Published in final edited form as:

Stat Med. 2019 March 30; 38(7): 1245–1261. doi:10.1002/sim.8038.

There is no impact of exposure measurement error on latency estimation in linear models

S.B. Peskoe¹, D. Spiegelman^{1,2,3,4,5}, and M. Wang^{1,2,6}

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

²Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts

³Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, Massachusetts

⁴Department of Global Health, Harvard T.H. Chan School of Public Health, Boston, Massachusetts

⁵Center for Methods in Implementation and Prevention Science, Yale School of Public Health, New Haven, Connecticut

⁶Channing Division of Network Medicine, Department of Medicine, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts

Abstract

Identification of the latency period for the effect of a time-varying exposure is key when assessing many environmental, nutritional, and behavioral risk factors. A pre-specified exposure metric involving an unknown latency parameter is often used in the statistical model for the exposure-disease relationship. Likelihood-based methods have been developed to estimate this latency parameter for generalized linear models but do not exist for scenarios where the exposure is measured with error, as is usually the case. Here, we explore the performance of naive estimators for both the latency parameter and the regression coefficients, which ignore exposure measurement error, assuming a linear measurement error model. We prove that, in many scenarios under this general measurement error setting, the least squares estimator for the latency parameter remains consistent, while the regression coefficient estimates are inconsistent as has previously been found in standard measurement error models where the primary disease model does not involve a latency parameter. Conditions under which this result holds are generalized to a wide class of covariance structures and mean functions. The findings are illustrated in a study of body mass index in relation to physical activity in the Health Professionals Follow-Up Study.

Correspondence: S.B. Peskoe, Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115; or Channing Division of Network Medicine, Department of Medicine, Harvard Medical School and Brigham and Women's Hospital, Boston, MA 02115. speskoe@fas.harvard.edu, M. Wang, Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115; or Channing Division of Network Medicine, Department of Medicine, Harvard Medical School and Brigham and Women's Hospital, 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.

Keywords

bias evaluation; consistency; exposure measurement error; latency estimation; least squares estimation; life course epidemiology; susceptibility

1 | INTRODUCTION

Identification of the latency period of a time-varying exposure can be key when assessing the effects of many environmental, nutritional, and behavioral risk factors. The latency period, sometimes referred to as an exposure window of susceptibility, is the time window leading up to the occurrence of an outcome during which time the exposure predicts the outcome.¹ Outside the latency period, the exposure has no effect. Latency periods have been discussed in many different frameworks. Caplan et al² provided a general formula for weighting exposure metrics in assessing dose-response associations in epidemiological studies. Salvan et al³ proposed methodology for selecting a lag time, which reflects a delay between the time of the exposure and the occurrence of the outcome, in models for relative risk. Hauptmann et al⁴ proposed the use of sliding time windows, which fits a series of risk models that contain cumulative exposure during fixed times intervals, for the exploratory analysis of temporal effects of smoking histories on lung cancer risk. Others have considered spline functions and lag models to estimate latency patterns.^{5–7} Wang et al¹ derived likelihood-based methods to estimate latency parameters for survival models for a range of latency and exposure metrics. Although there have been methods proposed for estimating a latency function or period of susceptibility, thus far, none has considered the performance of these estimators in the presence of exposure measurement error.

Measurement error is a broad term that generally refers to the deviation of some measured value from its true value. It can be present in exposure, confounding, and outcome variables. It can simply be random noise, or it can be dependent on any number of factors, including the true exposure itself and outcome variables. In many cases, measurement error leads to bias.⁸ Methods have been developed to estimate and correct for this bias in a number of settings, including linear and logistic regression,^{8–14} and other nonlinear models.¹⁴ In a regression setting, these methods often implement a regression calibration approach, which requires estimation of, or historical knowledge of, the underlying measurement error model. None of these methods, however, has been applied to a setting where the model includes one or more latency parameters. Therefore, the effect of exposure measurement error on the estimation of a latency parameter, and subsequent regression coefficients, is unknown.

In this paper, we explore the performance of naive least squares estimators for both the latency parameter and the regression coefficients in linear models, when exposure measurement error is present but ignored in the analysis, assuming a linear measurement error model. The findings are illustrated in a study of physical activity in relation to body mass index (BMI) in the Health Professionals Follow-Up Study (HPFS).

2 | METHODS

2.1 | Latency metric

When studying the effect of timing of exposure, for example, a latency period or age-related susceptibility, an appropriate exposure metric needs to be specified. Some of these include mid-life or later-life-related susceptibility windows, exposure during a critical period of susceptibility, and age-related or time-related moving exposure with a lag.¹ In Table 1, we define some time-varying exposure metrics for an individual, i , that include a single latency parameter, a , and a time-varying exposure, $X_i(t)$, which have previously been considered in epidemiologic research.

Though results of this paper can be generalized to all exposure metrics presented in Table 1, we focus on the recent moving average. The recent moving average at a given time t is defined as the average exposure over $(t - a, t)$. With a continuously measured exposure at time t , this can be written as

$$h_i(a, t) = \frac{\int_{t-a}^t X_i(s) ds}{a}.$$

This function is not observed; rather, it is measured at discrete points in time, $X_i(s)$ for $s = s_0, s_1, \dots$. Hence, we represent $h_i(a, t)$ by its discrete approximation using empirical data

$$h_i(a, t) = \frac{\sum_{s=t-a}^t X_i(s)}{a+1}.$$

Recent moving average is often used in air pollution epidemiology (eg, the works of Puett et al¹⁵ and Laden et al¹⁶) where a defines the beginning of the recent exposure susceptibility period. For example, the 12-month (ie, $a = 12$ months) recent moving average of PM_{2.5} has been used as the exposure metric in studying the association of air pollution with all-cause mortality (eg, the work of Hart et al¹⁷). Using this exposure metric, researchers found that increased exposure to PM_{2.5} was associated with an increased risk of mortality.¹⁷ The 72-months (ie, $a = 72$ months) recent moving average was used in a study of air pollution in relation to lung cancer, and the findings using this exposure metric supported those from other studies, indicating an increased risk of incident lung cancer in relation to ambient PM exposures, especially among never- and long-term former smokers (eg, the work of Puett et al¹⁸). Statistical methods have been developed for estimating the point and interval estimates of the latency parameters.¹ This paper considers the behavior of these estimates in the presence of exposure measurement error, a common feature of epidemiological studies. While this paper focuses on the recent moving average, any single-parameter latency metric can be used as $h_i(a, t)$.

2.2 | Exposure-outcome model and linear measurement error

Assume the exposure-outcome model, based on the recent moving average and the latency parameter, is

$$E[Y_i(t_i) | \tilde{X}_i, t_i, C_i] = \beta_0 + \beta_1 h_i(a, t_i) + \beta_2^T C_i \quad (1)$$

where $Y_i(t_i)$ is the continuous outcome measure for individual i at time t_i and $\tilde{X}_i$ represents the history of exposure $X_i(s)$ up to the time of the outcome measurement, $s = t_i$ and C_i is a p -dimensional vector of time-invariant confounders measured without error. All vectors in this paper are column vectors. Define $Z_i(s)$ as the exposure measured with error, with the assumed linear measurement error model

$$X_i(s) = \gamma_0 + \gamma_1 Z_i(s) + \epsilon_i(s), \quad (2)$$

where $\text{var}(\epsilon_i) = \sigma_\epsilon^2$ and $E[\epsilon_i(s)] = 0$, and $E[X_i(s) | Z_i(s), C_i] = \gamma_0 + \gamma_1 Z_i(s)$. As shown in Appendix A.1, it follows that the exposure-outcome model based on $\tilde{Z}_i$, where $\tilde{Z}_i$ represents the history of exposure $Z_i(s)$ up to the time of the outcome measurement, t_i is

$$E[Y_i(t_i) | \tilde{Z}_i, t_i, C_i] = \alpha_0 + \alpha_1 h_i^*(b, t_i) + \alpha_2^T C_i \quad (3)$$

with $h_i^*(b, t) = \frac{\int_{t-b}^t Z_i(s) ds}{b}$ and $\text{var}[Y_i(t_i) | \tilde{Z}_i, t_i, C_i] = \sigma_y^2$. Our goal is to evaluate the bias in the estimate of the true parameters, a , β_0 , β_1 , and β_2 in (1) due to exposure measurement error, should we estimate parameters for Equation (3) from mismeasured exposure data rather than from (1) with perfectly measured exposure data.

It is important to note that the measurement error model given by (2) incorporates a large class of models where Z can depend on X , including the classical additive measurement error model

$$Z_i(s) = X_i(s) + e_i.$$

When, in addition, $e_i \sim N(0, \sigma_e^2)$, $X_i(s) \sim N(\mu_x, \sigma_x^2)$, and $X_i(s)$ and e_i are independent, $X_i(s) = \mu_x(1 - \rho_I) + \rho_I Z_i(s) + \epsilon_i$, $\text{var}(\epsilon_i) = (1 - \rho_I)\sigma_x^2$, and $\rho_I = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}$. The full derivation of this point is given in Appendix A.2.

2.3 | Least squares estimation

We consider a least squares approach for estimating b , α_0 , α_1 , α_2 . This corresponds to minimizing $L = \frac{1}{n} \sum_{i=1}^n l_i$, where n is the number of independent individuals in the study and

$$l_i = \left\{ Y_i(t_i) - \alpha_0 - \alpha_1 h_i^*(b, t_i) - \alpha_2^T C_i \right\}^2.$$

By the Weak Law of Large Numbers,¹⁹ as the sample size (ie, the number of individuals for a fixed amount of follow-up time) goes to infinity, L converges in probability to

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n E_{Y_i, \tilde{Z}_i, C_i} [l_i]. \text{ Define } e(b, \alpha_0, \alpha_1, \alpha_2) \text{ as } E_{Y, \tilde{Z}, C} [L] = \frac{1}{n} \sum_{i=1}^n E_{Y_i, \tilde{Z}_i, C_i} [l_i],$$

where $\tilde{Z} = (\tilde{Z}_1, \tilde{Z}_2, \dots, \tilde{Z}_n)$, $Y = (Y_1, \dots, Y_n)$, and $C = (C_1, \dots, C_n)$. As shown in Appendix A.3,

$$\hat{b}, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2 \xrightarrow{P} \operatorname{argmin} \left[\lim_{n \rightarrow \infty} e(b, \alpha_0, \alpha_1, \alpha_2) \right],$$

where $\xrightarrow{P}$ denotes convergence in probability as $n \rightarrow \infty$, and $\hat{b}, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2$ are the least squares estimators for $b, \alpha_0, \alpha_1, \alpha_2$, respectively. We show in Appendix A.4 that, for any fixed b ,

$$\begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_2 \\ \hat{\alpha}_1 \end{pmatrix} = Q_{(p+2) \times (p+2)}^{-1} \Delta_{(p+2) \times 1},$$

$$\text{where } Q = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} 1 \\ C_i \\ h_i^*(b, t_i) \end{pmatrix} \begin{pmatrix} 1 & C_i^T h_i^*(b, t_i) \end{pmatrix} \text{ and } \Delta = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} 1 \\ C_i \\ h_i^*(b, t_i) \end{pmatrix} \begin{bmatrix} 1 & C_i^T h_i^*(b, t_i) \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix}.$$

Define $g_a = \frac{1}{n} \sum_{i=1}^n h_i^*(a, t_i)$ and $g_b = \frac{1}{n} \sum_{i=1}^n h_i^*(b, t_i)$. Using the results above, we show in Appendix A.5 that, under the surrogacy assumption, which states that conditional on $\tilde{X}_i$ and C_i , $\tilde{Z}_i$ does not provide any additional information about Y_i , the minimum of $e(b, \alpha_0, \alpha_1, \alpha_2)$ is equivalent to the maximum of

$$e^*(b, \alpha_0, \alpha_1, \alpha_2) = E_{\tilde{Z}, C} \left[\frac{\overline{\operatorname{Cov}}_{ab}^2 + 2 \overline{\operatorname{Cov}}_{ab} \overline{\operatorname{Cov}}_{Cb}^T \tilde{C}^{-1} \overline{\operatorname{Cov}}_{Ca} + \overline{\operatorname{Var}}_b \overline{\operatorname{Cov}}_{Ca}^T \tilde{C}^{-1} \overline{\operatorname{Cov}}_{Ca}}{\overline{\operatorname{Var}}_b - \overline{\operatorname{Cov}}_{Cb}^T \tilde{C}^{-1} \overline{\operatorname{Cov}}_{Cb}} \right],$$

where $\overline{\operatorname{Cov}}_{ab} = \frac{1}{n} \sum_{i=1}^n h_i^*(a, t_i) h_i^*(b, t_i) - \frac{1}{n} \sum_{i=1}^n h_i^*(a, t_i) \left(\frac{1}{n} \sum_{i=1}^n h_i^*(b, t_i) \right)$ is the empirical covariance of $h_i^*(a, t_i), h_i^*(b, t_i)$, $\overline{\operatorname{Cov}}_{Cb} = \frac{1}{n} \sum_{i=1}^n C_i h_i^*(b, t_i) - \frac{1}{n} \sum_{i=1}^n C_i \left(\frac{1}{n} \sum_{i=1}^n h_i^*(b, t_i) \right)$ is a column vector of the empirical covariance of $C_i, h_i^*(b, t_i)$,

$\overline{\operatorname{Cov}}_{Ca} = \frac{1}{n} \sum_{i=1}^n C_i h_i^*(a, t_i) - \frac{1}{n} \sum_{i=1}^n C_i \left(\frac{1}{n} \sum_{i=1}^n h_i^*(a, t_i) \right)$ is a column vector of the empirical covariance of $C_i, h_i^*(a, t_i)$, $\overline{\operatorname{Var}}_b = \frac{1}{n} \sum_{i=1}^n h_i^*(b, t_i)^2 - \left(\frac{1}{n} \sum_{i=1}^n h_i^*(b, t_i) \right)^2$ is the scalar empirical

variance of $h_i^*(b, t_i)$, and $\tilde{\Sigma}_C = \frac{1}{n} \sum_{i=1}^n C_i C_i^T - \frac{1}{n} \sum_{i=1}^n C_i \left(\frac{1}{n} \sum_{i=1}^n C_i \right)^T$ is the empirical variance-covariance matrix of C_i .

2.4 | Consistency of the least squares estimator of the latency parameter

We can see simply that, in the absence of confounders (ie, $p = 0$), e^* reduces to

$$E_Z \left[\frac{\overline{\text{Cov}}_{ab}^2}{\overline{\text{Var}}_b} \right] = E_Z \left[\frac{\overline{\text{Var}}_b \overline{\text{Var}}_a \tilde{\rho}_{ab}}{\overline{\text{Var}}_b} \right] = E_Z [\overline{\text{Var}}_a \tilde{\rho}_{ab}], \text{ where}$$

$\overline{\text{Var}}_a = \frac{1}{n} \sum_{i=1}^n h_i^*(a, t_i)^2 - \left(\frac{1}{n} \sum_{i=1}^n h_i^*(a, t_i) \right)^2$ is the scalar empirical variance of $h_i^*(a, t_i)$ and is

not a function of b , and $\tilde{\rho}_{ab} = \frac{\overline{\text{Cov}}_{ab}}{\sqrt{\overline{\text{Var}}_a \overline{\text{Var}}_b}}$ is the empirical correlation between $h_i^*(a, t_i)$ and

$h_i^*(b, t_i)$. Given that, any correlation is bounded between -1 and 1 , and that $\overline{\text{Var}}_a > 0$, the maximum of e^* is equal to the maximum of $\tilde{\rho}_{ab}$, which is 1 . Under the assumption that $\tilde{\rho}_{ab} = 1$ if and only if $a = b$, this implies that e^* is maximized at $b = a$, and the least squares estimator, $\hat{b}$, is consistent for the true latency parameter for the exposure measured without error, a .

We further show in Appendix A.6 that, when the empirical correlation between C_i and $h_i^*(b, t_i)$ is constant over time (ie, $\overline{\text{Cov}}_{Cb}$ is not a function of b), the expression for $e^*(b, \alpha_0, \alpha_1, \alpha_2)$ in Section 2.3 is again maximized when $\tilde{\rho}_{ab} = 1$. Under the same assumption that $\tilde{\rho}_{ab} = 1$ if and only if $a = b$, $e^*(b, \alpha_0, \alpha_1, \alpha_2)$ is maximized at $b = a$, and the least squares estimator for the latency parameter is consistent for the true latency parameter, a .

Conditions under which $\hat{b} \xrightarrow{P} a$ regarding covariances can empirically verified in any given dataset. They also allow for a wide class of covariance structures, including but not limited to, the compound symmetric covariance structure where the covariance of Z_i for repeated measures of Z within an individual is constant at $\rho < 1$ for discrete time points, and AR(1) covariance structures, where the correlation in Z_i between two time points is of the form $\rho^{|t_1 - t_2|} < 1$ where $\rho < 1$ for discrete and continuous time points. Proofs can be found in the web supplement.

Although the least squares estimator for the latency parameter is consistent in the presence of exposure measurement error, the estimated regression coefficients of the exposure-outcome association are not necessarily consistent. When the latency parameter is known, the least squares regression of the outcome model in (3) reduces to ordinary linear regression with $h_i^*(a, t_i)$ as the exposure of interest, which is known to be biased in the presence of exposure measurement error.¹⁴ If we know or can estimate the linear model for the measurement error in the exposure (eg, by a validation study), we can use this information to derive the measurement error model for the latency metric. For example, according to the recent moving average exposure metric, if $X_i(s) = \gamma_0 + \gamma_1 Z_i(s) + \epsilon_i(s)$, where γ_0 and γ_1 are

known or can be estimated and $\epsilon_i(s)$ are independent, identically distributed with mean 0 and finite variance, then it follows that

$$h_i(a, t_i) = \frac{\int_{t_i-a}^{t_i} X_i(s) ds}{a} = \frac{\int_{t_i-a}^{t_i} \gamma_0 + \gamma_1 Z_i(s) + \epsilon_i(s) ds}{a} = \gamma_0 + \gamma_1 h_i^*(a, t_i) + \tilde{\epsilon}_i$$

where $\tilde{\epsilon}_i = \frac{1}{a} \int_{t_i-a}^{t_i} \epsilon_i(s) ds$ has mean 0 and finite variance. Thus, we are able to derive the measurement error model for the recent moving average. Note that a similar approach can be taken for any of the exposure metrics. This will allow us to correct for any bias in the least squares estimates of α_0 and α_1 , using regression calibration-type methods,²² ultimately giving us consistent estimates of the latency parameter and regression coefficients.

Regarding the variance estimates, we can first use the modified Hessian matrix method¹ to estimate the variance of the naive estimates of α_0 , α_1 , and α_2 , ignoring measurement error but accounting for the variation due to estimating the latency parameters, and then, we can use the Delta method to obtain the variance of the corrected estimates as used in the regression calibration method.¹³

3 | ILLUSTRATIVE EXAMPLE

We used the results above to examine the timing of physical activity history in relation to BMI in the HPFS²³ and to evaluate the performance of the least squares estimator for a single latency parameter, which estimates how far in the past physical activity affects BMI, in the presence of exposure measurement error in physical activity.

Study population

We included 16 731 men who were participants in the HPFS with complete BMI information in 2006, complete physical activity history between 1986 and 2006, and reported age and race. HPFS is a prospective cohort study that began in 1986, enrolling 51 529 men, ages 40 to 75 years.²⁴ Baseline questionnaires were mailed to all participants to collect information on demographics, medical history, and lifestyle factors. A follow-up questionnaire was mailed to participants every 2 years collecting information, including current physical activity and BMI. As shown in Table 2, the men in the HPFS were primarily white, reflecting the demographics of male health professionals during the era in which they trained. Mean reported physical activity increased over time, which is consistent with previous studies of the long-term physical activity among health professionals.²⁵

Physical activity and BMI assessment

Physical activity is defined as total activity MET-hrs/wk. Participants reported hours spent performing a list of physical activities on each questionnaire. Weekly expenditure of metabolic equivalents (METs-hrs/wk) for each of these activities and total weekly expenditure were calculated. Last observed exposure was carried forward when activity was missing for a given questionnaire or if the questionnaire was not returned. Physical activity was treated as a continuous variable in this analysis. The outcome of interest in this analysis

was BMI reported in 2006. Men who died prior to 2007, men who were missing age in 1986, and men who were missing race information were excluded from the analysis.

Latency estimation

We fit a linear regression model with BMI in 2006 as the dependent variable and the recent moving average of physical activity $h_i(a) = \frac{\sum_{t=2006-a}^{2006} X_i(t)}{a+1}$, where $X_i(t)$ is the physical activity for individual i at time t , as the outcome, adjusting for race and age. The latency parameter was selected as that corresponding to the recent moving average that, when included in the linear regression model, minimized the mean squared error (MSE). In Figure 1, we see the MSE plotted for each possible value of the latency parameter. We conclude the “true” latency parameter is 10 years.

4 | MEASUREMENT ERROR SIMULATION

To assess the impact of measurement error on the estimation of the latency parameter, we performed a simulation to add in varying amounts of measurement error, based on these data from HPFS. For each simulation, we sampled from 16 731 men, with replacement, augmented their physical activity history, and estimated the latency parameter by the same procedure as the full cohort. Measurement error was added using the following approach:

Step 1: sample size N from full cohort: $\mathbf{X}_{N \times t}$ matrix of physical activity history, where N ranged from 50 to 1 million;

Step 2: generate $\mathbf{E}_{N \times t}$ error matrix;

Step 3: estimate latency parameter using $\mathbf{Z}_{N \times t} = \mathbf{X}_{N \times t} + \mathbf{E}_{N \times t}$.

The error matrix, $\mathbf{E}_{N \times t}$ was generated by sampling n times from a multivariate normal with mean μ and covariance matrix Σ . These parameters were varied such that $\mu = 0, t$, or $\sin(t\pi)$; the compound symmetric covariance matrix Σ had $\text{diag}(\Sigma) = \sigma^2$ and $\text{off-diag}(\Sigma) = \rho\sigma^2$, where $\sigma = 10, 100, 1000$, or 2000 and $\rho = 0.2$ or 0.4 ; and $N \in [50, 1\,000\,000]$. We also included a simulation where no error was added for comparison. Results are shown in Table 3.

We see that, for all covariance structures considered and all mean functions for the error matrix, the percent bias in the mean estimated latency parameter approaches 0 as sample size increases. As expected, when the sample size is small, the percent bias in the mean estimated latency parameter is high and unstable; however, this appears to be primarily driven by the sample size rather than the measurement error, as can be seen by comparing to the no-error case in the top row. As the measurement error increases (eg, larger variability in errors), the percent bias also increases relative to simulation with no measurement error. Note that these simulations extend beyond the linear measurement error model in (2) as they include a possible time-varying mean, but we observe that the mean of the measurement error has little effect on the bias in the latency parameter in all simulations. However, this

mean would affect the least squares estimator of the regression coefficients, as has been consistently reported in the measurement error literature (eg, the work of Carroll et al¹⁴).

5 | DISCUSSION

As seen from the analytic results, the naive least squares estimator for the latency parameter is consistent under the assumption that $\text{Corr}[h_i^*(a, t_i), h_i^*(b, t_i)] = 1$ if and only if $a = b$. This applies to linear measurement error models, where the errors in the outcome model are independently, identically distributed with mean 0 and finite variance, and the empirical correlation between C_i and $h_i^*(b, t_i)$ does not depend on b . These results were found to hold in a simulation study, as demonstrated by augmenting physical activity data with additional synthetic error in the HPFS. If we know or can estimate the model for the measurement error in the exposure by a validation study, we can use extrapolate using the estimate of the latency parameter and the functional form of the latency metric to determine the measurement error model in the latency metric. For instance, when the measurement error is linear, the measurement error in the recent moving average will also be linear. This will allow us to correct for any bias in the least squares estimates of the regression coefficients.

The primary limitation of the key result is that correlation between C_i and $h_i^*(b, t_i)$ is fixed, or independent of b . While this may be a reasonable assumption in many settings, for instance, when demographics or sex is the confounder, there are certainly situations in which this may not be a reasonable assumption. For example, consider the situation where C is a continuous baseline exposure. It is plausible that C is strongly correlated with $Z(0)$, eg, at baseline, but less correlated with $Z(20)$, eg, after 20 years of follow-up. This would imply that the correlation between C_i and $h_i^*(b, t_i)$ changes as a function of b . Further investigation to the impact of this heteroskedasticity is required to determine the effect it may have on the consistency of the least squares estimator when the correlation between C and $h_i^*(b, t_i)$ changes with time.

In future research, we hope to investigate the extent to which these results hold in generalized linear models and survival models. In this paper, a linear relationship between the true measurement, X , and the measurement with error, Z , is required, as in measurement error model (2), for there to be consistency of the latency parameter estimates in the presence of measurement error. This modeling assumption has been found empirically to be reasonable in many epidemiologic applications in nutrition and environmental health (eg, the works of Horick et al²⁷ and Rosner et al²⁸). Future work will study the impact of a nonlinear measurement error model on the behavior of the latency parameter estimates.

In summary, the naive least squares estimator for the latency parameter is consistent. When the model for the measurement error in the exposure is known or can be estimated, this information can be used to model the measurement error in the latency metric, namely, the recent moving average, and existing approaches for bias correction can be implemented to achieve unbiased regression coefficient estimators.

ACKNOWLEDGEMENT

This research was supported by NIH grants DP1ES025459, R01ES026246, and U01CA167552 (for HPFS).

Funding information

NIH, Grant/Award Number: DP1ES025459, R01ES026246, and U01CA167552 (for HPFS)

APPENDIX

A.1 | Derivation of the exposure-outcome model based on $\tilde{Z}_i$

By the surrogacy assumption, $E_{Y_i|\tilde{Z}_i, C_i} = E_{\tilde{X}_i|\tilde{Z}_i, C_i} E_{Y_i|\tilde{Z}_i, \tilde{X}_i, C_i} = E_{\tilde{X}_i|\tilde{Z}_i, C_i} E_{Y_i|\tilde{X}_i, C_i}$; thus we can write

$$\begin{aligned} E_{Y_i|\tilde{Z}_i, C_i} &= E_{\tilde{X}_i|\tilde{Z}_i, C_i} E_{Y_i|\tilde{X}_i, C_i} \\ &= E_{\tilde{X}_i|\tilde{Z}_i, C_i} \left[\beta_0 + \beta_1 h_i(a, t_i) + \beta_2^T C_i \right] \\ &= \beta_0 + \beta_1 E_{\tilde{X}_i|\tilde{Z}_i, C_i} [h_i(a, t_i)] + \beta_2^T C_i. \end{aligned}$$

Given that h_i can be written as an integral, and the expectation with respect to $\tilde{X}_i$ is an integral, we can exchange the order of integration, and indeed, $E_{Y_i|\tilde{Z}_i, C_i}$ is a linear function of the latency metric using $\tilde{Z}_i$ and C_i .

A.2 | Classical additive measurement error model

Assume the classical additive measurement error model

$$Z_i(s) = X_i(s) + e_i,$$

where $e_i \sim N(0, \sigma_e^2)$, $X_i \sim N(\mu_x, \sigma_x^2)$, and X_i and e_i are independent. We can express this as a specific case of the measurement error model in (2); that is,

$$X_i(s) = \gamma_0 + \gamma_1 Z_i(s) + e_i.$$

In the following, we derive the expressions of γ_0 , γ_1 , and $\text{var}(e_i)$.

$$\text{since } \begin{pmatrix} X_i \\ Z_i \end{pmatrix} \sim N_2 \left(\begin{bmatrix} \mu_x \\ \mu_x \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 + \sigma_e^2 \end{bmatrix} \right),$$

$$X_i \left| Z_i \sim N \left(\mu_x + \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2} (Z_i - \mu_x), \sigma_x^2 - \frac{\sigma_x^4}{\sigma_x^2 + \sigma_e^2} \right).$$

It follows that $E[X_i | Z_i] = \mu_x(1 - \rho_I) + \rho_I Z_i$,

$$\text{and } \text{Var}[e_i] = (1 - \rho_I)\sigma_x^2.$$

$$\text{Therefore, } \gamma_0 = \mu_x(1 - \rho_I)$$

$$\gamma_1 = \rho_I.$$

A.3 | Proof that $\hat{b}, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2 \xrightarrow{P} \text{argmin}[\lim_{n \rightarrow \infty} e(b, \alpha_0, \alpha_1, \alpha_2)]$

Let $(a_0, a_1, a_2, b) \equiv \xi_0$. That is, for the sequence of real-valued outcomes, Y_i has the structure

$$Y_i = f_i(\xi_0) + e_i,$$

where $f_i(\xi_0) = \alpha_0 + \alpha_1 h_i^*(b, t_i) + \alpha_2^T C_i$ is a known, continuous function. In order to estimate ξ_0 , we use $e = E_{Y, \tilde{Z}, C}(L)$ and denote

$$q(\xi) = \frac{\partial L}{\partial \xi}$$

the estimating function from L . We know that $E[q(\xi)]|_{\xi = \xi_0} = 0$ because ξ_0 is the true value.

So, we can write

$$\begin{aligned} \frac{\partial e}{\partial \xi} &= E_{Y, \tilde{Z}, C} \left[\frac{\partial L}{\partial \xi} \right] \\ &= E[q(\xi)], \end{aligned}$$

where we exchange between expectation and derivatives under regularity conditions. That is,

$$\frac{\partial e}{\partial \xi}(\xi_0) = 0.$$

If e has a unique global minimum, this implies that $\text{argmin}[e(b, \alpha_0, \alpha_1, \alpha_2)] = \xi_0$, which holds as $n \rightarrow \infty$. Finally, we can use the result shown in the work of Jennrich²⁰ that, under regularity conditions, the least squares estimators for nonlinear equations are consistent. These regularity conditions are met for (3), which proves

$$\hat{b}, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2 \xrightarrow{P} \xi_0 = \text{argmin}[\lim_{n \rightarrow \infty} e(b, \alpha_0, \alpha_1, \alpha_2)].$$

A.4 | Derivation of equations for $\hat{\alpha}_0$, $\hat{\alpha}_1$, and $\hat{\alpha}_2$

Given b , and the linear nature of model for $Y_i | \tilde{Z}_i, C_i$, we can write $\hat{\alpha}_0$, $\hat{\alpha}_1$, and $\hat{\alpha}_2$ as functions of b using the usual estimating equations

$$0 = \frac{1}{n} \sum_{i=1}^n E_{Y_i | \tilde{Z}_i, C_i} \left[\begin{pmatrix} 1 \\ C_i \\ h_i^*(b, t_i) \end{pmatrix} \left[y_i - (\hat{\alpha}_0 + \hat{\alpha}_1 h_i^*(b, t_i) + \hat{\alpha}_2^T C_i) \right] \right].$$

By the surrogacy assumption, we again have

$$0 = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} 1 \\ C_i \\ h_i^*(b, t_i) \end{pmatrix} \left[(\beta_0 + \beta_1 \gamma_0 + \beta_1 \gamma_1 h_i^*(a, t_i) + \beta_2^T C_i) - (\hat{\alpha}_0 + \hat{\alpha}_1 h_i^*(b, t_i) + \hat{\alpha}_2^T C_i) \right].$$

$$\text{Define } \Delta = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} 1 \\ C_i \\ h_i^*(b, t_i) \end{pmatrix} (\beta_0 + \beta_1 \gamma_0 + \beta_1 \gamma_1 h_i^*(a, t_i) + \beta_2^T C_i).$$

$$\text{It follows that } \Delta = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} 1 \\ C_i \\ h_i^*(b, t_i) \end{pmatrix} \begin{bmatrix} 1 & C_i^T & h_i^*(b, t_i) \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix}.$$

$$\text{Therefore, } \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_2 \\ \hat{\alpha}_1 \end{pmatrix} = Q^{-1} \times (p+2) \times \Delta \times (p+2) \times 1, \quad \text{where } Q = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} 1 \\ C_i \\ h_i^*(b, t_i) \end{pmatrix} \begin{bmatrix} 1 & C_i^T & h_i^*(b, t_i) \end{bmatrix}.$$

A.5 | Minimizing $e(b, \alpha_0, \alpha_1, \alpha_2)$

By the surrogacy assumption, we can rewrite

$$e = E_{Y_i, \tilde{Z}_i, C_i} [l_i] \text{ as } e = E_{\tilde{Z}_i, C_i} \left\{ E_{\tilde{X}_i | \tilde{Z}_i, C_i} E_{Y_i | \tilde{X}_i, C_i} l_i \right\}. \text{ Therefore, we can write out } e(b, \alpha_0, \alpha_1, \alpha_2) \text{ as}$$

$$\begin{aligned}
& e(b, \alpha_0, \alpha_1, \alpha_2) \\
&= E_{Y, \tilde{Z}, \tilde{C}} \left[\frac{1}{n} \sum_{i=1}^n l_i \right] \\
&= \frac{1}{n} \sum_{i=1}^n E_{\tilde{Z}_i, C_i} \left[E_{\tilde{X}_i | \tilde{Z}_i, C_i} \left[E_{Y_i | \tilde{X}_i, C_i} [l_i] \right] \right] \\
&= \frac{1}{n} \sum_{i=1}^n E_{\tilde{Z}_i, C_i} \left[E_{\tilde{X}_i | \tilde{Z}_i, C_i} \left\{ \left[\alpha_0 + \alpha_1 h_i^*(b, t_i) + \alpha_2^T C_i \right]^2 - 2 \left[\beta_0 + \beta_1 h_i(a, t_i) + \beta_2^T C_i \right] \left[\alpha_0 + \alpha_1 h_i^*(b, t_i) + \alpha_2^T C_i \right] \right\} \right] \\
&= \frac{1}{n} \sum_{i=1}^n E_{\tilde{Z}_i, C_i} \left\{ \left(\alpha_0 \alpha_2^T \alpha_1 \right) \left(\begin{array}{c} 1 \\ C_i \\ h_i^*(b, t_i) \end{array} \right) \right\}^2 - 2 \left[\beta_0 + \beta_1 \gamma_0 + \beta_1 \gamma_1 h_i^*(a) + \beta_2^T C_i \right] \left(\alpha_0 \alpha_2^T \alpha_1 \right) \left(\begin{array}{c} 1 \\ C_i \\ h_i^*(b, t_i) \end{array} \right) \right\} \\
&= E_{\tilde{Z}, C} \left\{ \frac{1}{n} \sum_{i=1}^n \left[\Delta^T Q^{-1} \left[1 \ C_i^T \ h_i^*(b, t_i) \right]^T \right]^2 - 2 \frac{1}{n} \sum_{i=1}^n \Delta^T Q^{-1} \left(\begin{array}{c} 1 \\ C_i \\ h_i^*(b, t_i) \end{array} \right) \left[\beta_0 + \beta_1 \gamma_0 + \beta_1 \gamma_1 h_i^*(a) + \beta_2^T C_i \right] \right\} \\
&= E_{\tilde{Z}, C} \left\{ \frac{1}{n} \sum_{i=1}^n \left[1 \ C_i^T \ h_i^*(b, t_i) \right] Q^{-1} \Delta \Delta^T Q^{-1} \left[1 \ C_i^T \ h_i^*(b, t_i) \right]^T - 2 \Delta^T Q^{-1} \Delta \right\} \\
&= E_{\tilde{Z}, C} \left\{ \text{trace} \left(\frac{1}{n} \sum_{i=1}^n \left[1 \ C_i^T \ h_i^*(b, t_i) \right] Q^{-1} \Delta \Delta^T Q^{-1} \left[1 \ C_i^T \ h_i^*(b, t_i) \right]^T \right) - 2 \Delta^T Q^{-1} \Delta \right\} \\
&= E_{\tilde{Z}, C} \left\{ \text{trace} \left(\frac{1}{n} \sum_{i=1}^n \left[1 \ C_i^T \ h_i^*(b, t_i) \right]^T \left[1 \ C_i^T \ h_i^*(b, t_i) \right] Q^{-1} \Delta \Delta^T Q^{-1} \right) - 2 \Delta^T Q^{-1} \Delta \right\} \\
&= E_{\tilde{Z}, C} \left\{ \text{trace} \left(Q Q^{-1} \Delta \Delta^T Q^{-1} \right) - 2 \Delta^T Q^{-1} \Delta \right\} \\
&= E_{\tilde{Z}, C} \left\{ \text{trace} \left(\Delta^T Q^{-1} \Delta \right) - 2 \Delta^T Q^{-1} \Delta \right\} \text{ (using the cyclic property of the function)} \\
&= E_{\tilde{Z}, C} \left\{ -\Delta^T Q^{-1} \Delta \right\}.
\end{aligned}$$

Define

$$\begin{aligned}
g_a &= \frac{1}{n} \sum_{i=1}^n h_i^*(a, t_i), & g_b &= \frac{1}{n} \sum_{i=1}^n h_i^*(b, t_i), \\
\tilde{g} &= \frac{1}{n} \sum_{i=1}^n h_i^*(a, t_i) h_i^*(b, t_i), & \tilde{g}_b &= \frac{1}{n} \sum_{i=1}^n h_i^*(b, t_i)^2, \\
\bar{C} &= \frac{1}{n} \sum_{i=1}^n C_i & & (p \times 1 \text{ vector}), \\
C_a &= \frac{1}{n} \sum_{i=1}^n C_i h_i^*(a, t_i) & & (p \times 1 \text{ vector}), \\
C_b &= \frac{1}{n} \sum_{i=1}^n C_i h_i^*(b, t_i) & & (p \times 1 \text{ vector}), \\
\text{and } \overline{C^2} &= \frac{1}{n} \sum_{i=1}^n C_i C_i^T & & (p \times p \text{ matrix}).
\end{aligned}$$

Let $\psi_0 = \beta_0 + \beta_1 \gamma_0 + \beta_2^T \bar{C}$ and $\psi_1 = \beta_1 \gamma_1$. Note that $\Delta = \begin{pmatrix} \psi_0 + \psi_1 g_a \\ \bar{C} \psi_0 + \psi_1 C_a \\ \psi_0 g_b + \psi_1 \tilde{g} \end{pmatrix}$, which we denote as

$\begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix}$. Define

$$\begin{aligned}
Q_{(p+2) \times (p+2)} &= \begin{pmatrix} 1 & \bar{C}^T & g_b \\ \bar{C} & \bar{C}^2 & C_b \\ g_b & C_b^T & \bar{g}_b \end{pmatrix} = \begin{pmatrix} A = \begin{bmatrix} 1 & \bar{C}^T \\ \bar{C} & \bar{C}^2 \end{bmatrix} & B = \begin{bmatrix} g_b \\ C_b \end{bmatrix} \\ B^T = \begin{bmatrix} g_b & C_b^T \end{bmatrix} & D = \begin{bmatrix} \bar{g}_b \end{bmatrix} \end{pmatrix}, \\
Q^{-1} &= \begin{pmatrix} A^{-1} + A^{-1}B(D - B^T A^{-1}B)^{-1}B^T A^{-1} & -A^{-1}B(D - B^T A^{-1}B)^{-1} \\ -(D - B^T A^{-1}B)^{-1}B^T A^{-1} & (D - B^T A^{-1}B)^{-1} \end{pmatrix} \\
&= \begin{pmatrix} A^{-1} + \frac{1}{k}A^{-1}BB^T A^{-1} & -\frac{1}{k}A^{-1}B \\ -\frac{1}{k}B^T A^{-1} & \frac{1}{k} \end{pmatrix} \\
&= \frac{1}{k} \begin{pmatrix} kA^{-1} + A^{-1}BB^T A^{-1} & -A^{-1}B \\ -B^T A^{-1} & 1 \end{pmatrix},
\end{aligned}$$

where $k = D - B^T A^{-1}B$ is a scalar,

$$\begin{aligned}
\text{and } A^{-1} &= \begin{pmatrix} 1 + 1^{-1}\bar{C}^T(\bar{C}^2 - \bar{C}1^{-1}\bar{C}^T)^{-1}\bar{C}1^{-1} & -1^{-1}\bar{C}^T(\bar{C}^2 - \bar{C}1^{-1}\bar{C}^T)^{-1} \\ -(\bar{C}^2 - \bar{C}1^{-1}\bar{C}^T)^{-1}\bar{C}1^{-1} & (\bar{C}^2 - \bar{C}1^{-1}\bar{C}^T)^{-1} \end{pmatrix} \\
&= \begin{pmatrix} 1 + \bar{C}^T(\bar{C}^2 - \bar{C}\bar{C}^T)^{-1}\bar{C} & -\bar{C}^T(\bar{C}^2 - \bar{C}\bar{C}^T)^{-1} \\ -(\bar{C}^2 - \bar{C}\bar{C}^T)^{-1}\bar{C} & (\bar{C}^2 - \bar{C}\bar{C}^T)^{-1} \end{pmatrix} \\
&= \begin{pmatrix} 1 + \bar{C}^T K^{-1}\bar{C} & -\bar{C}^T K^{-1} \\ -K^{-1}\bar{C} & K^{-1} \end{pmatrix}, \text{ where } K_{p \times p} = \bar{C}^2 - \bar{C}\bar{C}^T.
\end{aligned}$$

Denote $F = A^{-1} = \begin{pmatrix} f_{11(1 \times 1)} & f_{12(1 \times p)} \\ f_{12(1 \times p)}^T & f_{22(p \times p)} \end{pmatrix}$, where $f_{22} = K_p^{-1}$. Note that A , and subsequently,

K and F , do not depend on b or Z . Minimizing $e = E_{\tilde{Z}_i C_i} \{-\Delta^T Q^{-1} \Delta\}$ is equivalent to

maximizing $\tilde{e} = E_{\tilde{Z}_i C_i} \{\Delta^T Q^{-1} \Delta\}$, and

$$\begin{aligned}
\tilde{e} &= E_{\tilde{Z}, C} \left\{ \Delta^T \begin{pmatrix} kF + FBB^T F & -FB \\ -B^T F & 1 \end{pmatrix} \begin{pmatrix} \delta \\ \delta_3 \end{pmatrix} \frac{1}{k} \right\}, \text{ where } \delta_{(p+1) \times 1} = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} \\
&= E_{\tilde{Z}, C} \left\{ \begin{pmatrix} \delta^T & \delta_3 \end{pmatrix} \begin{pmatrix} kF\delta + FBB^T F\delta & -FB\delta_3 \\ -B^T F\delta & \delta_3 \end{pmatrix} \frac{1}{k} \right\} \\
&= E_{\tilde{Z}, C} \left\{ \frac{\delta^T kF\delta + \delta^T FBB^T F\delta - \delta^T FB\delta_3 - \delta_3 B^T F\delta + \delta_3^2}{k} \right\} \\
&= E_{\tilde{Z}, C} \left\{ \frac{\delta^T kF\delta + \delta^T FBB^T F\delta - 2\delta^T FB\delta_3 + \delta_3^2}{k} \right\} \\
&= E_{\tilde{Z}, C} \left\{ \frac{kG\delta + GBB^T G^T - 2GB\delta_3 + \delta_3^2}{\tilde{g}_b - B^T FB} \right\},
\end{aligned}$$

where $G = \delta^T F$ is a $1 \times (p+1)$ vector. Note that $E_{\tilde{Z}, C} [kG\delta/\tilde{g}_b - B^T FB] = E_{\tilde{Z}, C} [G\delta]$ is not a function of b , so maximizing $\tilde{e}$ with respect to b is equivalent to maximizing

$$\tilde{e} = E_{\tilde{Z}, C} \left\{ \frac{GBB^T G^T - 2GB\delta_3 + \delta_3^2}{\tilde{g}_b - B^T FB} \right\}.$$

We have

$$\begin{aligned}
\delta_1^2 &= (\psi_0 + \psi_1 g_a)^2 &= \psi_0^2 + 2\psi_0\psi_1 g_a + \psi_1^2 g_a^2 \\
\delta_2^T \delta_2 &= (\bar{C}\psi_0 + \psi_1 C_a)^T (\bar{C}\psi_0 + \psi_1 C_a) &= \psi_0^2 \bar{C}^T \bar{C} + 2\psi_0\psi_1 \bar{C}^T C_a + \psi_1^2 C_a^T C_a \\
\delta_3^2 &= (\psi_0 g_b + \psi_1 \tilde{g})^2 &= \psi_0^2 g_b^2 + 2\psi_0\psi_1 g_b \tilde{g} + \psi_1^2 \tilde{g}^2 \\
\delta_1 \delta_2 &= (\psi_0 + \psi_1 g_a)(\bar{C}\psi_0 + \psi_1 C_a) &= \psi_0^2 \bar{C} + \psi_1^2 g_a C_a + \psi_0\psi_1 (C_a + g_a \bar{C}) \\
\delta_1 \delta_3 &= (\psi_0 + \psi_1 g_a)(\psi_0 g_b + \psi_1 \tilde{g}) &= \psi_0^2 g_b + \psi_1^2 g_a \tilde{g} + \psi_0\psi_1 (g_a g_b + \tilde{g}) \\
\delta_2 \delta_3 &= (\bar{C}\psi_0 + \psi_1 C_a)(\psi_0 g_b + \psi_1 \tilde{g}) &= \psi_0^2 \bar{C} g_b + \psi_1^2 C_a \tilde{g} + \psi_0\psi_1 (C_a g_b + \bar{C} \tilde{g}) \\
GB &= \begin{pmatrix} g_1 & g_{2(1 \times p)} \end{pmatrix} \begin{pmatrix} g_b \\ C_b \end{pmatrix} &= g_1 g_b + g_{2(1 \times p)} C_b^T.
\end{aligned}$$

So, we can write

$$\begin{aligned}
 \tilde{e} &= E_{\tilde{Z}, C} \left\{ \frac{GBB^T G^T - 2GB\delta_3 + \delta_3^2}{\tilde{g}_b - B^T F B} \right\} \\
 &= E_{\tilde{Z}, C} \left\{ \frac{(g_1 g_b + g_2 C_b)^2 - 2(g_1 g_b + g_2 C_b)(\psi_0 g_b + \psi_1 \tilde{g}) + \psi_0^2 g_b^2 + 2\psi_0 \psi_1 g_b \tilde{g} + \psi_1^2 \tilde{g}^2}{\tilde{g}_b - (g_b \ C_b^T) \begin{pmatrix} f_{11} & f_{12} \\ f_{12}^T & f_{22} \end{pmatrix} \begin{pmatrix} g_b \\ C_b \end{pmatrix}} \right\}.
 \end{aligned} \tag{A1}$$

We can rewrite the components of the block matrix in F as

$$\begin{aligned}
 f_{11} &= 1 + \bar{C}^T f_{22} \bar{C}, \\
 f_{12} &= -\bar{C}^T f_{22}, \\
 f_{12}^T &= -f_{22} \bar{C}, \\
 \text{and } f_{22} &= K^{-1},
 \end{aligned}$$

which does not involve b . It follows that the components of matrix G can be written as

$$\begin{aligned}
G_{1 \times (1+p)} &= \begin{bmatrix} g_{1(1 \times 1)} & g_{2(1 \times p)} \end{bmatrix} = (\psi_0 + \psi_1 g_a \bar{\psi}_0 \bar{C}^T + \psi_1 C_a^T) \begin{pmatrix} f_{11} & f_{12} \\ f_{12}^T & f_{22} \end{pmatrix}, \\
g_1 &= (\psi_0 + \psi_1 g_a) f_{11} + (\psi_0 \bar{C}^T + \psi_1 C_a^T) f_{12}^T \\
&= (\psi_0 + \psi_1 g_a) (1 + \bar{C}^T f_{22} \bar{C}) - (\psi_0 \bar{C}^T + \psi_1 C_a^T) f_{22} \bar{C} \\
&= \psi_0 + \psi_1 g_a + \psi_0 \bar{C}^T f_{22} \bar{C} + \psi_1 g_a \bar{C}^T f_{22} \bar{C} - \psi_0 \bar{C}^T f_{22} \bar{C} - \psi_1 C_a^T f_{22} \bar{C} \\
&= \psi_0 + \psi_1 g_a + \psi_1 g_a \bar{C}^T f_{22} \bar{C} - \psi_1 C_a^T f_{22} \bar{C} \\
&= \psi_0 + \psi_1 [g_a + g_a \bar{C}^T f_{22} \bar{C} - C_a^T f_{22} \bar{C}], \\
\text{and } g_2 &= (\psi_0 + \psi_1 g_a) f_{12} + (\psi_0 \bar{C}^T + \psi_1 C_a^T) f_{22} \\
&= \psi_0 f_{12} + \psi_1 g_a f_{12} + \psi_0 \bar{C}^T f_{22} + \psi_1 C_a^T f_{22} \\
&= -\psi_0 \bar{C}^T f_{22} - \psi_1 g_a \bar{C}^T f_{22} + \psi_0 \bar{C}^T f_{22} + \psi_1 C_a^T f_{22} \\
&= -\psi_0 \bar{C}^T f_{22} - \psi_1 g_a \bar{C}^T f_{22} + \psi_0 \bar{C}^T f_{22} + \psi_1 C_a^T f_{22} \\
&= -\psi_1 g_a \bar{C}^T f_{22} + \psi_1 C_a^T f_{22} \\
&= \psi_1 [C_a^T - g_a \bar{C}^T] f_{22}.
\end{aligned}$$

We also have

$$\begin{aligned}
g_1^2 &= \left\{ \psi_0 + \psi_1 [g_a + g_a \bar{C}^T f_{22} \bar{C} - C_a^T f_{22} \bar{C}] \right\}^2 \\
&= \psi_0^2 + \psi_1^2 [g_a + g_a \bar{C}^T f_{22} \bar{C} - C_a^T f_{22} \bar{C}]^2 + 2\psi_0 \psi_1 [g_a + g_a \bar{C}^T f_{22} \bar{C} - C_a^T f_{22} \bar{C}], \\
g_2 C_b &= \psi_1 [C_a^T - g_a \bar{C}^T] f_{22} C_b, \\
(g_2 C_b)^2 &= \psi_1^2 [C_a^T - g_a \bar{C}^T] f_{22} C_b^2, \\
\text{and } g_1 (g_2 C_b) &= \left\{ \psi_0 + \psi_1 [g_a + g_a \bar{C}^T f_{22} \bar{C} - C_a^T f_{22} \bar{C}] \right\} \psi_1 [C_a^T - g_a \bar{C}^T] f_{22} C_b \\
&= \psi_0 \psi_1 [C_a^T - g_a \bar{C}^T] f_{22} C_b + \psi_1^2 [g_a + g_a \bar{C}^T f_{22} \bar{C} - C_a^T f_{22} \bar{C}] [C_a^T - g_a \bar{C}^T] f_{22} C_b.
\end{aligned}$$

We can rewrite the denominator k as

$$\begin{aligned}
k &= \bar{g}_b - \begin{pmatrix} g_b & C_b^T \end{pmatrix} \begin{pmatrix} f_{11} & f_{12} \\ f_{12}^T & f_{22} \end{pmatrix} \begin{pmatrix} g_b \\ C_b \end{pmatrix} \\
&= \bar{g}_b - \begin{pmatrix} g_b & C_b^T \end{pmatrix} \begin{pmatrix} f_{11}g_b + f_{12}C_b \\ f_{12}^Tg_b + f_{22}C_b \end{pmatrix} \\
&= \bar{g}_b - g_b(f_{11}g_b + f_{12}C_b) - C_b^T(f_{12}^Tg_b + f_{22}C_b) \\
&= \bar{g}_b - g_b^2f_{11} - g_bf_{12}C_b - C_b^Tf_{12}^Tg_b - C_b^Tf_{22}C_b \\
&= \bar{g}_b - g_b(f_{11}g_b + f_{12}C_b) - C_b^T(f_{12}^Tg_b + f_{22}C_b) \\
&= \bar{g}_b - g_b^2f_{11} - 2g_bf_{12}C_b - C_b^Tf_{22}C_b \\
&= \bar{g}_b - g_b^2 - g_b^2\bar{C}^Tf_{22}\bar{C} + 2g_b\bar{C}^Tf_{22}C_b - C_b^Tf_{22}C_b \\
&= \bar{g}_b - g_b^2 - (g_b\bar{C} - C_b)^Tf_{22}(g_b\bar{C} - C_b).
\end{aligned}$$

Focusing on the numerator inside the expectation in (3), we have

$$\begin{aligned}
\text{numerator} &= (g_1g_b + g_2C_b)^2 - 2(g_1g_b + g_2C_b)(\psi_0g_b + \psi_1\bar{g}) + \psi_0^2g_b^2 + 2\psi_0\psi_1g_b\bar{g} + \psi_1^2\bar{g}^2 \\
&= g_1^2g_b^2 + (g_2C_b)^2 + 2g_1g_b(g_2C_b) - 2[g_1g_b^2\psi_0 + g_1g_b\bar{g}\psi_1 + g_b(g_2C_b)\psi_0 + \bar{g}(g_2C_b)\psi_1] + \psi_0^2g_b^2 + 2\psi_0\psi_1g_b\bar{g} + \psi_1^2\bar{g}^2 \\
&= g_1^2g_b^2 + (g_2C_b)^2 + 2g_1g_b(g_2C_b) - 2[g_1g_b^2\psi_0 + g_1g_b\bar{g}\psi_1 + g_b(g_2C_b)\psi_0 + \bar{g}(g_2C_b)\psi_1] + \psi_0^2g_b^2 + 2\psi_0\psi_1g_b\bar{g} + \psi_1^2\bar{g}^2 \\
&= \psi_1^2\{[g_ag_b - \bar{g}]^2 + g_a^2g_b^2(\bar{C}^Tf_{22}\bar{C})^2 + g_b^2(C_a^Tf_{22}\bar{C})^2 - 2g_ag_b^2\bar{C}^Tf_{22}\bar{C}C_a^Tf_{22}\bar{C} + 2g_a^2g_b^2\bar{C}^Tf_{22}\bar{C} \\
&\quad - 2g_ag_b^2C_a^Tf_{22}\bar{C} + (C_a^Tf_{22}C_b)^2 + g_a^2(\bar{C}^Tf_{22}C_b)^2 - 2g_aC_a^Tf_{22}C_b\bar{C}^Tf_{22}C_b + 2g_bg_aC_a^Tf_{22}C_b \\
&\quad - 2g_bg_ag_a\bar{C}^Tf_{22}C_b + 2g_bg_a\bar{C}^Tf_{22}\bar{C}C_a^Tf_{22}C_b - 2g_bg_a^2\bar{C}^Tf_{22}\bar{C}\bar{C}^Tf_{22}C_b - 2g_bC_a^Tf_{22}\bar{C}C_a^Tf_{22}C_b \\
&\quad + 2g_bg_aC_a^Tf_{22}\bar{C}\bar{C}^Tf_{22}C_b - 2g_ag_b\bar{g}\bar{C}^Tf_{22}\bar{C} + 2g_b\bar{g}C_a^Tf_{22}\bar{C} - 2\bar{g}C_a^Tf_{22}C_b + 2\bar{g}g_a\bar{C}^Tf_{22}C_b\}
\end{aligned}$$

Note that ψ_1^2 is positive and does not depend on b , so we only need to maximize the other terms. We can show that

$$\begin{aligned}
& (g_a g_b - \tilde{g})(C_b - \bar{C}_{g_b})^T f_{22}(C_a - \bar{C}_{g_a}) = (g_a g_b - \tilde{g})(C_b^T f_{22} - \bar{C}^T g_b f_{22})(C_a - \bar{C}_{g_a}) \\
& = (g_a g_b - \tilde{g})(C_b^T f_{22} C_a - g_a C_b^T f_{22} \bar{C} - g_b \bar{C}^T f_{22} C_a + g_a g_b \bar{C}^T f_{22} \bar{C}) \\
& = g_a g_b C_b^T f_{22} C_a - g_a^2 g_b C_b^T f_{22} \bar{C} - g_a g_b^2 \bar{C}^T f_{22} C_a + g_a^2 g_b^2 \bar{C}^T f_{22} \bar{C} - \tilde{g} C_b^T f_{22} C_a + \tilde{g} g_a C_b^T f_{22} \bar{C} + \tilde{g} g_b \bar{C}^T f_{22} C_a \\
& \quad - \tilde{g} g_a g_b \bar{C}^T f_{22} \bar{C}, \\
& \text{and } [(C_b - \bar{C}_{g_b})^T f_{22}(C_a - \bar{C}_{g_a})]^2 = (C_b^T f_{22} C_a - g_a C_b^T f_{22} \bar{C} - g_b \bar{C}^T f_{22} C_a + g_a g_b \bar{C}^T f_{22} \bar{C})^2 \\
& = (C_b^T f_{22} C_a)^2 + g_a^2 (C_b^T f_{22} \bar{C})^2 + g_b^2 (\bar{C}^T f_{22} C_a)^2 + g_a^2 g_b^2 (\bar{C}^T f_{22} \bar{C})^2 - 2g_a C_b^T f_{22} C_a C_b^T f_{22} \bar{C} \\
& \quad - 2g_b C_b^T f_{22} C_a \bar{C}^T f_{22} C_a + 2g_a g_b C_b^T f_{22} C_a \bar{C}^T f_{22} \bar{C} + 2g_a g_b C_b^T f_{22} \bar{C} \bar{C}^T f_{22} C_a - 2g_a^2 g_b C_b^T f_{22} \bar{C} \bar{C}^T f_{22} \bar{C} \\
& \quad - 2g_a g_b^2 \bar{C}^T f_{22} C_a \bar{C}^T f_{22} \bar{C}.
\end{aligned}$$

Thus, the numerator inside the expectation (3) above reduces to

$$\text{numerator} = [g_a g_b - \tilde{g}]^2 + [(C_b - \bar{C}_{g_b})^T f_{22}(C_a - \bar{C}_{g_a})]^2 + 2(g_a g_b - \tilde{g})(C_b - \bar{C}_{g_b})^T f_{22}(C_a - \bar{C}_{g_a}).$$

Now, we can again look at $\tilde{e}$ as

$$\tilde{e} = E_{\tilde{Z}, C} \left[\frac{[g_a g_b - \tilde{g}]^2 + [(C_b - \bar{C}_{g_b})^T f_{22}(C_a - \bar{C}_{g_a})]^2 + 2(g_a g_b - \tilde{g})(C_b - \bar{C}_{g_b})^T f_{22}(C_a - \bar{C}_{g_a})}{\tilde{g}_b - g_b^2 - (g_b \bar{C} - C_b)^T f_{22}(g_b \bar{C} - C_b)} \right].$$

We can define the following, which are empirical variances, covariances, and variance-covariance matrices of the exposure, confounders, and outcome data:

$$\begin{aligned}
\overline{\text{Var}}_b &= \tilde{g}_b - g_b^2, \\
\overline{\text{Cov}}_{ab} &= \tilde{g} - g_a g_b, \\
\overline{\text{Cov}}_{Cb} &= (C_b - \bar{C}_{g_b})^T, \\
\overline{\text{Cov}}_{Ca} &= (C_a - \bar{C}_{g_a})^T, \\
\text{and } \tilde{\Sigma}_C &= C^2 - \bar{C} \bar{C}^T.
\end{aligned}$$

This allows us to rewrite $\tilde{e}$ as

$$\begin{aligned}
\tilde{e} &= E_{\tilde{Z}, C} \left[\frac{\overline{\text{Cov}}_{ab}^2 + [\overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca}]^2 + 2\overline{\text{Cov}}_{ab} \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca}}{\overline{\text{Var}}_b - \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Cb}} \right] \\
&= E_{\tilde{Z}, C} \left[\frac{(\overline{\text{Cov}}_{ab} + \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca})^2}{\overline{\text{Var}}_b - \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Cb}} \right].
\end{aligned}$$

Define $e^* = \tilde{e} + E_{\tilde{Z}, C}[\overline{\text{Cov}}_{Ca}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca}]$, where $e^* - \tilde{e}$ does not depend on b , $\mathbf{a}_1$, and $\mathbf{a}_2$. We have

$$\begin{aligned} e^* &= E_{\tilde{Z}, C} \left[\frac{(\overline{\text{Cov}}_{ab} + \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca})^2}{\overline{\text{Var}}_b - \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Cb}} + \overline{\text{Cov}}_{Ca}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca} \right] \\ &= E_{\tilde{Z}, C} \left[\frac{\overline{\text{Cov}}_{ab}^2 + 2\overline{\text{Cov}}_{ab} \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca} + \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca} \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca} + \overline{\text{Cov}}_{Ca}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca}}{\overline{\text{Var}}_b - \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Cb}} + \overline{\text{Cov}}_{Ca}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca} \right] \\ &= E_{\tilde{Z}, C} \left[\frac{\overline{\text{Cov}}_{ab}^2 + 2\overline{\text{Cov}}_{ab} \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca} + \overline{\text{Var}}_b \overline{\text{Cov}}_{Ca}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca}}{\overline{\text{Var}}_b - \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Cb}} \right]. \end{aligned}$$

A.6 | $e(b, \mathbf{a}_0, \mathbf{a}_1, \mathbf{a}_2)$ is minimized when $\tilde{\rho}_{ab} = 1$

We can see simply that, in the absence of confounders (ie, $p = 0$), e^* reduces to $E_{\tilde{Z}, C} \left[\frac{\overline{\text{Cov}}_{ab}^2}{\overline{\text{Var}}_b} \right]$,

which is maximized at $b = a$ if the expected empirical correlation between $h_i^*(a, t_i)$ and $h_i^*(a, t_i)$ is 1. Now, we will show that the same holds when $p > 0$.

If we denote $\overline{\text{Cov}}_{C(j)b}$ as the j th element of $\overline{\text{Cov}}_{Cb}$ and $\rho_{C(j)b} = \frac{\overline{\text{Cov}}_{C(j)b}}{\sqrt{\sum_{C(j,j)} \overline{\text{Var}}_b}}$, where $\tilde{\Sigma}_{C(j,j)}$ is the (j, j) th element of $\tilde{\Sigma}_C$. $\overline{\text{Cov}}_{Cb}$ and $\overline{\text{Cov}}_{Ca}$ can be written as

$$\begin{aligned} \overline{\text{Cov}}_{Cb} &= [\overline{\text{Cov}}_{C(1)b}, \dots, \overline{\text{Cov}}_{C(p)b}]^T \\ &= [\tilde{\rho}_{C(1)b} \sqrt{\tilde{\Sigma}_{C(1,1)} \overline{\text{Var}}_b}, \dots, \tilde{\rho}_{C(p)b} \sqrt{\tilde{\Sigma}_{C(p,p)} \overline{\text{Var}}_b}]^T \\ &= \sqrt{\overline{\text{Var}}_b} [\tilde{\rho}_{C(1)b} \sqrt{\tilde{\Sigma}_{C(1,1)}}, \dots, \tilde{\rho}_{C(p)b} \sqrt{\tilde{\Sigma}_{C(p,p)}}]^T, \\ \text{and } \overline{\text{Cov}}_{Ca} &= \sqrt{\overline{\text{Var}}_a} [\tilde{\rho}_{C(1)a} \sqrt{\tilde{\Sigma}_{C(1,1)}}, \dots, \tilde{\rho}_{C(p)a} \sqrt{\tilde{\Sigma}_{C(p,p)}}]^T. \end{aligned}$$

Define

$$\begin{aligned} \tilde{\rho}_{ab} &= \frac{\overline{\text{Cov}}_{ab}}{\sqrt{\overline{\text{Var}}_a \overline{\text{Var}}_b}}, \\ \tilde{\rho}_a &= \frac{\overline{\text{Cov}}_{Ca}}{\sqrt{\overline{\text{Var}}_a}} = [\tilde{\rho}_{C(1)a} \sqrt{\tilde{\Sigma}_{C(1,1)}}, \dots, \tilde{\rho}_{C(p)a} \sqrt{\tilde{\Sigma}_{C(p,p)}}]^T \\ \text{and } \tilde{\rho}_b &= \frac{\overline{\text{Cov}}_{Cb}}{\sqrt{\overline{\text{Var}}_b}} = [\tilde{\rho}_{C(1)b} \sqrt{\tilde{\Sigma}_{C(1,1)}}, \dots, \tilde{\rho}_{C(p)b} \sqrt{\tilde{\Sigma}_{C(p,p)}}]^T. \end{aligned}$$

This allows us to write

$$\begin{aligned}\overline{\text{Cov}}_{Ca}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca} &= \overline{\text{Var}}_a \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a, \\ \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Cb} &= \overline{\text{Var}}_b \tilde{\rho}_b^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b, \\ \overline{\text{Cov}}_{Ca}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Cb} &= \sqrt{\overline{\text{Var}}_a \overline{\text{Var}}_b} \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b.\end{aligned}$$

Therefore, e^* reduces to

$$\begin{aligned}e^* &= E_{\tilde{Z}, C} \left[\frac{\overline{\text{Cov}}_{ab}^2 + 2\overline{\text{Cov}}_{ab} \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca} + \overline{\text{Var}}_b \overline{\text{Cov}}_{Ca}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Ca}}{\overline{\text{Var}}_b - \overline{\text{Cov}}_{Cb}^T \tilde{\Sigma}_C^{-1} \overline{\text{Cov}}_{Cb}} \right] \\ &= E_{\tilde{Z}, C} \left[\frac{\overline{\text{Cov}}_{ab}^2 + 2\overline{\text{Cov}}_{ab} \sqrt{\overline{\text{Var}}_a \overline{\text{Var}}_b} \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b + \overline{\text{Var}}_b \overline{\text{Var}}_a \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a}{\overline{\text{Var}}_b - \overline{\text{Var}}_b \tilde{\rho}_b^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b} \right] \\ &= E_{\tilde{Z}, C} \left[\frac{\tilde{\rho}_{ab}^2 + \overline{\text{Var}}_a \overline{\text{Var}}_b + 2\tilde{\rho}_{ab} \overline{\text{Var}}_a \overline{\text{Var}}_b \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b + \overline{\text{Var}}_b \overline{\text{Var}}_a \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a}{\overline{\text{Var}}_b - \overline{\text{Var}}_b \tilde{\rho}_b^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b} \right] \\ &= E_{\tilde{Z}, C} \left[\frac{\tilde{\rho}_{ab}^2 + \overline{\text{Var}}_a + 2\tilde{\rho}_{ab} \overline{\text{Var}}_a \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b + \overline{\text{Var}}_a \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a}{1 - \tilde{\rho}_b^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b} \right].\end{aligned}$$

Here, we make the assumption that $\tilde{\rho}_b$ is not a function of b . Note that all quantities in the above expression, aside from $\tilde{\rho}_{ab}$, are therefore considered fixed with respect to b , conditional on $\tilde{Z}$ and C . We now examine how the function inside the expectation in e^* changes across values of $\tilde{\rho}_{ab}$. We have

$$\frac{\partial}{\partial \tilde{\rho}_{ab}} \left[\frac{\tilde{\rho}_{ab}^2 \overline{\text{Var}}_a + 2\tilde{\rho}_{ab} \overline{\text{Var}}_a \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b + \overline{\text{Var}}_a \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a}{1 - \tilde{\rho}_b^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b} \right] = \frac{2\tilde{\rho}_{ab} \overline{\text{Var}}_a + 2\overline{\text{Var}}_a \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b}{1 - \tilde{\rho}_b^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b}.$$

Because $\tilde{\rho}_b$ is not a function of b , we know $\tilde{\rho}_b = \tilde{\rho}_a$. Hence, this derivative simplifies to

$$2\overline{\text{Var}}_a \left[\frac{\tilde{\rho}_{ab} + \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a}{1 - \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a} \right].$$

Note that

$$\begin{aligned}
\tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a &= \left[\tilde{\rho}_{C(1)a} \sqrt{\tilde{\Sigma}_{C(1,1)}} \cdots \tilde{\rho}_{C(p)a} \sqrt{\tilde{\Sigma}_{C(p,p)}} \right] \tilde{\Sigma}_C^{-1} \left[\tilde{\rho}_{C(1)a} \sqrt{\tilde{\Sigma}_{C(1,1)}} \cdots \tilde{\rho}_{C(p)a} \sqrt{\tilde{\Sigma}_{C(p,p)}} \right]^T \\
&= \left[\tilde{\rho}_{C(1)a} \cdots \tilde{\rho}_{C(p)a} \right] \begin{bmatrix} \sqrt{\tilde{\Sigma}_{C(1,1)}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\tilde{\Sigma}_{C(p,p)}} \end{bmatrix} \tilde{\Sigma}_C^{-1} \begin{bmatrix} \sqrt{\tilde{\Sigma}_{C(1,1)}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\tilde{\Sigma}_{C(p,p)}} \end{bmatrix} \begin{bmatrix} \tilde{\rho}_{C(1)a} \\ \vdots \\ \tilde{\rho}_{C(p)a} \end{bmatrix} \\
&= \left[\tilde{\rho}_{C(1)a} \cdots \tilde{\rho}_{C(p)a} \right] \left(\begin{bmatrix} 1 & \cdots & 0 \\ \sqrt{\tilde{\Sigma}_{C(1,1)}} & & \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 & \cdots & 0 \\ \sqrt{\tilde{\Sigma}_{C(1,1)}} & & \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} \tilde{\rho}_{C(1)a} \\ \vdots \\ \tilde{\rho}_{C(p)a} \end{bmatrix} \\
&= \left[\tilde{\rho}_{C(1)a} \cdots \tilde{\rho}_{C(p)a} \right] \tilde{\Sigma}_C^{-1} \begin{bmatrix} \tilde{\rho}_{C(1)a} \\ \vdots \\ \tilde{\rho}_{C(p)a} \end{bmatrix},
\end{aligned}$$

where Σ_{CC} is the observed correlation matrix for C_i , $i = 1, \dots, n$. Therefore, $\tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a$ is equal to the coefficient of multiple determination,²⁶ which is guaranteed to be $\in [0, 1]$. For $\tilde{\rho}_{ab} \in [0, 1]$, we see that

$$2\sqrt{\text{Var}}_a \left[\frac{\tilde{\rho}_{ab} + \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a}{1 - \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a} \right] > 0.$$

Therefore, $\frac{\tilde{\rho}_{ab}^2 \sqrt{\text{Var}}_a + 2\tilde{\rho}_{ab} \sqrt{\text{Var}}_a \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b + \text{Var}(g_a) \tilde{\rho}_a^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_a}{1 - \tilde{\rho}_b^T \tilde{\Sigma}_C^{-1} \tilde{\rho}_b}$ is maximized when $\tilde{\rho}_{ab} = 1$.

REFERENCES

1. Wang M, Liao X, Laden F, Spiegelman D. Quantifying risk over the life course - latency, age-related susceptibility, and other time-varying exposure metrics. *Statist Med*. 2016;35:2283–2295.
2. Caplan RJ, Marsh GM, Enterline PE. A generalized effective exposure modeling program for assessing dose-response in epidemiologic investigations. *Comput Biomed Res*. 1983;16(6):587–596. [PubMed: 6689145]
3. Salvan A, Stayner L, Steenland K, Smith R. Selecting an exposure lag period. *Epidemiology*. 1995;6(4):387–390. [PubMed: 7548346]
4. Hauptmann M, Lubin JH, Rosenberg P, Wellmann J, Kreienbrock L. The use of sliding time windows for the exploratory analysis of temporal effects of smoking histories on lung cancer risk. *Statist Med*. 2000;19(16):2185–194.
5. Hauptmann M, Berhane K, Langholz B, Lubin JH. Using splines to analyse latency in the Colorado Plateau uranium miners cohort. *J Epidemiol Biostat*. 2001;6(6):417–424. [PubMed: 11831677]

6. Richardson DB, Cole SR, Chu H, Langholz B. Lagging exposure information in cumulative exposure-response analyses. *Am J Epidemiol.* 2011;174(12):1416–1422. [PubMed: 22047823]
7. Heaton MJ, Peng RD. Extending distributed lag models to higher degrees. *Biostatistics.* 2014;15:398–412. [PubMed: 23990524]
8. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statist Med.* 1989;8(9):1051–1069.
9. 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;132:734–745. [PubMed: 2403114]
10. 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;136:1400–1413. [PubMed: 1488967]
11. Carroll RJ, Wand MP. Semiparametric estimation in logistic measurement error models. *J R Stat Soc Ser B Stat Methodol.* 1991;53:573–585.
12. Kuha J. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Statist Med.* 1994;13:1135–1148.
13. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr.* 1997;65:1179S–1186S. [PubMed: 9094918]
14. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective.* Boca Raton, FL: Chapman & Hall/CRC; 2006.
15. Puett RC, Hart JE, Yanosky JD, et al. Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses' Health Study. *Environ Health Perspect.* 2009;117:1697–1701. [PubMed: 20049120]
16. Laden F, Schwartz J, Speizer FE, Dockery DW. Reduction in fine particulate air pollution and mortality: extended follow-up of the Harvard six cities study. *Am J Respir Crit Care Med.* 2006;173:667–672. [PubMed: 16424447]
17. Hart JE, Liao X, Hong B, et al. The association of long-term exposure to PM_{2.5} on all-cause mortality in the Nurses' Health Study and the impact of measurement-error correction. *Environ Health.* 2015;14:38. [PubMed: 25926123]
18. Puett RC, Hart JE, Yanosky JD, et al. Particulate matter air pollution exposure, distance to road, and incident lung cancer in the Nurses' Health Study cohort. *Environ Health Perspect.* 2014;122(9):926–932. [PubMed: 24911062]
19. Loève M *Probability Theory 1.* 4th ed New York, NY: Springer-Verlag; 1977.
20. Jennrich RI. Asymptotic properties of non-linear least squares estimators. *Ann Math Stat.* 1969;40(2):633–643.
21. Hauptmann M, Pohlabeln H, Lubin JH, et al. The exposure-time-response relationship between occupational asbestos exposure and lung cancer in two German case-control studies. *Am J Ind Med American.* 2002;41(2):89–97.
22. Liao X, Zucker DM, Li Y, Spiegelman D. Survival analysis with error-prone time-varying covariates: a risk set calibration approach. *Biometrics.* 2010;67(1):50–58.
23. Grobbee DE, Rimm EB, Giovannucci E, Colditz G, Stampfer M, Willett W. Coffee, caffeine, and cardiovascular disease in men. *N Engl J Med.* 1990;323(15):1026–1032. [PubMed: 2215561]
24. Heaphy CM, Yoon GS, Pescoe SB, et al. Prostate cancer cell telomere length variability and stromal cell telomere length as prognostic markers for metastasis and death. *Cancer Discov.* 2013;3(10):1130–1041. [PubMed: 23779129]
25. Jardim TV, Sousa ALL, Pova TIR, et al. The natural history of cardiovascular risk factors in health professionals: 20-year follow-up. *BMC Public Health.* 2015;15(1):1111. [PubMed: 26558824]
26. Kutner MH, Nachtsheim CJ, Neter J, Li W. *Applied Linear Statistical Models.* New York, NY: McGraw-Hill; 2005.

27. Horick N, Weller E, Milton DK, Gold DR, Li R, Spiegelman D. Home endotoxin exposure and wheeze in infants: correction for bias due to exposure measurement error. *Environ Health Perspect.* 2006;114(1):135–140. [PubMed: 16393671]
28. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statist Med.* 1989;8(9): 1051–1069.

FIGURE 1.
Mean squared error (MSE) for each possible latency parameter value

TABLE 1

Exposure metrics with a single latency parameter

Metric	$h_i(a,t)$			
	Continuous		Discrete	
	Average	Total	Average	Total
Recent moving	$\frac{\int_{t-a}^t X_i(s)ds}{a}$	$\int_{t-a}^t X_i(s)ds$	$\frac{\sum_{s=t-a}^t X_i(s)}{a+1}$	$\sum_{s=t-a}^t X_i(s)$
Mid-life or later-life-related susceptibility window	$\frac{\int_a^t X_i(s)ds}{t-a}$	$\int_a^t X_i(s)ds$	$\frac{\sum_{s=a}^t X_i(s)}{t-a+1}$	$\sum_{s=a}^t X_i(s)$
Exposure during critical period of susceptibility	$\frac{\int_{t_0}^a X_i(s)ds}{a-t_0}$	$\int_{t_0}^a X_i(s)ds$	$\frac{\sum_{s=t_0}^a X_i(s)}{a-t_0+1}$	$\sum_{s=t_0}^a X_i(s)$
Age-related or time-related moving exposure with a lag	$\frac{\int_{t_0}^{t-a} X_i(s)ds}{t-a-t_0}$	$\int_{t_0}^{t-a} X_i(s)ds$	$\frac{\sum_{s=t_0}^{t-a} X_i(s)}{t-a-t_0+1}$	$\sum_{s=t_0}^{t-a} X_i(s)$
One-time exposure effect			$X_i(a)$	
A lag-type model			$X_i(t-a)$	

TABLE 2

Characteristics of 16 371 men in the HPFS

Race, %	
White	92.4
Black	0.4
Asian	1.3
Other	5.9
BMI in 2006, mean (sd)	25.9 (3.6)
Age in 1986, mean (sd)	52.0 (8.5)
Physical activity in 1986 in MET-hrs/wk, mean (sd)	22.0 (29.5)
Physical activity in 1996 in MET-hrs/wk, mean (sd)	37.6 (39.9)
Physical activity in 2006 in MET-hrs/wk, mean (sd)	43.9 (45.5)

Abbreviations: BMI, body mass index; HPFS, Health Professionals Follow-Up Study.

TABLE 3

Mean % bias in latency estimation based on 1000 simulation replicates for each setting

μ	σ^2	ρ	Mean % Bias					
			$N = 50$	$N = 1K$	$N = 10K$	$N = 40K$	$N = 100K$	$N = 1MIL$
No error	–	–	–17	15	14	5	2	0
		0.2	–13	10	16	6	2	0
		0.4	–16	16	12	5	2	0
	100	0.2	–16	11	13	5	2	0
		0.4	–15	9	7	3	2	0
		0.2	–21	6	8	6	6	1
	1000	0.4	–21	–8	–7	–3	–2	0
		0.2	–25	–8	8	8	7	3
		0.4	–28	–11	–11	–8	–7	0
	2000	0.2	–13	10	16	6	2	0
		0.4	–16	16	12	5	2	0
		0.2	–16	11	13	5	2	0
t	10	0.4	–15	9	7	3	2	0
		0.2	–21	6	8	6	6	1
		0.4	–21	–8	–7	–3	–2	0
	1000	0.2	–25	8	8	8	7	3
		0.4	–28	–11	–11	–8	–7	–1
		0.2	–18	13	12	5	2	0
	100	0.4	–18	13	11	5	2	0
		0.2	–15	12	12	5	3	0
		0.4	–16	10	8	3	2	0
	1000	0.2	–15	5	9	7	5	1
		0.4	–16	–9	–7	–4	–1	0
		0.2	–18	6	9	7	7	3
$\sin(t/\pi)$	2000	0.4	–24	–13	–11	–8	–6	0