American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2023.

Vol. 192, No. 8
https://doi.org/10.1093/aje/kwad098
Advance Access publication:
April 21, 2023

Practice of Epidemiology

Issues in Implementing Regression Calibration Analyses

Lillian A. Boe*, Pamela A. Shaw, Douglas Midthune, Paul Gustafson, Victor Kipnis, Eunyoung Park, Daniela Sotres-Alvarez, and Laurence Freedman, on behalf of the Measurement Error and Misclassification Topic Group (TG4) of the STRATOS Initiative

* Correspondence to Dr. Lillian Boe, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 633 3rd Avenue, 3rd Floor, New York, NY 10017 (e-mail: boel@mskcc.org).

Initially submitted August 2, 2022; accepted for publication April 13, 2023.

Regression calibration is a popular approach for correcting biases in estimated regression parameters when exposure variables are measured with error. This approach involves building a calibration equation to estimate the value of the unknown true exposure given the error-prone measurement and other covariates. The estimated, or calibrated, exposure is then substituted for the unknown true exposure in the health outcome regression model. When used properly, regression calibration can greatly reduce the bias induced by exposure measurement error. Here, we first provide an overview of the statistical framework for regression calibration, specifically discussing how a special type of error, called Berkson error, arises in the estimated exposure. We then present practical issues to consider when applying regression calibration, including: 1) how to develop the calibration equation and which covariates to include; 2) valid ways to calculate standard errors of estimated regression coefficients; and 3) problems arising if one of the covariates in the calibration model is a mediator of the relationship between the exposure and outcome. Throughout, we provide illustrative examples using data from the Hispanic Community Health Study/Study of Latinos (United States, 2008–2011) and simulations. We conclude with recommendations for how to perform regression calibration.

Berkson error; bias (epidemiology); calibration equation; measurement error; nutritional epidemiology; regression calibration; STRATOS initiative; validation studies

Abbreviations: BMI, body mass index; CI, confidence interval; HCHS/SOL, Hispanic Community Health Study/Study of Latinos; OR, odds ratio; SE, standard error; SOLNAS, Study of Latinos: Nutrition and Physical Activity Assessment Study; STRATOS, Strengthening Analytical Thinking for Observational Studies.

Measurement error in exposures that are important to public health, such as dietary intakes, physical activity, and smoking, is becoming increasingly recognized as a problem (1–5), and methods for mitigating its effects are being researched and applied (6–11). A recent review found regression calibration to be the most commonly used method to adjust for bias in estimates of an exposure-health outcome association resulting from measurement error (11). However, implementing regression calibration requires care and, crucially, information regarding the errors in measuring the exposure that is generally acquired from a validation study. Here, we highlight issues that commonly arise in its implementation and illustrate them with examples drawn from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) (12) and simulations. This work is part of the Strengthening Analytical Thinking for Observational Studies (STRATOS) Initiative, an international effort aiming to provide accessible biostatistical guidance to elevate current practice in analyzing observational studies (13, 14). STRATOS comprises several working groups, including the Measurement Error and Misclassification Topic Group (TG4), focusing on areas of statistics where gaps exist between currently available statistical methodology and practice (13, 15).

We focus in this paper on the version of regression calibration described by Carroll et al. (6). The regression calibration of Rosner et al. (16) gives mathematically equivalent estimates for linear predictors (17), and is mainly used with

external validation studies (discussed below). Regression calibration (6, 18) involves developing a calibration equation that estimates the true exposure value for an individual based on their error-prone measured value and other covariates. This calibration estimate is then used in place of the unknown true exposure in the health outcome regression model. We illustrate the concepts in this paper with a simplified model that assumes the exposure is a linear effect with no interaction terms, but note that regression calibration may be applied more broadly to any outcome model. In The Basics, below, we explain the statistical basis for this approach, noting that the calibration estimate is itself an error-prone measure of the exposure but one with a particular type of error, called Berkson error (19). The method requires that this Berkson error be uncorrelated with the other outcome-model covariates.

We then consider 3 main topics:

1. The calibration equation—the validation study data required and choice of covariates. We emphasize the relationship between the covariates in the health outcome model and those in the calibration model.
2. The outcome model and appropriate methods for computing the standard error (SE) of the estimated association parameter.
3. The special problem when one of the calibration-model covariates is a mediator of the relationship between exposure and health outcome.

We conclude with a checklist of the issues raised.

THE BASICS

Regression calibration

Say we aim to learn the association between exposure X and health outcome Y, adjusted for confounding covariates Z. For instance, Y could be hypertension, X potassium intake, and Z age and sex. With data available on (Y, X, Z), fitting a regression model of Y on X and Z (the outcome model) allows examination of this association. For ease of presentation, we assume the model in question is a generalized linear model (including both linear and logistic regression as possibilities) or a Cox regression model, and that the correctly specified outcome regression model includes X as a linear effect with no interactions. Regression calibration may be applied more broadly to any outcome regression model. For example, Murad and Freedman (20) consider regression calibration applied to outcome models that include interactions between two continuously measured, error-prone covariates. Estimating the regression coefficient of X then accomplishes the task. We focus on X being a single exposure measure, but the concepts described can be extended to several exposures considered simultaneously (16).

When X is measured with error, as occurs when potassium intake is self-reported, more statistical effort is required. The available data are now (Y, X^*, Z), where X^* is an error-prone measurement of X. We assume the measurement error is nondifferential with respect to the outcome, meaning that X^* carries no more information about Y than is already provided by X and Z. In our example, this occurs if people with and without hypertension misreport similarly their potassium intake, a reasonable assumption if diet is reported before any diagnosis of hypertension. Even under this condition, plugging the (Y, X^*, Z) data into the outcome model and estimating the regression coefficient of X^* gives a biased estimate of the association of hypertension with potassium intake (3, 5, 8).

A popular method for avoiding this bias, regression calibration, is based on estimating each participant's unobserved X value given their observed X^* and Z values. In our example, we estimate true potassium intake from self-reported intake and other covariates. The outcome model is now fitted to $(Y, \hat{X}, Z)$ data, where $\hat{X}$ is the said estimate. Underlying theory prescribes that this estimate be formed as the conditional mean of X given X^* and Z, that is, $\hat{X} = E(X|X^*, Z)$. This expression is called the calibration equation. We call the resulting value of $\hat{X}$ the calibration estimate of X. In some linear outcome models, regression calibration exactly removes the bias, but more generally, this is only approximately true, including in our example, where hypertension, Y, a binary variable, is modeled using logistic regression. For logistic and Cox regression models, the remaining bias is small when there is only a modest association between X and Y, a relatively small amount of measurement error in X^*, or a rare outcome Y (6, 18, 21).

Using the calibration estimate $\hat{X}$ in the outcome model does lead to a higher variance of the estimated regression coefficient of X. In general, the lower the correlation between $\hat{X}$ and X, the greater the increase in variance of the estimated regression coefficient and the lower the resultant statistical power to detect the association (22).

Usually the calibration equation, $\hat{X} = E(X|X^*, Z)$ is unknown, and it must be estimated from validation data. These could be a sample of individuals for which (X, X^*, Z) are observed. The estimated regression model of X on X^* and Z approximates the needed equation for $\hat{X}$, but the extra uncertainty involved in this estimation should be accounted for in the subsequent analysis (see The Variance of the Estimated Exposure-Outcome Association, below).

Ideally, the validation data would consist of observations on (X, X^*, Z)—in our example, exact potassium intake, self-reported potassium intake, and age and sex. However, sometimes we might have only (X^{**}, X^*, Z), where X^{**} is an unbiased measurement of X with random error, so that $E(X^{**}|X^*, Z) = E(X|X^*, Z)$. Then the estimation procedure is still valid. In our example, X^{**} could be 24-hour urinary potassium, which is considered an unbiased measurement of potassium intake (3, 8). Sometimes, X^* may itself be an unbiased measurement of X with random error, in which case replicate measures of X^* provide sufficient data to form a calibration equation (see Web Appendix 1, available at https://doi.org/10.1093/aje/kwad098).

Berkson error

With most error-prone measurements X^*, the error in the measurement is correlated with X^*. In contrast, any surrogate measure $\hat{X}$ of an exposure X, is said to exhibit Berkson

error if $X = \hat{X} + U$, where the random error U has mean 0 and is independent of $\hat{X}$. Berkson error occurs in occupational health studies, when individuals are assigned exposures equal to the mean exposure of their occupational subgroup. Berkson error is generally viewed as not causing bias in estimating associations, which is sometimes true. If error U is nondifferential with respect to outcome Y and is uncorrelated with all confounders Z, then using $\hat{X}$ in place of X in the outcome model does yield an unbiased estimate of the regression coefficient (23). As explained below, this fact is the basis for the regression calibration method.

Regression calibration through the Berkson error lens

A natural question arises about why regression calibration works. When we cannot measure X exactly but have an error-prone measurement X^*, we are told that substituting X^* for X in the regression of Y on X and Z gives a biased estimate of the association of X with Y. Regression calibration tells us to use $\hat{X} = E(X|X^*, Z)$ in place of X. But $\hat{X}$ itself is an error-prone measurement of X, so how have we improved our situation? The answer is that $\hat{X}$ has Berkson error, whereas X^* (usually) does not. Moreover, by defining $\hat{X} = E(X|X^*, Z)$, the Berkson error term, U, is guaranteed to be uncorrelated with the confounding covariates Z.

Viewing regression calibration this way reinforces the hindrances placed upon the study investigator. We must estimate the exposure using $\hat{X} = E(X|X^*, Z)$; that is, we must use the same variables Z in the calibration equation as in the outcome model. This prohibits using a single all-purpose calibration equation for an exposure. In our example, estimating potassium intake as a function of self-reported potassium intake, age, and sex is appropriate only for assessing the association between potassium intake and hypertension given age and sex. If we adjust the association for another confounder, such as socioeconomic status, then that confounder needs to be included in the calibration equation. Other principles in forming the calibration equation are found in The Calibration Equation section, below.

FORMING THE CALIBRATION EQUATION

The required data

To form the calibration equation $E(X|X^*, Z)$, data are needed on X^*, Z, and either X or an unbiased measure of X, denoted by X^{**} (see Regression Calibration, above). Where possible, these data should be collected in an internal validation study (i.e., the participants should be a subgroup of the main study cohort). Internal validation is preferable for several reasons. First, covariates Z should include the confounders of the outcome model. With internal validation, these are naturally available. Second, the method of measuring X^* should be the same in the validation study as in the main study.

Third, the calibration equation derived from the validation study should be transportable (i.e., applicable) to the main study data (Sections 2.2.4–2.2.5 in Carroll et al. (6)). With an internal validation study, this property will likely hold. For external validation studies, there are several issues that may lead to the calibration equation not being transportable. First, the instrument used to measure X^* in the external study may not be identical to that used in the main study. For example, in dietary studies the questionnaires used to capture dietary intake may differ between the external and main studies. Second, the populations of the 2 studies may differ in the way they report their dietary intake. Even when the same instrument for X^* is used in both studies, and the populations are similar, thus supporting the assumption that the measurement error models $(X^*|X, Z)$ are the same in the external and main studies, calibration equations $E(X|X^*, Z)$ of the 2 studies will coincide only when, in addition, the distribution of exposure conditional on covariates $(X|Z)$ is the same in both studies (Section 2.2.5 in Carroll et al. (6)). Thus, in our example, a difference in the distribution of dietary intakes (adjusted for covariates) between the 2 studies could lead to nontransportability. In summary, with external validation studies, extra care in applying regression calibration is needed (24).

The design and sample size of validation studies are discussed in the section Size and Design of a Validation Study.

The calibration equation

The Basics, above, provides basic principles for constructing the calibration equation. Here we discuss implementation details and provide examples.

When the outcome model is specified using a transformed exposure (e.g., logarithmic scale), the calibration equation should be specified to estimate the transformed exposure directly (Section 4 in Carroll et al. (6)). It is generally inappropriate to first estimate the untransformed exposure and then transform the estimate. This applies equally when using a spline for modeling the exposure-outcome relationship (e.g., Harrell et al. (25)). See Web Appendix 2 for details.

The rule of including all outcome-model confounders Z in the calibration equation can be waived only when certain confounders demonstrably do not contribute to the estimation of X. In practice, one may test statistically whether confounders contribute to estimating X, and omit those that are seemingly unimportant. See Heinze et al. (26), for example, for guidance on selecting covariates.

Suppose an additional variable $\tilde{Z}$, not included among the outcome model confounders Z, improves the estimation of X. Its inclusion in the calibration equation, called enhanced regression calibration, will increase the correlation of the resulting $\hat{X}$ with X, thereby increasing the power to detect the association of X with Y (27). Theory allows use of $\tilde{Z}$ to help estimate X if it provides no extra information about the outcome beyond that provided by X and Z. In our example, $\tilde{Z}$ might be an indicator of whether the self-report was made on a weekday or weekend, allowing adjustment of estimated potassium intake accordingly. However, if $\tilde{Z}$ provides extra information about the outcome, then the investigator needs to add $\tilde{Z}$ to the confounders Z already in the outcome model. In practice, one may test whether $\tilde{Z}$ should be selected for the outcome model using the methods of Heinze et al. (26).

Table 1. The Mean Estimate, Empirical Standard Error of the Mean, and Mean Percent (%) Bias for 1,000 Simulated Data Sets

Method[a]	Mean of Log Odds Ratio Estimates	Empirical Standard Error of Mean[b]	% Bias
Uncorrected	0.201	0.002	−50.3
Correct RC	0.407	0.004	0.3
RC, nonaligned outcome model[c]	0.912	0.006	125.0
RC, nonaligned calibration model[d]	0.366	0.004	−9.7

Abbreviation: RC, regression calibration

[a] Results presented for the log odds ratio (β) of X from uncorrected logistic regression, RC correctly performed, RC with a nonaligned outcome model, and RC with a nonaligned calibration model. True value of β.

[b] Empirical standard error of log odds ratio estimate/$\sqrt{}$(number of simulations).

[c] Outcome model that omits the confounder V.

[d] Calibration model that omits the confounder V to produce a nonaligned calibration estimate $\hat{X}$.

Including $\tilde{Z}$ in the calibration is useful only when it provides information about X over and above that provided by X^* and Z (27). In the example above, inclusion of the weekday/weekend report variable should contribute enough to the calibration model for potassium intake to justify its selection (26).

To demonstrate the principle that the calibration and outcome models need to include the same confounders, we conducted simulations of a logistic regression outcome model with multivariate normal covariates X^*, X, Z, and an additional confounder V. Error-prone exposure X^* was correlated with X, Z and V. A validation subset included an unbiased biomarker X^{**} with independent random error. Full details are provided in Web Appendix 3 and Web Table 1. For each simulation we fitted the outcome model with X replaced by: 1) the unadjusted X^*; 2) the calibration estimate $\hat{X}$ calculated from the correct calibration model; 3) as in (2) but with a nonaligned outcome model that omitted V from the covariates; and 4) the calibration estimate $\hat{X}$ calculated from a nonaligned calibration model that omitted V from the covariates,. The resulting estimated exposure coefficients are summarized in Table 1. Only the correctly performed regression calibration method (2) yielded an (approximately) unbiased estimate.

We illustrate the same principle with a real example. The HCHS/SOL is a large community-based cohort study of 16,415 Hispanic/Latinos in the United States, recruited using a complex survey design (12). For details, see Web Appendix 4. We examined the association between potassium intake and hypertension. Log potassium intake averaged over two 24-hour recalls was the error-prone (28) exposure measure (X^*). We used data from an HCHS/SOL internal validation substudy, Study of Latinos: Nutrition and Physical Activity Assessment Study (SOLNAS) (28), on 24-hour urinary excretion of potassium (X^{**}), to develop a calibration equation for log potassium intake. The calibration equation included X^* and all outcome model confounders, Z : age, sex, Hispanic/Latino background, education, income, current smoking, body mass index (BMI), and supplement use (yes/no). Using the calibration estimate $\hat{X}$ in the logistic regression outcome model, we estimated the odds ratio (OR) of hypertension for a 20% increase in potassium intake, controlling for confounders. We fitted the outcome model first including all confounders, Z, and then including all except supplement use, to assess the impact of omitting a calibration equation covariate.

With supplement use included in the outcome model, the estimated OR was 0.76 (95% confidence interval (CI): 0.60, 0.96), compared with 0.90 (95% CI: 0.75, 1.07) when omitted (see Table 2). Thus, incorrectly omitting this calibration-model covariate from the outcome model gave an OR much closer to 1.0 and no longer statistically significant.

Table 2. Data Example, Presenting Estimates of the Odds Ratio of Hypertension Associated With a 20% Increase in Potassium Intake When Supplement Use Is Either Included or Excluded From the Outcome Model, Hispanic Community Health Study/Study of Latinos, United States, 2008–2011

Method of Estimation	OR	95% CI[a]
Including supplement use in outcome model	0.76	0.60, 0.96
Omitting supplement use from outcome model	0.90	0.75, 1.07

Abbreviations: CI, confidence interval; OR, odds ratio.

[a] Based on a multiple imputation procedure described by Baldoni et al. (33), with 25 imputations; see Web Appendix 8.

Table 3. Regression Calibration Simulation Study Showing the Mean Estimate, Mean Percent (%) Bias, Empirical Standard Error, Median Estimated Standard Error, and Coverage Probabilities of the Estimated 95% Confidence Interval for the Log Odds Ratio (β) of X[a]

Method	Mean of Log Odds Ratio Estimates	% Bias	Empirical Standard Error of Log Odds Ratio Estimate	Average Estimated Standard Error	Coverage Probability
Model-based	0.407	0.3	0.136	0.113	0.915
Bootstrap-based				0.140	0.954

[a] Results based on 1,000 simulated data sets. True $\beta = \log(1.5) = 0.4055$. Bootstrap sampling was stratified on membership in the validation substudy. We performed 1,000 bootstrap iterations. Bootstrap confidence intervals are based on the percentile bootstrap.

THE VARIANCE OF THE ESTIMATED EXPOSURE-OUTCOME ASSOCIATION

Correctly estimating uncertainty of the association

Because regression calibration uses a calibration estimate, $\hat{X}$, of exposure, there is more uncertainty in the estimated exposure-outcome association compared with using X. The extra uncertainty derives from 2 sources. The main source is the imperfect correlation between $\hat{X}$ and X arising from the measurement error in X^*. The second source is the finite sample available for estimating the calibration equation, leading to error in the estimated coefficients. Fitting the outcome model with $(Y, \hat{X}, Z)$, the SEs of the estimated outcome model coefficients reported by standard statistical software, such as *glm* in R or *genmod* in SAS (SAS Institute, Inc., Cary, North Carolina), do not incorporate this second source of uncertainty, so are generally too small. This results in 95% CIs being too narrow and having less than 95% coverage probability.

To demonstrate this problem, we used the simulation study described in Forming the Calibration Equation, above. For correctly formulated regression calibration, we compared the SEs reported by *glm* in R (R Foundation for Statistical Computing, Vienna, Austria) with those based on a nonparametric bootstrap procedure accounting for the uncertainty in the calibration equation coefficients. We also compared the coverage of their 95% CIs.

Table 3 shows that the model-based SEs (hereafter called "unadjusted") were too small, as judged by the empirical SE, and resulted in less than 95% coverage probability. The bootstrap-based SE was close to the empirical SE, providing close to 95% coverage.

In the HCHS/SOL example, described above, the unadjusted SE estimate of the regression coefficient for log potassium was 11% smaller than the SE estimate adjusted for calibration equation uncertainty (unadjusted SE = 0.59, corrected SE = 0.66). Consequently, the unadjusted 95% CI for the OR for a 20% increase in potassium intake was (0.61, 0.94), compared with the adjusted 95% CI (0.60, 0.96). Here, the difference between the CIs was small; in other settings (e.g., with smaller validation studies) more appreciable differences are seen.

Two general methods to calculate a valid SE of regression calibration-based estimates are 1) the bootstrap and 2) the "stacked estimation equations" methods. The bootstrap is usually easier to apply, with readily available statistical software, and widely applicable, but it is computationally intensive; it requires fitting the calibration model and then the outcome model on many (generally ≥1,000) bootstrap samples (29). With an internal validation study, bootstrap sampling should be stratified by validation study participation.

Details of the "stacked estimation equations" method (30) are provided in Web Appendix 5. In some settings, the analytical formulas of Rosner et al. (16, 31, 32) may be appropriate, and in studies with complex survey designs other methods are needed, as in our use of the multiple imputation variance estimator of Baldoni et al. (33) for the HCHS/SOL study.

Size and design of a validation study

Uncertainty in the coefficients of the calibration equation increases the uncertainty in the estimated exposure-outcome association. One way of reducing uncertainty in estimating the calibration equation is to increase the validation study sample size. Keogh et al. (23) provide guidelines for choosing the validation study size, including a sample size formula. Alternatively, simulations may be conducted, as in designing the SOLNAS validation study. See Web Appendix 6.

Sampling design issues arise when planning internal validation studies. A simple random sample from the parent cohort is the simplest valid option. SOLNAS participants were recruited from the parent cohort using stratified sampling to ensure an equal distribution of participants from each of the 4 field centers. Consideration was given to adequate representation from each of the 6 Hispanic/Latino ethnicities (Cuban, Dominican, Mexican, Puerto Rican, Central American, and South American) and BMI categories. Increasing the sampling probabilities of individuals from small strata defined by calibration model covariates can improve the precision of calibration equation coefficients, while conditioning on the design variables in the calibration equation ensures valid estimates (34). Alternatively, inverse-probability weighting could be used to obtain valid calibration regression coefficients under more general validation selection designs, where the selection depends on additional factors beyond the calibration covariates; however, weighted

estimators can result in a loss of efficiency relative to simple random sampling (35). Failure to fully account for the selection, either by adjusting for the design variables or inverse-probability weighting can result in bias. In the setting where selection may depend on unobserved factors, perhaps through refusal to participate, it may be challenging to estimate a correct adjustment for selection.

THE DILEMMA OF MEDIATION OF A COVARIATE IN THE CALIBRATION EQUATION

Covariates in the calibration model should usually be included as covariates in the outcome model (The Basics, above). However, a methodological problem can arise.

In outcome models, mediators (variables believed to lie on a causal pathway between exposure and outcome) should not be entered as covariates (36). Doing so makes the exposure regression coefficient represent not the total effect of exposure on outcome (as required), but only the nonmediated part of the effect. With regression calibration, this can create a dilemma. If a mediator variable is also important for estimating exposure, and is included in the calibration model, then it should be included in the outcome model to avoid biased estimation, but, as a mediator, the variable should not be included.

This dilemma was first encountered in nutritional epidemiology when studying the association of cancer with total energy intake (7). Total energy intake is poorly reported using self-report instruments (37), but including BMI in the calibration model greatly improves the R^2. In that case, BMI should be included in the outcome model. This creates 2 problems, one philosophical and one practical.

The philosophical problem is whether to regard BMI as a confounder or a mediator of the energy intake–disease association. Does high BMI cause higher energy intake, or does high energy intake cause higher BMI? If the former, then BMI is a confounder and should be included in the outcome model. If the latter, then BMI is a mediator and should be excluded.

Assuming BMI is a confounder allows its inclusion in the outcome model, but this still leads to a practical problem. BMI, being the most important covariate in the calibration equation, is highly correlated with calibrated energy, so entering both in the outcome model leads to collinearity and difficulty with estimating their separate associations with disease (9).

Assuming BMI is a mediator (which seems more likely) leads to the dilemma described above, and is the issue that we now focus on. To reiterate, when one of the principle covariates in the calibration equation is also a mediator in the outcome model, how should we estimate the association? In reference (38), Douglas Midthune, proposed a solution: BMI (the mediator) and calibrated energy intake are both included in the outcome model, and their coefficients are estimated and then combined linearly, accounting for the mediation and yielding a consistent estimate (see Web Appendix 7). Collinearity is partly overcome in this approach, because the linear combination of the 2 parameters can be estimated more precisely than each association separately, due to the large negative correlation between them.

Table 4. Data Example, Presenting Estimates of the Odds Ratio of at Least One of 4 Metabolic Syndrome Risk Factors[a] Associated With a 20% Increase in Energy Intake, Using 3 Different Methods, Hispanic Community Health Study/Study of Latinos, United States, 2008–2011

Method of Estimation	OR	95% CI[b]
Including BMI in outcome model	0.85	0.47, 1.53
Omitting BMI from outcome model	3.76	3.06, 4.62
Midthune's method	1.52	1.01, 2.27

Abbreviations: CI, confidence interval; OR, odds ratio.
[a] Hypertension, hyperlipidemia, hypercholesterolemia, or hyperglycemia.
[b] Method for 95% CIs is described in Web Appendix 8.

We illustrate this method with another example from HCHS/SOL. We considered the association between energy intake and an outcome related to metabolic syndrome (39), adjusted for confounders, Z: age, sex, Hispanic/Latino background, education, income, and current smoking. Calibrated log energy intake, $\hat{X}$, based on SOLNAS data, was a linear combination of the log average of two 24-hour recall energy intakes, confounders Z, and BMI. Note that we assume that participants' usual diet and weight were stable at the time of data collection.

We examined 3 methods of estimating the OR per 20% increase in energy intake.

1. Including BMI in the outcome model: that is, regressing outcome on $\hat{X}$, BMI, Z (assuming BMI is a confounder).
2. Omitting BMI from the outcome model: that is, regressing outcome on $\hat{X}$, Z (assuming BMI is a mediator—but known to give a biased estimate).
3. Midthune's method (assuming BMI is a mediator—see Web Appendix 7).

Table 4 shows that the 3 methods yielded widely different estimates. Including BMI in the model yielded an OR estimate of less than 1, with the 95% CI including the null value. Omitting BMI from the outcome model gave a large, highly statistically significant OR (3.76) but one that is biased. Midthune's method yielded a statistically significant, but much smaller OR (1.52), suggesting a positive total association between energy intake and the outcome. The 95% CIs for these latter 2 methods did not even overlap. This demonstrates that the estimate obtained from omitting a mediator from the outcome model can be quite misleading when the mediator is itself used in the calibration equation.

DISCUSSION

After explaining the statistical principles involved, we have here considered practical issues arising when implementing regression calibration for addressing exposure error in epidemiologic studies. To conclude, we provide a checklist of 7 main points made in the earlier sections.

1. To avoid bias, the calibration equation should include all confounders included in the outcome model. Exceptions to this rule occur if a particular confounder does not contribute to estimation of the exposure. This can be investigated while building the calibration equation using variable selection methods (e.g., see Heinze et al. (26)). This principle means that, for any given exposure, there is no single calibration equation that is appropriate for all analyses.
2. If at all possible, a validation study should be conducted internally. The simplest valid design for an internal validation study is a simple random sample of participants in the main study. Sometimes more complex sampling designs can improve efficiency (40–42).
3. The validation study should be large enough that the uncertainty in the calibration equation plays only a minor role in the precision of the estimated association. A sample size formula is provided in Web Appendix 6.
4. When constructing the calibration equation, the equation's dependent variable should have the same functional form of the exposure that is used in the outcome model. For example, if log exposure is used in the outcome model, the dependent variable of the calibration model should be log exposure.
5. To avoid bias, the outcome model should include as covariates not only the relevant confounding variables, but also any extra covariates that are in the regression calibration model. An exception is when there is evidence that the extra covariate is independent from the outcome, conditional on the other covariates in the model. This evidence may include showing that the covariate would not be selected for the outcome model (26).
6. When regression calibration is used, SEs must be adjusted to account for the uncertainty in the estimation of the calibration equation. Approaches for SE estimation in the presence of regression calibration include the bootstrap and stacked estimating equations methods, and, for external validation studies, Rosner et al.'s analytical formulas (16, 31, 32).
7. When a calibration model covariate mediates the exposure-outcome relationship, Midthune's method of estimating the association parameter should be used.

We have here reviewed common issues encountered when applying regression calibration and have made recommendations for how to address them. Regression calibration is an intuitive approach for addressing measurement error in covariates, but its implementation requires care. When properly applied, it greatly reduces the bias in estimated association parameters caused by exposure measurement error (43). Its wider use in observational epidemiology is recommended.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, United States (Lillian A. Boe, Pamela A. Shaw, Eunyoung Park); Biostatistics Unit, Kaiser Permanente Washington Health Research Institute, Seattle, Washington, United States (Pamela A. Shaw); Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, Maryland, United States (Douglas Midthune, Victor Kipnis); Department of Statistics, The University of British Columbia, Vancouver, British Columbia, Canada (Paul Gustafson); Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, Chapel Hill, North Carolina, United States (Daniela Sotres-Alvarez); Collaborative Studies Coordinating Center, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, Chapel Hill, North Carolina, United States (Daniela Sotres-Alvarez); Biostatistics and Biomathematics Unit, Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Tel Hashomer, Israel (Laurence Freedman); and Information Management Services, Inc., Rockville, Maryland, United States (Laurence Freedman).

This work was supported in part by the National Institutes of Health (grant R01-AI131771 (P.A.S., L.A.B.)) and by the National Heart, Lung, and Blood Institute (contract 75N92019D00010 (D.S.A.)). The Hispanic Community Health Study/Study of Latinos was carried out as a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute to the University of North Carolina (N01-HC65233), University of Miami (N01-HC65234), Albert Einstein College of Medicine (N01-HC65235), Northwestern University (N01-HC65236), and San Diego State University (N01-HC65237). This work was conducted on behalf of the STRATOS Measurement Error and Misclassification Topic Group (Topic Group 4). Membership of Topic Group 4 can be found at https://www.stratos-initiative.org/en/group_4.

The data used in this paper was obtained through submission and approval of a manuscript proposal to the Hispanic Community Health Study/Study of Latinos Publications Committee, as described on the HCHS/SOL website. For more details, see https://sites.cscc.unc.edu/hchs/publications-pub. R Code used for conducting the analyses described in this paper, including for implementing a bootstrap analysis, together with simulated data similar to those used in the analyses, may be accessed at the website: https://github.com/PamelaShaw/STRATOS-TG4-RC.

The authors acknowledge the investigators, the staff, and the participants of Hispanic Community Health Study/Study of Latinos and Study of Latinos: Nutrition and Physical Activity Assessment Study for their dedication and commitment to the success of this study. Investigator website: http://www.cscc.unc.edu/hchs/

A preprint of this article has been published online. Boe L, Shaw PA, Midthune D, et al. Issues in Implementing Regression Calibration Analyses. *arXiv*. 2022. https://arxiv.org/abs/2209.12304.

Conflict of interest: none declared.

REFERENCES

1. Murray RP, Connett JE, Lauger GG, et al. Error in smoking measures: effects of intervention on relations of cotinine and carbon monoxide to self-reported smoking. The Lung Health Study Research Group. *Am J Public Health.* 1993;83(9):1251–1257.
2. Coggon D, Rose G, DJP B, eds. Measurement error and bias. In: *Epidemiology For the Uninitiated*. 5th ed. London: BMJ Publishing Group; 2003:21–28.
3. Thiébaut AC, Freedman LS, Carroll RJ, et al. Is it necessary to correct for measurement error in nutritional epidemiology? *Ann Intern Med.* 2007;146(1):65–67.
4. Ferrari P, Friedenreich C, Matthews CE. The role of measurement error in estimating levels of physical activity. *Am J Epidemiol.* 2007;166(7):832–840.
5. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Stat Med.* 2014;33(12):2137–2155.
6. Carroll RJ, Ruppert D, Stefanski LA, et al. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. New York, NY: Chapman and Hall/CRC; 2006.
7. Prentice RL, Shaw PA, Bingham SA, et al. Biomarker-calibrated energy and protein consumption and increased cancer risk among postmenopausal women. *Am J Epidemiol.* 2009;169(8):977–989.
8. Freedman LS, Schatzkin A, Midthune D, et al. Dealing with dietary measurement error in nutritional cohort studies. *J Natl Cancer Inst.* 2011;103(14):1086–1092.
9. Prentice RL, Huang Y, Kuller LH, et al. Biomarker-calibrated energy and protein consumption and cardiovascular disease risk among postmenopausal women. *Epidemiology.* 2011;22(2):170–179.
10. Mossavar-Rahmani Y, Shaw PA, Wong WW, et al. Applying recovery biomarkers to calibrate self-report measures of energy and protein in the Hispanic Community Health Study/Study of Latinos. *Am J Epidemiol.* 2015;181(12):996–1007.
11. Shaw PA, Deffner V, Keogh RH, et al. Epidemiologic analyses with error-prone exposures: review of current practice and recommendations. *Ann Epidemiol.* 2018;28(11):821–828.
12. Sorlie PD, Avilés-Santa LM, Wassertheil-Smoller S, et al. Design and implementation of the Hispanic Community Health Study/Study of Latinos. *Ann Epidemiol.* 2010;20(8):629–641.
13. Sauerbrei W, Abrahamowicz M, Altman DG, et al. STRATOS initiative. STRengthening analytical thinking for observational studies: the STRATOS initiative. *Stat Med.* 2014;33(30):5413–5432.
14. The STRATOS Initiative. STRengthening Analytical Thinking for Observational Studies. Topic Groups. https://www.stratos-initiative.org/en/groups. Accessed June 2, 2022.
15. The STRATOS Initiative. Topic Group 4 of the STRATOS Initiative: Resources. http://www.stratostg4.statistik.uni-muenchen.de/Resources.html. Accessed June 22, 2022.
16. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am J Epidemiol.* 1990;132(4):734–745.
17. Thurston SW, Spiegelman D, Ruppert D. Equivalence of regression calibration methods in main study/external validation study designs. *J Stat Plan Inference.* 2003;113(2):527–539.
18. Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika.* 1982;69(2):331–342.
19. Berkson J. Are there two regressions? *J Am Stat Assoc.* 1950;45(250):164–180.
20. Murad H, Freedman LS. Estimating and testing interactions in linear regression models when explanatory variables are subject to classical measurement error. *Stat Med.* 2007;26(23):4293–4310.
21. Wang CY, Hsu L, Feng ZD, et al. Regression calibration in failure time regression. *Biometrics.* 1997;53(1):131–145.
22. Lagakos SW. Effects of mismodeling and mismeasuring explanatory variables on tests of their association with a response variable. *Stat Med.* 1988;7(1–2):257–274.
23. Keogh RH, Shaw PA, Gustafson P, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: part 1—basic theory and simple methods of adjustment. *Stat Med.* 2020;39(16):2197–2231.
24. Shaw PA, Gustafson P, Carroll RJ, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: part 2—more complex methods of adjustment and advanced topics. *Stat Med.* 2020;39(16):2232–2263.
25. Harrell FE Jr, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. *J Natl Cancer Inst.* 1988;80(15):1198–1202.
26. Heinze G, Wallisch C, Dunkler D. Variable selection—a review and recommendations for the practicing statistician. *Biom J.* 2018;60(3):431–449.
27. Kipnis V, Midthune D, Buckman DW, et al. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics.* 2009;65(4):1003–1010.
28. Mossavar-Rahmani Y, Sotres-Alvarez D, Wong WW, et al. Applying recovery biomarkers to calibrate self-report measures of sodium and potassium in the Hispanic Community Health Study/Study of Latinos. *J Hum Hypertens.* 2017;31(7):462–473.
29. Efron B. Better bootstrap confidence intervals. *J Am Stat Assoc.* 1987;82(397):171–185.
30. Stefanski LA, Boos DD. The calculus of M-estimation. *Am Stat.* 2002;56(1):29–38.
31. 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.
32. 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.
33. Baldoni PL, Sotres-Alvarez D, Lumley T, et al. On the use of regression calibration in a complex sampling design with application to the Hispanic Community Health Study/Study of Latinos. *Am J Epidemiol.* 2021;190(7):1366–1376.
34. Cochran WG. *Sampling Techniques*. 3rd ed. New York, NY: John Wiley & Sons; 1977.
35. Korn EL, Graubard BI. *Analysis of Health Surveys*. New York, NY: John Wiley & Sons; 2011.
36. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol.* 1986;51(6):1175–1182.
37. Freedman LS, Commins JM, Moler JE, et al. Pooled results from five validation studies of dietary self-report instruments using recovery biomarkers for energy and protein intake. *Am J Epidemiol.* 2014;180(2):172–188.

38. Freedman LS, Midthune D, Carroll RJ, et al. Using regression calibration equations that combine self-reported intake and biomarker measures to obtain unbiased estimates and more powerful tests of dietary associations. *Am J Epidemiol.* 2011;174(11):1238–1245.
39. Donato KA, Eckel RH, Franklin BA, et al. Diagnosis and management of the metabolic syndrome. *Circulation.* 2005; 112(17):2735–2740.
40. Pepe MS, Reilly M, Fleming TR. Auxiliary outcome data and the mean score method. *J Stat Plan Inference.* 1994;42(1–2): 137–160.
41. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *J R Stat Soc Ser C.* 1999;48(4):457–468.
42. Sarndal CE, Swensson B, Wretman J. *Model Assisted Survey Sampling*. New York, NY: Springer Science & Business Media; 2003.
43. Buonaccorsi JP. *Measurement Error: Models, Methods, and Applications*. New York, NY: Chapman and Hall/CRC; 2010.