Practice of Epidemiology

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 8623P, Washington, DC 20460 (e-mail: bateson.thomas@epa.gov).

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 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.

bias (epidemiology); environmental exposure; epidemiologic methods; measurement error; misclassification; regression analysis; surrogate marker; water pollution

Abbreviations: DBPs, disinfection by-products; HAA5, 5 haloacetic acids.

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

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 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 nondifferential 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 (σ_B^2). The between-locality variance was a function of the system-wide variance (0.4123 μg/L) and the specified ratio of within-locality variance (σ_W^2) to between-locality variance, such that $\sigma_B^2 = [0.4123/(1 + \rho)]$, where $\rho = \sigma_W^2/\sigma_B^2$. 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 (σ_W^2) was assumed to follow a 0-mean normal distribution with variance σ_{Wj}^2.

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
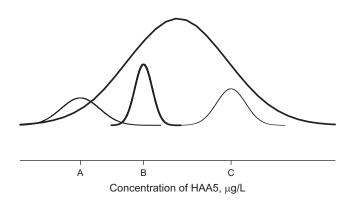
Figure 1. Simulated system-wide and locality-specific distributions of concentrations of 5 haloacetic acids (HAA5). HAA5 is the sum of concentrations of mono-, di-, and trichloroacetic acids and mono- and dibromoacetic acids. The largest line represents the simulated system-wide distribution. The 3 smaller lines represent HAA5 distributions for 3 different theoretical localities (towns A, B, and C).

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
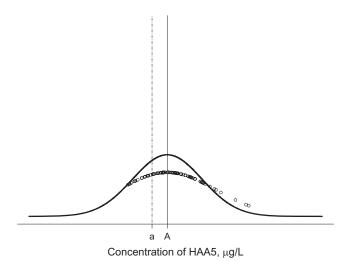

Figure 2. Underlying distribution of local concentrations of 5 haloacetic acids (HAA5), with superimposed sample values for 1 locality. HAA5 is the sum of concentrations of mono-, di-, and trichloroacetic acids and mono- and dibromoacetic acids. The taller and thicker density function shows the true underlying distribution of ln(HAA5) concentrations in that locality. The shorter and thinner density function shows the true maternal exposures with Berkson error. a: Observed sample mean of 4 random HAA5 samples. A: True assigned mean of the hypothetical HAA5 distribution. The large circles represent 4 randomly selected HAA5 samples; the small circles represent the true assigned exposure for 100 subjects.

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,

$$\hat{T} = \hat{\sigma}^2_{\text{Between-town}} \big/ \Big(\hat{\sigma}^2_{\text{Between-town}} + \frac{\hat{\sigma}^2_{\text{Within-town}}}{n_{\text{Samples}}} \Big),$$

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.

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[a]

$\frac{Var(W)_b}{Var(B)}$	Naive				Calibrated			
	Mean $\hat{\beta}_{Obs}$[c]	Mean Model-based $Var(\hat{\beta}_{Obs})$[d]	Empirical Variance $\hat{\beta}_{Obs}$[e]	Coverage, %	Mean $\hat{\beta}_{Obs}$	Mean Model-based $Var(\hat{\beta}_{Obs})$	Empirical Variance $\hat{\beta}_{Obs}$	Coverage, %
0	0.694	0.0042	0.0039	96.2	0.694	0.0042	0.0039	96.2
0.25	0.653	0.0049	0.0051	90.8	0.695	0.0056	0.0058	94.8
0.33	0.641	0.0052	0.0054	89.0	0.695	0.0061	0.0064	94.2
0.50	0.616	0.0056	0.0060	80.9	0.696	0.0072	0.0078	93.9
0.67	0.594	0.0060	0.0066	74.4	0.697	0.0083	0.0096	93.0
1.00	0.554	0.0067	0.0077	59.4	0.698	0.0108	0.0133	91.0
1.50	0.503	0.0077	0.0090	39.5	0.702	0.0153	0.0204	90.7
2.00	0.461	0.0084	0.0100	28.7	0.706	0.0207	0.0297	89.8
3.00	0.395	0.0097	0.0116	16.3	0.723	0.0367	0.0748	89.6
4.00	0.347	0.0106	0.0127	9.4	0.734	0.0821	0.1900	88.6

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 β 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.

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 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[a]

$\frac{Var(W)_b}{Var(B)}$	Naive				Calibrated			
	Mean $\hat{\beta}_{Obs}$[c]	Mean Model-based Var($\hat{\beta}_{Obs}$)[d]	Empirical Variance $\hat{\beta}_{Obs}$[e]	Coverage, %	Mean $\hat{\beta}_{Obs}$	Mean Model-based Var($\hat{\beta}_{Obs}$)	Empirical Variance $\hat{\beta}_{Obs}$	Coverage, %
0	0.693	0.0084	0.0081	95.0	0.693	0.0084	0.0081	95.0
0.25	0.652	0.0099	0.0100	92.7	0.694	0.0113	0.0114	94.6
0.33	0.640	0.0104	0.0105	92.2	0.694	0.0123	0.0125	94.4
0.50	0.615	0.0113	0.0117	88.3	0.695	0.0145	0.0151	95.0
0.67	0.593	0.0121	0.0127	85.0	0.696	0.0168	0.0180	94.4
1.00	0.554	0.0135	0.0147	77.3	0.698	0.0219	0.0245	93.6
1.50	0.504	0.0154	0.0172	66.5	0.703	0.0308	0.0367	92.7
2.00	0.462	0.0170	0.0190	56.5	0.707	0.0416	0.0516	92.3
3.00	0.397	0.0194	0.0221	42.9	0.725	0.0737	0.1124	91.4
4.00	0.348	0.0213	0.0240	33.4	0.730	0.1661	0.2820	91.1

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 β 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 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[a]

$\frac{Var(W)_b}{Var(B)}$	Naive				Calibrated			
	Mean $\hat{\beta}_{Obs}$[c]	Mean Model-based Var($\hat{\beta}_{Obs}$)[d]	Empirical Variance $\hat{\beta}_{Obs}$[e]	Coverage, %	Mean $\hat{\beta}_{Obs}$	Mean Model-based Var($\hat{\beta}_{Obs}$)	Empirical Variance $\hat{\beta}_{Obs}$	Coverage, %
0	0.694	0.0021	0.0020	95.4	0.694	0.0021	0.0020	95.4
0.25	0.652	0.0024	0.0027	84.4	0.694	0.0028	0.0031	93.4
0.33	0.639	0.0025	0.0029	79.4	0.694	0.0030	0.0035	93.4
0.50	0.615	0.0028	0.0032	65.5	0.694	0.0035	0.0043	92.4
0.67	0.592	0.0030	0.0036	52.1	0.695	0.0041	0.0053	91.0
1.00	0.553	0.0033	0.0042	32.3	0.697	0.0054	0.0076	89.6
1.50	0.502	0.0038	0.0051	16.3	0.701	0.0075	0.0124	87.8
2.00	0.461	0.0042	0.0056	8.2	0.707	0.0102	0.0187	86.6
3.00	0.395	0.0048	0.0066	3.0	0.722	0.0181	0.0570	85.0
4.00	0.345	0.0052	0.0071	1.0	0.735	0.0405	0.1757	84.4

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 β 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.

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[a]

$\frac{\text{Var}(W)_b}{\text{Var}(B)}$	Naive				Calibrated			
	Mean $\hat{\beta}_{Obs}$ [c]	Mean Model-based $\text{Var}(\hat{\beta}_{Obs})$ [d]	Empirical Variance $\hat{\beta}_{Obs}$ [e]	Coverage, %	Mean $\hat{\beta}_{Obs}$	Mean Model-based $\text{Var}(\hat{\beta}_{Obs})$	Empirical Variance $\hat{\beta}_{Obs}$	Coverage, %
0	0.689	0.0041	0.0037	95.5	0.689	0.0041	0.0037	95.5
0.25	0.648	0.0048	0.0049	89.3	0.6896	0.0054	0.0056	94.9
0.33	0.636	0.0052	0.0050	86.4	0.6902	0.0063	0.0059	94.1
0.50	0.612	0.0054	0.0059	78.6	0.6908	0.0070	0.0077	93.4
0.67	0.589	0.0058	0.0065	71.7	0.6915	0.0081	0.0093	93.3
1.00	0.550	0.0065	0.0075	57.1	0.6937	0.0105	0.0130	92.1
1.50	0.500	0.0074	0.0088	38.5	0.6977	0.0148	0.0201	90.9
2.00	0.458	0.0082	0.0099	27.2	0.7017	0.0201	0.0300	90.0
3.00	0.393	0.0094	0.0114	14.8	0.7183	0.0356	0.0810	89.2
4.00	0.344	0.0103	0.0124	8.4	0.7272	0.0797	0.1897	89.1

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 β 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.

home; percentage of boiled, filtered, or bottled water consumed; and effectiveness of water filtration equal to system-wide sample variance. When combined with classical measurement error, Berkson error consistently decreases both the naive and calibrated effect estimates and their coverage proportions by less than 1%. Table 5 shows the results of adding Berkson error equal to twice the system-wide sample variance on the log scale. The introduction of more Berkson error attenuated both the naive and calibrated effect estimates by approximately 4%. Tables 4 and 5 also show that additional Berkson error does not influence the model-based variances by more than the same 1%–4%.

Tables 1–5 show the mean of the 1,000 model-based variances as well as the empirical variances of the estimated β coefficients. The mean model-based variances from the naive analyses were smaller than the empirical variances, and the differences increased with additional classical measurement error. The same pattern of differences was observed for the calibrated results. However, when the ratio of the within-locality variance to the between-locality variance in exposure was smaller (≤0.5), the mean model-based variances were within 10% of the empirical variances.

DISCUSSION

Many environmental epidemiologic studies estimate individuals' personal exposures using local area samples, which are considered indirect surrogate measures of true local pollutant concentrations. This results in classical nondifferential measurement error, which is known to usually bias regression results toward the null. Although the magnitude of bias is often unknown, most investigators simply state that had there been no measurement error the estimated

effect estimate would have been larger. In only 1 of 57 studies published in 3 prominent epidemiology journals over a 1-year period did the researchers quantify the likely 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 5. Regression Calibration Results With Classical Measurement Error and Berkson Error Equal to Twice the System-Wide Sample Variance in a Simulated Water Distribution System With System-Wide Log-Scale Sample Variance of 0.412 μg/L[a]

$\frac{\text{Var}(W)_b}{\text{Var}(B)}$	Naive				Calibrated			
	Mean $\hat{\beta}_{Obs}$[c]	Mean Model-based Var($\hat{\beta}_{Obs}$)[d]	Empirical Variance $\hat{\beta}_{Obs}$[e]	Coverage, %	Mean $\hat{\beta}_{Obs}$	Mean Model-based Var($\hat{\beta}_{Obs}$)	Empirical Variance $\hat{\beta}_{Obs}$	Coverage, %
0	0.669	0.0035	0.0038	92.7	0.669	0.0035	0.0038	92.7
0.25	0.629	0.0044	0.0045	82.3	0.669	0.0050	0.0051	92.8
0.33	0.617	0.0046	0.0048	79.0	0.670	0.0054	0.0057	93.1
0.50	0.593	0.0050	0.0053	70.8	0.670	0.0064	0.0071	93.0
0.67	0.571	0.0054	0.0059	61.3	0.670	0.0075	0.0086	92.1
1.00	0.533	0.0060	0.0067	44.8	0.672	0.0097	0.0118	91.6
1.50	0.485	0.0068	0.0078	29.5	0.676	0.0136	0.0181	90.8
2.00	0.443	0.0075	0.0086	19.5	0.680	0.0184	0.0266	90.5
3.00	0.380	0.0086	0.0098	9.6	0.696	0.0327	0.0353	89.5
4.00	0.332	0.0094	0.0108	5.4	0.704	0.073	0.1787	88.0

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 β 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 on the basis of the locality-level distribution system data, this could skew any measurement error and cause bias. The results presented here assume that the mean error is zero and the behavior of this method of regression calibration may change under alternative conditions. The potential impact of this kind of measurement error is much more difficult to predict and correct for in epidemiologic studies.

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 var-

iance 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 $\hat{\beta}_{Obs}$: 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.

REFERENCES

1. Whitaker HJ, Nieuwenhuijsen MJ, Best NG. The relationship between water concentrations and individual uptake of chloroform: a simulation study. *Environ Health Perspect.* 2003; 111(5):688–694.
2. Wright JM, Bateson TF. A sensitivity analysis of bias in relative risk estimates due to disinfection by-product exposure misclassification. *J Expo Anal Environ Epidemiol.* 2005; 15(3):212–216.
3. Waller K, Swan SH, Windham GC, et al. Influence of exposure assessment methods on risk estimates in an epidemiologic study of total trihalomethane exposure and spontaneous abortion. *J Expo Anal Environ Epidemiol.* 2001;11(6): 522–531.
4. King WD, Dodds L, Armson BA, et al. Exposure assessment in epidemiologic studies of adverse pregnancy outcomes and disinfection byproducts. *J Expo Anal Environ Epidemiol.* 2004;14(6):466–472.
5. Rothman K, Greenland S, eds. *Modern Epidemiology.* 2nd ed. Philadelphia, PA: Lippincott-Raven Publishers; 1998.
6. Zeger SL, Thomas D, Dominici F, et al. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ Health Perspect.* 2000;108(5): 419–426.
7. Carroll RJ, Ruppert D, Stefanski LA, et al. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. London, United Kingdom: Chapman & Hall Ltd; 2006.
8. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med.* 1998;55(10):651–656.
9. Heid IM, Küchenhoff H, Miles J, et al. Two dimensions of measurement error: classical and Berkson error in residential radon exposure assessment. *J Expo Anal Environ Epidemiol.* 2004;14(5):365–377.
10. Kramer MD, Lynch CF, Isacson P, et al. The association of waterborne chloroform with intrauterine growth retardation. *Epidemiology.* 1992;3(5):407–413.
11. Bove FJ, Fulcomer MC, Klotz JB, et al. Public drinking water contamination and birth outcomes. *Am J Epidemiol.* 1995; 141(9):850–862.

12. Dodds L, King W, Woolcott C, et al. Trihalomethanes in public water supplies and adverse birth outcomes. *Epidemiology.* 1999;10(3):233–237.

13. Wright JM, Schwartz J, Dockery DW. Effect of trihalomethane exposure on fetal development. *Occup Environ Med.* 2003;60(3):173–180.

14. Wright JM, Schwartz J, Dockery DW. The effect of disinfection by-products and mutagenic activity on birth weight and gestational duration. *Environ Health Perspect.* 2004;112(8):920–925.

15. Hinckley AF, Bachand AM, Reif JS. Late pregnancy exposures to disinfection by-products and growth-related birth outcomes. *Environ Health Perspect.* 2005;113(12):1808–1813.

16. Porter CK, Putnam SD, Hunting KL, et al. The effect of trihalomethane and haloacetic acid exposure on fetal growth in a Maryland county. *Am J Epidemiol.* 2005;162(4):334–344.

17. Lewis C, Suffet IH, Ritz B. Estimated effects of disinfection by-products on birth weight in a population served by a single water utility. *Am J Epidemiol.* 2006;163(1):38–47.

18. Yang CY, Xiao ZP, Ho SC, et al. Association between trihalomethane concentrations in drinking water and adverse pregnancy outcome in Taiwan. *Environ Res.* 2007;104(3):390–395.

19. Perera FP, Rauh V, Tsai WY, et al. Effects of transplacental exposure to environmental pollutants on birth outcomes in a multiethnic population. *Environ Health Perspect.* 2003;111(2):201–205.

20. Ritz B, Wilhelm M, Hoggatt KJ, et al. Ambient air pollution and preterm birth in the environment and pregnancy outcomes study at the University of California, Los Angeles. *Am J Epidemiol.* 2007;166(9):1045–1052.

21. Srám RJ, Binková B, Dejmek J, et al. Ambient air pollution and pregnancy outcomes: a review of the literature. *Environ Health Perspect.* 2005;113(4):375–382.

22. Berkson J. Are there two regressions? *J Am Stat Assoc.* 1950;45(250):164–180.

23. Burr D. On errors-in-variables in binary regression—Berkson case. *J Am Stat Assoc.* 1988;83(403):739–743.

24. Reeves GK, Cox DR, Darby SC, et al. Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Stat Med.* 1998;17(19):2157–2177.

25. Kim HM, Yasui Y, Burstyn I. Attenuation in risk estimates in logistic and Cox proportional-hazards models due to group-based exposure assessment strategy. *Ann Occup Hyg.* 2006;50(6):623–635.

26. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr.* 1997;65(4 suppl):1179S–1186S.

27. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med.* 1989;8(9):1051–1069.

28. Willett WC. Correction for the effects of measurement error. In: *Nutritional Epidemiology.* New York, NY: Oxford University Press; 1990:272–291.

29. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *Am J Epidemiol.* 1992;136(11):1400–1413.

30. Hunter DJ, Spiegelman D, Adami HO, et al. Cohort studies of fat intake and the risk of breast cancer—a pooled analysis. *N Engl J Med.* 1996;334(6):356–361.

31. Tanasescu M, Leitzmann MF, Rimm EB, et al. Exercise type and intensity in relation to coronary heart disease in men. *JAMA.* 2002;288(16):1994–2000.

32. Bingham SA, Day NE, Luben R, et al. Dietary fibre in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC): an observational study. *Lancet.* 2003;361(9368):1496–1501.

33. Beulens JW, Rimm EB, Ascherio A, et al. Alcohol consumption and risk for coronary heart disease among men with hypertension. *Ann Intern Med.* 2007;146(1):10–19.

34. Wilson WE, Mage DT, Grant LD. Estimating separately personal exposure to ambient and nonambient particulate matter for epidemiology and risk assessment: why and how? *J Air Waste Manag Assoc.* 2000;50(7):1167–1183.

35. Sarnat JA, Long CM, Koutrakis P, et al. Using sulfur as a tracer of outdoor fine particulate matter. *Environ Sci Technol.* 2002;36(24):5305–5314.

36. Wilson WE, Brauer M. Estimation of ambient and non-ambient components of particulate matter exposure from a personal monitoring panel study. *J Expo Sci Environ Epidemiol.* 2006;16(3):264–274.

37. Dominici F, Zeger SL, Samet JM. A measurement error model for time-series studies of air pollution and mortality. *Biostatistics.* 2000;1(2):157–175.

38. Strand M, Vedal S, Rodes C, et al. Estimating effects of ambient PM(2.5) exposure on health using PM(2.5) component measurements and regression calibration. *J Expo Sci Environ Epidemiol.* 2006;16(1):30–38.

39. Källén BA, Robert E. Drinking water chlorination and delivery outcome—a registry-based study in Sweden. *Reprod Toxicol.* 2000;14(4):303–309.

40. Jaakkola JJ, Magnus P, Skrondal A, et al. Foetal growth and duration of gestation relative to water chlorination. *Occup Environ Med.* 2001;58(7):437–442.

41. Infante-Rivard C. Drinking water contaminants, gene polymorphisms, and fetal growth. *Environ Health Perspect.* 2004;112(11):1213–1216.

42. Hoffman CS, Mendola P, Savitz DA, et al. Drinking water disinfection by-product exposure and fetal growth. *Epidemiology.* 2008;19(5):729–737.

43. Wright JM, Murphy PA, Nieuwenhuijsen MJ, et al. The impact of water consumption, point-of-use filtration and exposure categorization on exposure misclassification of ingested drinking water contaminants. *Sci Total Environ.* 2006;366(1):65–73.

44. Zender R, Bachand AM, Reif JS. Exposure to tap water during pregnancy. *J Expo Anal Environ Epidemiol.* 2001;11(3):224–230.

45. Jurek AM, Maldonado G, Greenland S, et al. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *Eur J Epidemiol.* 2006(12);21:871–876.

46. Hinckley AF, Bachand AM, Nuckols JR, et al. Identifying public water facilities with low spatial variability of disinfection by-products for epidemiological investigations. *Occup Environ Med.* 2005;62(7):494–499.

47. Villanueva CM, Cantor KP, Grimalt JO, et al. Assessment of lifetime exposure to trihalomethanes through different routes. *Occup Environ Med.* 2006;63(4):273–277.

48. Keegan T, Whitaker H, Nieuwenhuijsen MJ, et al. Use of routinely collected data on trihalomethane in drinking water for epidemiological purposes. *Occup Environ Med.* 2001;58(7):447–452.

49. Savitz DA, Singer PC, Herring AH, et al. Exposure to drinking water disinfection by-products and pregnancy loss. *Am J Epidemiol.* 2006;164(11):1043–1051.