Published in final edited form as:

Stat Med. 2019 April 15; 38(8): 1303–1320. doi:10.1002/sim.8052.

Design and analysis considerations for combining data from multiple biomarker studies

Abigail Sloan¹, Yue Song¹, Mitchell H. Gail², Rebecca Betensky¹, Bernard Rosner^{1,3}, Regina G. Ziegler², Stephanie A. Smith-Warner^{4,5}, Molin Wang^{1,3,5}

¹Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts

²Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland ³Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts ⁴Department of Nutrition, Harvard T. H. Chan School of Public Health, Boston, Massachusetts ⁵Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts

Abstract

Pooling data from multiple studies improves estimation of exposure-disease associations through increased sample size. However, biomarker exposure measurements can vary substantially across laboratories and often require calibration to a reference assay prior to pooling. We develop two statistical methods for aggregating biomarker data from multiple studies: the full calibration method and the internalized method. The full calibration method calibrates all biomarker measurements regardless of the availability of reference laboratory measurements while the internalized method calibrates only non-reference laboratory measurements. We compare the performance of these two aggregation methods to two-stage methods. Furthermore, we compare the aggregated and two-stage methods when estimating the calibration curve from controls only or from a random sample of individuals from the study cohort. Our findings include the following: (1) Under random sampling for calibration, exposure effect estimates from the internalized method have a smaller mean squared error than those from the full calibration method. (2) Under the controls-only calibration design, the full calibration method yields effect estimates with the least bias. (3) The two-stage approaches produce average effect estimates that are similar to the full calibration method under a controls only calibration design and the internalized method under a random sample calibration design. We illustrate the methods in an application evaluating the relationship between circulating vitamin D levels and stroke risk in a pooling project of cohort studies.

Correspondence Abigail Sloan, Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA 02115. asloan@g.harvard.edu, Molin Wang, Departments of Biostatistics and Epidemiology, Harvard T. H. Chan School of Public Health, and Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115. stmow@channing.harvard.edu.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

CONFLICT OF INTEREST

None to report.

Keywords

aggregation; between-study variability; calibration; pooling project; two-stage method

1 | INTRODUCTION

For many health outcomes, it has become increasingly common to combine biomarker data from different studies (“pooling”) to evaluate risk profiles. Throughout this paper, we use the term pooling to indicate data combination rather than physical biospecimen combination. The benefits of pooling stem primarily from the increased sample size and include more precise estimation of the biomarker-disease association, the ability to evaluate the dose-response relationship over a wide range of biomarker values, and the ability to perform subgroup and tumor subtype analyses.^{1–3} Many pooling consortia examine biomarker-disease associations, including the Circulating Biomarkers and Breast and Colorectal Cancer,⁴ the Cohort Consortium Vitamin D Pooling Project of Rarer Cancers,⁵ the Endogenous Hormones, Nutritional Biomarkers, and Prostate Cancer Collaborative Group,^{6,7} and the Breast and Prostate Cancer Cohort Consortium.⁸

There are two major approaches for analyzing pooled data. The first, named the two-stage method, involves completing study-specific analyses in the first stage and performing meta-analysis in the second stage.² The second approach, named the aggregated method, involves pooling primary data from multiple studies and performing statistical analyses on the collective data set. Although the aggregated approach typically requires stronger assumptions than the two-stage method, in many settings, the two-stage method faces critical limitations that the aggregated data approach can address. For example, the two-stage approach can be limited in its examination of tumor subtypes, population subgroups, or rare diseases due to data sparsity in the first stage analysis. In one or more individual studies, the data may be so sparse that no cases exist for a given exposure category, thus impeding estimation of effects. Aggregated analyses are less likely to encounter these data sparsity issues because the analyses are performed on the pooled data.

Between-study variability exists in biomarker data. For example, circulating 25-hydroxyvitamin D measurements (25(OH)D) can vary up to 40% between assays and laboratories.^{9–11} Many hormones, such as estrone and testosterone, have highly variable measurements across assays and laboratories.^{1,3} In order to use the two-stage or aggregated data method for these biomarkers, the design and analysis should incorporate a calibration component to account for between-study measurement variability. The objective of calibration is to transform the original data so that all the calibrated measurements are comparable. In practice, calibration is performed by designating a reference laboratory and reassaying a subset of biospecimens from each study at the reference laboratory. In these subsets, we can estimate a study-specific model relating local laboratory measurements to reference laboratory measurements. This model is then used to calibrate other biomarker measurements from the same study. The reassayed biospecimens may be selected at random from the study cohort (“random sample calibration”). Owing to potential concerns about biomarker measurements in cases or the availability of case biospecimens, investigators

typically use only noncases from the cohort or controls from a case-control subsample (“controls-only calibration”).¹²

These statistical issues can also be framed as a covariate measurement error problem. We can envision the reference and local laboratory measurements as the true and surrogate¹³ biomarker values, respectively. Different measurement error models are implemented across different subsets of the data (ie, the different studies contributing to the final analysis) to address variation in the measurement error models. Furthermore, the measurement error models can be estimated from different data subsets, including random sample or controls-only subsamples. We extend an existing measurement error strategy, regression calibration,^{13,14} as the basis of our proposed methods. While we differentiate between the two-stage and aggregated methods for clarity of discussion, it should be noted that both approaches effectively implement regression calibration. Whether we use an aggregated “one-stage” approach or a two-stage approach, we calibrate and combine all data.

In this paper, we develop two aggregation approaches for pooled data that accommodate calibration and compare them to several two-stage approaches. Specifically, we focus on continuous biomarker data pooled from multiple studies for evaluating the log odds ratio (OR) of a binary disease outcome. Section 2 presents the models and statistical methods. Section 3 compares the methods via simulation, and Section 4 illustrates the methods in an example relating circulating vitamin D levels to stroke incidence in the Nurses’ Health Study (NHS1), Nurses’ Health Study II (NHS2), and Health Professionals Follow-up Study (HPFS). Section 5 offers concluding remarks.

2 | METHODS

2.1 | Approximate likelihood under aggregation

Let $s = 1, \dots, S$ index the studies contributing to the aggregation, where the first Q require calibration. For individual i in study s , let Y_{si} be the binary disease outcome, X_{si} be the biomarker exposure measurement from the reference laboratory, W_{si} be the biomarker measurement from a study-specific local laboratory, and Z_{si} be a vector of other covariates. Without further specification, all vectors are column vectors. We assume at least one study used a local laboratory to measure the biomarker and that X_{si} is not available for all subjects in that study. However, we can observe W_{si} for all subjects in studies with measurements from local laboratories. If between-study measurement variability exists, substituting for X_{si} without making any adjustments leads to biased estimates of the biomarker-outcome relationship.

We propose a likelihood-based method for the estimation of exposure effect using aggregated data. The logistic regression model for the biomarker-disease association is

$$\text{logit}(P(Y_{si} = 1 | X_{si}, Z_{si})) = \beta_{0s} + \beta_x X_{si} + \beta_z^T Z_{si},$$

(1)

where β_{0s} is a study-specific intercept. We aim to provide point and interval estimates for β_x , the log OR describing the biomarker-disease association. The corresponding exact likelihood is

$$L = \prod_s \prod_i \frac{\exp\left\{\mathbb{I}(Y_{si} = 1)(\beta_{0s} + \beta_x X_{si} + \beta_z^T Z_{si})\right\}}{\exp\left\{\beta_{0s} + \beta_x X_{si} + \beta_z^T Z_{si}\right\} + 1},$$

where $\mathbb{I}(\cdot)$ is the indicator function. For studies with local laboratory measurements, the exact likelihood contribution cannot be computed because not all subjects have X_{si} available. Using the surrogacy assumption, we develop an approximate likelihood that uses an estimate of X_{si} in place of X_{si} itself. The surrogacy assumption states that the distribution of Y_{si} given (X_{si}, Z_{si}, W_{si}) depends only on (X_{si}, Z_{si}) . In other words, the local laboratory biomarker measurement contributes no additional information about the likelihood of disease beyond that which is provided by the reference laboratory measurement and covariates. Under aggregation, the likelihood contribution, l_{si} , from a subject with only a local laboratory measurement is

$$\begin{aligned} l_{si} &= P(Y_{si} | W_{si}, Z_{si}) \\ &= \int P(Y_{si} | X_{si}, W_{si}, Z_{si}) f(X_{si} | W_{si}, Z_{si}) dX \\ &= \int \frac{\exp\left\{\mathbb{I}(Y_{si} = 1)(\beta_{0s} + \beta_x X_{si} + \beta_z^T Z_{si})\right\}}{1 + \exp\left\{\beta_{0s} + \beta_x X_{si} + \beta_z^T Z_{si}\right\}} f(X_{si} | W_{si}, Z_{si}) dX \\ &= E_{X_{si} | W_{si}, Z_{si}} \left[\frac{\exp\left\{\mathbb{I}(Y_{si} = 1)(\beta_{0s} + \beta_x X_{si} + \beta_z^T Z_{si})\right\}}{1 + \exp\left\{\beta_{0s} + \beta_x X_{si} + \beta_z^T Z_{si}\right\}} \right], \end{aligned}$$

where we have used the surrogacy assumption in the third line. A second order Taylor series approximation with respect to X_{si} about $E[X_{si} | W_{si}, Z_{si}]$ yields the approximate contribution

$$\tilde{l}_{si} = \frac{\exp\left\{\mathbb{I}(Y_{si} = 1)(\beta_0 + \beta_x E[X_{si} | W_{si}, Z_{si}] + \beta_z^T Z_{si})\right\}}{\exp\left\{\beta_0 + \beta_x E[X_{si} | W_{si}, Z_{si}] + \beta_z^T Z_{si}\right\} + 1}.$$

(2)

The approximate likelihood performs best when $\text{Var}(X_{si} | W_{si}, Z_{si})$ is small, ie, there is not much noise in the relationship between X_{si} and (W_{si}, Z_{si}) , and/or the association between Y_{si} and X_{si} is not too strong. Further details on these conditions and the derivation are available in Appendix A. The presence of $E[X_{si} | W_{si}, Z_{si}]$ in the approximate likelihood motivates the development of the calibration model.

2.2 | Calibration model

To evaluate $\tilde{l}_{si}$ in the aggregated approach, we develop study-specific calibration models $E[X_{si}|W_{si}, Z_{si}]$ for those studies with local laboratory measurements. In each of those studies, the calibration models are estimated from the subset of individuals who have both X_{si} and W_{si} available. The resulting model is then used to estimate the reference laboratory measurement for individuals in that study. We assume that $E[X_{si}|W_{si}, Z_{si}] \approx E[X_{si}|W_{si}]$ to simplify the model in what we call the “calibration approximation.” Furthermore, we assume the reference and local laboratory measurements are linearly related such that

$$E[X_{si}|W_{si}] = a_s + b_s W_{si},$$

(3)

where a_s and b_s are the calibration parameters for $s = 1, \dots, Q$. The calibration approximation states that the mean value of X_{si} given W_{si} does not depend on other covariates. However, this approximation could be relaxed by instead fitting $E[X_{si}|W_{si}, Z_{si}] = a_s + b_s W_{si} + c_s^T Z_{si}$. denote (3) with estimated values $\hat{a}_s$ and $\hat{b}_s$. We will return to a discussion of the calibration approximation and the sensitivity of the method to failures of it in the simulation studies in Section 3.1.

Let $\tilde{X}_{si}$ be the biomarker measurement used in the aggregated analysis. Its value depends on the calibration method being used and the availability of the reference laboratory measurement. For studies that use the reference laboratory for all biomarker measurements, $\tilde{X}_{si} = X_{si}$ for all study participants and no calibration occurs. For studies that use a local laboratory, the *full calibration* and *internalized* methods are defined as follows:

$$\begin{aligned} \text{Full calibration method : } \tilde{X}_{si} &= \hat{E}[X_{si}|W_{si}] \\ \text{Internalized method : } \tilde{X}_{si} &= \begin{cases} X_{si} & \text{if reference laboratory measurement available,} \\ \hat{E}[X_{si}|W_{si}] & \text{otherwise.} \end{cases} \end{aligned}$$

Thus, for studies with local laboratory measurements, the biomarker measurement for each participant is calibrated under the full calibration method. Under the internalized method, the measurement is only calibrated for biospecimens without a reference laboratory measurement outside the calibration subset. We applied one of these methods to all studies that used local laboratories in the pooled analysis.

In the controls-only calibration study (COCS), the calibration subset consists of a random sample from controls. In the random sample calibration study (RSCS), the calibration subset is a random sample from the entire population and includes both cases and controls.

2.3 | Parameter estimation

Recall that S total studies contribute to the pooled analysis, Q of which require calibration. Define $\beta_0 = [\beta_{01}, \dots, \beta_{0S}]$, $\beta = [\beta_0, \beta_x, \beta_z]$, $a = [a_1, \dots, a_Q]$, and $b = [b_1, \dots, b_Q]$. Thus, the collective set of parameters to be estimated is $\theta = [a, b, \beta]$. We define estimating functions $\Psi(\theta) = [\psi_a, \psi_b, \psi_{\beta_0}, \psi_{\beta_x}, \psi_{\beta_z}]$, where $\psi_a, \psi_b, \psi_{\beta_0}, \psi_{\beta_x}$, and ψ_{β_z} are approximately unbiased estimating functions for their respective parameters (see Appendix B for their forms).

Estimates of β are obtained through a two step process in the spirit of pseudo-maximum likelihood estimation.¹⁵ First, we obtain estimates of the calibration parameters $\hat{a}$ and $\hat{b}$ by fitting the linear regression calibration models, through solving $[\psi_a, \psi_b] = \mathbf{0}$. Second, to obtain $\hat{\beta}$, we solve the set of estimating equations $[(\psi_{\beta_0}(\hat{a}, \hat{b}), \psi_{\beta_x}(\hat{a}, \hat{b}), \psi_{\beta_z}(\hat{a}, \hat{b}))] = \mathbf{0}$, where we have replaced instances of a and b with their estimated version in $\psi_{\beta_0}, \psi_{\beta_x}$, and ψ_{β_z} . Note that we would obtain the same numerical estimates had we obtained $\hat{\theta}$ by solving the entire set of estimating equations, $\left[(\psi_{\beta_0}(\hat{a}, \hat{b}), \psi_{\beta_x}(\hat{a}, \hat{b}), \psi_{\beta_z}(\hat{a}, \hat{b})) \right] = \mathbf{0}$.

To obtain valid estimates of $\widehat{\text{Var}}(\hat{\beta})$, we must account for uncertainty in the estimation of $\hat{a}$ and $\hat{b}$. As such, we obtain $\widehat{\text{Var}}(\hat{\theta})$ by using the sandwich variance formula over the entire set of estimating equations $\Psi(\theta) = \mathbf{0}$. Variance estimates for $\hat{\beta}$ are found by selecting the appropriate diagonal term of the sandwich variance matrix (see Appendix C).

2.4 | Two-stage methods

The two-stage approach uses the regression calibration method,^{13,14,16,17} which was developed for measurement error problems, to adjust for calibration in the first stage study-specific analyses. It then combines study-specific estimates using fixed effects meta-analysis in the second stage. The regression calibration method uses the surrogacy assumption and one of the following two assumptions: (i) the noise in the calibration model is moderate and/or the association between Y_{si} and X_{si} is not too strong or (ii) the disease is rare and X_{si} given (W_{si}, Z_{si}) has a multivariate normal distribution.^{14,18} We briefly describe the formulation of regression calibration to facilitate comparison to the aggregated methods in later sections.

In the studies using the reference laboratory, we directly compute an effect estimate and its variance from a logistic regression model. For each of the Q studies using a local laboratory, define $\hat{\beta}_{x,q}$ as the adjusted effect estimate from study q obtained from the regression calibration method. Under regression calibration,

$$\hat{\beta}_{x,q} = \hat{\beta}_{w,q} / \hat{b}_q,$$

$$\widehat{\text{Var}}(\hat{\beta}_{x,q}) = \hat{b}_q^{-2} \widehat{\text{Var}}(\hat{\beta}_{w,q}) + \hat{b}_q^{-4} \hat{\beta}_{w,q}^2 \widehat{\text{Var}}(\hat{b}_q),$$

(4)

where $\hat{b}_q$ is obtained from fitting model (3) in the calibration study and $\hat{\beta}_{w,q}$ and $\widehat{\text{Var}}(\hat{\beta}_{w,q})$ are the naive point and variance estimates of the biomarker-disease association from a logistic regression model using the local laboratory measurements of the individuals not included in the calibration study in study q .¹⁴ We combine the study-specific estimates using fixed-effects meta-analysis, which uses inverse variance weights to obtain a final effect and variance estimate for $\hat{\beta}_x$. We name this formulation the two-stage approach and implement it under both the COCS and RSCS designs.

Under the RSCS design, we also consider a two-stage weighted approach, where in the study-specific analysis stage, the above estimate is combined with the effect estimate obtained from running a logistic regression of Y_{si} on X_{si} and Z_{si} in the calibration subset. We use inverse variance weights to combine these estimates per Spiegelman et al.¹⁷ We can apply this method when a sufficient number of cases exist in the calibration subset. An analogous two-stage weighted approach does not exist for the COCS design because the logistic regression model cannot be defined in the controls-only subset.

It is worth noting that the aggregated approaches are effectively one-stage implementations of regression calibration, but with different calibration models for different observations depending on their study of origin. For the two-stage method, we could also rewrite the approximate likelihood in a form that omits the product over s , maximize the likelihood one study at a time, and still obtain the same study-specific point estimates as in (4).

2.5 | Calibration study designs

We evaluated the bias in the intercept and slope estimates in Equation (3) for a single study when using the RSCS and COCS designs. Under a RSCS design, the linear regression estimates obtained from the calibration study are unbiased for the parameters in Equation (3). That is, when we fit the model $\hat{E}[X_{si}|W_{si}] = \hat{a}_s + \hat{b}_s W_{si}$, the estimated parameters are unbiased for a_s and b_s . Therefore, the calibrated value of the reference laboratory measurement is a valid approximation of the true reference laboratory measurement under the RSCS design.

Under the COCS design, we fit the model $E[X_{si}|W_{si}, Y_{si} = 0] = \hat{a}_{s,co} + \hat{b}_{s,co} X_{si}$ using the subscript co to indicate usage of only controls in the calibration subset. Generally, $\hat{a}_{s,co}$ and $\hat{b}_{s,co}$ are not consistent for a_s and b_s . However, under some conditions, we can show $\hat{a}_{s,co} \approx a_s$ and $\hat{b}_{s,co} \approx b_s$ (see Appendix C for details). For example, if $\beta_x \approx 0$, the reference

laboratory measurement is nearly independent of disease outcome and the calibration model among controls approximates that of random sample calibration. Furthermore,

$$\begin{aligned}\hat{b}_{s,co} &\approx b_s \text{ when } \text{Var}(X_{si}|Y_{si}=0) \approx \text{Var}(X_{si}), \\ \hat{a}_{s,co} &\approx a_s \text{ when } \begin{cases} \text{Var}(X_{si}|Y_{si}=0) \approx \text{Var}(X_{si}) \\ E[X_{si}|Y_{si}=0] \approx E[X_{si}]. \end{cases}\end{aligned}$$

As a result, the slope term under a COCS design is typically less biased than the intercept term. Additionally, $\hat{b}_{s,co} \approx b_s$ and $\hat{a}_{s,co} \approx a_s$ when $\text{Var}(W_{si}|X_{si}) \approx 0$ (see Appendix D).

We plotted the mean percent bias in calibration parameter estimates against the OR under each calibration study design over 1000 simulation replicates under 5% and 50% disease prevalence (Figure 1). As expected, the parameters under the RSCS design were nearly unbiased for the parameters in Equation (3) regardless of the disease prevalence. However, when a COCS design was used, the estimated parameter estimates were increasingly different from the parameters in Equation (3) with increasing effect size. Furthermore, the intercept estimate was more disparate than the slope estimate. The departure was not as severe for simulations with 5% disease prevalence; this is expected given that the RSCS subset resembles the COCS subset with rare disease.

Although the COCS design yields biased calibration parameters that manifest as bias in $\hat{\beta}_x$, we will see in simulation studies in Section 3 that bias in $\hat{\beta}_x$ for COCS designs is still much smaller than bias of estimates from a naive approach. For instance, we will observe that using all local laboratory measurements directly without performing calibration (the naive method) will result in estimates with approximately 30% downward bias (Table 1).

The RSCS design is the preferred calibration study design since it yields unbiased $\hat{a}_s$ and $\hat{b}_s$ estimates. However, biospecimens for cases are typically less accessible than those of controls, and a COCS design is commonly used in practice. We propose using the full calibration method under the COCS design because it avoids differential calibration of subjects. In particular, the bias in $\hat{a}_{s,co}$ does not impact the estimation of $\hat{\beta}_x$ because $\hat{a}_{s,co}$ becomes encapsulated in the biomarker estimate for all subjects (Appendix E). Under the internalized approach for COCS designs, some controls contribute $\tilde{X}_{si}$ to the likelihood while others contribute X_{si} . Because only subjects outside the calibration subset incorporate the biased $\hat{a}_{s,co}$ term, additional bias manifests in $\hat{\beta}_x$ after full optimization of the likelihood. We will see in the simulation study that the bias of $\hat{\beta}_x$ increases as more controls are included in the calibration subset due to increasingly differential calibration among controls.

3 | SIMULATION STUDY

Recall that X_{si} , W_{si} , and Y_{si} are the reference laboratory measurement, local laboratory measurement, and disease indicator for individual i in study s , respectively. Let e_{si} be the

error term in the linear regression of X_{si} on W_{si} . We generate (X_{si}, W_{si}, e_{si}) assuming joint multivariate normality such that

$$\begin{pmatrix} X_{si} \\ W_{si} \\ e_{si} \end{pmatrix} \sim \text{MVN} \left(\begin{pmatrix} \mu_x \\ (\mu_x - a_s)/b_s \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & b_s \sigma_{ws}^2 & \sigma_x^2 - b_s^2 \sigma_{ws}^2 \\ \cdot & \sigma_{ws}^2 & 0 \\ \cdot & \cdot & \sigma_x^2 - b_s^2 \sigma_{ws}^2 \end{pmatrix} \right).$$

Under this formulation, $E[X_{si}|W_{si}] = a_s + b_s W_{si}$. This data structure also induces $\text{Cov}(W_{si}, e_{si}) = 0$; that is, the local laboratory measurement is independent of the error in the calibration model. This is a standard convention in the measurement error literature. We considered a simple risk model without covariates, $\text{logit}(P(Y_{si} = 1|X_{si})) = \beta_{0s} + \beta_x X_{si}$.

We set $\mu_x = 0$ and $\sigma_x^2 = 1$. We assumed four studies contributed to the final analysis, each with 1000 total subjects and 100 calibration participants, and assigned calibration parameters $a = [-3, 1, -1, 3]$ and $b = [0.5, 0.75, 1.25, 1.5]$. We considered disease prevalences of 5%, 25%, and 50%. At each OR considered, including (1.25, 1.5, 1.75, 2, 2.25, 2.5), we completed 1000 simulation replicates.

We compared the methods with regard to three operating characteristics: mean percent bias over the simulation replicates, mean squared error (MSE), and coverage rate, defined as the proportion of simulation replicates whose 95% confidence intervals covered the true β_x . We assumed that increasing levels of the biomarker correspond to an increased level of disease risk. However, in other simulation studies (not shown here), similar results were obtained if the biomarker was instead assumed to have a protective effect. For purposes of comparison, we included a naive method that uses only local laboratory measurements. We present tables of results in the main paper and include graphical versions of these tables in Online Supplement Section S2.

We first considered the COCS design (Table 1). The naive method performed poorly for all three operating characteristics regardless of disease prevalence and effect size, indicating the necessity of calibration. At every OR considered, the percent bias of the naive effect estimate exceeded 29%. Among the aggregated methods, percent bias of the effect estimate was generally minimized by the full calibration approach. For all considered diseases prevalences and ORs, the internalized estimate was biased downwards by roughly 3% to 4% while the full calibration estimate had less than 1% bias. The coverage rates for the full calibration method were closer to the nominal 95% than the internalized method, particularly with large effect sizes. With 50% disease prevalence, the internalized method had coverage rates less than 90% for ORs greater than or equal to 2, due to strong violation of the rare disease and small effect assumptions. For all disease prevalences, the coverage rate for the internalized method was between 0.91 and 0.95 for ORs less than 2. The relative biases for the two-stage method under the COCS design were slightly larger than the full calibration method but less than the internalized method for all prevalences, and the two-stage method often yielded estimates with slightly larger standard errors. However, the two-

stage coverage rates consistently approximated 95% regardless of effect size or disease prevalence.

Under the RSCS design, the OR estimates from the internalized method were typically slightly less biased than those from the full calibration method, particularly as the disease prevalence increased (Table 2). Additionally, the standard errors and MSE estimates under the internalized method were slightly smaller than those under the full calibration method. The two-stage methods typically had comparable biases but larger standard errors and MSE than the aggregated methods. The two-stage weighted approach typically yielded less biased effect estimates with smaller standard errors and MSE than the unweighted two-stage method and offered comparable MSEs to the full calibration method. As disease prevalence or effect size decreased, the coverage rates for the aggregated methods improved.

The methods were insensitive to changes in the number of contributing studies, the magnitude of the calibration parameters, and distribution parameters for X_{si} and W_{si} . Method performance was most significantly impacted when reducing the variance of the error term or increasing the percentage of subjects involved in the calibration study. Smaller values of $\text{Var}(e_{si})$ yielded less biased point estimates due to reduced noise in the X_{si} and W_{si} relationship. Under the COCS design with 25% disease prevalence using the internalized method, increasing the percentage of subjects in the calibration subset further attenuated the effect estimate and increased MSE (Figures 2A-F). This is not observed with the full calibration method, where increasing the percentage of subjects slightly improved the MSE. Under the RSCS designs, increasing the percentage of subjects in the calibration study only slightly reduced the bias and MSE of the aggregated and weighted and unweighted two-stage methods (Figures 2G-L).

3.1 | Reviewing the calibration approximation

Recall that we adopted the optional calibration approximation, specifically $E[X_{si}|W_{si}, Z_{si}] \approx E[X_{si}|W_{si}]$, to simplify the calibration model. If the calibration approximation $E[X_{si}|W_{si}, Z_{si}] \approx E[X_{si}|W_{si}]$ does not hold, investigators may instead fit the full model $E[X_{si}|W_{si}, Z_{si}] = a_s + b_s W_{si} + c_s^T Z_{si}$. We investigated the consequences of violating the calibration approximation and compared bias in $\hat{\beta}_x$ estimates from the simple model using the calibration approximation against estimates from the full model in a simulation study. We assumed the underlying data structure involved a covariate Z_{si} that was correlated with the biomarker and predictive of the outcome such that

$$\begin{pmatrix} W_{si} \\ Z_{si} \\ e_{si} \end{pmatrix} \sim \text{MVN} \left(\begin{pmatrix} \mu_{ws} \\ \mu_z \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{ws}^2 & \sigma_{ws} & 0 \\ \cdot & \sigma_z^2 & 0 \\ \cdot & \cdot & \sigma_e^2 \end{pmatrix} \right),$$

$$X_{si} = a_s + b_s W_{si} + c_s^T Z_{si} = e_{si}$$

where $\mu_{ws} = (\mu_x - a_s - c_s \mu_z)/b_s$, $\sigma_{ws} = \text{Cov}(W_{si}, Z_{si})$, and $\sigma_{ws}^2 = (\sigma_x^2 - c_s^2 \sigma_z^2 - 2b_s c_s \sigma_{ws} - \sigma_e^2)/b_s^2$.

The risk model was $\text{logit}(P(Y_{si} = 1 | X_{si}, Z_{si})) = \beta_{0s} + \beta_x X_{si} + \beta_z Z_{si}$, where we assumed 25% underlying disease prevalence and $\beta_z = \log(1.25)$. We set

$$\mu_x = \mu_z = 0, \sigma_x^2 = \sigma_z^2 = 1, a = (-3, 1, -1, 3), b = (0.5, 0.75, 1.25, 1.5), \text{ and } c = (0.15, 0.15, 0.15, 0.15).$$

As before, we assumed four studies contributed to the analysis, each with 1000 total subjects and 100 individuals in the calibration subset. We chose $\text{Corr}(X_{si}, Z_{si}) = 0.2$, which in turn induced σ_{ws}^2 and σ_{ws} for each study. Additional simulations are available in the Online Supplement Section S3.

We compared two calibration models for this data generating mechanism. The first, Model Z, includes the covariate and thus bypasses the calibration approximation. The second, Model CA, leverages the calibration approximation in order to fit a simpler model.

$$\begin{aligned} \text{Model Z: } \hat{E}[X_{si} | W_{si}, Z_{si}] &= \hat{a}_s + \hat{b}_s W_{si} + \hat{c}_s Z_{si} \\ \text{Model CA: } \hat{E}[X_{si} | W_{si}] &= \tilde{a}_s + \tilde{b}_s W_{si} \end{aligned}$$

In both scenarios, we included the covariate Z_{si} in the final logistic regression model. We compared point and variance estimates of $\hat{\beta}_x$ across 1000 simulation replicates for both models for each method and calibration study subtype (Tables 3 and 4).

Although Model CA yielded more bias in $\hat{\beta}_x$ than Model Z as expected, the difference between the two models was less than 5% over the range of β_x considered for both the COCS and RSCS designs. Under the RSCS design, the bias in the $\hat{\beta}_x$ for Model CA was smallest near null values for all methods. Further simulations (included in Online Supplement Section S3) demonstrated that increasing $\text{Corr}(X_{si}, Z_{si})$ (ie, increasing the degree to which we violate the calibration approximation) yielded larger biases in the Model CA β_x estimates. However, even at a relatively large correlation of 0.3, the difference in percent bias between Model CA and Model Z was less than 8% for both the COCS and RSCS designs. The model performances were largely invariant to the specific value of c_s and β_z . Thus, using the calibration approximation only mildly impacts the resulting effect estimates.

4 | APPLIED EXAMPLE

We applied the two-stage and aggregated methods to data from three cohort studies to investigate the link between circulating vitamin D (25(OH)D) levels and risk of stroke. Previous work suggests that individuals with 25(OH)D measurements in the top tertile of the population have reduced risk of stroke relative to those in the bottom tertile.¹⁹ We pooled data from HPFS,²⁰ NHS1,²¹ and NHS2,²² three large prospective cohort studies in the United States. The HPFS enrolled 51,529 male health professionals in 1986 aged 40 to 75 years at baseline. The NHS1 cohort began enrollment in 1976 and includes 121,701 female

nurses aged 30 to 55 years at baseline. In 1989, the NHS2 was established with the enrollment of 116,671 female nurses, aged 25 to 42 years at baseline. In each of the three cohorts, participants have completed questionnaires reporting information about diet, medical history, and lifestyle conditions every two years since baseline. Further information on each cohort has been discussed previously elsewhere.^{19–21}

Between 1989 and 1997, each study obtained blood samples from a subset of participants and completed assays for a host of biomarkers, including 25(OH)D, prior to diagnosis (Table 5). Subjects who had previously reported a cancer diagnosis were not eligible for random selection. Individuals were excluded from the pooled analysis if they did not have 25(OH)D measurements available or stroke outcome data. A total of 3,768 subjects were included in the pooled analysis. We regard these 3,768 subjects as arising from three cohorts that were sampled from the much larger HPFS, NHS1, and NHS2 cohorts. Participants were considered stroke cases if they had a nonfatal or fatal suspected or confirmed stroke event during study follow-up after blood draw. The cumulative incidence of stroke in the HPFS, NHS1, NHS2, and final pooled prospective cohort subset were 5.8%, 9.6%, 1.5%, and 4.8%, respectively. A RSCS design was implemented by randomly selecting 29 individuals from each cohort and reassaying their blood samples at Heartland Assays, LLC (Ames, IA) during 2011–2013. We also simulated a COCS design by excluding cases from the original calibration subset. Each calibration subset indicated a linear relationship between the local laboratory assay and reference Heartland assay with non-null intercept and slope estimates (Table 6).

We applied the internalized method, full calibration method, and the two-stage approach to the final pooled data set under the RSCS and COCS designs, adjusting for age at blood draw, years of follow-up after blood draw, smoking (never/ever), family history of myocardial infarction (yes/no), hypertension (yes/no), diabetes (yes/no), and BMI (greater than or less than or equal to 25 kg/m²) (Table 5). Circulating 25(OH)D, measured in nmol/L, was standardized to 20 unit increments. All analytic approaches demonstrate that increasing 25(OH)D levels correlate with a small, nonsignificant protective effect against stroke (Table 7). The estimated OR and its 95% confidence interval under RSCS for all three methods is 0.95 (0.78, 1.16). Under the COCS design, the estimated OR for the internalized and full calibration method is approximately 0.96 while the estimated OR under the two-stage approach is 0.95. In this example, the exposure effect is close to the null and the outcome is rare. These conditions lead to less bias in both the intercept and slope estimates of the calibration models in the COCS design as discussed in Section 2.4, and therefore, it is expected that the internalized and full calibration methods in the COCS design result in similar estimates and also similar to those from the RSCS design. For comparison, we also conducted a naive analysis where all local laboratory biomarker values were used directly in a simple logistic regression model without calibration. The resulting naive OR and its 95% confidence interval was 0.98 (0.84, 1.14). Compared to the calibration methods, the naive point estimate is closer to the null and has a smaller confidence interval and smaller standard error.

5 | DISCUSSION

In this paper, we compared aggregated and two-stage methods using COCS and RSCS designs for pooled biomarker data that were measured using different assays and laboratories over different time periods. Several practical recommendations and conclusions follow from our work. First, under the RSCS design, the internalized method provides slightly less biased and more efficient estimates than the full calibration method because it maximally utilizes available information in the calibration subset. Second, the COCS design does not yield slope and intercept estimates that are consistent for the values in parameters in Equation (3) outside of the null effect. However, the bias in calibration parameter estimates, especially that of the slope estimate, is minimal when the effect is small or the disease prevalence is small. Third, under the COCS design, the full calibration method has less bias in point estimates but larger standard errors than the internalized method. The internalized method performs increasingly poorly as the percentage of subjects in the calibration study increases. Fourth, the two-stage approach provides similar effect estimates to the full calibration method under the COCS design or the internalized method under the RSCS design but larger MSE. If implementing an RSCS design with sufficient cases in the calibration subset, the two-stage weighted approach is preferred to the two-stage method. We elaborate further on these points.

Under the COCS design, the aggregated methods exhibit a bias-variance trade-off, where the estimates from the internalized method are more biased but less variable than estimates from the full calibration method. The two-stage approach is not as efficient as the aggregated methods because the calibration subset under regression calibration does not contribute to the effect estimates. However, under an RSCS design using the two-stage weighted approach, the calibration subset does contribute to the effect estimates and the approach is as efficient as the aggregated methods.

This work relies on calibration data. Investigators must have the means to reassay biospecimens at a reference laboratory in order to obtain calibration parameters. In practice, the choice of reference laboratory will be dictated by financial and logistical considerations in the design stage. Resources permitting, we recommend choosing a reference laboratory that maximizes the number of subjects in the calibration subsets. Our methods do not require a specific proportion of subjects in the calibration subsets, but the number of subjects in the subset needs to be sufficiently large to obtain stable calibration model estimates.

Although we compared the two-stage and aggregated methods throughout this paper, it is important to note that an aggregated method may be better suited to particular analyses. For example, if one wishes to perform subgroup analyses, an aggregated method reduces problems from data sparsity. As previously shown in scenarios where no calibration process is necessary or performed for the exposure measurement, if the main exposure effect is homogeneous across studies and some of the other covariate effects are also homogeneous across studies, the aggregated method may offer additional improvements in efficiency for the main exposure effect compared to the two-stage method, which allows for heterogeneity in the other covariate effects.²³

This paper focuses on the scenario where the relationship between the reference and local laboratory measurements is described by a linear regression model. We will consider a nonlinear calibration model in the future. The methods extend readily to multiple pooled biomarkers requiring calibration. Although we used the calibration approximation to simplify the models, it can be relaxed by including additional covariates in the calibration model. Nonetheless, using the simple calibration model when the full model is more appropriate only mildly impacts the final effect estimate.

All methods discussed can be applied to retrospective case-control studies or prospective studies with a binary outcome. When applying the methods to a case-control sample, disease prevalence in our recommendation refers to the case percentage in the case-control sample. The code to implement these analyses is available upon request.

To summarize, the full calibration method provides the best estimates for the COCS design because it avoids the differential bias introduced by the internalized method. In contrast, under the RSCS design, the internalized method is the preferred aggregated approach. Additionally, the calibration parameters from a COCS study are minimally biased when the effect is small or disease is rare. These observations can assist investigators in selecting the best calibration study structures and appropriate analyses to meet their goals.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank the reviewers for their constructive feedback and Kathryn Rexrode, Chao Cheng, and Xiao Wu for helpful comments and suggestions. We are grateful to Tao Hou and Shiao-Shyuan (Sherry) Yaun for their assistance in accessing the data. We also thank the Circulating Biomarkers and Breast and Colorectal Cancer Consortium team (R01CA152071, PI: Stephanie Smith-Warner; Intramural Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute: Regina Ziegler) for conducting the calibration study in the vitamin D example. Abigail Sloan was supported by the National Institutes of Health under grant T32-NS048005 and Molin Wang was supported in part by NIH/NCI grant R03CA212799. This work was supported in part by the Intramural Program of the National Cancer Institute, Division of Cancer Epidemiology and Genetics.

Funding information

NIH/NCI, Grant/Award Numbers: R01CA152071 and R03CA212799; Intramural Research Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute; NIH, Grant/Award Number: T32-NS048005

APPROXIMATE LIKELIHOOD DERIVATION

The exact likelihood corresponding to the logistic regression model in Equation (1) is

$$L = \prod_s \prod_i \frac{\exp\left\{\mathbb{I}(Y_{si} = 1) \left(\beta_{0s} + \beta_x X_{si} + \beta_z^T Z_{si} \right)\right\}}{\exp\left\{\beta_{0s} + \beta_x X_{si} + \beta_z^T Z_{si}\right\} + 1}.$$

For studies using a local laboratory, the exact likelihood contribution cannot be computed because X_{si} is not available for all subjects. Under the surrogacy assumption

$$P(Y_{si} | X_{si}, W_{si}, Z_{si}) = P(Y_{si} | X_{si}, Z_{si}) \text{ and calibration approximation } E[X_{si} | W_{si}, Z_{si}] \approx E[X_{si} | W_{si}],$$

the likelihood contribution from an individual with a local laboratory measurement, suppressing subscripts for brevity, henceforth can be approximated as

$$\begin{aligned}
 l &= P(Y|W, Z) \\
 &= \int P(Y|X, W, Z) f(X|W, Z) dX \\
 &= \int \frac{\exp\{\mathbb{I}(Y=1)(\beta_{0s} + \beta_x X + \beta_z^T Z)\}}{1 + \exp\{\beta_{0s} + \beta_x X + \beta_z^T Z\}} f(X|W, Z) dX \\
 &= E_{X|W} \left[\frac{\exp\{\mathbb{I}(Y=1)(\beta_{0s} + \beta_x X + \beta_z^T Z)\}}{1 + \exp\{\beta_{0s} + \beta_x X + \beta_z^T Z\}} \right],
 \end{aligned}$$

where the surrogacy assumption and calibration approximation are used in the third and fourth lines, respectively. An approximate likelihood contribution is obtained by replacing the expectation operand with a second order Taylor series expansion with respect to X about $E[X|W]$ such that

$$\begin{aligned}
 &\frac{\exp\{\mathbb{I}(Y=1)(\beta_{0s} + \beta_x X + \beta_z^T Z)\}}{1 + \exp\{\beta_{0s} + \beta_x X + \beta_z^T Z\}} \approx \frac{\exp\{\mathbb{I}(Y=1)(\beta_{0s} + \beta_x E[X|W] + \beta_z^T Z)\}}{1 + \exp\{\beta_{0s} + \beta_x E[X|W] + \beta_z^T Z\}} \Bigg|_{X=E[X|W]} \\
 &+ \frac{\partial}{\partial X} \left(\frac{\exp\{\mathbb{I}(Y=1)(\beta_{0s} + \beta_x X + \beta_z^T Z)\}}{1 + \exp\{\beta_{0s} + \beta_x X + \beta_z^T Z\}} \right) \Bigg|_{X=E[X|W]} \Delta \\
 &+ \frac{\partial^2}{\partial^2 X} \left(\frac{\exp\{\mathbb{I}(Y=1)(\beta_{0s} + \beta_x X + \beta_z^T Z)\}}{1 + \exp\{\beta_{0s} + \beta_x X + \beta_z^T Z\}} \right) \Bigg|_{X=E[X|W]} \frac{\Delta^2}{2} + \text{remainder},
 \end{aligned}$$

where $\Delta = X - E[X|W]$. Replacing this expansion in the expectation causes the first order term to disappear, and we rewrite the second order term as a function of the $\text{Var}(X|W)$ and β_x^2 such that

$$\begin{aligned}
 l &\approx \frac{\exp\{\mathbb{I}(Y=1)(\beta_{0s} + \beta_x E[X|W] + \beta_z^T Z)\}}{1 + \exp\{\beta_{0s} + \beta_x E[X|W] + \beta_z^T Z\}} \Bigg|_{X=E[X|W]} \\
 &+ \frac{\beta_x^2 (-1) \mathbb{I}(Y=1) (\exp\{2(\beta_{0s} + \beta_x E[X|W] + \beta_z^T Z)\} - \exp\{\beta_{0s} + \beta_x E[X|W] + \beta_z^T Z\})}{(1 + \exp\{\beta_{0s} + \beta_x E[X|W] + \beta_z^T Z\})^3} \Bigg|_{X=E[X|W]} \frac{\text{Var}(X|W)}{2}.
 \end{aligned}$$

If $|\beta_x|$ or $\text{Var}(X|W)$ is small, the second-order term also approaches zero. Thus, for subject i in study s , the likelihood contribution can be approximated by

$$\tilde{l}_{si} = \frac{\exp\{\mathbb{I}(Y_{si}=1)(\beta_{0s} + \beta_x E[X_{si}|W_{si}] + \beta_z^T Z_{si})\}}{\exp\{\beta_{0s} + \beta_x E[X_{si}|W_{si}] + \beta_z^T Z_{si}\} + 1}.$$

APPENDIX B

ESTIMATING EQUATIONS

Define $\text{expit}(x) = \exp(x)/(1 + \exp(x))$ and let $\beta^T \tilde{X}_{si} = \beta_{0s} + \beta_x \tilde{X}_{si} + \beta_z^T Z_{si}$. Let $s = 1, \dots, S$ index the studies contributing to the aggregation, where the first Q require calibration. Define M_s as the set of subjects in study s and define C_{si} as an indicator variable that takes value 1 if subject i is in study s and is part of that study's calibration subset and 0 otherwise. Suppose study s contains n_s subjects such that $n = \sum_{s=1}^S n_s$ total individuals contribute to the pooled analysis with p additional covariates in the model. Let Z_{sip} denote the value covariate p for individual i in study s . The final set of estimating equations are defined as

$\left[\psi_a, \psi_b, \psi_{\beta_0}, \psi_{\beta_x}, \psi_{\beta_z} \right] = \mathbf{0}$, where ψ_a and ψ_b take the forms

$$\begin{aligned} \psi_a &= \left(\sum_{i \in M_1} \mathbb{I}(C_{1i}=1)(X_{1i} - a_1 - b_1 W_{1i}), \dots, \sum_{i \in M_Q} \mathbb{I}(C_{Qi}=1)(X_{Qi} - a_Q - b_Q W_{Qi}) \right) \\ &= \left(\sum_{i \in M_1} \psi_{a_{1i}}, \dots, \sum_{i \in M_Q} \psi_{a_{0i}} \right) \\ \psi_b &= \left(\sum_{i \in M_1} \mathbb{I}(C_{1i}=1)(X_{1i} - a_1 - b_1 W_{1i})W_{1i}, \dots, \sum_{i \in M_Q} \mathbb{I}(C_{Qi}=1)(X_{Qi} - a_Q - b_Q W_{Qi})W_{Qi} \right) \\ &= \left(\sum_{i \in M_1} \psi_{b_{1i}}, \dots, \sum_{i \in M_Q} \psi_{b_{0i}} \right). \end{aligned}$$

Additionally, $\psi_{\beta_0}, \psi_{\beta_x}, \psi_{\beta_z}$ take the forms

$$\begin{aligned} \psi_{\beta_0} &= \left(\sum_{i \in M_1} \frac{\partial}{\partial \beta_{01}} \log \tilde{l}_{1i}, \dots, \sum_{i \in M_S} \frac{\partial}{\partial \beta_{0S}} \log \tilde{l}_{Si} \right) \\ &= \left(\sum_{i \in M_1} \psi_{\beta_{01i}}, \dots, \sum_{i \in M_S} \psi_{\beta_{0Si}} \right) \\ \psi_{\beta_x} &= \sum_{s=1}^S \sum_{i=1}^{n_s} \frac{\partial}{\partial \beta_x} \log \tilde{l}_{si} \\ &= \sum_{s=1}^S \sum_{i=1}^{n_s} \psi_{xsi}, \\ \psi_{\beta_z} &= \left(\sum_{s=1}^S \sum_{i=1}^{n_s} \frac{\partial}{\partial \beta_{z1}} \log \tilde{l}_{si}, \dots, \sum_{s=1}^S \sum_{i=1}^{n_s} \frac{\partial}{\partial \beta_{zp}} \log \tilde{l}_{si} \right) \\ &= \left(\sum_{s=1}^S \sum_{i=1}^{n_s} \psi_{zsi1}, \dots, \sum_{s=1}^S \sum_{i=1}^{n_s} \psi_{zsi p} \right) \end{aligned}$$

$$\text{where } \frac{\partial}{\partial \beta_k} \log \tilde{l}_{s1} = \mathbb{I}(Y_{si} = 1) - \text{expit}(\beta^T \tilde{X}_{si}), \frac{\partial}{\partial \beta_k} \log \tilde{l}_{si} = \tilde{X}_{si}(\mathbb{I}(Y_{si} = 1) - \text{expit}(\beta^T \tilde{X}_{si})), \text{ and } \frac{\partial}{\partial \beta_{z_p}} \\ = Z_{sip}(\mathbb{I}(Y_{si} = 1) - \text{expit}(\beta^T \tilde{X}_{si})).$$

APPENDIX C

SANDWICH VARIANCE MATRIX

Recall that $\beta_0 = [\beta_{01}, \dots, \beta_{0S}]$, $\beta = [\beta_0, \beta_x, \beta_z]$, $a = [a_1, \dots, a_Q]$, and $b = [b_1, \dots, b_Q]$. Thus, the collective set of parameters to be estimated is $\theta = [a, b, \beta]$.

The sandwich estimator matrix is given by $\hat{V}(\theta) = \hat{B}(\theta)^{-1} \hat{A}(\theta) (\hat{B}(\theta)^{-1})^T$, and $\hat{A}(\theta)$ and $\hat{B}(\theta)$ are square matrices of dimension $2Q + S + p + 1$ with forms given in the following. By evaluating $V(\theta)$ at $\theta = \hat{\theta} = [\hat{a}, \hat{b}, \hat{\beta}]^T$, we can estimate $\widehat{\text{Var}}(\hat{\beta}_x)$ by the appropriate diagonal element of $\hat{V}(\hat{\theta})$.

Let $\Psi_{\theta si} = \left(\psi_{a_{1i}}, \dots, \psi_{a_{Qi}}, \psi_{b_{1i}}, \dots, \psi_{b_{Qi}}, \psi_{\beta_{01i}}, \dots, \psi_{\beta_{0Si}}, \psi_{x_{si}}, \psi_{z_{si1}}, \dots, \psi_{z_{sip}} \right)^T$. The matrix components of $\hat{V}(\theta)$

$$\hat{A}(\theta) = \sum_{s=1}^S \sum_{i=1}^{n_s} \Psi_{\theta si} \Psi_{\theta si}^T$$

$$\hat{B}(\theta) = \sum_{s=1}^S \sum_{i=1}^{n_s} \frac{\partial \Psi_{\theta si}}{\partial \theta^T}.$$

Note that the sandwich matrices for the internalized or full calibration methods will be different due to differential use of X_{si} and $\tilde{X}_{si}$. The second derivatives of the $\hat{B}(\theta)$ matrix are included in the Online Supplement Section S1.

APPENDIX D

CALIBRATION PARAMETERS UNDER CONTROLS-ONLY CALIBRATION

Under the COCS design, we fit the calibration model $E[X|W, Y = 0] = \hat{a}_{co} + \hat{b}_{co}W$ (some subscripts suppressed for brevity). Generally, $\hat{a}_{co}$ and $\hat{b}_{co}$ are not consistent for a and b from Equation (3). However, under some conditions, $\hat{a}_{co} \approx a$. The first condition involves small or null exposure effect. Under the calibration approximation, the distribution of $f(X|W, Z, Y = 0)$ can be rewritten as

$$f(X|W, Z, Y = 0) = \frac{f(X, W, Z, Y = 0)}{f(W, Z, Y = 0)} = \frac{f(Y = 0|X, Z)f(X, W|Z)}{\int f(Y = 0|X, Z, W)f(X, W|Z)dx} = \frac{f(Y = 0|X, Z)f(X, W|Z)}{\int f(Y = 0|X, Z)f(X, W|Z)dx}.$$

Under small effect, $\beta_x \approx 0$, $f(Y = 0|X, Z) \approx f(Y = 0|Z)$, and it follows that

$$\begin{aligned} E[X|W, Z, Y = 0] &= \int \frac{Xf(Y = 0|X, Z)f(X, W|Z)}{\int f(Y = 0|X, Z)f(X, W|Z)dx} dx \approx \int \frac{Xf(Y = 0|Z)f(X, W|Z)}{\int f(Y = 0|Z)f(X, W|Z)dx} dx \\ &= \int \frac{Xf(X, W|Z)}{\int f(X, W|Z)dx} dx = \int Xf(X|W, Z)dx = E[X|W, Z]. \end{aligned}$$

Therefore, $a_{co} \approx a$ and $b_{co} \approx b$.

We can also see $b_{co} \approx b$ under no exposure effect or small disease prevalence through an alternative argument. In either setting, $\text{Cov}(X, W|Y = 0) \approx \text{Cov}(X, W)$ and $\text{Var}(X|Y = 0) \approx \text{Var}(X)$, leading to $b_{co} \approx b$. To be more precise, without making any assumptions on the joint distributions,

$$\begin{aligned} b_{co} &= \frac{\text{Cov}(X, W|Y = 0)}{\text{Var}(W|Y = 0)} = \frac{\text{Cov}(X, W|Y = 0)(1 - \pi)}{\text{Var}(W|Y = 0)(1 - \pi)} \\ &= \frac{\text{Cov}(X, W) - \pi\text{Cov}(X, W|Y = 1) - \text{Cov}(E[X|Y], E[W|Y])}{\text{Var}(W) - \pi\text{Var}(W|Y = 1) - \text{Var}(E[W|Y])}, \end{aligned}$$

where π is the disease prevalence. As $\pi \rightarrow 0$, Y becomes degenerate at 0 and b_{co} tends to $\frac{\text{Cov}(X, W)}{\text{Var}(W)} = b$. Likewise, if there is no association between Y and X ,

$$b_{co} = \frac{\text{Cov}(X, W) - \pi\text{Cov}(X, W)}{\text{Var}(W) - \pi\text{Var}(W)} = b.$$

We now find the analytic forms for a_{co} and b_{co} and compare them to the intercept and slope terms under the marginal linear model $E[X|W] = a + bW$. To this end, we restate the surrogacy assumption as

$$f(Y|X, W) = f(Y|X) \Leftrightarrow \frac{f(Y, X, W)}{f(X, W)} = \frac{f(X, Y)}{f(X)} \Leftrightarrow \frac{f(W|X, Y)f(X, Y)}{f(W|X)f(X)} = \frac{f(X, Y)}{f(X)} \Leftrightarrow f(W|X, Y) = f(W|X).$$

Suppose $E(W|X) = \alpha_0 + \alpha_1 X$. A sufficient, but not necessary, condition to produce this linear conditional mean is multivariate normality of (X, W) . Under this normality, the induced conditional distribution is $W|X \sim N(\alpha_0 + \alpha_1 X, \sigma_{w|x}^2)$, where

$$\alpha_0 = \frac{\sigma_x^2(\mu_x - a) - b^2\sigma_w^2\mu_x}{b\sigma_x^2}, \alpha_1 = \frac{b\sigma_w^2}{\sigma_x^2}, \text{ and } \sigma_{w|x}^2 = \frac{\sigma_w^2(\sigma_x^2 - b^2\sigma_w^2)}{\sigma_x^2}. \text{ Using the restated surrogacy}$$

assumption, we note the following equivalences:

$$\begin{aligned} E[W|Y = 0] &= \int E[W|X = x, Y = 0]f(x|Y = 0)dx = \int E[W|X = x]f(x|Y = 0)dx, \text{ by surrogacy assumption} \\ &= \int (\alpha_0 + \alpha_1 x)f(x|Y = 0)dx = \alpha_0 + \alpha_1 E[X|Y = 0] \end{aligned}$$

$$\begin{aligned} E[XW|Y = 0] &= \int E[XW|X = x, Y = 0]f(x|Y = 0)dx = \int xE[W|X = x]f(x|Y = 0)dx \text{ by surrogacy assumption} \\ &= \int (\alpha_0 x + \alpha_1 x^2)f(x|Y = 0)dx = \alpha_0 E[X|Y = 0] + \alpha_1 E[X^2|Y = 0] \end{aligned}$$

$$\text{Cov}(X, W|Y=0) = E[XW|Y=0] - E[X|Y=0]E[W|Y=0] = \alpha_1(E[X^2|Y=0] - E[X|Y=0]^2) = \alpha_1 \text{Var}(X|Y=0)$$

$$E[W^2|Y=0] = \int E[W^2|x]f(x|Y=0)dx = \int (\sigma_{w|x}^2 + (\alpha_0 + \alpha_1 x)^2)f(x|Y=0)dx = \sigma_{w|x}^2 + \alpha_0^2 + 2\alpha_0\alpha_1 E[X|Y=0] + \alpha_1^2 E[X^2|Y=0]$$

$$\text{Var}(W|Y=0) = E[W^2|Y=0] - E[W|Y=0]^2 = \sigma_{w|x}^2 + \alpha_1^2 \text{Var}(X|Y=0).$$

Thus, the slope and intercept terms obtained from fitting a linear regression model under a COCS design can be expressed as

$$\begin{aligned}\hat{b}_{co} &\rightarrow \frac{\text{Cov}(X, W|Y=0)}{\text{Var}(W|Y=0)} = \frac{\alpha_1 \text{Var}(X|Y=0)}{\sigma_{w|x}^2 + \alpha_1^2 \text{Var}(X|Y=0)} \\ \hat{a}_{co} &\rightarrow E[X|Y=0] - \hat{b}_{co}(\alpha_0 + \alpha_1 E[X|Y=0]).\end{aligned}$$

Therefore, $b_{co} \approx b$ when $\text{Var}(X) \approx \text{Var}(X|Y=0)$, and $a_{co} \approx a$ when $\text{Var}(X) \approx \text{Var}(X|Y=0)$ and $E[X] \approx E[X|Y=0]$. Furthermore, the above expressions also make it straightforward to observe that, if $\sigma_{w|x}^2 \approx 0$, then $a_{co} \approx a$ and $b_{co} \approx b$. Formally, we can express the bias in $\hat{a}_{co}$ and $\hat{b}_{co}$ in terms of the original parameters as follows:

$$\begin{aligned}\text{Bias}(a_{co}) &= \frac{(\sigma_x^2 - b^2 \sigma_w^2)(\sigma_x^2 E[X|Y=0] - \mu_x \text{Var}(X|Y=0) + a(\text{Var}(X|Y=0) - \sigma_x^2))}{b^2 \sigma_w^2 (\text{Var}(X|Y=0) - \sigma_x^2) + \sigma_x^4} \\ \text{Bias}(b_{co}) &= \frac{b(\text{Var}(X|Y=0) - \sigma_x^2)(\sigma_x^2 - b^2 \sigma_w^2)}{b^2 \sigma_w^2 (\text{Var}(X|Y=0) - \sigma_x^2) + \sigma_x^4}.\end{aligned}$$

APPENDIX E

THE ROLE OF $\hat{a}_s$ IN THE FULL CALIBRATION METHOD

Suppose we reparameterize the linear combination of the logistic regression model for a calibrated measurement, namely,

$$\beta_{0s} + \beta_x \tilde{X}_{si} + \beta_z^T Z_{si}, \text{ as } \beta_{0s}^* + \beta_x \hat{b}_s W_{si} + \beta_z^T Z_{si}, \text{ where } \beta_{0s}^* = \beta_{0s} + \hat{a}_s \beta_x. \text{ For presentational}$$

simplicity, we assume that all contributing studies have a calibration process. We demonstrate that solving the estimating equations corresponding to the updated

parameterization, namely, $\left[\Psi_{\beta_0}^*(\hat{a}, \hat{b}), \Psi_{\beta_x}^*(\hat{a}, \hat{b}), \Psi_{\beta_z}^*(\hat{a}, \hat{b}) \right] = \mathbf{0}$, is equivalent to solving the

previously defined estimating equations $\left[\left(\psi_{\beta_0}(\hat{a}, \hat{b}), \psi_{\beta_x}(\hat{a}, \hat{b}), \psi_{\beta_z}(\hat{a}, \hat{b}) \right) \right] = \mathbf{0}$. Similar to the derivation in Appendix B, the estimating equation $\psi_{\beta_x}^*$ is

$$\psi_{\beta_x}^* = \sum_{s=1}^S \sum_{i=1}^{n_s} \hat{b}_s W_{si} \Delta_{si},$$

where $\Delta_{si} = \mathbb{I}(Y_{si} = 1) - \text{expit}(\beta^T \tilde{X}_{si})$. It follows that the estimating equation corresponding to

β_x in the updated parameterization is $\psi_{\beta_x}^* = \psi_{\beta_x} - \sum_{s=1}^S \hat{a}_s \psi_{\beta_{0s}} = \psi_{\beta_x} - \hat{a}^T \psi_{\beta_0}$, where

$\psi_{\beta_{0s}} = \sum_{i=1}^{n_s} \Delta_{si}$. The estimating equations $\psi_{\beta_0}^*$ corresponding to β_0^* are the same as ψ_{β_0} and

similarly $\psi_{\beta_z}^* = \psi_{\beta_z}$. Therefore, the solutions for $(\beta_0, \beta_x, \beta_z)$ obtained by solving

$\left[\psi_{\beta_0}^*(\hat{a}, \hat{b}), \psi_{\beta_x}^*(\hat{a}, \hat{b}), \psi_{\beta_z}^*(\hat{a}, \hat{b}) \right] = \mathbf{0}$ are the same as those obtained by solving

$\left[\left(\psi_{\beta_0}(\hat{a}, \hat{b}), \psi_{\beta_x}(\hat{a}, \hat{b}), \psi_{\beta_z}(\hat{a}, \hat{b}) \right) \right] = \left[\psi_{\beta_x}(\hat{a}, \hat{b}) - \hat{a}^T \psi_{\beta_0}(\hat{a}, \hat{b}), \psi_{\beta_x}(\hat{a}, \hat{b}), \psi_{\beta_z}(\hat{a}, \hat{b}) \right] = \mathbf{0}$. Therefore,

the bias in $\hat{a}_s$ is “absorbed” by the estimated study-specific intercept in the logistic

regression and does not impact the estimation of $\hat{\beta}_x$ or $\hat{\beta}_z$.

REFERENCES

1. Key T, Appleby P, Allen N, Reeves G. Pooling biomarker data from different studies of disease risk, with a focus on endogenous hormones. *Cancer Epidemiol Biomark Prev* 2010;19(4):960–965.
2. Smith-Warner S, Spiegelman D, Ritz J, et al. Methods for pooling results of epidemiologic studies the pooling project of prospective studies of diet and cancer. *Am J Epidemiol* 2006;163(11):1053–1064. [PubMed: 16624970]
3. Tworoger S, Hankinson S. Use of biomarkers in epidemiologic studies: minimizing the influence of measurement error in the study design and analysis. *Cancer Causes Control*. 2006;17(7):889–899. [PubMed: 16841256]
4. McCullough ML, Zoltick ES, Weinstein SJ, et al. Circulating vitamin D and colorectal cancer risk: an international pooling project of 17 cohorts. *J Natl Cancer Inst* 2018.
5. Gallicchio L, Helzlsouer K, Chow W, et al. Circulating 25-hydroxyvitamin D and the risk of rarer cancers: design and methods of the cohort consortium vitamin D pooling project of rarer cancers. *Am J Epidemiol* 2010;172(1):10–20. [PubMed: 20562188]
6. Travis R, Crowe F, Appleby P, et al. Circulating fatty acids and prostate cancer risk: individual participant meta-analysis of prospective studies. *J Natl Cancer Inst* 2014;106(9).
7. Key T, Appleby P, Travis R, et al. Carotenoids, retinol, tocopherols, and prostate cancer risk: pooled analysis of 15 studies. *Am J Clin Nutr* 2015;102(5):1142–1157. [PubMed: 26447150]
8. Tsilidis K, Travis R, Appleby P, et al. Insulin-like growth factor pathway genes and blood concentrations, dietary protein and risk of prostate cancer in the NCI Breast and Prostate Cancer Cohort Consortium (BPC3). *Int J Cancer*. 2013;133(2):495–504. [PubMed: 23341348]
9. Barake M, Daher R, Rose T, Salti I, et al. 25-Hydroxyvitamin D assay variations and impact on clinical decision making. *J Clin Endocrinol Metabol* 2012;97(3):835–843.
10. Lai J, Lucas R, Banks E, et al. Variability in vitamin D assays impairs clinical assessment of vitamin D status. *Intern Med J*. 2012;42(1):43–50. [PubMed: 21395958]

11. Snellman G, Melhus H, Gedeberg R, et al. Determining vitamin D status: a comparison between commercially available assays. *PloS One*. 2010;5(7):e11555.
12. Gail M, Wu J, Wang M, et al. Calibration and seasonal adjustment for matched case-control studies of vitamin D and cancer. *Statist Med* 2016;35(13):2133–2148.
13. Carroll R, Ruppert D, Stefanski L, Crainiceanu C. *Measurement Error in Nonlinear Models: A Modern Perspective*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC; 2006 <https://cds.cern.ch/record/1012043>
14. Rosner B, Spiegelman D, Willett W. 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. [PubMed: 2403114]
15. Gong G, Samaniego F. Pseudo maximum likelihood estimation: theory and applications. *Ann Stat* 1981;861–869.
16. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr* 1997;65(4):1179S–1186S.
17. Spiegelman D, Carroll R, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statist Med* 2001;20(1):139–160.
18. Kuha J. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Statist Med* 1994;13(11):1135–1148.
19. Sun Q, Pan A, Hu F, Manson J, Rexrode K. 25-Hydroxyvitamin D levels and the risk of stroke. *Stroke*. 2012;43(6):1470–1477. [PubMed: 22442173]
20. Wu K, Feskanich D, Fuchs CS, et al. Interactions between plasma levels of 25-hydroxyvitamin D, insulin-like growth factor (IGF)-1 and C-peptide with risk of colorectal cancer. *PLoS One*. 2011;6(12):e28520.
21. Eliassen AH, Warner ET, Rosner B, et al. Plasma 25-hydroxyvitamin D and risk of breast cancer in women followed over 20 years. *Cancer Res* 2016;76(18):5423–5430. [PubMed: 27530324]
22. Eliassen AH, Spiegelman D, Hollis BW, Horst RL, Willett WC, Hankinson SE. Plasma 25-hydroxyvitamin D and risk of breast cancer in the Nurses' Health Study II. *Breast Cancer Res* 2011;13(3):R50. [PubMed: 21569367]
23. Lin D, Zeng D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*. 2010;97(2).

FIGURE 1.

Comparison of percent bias in average calibration parameter estimates over 1000 simulations for a single study when fitting the model $\hat{E}[X_{si}|W_{si}] = \hat{a}_s + \hat{b}_s W_{si}$, where the true $(a_s, b_s) = (3, 0.8)$ and other parameters set at levels as described in Section 3. We consider a common disease with 50% prevalence (panels A and B) and a rare disease with 5% prevalence (panels C and D)

FIGURE 2.

Comparison of methods as the percentage of participants in the calibration study increases. The sample size of each study remains fixed at 1000 with a calibration study participation rate of 5%, 15%, or 25%. Panels A to F use the controls-only calibration study (COCS) design while panels G to L use the random sample calibration study (RSCS) design

TABLE 1

Comparison of operating characteristics under a COCS design for naive ($\hat{\beta}_N$), internalized ($\hat{\beta}_{IN}$), full calibration ($\hat{\beta}_{FC}$), and two-stage ($\hat{\beta}_{TS}$) methods

Prevalence	β_*	Percent Bias (Standard Error)				MSE				Coverage Rate			
		$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$
5	log(1.25)	-29.9 (0.064)	-3.3 (0.081)	-0.1 (0.084)	-0.9 (0.083)	0.0085	0.0066	0.007	0.007	0.82	0.94	0.95	0.95
	log(1.5)	-29.4 (0.064)	-3.0 (0.080)	0.3 (0.083)	-0.6 (0.082)	0.0182	0.0065	0.0068	0.0068	0.53	0.95	0.95	0.96
	log(1.75)	-29.7 (0.064)	-3.4 (0.080)	-0.2 (0.083)	-1.0 (0.083)	0.0319	0.0068	0.0069	0.0069	0.25	0.94	0.95	0.95
	log(2)	-29.5 (0.065)	-3.3 (0.080)	-0.1 (0.083)	-1.0 (0.083)	0.0461	0.007	0.0069	0.007	0.11	0.93	0.94	0.95
	log(2.25)	-29.4 (0.066)	-3.6 (0.081)	-0.3 (0.084)	-1.2 (0.084)	0.0613	0.0073	0.007	0.0071	0.05	0.92	0.94	0.95
	log(2.5)	-29.5 (0.067)	-3.9 (0.081)	-0.6 (0.084)	-1.5 (0.085)	0.0773	0.0078	0.0072	0.0074	0.02	0.90	0.93	0.94
25	β_*	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$
	log(1.25)	-29.7 (0.033)	-3.0 (0.041)	0.2 (0.043)	-0.5 (0.043)	0.0054	0.0018	0.0018	0.0019	0.47	0.94	0.95	0.95
	log(1.5)	-29.8 (0.034)	-3.2 (0.044)	-0.0 (0.045)	-0.8 (0.046)	0.0157	0.0021	0.0021	0.0021	0.06	0.93	0.94	0.95
	log(1.75)	-29.8 (0.036)	-3.4 (0.046)	-0.2 (0.048)	-0.9 (0.049)	0.0292	0.0025	0.0023	0.0024	0.00	0.91	0.94	0.95
	log(2)	-29.7 (0.038)	-3.4 (0.049)	-0.1 (0.051)	-0.9 (0.052)	0.0437	0.003	0.0026	0.0028	0.00	0.90	0.93	0.94
	log(2.25)	-29.7 (0.040)	-3.6 (0.051)	-0.3 (0.054)	-1.1 (0.055)	0.0597	0.0035	0.0029	0.0031	0.00	0.87	0.93	0.94
50	log(2.5)	-29.9 (0.041)	-3.8 (0.054)	-0.5 (0.057)	-1.3 (0.058)	0.0765	0.0041	0.0032	0.0034	0.00	0.86	0.92	0.94
	β_*	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$
	log(1.25)	-29.9 (0.028)	-3.2 (0.036)	0.0 (0.038)	-0.7 (0.040)	0.0053	0.0014	0.0014	0.0016	0.35	0.94	0.95	0.95
	log(1.5)	-29.9 (0.030)	-3.3 (0.039)	-0.0 (0.041)	-0.8 (0.043)	0.0156	0.0017	0.0017	0.0018	0.02	0.93	0.94	0.95
	log(1.75)	-29.7 (0.032)	-3.3 (0.042)	0.0 (0.044)	-0.7 (0.046)	0.0287	0.0021	0.0019	0.0021	0.00	0.91	0.94	0.95
	log(2)	-29.8 (0.034)	-3.4 (0.045)	0.1 (0.047)	-0.7 (0.049)	0.0438	0.0026	0.0022	0.0025	0.00	0.88	0.93	0.95
	log(2.25)	-29.9 (0.035)	-3.6 (0.048)	-0.1 (0.051)	-0.8 (0.053)	0.0599	0.0032	0.0026	0.0029	0.00	0.86	0.92	0.94
	log(2.5)	-30.0 (0.037)	-3.7 (0.051)	-0.1 (0.054)	-0.9 (0.056)	0.0772	0.0037	0.0029	0.0032	0.00	0.85	0.92	0.94

* Percent bias and MSE are computed by $(\hat{\beta} - \beta)/\beta$ and $(\hat{\beta} - \beta)^2$, respectively, and the reported value is the average over 1000 simulations. Standard error is the square root of the empirical variance over all replicates.

Abbreviations: COCS, controls-only calibration study; MSE, mean squared error.

TABLE 2

Comparison of operating characteristics under an **RSCS** design for naive ($\hat{\beta}_N$), internalized ($\hat{\beta}_{IN}$), full calibration ($\hat{\beta}_{FC}$), two-stage ($\hat{\beta}_{TS}$), and two-stage weighted ($\hat{\beta}_{TSW}$) methods

Prevalence	β_x	Percent Bias (Standard Error)					MSE					Coverage Rate				
		$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$
5	log(1.25)	-29.9 (0.064)	-0.2 (0.082)	-0.2 (0.084)	-0.9 (0.088)	-0.9 (0.082)	0.0085	0.0068	0.0070	0.0077	0.0067	0.82	0.95	0.95	0.95	0.95
	log(1.5)	-29.4 (0.064)	0.0 (0.081)	0.1 (0.082)	-0.7 (0.086)	-0.7 (0.081)	0.0182	0.0066	0.0068	0.0074	0.0065	0.53	0.95	0.95	0.96	0.95
	log(1.75)	-29.7 (0.064)	-0.7 (0.081)	-0.7 (0.082)	-1.4 (0.086)	-1.4 (0.080)	0.0319	0.0065	0.0068	0.0075	0.0065	0.25	0.95	0.95	0.95	0.95
	log(2)	-29.5 (0.065)	-0.8 (0.081)	-0.8 (0.082)	-1.6 (0.086)	-1.4 (0.080)	0.0461	0.0066	0.0068	0.0075	0.0066	0.11	0.94	0.94	0.95	0.95
	log(2.25)	-29.4 (0.066)	-1.1 (0.081)	-1.1 (0.082)	-1.9 (0.086)	-1.7 (0.081)	0.0613	0.0066	0.0068	0.0076	0.0067	0.05	0.94	0.94	0.95	0.94
	log(2.5)	-29.5 (0.067)	-1.7 (0.081)	-1.8 (0.082)	-2.6 (0.087)	-2.3 (0.081)	0.0773	0.0068	0.0070	0.0081	0.0070	0.02	0.93	0.93	0.93	0.94
	β_x	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$
25	log(1.25)	-29.7 (0.033)	-0.1 (0.042)	-0.1 (0.043)	-0.8 (0.045)	-0.8 (0.042)	0.0054	0.0018	0.0018	0.0020	0.0018	0.47	0.95	0.95	0.95	0.95
	log(1.5)	-29.8 (0.034)	-0.7 (0.044)	-0.8 (0.045)	-1.6 (0.047)	-1.3 (0.044)	0.0157	0.0019	0.0020	0.0022	0.0019	0.06	0.94	0.94	0.95	0.95
	log(1.75)	-29.8 (0.036)	-1.4 (0.046)	-1.5 (0.047)	-2.2 (0.050)	-1.9 (0.046)	0.0292	0.0022	0.0023	0.0026	0.0022	0.00	0.94	0.94	0.94	0.94
	log(2)	-29.7 (0.038)	-1.9 (0.048)	-2.2 (0.049)	-2.9 (0.052)	-2.4 (0.048)	0.0437	0.0025	0.0026	0.0031	0.0026	0.00	0.92	0.92	0.93	0.93
	log(2.25)	-29.7 (0.040)	-2.6 (0.050)	-3.0 (0.051)	-3.7 (0.054)	-3.2 (0.050)	0.0597	0.0030	0.0032	0.0038	0.0032	0.00	0.90	0.89	0.90	0.91
	log(2.5)	-29.9 (0.041)	-3.3 (0.053)	-3.7 (0.053)	-4.4 (0.056)	-3.8 (0.053)	0.0765	0.0037	0.0040	0.0048	0.0040	0.00	0.87	0.86	0.88	0.89
	β_x	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$
50	log(1.25)	-29.9 (0.028)	-0.2 (0.037)	-0.2 (0.037)	-0.9 (0.039)	-0.8 (0.037)	0.0053	0.0014	0.0014	0.0016	0.0013	0.35	0.95	0.95	0.95	0.95
	log(1.5)	-29.9 (0.030)	-0.8 (0.039)	-1.0 (0.040)	-1.7 (0.042)	-1.4 (0.039)	0.0156	0.0016	0.0016	0.0018	0.0016	0.02	0.94	0.94	0.95	0.94
	log(1.75)	-29.7 (0.032)	-1.5 (0.042)	-1.7 (0.042)	-2.4 (0.044)	-2.1 (0.041)	0.0287	0.0018	0.0019	0.0021	0.0019	0.00	0.93	0.93	0.94	0.94
	log(2)	-29.8 (0.034)	-2.3 (0.045)	-2.6 (0.045)	-3.3 (0.048)	-2.9 (0.044)	0.0438	0.0023	0.0024	0.0028	0.0024	0.00	0.91	0.91	0.91	0.92
	log(2.25)	-29.9 (0.035)	-3.0 (0.046)	-3.4 (0.047)	-4.1 (0.050)	-3.5 (0.046)	0.0599	0.0027	0.0030	0.0036	0.0029	0.00	0.88	0.87	0.89	0.90
	log(2.5)	-30.0 (0.037)	-3.8 (0.048)	-4.3 (0.049)	-5.0 (0.052)	-4.3 (0.048)	0.0772	0.0036	0.0040	0.0048	0.0039	0.00	0.84	0.83	0.85	0.86
	β_x	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$

Abbreviations: RSCS, random sample calibration study; MSE, mean squared error.

TABLE 3

Comparison of operating characteristics for Model CA and Model Z under a **COCS** design for naive ($\hat{\beta}_N$), internalized ($\hat{\beta}_{IN}$), full calibration ($\hat{\beta}_{FC}$), and two-stage ($\hat{\beta}_{TS}$) methods

Model	β_x	Percent Bias (Standard Error)				MSE				Coverage Rate			
		$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$
CA	log(1.25)	-30.0 (0.066)	-8.4 (0.082)	-3.6 (0.084)	-3.7 (0.087)	0.0088	0.0070	0.0071	0.0076	0.82	0.94	0.95	0.95
	log(1.5)	-29.6 (0.067)	-7.0 (0.081)	-2.9 (0.083)	-2.9 (0.086)	0.0188	0.0073	0.0071	0.0076	0.54	0.93	0.94	0.95
	log(1.75)	-30.0 (0.066)	-7.4 (0.080)	-3.4 (0.082)	-3.8 (0.085)	0.0325	0.0081	0.0072	0.0077	0.26	0.91	0.94	0.94
	log(2)	-29.9 (0.066)	-7.3 (0.078)	-3.4 (0.081)	-3.9 (0.084)	0.0473	0.0086	0.0071	0.0078	0.12	0.89	0.94	0.94
	log(2.25)	-29.7 (0.069)	-7.3 (0.082)	-3.5 (0.086)	-3.9 (0.088)	0.0628	0.0102	0.0081	0.0087	0.06	0.85	0.91	0.92
	log(2.5)	-29.8 (0.068)	-7.6 (0.082)	-3.8 (0.085)	-4.2 (0.087)	0.0794	0.0116	0.0085	0.0090	0.02	0.82	0.90	0.92
Z	β_x	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$
	log(1.25)	-30.0 (0.066)	-4.0 (0.084)	-0.6 (0.087)	-0.9 (0.089)	0.0088	0.0071	0.0075	0.0079	0.82	0.94	0.95	0.95
	log(1.5)	-29.6 (0.067)	-3.3 (0.083)	0.1 (0.086)	-0.2 (0.089)	0.0188	0.0070	0.0074	0.0078	0.54	0.95	0.95	0.96
	log(1.75)	-30.0 (0.066)	-4.0 (0.082)	-0.6 (0.085)	-1.2 (0.088)	0.0325	0.0072	0.0072	0.0078	0.26	0.94	0.94	0.95
	log(2)	-29.9 (0.066)	-4.1 (0.079)	-0.7 (0.083)	-1.3 (0.086)	0.0473	0.0071	0.0069	0.0076	0.12	0.93	0.95	0.96
	log(2.25)	-29.7 (0.069)	-4.2 (0.084)	-0.7 (0.088)	-1.3 (0.090)	0.0628	0.0082	0.0077	0.0083	0.06	0.91	0.93	0.94
	log(2.5)	-29.8 (0.068)	-4.5 (0.084)	-1.2 (0.088)	-1.5 (0.090)	0.0794	0.0088	0.0078	0.0082	0.02	0.90	0.93	0.95

TABLE 4

Comparison of operating characteristics for Model CA and Model Z under a **RSCS** design for naive ($\hat{\beta}_N$), internalized ($\hat{\beta}_{IN}$), full calibration ($\hat{\beta}_{FC}$), two-stage ($\hat{\beta}_{TS}$), and two-stage weighted ($\hat{\beta}_{TSW}$) methods

Model	β_x	Percent Bias (Standard Error)					MSE					Coverage Rate				
		$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$
CA	log(1.25)	-30.0 (0.066)	-3.4 (0.082)	-3.7 (0.084)	-3.9 (0.091)	-3.5 (0.085)	0.0088	0.0068	0.0071	0.0083	0.0072	0.82	0.95	0.95	0.95	0.95
	log(1.5)	-29.6 (0.067)	-2.9 (0.082)	-3.1 (0.083)	-3.0 (0.091)	-2.6 (0.084)	0.0188	0.0068	0.0071	0.0083	0.0072	0.54	0.94	0.95	0.95	0.95
	log(1.75)	-30.0 (0.066)	-3.7 (0.080)	-3.9 (0.082)	-4.3 (0.088)	-3.7 (0.082)	0.0325	0.0068	0.0072	0.0083	0.0072	0.26	0.94	0.94	0.95	0.95
	log(2)	-29.9 (0.066)	-4.0 (0.078)	-4.1 (0.080)	-4.5 (0.087)	-4.0 (0.081)	0.0473	0.0068	0.0072	0.0086	0.0073	0.12	0.93	0.94	0.93	0.94
	log(2.25)	-29.7 (0.069)	-4.5 (0.082)	-4.7 (0.084)	-5.0 (0.090)	-4.4 (0.084)	0.0628	0.0081	0.0084	0.0098	0.0083	0.06	0.90	0.90	0.91	0.91
	log(2.5)	-29.8 (0.068)	-5.0 (0.082)	-5.2 (0.084)	-5.5 (0.090)	-4.8 (0.084)	0.0794	0.0089	0.0093	0.0106	0.0089	0.02	0.88	0.88	0.90	0.91
Z	β_x	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$	$\hat{\beta}_N$	$\hat{\beta}_{IN}$	$\hat{\beta}_{FC}$	$\hat{\beta}_{TS}$	$\hat{\beta}_{TSW}$
	log(1.25)	-30.0 (0.066)	-0.6 (0.085)	-0.5 (0.087)	-1.1 (0.095)	-1.0 (0.087)	0.0088	0.0073	0.0075	0.0090	0.0075	0.82	0.95	0.95	0.95	0.96
	log(1.5)	-29.6 (0.067)	-0.1 (0.085)	-0.1 (0.086)	-0.3 (0.095)	-0.4 (0.087)	0.0188	0.0072	0.0074	0.0090	0.0076	0.54	0.95	0.95	0.95	0.95
	log(1.75)	-30.0 (0.066)	-0.9 (0.083)	-0.9 (0.084)	-1.4 (0.092)	-1.3 (0.085)	0.0325	0.0069	0.0071	0.0084	0.0073	0.26	0.94	0.94	0.95	0.95
	log(2)	-29.9 (0.066)	-1.3 (0.080)	-1.3 (0.082)	-2.1 (0.089)	-1.7 (0.083)	0.0473	0.0065	0.0068	0.0082	0.0070	0.12	0.95	0.95	0.95	0.96
	log(2.25)	-29.7 (0.069)	-1.7 (0.084)	-1.7 (0.085)	-2.3 (0.093)	-1.9 (0.086)	0.0628	0.0072	0.0075	0.0089	0.0076	0.06	0.93	0.93	0.94	0.94
	log(2.5)	-29.8 (0.068)	-2.1 (0.084)	-2.2 (0.086)	-2.7 (0.092)	-2.3 (0.087)	0.0794	0.0075	0.0078	0.0092	0.0079	0.02	0.93	0.92	0.94	0.94

TABLE 5

Descriptive characteristics of studies participating in the pooled analysis

Study **	Cases/Controls (Prevalence)	Median 25(OH)D, nmol/L (10%–90%)	Year of Blood Draw	Median Age at Blood Draw, Years (10%–90%) *
HPFS	49/797 (5.8%)	71 (44, 99)	1993–1995	69 (53, 76)
NHS1	103/969 (9.6%)	61 (35, 93)	1989–1991	62 (50, 69)
NHS2	27/1823 (1.5%)	67 (38, 95)	1995–1997	42 (36, 47)
Total	179/3589 (4.8%)	66 (38, 96)	1989–1997	49 (40, 70)

* Quantiles of circulating 25(OH)D levels were obtained from the distribution of calibrated measurements.

** Subsets of size 49+797=846, 103+969=1072, and 27+1823=1850 subjects with 25(OH)D measurements were taken from the HPFS, NHS1, and NHS2 cohorts, respectively.

Abbreviations: HPFS, Health Professionals Follow-up Study; NHS1, Nurses' Health Study; NHS2, Nurses' Health Study II

TABLE 6

Properties of calibration subsets for each participating study under both calibration study designs

Study	Cases/Controls	RSCS		Controls	COCS	
		$\hat{a}(\text{SE})$	$\hat{b}(\text{SE})$		$\hat{a}(\text{SE})$	$\hat{b}(\text{SE})$
HPFS	25-Apr	2.65 (2.85)	0.97 (0.04)	25	4.35 (2.80)	0.94 (0.04)
NHS1	27-Feb	6.03 (2.74)	0.84 (0.04)	27	6.50 (2.88)	0.84 (0.04)
NHS2	28-Jan	8.94 (4.66)	0.95 (0.07)	28	9.01 (4.88)	0.95 (0.07)

Abbreviations: COCS, controls-only calibration study; HPFS, Health Professionals Follow-up Study; NHS1, Nurses' Health Study; NHS2, Nurses' Health Study II; RSCS, random sample calibration study; SE, standard error.

TABLE 7

Pooled parameter and OR estimates for circulating 25(OH)D and stroke utilizing different calibration study designs and methods

Method	RSCS		COCS	
	$\hat{\beta}_x(\text{SE})^*$	OR (95% CI)	$\hat{\beta}_x(\text{SE})^*$	OR (95% CI)
Internalized	−0.050 (0.100)	0.951 (0.782, 1.157)	−0.044 (0.101)	0.957 (0.785, 1.166)
Full calibration	−0.050 (0.101)	0.951 (0.780, 1.159)	−0.044 (0.102)	0.957 (0.784, 1.169)
Two-stage	−0.054 (0.102)	0.947 (0.776, 1.157)	−0.047 (0.102)	0.953 (0.781, 1.165)
Naïve	−0.020 (0.077)	0.980 (0.842, 1.140)		

* Point and OR estimates correspond to a 20 nmol/L increase in circulating 25(OH)D. Results are adjusted for age at blood draw, years of follow-up after blood draw, smoking (never/ever), family history of myocardial infarction (yes/no), hypertension (yes/no), diabetes (yes/no), and BMI (greater than or less than or equal to 25 kg/m²).

Abbreviations: CI, confidence interval; COCS, controls-only calibration study; OR, odds ratio; RSCS, random sample calibration study; SE, standard error.