Regression Calibration for Classical Exposure Measurement Error in Environmental Epidemiology Studies Using Multiple Local Surrogate Exposures

Thomas F. Bateson* and J. Michael Wright

* Correspondence to Dr. Thomas F. Bateson, National Center for Environmental Assessment, Environmental Protection Agency, 1200 Pennsylvania Avenue NW, Mail Code 6623P, Washington, DC 20460 (e-mail: bateson.thomas@epa.gov).

Initially submitted December 9, 2008; accepted for publication April 13, 2010.

Environmental epidemiologic studies are often hierarchical in nature if they estimate individuals' personal exposures using ambient metrics. Local samples are indirect surrogate measures of true local pollutant concentrations which estimate true personal exposures. These ambient metrics include classical-type nondifferential measurement error. The authors simulated subjects' true exposures and their corresponding surrogate exposures as the mean of local samples and assessed the amount of bias attributable to classical and Berkson measurement error on odds ratios, assuming that the logit of risk depends on true individual-level exposure. The authors calibrated surrogate exposures using scalar transformation functions based on observed within- and between-locality variances and compared regression-calibrated results with naive results using surrogate exposures. The authors further assessed the performance of regression calibration in the presence of Berkson-type error. Following calibration, bias due to classical-type measurement error, resulting in as much as 50% attenuation in naive regression estimates, was eliminated. Berkson-type error appeared to attenuate logistic regression results less than 1%. This regression calibration method reduces effects of classical-type measurement error that are typical of epidemiologic studies using multiple local surrogate exposures as indirect surrogate exposures for unobserved individual exposures. Berkson-type error did not alter the performance of regression calibration. This regression calibration method does not require a supplemental validation study to compute an attenuation factor.

Accurate exposure assessment is a major challenge for environmental epidemiologic studies that lack individual-level exposure assessment metrics. For example, many studies of waterborne contaminants use the mean concentration of a limited number of spatially distributed water quality samples as an indirect surrogate for the unobserved true local mean exposure. The mean of these multiple local samples serves as an estimate of individuals’ contemporaneous personal exposures within that locale. Using local area data to assign individual-level exposures results in measurement errors that can lead to imprecise effect estimates that are conservatively biased towards the null (1–4).

Differential exposure measurement error that is associated with the outcome can cause bias in an effect estimate towards or away from the null, while nondifferential exposure error typically results in bias towards the null (5). Nondifferential measurement error in a continuous exposure can be of the classical or Berkson type and typically arises in environmental and occupational settings as a mixture of the 2 forms (6). Classical error occurs when true exposures are measured with additive error (7) and the average of many replicate measurements, conditional on the true value, equals the true exposure (8). This error is statistically independent of the true exposure that is being measured and attenuates true linear effects of exposure, resulting in effect estimates in epidemiologic studies that are biased towards the null (6, 8, 9). Such errors occur when the mean values of multiple local air and water

Abbreviations: DBPs, disinfection by-products; HAA5, 5 haloacetic acids.
pollutant samples are used to estimate the true underlying mean exposures (10–21).

Berkson measurement error is independent of the surrogate measure of exposure (9, 22), and is present when the average of individuals’ true exposures, conditional on the assigned measurement, equals the assigned measurement. Berkson measurement error can arise from the use of local area mean exposures to represent the individual exposures of people in that area—even when the estimated area mean is equal to the true underlying mean (i.e., no classical error). Examples of random variability in personal behavior that may produce Berkson-type error in personal exposure estimates include the volume of water consumed per day; the percentage of water consumed at or away from home; the percentage of boiled, filtered, or bottled water consumed; and the effectiveness of water filtration at removing waterborne contaminants. Nonrandom sources of variability in personal exposure can occur if the error is associated with the true underlying mean exposure (e.g., if behaviors are modified on the basis of water quality perception). In some analytic contexts such as Poisson regression, Berkson error does not bias effect estimates but rather increases the standard errors of effect estimates (6). However, many epidemiologic studies use logistic regression, and studies suggest that Berkson error can produce bias towards the null in logistic regression analyses (23–25).

Regression calibration is a statistical method for adjusting point and confidence interval estimates of effects obtained from regression models for bias due to exposure measurement error (26). This method typically uses validation data sets of the measured surrogate exposure compared with more accurate exposure measurements to estimate the magnitude of any bias derived from using the surrogate exposure metric. The original naive epidemiologic effect estimate based on the surrogate is then scaled or transformed by inverting the estimated attenuation factor, and confidence intervals are adjusted to reflect variability in the estimated attenuation factor. Regression calibration methods have been developed and widely used in nutritional epidemiologic studies (26–33). Regression calibration methods for estimating attenuation factors have been used in air pollution studies with external (34–36) and contemporaneously based on the surrogate is then scaled or transformed by inverting the estimated attenuation factor, and confidence intervals are adjusted to reflect variability in the estimated attenuation factor. Regression calibration methods have been developed and widely used in nutritional epidemiologic studies (26–33). Regression calibration methods for estimating attenuation factors have been used in air pollution studies with external (34–36) and contemporaneously collected (37, 38) validation data sets.

Measurement error is a significant concern in disinfection by-product (DBP) studies of the effects of water quality and may explain some inconsistent findings. Only 2 of 13 studies examining the impact of DBPs on smalls-for-gestational-age birth collected data on individual water use (10–18, 39–42). In previous DBP studies, investigators have evaluated potential bias from spatial variability resulting from using local average concentrations to estimate individual-level exposures (2, 3), while others have examined potential exposure misclassification resulting from interindividual variability in water-use patterns (1, 4, 43, 44). These investigators have not attempted to correct for bias attributable to exposure measurement error in epidemiologic studies of waterborne contaminants.

We investigated the quantitative nature of bias from non-differential measurement error of the classical and Berkson types as they may occur in epidemiologic studies of DBPs. Unlike many regression calibration methods currently employed in epidemiologic analyses, our proposed method does not require the use of a validation data set and is amenable to implementation using only routinely collected data.

MATERIALS AND METHODS

We simulated data to represent exposure to haloacetic acids, a type of nonvolatile DBP, specifically the metric called HAA5, which is the sum of concentrations of mono-, di-, and trichloroacetic acids and mono- and dibromoacetic acids. The HAA5 exposures were intended to represent exposures from public drinking water in 100 fictional localities within a single distribution system. The framework for the data simulations assumed considerable variability in HAA5 concentrations within and between localities, as has been described in other DBP studies (14–16). Simulated values of true underlying locality-specific mean HAA5 levels were based on observed distributions of haloacetic acid samples collected across several drinking-water distribution systems in Massachusetts during 1997–1998 (14) (mean = 35 µg/L; standard deviation, 25).

Since these data were approximately lognormally distributed, we used a corresponding normal distribution with a mean concentration of 3.35 µg/L and a standard deviation of 0.642 µg/L to represent the log-transformed HAA5 concentrations in our simulated distribution system. We constrained our simulations such that the overall variance of water samples within the system as a whole remained constant, while the ratio of within- to between-locality variances changed. The true but unobservable mean ln(HAA5) concentrations in each of 100 localities were randomly selected from a specific distribution. Each locality’s mean concentration of HAA5 was randomly assigned from the system-wide distribution, with mean 3.35 µg/L and variance equal to the simulated between-locality variance on the log scale (σ²w). The between-locality variance was a function of the system-wide variance (0.4123 µg/L) and the specified ratio of within-locality variance (σ²j) to between-locality variance, such that σ²w = (σ²j/(1 + ρ)), where ρ = σ²w/σ²j. Each locality’s simulated variance was allowed to vary around the assigned variance; the distribution of the difference between the assigned level for locality j and the true mean level (σ²w) was assumed to follow a 0-mean normal distribution with variance σ²w.

Figure 1 shows how 3 localities’ means and variances might be represented. The wide distribution reflects the true underlying water distribution system ln(HAA5) concentrations alongside narrower distributions reflecting 3 localities’ randomly assigned true underlying means (A, B, and C) from the system-wide distribution with their randomly assigned within-locality variances, such that the specified ratio of within-locality variance to between-locality variance was maintained. We used each locality’s true mean ln(HAA5) concentration (e.g., A, B, and C from Figure 1) and its variance to generate 4 random ln(HAA5) samples for each locality. Each of the 100 simulated subjects per locality was independently assigned a personal ln(HAA5) concentration that was randomly chosen on the basis of the true mean of...
ln(HAA5) concentrations in the individual’s locality and additional random Berkson variability representing personal water-usage patterns. The Berkson error was either zero, once, or twice the system-wide variability (i.e., geometric standard deviation of 0.642 µg/L).

Figure 2 shows the true underlying distribution of ln(HAA5) concentrations in 1 locality with the true mean and sample mean based on 4 random samples (thicker gray density function). “A” denotes the true assigned mean of this distribution, while “a” denotes the average of 4 ln(HAA5) samples (large circles) randomly selected from this distribution and represents the observed sample mean, which served as the surrogate measurement for each individual’s true exposure. The difference between “A” and “a” shows the classical measurement error. Figure 2 also shows the 100 randomly sampled values from around 1 locality’s mean, representing the 100 subjects’ true ln(HAA5) exposures (thinner black density function). The small circles show each of 100 subjects’ true exposures. Differences between each small circle and “A” reflect the Berkson measurement error.

The health outcome event was assumed to be binary and a logistic function of the natural logarithm of true maternal exposure, the baseline risk of an adverse effect, and the magnitude of simulated effects. The baseline probability of each simulated subject’s experiencing an adverse birth outcome was 5% in the absence of exposure, and we simulated an arbitrary odds ratio of 2.00 per ln(20 µg/L) of true maternal HAA5 exposure. Logistic regression analyses were conducted using PROC LOGISTIC in SAS, version 9.1 (SAS Institute Inc., Cary, North Carolina). We simulated system-wide variability equivalent to ½, 1, and 2 times the variability reported across 17 towns in Massachusetts (14). We simulated different ratios of within-locality variance to between-locality variance ranging from 1:4 to 4:1.

Classical error variance in the mean of several replicate samples is equal to the within-locality exposure variance divided by the within-locality sample size that was used to develop the surrogate measure (24). In order to correct for this classical measurement error in our naive simulation results, we calibrated the observed local mean exposure metrics using estimates of within- and between-locality variance which were obtained using PROC VARCOMP (with type = REML). We propose using the following formula (adapted from Reeves et al. (24)) to calibrate the surrogate exposure metric (i.e., mean of the log-transformed local area samples), such that the effect of exposure estimated through logistic regression analysis of these transformed data will not be biased by classical exposure measurement error:

\[ X_{\text{Calibrated}} = \bar{X}_{\text{Sample}} \times \hat{T}, \]

where \( \bar{X}_{\text{Sample}} = \) local sample mean, 
\[ \hat{T} = \frac{\sigma^2_{\text{Between-town}}}{(\sigma^2_{\text{Between-town}} + \frac{\sigma^2_{\text{Within-town}}}{n_{\text{Samples}}})}, \]

and \( X_{\text{Calibrated}} = \) the transformed surrogate exposure metric.

We simulated 1,000 iterations of each combination of classical and Berkson error patterns. We present mean logistic regression coefficients (β), which represent log odds ratios, and their standard errors as estimates of the effect of HAA5 on adverse birth outcomes under various simulated scenarios. We present the ratio of observed and calibrated effect estimates to the true effect (i.e., odds ratio = 2.00; β = 0.693), as well as coverage proportions for both observed and calibrated results. Coverage proportions are the proportion of estimated 2-sided 95% confidence intervals that include the true effect of exposure.
RESULTS

The results shown in Table 1 reflect the effects of classical measurement error, with Berkson error set to zero, that we observed in our simulations, based on a system-wide sample arithmetic mean of 35 μg/L and a standard deviation of 25 μg/L, which were converted to the log scale. These data, based on 1,000 simulated iterations, show the mean estimated β coefficient (log of the odds ratio), the mean of the model-based estimated variances, the empirical variance of the estimated β’s, and coverage proportions for several different ratios of variances within and between individual local sampling areas (i.e., localities or towns) that comprise the distribution system. In the absence of within-locality variability, the ratio of within-locality variance to between-locality variance is zero and the sample mean equals the true locality mean.

All of the tables show that in the absence of classical measurement error, the estimated β’s are unbiased and the model-based variances approximate the empirical variances. When the ratio of within-locality variance to between-locality variance was 0.25, bias toward the null attributable to classical exposure measurement error was 6%, with the mean model-based variance of the estimated β’s being overestimated by 18% as compared with the mean model-based variance in the absence of classical measurement error shown in the first row of Table 1 and a coverage proportion of 91%. When the ratio of within-locality variance to between-locality variance was 1.00, bias toward the null was 20%, with the mean model-based variance being overestimated by 60% and a coverage probability of 59%. When the ratio of within-locality variance to between-locality variance was 4.00, bias toward the null was 50%, with the mean model-based variance being inflated 2.5-fold and a coverage probability of 9%. As expected (8), the biased results were equal to the product of the true effect and the square of the correlation between the subject’s true exposure and the assigned surrogate exposure (also known as the coefficient of reliability) and were independent of the magnitude of the true effect (results not shown).

Table 1 also shows the results of regression calibration analyses using observed sample data transformed by the function of the estimated within- and between-locality variances using our formula adapted from Reeves et al. (24). In these results, when the ratio of within- to between-locality variance was 0.25, there was no apparent bias in the estimated β’s and the associated coverage probability was 95%; however, the mean model-based variance was inflated by 34% as compared with the mean model-based variance in the absence of measurement error shown in the first row of Table 1. When the ratio of within- to between-locality variance was 1.00, the calibrated results showed a slight overestimation of 1% relative to the true effect, with an associated coverage probability of 91% with a 2.6-fold inflated mean model-based variance. When the ratio of within- to between-locality variance was 4.00, the calibrated results overestimated the true effect by 6% and had a coverage probability of 89% with a 20-fold inflated mean model-based variance.

We also examined how these findings depended on the relative magnitude of homoscedastic system-wide sample variance. The results shown in Table 2 reflect the effects of classical measurement error, again with Berkson error set to zero, based on a system-wide sample variance one-half of that which was simulated in Table 1, while the results in Table 3 reflect a system-wide variance double that in Table 1. Compared with the naive coverage estimates of 91%, 59%, and 9% (Table 1) for ratios of within- to between-locality variance of 0.25, 1.00, and 4.00, respectively, Table 2 shows

---

Table 1. Regression Calibration Results With Classical Measurement Error and No Berkson Error in a Simulated Water Distribution System With System-Wide Log-Scale Sample Variance of 0.412 μg/L

<table>
<thead>
<tr>
<th>Var ( W / B )</th>
<th>( \beta_{\text{Obs}} )</th>
<th>( \beta_{\text{Obs}} )</th>
<th>Empirical Variance</th>
<th>Coverage, %</th>
<th>( \beta_{\text{Obs}} )</th>
<th>( \beta_{\text{Obs}} )</th>
<th>Empirical Variance</th>
<th>Coverage, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.694</td>
<td>0.0042</td>
<td>0.0039</td>
<td>96.2</td>
<td>0.694</td>
<td>0.0042</td>
<td>0.0039</td>
<td>96.2</td>
</tr>
<tr>
<td>0.25</td>
<td>0.653</td>
<td>0.0049</td>
<td>0.0051</td>
<td>90.8</td>
<td>0.695</td>
<td>0.0056</td>
<td>0.0058</td>
<td>94.8</td>
</tr>
<tr>
<td>0.33</td>
<td>0.641</td>
<td>0.0052</td>
<td>0.0054</td>
<td>89.0</td>
<td>0.695</td>
<td>0.0061</td>
<td>0.0064</td>
<td>94.2</td>
</tr>
<tr>
<td>0.50</td>
<td>0.616</td>
<td>0.0056</td>
<td>0.0060</td>
<td>80.9</td>
<td>0.696</td>
<td>0.0072</td>
<td>0.0078</td>
<td>93.9</td>
</tr>
<tr>
<td>0.67</td>
<td>0.594</td>
<td>0.0060</td>
<td>0.0066</td>
<td>74.4</td>
<td>0.697</td>
<td>0.0083</td>
<td>0.0096</td>
<td>93.0</td>
</tr>
<tr>
<td>1.00</td>
<td>0.554</td>
<td>0.0067</td>
<td>0.0077</td>
<td>59.4</td>
<td>0.698</td>
<td>0.0108</td>
<td>0.0133</td>
<td>91.0</td>
</tr>
<tr>
<td>1.50</td>
<td>0.503</td>
<td>0.0077</td>
<td>0.0090</td>
<td>39.5</td>
<td>0.702</td>
<td>0.0153</td>
<td>0.0204</td>
<td>90.7</td>
</tr>
<tr>
<td>2.00</td>
<td>0.461</td>
<td>0.0084</td>
<td>0.0100</td>
<td>28.7</td>
<td>0.706</td>
<td>0.0207</td>
<td>0.0297</td>
<td>89.8</td>
</tr>
<tr>
<td>3.00</td>
<td>0.395</td>
<td>0.0097</td>
<td>0.0116</td>
<td>16.3</td>
<td>0.723</td>
<td>0.0367</td>
<td>0.0748</td>
<td>89.6</td>
</tr>
<tr>
<td>4.00</td>
<td>0.347</td>
<td>0.0106</td>
<td>0.0127</td>
<td>9.4</td>
<td>0.734</td>
<td>0.0821</td>
<td>0.1900</td>
<td>88.6</td>
</tr>
</tbody>
</table>

Abbreviation: HAA5, 5 haloacetic acids (sum of concentrations of mono-, di-, and trichloroacetic acids and mono- and dibromoacetic acids).

a Results are based on 1,000 simulations using a true odds ratio of 2.0, represented as a true logistic regression coefficient of 0.693 per ln(20-μg/L) change in HAA5.
b \( Var(W) \): within-locality variance; \( Var(B) \): between-locality variance.
c Mean of 1,000 estimated β coefficients.
d Mean of 1,000 estimated variances of β coefficients.
e Variance of 1,000 estimated β coefficients.
larger coverage estimates of 93%, 77%, and 33%, while Table 3 shows smaller coverage estimates of 84%, 32%, and 1%. The calibrated coverage estimates were less sensitive to large changes in the system-wide sample variance. The naive and calibrated model-based variances varied in inverse proportion to the relative multiplicative change in system-wide variability (Tables 1–3).

The results of adding different magnitudes of Berkson measurement error to the classical error simulation results based on the sample variance of 0.4123 μg/L (from Table 1) are presented in Table 4 and Table 5. Table 4 shows results from adding Berkson error to represent interindividual variability in factors such as volume of water consumed per day; percentage of water consumed at or away from

Table 2. Regression Calibration Results With Classical Measurement Error and No Berkson Error in a Simulated Water Distribution System With System-Wide Log-Scale Sample Variance of 0.206 μg/L

<table>
<thead>
<tr>
<th>Var(W) Var(B)</th>
<th>Variance, %</th>
<th>Coverage, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean b(b) Obs</td>
<td>Mean Model-based var(b) Obs</td>
<td>Empirical Variance b(b) Obs</td>
</tr>
<tr>
<td>Naive</td>
<td>Calibrated</td>
<td>Naive</td>
</tr>
<tr>
<td>0.0</td>
<td>0.694</td>
<td>0.0021</td>
</tr>
<tr>
<td>0.25</td>
<td>0.662</td>
<td>0.0024</td>
</tr>
<tr>
<td>0.33</td>
<td>0.639</td>
<td>0.0025</td>
</tr>
<tr>
<td>0.50</td>
<td>0.615</td>
<td>0.0028</td>
</tr>
<tr>
<td>0.67</td>
<td>0.592</td>
<td>0.0030</td>
</tr>
<tr>
<td>1.00</td>
<td>0.553</td>
<td>0.0033</td>
</tr>
<tr>
<td>1.50</td>
<td>0.502</td>
<td>0.0038</td>
</tr>
<tr>
<td>2.00</td>
<td>0.461</td>
<td>0.0042</td>
</tr>
<tr>
<td>3.00</td>
<td>0.395</td>
<td>0.0048</td>
</tr>
<tr>
<td>4.00</td>
<td>0.345</td>
<td>0.0052</td>
</tr>
</tbody>
</table>

Abbreviation: HAA5, 5 haloacetic acids (sum of concentrations of mono-, di-, and trichloroacetic acids and mono- and dibromoacetic acids).

Results are based on 1,000 simulations using a true odds ratio of 2.0, represented as a true logistic regression β coefficient of 0.693 per ln(20-μg/L) change in HAA5.

Var(W): within-locality variance; Var(B): between-locality variance.

Mean of 1,000 estimated β coefficients.

Mean of 1,000 estimated variances of β coefficients.

Variance of 1,000 estimated β coefficients.

Table 3. Regression Calibration Results With Classical Measurement Error and No Berkson Error in a Simulated Water Distribution System With System-Wide Log-Scale Sample Variance of 0.908 μg/L

<table>
<thead>
<tr>
<th>Var(W) Var(B)</th>
<th>Variance, %</th>
<th>Coverage, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean b(b) Obs</td>
<td>Mean Model-based var(b) Obs</td>
<td>Empirical Variance b(b) Obs</td>
</tr>
<tr>
<td>Naive</td>
<td>Calibrated</td>
<td>Naive</td>
</tr>
<tr>
<td>0.0</td>
<td>0.694</td>
<td>0.0021</td>
</tr>
<tr>
<td>0.25</td>
<td>0.662</td>
<td>0.0024</td>
</tr>
<tr>
<td>0.33</td>
<td>0.639</td>
<td>0.0025</td>
</tr>
<tr>
<td>0.50</td>
<td>0.615</td>
<td>0.0028</td>
</tr>
<tr>
<td>0.67</td>
<td>0.592</td>
<td>0.0030</td>
</tr>
<tr>
<td>1.00</td>
<td>0.553</td>
<td>0.0033</td>
</tr>
<tr>
<td>1.50</td>
<td>0.502</td>
<td>0.0038</td>
</tr>
<tr>
<td>2.00</td>
<td>0.461</td>
<td>0.0042</td>
</tr>
<tr>
<td>3.00</td>
<td>0.395</td>
<td>0.0048</td>
</tr>
<tr>
<td>4.00</td>
<td>0.345</td>
<td>0.0052</td>
</tr>
</tbody>
</table>

Abbreviation: HAA5, 5 haloacetic acids (sum of concentrations of mono-, di-, and trichloroacetic acids and mono- and dibromoacetic acids).

Results are based on 1,000 simulations using a true odds ratio of 2.0, represented as a true logistic regression β coefficient of 0.693 per ln(20-μg/L) change in HAA5.

Var(W): within-locality variance; Var(B): between-locality variance.

Mean of 1,000 estimated β coefficients.

Mean of 1,000 estimated variances of β coefficients.

Variance of 1,000 estimated β coefficients.
Discussions as well as the empirical variances of the estimated variance by more than the same 1%–4%.

Tables 4 and 5 also show Berkson error attenuated both the naive and calibrated effect estimates by approximately 4%. The introduction of more measurement error, Berkson error consistently decreases sample variance on the log scale. The impact of exposure measurement error on results, while only 29% qualitatively described the possible effects (45). The impact of exposure measurement error in epidemiologic studies appears to be ignored in practice.

We have shown that bias that results from classical measurement error can be reduced using variance parameters directly estimable from observed data when there are a fixed number of multiple exposure measurements within multiple local areas (e.g., homes, neighborhoods, or towns). This assumes that the available monitoring data adequately captured the true within-area variance. Observed within- and between-area variances can be used to compute study-specific transformations of surrogate exposure measures that yield the true effect of exposure, akin to regression calibration using an external validation sample. We have also shown, in Tables 4 and 5, how bias that results from Berkson measurement error, while not directly estimable from the observed data, behaves in combination with classical error.

Nondifferential measurement error, including classical and Berkson error, can result in misclassification of exposures and outcomes in epidemiologic studies. These simulations show that regression calibration techniques can be used to reduce the bias from classical error. On the basis of these results, we have shown that classical measurement error attenuates effect estimates considerably, while Berkson measurement error attenuation is minimal under the specified conditions used in the simulations. These data suggest that random interindividual variability in water intake habits with respect to the true local exposures and within a range typical of personal variability in DBP

### Table 4. Regression Calibration Results With Classical Measurement Error and Berkson Error Equal to the System-Wide Sample Variance in a Simulated Water Distribution System With System-Wide Log-Scale Sample Variance of 0.412 μg/L

<table>
<thead>
<tr>
<th>Var(W) W/Var(B)</th>
<th>Naive Mean β̂_Obs</th>
<th>Naive Mean Model-based Variance Var(β̂_Obs)</th>
<th>Empirical Variance Var(β̂_Obs)</th>
<th>Coverage, %</th>
<th>Calibrated Mean β̂_Obs</th>
<th>Calibrated Mean Model-based Variance Var(β̂_Obs)</th>
<th>Empirical Variance Var(β̂_Obs)</th>
<th>Coverage, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.689</td>
<td>0.0041</td>
<td>0.0037</td>
<td>95.5</td>
<td>0.689</td>
<td>0.0041</td>
<td>0.0037</td>
<td>95.5</td>
</tr>
<tr>
<td>0.25</td>
<td>0.648</td>
<td>0.0048</td>
<td>0.0049</td>
<td>89.3</td>
<td>0.6906</td>
<td>0.0054</td>
<td>0.0056</td>
<td>94.9</td>
</tr>
<tr>
<td>0.33</td>
<td>0.636</td>
<td>0.0052</td>
<td>0.0050</td>
<td>86.4</td>
<td>0.6902</td>
<td>0.0063</td>
<td>0.0059</td>
<td>94.1</td>
</tr>
<tr>
<td>0.50</td>
<td>0.612</td>
<td>0.0054</td>
<td>0.0059</td>
<td>78.6</td>
<td>0.6908</td>
<td>0.0070</td>
<td>0.0077</td>
<td>93.4</td>
</tr>
<tr>
<td>0.67</td>
<td>0.589</td>
<td>0.0058</td>
<td>0.0065</td>
<td>71.7</td>
<td>0.6915</td>
<td>0.0081</td>
<td>0.0093</td>
<td>93.3</td>
</tr>
<tr>
<td>1.00</td>
<td>0.550</td>
<td>0.0065</td>
<td>0.0075</td>
<td>57.1</td>
<td>0.6937</td>
<td>0.0105</td>
<td>0.0130</td>
<td>92.1</td>
</tr>
<tr>
<td>1.50</td>
<td>0.500</td>
<td>0.0074</td>
<td>0.0088</td>
<td>38.5</td>
<td>0.6977</td>
<td>0.0148</td>
<td>0.0201</td>
<td>90.9</td>
</tr>
<tr>
<td>2.00</td>
<td>0.458</td>
<td>0.0082</td>
<td>0.0099</td>
<td>27.2</td>
<td>0.7017</td>
<td>0.0201</td>
<td>0.0300</td>
<td>90.0</td>
</tr>
<tr>
<td>3.00</td>
<td>0.393</td>
<td>0.0094</td>
<td>0.0114</td>
<td>14.8</td>
<td>0.7183</td>
<td>0.0356</td>
<td>0.0810</td>
<td>89.2</td>
</tr>
<tr>
<td>4.00</td>
<td>0.344</td>
<td>0.0103</td>
<td>0.0124</td>
<td>8.4</td>
<td>0.7272</td>
<td>0.0797</td>
<td>0.1897</td>
<td>89.1</td>
</tr>
</tbody>
</table>

Abbreviation: HAA5, 5 haloacetic acids (sum of concentrations of mono-, di-, and trichloroacetic acids and mono- and dibromoacetic acids).

a Results are based on 1,000 simulations using a true odds ratio of 2.0, represented as a true logistic regression coefficient of 0.693 per ln(20-μg/L) change in HAA5.

b Var(W): within-locality variance; Var(B): between-locality variance.

c Mean of 1,000 estimated β coefficients.

d Mean of 1,000 estimated variances of β coefficients.

e Variance of 1,000 estimated β coefficients.
exposures may not result in much bias. Therefore, boiling, filtering, and consumption volumes are unlikely to have any noticeable impact on the observed effect estimates unless they produce systematic error that is related to the true underlying locality mean concentration. For example, subjects who perceive that their water quality is poor may consume only bottled water, which generally has low levels of chlorination by-products. Since these subjects would be truly unexposed to DBPs via ingestion but classified as exposed, this could skew any measurement error and cause bias. The results presented here assume that the mean error is zero and this could affect any measurement error and cause bias.

We examined variation in personal exposure up to 2 times as high as the variability in haloacetic acid concentrations across localities on the log scale. The maximum simulated untransformed maternal HAA5 exposure concentrations were 270 µg/L and 73,130 µg/L when the Berkson errors were equal to and twice the size of the variability in the system-wide distribution, respectively. This extremely large range of simulated maternal exposures should capture the exposure experience of individuals in the outer tails of most distributions. Our data suggest that investigators in DBP studies that do not have personal exposure measurements can still accurately assess putative effects, as long as there is no systematic exposure measurement error.

We based our baseline simulations on a generally homoscedastic system-wide sample geometric standard deviation of 0.642 µg/L to illustrate the effects of classical measurement error. This was based on the combined variance of several water distribution systems in Massachusetts but is also consistent with previous DBP studies (4, 46–48). Our simulations suggest that the absolute magnitude of system-wide variance does not determine the magnitude of bias for exposures based on group-level surrogate data. Rather, it is the ratio of within- to between-locality variances that predicts the bias. However, higher or lower system-wide variability does affect coverage proportions, with lower variability being associated with more accurate coverage. Most public water distribution systems typically have much less within-system variability relative to variability between localities. Some epidemiologic studies have targeted water systems with low spatial variability, including those utilizing certain types of disinfection such as chloramination to reduce DBP formation (15, 49). In contrast, larger studies examining many different water systems have greater between-system variability due to variable disinfection practices and drinking water sources (i.e., groundwater, surface water, or mixed supplies). The potential for calibration of measurement error reinforces the need to adequately characterize within- and between-locality variability. Increasing the number of samples per locality to 10 (instead of 4) generally improved the performance of the calibration method—especially when the ratio of within-locality variance to between-locality variance was high. We compared the results in Table 1 (no Berkson error, 4 samples per locality) for different scenarios. Results were substantially unchanged when the ratio of within- to between-locality variance was small (results not shown). However, in scenarios where the within-locality variance was 4 times larger than the between-locality variance, a larger sample size per locality yielded substantial improvement in the performance of our
regression calibration method. Compared with the results shown in Table 1 for a within:between-locality ratio of 4, the calibrated results based on more samples per locality were less biased (mean $\hat{\beta}_{\text{Obs}}$: 0.703 vs. 0.734) and had improved coverage (92.5% vs. 88.6%).

One artifact of this analysis occurred when the ratio of within- to between-locality variability was very high; the calibrated results showed a tendency to inflate the estimated effect of exposure. This resulted from estimated within- and between-locality variances from PROC VARCOMP in SAS, version 9.1, that were slightly different than was specified in the simulation design. The variance estimates tended to overestimate the true within-locality variance while slightly underestimating the between-locality variance. This led to an underestimate of the transformation scalar of the surrogate exposure, which had the impact of slightly overestimating the associated risk measures. Supplemental simulations that used the true variances rather than those estimated from the observed data did not produce higher calibrated effect estimates at higher ratios of within- to between-locality error (data not shown). Alternative computational procedures that more accurately estimate the within- and between-locality variances would diminish this trend and might correspondingly improve the coverage proportions of the calibrated results. However, the relevant range of this important ratio is at the lower end of our simulations—typically less than unity for DBPs such as trihalomethanes (48). Within this range, our method provides calibrated results that are nearly unbiased.

Risk assessment seeks to characterize the true, underlying effect of exposures that may be biased or otherwise obscured by various uncertainties; therefore, it is critical that investigators make use of all possible tools that aid estimation of this risk. We have presented a method with the potential to reduce bias due to classical and Berkson measurement error. This method should allow for hierarchically designed investigations of environmental exposures based on local area means of several samples to be calibrated for this bias, thereby providing effect estimates that are less biased and less uncertain. Comparing the results without Berkson error (Table 1) to those with Berkson error (Tables 4 and 5), we have shown that unmeasured Berkson measurement error induces only a small bias in the direction of the null.

The central strength of our regression calibration method of correcting for attenuation of effect estimates due to classical exposure measurement error is that it does not require the use of a supplemental validation study. This method can be implemented using routinely collected local surrogate exposure data. Since our method depends on estimating the within- and between-locality variability in a water distribution system, it is likely to be more effective when there are more exposure measures within each locality. While this method may be most useful in primary data analyses to show the extent of the magnitude of bias from measurement error, previously published results may be amenable to calibration if estimates of the within- and between-locality variances are available. This seems particularly relevant for epidemiologic studies of DBPs, since only 2 studies examining the impact of DBPs on fetal growth have incorporated data on individual water use (10–18, 39–42). Since the true magnitude of associations may have been underestimated in these studies, assessing this potential bias and other sources of uncertainty is essential to estimating any causal effect of exposure.

ACKNOWLEDGMENTS

Author affiliations: National Center for Environmental Assessment, Office of Research and Development, Environmental Protection Agency, Washington, DC (Thomas F. Bateson) and Cincinnati, Ohio (J. Michael Wright).

This work was supported by the Environmental Protection Agency.

The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the Environmental Protection Agency.

Conflict of interest: none declared.

REFERENCES


