and  $f_2$  are the likelihoods corresponding to  $\ell_1$  and  $\ell_2$  in (7). The observed information for  $\Theta$  based on the likelihood above is  $I_{\Theta}(\hat{\Theta}) = -(\partial^2 \log f_{\text{obs}} / \partial \Theta \partial \Theta^T)(\mathcal{Y}, \mathcal{W}, \mathcal{Z}, \hat{\Theta})$ . Then the distribution of the point estimator  $\hat{\delta}$  given by EM algorithm can be approximated by  $\text{Normal}(\hat{\delta}, \hat{\sigma}_{\delta}^2)$ , where  $\hat{\sigma}_{\delta}^2$  is the diagonal entry in  $I_{\Theta}^{-1}(\hat{\Theta})$  corresponding to  $\delta$ . By this result, we can get an approximate confidence interval for  $\delta$  by  $\hat{\delta} \pm Z_{\alpha/2} \times \hat{\sigma}_{\delta}$ , and then exponentiate to get a confidence interval for the excess relative risk per gray parameter  $\theta$ .

Following Louis (1982), we have

$$\begin{aligned}
 I_{\Theta}(\hat{\Theta}) &= E\{V(\mathcal{Y}, \mathcal{X}, \mathcal{L}, \hat{\Theta}) | \mathcal{Y}, \mathcal{W}, \hat{\Theta}\} \\
 &\quad - E\{S(\mathcal{Y}, \mathcal{X}, \mathcal{L}, \hat{\Theta}) S^T(\mathcal{Y}, \mathcal{X}, \mathcal{L}, \hat{\Theta}) | \mathcal{Y}, \mathcal{W}, \hat{\Theta}\}, \tag{8}
 \end{aligned}$$

where  $S(\mathcal{Y}, \mathcal{X}, \mathcal{L}, \Theta) = (\partial \log f_1 / \partial \Theta)(\mathcal{Y}, \mathcal{X}, \mathcal{L}, \Theta)$ ,  $V(\mathcal{Y}, \mathcal{X}, \mathcal{L}, \Theta) = -(\partial^2 \log f_1 / \partial \Theta \partial \Theta^T)(\mathcal{Y}, \mathcal{X}, \mathcal{L}, \Theta)$ . The formulae for the derivatives of  $f_1$  are given in Web Appendix A.

To evaluate  $I_{\Theta}(\hat{\Theta})$ , we take samples from the conditional distribution  $(\mathcal{X}, \mathcal{L} | \mathcal{Y}, \mathcal{W}, \hat{\Theta})$ , and replace the conditional expectations by the mean over the random samples. We used the block sampler described in the previous sections to generate a Markov Chain with the desired conditional distribution being its stationary distribution. The algorithm is the same as the one used in the MCEM.

4.2.3 *Implementation.* One important issue in the implementation of the MCEM algorithm is the stopping rule, i.e., when to stop the algorithm. A standard stopping rule for a deterministic EM algorithm is to stop when

$$\max_i \left( \frac{|\Theta_{i,\text{new}} - \Theta_{i,\text{curr}}|}{|\Theta_{i,\text{curr}} + \lambda_1|} \right) < \lambda_2, \tag{9}$$

where  $\lambda_1$  and  $\lambda_2$  are predetermined constants. However, as Booth and Hobert (1999) point out, the MCEM algorithm may never converge according to the deterministic stopping rule, because of the Monte Carlo errors.

In the analysis of the NTS data, we used the following settings. We stop the algorithm by stopping rule (9) with  $\lambda_1 = 0.001$  and  $\lambda_2 = 0.005$ , numbers in the range suggested by Booth and Hobert (1999). We let the number of Monte Carlo sample  $B$  increase as the iteration number  $k$  increases. We let  $B$  slowly increase from 2000 to 20,000 in the iterations.

We also constructed a confidence interval for  $\theta$  by computing the observed information described in Section 4.2.2. To compute the conditional expectation (8), we generated 100,000 samples from the conditional distribution  $[\mathcal{X}, \mathcal{L} | \mathcal{Y}, \mathcal{W}, \hat{\Theta}]$  using the block sampler.

### 5. Analysis of the NTS Data

We ran the Gibbs sampler for 100,000 iterations. We threw out the first 10,000 iterations as burn-in.

In addition to the Jeffreys prior described in Section 3.3, we also tried other proper priors on  $\theta$ , such as the log-normal distribution and, following Mallick et al. (2002), the truncated normal prior distribution that is the Gaussian distribution truncated at 0. For the terms  $\mu_s$  and  $\beta$ , we used independent  $\text{Normal}(0, 1000)$  priors, while for the variance terms  $\sigma_s^2$ , we used an inverse gamma prior with shape 0.05 and scale 3.0.

For the truncated normal priors, we used the independent sampler, described as scheme ( $\theta.1$ ) in Web Appendix A.2 of the technical complements, and the acceptance rate is about

**Table 1**

*Thyroiditis data: posterior distribution for the excess relative risk per gray ( $\theta$ ). Results for two priors for  $\theta$  are displayed: the truncated normal distribution and our rare-event approximation to Jeffreys prior. Confidence intervals for regression calibration are via the bootstrap, those for the Bayesian methods are upper and lower 2.5%-percentiles from the MCMC, and those for MCEM are derived in Section 4.2.2*

Model	Method	Estimate or posterior mean	Posterior median	95%	C.I.
Berkson and classical uncertainties, shared	MCEM	8.05		3.25	19.95
	Bayes- $TN(8.5, 20^2)$	10.12	9.39	3.85	20.04
	Bayes-Jeffreys	9.35	8.60	3.11	19.39
Berkson and classical uncertainties, unshared	MCEM	8.43		3.70	19.21
	Regression calibration	7.46		2.94	15.12
	Bayes- $TN(8.5, 20^2)$	9.89	9.25	3.99	19.42
	Bayes-Jeffreys	9.10	8.47	3.37	18.22
All Berkson, shared	MCEM	5.42		2.14	13.75
	Bayes- $TN(8.5, 20^2)$	7.31	6.88	2.52	15.31
	Bayes-Jeffreys	6.42	5.91	2.15	13.28
All Berkson	MCEM	5.53		2.41	12.71
	Regression calibration	4.90		1.81	9.37
	Bayes- $TN(8.5, 20^2)$	7.11	6.70	2.73	14.24
	Bayes-Jeffreys	6.49	6.05	2.55	12.61

18%. For Jeffreys prior, we use scheme ( $\theta.2$ ) in Web Appendix A.2 in the technical complements to update  $\theta$ , and the acceptance rate is about 67%.

We use the block sampler in Section 3.2 to update  $\mathcal{X}_s$ . In general, bigger strata tend to have lower acceptance rates. For the full model (4), the acceptance rates are all above 30%.

There are three important subcases that we also ran. The first is the mixture of Berkson and classical error models without correlated Berkson uncertainties, obtained by setting  $\rho_s = 0$ . The second is the Berkson model with shared uncertainty, obtained by setting  $\gamma_C = 0$  but still allowing correlated Berkson error within each stratum. The third is the pure Berkson model, obtained by setting  $\gamma_C = 0$  and  $\rho_s = 0$ .

### 5.1 Thyroiditis

For thyroiditis, along with our rare-event approximation to Jeffreys prior for the excess relative risk per gray, we used a truncated-Normal prior with mean 8.5 and standard deviation 20. Since there were 123 cases of thyroiditis in the NTS data, we would not expect a great deal of sensitivity to the choice of the prior distribution for  $\theta$ .

The results are given in Table 1, with trace plots for  $\theta$  and the gender effect given in Web Figure W.1: the figure shows reasonable mixing. In Figure 1, we plot estimates of the posterior density for the approximate Jeffreys prior for the pure Berkson and shared Berkson-classical models, both with shared Berkson uncertainties. Also plotted for illustration are the MCEM estimates and the posterior medians.

As is no surprise, the difference between the results when the uncertainties are assumed to be entirely of Berkson form and the results when classical uncertainty is also allowed are striking: the various point estimates of  $\theta$  increase, and the upper end of the 95% credible/confidence intervals increases

even more. The results for the two choices of prior for  $\theta$  are similar. Figure 1 shows that the posterior density is only slightly skewed, so that the posterior median and the MCEM estimates are roughly the same.

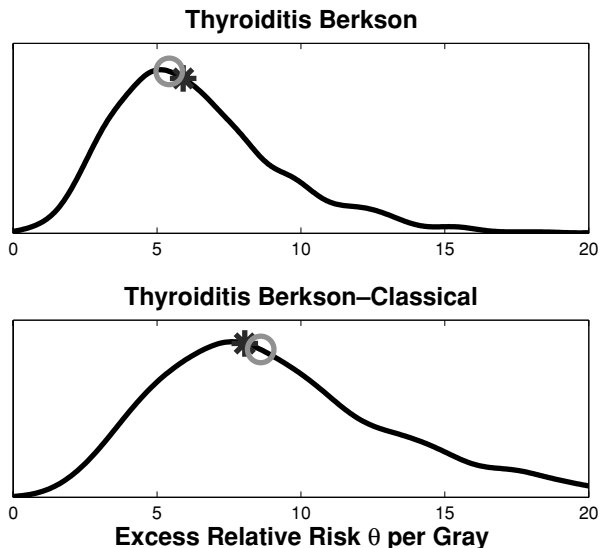
Overall, the message is rather clear: the point estimate for ERR/Gy is  $\approx 8.0$ , with an upper end of 95% credible/confidence limits being  $\approx 20$ . Treating all sources of dose uncertainties explicitly raises both the point estimate and the upper limits of the 95% confidence interval by about a factor of two above the special case when regression calibration is used assuming all dose uncertainty is 100% unshared Berkson errors.

In this case, the difference between the analysis accounting and not accounting for correlation of the Berkson uncertainties is minor.

### 5.2 Thyroid Neoplasms

For thyroid neoplasms, along with our rare-event approximation to Jeffreys prior, we used a truncated-Normal prior with mean 20 and standard deviation 100. We also used a log-normal prior with mean 26 and standard deviation 64. The prior means are roughly what evidence from Berkson models suggest for thyroid cancers, while the prior standard deviations are also roughly in accord with standard deviations from other analyses. Since there were only 20 cases of thyroid neoplasm in the NTS data, we expect greater sensitivity to the choice of the prior distribution for the ERR/Gy,  $\theta$ , along with much wider credible/confidence intervals.

The results are given in Table 2, with trace plots for  $\theta$  and the gender effect given in Web Figure W.2: the figure again shows reasonable mixing. The results clearly indicate the far smaller amount of information for  $\theta$  for thyroid neoplasms than for thyroiditis. The credible/confidence intervals



**Figure 1.** Estimates of the posterior densities for thyroiditis excess relative risk per gray ( $\theta$ ) using the approximate Jeffreys prior. Top plot: the pure Berkson model with shared uncertainties. Bottom plot: the mixed Berkson-classical model with shared uncertainties. The star indicates the MCEM estimate, while the circle indicates the posterior median.

are much wider, with the Bayesian results being somewhat but not ridiculously sensitive to the prior distribution chosen. Basically, what is going on here is that while frequentist and Bayesian methods agree that there is an effect of radiation on

thyroid neoplasm, the small number of disease cases means that the methods have a difficult time finding the upper end of the credible/confidence intervals of excess relative risk per Gy. The largest difference occurs when comparing the log-normal prior with the others: the log normal has much lighter tails, and this is reflected in much smaller upper limits, posterior means and even posterior medians. Once again, the effect of the shared Berkson uncertainties is rather modest. We conclude that the excess relative risk per gray for thyroid neoplasm is much greater than it is for thyroiditis, with a point estimate that is described roughly in the range of 20–30 ERR/Gy.

It is interesting to consider Figure 2, which helps explain the large numerical differences between the MCEM estimates and the posterior medians/means for the approximately Jeffreys prior. As seen in that figure, the posterior densities are highly skewed. The MCEM estimate targets the posterior mode remarkably well, but of course the posterior median is clearly much greater because of the skewness, and the posterior mean even larger.

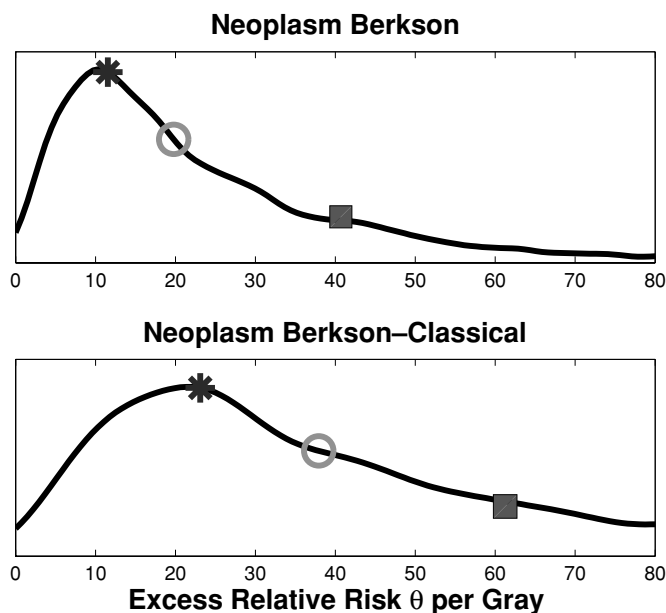
5.3 Model Robustness

Mallick et al. (2002) describe a much more complex framework than that described here. They allow the latent intermediate values  $L_{ij}$  to have a nearly nonparametric distribution within each stratum, while we have focused on the normal distribution. The major reason for this change was that the estimated dose data we use is very much different from that used by Mallick et al. (2002). Indeed, the dosimetry was totally redone, and the net effect was that everyone had a positive dose, outliers were removed, etc. In addition, the strata used

**Table 2**

*Neoplasm data: posterior distribution for the excess relative risk per gray ( $\theta$ ). Results for three priors for  $\theta$  are displayed: the truncated normal distribution, a lognormal distribution with the indicated mean and variance, and our rare-event approximation to Jeffreys prior. Confidence intervals for regression calibration are via the bootstrap, those for the Bayesian methods are upper and lower 2.5%-percentiles from the MCMC, and those for MCEM are derived in Section 4.2.2*

Model	Method	Estimate or posterior mean	Posterior median	95%	C.I.
Berkson and classical uncertainties, shared	MCEM	23.10		4.40	121.30
	Bayes- $TN(20, 100^2)$	63.83	47.52	8.55	199.28
	Bayes- $LN(26, 64^2)$	24.28	17.11	3.88	79.19
	Bayes-Jeffrey	60.85	37.93	5.39	240.32
Berkson and classical uncertainties, unshared	MCEM	24.14		4.74	123.05
	Regression Calibration	22.92		4.30	184.27
	Bayes- $TN(20, 100^2)$	56.37	45.33	7.91	163.92
	Bayes- $LN(26, 64^2)$	25.08	18.49	4.08	83.68
	Bayes-Jeffreys	60.94	38.83	6.04	243.26
All Berkson, shared	MCEM	11.55		2.45	54.30
	Bayes- $TN(20, 100^2)$	38.76	24.73	4.86	165.22
	Bayes- $LN(26, 64^2)$	15.73	12.51	2.64	51.23
	Bayes-Jeffreys	40.73	19.77	3.21	224.66
All Berkson	MCEM	12.02		2.65	54.43
	Regression Calibration	11.35		2.10	59.24
	Bayes- $TN(20, 100^2)$	40.46	28.01	5.02	142.29
	Bayes- $LN(26, 64^2)$	15.99	11.77	2.55	53.75
	Bayes-Jeffreys	35.17	19.51	3.41	187.81



**Figure 2.** Estimates of the posterior densities for thyroid neoplasm of the excess relative risk per gray ( $\theta$ ) using the approximate Jeffreys prior. Top plot: the pure Berkson model with shared uncertainties. Bottom plot: the mixed Berkson–classical model with shared uncertainties. The star indicates the MCEM estimate, the circle indicates the posterior median and the square indicates the posterior mean.

by Mallick et al. (2002) were simply the state of residence and gender, while our strata were more finely tuned to reflect how the dosimetry was actually done.

The net effect of redoing the dosimetry and building more relevant strata greatly improved the quality of the data, and actually made  $L_{ij}$  appear to be much more normally distributed within strata. For example, we took each of the strata with more than 50 individuals, and did q–q plots for the gamma and normal distributions in dose and log dose. The dose data appear roughly gamma distributed, with varying shape and scale. For the normal distribution, there was some modest left skewness, averaging about  $-0.40$ , and the observed log doses tended to have kurtosis *smaller* than the normal distribution, i.e., without heavy tails. In general then, we do not think the complex mixture model for  $L$  proposed by Mallick et al. is really needed with these new data.

As an added robustness check, we followed a prescription of Huang, Stefanski, and Davidian (2006), wherein we assumed that 40% of the error was classical, and added double this amount of normally distributed noise to the observed log doses. That is, remembering for an individual that  $\sigma_{is,B}^2 + \sigma_{is,C}^2 = \sigma_{is,total}^2$ , we took  $\sigma_{is,C}^2 = 0.4 \sigma_{is,total}^2$ , then added normal noise with variance  $0.8 \sigma_{is,total}^2$ , thus making the fraction of classical error to be  $2/3$ , i.e., we tripled the amount of classical error. We recomputed our Bayesian estimators (mixture of classical and Berkson errors, with shared uncertainties, using the truncated normal prior for  $\theta$ ) in the thyroiditis data 200 times. For  $\theta$ , the posterior mean and confidence interval without added noise were 10.1 and (3.8, 20.0), while their

averages with the increased classical noise were 7.9 and (2.4, 17.2), which is hardly any change practically. This suggests that our analysis is reasonably robust to the assumption that  $L_{is}$  is normally distributed within strata.

A second difference is that we have focused on the excess relative risk model (3), while Mallick et al. use a form of very flexible, near-nonparametric regression. Our reasons for this are two. First, in radiation dosimetry, almost all data analyses use either the excess relative risk model or a linear dose–response model (Stram and Kopecky, 2003). It is unlikely that much more complex dose–response models will be accepted in the literature. Secondly, as seen in Tables 1 and 2 of Mallick et al. (2002), even with more extreme data with zero estimated doses and outliers, there was little practical difference between the analysis using the excess relative risk model and the nearly nonparametric dose–response model.

#### 5.4 Model Identifiability and Sensitivity Analysis

There are two parts of the modeling effort that are not identified in the technical frequentist sense: the amount of uncertainty that is Berkson, and the correlation of the Berkson uncertainties. Thus, while we know from the science that there is classical measurement error in the dosimetry, and that the Berkson uncertainties are correlated, the data themselves are not sufficient to test for either. In Section 2.2, we described how we specified the ranges of these parameters for the NTS fallout data.

To check how sensitive the results are to the specifications of these parameters, we reran the Bayesian analysis for the thyroiditis data under the Classical–Berkson mixture model with shared Berkson uncertainty for the following three cases: (a) a wider range for the percentage of classical uncertainty  $\gamma_C$ , namely [0.2, 0.6]; (b) higher shared Berkson correlation, where we add 0.2 to all ranges of the correlations; and (c) lower correlations, where we subtract 0.1 from all ranges of the shared Berkson correlations. The posterior distributions of the excess relative risk per Gy,  $\theta$ , under the three cases are summarized in Table 3. The results should be compared with those in Table 1, obtained by assigning a truncated normal prior to  $\theta$ . As one can see, the ranges of these parameters do affect the results, but the effects are rather small.

In the frequentist analyses, it would be more desirable to make no assumptions about these parameters. However, the lack of identifiability precludes this, and thus we opted for an approach using the midpoints of assigned intervals.

**Table 3**

*Sensitivity analysis for the Bayesian method. Posterior estimate for  $\theta$  for different specification of the unidentifiable parameters. Results were obtained using the thyroiditis data, under the classical–Berkson mixture model with shared Berkson uncertainties.  $\theta$  is assigned a truncated normal prior as in Table 1*

	Posterior mean	Posterior median	95%	C.I.
Wider range for $\gamma_C$	9.23	8.66	3.49	18.26
Higher correlations	9.82	9.00	3.59	21.26
Lower correlations	9.93	9.20	3.96	19.70



**6. Why Consider Correlated Berkson Uncertainties?**

Stram and Kopecky (2003) and Hoffman et al. (2006) use different models and show that the existence of shared/correlated Berkson uncertainty will lead to a loss of power for dose-response testing using regression calibration methods, and that, as Stram and Kopecky note on page 414, “power will be overstated if shared dosimetry errors are ignored.” It is less clear what will be the effect of accounting for the shared Berkson uncertainty versus simply ignoring it. Stram and Kopecky (p. 410) conclude that “ignoring shared dosimetry error when constructing confidence intervals will result in confidence intervals that are too narrow.”

In general, if one starts with a generalized nonlinear fixed effects model, the effect of independent Berkson uncertainties is to change the form of the model for the observed data, but it remains a generalized nonlinear fixed effects model.

Correlated (shared) Berkson uncertainties change much more. As in our case, the effect of correlated Berkson uncertainties is to change the model for the observed data to a generalized nonlinear *mixed* effects model with stratum-specific random effects. As a general matter, consideration of such mixed effects can be crucial both for efficiency in estimating fixed effects and for validity of inferences about those fixed effects. In this section, we consider two examples of this phenomenon.

*6.1 The Linear Model*

Consider the linear model that  $Y_{is} = \beta_0 + \beta_1 X_{is} + \epsilon_{is}$ , where  $\epsilon_{is} = \text{Normal}(0, \sigma_\epsilon^2)$ , and suppose that the model is purely Berkson with  $X_{is} = W_{is} + U_{b,is}$ , where  $U_{b,is} = \text{Normal}(0, \sigma_{b,is}^2)$  and the Berkson uncertainties are shared within strata, so that  $\text{corr}(U_{b,is}, U_{b,ks}) = \rho$  for  $i \neq k$ . Then the actual observed data follow a heteroscedastic *mixed* model, with  $E(Y_{is} | W_{is}) = \beta_0 + \beta_1 W_{is}$ ,  $\text{var}(Y_{is} | W_{is}) = \sigma_\epsilon^2 + \beta_1^2 \sigma_{b,is}^2$  and

$$\text{cov}(Y_{is}, Y_{ks} | W_{is}, W_{ks}) = \rho \beta_1^2 \sigma_{b,is} \sigma_{b,ks}. \tag{10}$$

Thus the observed data follow a linear mixed model. Accounting for the correlations can greatly improve efficiency, while not accounting for them can affect the validity of inferences.

To see this, we did the following simulation. We had a total of 300 observations, and allowed there to be 5, 10, 20, and 50 strata, all of the same size. We set  $\rho = 0.5$ ,  $\sigma_\epsilon^2 = 1$ ,  $\beta_1 = 2$ , and  $\sigma_{u,ij}^2 \equiv 2$ , so that the Berkson uncertainty was homoscedastic. This simulation was set up so that there was a large signal, that the observed data have homoscedastic errors but that the errors are correlated within strata, and the amount of the Berkson errors is known as well as is their correlation. In this setup, most of the variability is due to the Berkson uncertainties. Simulations indicate that in this case, a weighted least squares analysis that accounts for the Berkson correlations had about a 50% gain in mean squared error efficiency when compared to the analysis that ignores the correlations.

Inspection of (10) shows an important point: if  $\beta_1 = 0$  and there is no relationship between  $Y$  and  $X$ , then there is also no effect of shared Berkson uncertainty. Thus, we can expect to see some effects of shared uncertainties when the Berkson errors are substantial and the signal is substantial. If there is

no signal, then of course that the Berkson errors are shared is irrelevant.

*6.2 Excess Relative Risk Model*

The same considerations also apply to the excess relative risk model (3). Let  $\text{var}(U_{is}) = \sigma_u^2$ ,  $U_{is} = \tau_s + \zeta_{is}$ ,  $\text{var}(\zeta_{is}) = (1 - \rho)\sigma_u^2$  and  $\text{var}(\tau_s) = \rho\sigma_u^2$ . For simplicity, make a rare disease assumption so that  $\text{pr}(Y_{is} = 1 | Z_{is}, X_{is}) \approx \exp(\beta_0 + \beta_1 Z_{is})\{1 + \theta \exp(X_{is})\}$ . Then

$$\begin{aligned} &\text{pr}(Y_{is} = 1 | Z_{is}, W_{is}, \text{stratum} = s) \\ &\approx \exp(\beta_0 + \beta_1 Z_{is}) [1 + \theta \exp\{W_{is} + \tau_s + (1 - \rho)\sigma_u^2/2\}] \\ &\approx H(\beta_0 + \beta_1 Z_{is} + \log[1 + \theta \exp\{W_{is} + \tau_s + (1 - \rho)\sigma_u^2/2\}]). \end{aligned} \tag{11}$$

Thus, the effect of shared Berkson uncertainties is to induce a generalized nonlinear mixed model for the observed data, with correlations among data within the same stratum. Of course if there is no effect of  $X$ , i.e., if  $\theta = 0$ , then there will be no effect due to the shared uncertainties.

An interesting special case occurs when the correlations in the Berkson uncertainties are large, i.e.,  $\rho \approx 1$ . Then if the Berkson uncertainties  $\tau_s$  have large variance relative to the variability of the calculated log doses  $W_{is}$ , and if the excess relative risk  $\theta$  is also large, then we can expect strong correlations among the binary responses within some of the strata.

To illustrate these phenomena, we simulated 200 data sets in three different scenarios, as follows. All uncertainty was Berkson, and the truncated normal prior for  $\theta$  was used. Each data set had 3000 observations, with 50 equal-sized strata, with common correlation  $\rho = 0.5$  within each stratum. We set  $Z_{is}$  to be 0 or 1 with equal frequency in each stratum. In Scenario 1, that of Hoffman et al. (2006), we set  $\text{logit}(\beta_0) = 0.0049$ ,  $\text{logit}(\beta_0 + \beta_1) = 0.0098$ ,  $\theta = 4.0$ ,  $E(W) = \log(0.10)$ ,  $\text{var}(W) = \log(2.70)^2$ , and the Berkson uncertainty had variance  $\log(2.3)^2$ . In Scenario 2, we set  $\text{logit}(\beta_0) = 0.0049$ ,  $\text{logit}(\beta_0 + \beta_1) = 0.0098$ ,  $\theta = 7$ ,  $E(W) = \log(0.10)$ ,  $\text{var}(W) = \log(2.70)^2$  and the Berkson uncertainty had variance  $\log(3.6)^2$ , i.e., larger dose response and more Berkson uncertainty. Finally, in Scenario 3, we set  $\text{logit}(\beta_0) = 0.0098$ ,  $\text{logit}(\beta_0 + \beta_1) = 0.0196$ ,  $\theta = 14$ ,  $E(W) = \log(1.0)$ ,  $\text{var}(W) = \log(2.70)^2$  and the Berkson uncertainty had variance  $\log(3.6)^2$ , i.e., even larger dose response, higher rates of disease, more Berkson uncertainty and much higher overall doses. We reran the MCMC analysis, with 50,000 MCMC samples, ignoring and taking into account the correlations in Berkson uncertainties. There was almost no difference accounting for correlation in Scenarios 1 or 2, with both methods having some upward bias and near-nominal (Scenario 1) or over-nominal (Scenario 2) coverage of the credible intervals, see Table 4. Finally, in Scenario 3, we found the analysis accounting for correlation had approximately 20% smaller variability in the posterior mean, and that the analysis without shared uncertainty had a biased posterior mean (approximately 10, rather than the true value  $\theta = 14$ ). Also, differences occurred in the 95% credible intervals. The analysis with shared uncertainty had longer mean width (25.5 vs. 22.3), and the analysis that did not account for correlations had 95% credible intervals that covered the

**Table 4**

Results for the simulation described in Section 6.2. The scenarios are described in that section. “Shared” refers to an analysis that accounts for shared Berkson uncertainty, while “Unshared” refers to an analysis that accounts for the Berkson error but ignores the correlation

Scenario	Shared			Unshared		
	1	2	3	1	2	3
Excess relative risk	4.0	7.0	14.0	4.0	7.0	14.0
Posterior mean	7.3	10.9	14.1	7.4	10.9	9.5
Posterior variance	12.4	11.3	11.5	12.5	11.1	13.8
Mean posterior CI length	16.3	20.2	25.5	16.7	21.3	22.3
Frequentist coverage	0.94	0.96	0.98	0.92	0.99	0.88

true value only 87.5% of the time, versus 98% of the time for those accounting for correlations.

As described here, we see no large effect due to the shared Berkson uncertainties in our actual data. Our simulations also suggest that the effects of ignoring the correlations in the analysis will sometimes be minor, although with high doses and substantial dose response, we observed the phenomenon of narrower confidence intervals that Stram and Kopecky (2003) conclude should arise. It is important to emphasize, however, that the shared nature of the Berkson uncertainties is well established, see e.g., Stram and Kopecky (2003), and it is of scientific importance that analyses account for this in order to have some chance being accepted.

## 7. Concluding Remarks

We have considered the problem of risk estimation in measurement error problems where there is a complex mixture of classical and Berkson errors, and in which the Berkson errors are correlated within distinct strata that are of sometimes large size. Using a block-sampling scheme, we have derived and implemented both a Bayesian MCMC analysis and a MCEM analysis. In our examples, the two methods give roughly similar answers, perhaps as to be expected, but reassuring nonetheless.

The problem we have considered arises naturally in studies of radiation and disease, because of the complex nature of the process of radiation dosimetry. The NTS Thyroid Disease Study is a good but not the only example of this type of problem. Besides the model formulation, a key empirical finding is that block sampling can be implemented successfully in either the frequentist or Bayesian paradigms in problems of complex mixtures of Berkson and classical errors.

Scientifically, we have found that for a relatively common condition, thyroiditis, there is a strong dose response with relatively tight confidence intervals for excess relative risk. For the much rarer thyroid neoplasms, the dose response is very much less well articulated, with upper ends of confidence intervals that reflect the lesser amount of information.

For thyroiditis, full treatment of all sources of uncertainty increased both the point estimate and the upper limit of the 95% confidence interval by a factor  $\approx 2$  from the case whereby all uncertainties were assumed to be 100% unshared Berkson

and regression calibration used with the expectation value of individual dose. For neoplasms, the point estimate and the lower and upper limits of the 95% confidence interval of the dose–response relationship increased by a factor of  $\approx 2$ –4, depending on the method used. Thus, we conclude that ignoring mixtures of Berkson and classical errors in the reconstruction of individual doses can result in a substantial underestimate of the dose–response relationship in an epidemiological study, and can produce misleading conclusions concerning the relevance and uniqueness of the findings when compared with other studies.

In simulation studies (Hoffman et al., 2006), we found that shared Berkson uncertainties have the potential to decrease statistical power by 10% or more. Finally, the effect of the correlation of the Berkson uncertainties being shared within strata was relatively modest in these data. However, the correlation clearly exists and it is important for scientific credibility to account for it. Our methodology shows one way to do this.

## 8. Supplementary Materials

Web Appendix A and Web Figures W.1–W.2 are available at the last author’s website and under the Paper Information link at the *Biometrics* website <http://www.tibs.org/biometrics>.

## ACKNOWLEDGEMENTS

Our research was supported by a grant from the National Cancer Institute (CA-57030), and by the Texas A& M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106). We thank Dr J. L. Lyon of the University of Utah for making the updated NTS data available to us.

## REFERENCES

- Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society Series B*, **61**, 265–285.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Craineceanu, C. M. (2006). *Measurement Error in Nonlinear Models*, 2nd edition. Boca Raton, Florida: Chapman and Hall CRC Press.
- Hoffman, F. O., Ruttenber, A. J., Apostoaei, A. I., Carroll, R. J., and Greenland, S. (2006). The Hanford Thyroid Disease Study: An alternative view of the findings. *Health Physics* **92**, 99–111.
- Huang, X., Stefanski, L. A., and Davidian, M. (2006). Latent-model robustness in structural measurement error models. *Biometrika* **93**, 53–64.
- Kerber, R. L., Till, J. E., Simon, S. L., Lyon, J. L. Thomas, D. C., Preston-Martin, S., Rollison, M. L., Lloyd, R. D., and Stevens, W. (1993). A cohort study of thyroid disease in relation to fallout from nuclear weapons testing. *Journal of the American Medical Association* **270**, 2076–2083.
- Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B*, **44**, 226–233.
- Lubin, J. H., Schafer, D. W., Ron, E., Stovall, M., and Carroll, R. J. (2004). A reanalysis of thyroid

- neoplasms in the Israeli tinea capitis study accounting for dose uncertainties. *Radiation Research* **161**, 359–368.
- Lyon, J. L., Alder, S. C., Stone, M. B., et al. (2006). Thyroid disease associated with exposure to the Nevada Test Site radiation: A reevaluation based on corrected dosimetry and examination data. *Epidemiology* **17**, 604–614.
- Mallick, B., Hoffman, F. O., and Carroll, R. J. (2002). Semi-parametric regression modelling with mixtures of Berkson and classical error, with application to fallout from the Nevada Test Site. *Biometrics* **58**, 13–20.
- McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–170.
- Reeves, G. K., Cox, D. R., Darby, S. C., and Whitley, E. (1998). Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine* **17**, 2157–2177.
- Ron, E. and Hoffman, F. O. (1999). *Uncertainties in Radiation Dosimetry and Their Impact on Dose response Analysis*. Bethesda, Maryland: National Cancer Institute Press.
- Schafer, D. W. and Gilbert, E. S. (2006). Some statistical implications of dose uncertainty in radiation dose–response analyses. *Radiation Research* **166**, 303–312.
- Schafer, D. W., Lubin, J. H., Ron, E., Stovall, M., and Carroll, R. J. (2001). Thyroid cancer following scalp irradiation: A reanalysis accounting for uncertainty in dosimetry. *Biometrics* **57**, 689–697.
- Simon, S. L., Till, J. E., Lloyd, R. D., Kerber, R. L., Thomas, D. C., Preston–Martin, S., Lyon, J. L. and Stevens, W. (1995). The Utah Leukemia case–control study: Dosimetry methodology and results. *Health Physics* **68**, 460–471.
- Simon, S. L., Anspaugh, L. R., Hoffman, F. O., Scholl, A. E., Stone, M. B., Thomas, B. A., and Lyon, J. L. (2006). 2004 update of dosimetry for the Utah Thyroid Cohort Study. *Radiation Research* **165**, 208–222.
- Stevens, W., Thomas, D. C., Lyon, J. L. Till, J. E., Kerber, R. A., Simon, S. L., Lloyd, R. D., Elghary, N. A., and Preston–Martin, S. (1992). Assessment of leukemia and thyroid disease in relation to fallout in Utah: Report of a cohort study of thyroid disease and radioactive fallout from the Nevada test site. Technical Paper University of Utah, Salt Lake City, Utah.
- Stram, D. O., and Kopecky, K. J. (2003). Power and uncertainty analysis of epidemiological studies of radiation-related disease risk in which dose estimates are based on a complex dosimetry system: Some observations. *Radiation Research* **160**, 408–417.

Received June 2006. Revised December 2006.

Accepted January 2007.