Correcting for Biases Arising in Epidemiologic Research

Correcting for Biases Arising in Epidemiologic Research

A DISSERTATION PRESENTED
BY
SARAH BANCROFT PESKOE
TO
THE DEPARTMENT OF BIOSTATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOSTATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
NOVEMBER 2017

Dissertation advisor: Professor Donna Spiegelman Sarah Bancroft Peskoe

Correcting for Biases Arising in Epidemiologic Research

Abstract

In chapter 1, we explore the performance of naive least squares estimators for latency parameters in linear models in the presence of measurement error. We prove that in many scenarios under a 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

In chapter 2, we extend the results obtained in chapter 1 to the survival setting when the exposure of interest is a time-varying recent-moving cumulative average. We show that when the disease outcome is rare, the latency parameter for a surrogate exposure is approximately the same as the latency parameter for the corresponding true exposure. We show these results in a series of simulations and illustrate the findings in a study of air pollution and incidence lung cancer in the Nurses Health Study.

In chapter 3, we specificy a statistical framework for estimation and inference based on inverse probability weighting (IPW) to adjust for selection bias in EHR-based research that allows for a hierarchy of missingness mechanisms to better align with the complex nature of electronic health record (EHR) data. We show that this estimator is consistent and asymptotically Normal, and we derive the form of the asymptotic variance. We use simulations to highlight the potential for bias in EHR studies when standard approaches are used to account for selection bias. We use this approach to adjust for selection in an on-going, multi-site EHR-based study of bariatric surgery on BMI.

Contents

Acknowledgments

Thanks to my committee, my friends, and my family for your love, your inspiration, and your unwavering support.

1

There is No Impact of Exposure Measurement Error on Latency Estimation in Linear Models

Sarah Bancroft Peskoe

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Donna Spiegelman

Departments of Epidemiology, Biostatistics, Nutrition and Global Health

Harvard TH Chan School of Public Health

Molin Wang

Departments of Epidemiology and Biostatistics

Harvard TH Chan School of Public Health

1.1 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 is the time window leading up to the occurrence of an outcome during which time the exposure predicts the outcome. Outside the latency period, the exposure has no effect. Latency periods have been discussed in many different frameworks. Caplan et al[1] provided a general formula for weighting exposure metrics in assessing dose-response associations in epidemiological studies. Salvan et al[2] proposed methodology for selecting a lag time, after which the exposure has no effect, in models for relative risk. Hauptmann et al[3] proposed the use of sliding time windows, which fits a series of risk models which 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[4,5,6]. Wang et al[7] 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[8,9,10,11,12,13,14] 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).

1.2 Methods

1.2.1 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[7]. In Table 1.1, 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.

Table 1.1. Exposure metrics with a single latency parameter

	$h_i(a,t)$			
	Continuous		Discrete	
Metric	Average	Total	Average	Total
Recent Moving	$\frac{\int_{t-a}^{t} X_i(s)ds}{a}$	$\int_{t-a}^{t} X_i(s)ds$	$\frac{\sum_{s=t-a}^{t} X_i(s)}{a+1}$	$\sum_{s=t-a}^{t} X_i(s)$
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 period of susceptibility	$\frac{\int_{t_0}^{a} X_i(s)ds}{a-t_0}$	$\int_{t_0}^{a} X_i(s)ds$	$\frac{\sum_{s=t_0}^{a} X_i(s)}{a-t_0+1}$	$\sum_{s=t_0}^{a} X_i(s)$
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	$X_i(t-a)$			

Though results of this paper can be generalized to all exposure metrics presented in Table 1.1, we focus on the recent moving cumulative average. The recent moving cumulative 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 $s = s_0, s_1, \ldots$.

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 cumulative average is often used in air pollution epidemiology (e.g.,[15,16]) where a defines the beginning of the recent exposure susceptibility period. While this paper focuses on the recent moving cumulative average, any single-parameter latency metric can be used as $h_i(a, t)$.

1.2.2 Least Squares Estimation

We consider a least squares approach for estimating b, α_0, α_1, and 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

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

By the Weak Law of Large Numbers (WLLN)[17], as the sample size (i.e. the number of individuals for a fixed amount of follow-up time) goes to infinity, L converges in probability to $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} E_{Y_i,\tilde{Z}_i,C_i}[l_i]$. Define $e\left(b, \alpha_0, \alpha_1, \alpha_2\right)$ 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 $\tilde{Z} = \left(\tilde{Z}_1, \tilde{Z}_2, ..., \tilde{Z}_n\right)$, $Y = (Y_1, ..., Y_n)$, and $C = (C_1, ..., C_n)$. As shown in Appendix A.2,

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

where $\xrightarrow{P}$ denotes convergence in probability as $n \to \infty$, and $\hat{b}$, $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\alpha}_2$ are the least squares estimators for b, α_0, α_1, α_2, respectively. We show in Appendix A.3 that for any fixed b,

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

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

Define $g_a = \frac{1}{n}\sum_{i=1}^{n} h_i^{\star}(a,t_i)$ and $g_b = \frac{1}{n}\sum_{i=1}^{n} h_i^{\star}(b,t_i)$. Using the results above, we show in Appendix A.4 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 $e\left(b, \alpha_0, \alpha_1, \alpha_2\right)$ is equivalent to the maximum of

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

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

is a column vector of the empirical covariance of C_i, $h_i^\star(a, t_i)$, $\widetilde{Var_b} = \frac{1}{n}\sum_{i=1}^{n} h_i^\star(b, t_i)^2 - \left(\frac{1}{n}\sum_{i=1}^{n} h_i^\star(b, t_i)\right)^2$ is the scalar empirical variance of $h_i^\star(b, t_i)$, and $\widetilde{\Sigma}_C = \frac{1}{n}\sum_{i=1}^{n} C_i {C_i}^T - \frac{1}{n}\sum_{i=1}^{n} C_i \left(\frac{1}{n}\sum_{i=1}^{n} C_i\right)^T$ is the empirical variance-covariance matrix of C_i.

1.2.3 Consistency of the Least Squares Estimator of the Latency Parameter

We can see simply that in the absence of confounders (i.e. $p = 0$), $e^\star$ reduces to $E_{\tilde{Z}}\left[\frac{\widetilde{Cov}^2_{ab}}{\widetilde{Var_b}}\right] = E_{\tilde{Z}}\left[\frac{\widetilde{Var_b}\widetilde{Var_a}\,\widetilde{\ }_{ab}}{\widetilde{Var_b}}\right] = E_{\tilde{Z}}\left[\widetilde{Var_a}\widetilde{\varrho}_{ab}\right]$, where $\widetilde{Var_a} = \frac{1}{n}\sum_{i=1}^{n} h_i^\star(a, t_i)^2 - \left(\frac{1}{n}\sum_{i=1}^{n} h_i^\star(a, t_i)\right)^2$ is the scalar empirical variance of $h_i^\star(a, t_i)$ and is not a function of b, and $\widetilde{\varrho}_{ab} = \frac{\widetilde{Cov}_{ab}}{\sqrt{\widetilde{Var_a}\widetilde{Var_b}}}$ is the empirical correlation between $h_i^\star(a, t_i)$ and $h_i^\star(b, t_i)$. Given that any correlation is bounded between -1 and 1, and that $\widetilde{Var_a} > 0$, the maximum of $e^\star$ is equal to the maximum of $\widetilde{\varrho}_{ab}$, which is 1. Under the assumption that $\widetilde{\varrho}_{ab} = 1$ if and only if $a = b$, this implies that $e^\star$ 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.5 that when the empirical correlation between C_i and $h_i^\star(b, t_i)$ is constant over time (i.e. $\widetilde{Cov}_{Cb}$ is not a function of b), the expression for $e^\star\ (b, \alpha_0, \alpha_1, \alpha_2)$ in (2.3) is again maximized when $\widetilde{\varrho}_{ab} = 1$. Under the same assumption that $\widetilde{\varrho}_{ab} = 1$ if and only if $a = b$, $e^\star\ (b, \alpha_0, \alpha_1, \alpha_2)$ 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 $\varrho < 1$ for discrete time points, and AR(1) covariance structures, where the correlation in Z_i between two time points is of the form $\varrho^{|t_1 - t_2|} < 1$ where $\varrho < 1$ 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 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 $h_i^\star(a, t_i)$ 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 (e.g. 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 cumulative average exposure metric, if $X_i(s) = \gamma_0 + \gamma_1 Z_i(s) + \varepsilon_i(s)$, where γ_0 and γ_1 are known or can be estimated and $\varepsilon_i(s)$ are independent, identically distributed with mean 0 and finite variance, then it follows that

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

where $\tilde{\varepsilon}_i = \frac{1}{a} \int_{t_i - a}^{t_i} \varepsilon_i$ has mean 0 and finite variance. Thus we are able to derive the measurement error model for the recent moving cumulative 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[20], ultimately giving us consistent estimates of the latency parameter and regression coefficients.

1.3 Illustrative Example

We used the results above to examine the relationship between physical activity history and BMI in the Health Professionals Follow-up Study (HPFS), and to evaluate the performance of the least squares estimator for the latency parameter in the presence of exposure measurement error.

1.3.1 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[21]. 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 1.2, 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[22].

Table 1.2. 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)

1.3.2 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 utcome of interest in this analysis was BMI reported in 2006. Men who died prior to 2007, were missing age in 1986, or were missing race information were excluded from the analysis.

1.3.3 Latency Estimation

We fit a linear regression model with BMI in 2006 as the dependent variable and the recent moving cumulative 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 cumulative average that, when included in the linear regression model, minimized the mean squared error (MSE). In Figure 1.1, we see the MSE plotted for each possible value of the latency parameter. We conclude the 'true' latency parameter is 10 years.

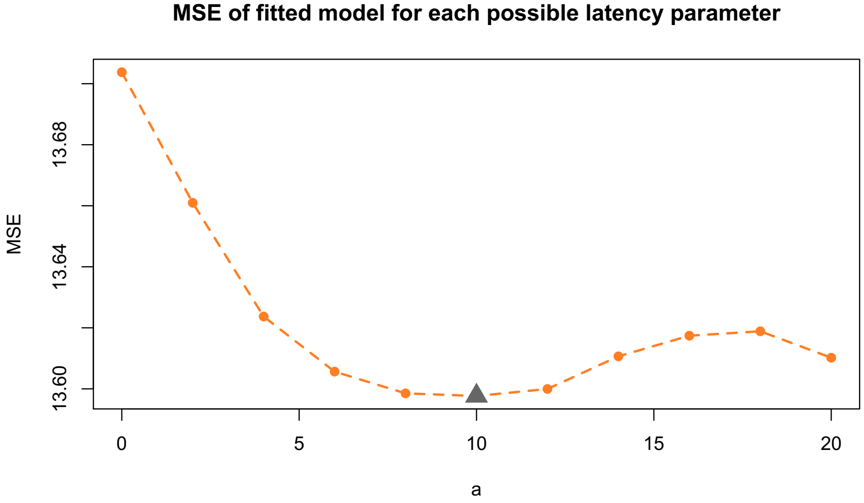


Figure 1.1. MSE for each possibly latency parameter value

1.3.4 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 the 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: $X_{n \times t}$ matrix of physical activity history, where N ranged from 50 to 1 million

Step 2: generate $E_{n \times t}$ error matrix

Step 3: estimate latency parameter using $Z_{n \times t} = X_{n \times t} + E_{n \times t}$

The error matrix, $E_{n \times t}$ was generated by sampling n times from a multivariate normal with mean μ and covariance matrix Σ. These parameters were varied such that $\mu = 0, \ t, \ \text{or} \ \sin(t/\pi)$; the compound symmetric covariance matrix Σ had diag(Σ) =σ^2 and off-diag(Σ) =$\varrho\sigma^2$, where $\sigma = 10, \ 100, \ 1000,$ or 2000 and $\varrho = 0.2$ 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 1.3.

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 (e.g. 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 (e.g.[14]).

Table 1.3. Mean % bias in latency estimation based on 1000 simulation replicates for each setting

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

1.4 Discussion

As seen from the analytic results, the naive least squares estimator for the latency parameter is consistent under the assumption that $Corr\left[h_i^{\star}(a, t_i), h_i^{\star}(b, t_i)\right] = 1$ 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 $h_i^{\star}(b, t_i)$ 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 Health Professionals Followup Study. 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 cumulative 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 $h_i^{\star}(b, t_i)$ 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 $Z(0)$, e.g. at baseline, but less correlated with $Z(20)$, e.g. after 20 years of follow-up. This would imply the correlation between C_i and $h_i^{\star}(b, t_i)$ 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_i and $h_i^\star(b, t_i)$ changes with time.

In future research, we hope to investigate the extent to which these results hold in generalized linear models and survival models. We also intend to define strategies for inference of the latency parameter in the presence of exposure measurement error, and derive equations for confidence intervals and investigate the impact of measurement error on the variance of $\hat{b}$.

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 cumulative average, and existing approaches for bias correction can be implemented to achieve unbiased regression coefficient estimators.

1.5 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

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

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

Let $(\alpha_0, \alpha_1, \alpha_2, b) \equiv \xi_0$. That is, for the sequence of real-valued outcomes, Y_i has the structure

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

where $f_i(\xi_0) = \alpha_0 + \alpha_1 h_i^\star(b, t_i) + \alpha_2^T C_i$ 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 $E\left[q(\xi)\right]|_{\ =\ _0} = 0$ because ξ_0 is the true value. So we can write

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

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

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

If e has a unique global minimum, this implies that *argmin* $\left[e\left(b, \alpha_0, \alpha_1, \alpha_2\right)\right] = \xi_0$, which holds as $n \rightarrow \infty$. Finally, we can use the result shown in Jennrich[18] that under regularity conditions, the least squares estimators for non-linear equations are consistent. These regularity conditions are met for (3), which proves $\hat{b}, \hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2 \xrightarrow{P} \xi_0 = argmin\left[\lim_{n\rightarrow\infty} e\left(b, \alpha_0, \alpha_1, \alpha_2\right)\right]$.

A.3 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:

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

By the surrogacy assumption, we again have

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

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

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

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

A.4 Minimizing $e\left(b,\alpha_0,\alpha_1,\alpha_2\right)$

By the surrogacy assumption, we can rewrite $e = E_{Y_i,\tilde{Z}_i,C_i}[l_i]$ as $e = E_{\tilde{Z}_i,C_i}\left\{E_{\tilde{X}_i|\tilde{Z}_i,C_i}E_{Y_i|\tilde{X}_i,C_i}l_i\right\}$.

Therefore, we can write out $e\left(b,\alpha_0,\alpha_1,\alpha_2\right)$ as

$$\begin{aligned}
&e\left(b,\alpha_0,\alpha_1,\alpha_2\right)\\
&\quad= E_{Y,\tilde{Z},\tilde{C}}\left[\frac{1}{n}\sum_{i=1}^{n} l_i\right]\\
&\quad= \frac{1}{n}\sum_{i=1}^{n} E_{\tilde{Z}_i,C_i}\left\{E_{\tilde{X}_i|\tilde{Z}_i,C_i}E_{Y_i|\tilde{X}_i,C_i}[l_i]\right\}\\
&\quad= \frac{1}{n}\sum_{i=1}^{n} E_{\tilde{Z}_i,C_i}E_{\tilde{X}_i|\tilde{Z}_i,C_i}\left\{\left[\alpha_0+\alpha_1 h_i^\star(b,t_i)+{\alpha_2}^T C_i\right]^2 - 2\left[\beta_0+\beta_1 h_i(a,t_i)+{\beta_2}^T C_i\right]\left[\alpha_0+\alpha_1 h_i^\star(b,t_i)+{\alpha_2}^T C_i\right]\right\}\\
&\quad= \frac{1}{n}\sum_{i=1}^{n} E_{\tilde{Z}_i,C_i}\left\{\left[\begin{pmatrix}\alpha_0 & {\alpha_2}^T & \alpha_1\end{pmatrix}\begin{pmatrix}1\\ C_i\\ h_i^\star(b,t_i)\end{pmatrix}\right]^2 - \right.\\
&\qquad\qquad \left. 2\left[\beta_0+\beta_1\gamma_0+\beta_1\gamma_1 h_i^\star(a)+{\beta_2}^T C_i\right]\begin{pmatrix}\alpha_0 & {\alpha_2}^T & \alpha_1\end{pmatrix}\begin{pmatrix}1\\ C_i\\ h_i^\star(b,t_i)\end{pmatrix}\right\}
\end{aligned}$$

$$
\begin{aligned}
&= E_{\tilde{Z},C}\left\{ \frac{1}{n}\sum_{i=1}^{n}\left[\Delta^T Q^{-1}\begin{bmatrix} 1 & C_i^T & h_i^\star(b,t_i)\end{bmatrix}^T\right]^2 - 2\frac{1}{n}\sum_{i=1}^{n}\Delta^T Q^{-1}\begin{pmatrix} 1 \\ C_i \\ h_i^\star(b,t_i)\end{pmatrix}\left[\beta_0+\beta_1\gamma_0+\beta_1\gamma_1 h_i^\star(a)+\beta_2{}^T C_i\right]\right\} \\
&= E_{\tilde{Z},C}\left\{ \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} 1 & C_i^T & h_i^\star(b,t_i)\end{bmatrix} Q^{-1}\Delta\Delta^T Q^{-1}\begin{bmatrix} 1 & C_i^T & h_i^\star(b,t_i)\end{bmatrix}^T - 2\Delta^T Q^{-1}\Delta\right\} \\
&= E_{\tilde{Z},C}\left\{ \text{trace}\left(\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} 1 & C_i^T & h_i^\star(b,t_i)\end{bmatrix} Q^{-1}\Delta\Delta^T Q^{-1}\begin{bmatrix} 1 & C_i^T & h_i^\star(b,t_i)\end{bmatrix}^T\right) - 2\Delta^T Q^{-1}\Delta\right\} \\
&= E_{\tilde{Z},C}\left\{ \text{trace}\left(\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} 1 & C_i^T & h_i^\star(b,t_i)\end{bmatrix}^T\begin{bmatrix} 1 & C_i^T & h_i^\star(b,t_i)\end{bmatrix} Q^{-1}\Delta\Delta^T Q^{-1}\right) - 2\Delta^T Q^{-1}\Delta\right\} \\
&= E_{\tilde{Z},C}\left\{ \text{trace}\left(QQ^{-1}\Delta\Delta^T Q^{-1}\right) - 2\Delta^T Q^{-1}\Delta\right\} \\
&= E_{\tilde{Z},C}\left\{ \text{trace}\left(\Delta^T Q^{-1}\Delta\right) - 2\Delta^T Q^{-1}\Delta\right\} \qquad \text{(using the cyclic property of the trace function)} \\
&= E_{\tilde{Z},C}\left\{ -\Delta^T Q^{-1}\Delta\right\}.
\end{aligned}
$$

Define

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

Let $\psi_0 = \beta_0 + \beta_1\gamma_0 + \beta_2{}^T\overline{C}$ and $\psi_1 = \beta_1\gamma_1$. Note that $\Delta = \begin{pmatrix} \psi_0 + \psi_1 g_a \\ \overline{C}\psi_0 + \psi_1 C_a \\ \psi_0 g_b + \psi_1 \tilde{g} \end{pmatrix}$, which we denote as $\begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix}$. Define

$$Q_{(p+2)\times(p+2)} = \begin{pmatrix} 1 & \overline{C}^T & g_b \\ \overline{C} & \overline{C^2} & C_b \\ g_b & C_b^T & \bar{g}_b \end{pmatrix} = \begin{pmatrix} A = \begin{bmatrix} 1 & \overline{C}^T \\ \overline{C} & \overline{C^2} \end{bmatrix} & B = \begin{bmatrix} g_b \\ C_b \end{bmatrix} \\ B^T = \begin{bmatrix} g_b & C_b^T \end{bmatrix} & D = \begin{bmatrix} \bar{g}_b \end{bmatrix} \end{pmatrix},$$

$$\begin{aligned} Q^{-1} &= \begin{pmatrix} A^{-1} + A^{-1}B(D - B^TA^{-1}B)^{-1}B^TA^{-1} & -A^{-1}B(D - B^TA^{-1}B)^{-1} \\ -(D - B^TA^{-1}B)^{-1}B^TA^{-1} & (D - B^TA^{-1}B)^{-1} \end{pmatrix} \\ &= \begin{pmatrix} A^{-1} + \frac{1}{k}A^{-1}BB^TA^{-1} & -\frac{1}{k}A^{-1}B \\ -\frac{1}{k}B^TA^{-1} & \frac{1}{k} \end{pmatrix} \\ &= \frac{1}{k}\begin{pmatrix} kA^{-1} + A^{-1}BB^TA^{-1} & -A^{-1}B \\ -B^TA^{-1} & 1 \end{pmatrix}, \end{aligned}$$

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

and $$A^{-1} = \begin{pmatrix} 1 + 1^{-1}\overline{C}^T(\overline{C^2} - \overline{C}1^{-1}\overline{C}^T)^{-1}\overline{C}1^{-1} & -1^{-1}\overline{C}^T(\overline{C^2} - \overline{C}1^{-1}\overline{C}^T)^{-1} \\ -(\overline{C^2} - \overline{C}1^{-1}\overline{C}^T)^{-1}\overline{C}1^{-1} & (\overline{C^2} - \overline{C}1^{-1}\overline{C}^T)^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} 1 + \overline{C}^T(\overline{C^2} - \overline{CC}^T)^{-1}\overline{C} & -\overline{C}^T(\overline{C^2} - \overline{CC}^T)^{-1} \\ -(\overline{C^2} - \overline{CC}^T)^{-1}\overline{C} & (\overline{C^2} - \overline{CC}^T)^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} 1 + \overline{C}^TK^{-1}\overline{C} & -\overline{C}^TK^{-1} \\ -K^{-1}\overline{C} & K^{-1} \end{pmatrix}, \quad \text{where } K_{p\times p} = \overline{C^2} - \overline{CC}^T.$$

Denote $F = A^{-1} = \begin{pmatrix} f_{11\,(1\times 1)} & f_{12\,(1\times p)} \\ f^T_{12\,(1\times p)} & f_{22\,(p\times p)} \end{pmatrix}$ where $f_{22} = K^{-1}_{p\times p}$. Note that A, and subsequently K and F, do not depend on b or Z. Minimizing $e = E_{\tilde{Z}_i C_i}\left\{-\Delta^T Q^{-1}\Delta\right\}$ is equivalent to maximizing $\tilde{e} = E_{\tilde{Z}_i C_i}\left\{\Delta^T Q^{-1}\Delta\right\}$, and

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

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

$$\tilde{\tilde{e}} = E_{\tilde{Z}_i C_i}\left\{ \frac{GBB^TG^T - 2GB\delta_3 + \delta_3^2}{\bar{g}_b - B^TFB} \right\}.$$

We have

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

So we can write

$$
\begin{aligned}
\tilde{\tilde{e}} &= E_{\tilde{Z},C}\left\{\frac{GBB^TG^T - 2GB\delta_3 + \delta_3^2}{\bar{g}_b - B^TFB}\right\} \\
&= E_{\tilde{Z},C}\left\{\frac{(g_1g_b + g_2C_b)^2 - 2\left(g_1g_b + g_2C_b\right)\left(\psi_0g_b + \psi_1\tilde{g}\right) + \psi_0^2g_b^2 + 2\psi_0\psi_1g_b\tilde{g} + \psi_1^2\tilde{g}^2}{\bar{g}_b - \begin{pmatrix} g_b & C_b^T \end{pmatrix}\begin{pmatrix} f_{11} & f_{12} \\ f_{12}^T & f_{22} \end{pmatrix}\begin{pmatrix} g_b \\ C_b \end{pmatrix}}\right\}.
\end{aligned}
\tag{1.1}
$$

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

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

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

$$G_{1\times(1+p)} = \begin{bmatrix} g_{1(1\times 1)} & g_{2(1\times p)} \end{bmatrix} = \begin{pmatrix} \psi_0 + \psi_1 g_a & \overline{\psi_0 C}^T + \psi_1 C_a^T \end{pmatrix} \begin{pmatrix} f_{11} & f_{12} \\ f_{12}^T & f_{22} \end{pmatrix},$$

$$\begin{aligned} g_1 &= (\psi_0 + \psi_1 g_a) f_{11} + \left(\psi_0 \overline{C}^T + \psi_1 C_a^T\right) f_{12}^T \\ &= (\psi_0 + \psi_1 g_a)\left(1 + \overline{C}^T f_{22} \overline{C}\right) - \left(\psi_0 \overline{C}^T + \psi_1 C_a^T\right) f_{22} \overline{C} \\ &= \psi_0 + \psi_1 g_a + \psi_0 \overline{C}^T f_{22} \overline{C} + \psi_1 g_a \overline{C}^T f_{22} \overline{C} - \psi_0 \overline{C}^T f_{22} \overline{C} - \psi_1 C_a^T f_{22} \overline{C} \\ &= \psi_0 + \psi_1 g_a + \psi_1 g_a \overline{C}^T f_{22} \overline{C} - \psi_1 C_a^T f_{22} \overline{C} \\ &= \psi_0 + \psi_1 \left[g_a + g_a \overline{C}^T f_{22} \overline{C} - C_a^T f_{22} \overline{C}\right], \end{aligned}$$

$$\begin{aligned} \text{and} \quad g_2 &= (\psi_0 + \psi_1 g_a) f_{12} + \left(\psi_0 \overline{C}^T + \psi_1 C_a^T\right) f_{22} \\ &= \psi_0 f_{12} + \psi_1 g_a f_{12} + \psi_0 \overline{C}^T f_{22} + \psi_1 C_a^T f_{22} \\ &= -\psi_0 \overline{C}^T f_{22} - \psi_1 g_a \overline{C}^T f_{22} + \psi_0 \overline{C}^T f_{22} + \psi_1 C_a^T f_{22} \\ &= -\psi_0 \overline{C}^T f_{22} - \psi_1 g_a \overline{C}^T f_{22} + \psi_0 \overline{C}^T f_{22} + \psi_1 C_a^T f_{22} \\ &= -\psi_1 g_a \overline{C}^T f_{22} + \psi_1 C_a^T f_{22} \\ &= \psi_1 \left[C_a^T - g_a \overline{C}^T\right] f_{22}. \end{aligned}$$

We also have

$$\begin{aligned} g_1^2 &= \left\{ \psi_0 + \psi_1 \left[g_a + g_a \overline{C}^T f_{22} \overline{C} - C_a^T f_{22} \overline{C} \right] \right\}^2 \\ &= \psi_0^2 + \psi_1^2 \left[g_a + g_a \overline{C}^T f_{22} \overline{C} - C_a^T f_{22} \overline{C} \right]^2 + 2 \psi_0 \psi_1 \left[g_a + g_a \overline{C}^T f_{22} \overline{C} - C_a^T f_{22} \overline{C} \right], \end{aligned}$$

$$g_2 C_b = \psi_1 \left[C_a^T - g_a \overline{C}^T \right] f_{22} C_b,$$

$$(g_2 C_b)^2 = \psi_1^2 \left\{ \left[C_a^T - g_a \overline{C}^T \right] f_{22} C_b \right\}^2,$$

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

We can rewrite the denominator k as:

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

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

$$
\begin{aligned}
\text{numerator} &= \left(g_1 g_b + g_2 C_b\right)^2 - 2\left(g_1 g_b + g_2 C_b\right)\left(\psi_0 g_b + \psi_1 \tilde{g}\right) + \psi_0^2 g_b^2 + 2\psi_0 \psi_1 g_b \tilde{g} + \psi_1^2 \tilde{g}^2 \\
&= g_1^2 g_b^2 + \left(g_2 C_b\right)^2 + 2 g_1 g_b \left(g_2 C_b\right) - 2\left[g_1 g_b^2 \psi_0 + g_1 g_b \tilde{g} \psi_1 + g_b \left(g_2 C_b\right) \psi_0 + \tilde{g}\left(g_2 C_b\right) \psi_1\right] \\
&\quad + \psi_0^2 g_b^2 + 2\psi_0 \psi_1 g_b \tilde{g} + \psi_1^2 \tilde{g}^2 \\
&= g_1^2 g_b^2 + \left(g_2 C_b\right)^2 + 2 g_1 g_b \left(g_2 C_b\right) - 2\left[g_1 g_b^2 \psi_0 + g_1 g_b \tilde{g} \psi_1 + g_b \left(g_2 C_b\right) \psi_0 + \tilde{g}\left(g_2 C_b\right) \psi_1\right] \\
&\quad + \psi_0^2 g_b^2 + 2\psi_0 \psi_1 g_b \tilde{g} + \psi_1^2 \tilde{g}^2
\end{aligned}
$$

$$
\begin{aligned}
&= \psi_1^2 \Big\{ \left[g_a g_b - \tilde{g}\right]^2 + g_a^2 g_b^2 \left(\overline{C}^T f_{22} \overline{C}\right)^2 + g_b^2 \left(C_a^T f_{22} \overline{C}\right)^2 - 2 g_a g_b^2 \overline{C}^T f_{22} \overline{C} C_a^T f_{22} \overline{C} \\
&\quad + 2 g_a^2 g_b^2 \overline{C}^T f_{22} \overline{C} - 2 g_a g_b^2 C_a^T f_{22} \overline{C} + \left(C_a^T f_{22} C_b\right)^2 + g_a^2 \left(\overline{C}^T f_{22} C_b\right)^2 - 2 g_a C_a^T f_{22} C_b \overline{C}^T f_{22} C_b \\
&\quad + 2 g_b g_a C_a^T f_{22} C_b - 2 g_b g_a g_a \overline{C}^T f_{22} C_b + 2 g_b g_a \overline{C}^T f_{22} \overline{C} C_a^T f_{22} C_b - 2 g_b g_a^2 \overline{C}^T f_{22} \overline{CC}^T f_{22} C_b \\
&\quad - 2 g_b C_a^T f_{22} \overline{C} C_a^T f_{22} C_b + 2 g_b g_a C_a^T f_{22} \overline{CC}^T f_{22} C_b \\
&\quad - 2 g_a g_b \tilde{g} \overline{C}^T f_{22} \overline{C} + 2 g_b \tilde{g} C_a^T f_{22} \overline{C} - 2 \tilde{g} C_a^T f_{22} C_b + 2 \tilde{g} g_a \overline{C}^T f_{22} C_b \Big\} .
\end{aligned}
$$

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

$$
\begin{aligned}
(g_a g_b - \tilde{g})\left(C_b - \overline{C} g_b\right)^T f_{22}\left(C_a - \overline{C} g_a\right) &= (g_a g_b - \tilde{g})\left(C_b^T f_{22} - \overline{C}^T g_b f_{22}\right)\left(C_a - \overline{C} g_a\right)\\
&= (g_a g_b - \tilde{g})\left(C_b^T f_{22} C_a - g_a C_b^T f_{22}\overline{C} - g_b \overline{C}^T f_{22} C_a + g_a g_b \overline{C}^T f_{22}\overline{C}\right)\\
&= g_a g_b C_b^T f_{22} C_a - g_a^2 g_b C_b^T f_{22}\overline{C} - g_a g_b^2 \overline{C}^T f_{22} C_a + g_a^2 g_b^2 \overline{C}^T f_{22}\overline{C}\\
&\quad - \tilde{g} C_b^T f_{22} C_a + \tilde{g} g_a C_b^T f_{22}\overline{C} + \tilde{g} g_b \overline{C}^T f_{22} C_a - \tilde{g} g_a g_b \overline{C}^T f_{22}\overline{C},
\end{aligned}
$$

$$
\begin{aligned}
\text{and}\quad \left[\left(C_b - \overline{C} g_b\right)^T f_{22}\left(C_a - \overline{C} g_a\right)\right]^2 &= \left(C_b^T f_{22} C_a - g_a C_b^T f_{22}\overline{C} - g_b \overline{C}^T f_{22} C_a + g_a g_b \overline{C}^T f_{22}\overline{C}\right)^2\\
&= \left(C_b^T f_{22} C_a\right)^2 + g_a^2\left(C_b^T f_{22}\overline{C}\right)^2 + g_b^2\left(\overline{C}^T f_{22} C_a\right)^2 + g_a^2 g_b^2\left(\overline{C}^T f_{22}\overline{C}\right)^2\\
&\quad - 2 g_a C_b^T f_{22} C_a C_b^T f_{22}\overline{C} - 2 g_b C_b^T f_{22} C_a \overline{C}^T f_{22} C_a + 2 g_a g_b C_b^T f_{22} C_a \overline{C}^T f_{22}\overline{C}\\
&\quad + 2 g_a g_b C_b^T f_{22}\overline{C C}^T f_{22} C_a - 2 g_a^2 g_b C_b^T f_{22}\overline{C C}^T f_{22}\overline{C} - 2 g_a g_b^2 \overline{C}^T f_{22} C_a \overline{C}^T f_{22}\overline{C}.
\end{aligned}
$$

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

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

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

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

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

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

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

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

Define $e^\star = \tilde{\tilde{e}} + E_{\tilde{Z},C}\left[\widetilde{Cov}_{Ca}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Ca}\right]$, where $e^\star - \tilde{\tilde{e}}$ does not depend on b, α_1, and α_2. We have

$$e^\star = E_{\tilde{Z},C}\left[\frac{\left(\widetilde{Cov}_{ab} + \widetilde{Cov}_{Cb}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Ca}\right)^2}{\widetilde{Var}_b - \widetilde{Cov}_{Cb}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Cb}} + \widetilde{Cov}_{Ca}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Ca}\right]$$

$$= E_{\tilde{Z},C}\left[\frac{\widetilde{Cov}_{ab}^{2} + 2\widetilde{Cov}_{ab}\widetilde{Cov}_{Cb}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Ca}}{\widetilde{Var}_b - \widetilde{Cov}_{Cb}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Cb}} + \frac{\widetilde{Cov}_{Cb}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Ca}\widetilde{Cov}_{Cb}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Ca}}{\widetilde{Var}_b - \widetilde{Cov}_{Cb}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Cb}} + \widetilde{Cov}_{Ca}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Ca}\right]$$

$$= E_{\tilde{Z},C}\left[\frac{\widetilde{Cov}_{ab}^{2} + 2\widetilde{Cov}_{ab}\widetilde{Cov}_{Cb}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Ca} + \widetilde{Var}_b\widetilde{Cov}_{Ca}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Ca}}{\widetilde{Var}_b - \widetilde{Cov}_{Cb}^{T}\widetilde{\Sigma}_C^{-1}\widetilde{Cov}_{Cb}}\right].$$

A.5 $e\left(b, \alpha_0, \alpha_1, \alpha_2\right)$ is minimized when $\widetilde{\varrho}_{ab} = 1$

We can see simply that in the absence of confounders (i.e. $p = 0$), $e^\star$ reduces to $E_{\tilde{Z},C}\left[\frac{\widetilde{Cov}_{ab}^{2}}{\widetilde{Var}_b}\right]$, which is maximized at $b = a$ if the expected empirical correlation between $h_i^\star(a, t_i)$ and $h_i^\star(b, t_i)$ is 1. Now we will show the same holds when $p > 0$.

If we denote $\widetilde{Cov}_{C(j)b}$ as the j^{th} element of $\widetilde{Cov}_{Cb}$ and $\varrho_{C(j)b} = \frac{\widetilde{Cov}_{C(j)b}}{\sqrt{\widetilde{\Sigma}_{C(j,j)}\widetilde{Var}_b}}$, where $\widetilde{\Sigma}_{C(j,j)}$ is the $(j,j)^{th}$

element of $\widetilde{\Sigma}_C$. $\widetilde{Cov}_{Cb}$ and $\widetilde{Cov}_{Ca}$ can be written as

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

Define

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

This allows us to write

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

Therefore, $e^\star$ reduces to

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

Here we make the assumption that $\widetilde{\varrho}_b$ is not a function of b. Note that all quantities in the above expression, aside from $\widetilde{\varrho}_{ab}$, are therefore considered fixed with respect to b, conditional on $\tilde{Z}$ and C.

We now examine how the function inside the expectation in $e^\star$ changes across values of $\widetilde{\varrho}_{ab}$. We have

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

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

$$2\widetilde{Var}_a \left[\frac{\widetilde{\varrho}_{ab} + \widetilde{\varrho}_a^T \widetilde{\Sigma}_C^{-1} \widetilde{\varrho}_a}{1 - \widetilde{\varrho}_a^T \widetilde{\Sigma}_C^{-1} \widetilde{\varrho}_a} \right].$$

Note that

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

where Σ_{CC} is the observed correlation matrix for C_i, $i = 1, \ldots, n$. Therefore, $\widetilde{\varrho}_a^T \widetilde{\Sigma}_C^{-1} \widetilde{\varrho}_a$ is equal to the coefficient of multiple determination[23], which is guaranteed to be $\in [0, 1]$. For $\widetilde{\varrho}_{ab} \in [0, 1]$, we see that

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

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

2

Latency Estimation and Inference in the Cox Proportional Hazards Model in the Presence of Exposure Measurement Error

Sarah Bancroft Peskoe

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

Donna Spiegelman

Departments of Epidemiology, Biostatistics, Nutrition and Global Health

Harvard TH Chan School of Public Health

Molin Wang

Departments of Epidemiology and Biostatistics

Harvard TH Chan School of Public Health

2.1 Introduction

Researchers are often interested in estimating the effect of time-varying exposure variables in relation to disease endpoints. In particular, interest lies in effects of cumulatively updated total or cumulatively updated average levels of time-varying exposures. For example, Gillies et al[24] investigated cumulative exposure of radiation and non-cancer mortality among nuclear workers on Europe. Barul et al[25] studied the cumulative level of occupational exposure to chlorinated solvents and the risk of head and neck cancer. Carroll et al[26] reported the effects of cumulative smoking and depressive symptoms on the risk of subclinical heart disease.

Not only are we interested in estimating he effects of these exposures, but also in determining the relevant, critical period of susceptibility. Belenchia et al[27] showed that the fetal period is an important window of susceptibility during which a vitamin D deficiency predisposes offspring to obesity and metabolic disease later in life. Lin et al[28] examined how exposure to amoxicylin in early pregnancy is associated with an increased risk of oral clefts. Identification of this critical period of susceptibility, often denoted a latency period, is an important component of assessing the public health impact of environmental, behavioral, or occupational exposures. Wang et al[7] developed approaches to both identify these periods of susceptibility for a range of exposure metrics and estimate the effects of exposures during this period.

One major obstacle in public health research is that many exposures are prone to measurement error. In many cases, exposure 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[8,9,10,11,12,13,14] and other nonlinear models[14]. The primary approach to correct for measurement error of exposures in survival models is regression calibration, which relies on the assumption of a rare disease and requires estimation of, or historical knowledge of, the underlying measurement error model[29,20]. Further approaches have been developed to correct for bias when the outcome is not rare, and therefore the measurement error model may change over time[20]. As shown in Chapter 1, we have explored the effect of exposure measurement error on estimating latency parameters in linear models, and we proved the naive least squares estimator for the latency parameter is consistent for the true latency parameter when the exposure is prone to linear measurement error. What has yet to be addressed, however, is the impact of exposure measurement error on the estimation of a latency parameter, which defines the susceptibility window, and the relevant effect in survival models. In this paper, we explore the impact of this when the disease is rare, and we demonstrate how to use previously proposed methods to estimate parameters and generate confidence intervals for both the latency parameter and regression coefficients. We apply this approach to air pollution data from the Nurses' Health Study (NHS).

2.2 Methods

2.2.1 Latency Metric

We focus on estimation of the recent moving cumulative average. For an individual, i, a latency parameter, a, and a time-varying exposure $X_i(t)$, recent moving cumulative average is defined as the average exposure over $(t-a,t)$. With a continuous exposure, this can be written as

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

With discrete exposure measurements, $s = 0, 1, 2, \ldots$, as is the case in empirical data, this can be expressed as a sum

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

2.2.2 Hazard Function and Linear Measurement Error

Assume the true hazard function, based on the recent moving cumulative average and the latency parameter, is

$$\lambda_i\left(t|\tilde{X}_i\right) = \lambda_x(t)\exp\left\{\beta h_i(a,t)\right\}$$

where $\lambda_x(t)$ is the baseline hazard function, and $\tilde{X}_i$ is the history of X_i or the vector of X_i. Define $Z_i(t)$ as the exposure measured with error, where the model for the error is linear:

$$X_i(t) = \gamma_0 + \gamma_1 Z_i(t) + \varepsilon_i(t)$$

where $var(\varepsilon_i) = \sigma^2$ and $E[\varepsilon_i] = 0$. Assume the true hazard function, based on the recent moving cumulative average of Z_i and the latency parameter, is

$$\lambda_i\left(t|\tilde{Z}_i\right) = \lambda_z(t)\exp\left\{\alpha h_i^\star(b,t)\right\}$$

with $h_i^\star(b,t) = \frac{\int_{t-b}^{t} Z_i(s)ds}{b}$ and $\lambda_Z(t)$ is the baseline hazard function, and $\tilde{Z}$ is the history of Z or vector of Z. We will show that this is either exactly true or approximately true in a number of scenarios. Our goal is to evaluate the measurement error-caused bias $\hat{b} - a$ and $\hat{\alpha} - \beta$, as the sample size goes to infinity, where b is the mis-measured latency parameter.

Prentice[29] shows that if the three following conditions hold,

(A.1) the proportional hazards model holds in the perfectly measured covariates

$$\lambda\left(t|\tilde{X}_i\right) = \lambda(t|h_i(a,t)) = \lambda_x(t)\exp\left\{\beta h_i(a,t)\right\}$$

(A.2) measurement error is nondifferential (isn't this also surrogacy assumption?)

$$\lambda\left(t|\tilde{X}_i,\tilde{Z}_i\right) = \lambda(t|h_i(a,t), h_i^\star(b,t)) = \lambda(t|h_i(a,t))$$

(A.3) there is random censorship conditional on the observed surrogate exposure

$$\lambda\left(t|\tilde{Z}_i,\ \text{no censorship in } [0,t)\right) = \lambda(t|h_i^\star(b,t),\ \text{no censorship in } [0,t)) = \lambda(t|h_i^\star(b,t))$$

then the hazard function for the surrogate exposure can be written as

$$\lambda\left(t|\tilde{Z}_i\right) = \lambda(t|h_i^\star(b,t)) = \lambda_x(t)E\left[\exp\left\{\beta h_i(a,t)\right\}|\tilde{Z}_i, T\geq t\right],$$

where T_i is the time-to-event for individual i.

2.2.3 Rare Disease Assumption

When a disease is rare, it implies that $P\left(T_i \geq t\right)$ is close to 1. In that case, we have the approximation

$$E\left[\exp\left\{\beta h_i(a,t)\right\} | \tilde{Z}_i, T_i \geq t\right] \approx E\left[\exp\left\{\beta h_i(a,t)\right\} | \tilde{Z}_i\right].$$

From this, we know the critical quantity is $E\left[\exp\left\{\beta h(a,t)\right\} | Z\right]$. If we want to evaluate this exactly, we need to assume a model for the full distribution of $X|Z$. Alternatively, we can use approximation by using moment assumptions only. That is,

$$\begin{aligned}
\lambda(t|Z) &\approx \lambda_x(t) E\left[\exp\left\{\beta h(a,t)\right\} | Z\right] && \text{rare disease assumption} \\
&\approx \lambda_x(t) \exp\left\{\beta E\left[h_i(a,t) | \tilde{Z}_i(t)\right]\right\} && \text{using a first-order approximation.}
\end{aligned}$$

This first order approximation holds when one of the following conditions holds:

(B.1) normally distributed errors for $h_i(a,t)|\tilde{Z}_i$ with constant variance (EXACT)

(B.2) normally distributed errors and small $\beta^2 var\left(h_i(a,t)|\tilde{Z}_i\right)$

(B.3) Small RR and small $\beta^2 var\left(h_i(a,t)|\tilde{Z}_i\right)$

(B.4) Small RR and $h_i(a,t)|\tilde{Z}_i$ has constant variance

Finally, we show in appendix 6.1 that we can substitute the measurement error model to arrive at

$$\lambda(t|\tilde{Z}_i) \approx \lambda_x(t) \exp\left\{\beta E\left[h_i(a,t) | \tilde{Z}_i(t)\right]\right\} = \lambda_x(t) \exp\left\{\beta\gamma_0\right\} \exp\left\{\beta\gamma_1 h_i^{\star}(a,t)\right\}. \tag{2.1}$$

As mentioned, this approximation is of the form $\lambda_i\left(t|\tilde{Z}_i\right) = \lambda_z(t)\exp\left\{\alpha h_i^\star(b,t)\right\}$. When the latency metric is the recent moving cumulative average, we see that there are only three scenarios under which these equations are equivalent:

(1) $\alpha = 0$ and $\beta = 0$ or $\gamma_1 = 0$. This is the trivial case of either not effect ($\alpha = 0$ and $\beta = 0$) or the surrogate exposure is independent of the true exposure ($\gamma_1 = 0$).

(2) $\tilde{X}$ or $\tilde{Z}$ is constant over time. This is the trivial case where the latency parameter is unidentifiable because there is no change in the exposure over time.

(3) $b \approx a$. This is the non-trivial case where the latency parameter for the true exposure is approximately the same as the latency parameter for the surrogate exposure.

Therefore, in the non-trivial case where there is an affect of the recent moving cumulative average, the surrogate exposure is not independent of the true exposure, and the latency parameter is identifiable, it must be the case that the latency parameter for the true exposure is approximately the same as the latency parameter for the surrogate exposure, and there is no effect of linear measurement error on the estimation of the latency parameter. We also note that in this non-trivial case, $\alpha \approx \beta\gamma_1$, implying that there is an effect of linear measurement error on the estimation of the regression coefficient.

2.2.4 Point Estimation

Given that the latency parameter for the surrogate exposure is approximately equal to the latency parameter for the true exposure, we propose a two-stage estimation approach wherein we first estimate the latency parameter for the surrogate exposure, then use this as a plug-in estimator for the latency parameter and correct for the bias in the regression coefficient using a validation data. All estimations are presented for discrete data points, as is the case for the majority of empirical data.

Estimation of Latency Parameter

Methods have been developed to estimate the latency parameter and regression coefficients via a generalized maximum partial likelihood estimator (MPLE) in the survival setting[7]. More specifically, when there are no ties, the partial likelihood based on $\tilde{X}_i$ is

$$L(a, \beta) = \prod_{i \in \mathcal{F}} \frac{\exp\left\{\beta h_i(a, T_i)\right\}}{\sum\limits_j Y_j(T_i) \exp\left\{\beta h_j(a, T_i)\right\}}$$

where $\mathcal{F}$ is a subset containing all the cases, and $Y_i(t)$ is the at-risk process for the ith individual, equal to 1 if at risk at time t and 0 otherwise. Let $N_i(t)$ be the counting process for the number of observed failures on $(0, t]$. The partial likelihood score function based on data available up to a specified time t is

$$D(a, \beta) = \sum_{i \in \mathcal{F}} \int_0^t \left\{ \begin{array}{c} \beta\left[\partial h_i(a, u)/\partial a\right] - \dfrac{\sum\limits_{j \in \mathcal{F}} \beta\left[\partial h_i(a, u)/\partial a\right] \times Y_j(u) \exp\left\{\beta h_j(a, u)\right\}}{\sum_j Y_j(u) \exp\{\quad h_j(a,u)\}} \\ \\ h_i(u, a) - \dfrac{\sum\limits_{j \in \mathcal{F}} h_j(a, u)}{\sum_j Y_j(u) \exp\{\quad h_j(a,u)\}} \end{array} \right\} dN_i(u)$$

We can also do this for the partial likelihood based on Z_i:

$$L^{\star}(b,\alpha) = \prod_{i \in \mathcal{F}} \frac{\exp\left\{\alpha h_i^{\star}(b, T_i)\right\}}{\sum_j Y_j(T_i) \exp\left\{\alpha h_j^{\star}(b, T_i)\right\}}$$

$$D^{\star}(b,\alpha) = \sum_{i \in \mathcal{F}} \int_{0}^{t} \left\{ \begin{array}{c} \alpha\left[\partial h_i^{\star}(b,u)/\partial b\right] - \dfrac{\sum\limits_{j \in \mathcal{F}} \alpha\left[\partial h_i^{\star}(b,u)/\partial b\right] \times Y_j(u) \exp\left\{\alpha h_j^{\star}(b,u)\right\}}{\sum_j Y_j(u) \exp\left\{\quad h_j^{\star}(b,u)\right\}} \\ \\ h_i(u,a) - \dfrac{\sum\limits_{j \in \mathcal{F}} h_j^{\star}(b,u)}{\sum_j Y_j(u) \exp\left\{\quad h_j^{\star}(b,u)\right\}} \end{array} \right\} dN_i(u)$$

To obtain the point estimate of the latency parameter, denoted $\hat{a}$, we combine a grid search and the Newton-Raphson approach. That is, we use Newton-Raphson to fit a Cox proportional hazard model for all possibly latency parameter values, then determine the estimator for the latency parameter as that which corresponds to the model that provides the maximum partial likelihood.

Estimation of Regression Coefficient

Using the MPLE for the latency parameter, we can proceed to correct for the bias in the regression coefficient due measurement error by performing ordinary regression calibration (ORC)[29,13,30]. Using validation data, we fit the linear measurement error model to obtain estimates for γ_0 and γ_1, denoted $\hat{\gamma}_0$ and $\hat{\gamma}_1$. We then proceed to predict $\tilde{X}_i$ for all individuals in the study, and fit the Cox proportional hazard model with the exposure

$$\hat{h}_i\left(\hat{a}, t\right) = \frac{\sum_{s=t-\hat{a}}^{t} \hat{X}_i(s)}{\hat{a}+1}.$$

2.2.5 Inference

Profile Likelihood Confidence Interval for the Latency Parameter

Given that the latency parameter for the surrogate exposure is approximately equal to the latency parameter for the true exposure, we can implement the method developed by Wang et al[7] to produce a profile likelihood confidence interval for the latency parameter when performing analyses on exposure data collected at discrete time points. In this approach, a $\tilde{\alpha}$-level profile likelihood confidence interval is determined as all possible values of a that satisfy

$$\log PL(a, \alpha_a) \geq \log PL\left(\hat{a}, \hat{\alpha}\right) - \frac{1}{2}\chi_1^2(1-\tilde{\alpha}),$$

where $PL\left(\hat{a}, \hat{\alpha}\right)$ is the partial likelihood evaluated at the point estimates $\hat{a}$ and $\hat{\beta}$, $PL(a, \alpha_a)$ is the partial likelihood evaluated at a given possible value of a and the corresponding α_a determined as the maximum partial likelihood estimator for the regression coefficient when the latency parameter is equal to a, and χ_1^2 is the cumulative function of the χ^2 distribution with 1 degree of freedom.

Confidence Interval for the Regression Coefficient

While we have demonstrated that the latency parameter for the surrogate exposure is approximately equal to the latency parameter for the true exposure, the regression coefficient is not. Therefore, adjustment for measurement error must be considered not only in the point estimate, but also in generating valid confidence intervals. We have seen in previous studies that while there is some variability involved in estimating the latency parameter, this estimation has very little effect impact on the variability in estimating the regression coefficient when there is no measurement error[7]. We argue that in this scenario, it may be unnecessary to adjust for the variability involved in estimating the latency parameter when generating confidence intervals for regression coefficients even in the presence of measurement error. Therefore, we propose treating the latency parameter as fixed and generating confidence intervals for regression parameters using regression calibration methods developed by Spiegelman et al[13] and Wang et al[7]. In this case, the asymptotic distribution of the regression coefficients is normal with mean zero. Software to calculate these confidence intervals is readily available using the %blinplus() SAS macro[31].

2.3 Simulation

We performed a series of finite-sample simulations to examine the performance of the naive estimator for the latency parameter when the time-varying exposure is subject to linear measurement error. This simulation is based off Liao et al[20], wherein investigators generated survival data with a recent moving cumulative average as the primary exposure.

2.3.1 Data Generation

The true and surrogate exposures, $\tilde{X}_i$ and $\tilde{Z}_i$ were generated as follows:

$$\tilde{Z}_i \sim MVN\left(\mu_z, \Sigma_z\right)$$
$$e_{it} \sim N(0, \sigma_e)$$
$$X_i(t) = \gamma_0 + \gamma_1 Z_i(t) + e_{it},$$

where μ_z is the mean vector for $\tilde{Z}$, and Σ_z is a compound symmetric covariance matrix with diagonal entries σ_z^2 and off-diagonal entries $\varrho_z \sigma_z^2$, for time points $t = 1, ..., 10$. Without loss of generality, we consider the simple data generation case with $\mu_z = 0$, $\sigma_z^2 = 0$, and $\varrho_z = 0.2$ and the linear measurement error model with $\gamma_0 = 1$ and $\gamma_1 = 0.5$. We varied the linear measurement error model by σ_e to be 1 or 2. For comparison, we also included a simulation with no measurement error where $X_i(t) = Z_i(t)$.

The true and surrogate recent moving cumulative average were calculated as follows:

$$h_i(a,t) = \frac{\sum\limits_{s=t-a}^{t} X_i(s)}{a+1}$$
$$h_i^\star(a,t) = \frac{\sum\limits_{s=t-a}^{t} Z_i(s)}{a+1},$$

for $a \in 1, \ldots, 10$. The survival data were generated according to the hazard model $\lambda(t|h_i(a,t)) = \lambda_0(t)\exp\{\beta h_i(a,t)\}$. For all simulations, we set the event rate to 0.03 to ensure the disease rare assumption applied.

2.3.2 Latency and Coefficient Estimation

To estimate the latency parameter, a and the coefficient β, we calculated $h_i^\star(a,t)$ for all values $a = 1, .., 10$, fit a Cox proportional hazards model using $h_i^\star(a,t)$ as the exposure, then selected $\hat{a}$ as the value that maximizes the profile likelihood. After selecting $\hat{a}$, we fit a Cox proportional hazards model using $h_i^\star(\hat{a},t)$ to find $\hat{\beta}$, the estimate for β.

2.3.3 Results

Mean estimated values for the latency parameter, $\hat{a}$ can be found in Table 2.1. We see that across all measurement error models, including the scenario with no measurement error, as the sample size increases, the estimated values for the latency parameter approach the true value, a. This is to be expected, as the latency parameter for the surrogate exposure is approximately equivalent to the latency parameter for the true exposure in this setting.

Mean estimated values for the regression coefficient, $\hat{\beta}$, can be found in Table 2.2. We wee that when there is no measurement error or the error is random (i.e. $\gamma_0 = 0$ and $\gamma_1 = 1$), the estimated regression coefficient approaches the true value, β. However, when the error is not random, the estimation regression coefficient approaches $\gamma_1\beta$. This is consistent with what we expect under linear measurement error, where we see the hazard model for the surrogate exposure is approximately

Table 2.1. Mean estimated latency parameter

	$\hat{a}$				
	$n = 500$	$n = 1{,}000$	$n = 5{,}000$	$n = 10{,}000$	$n = 25{,}000$
$a = 1$					
No Error	2.15	1.58	1.01	1.00	1.00
Random Error $\sigma = 1$	1.80	1.44	1.01	1.00	1.00
Random Error $\sigma = 2$	1.82	1.28	1.00	1.00	1.00
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 1$	3.52	2.76	1.48	1.13	1.02
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 2$	3.35	2.71	1.36	1.13	1.02
$a = 3$					
No Error	4.04	3.77	3.16	3.05	3.00
Random Error $\sigma = 1$	4.14	3.75	3.12	3.03	3.00
Random Error $\sigma = 2$	4.00	3.57	3.13	3.03	3.00
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 1$	4.35	4.34	3.76	3.47	3.07
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 2$	4.50	4.24	3.74	3.47	3.07
$a = 5$					
No Error	5.08	5.32	5.23	5.11	5.02
Random Error $\sigma = 1$	5.22	5.33	5.18	5.09	5.02
Random Error $\sigma = 2$	5.12	5.25	5.19	5.09	5.01
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 1$	4.66	4.86	5.19	5.36	5.09
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 2$	4.86	5.06	5.22	5.41	5.09
$a = 8$					
No Error	5.92	6.86	7.76	7.94	8.03
Random Error $\sigma = 1$	5.79	6.80	7.75	7.98	8.02
Random Error $\sigma = 2$	6.04	6.56	7.79	8.01	8.04
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 1$	5.08	5.51	6.62	7.22	7.79
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 2$	5.01	5.43	6.61	7.30	7.90
$a = 10$					
No Error	6.25	7.19	8.88	9.31	9.82
Random Error $\sigma = 1$	6.42	7.26	8.91	9.43	9.81
Random Error $\sigma = 2$	6.58	7.30	8.96	9.42	9.80
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 1$	5.12	5.49	7.18	7.92	8.90
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 2$	4.98	5.56	7.25	7.91	8.92

Table 2.2. Mean estimated regression coefficient

	$\hat{\beta}$				
	$n = 500$	$n = 1{,}000$	$n = 5{,}000$	$n = 10{,}000$	$n = 25{,}000$
$a = 1$					
No Error	1.12	1.06	1.00	1.00	1.00
Random Error $\sigma = 1$	1.08	1.03	0.99	0.99	0.99
Random Error $\sigma = 2$	1.07	1.01	0.98	0.98	0.98
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 1$	0.64	0.61	0.52	0.50	0.50
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 2$	0.62	0.59	0.51	0.50	0.50
$a = 3$					
No Error	1.17	1.09	1.01	1.01	1.00
Random Error $\sigma = 1$	1.14	1.08	1.01	1.00	1.00
Random Error $\sigma = 2$	1.12	1.06	1.00	1.00	0.99
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 1$	0.64	0.60	0.53	0.51	0.50
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 2$	0.62	0.61	0.53	0.51	0.50
$a = 5$					
No Error	1.11	1.08	1.01	1.00	1.00
Random Error $\sigma = 1$	1.11	1.06	1.01	1.00	1.00
Random Error $\sigma = 2$	1.08	1.05	1.00	1.00	0.99
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 1$	0.60	0.56	0.52	0.51	0.50
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 2$	0.62	0.58	0.52	0.51	0.50
$a = 8$					
No Error	1.03	1.08	1.01	1.00	1.00
Random Error $\sigma = 1$	1.11	1.06	1.01	1.00	1.00
Random Error $\sigma = 2$	1.06	1.01	1.01	1.00	1.00
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 1$	0.58	0.54	0.51	0.50	0.50
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 2$	0.56	0.53	0.51	0.50	0.52
$a = 10$					
No Error	1.02	1.00	1.00	0.99	1.00
Random Error $\sigma = 1$	1.03	1.00	0.99	0.99	1.00
Random Error $\sigma = 2$	1.03	1.00	0.99	0.99	1.00
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 1$	0.54	0.52	0.50	0.49	0.51
$\gamma_0 = 2, \gamma_1 = 0.5, \sigma = 2$	0.47	0.52	0.50	0.50	0.51

proportionate to $\exp\{\beta\gamma_1 h_i^\star(a,t)\}$, as shown in equation (2.1).

2.4 Illustrative Example

We applied the methodology above to estimate the latency parameter and regression coefficient for the recent moving cumulative average of ambient particulate matter $\leq 2.5\mu m$ ($PM_{2.5}$) exposure and its association with lung cancer in the NHS This analysis is based off a previously reported study by Puett et al[32], wherein investigators estimated the association between 72-month cumulative average $PM_{2.5}$ exposure, as measured by nearest monitor, and lung cancer incidence between 1994 and 2010.

2.4.1 Nurses' Health Study

The NHS is an ongoing prospective cohort of 121,700 female nurses who were enrolled between the ages of 30 and 55 in 1976. Participants complete biennial questionnaires by mail and provide information on potential risk factors and self-report new diagnoses of health outcomes. Included in this analysis are 103,650 women who were alive and free of cancer (except for non-melanoma skin cancer) before follow-up and had $PM_{2.5}$ information. Lung cancers were either, self-reported by the participants, reported by next of kin, or identified from death certificates and medical records. A number of potential confounders were selected *a priori* as known confounders or effect modifiers previously associated with lung cancer or exposure in the NHS. These confounders included geographic region of residence (Northeast, South, Midwest, West), body mass index (BMI; kilograms per meter squared, continuous), physical activity in metabolic equivalent hours per week (MET hr/week; < 3, 3 to < 18, ≥ 18), overall diet quality (Alternative Health Eating Index, continuous)[33], alcohol consumption (dichotomized at 0 g/day), smoking status (current, former, never), months since quitting for former smokers (continuous), pack-years (continuous), exposure to secondhand smoke at home, at work, and during childhood, median household income, and median house value. Table 2.3 shows the age-adjusted descriptive statistics of these potential confounders averaged over follow-up (1994-2010).

Table 2.3. Age-adjusted descriptive characteristics averaged over follow-up (1994-2010) among 103,650 participants in the Nurses' Health Study

Lung cancer cases	2,155
Person-years*	1,510,027
Age [years (mean ± SD)]*	67.0 ± 8.3
BMI [kg/m^2 (mean ± SD)]	25.6 ± 7.5
Pack-years of smoking (mean ± SD)	13.4 ± 20.0
Months since quit smoking (mean ± SD)	123.9 ± 178.8
Alternative healthy eating index (mean ± SD)	180.4 ± 108.5
Census-tract median household income (mean ± SD)	63,518 ± 24,491
Census-tract median home value (mean ± SD)	170,126 ± 125,261
Region (%)	
Northeast	51.1
Midwest	17.3
West	13.7
South	18.0
Alcohol category (%)	
Nondrinker (0 g/day)	15.0
Drinker	71.4
Missing	13.6

*Not age adjusted

Table 2.3 (continued). Age-adjusted descriptive characteristics averaged over follow-up (1994-2010) among 103,650 participants in the Nurses? Health Study

Physical activity	
<3 MET hr/week	21.5
3 to <18 MET hr/week	38.8
$\geq$ 18 MET hr/week	30.7
Missing	9.0
Second hand smoke during childhood (%)	
None	25.1
From mother	3.8
From father	33.9
From both parents	14.7
Missing	22.6
Home secondhand smoke (%)	
None	33.2
Occasional	18.6
Regular	17.1
Missing	31.2
Occupational secondhand smoke (%)	
None	15.0
Occasional	29.3
Regular	22.6
Missing/not working	33.1

*Not age adjusted

2.4.2 $PM_{2.5}$ Exposure in NHS

As part of the NHS biennial questionnaire, reported residential address information was updated every 2 years. All available addresses were geocoded to obtain the corresponding latitude and longitude, and monthly $PM_{2.5}$ exposure was estimated using models for the nearest air monitoring-station. These models and their previous use in assessing chronic PM exposures among the NHS cohort are described in detail elsewhere[?] [46,47,48]. This nearest monitor $PM_{2.5}$ exposure is certainly prone to measurement error in that nearest monitor may not perfectly reflect the true geography of each participant during each month.

2.4.3 $PM_{2.5}$ External Validation Study

The details of the selected validation study and corresponding populations have been described previously[34,39,36,37,38,39,40,41,42,43]. Both personal and ambient exposures of $PM_{2.5}$ were available from a series of short-term panel exposure studies performed in the US between 1999 and 2002. Reported personal $PM_{2.5}$ is considered the 'true' exposure and ambient $PM_{2.5}$ is considered the surrogate exposure.

2.4.4 Latency Estimation

Figure 2.1 shows the estimated partial log likelihood for the Cox proportional hazards model for lung cancer incidence, using age as the time scale, and using the surrogate exposure (recent moving cumulative average ambient $PM_{2.5}$) for all possible latency parameters between 6 and 120 months. The first plot shows the partial log likelihood adjusted for geographic region and calendar year (partially adjusted), and the second shows this for the model adjusted for all *a priori* selected confounders (fully adjusted). The point estimate for the latency parameter corresponds to the value of *a* that maximizes this log likelihood, which is 65 months in the partially adjusted model and 70 months for the fully adjusted model. We also calculated the profile likelihood confidence interval, which in fact included the entire range of possible latency parameters (6, 120) for both the partially adjusted and fully adjusted models. This is not unexpected as we have seen in previous studies that since air pollution exposure is highly correlated over time and the regression coefficient is close to zero, there is relatively low power to detect the latency parameter and thus particularly wide confidence intervals.

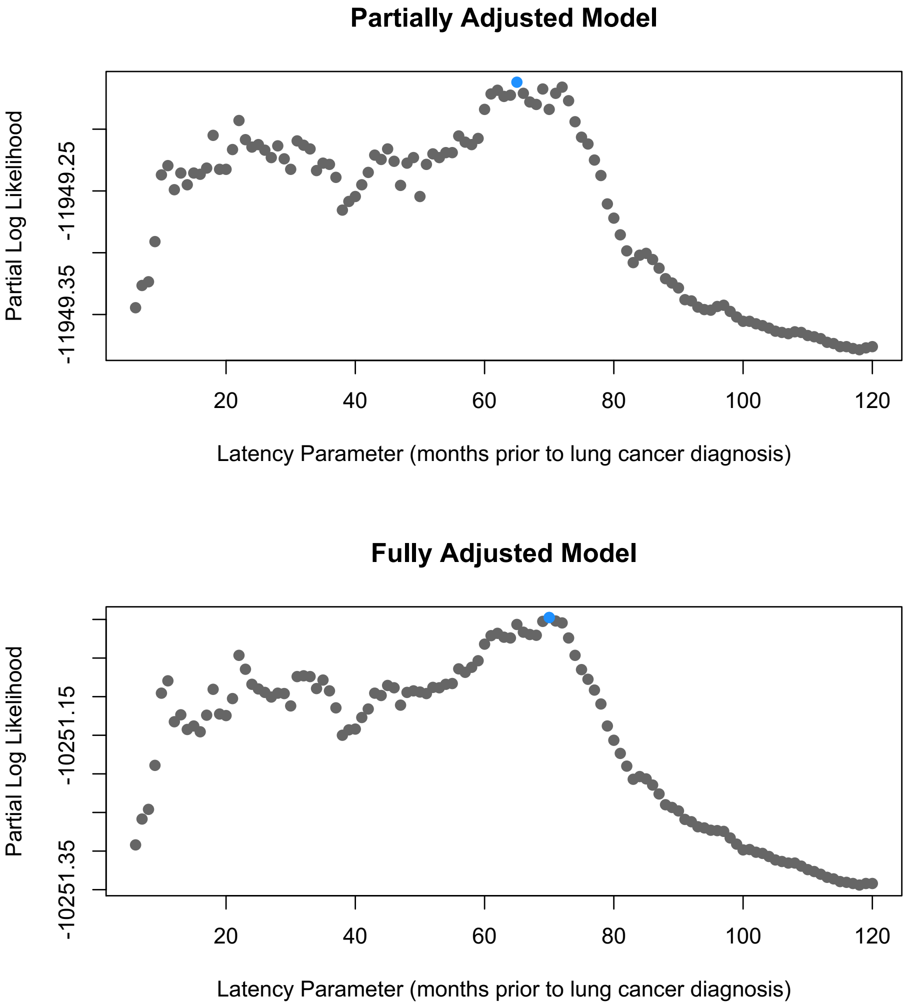


Figure 2.1: Estimated partial likelihood

2.4.5 Measurement Error Correction

We used ordinary regression calibration based on the external validation study to adjust the hazard ratio (HR) for $PM_{2.5}$ exposure. Table 2.4 shows the uncorrected and corrected HRs for both the partially adjusted and fully adjusted models.

Table 2.4. Latency parameter in months and HR (95% confidence intervals) of the association of incident lung cancer 1994-2010 per 10-$\mu g/m^3$ increase in recent moving cumulative average $PM_{2.5}$ exposures among 103,650 participants of the Nurses Health Study

Model	Latency parameter	Uncorrected HR	Corrected HR
Partially adjusted*	65 (6, 120)	1.08 (0.90, 1.30)	1.20 (0.77, 1.86)
Fully adjusted**	70 (6, 120)	1.09 (0.91, 1.32)	1.24 (0.79, 1.92)

* Adjusted for geographic region and calendar year

**Adjusted for geographic region, calendar year, BMI, alcohol consumption, physical activity, overall diet quality, smoking status (when not stratified by status) and pack-years, months since quitting smoking, secondhand smoke exposure at home, work, and during childhood, and census-tract median home value, and median income.

We see that overall, there appears to be an increased risk of lung cancer among women who were exposed to higher average levels of $PM_{2.5}$ during the prior 65 months of 70 months. Although this increased risk is not statistically significant at the 0.05 level, the estimated risk increases after correcting for measurement error.

2.5 Discussion

Public health researchers are often interested in estimating the effects of cumulatively updated total or cumulatively updated average levels of time-varying exposures and identifying critical windows of susceptibility. Although it is widely known that many measurements are prone to error, no one has yet considered its impact on the estimation of latency parameters in survival models. We have shown that when a time-varying exposure is subject to linear error, under the recent moving cumulative average latency metric, the latency parameter for the exposure with error is approximately equal to the latency parameter for the exposure without error. This was evident in a series of finite sample simulations. Although there appears to be no asymptotic impact of measurement error on the latency parameter, the regression coefficients may still be biased. For this, we recommend a two-step estimation approach wherein you first estimate the latency parameter by performing a grid search across all possible latency parameter values and selecting the value that maximizes the profile likelihood, then correcting the regression coefficient by performing regression calibration on the calculated latency metric. We used this approach to estimate the latency parameter for $PM_{2.5}$ exposure and its corresponding association to lung cancer incidence in the NHS.

One major drawback to the proposed approach is the reliance on a rare disease. When a disease is not rare (generally more than 5% of the target population develops the disease), we cannot directly apply the results from Prentice[29]. Instead, we have seen in preliminary simulations that the latency parameter may be approximately the same, but the regression coefficient will again be biased. To correct for measurement error, one must employ a risk set calibration approach such as the method proposed by Liao et al[20] to allow for a time-varying measurement error model.

In future research we hope to extend this to other exposure metrics and generalize to a wider class of measurement error models. Additionally, it may be beneficial to develop an analytic approach to generate confidence intervals for regression parameters that do account for the variability due to estimating the latency parameter. Ultimately, it is quite important to note that although validation data may be necessary to estimate associations between latency metrics and disease incidence, the estimation of the latency parameter itself is robust to exposure measurement error. Thus, we can gain a great deal of insight from time-varying data even in the presence of measurement error.

2.6 Appendix

2.6.1 First order approximation of the hazard function when the disease is rare

As shown by Prentice (1982), $E\left[\exp\left\{\beta h_i(a,t)\right\}|\tilde{Z}_i, T \geq t\right] \approx E\left[\exp\left\{\beta h_i(a,t)\right\}|\tilde{Z}_i\right]$ when the disease is rare. Using a first order approximation, we find

$$E\left[\exp\left\{\beta h_i(a,t)\right\}|\tilde{Z}\right] \approx \exp\left\{\beta E\left[h_i(a,t)|\tilde{Z}_i(t)\right]\right\}$$

We can substitute the linear measurement error model to obtain an expression in terms of Z:

$$
\begin{aligned}
X_i(t) &= \gamma_0 + \gamma_1 Z_i(t) + \varepsilon_i(t) \\
E\left[X_i(t)|Z_i(t)\right] &= E\left[\gamma_0 + \gamma_1 Z_i(t) + \varepsilon_i(t)\right] \\
&= \gamma_0 + \gamma_1 E\left[Z_i(t)\right] + E\left[\varepsilon_i\right] \\
&= \gamma_0 + \gamma_1 \mu_z(t) \\
E\left[h_i(a,t)\right] &= E\left[\frac{\int_{t-a}^{t} X_i(s)ds}{a}\right] \\
&= \frac{1}{a}\int_{t-a}^{t} E\left[X_i(s)ds\right] \\
&= \frac{1}{a}\int_{t-a}^{t} \gamma_0 + \gamma_1 Z_i(s)ds \\
&= \gamma_0 + \gamma_1 \frac{1}{a}\int_{t-a}^{t} Z_i(s)ds \\
&= \gamma_0 + \gamma_1 h_i^{\star}(a,t) \\
\exp\left\{\beta E\left[h_i(a,t)|\tilde{Z}_i(t)\right]\right\} &= \exp\left\{\beta\left(\gamma_0 + \gamma_1 h_i^{\star}(a,t)\right)\right\} \\
&= \exp\left\{\beta\gamma_0\right\}\exp\left\{\beta\gamma_1 h_i^{\star}(a,t)\right\}
\end{aligned}
$$

Thus we conclude $E\left[\exp\left\{\beta h_i(a,t)\right\}|\tilde{Z}\right] \approx exp\left\{\beta\gamma_0\right\}\exp\left\{\beta\gamma_1 h_i^{\star}(a,t)\right\}$ and

$$
\lambda(t|\tilde{Z}_i) \approx \lambda_x(t)\exp\left\{\beta E\left[h_i(a,t)|\tilde{Z}_i(t)\right]\right\} = \lambda_x(t)\exp\left\{\beta\gamma_0\right\}\exp\left\{\beta\gamma_1 h_i^{\star}(a,t)\right\}.
$$

3

Adjusting for Selection Bias in Electronic Health Records-based Research

Sarah Bancroft Peskoe

Department of Biostatistics

Harvard Graduate School of Arts and Sciences

David Arterburn

Kaiser Permanente Washington Health Research Institute

and University of Washington

Michael J Daniels

Department of Statistics & Data Sciences

University of Texas at Austin

Sebastien Haneuse

Department of Biostatistics

Harvard TH Chan School of Public Health

3.1 Introduction

Electronic Health Records (EHR) data provide a number of unique opportunities and substantial promise for public health research. They often include a broad range of collected information not always available in cohort studies, they house information for very large patient populations, they follow patients for a longer time-frame than many prospective cohorts can, and they have a relatively low cost because data are already collected[50,51,52,53]. It is for these reasons the Institute of Medicine recently called for a prioritization of EHR data in public health research[54].

Beyond the several advantages, there are many challenges faced by researchers using EHR data. Given that EHRs are typically developed for clinical and/or billing purposes, without a specific research question in mind, researchers using EHR data must therefore beg the question of whether or not the EHR data is in fact suitable for the research agenda. This includes consideration of whether or not the study population, which is by default rather than by design, is generalizable to the population of interest, whether all covariates relative to the research goals have been routinely collected, whether the covariates that have been measured are done so consistently across patients and time, and finally whether those measures are complete and reliable. Oftentimes the collection of these relevant data is left to the discretion of the physician or the patient. Without consideration of these issues, naive analyses may be subject to numerous biases, the most commonly cited of which is confounding bias due to a non-randomized study population[55]. There are several methods to adjust for confounding bias, including stratification by subgroup[54] and regression approaches[56,57,58,55,59].

Another important and often underappreciated type of bias that may arise in this setting is selection bias as a result of incomplete data. That is, patients who are identified as being eligible for inclusion in the study are found to have insufficient data to be included in the analysis[59]. This could be a result of missing baseline covariate information, missing treatment information, or missing outcome measurements during follow-up. This could also arise when patients disenroll from the health plan prior to the end of follow-up. It is important to note that selection bias is distinct from confounding bias in that the bias results from conditioning on a common effect rather than from the existence of common causes of exposure and outcome[59]. While there have been numerous papers dedicated to understanding the role of confounding bias in electronic health records research, there has been little to no focus on the impact of selection bias in EHR-based studies in the literature to date, as it is often cast as a simple missing data problem. In this paper, we focus on the issue of selection bias due to incomplete data.

Statistical methodology is flooded with approaches to handle missing data[60]. Once researchers have described what it means to have complete or incomplete data, they continue with one of several approaches: (1) assume data are missing completely at random and perform a complete data analysis, (2) assume data are missing at random and use a known approach like multiple imputation or inverse probability weighting (IPW), or (3) conclude data are not missing at random and either additional data need to be collected or the study is no longer feasible. Most relevant to the current data setting, Robins et al[61] developed methodology to specifically account for missing data when we have repeated outcomes measured. To date, most literature has focused on approaches to handle data that are missing at random by which researchers aver a particular 'missingness mechanism' and perform analyses that account for the missingness. Per Haneuse and Daniels (2016), we suggest reframing this as 'data provenance', or a process by which some data are complete and others are not. When framed as a missing-data problem, standard methods are often applied to control for selection bias. In EHR-based studies, however, data provenance involves the interplay of many clinical decisions made by patients, health care providers, and the health system; thus standard methods fail to capture the complexity of the data. In this paper, we build on Haneuse and Daniels' data provenance approach to develop a general framework for estimation and inference based on IPW that better aligns with the complex nature of EHR data.

Table 3.1. Baseline characteristics of patients in the DURABLE study who underwent bariatric surgery between 1997 and 2010

	All Patients N = 16,282			Complete Cases n=5,636		
	All	RYGB	VSG	All	RYGB	VSG
Age at surgery, mean (sd)	45.7 (10.7)	45.7 (10.7)	45.6 (10.9)	47.1 (10.7)	47.1 (10.7)	46.7 (10.6)
Year of surgery category, %						
1997-2005	25.1	27.9	0.4	5.1	6.0	0.0
2006-2008	37.3	40.3	10.5	42.8	48.6	11.8
2008-2010	37.6	31.8	89.1	52.1	45.4	88.2
Prior enrollment years, mean (sd)	8.4 (5.7)	8.2 (5.5)	10.7 (7.2)	10.3 (6.4)	10.0 (6.2)	11.7 (7.4)
Male, %	16.9	16.6	19.3	15.7	15.0	19.4
Site, %						
KW	9.2	10.1	0.9	7.2	8.5	0.3
KNC	43.6	48	4.1	29.5	34.2	3.7
KSC	47.2	41.8	95	63.4	57.3	95.9
Baseline BMI, mean (sd)	–	–	–	44.9 (7.3)	45.0 (7.3)	44.2 (6.9)
BMI pre slope, mean (sd)	–	–	–	-1.7 (2.6)	-1.8 (2.7)	-1.1 (2.0)
BMI at 5 years, mean (sd)	–	–	–	34.7 (7.0)	34.4 (6.9)	36.5 (7.3)
BMI post slope, mean (sd)	–	–	–	-10.1 (5.9)	-10.6 (5.8)	-7.7 (5.4)

We see immediately that there is a substantial amount of missing information. In particular, only $n = 5,363$ patients of the $N = 16,282$ who were identified as satisfying the eligibility criteria have complete information relevant for the research question. Figure 3.1 provides a summary of BMI-related information for 6 patients in DURABLE who underwent bariatric surgery between 1997 and 2010. There is substantial heterogeneity in the amount of information available for different patients. Some patients are missing baseline information, some disenroll from a participating health-care plan, and some are simply missing a BMI measurement at 5 years. All of these would results in 'incomplete data' for for a patient, but the reasons are quite distinct. Further, this missingness appears to be heterogenous across potential confounders. For instance, we see in Table 3.1 that patients who underwent surgery between 1997 and 2005 are less likely to have complete information than individuals who underwent surgery between 2008 and 2010. It is evident that whether or not an individual has complete data is more complicated than simple dropout, but rather the result of a complex series of decisions made by patients, health care providers, and the health system. Thus, even in a rich EHR-data setting, such as the DURABLE study, it is important to ask ourselves how do some people have complete information and others do not. In EHR data, this process is referred to as the *data provenance*[66].

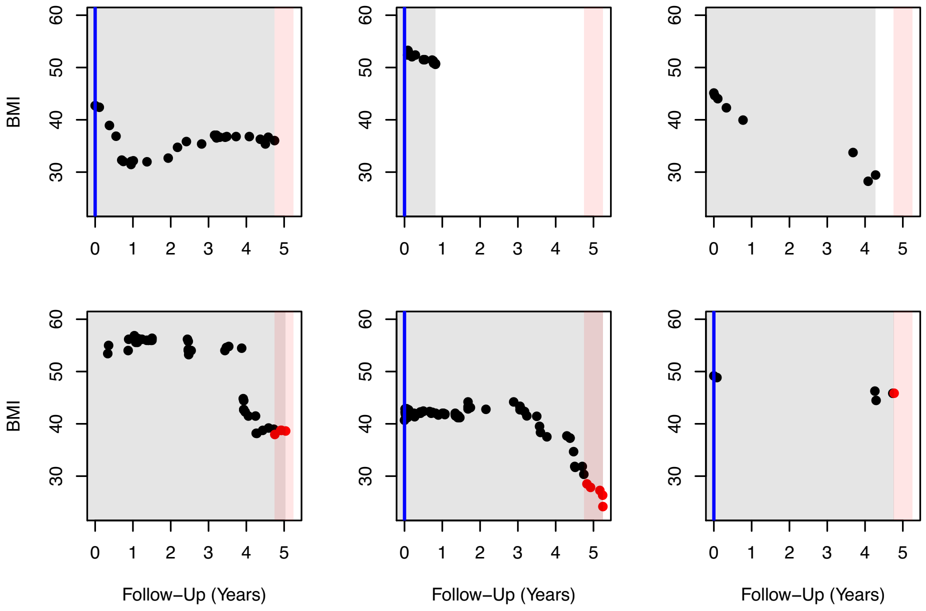


Figure 3.1: Summary of BMI-related information for 6 patients in DURABLE

In each panel, a blue vertical line indicates complete baseline information, including but not limited to pre-surgery BMI; grey shading indicates enrollment in one of the participating healthcare plans; black dots indicate a BMI measurement after surgery but before 5 years; red dots indicate a BMI measurement 5 years after surgery.

3.2 Characterizing the Observed Data Provenance

The importance of the data provenance in an EHR-based setting is the classification of the process by which data is observed. In an effort to capture the complex nature of EHR data, we use the novel data provenance framework proposed by Haneuse and Daniels (2016). This approach focuses on a modularization of the data provenance wherein investigators break down the series of decisions that cause a patient to have incomplete data with respect to a particular research question. For ease of illustration, we present this approach using the DURABLE data with the aim of examining the impact of bariatric surgery on BMI after 5 years.

3.2.1 Choosing the Question of Interest

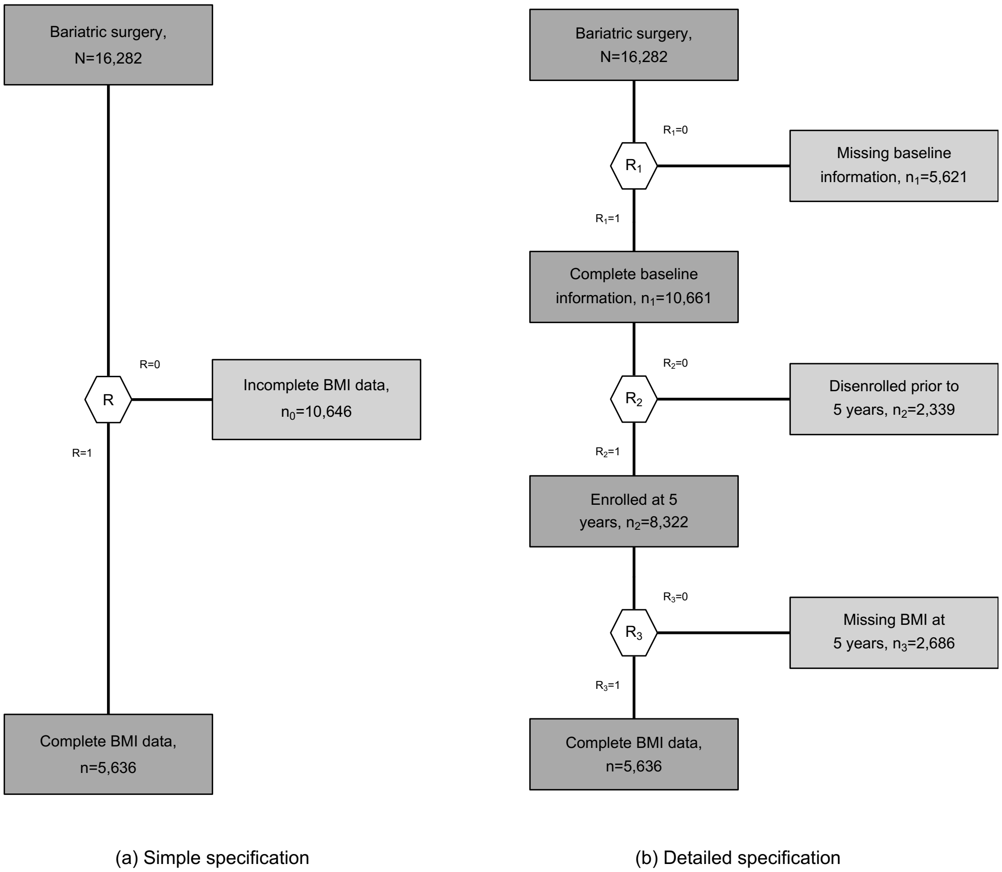
Suppose $N = 16,282$ patients are identified in the EHR as satisfying a set of pre-specified inclusion/exclusion criteria. Ideally, all N patients would have 'complete' information on the exposure, potential confounders and the outcome. If we cast selection bias as a missing data problem, there are many well-known approaches we can use moving forward. The concept of missing data, however, is only relevant in the context of a specific research question. We may be interested in the association between a single baseline exposure and a single outcome, the trajectory of a biological measure over time, or even time-to-event outcomes. In the context of the DURABLE study, this could be BMI at 5 years, trajectory of BMI over 5 years, or time to reduce BMI by some pre-specified amount. Only once the data question has been established can we identify what is necessary for an individual to have complete information and proceed with an approach to account for any missingness. For concreteness, we take the outcome to be BMI measured at 5 years in this paper, but recognize that there are numerous questions that could be asked of the DURABLE study data.

3.2.2 The Standard Single Mechanism Approach

Following inspection of the observed data, suppose only $n = 5,363 < N$ patients have complete BMI information at 5 years and complete baseline information on potential confounders; for simplicity we temporarily ignore the potential for incomplete data in the primary exposure. In principle one can cast the challenge of only having complete data on n patients as a missing data problem and make use of a broad range of existing methods. Regardless of the specific method chosen, analysts must first consider whether or not the data are missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). In practice, this is often operationalized by first defining a binary random variable, say, *R* which indicates whether or not a patient has complete data (0/1 = incomplete/complete); Figure 3.2(a) provides a graphical representation. Based on this, an analyst can work with subject-matter experts to identify covariates associated with *R* and, consider whether or not the data are MCAR, MAR or MNAR, ad eventually build a model for an IPW approach to adjust for selection if appropriate.

3.2.3 Modularization of Data Provenance

While reasonable in many research settings, the use of a single mechanism will be a gross oversimplification of reality in most EHR-based studies. Figure 3.1 illustrates this with EHR- derived information for six patients on BMI following bariatric surgery. If we are specifically interested in BMI at 5 years, we would mark four of these patients as having incomplete data: (c) and (d) are missing baseline information, (b) has complete baseline information but disenrolled from a participating health plan prior to 5 years, and (a) has complete baseline information, stayed enrolled in a participating health plan through 5 years, but failed to have a BMI measurement at 5 years.



(a) Simple specification

(b) Detailed specification

Figure 3.2: Alternative specifications for data provenance in hypothetical study

For EHR-based data such as that presented in Figure 3.1, naive application of standard missing data methods/strategies will generally result in inadequate adjustment for selection bias. For example, if a single logistic model is assumed to estimate IP weights, when in face the data provenance is more complex, as in Figure 3.1(b), the estimates from outcome model based on these weights are not guaranteed to be consistent. Towards ensuring as thorough control as possible, we propose a novel framework for selection bias based on modularizing the data provenance into a set of sub-mechanisms, each corresponding to a decision made by a patient and/or health care provider. To illustrate this, Figure 3.2(b) highlights the fact that for a patient to have 'complete' BMI data at 5 years they must have: (1) complete baseline information for all potential confounders, (2) been enrolled in the health plan/system at 5 years; and (3) had a BMI measurement recorded at 5 years. Intuitively, each of these requirements can be thought of as a distinct 'sub-decision' in the flow of decisions that make up the data provenance. Having performed this modularization, researchers bene- fit in a number of important ways. First, in collaboration with subject-matter experts, the consideration of missing data assumptions can be tailored to each sub-mechanism. It may be, for example, that MAR is plausible for one or two of the sub-mechanisms but MNAR suspected for the other(s). Second, statistical models can be built for each sub-mechanism, giving researchers the flexibility to consider different link functions, different sets of covariates and explicitly accommodate the fact that relevant covariates may vary over time (see Figure 3.1). Finally, sensitivity analyses can be tied specifically to those sub-mechanisms for which MNAR is suspected. Note that for the purposes of this paper, we focus on monotone data provenances. That is, in order for a patient to have 'complete' data at any particular sub-mechanism, the patient must have complete data for all prior sub-mechanisms. This restriction to monotone data provenances has been used by Robins et al[61] to perform approaches to handle missing data for repeated outcomes. Collectively, these benefits exploit the more realistic representation of data provenance that Figure 3.2(b) provides so that researchers can expect corresponding adjustments to be superior to those based on Figure 3.2(a)[49] .

3.3 Estimation and Inference

Returning briefly to the complete data setting, suppose, given a sample of complete data of size N, interest lies in estimating the components of a model for

$$\mu_i = E\left[Y_i | X_{1i}, X_{2i}, X_{3i}; \beta\right] = g^{-1}\left(X_{1i}, X_{2i}, X_{3i}; \beta\right), \tag{3.1}$$

where Y is the outcome of interest (BMI at 5 years), X_1 is a vector of exposures of interest (type of bariatric surgery), X_2 is a vector of potential confounders or effect modifiers that *are not* associated with selection, X_3 is a vector of potential confounders or effect modifiers that *are* associated with selection, and β is a vector of unknown regression coefficients. Interest may lie in main effects, namely the association between X_1 and Y, conditional on potential confounders, or in both main effects and effect modification, where the form of g allows for interactions between X_1 and potential effect modifiers.

Given complete data on the N individuals in the EHR, estimation could proceed by solving

$$\sum_{i=1}^{N} D_i^T V_I^{-1}\left(Y_i - \mu_i\right) = \sum_{i=1}^{N} U_i\left(\beta\right) = 0,$$

where $D_i = \partial\mu_i/\partial\beta$, $V_i = Var\left[Y_i | X_{1i}, X_{2i}, X_{3i}\right]$, and U_i is the i^{th} individual's contribution to score equation for β. Given complete data on solely $n < N$ patients, the analysis could proceed in one of several ways, inverse probability weighting (IPW), including saturated models, multiple imputation (MI) and pattern mixture models (PMM)[60,67].

3.3.1 Standard IPW

Suppose now that we wish to estimate the coefficients in the outcome model (1) above, but we are faced with incomplete data due to selection. We can proceed by using IPW to account for selection in the outcome model. Let $\pi_i \equiv P(R_i = 1)$ be the probability of having complete data, X_3 be a vector of potential confounders that affects selection as noted above, and X_4 be a vector of covariates associated with selection and with outcome, Y, but not with the main exposure, X_1. Note that this implies X_4 is not a confounder of the association between X_1 and Y when the data are complete. A consistent estimator of β is obtained by solving the usual weighted estimating equation

$$\sum_{i=1}^{N} R_i \hat{\pi}_i^{-1} U_i^{\star}(\alpha, \beta) = 0,$$

where $U_i^{\star}$ is the i^{th} individual's contribution to the score equation for weighted outcome model and $\hat{\pi}_i$ is consistent for π_i. We can take a standard IPW approach to estimate π_i[61]. That is, we specify a single regression model

$$\pi_i = g^{-1}\left(Z_i; \alpha\right),$$

where $Z = \{X_1, X_2, X_3\} \subset X$ is the vector of all covariates relevant to whether or not complete data is observed, some of which may be main effects or confounders in out main outcome model, and $g\quad(\cdot)$ is a single link function. We would proceed by substituting estimates of α into our estimating equation:

$$\sum_{i=1}^{N} R_i \left[g^{-1}\left(Z_i; \tilde{\alpha}\right)\right]^{-1} U_i^{\star}(\alpha, \beta) = 0, \tag{3.2}$$

which can be solved to obtain $\tilde{\beta}$, an estimate of β. The literature has shown that if $\pi_i = g^{-1}(Z_i;\alpha)$ is correctly specified, $\tilde{\beta}$ is a consistent estimator for β with a Multivariate Normal (MVN) asymptotic sampling distribution[60].

3.3.2 Modularized IPW

In most EHR-based settings, as argued in section 3.3, a single mechanism will be overly simplistic and insufficient to characterize the data provenance. Suppose instead that $\pi_i \neq g^{-1}(Z_i;\alpha)$, but rather the underlying data provenance has K distinct nested sub-mechanisms. Figure 3.2(b) shows how the data provenance for our bariatric surgery study can be decomposed into three sub-mechanisms. We outline an approach to perform IPW estimation when the data provenance is more complex than a single mechanism.

Model Specification

A key initial step to performing estimation is to specify the model for the data provenance. Essentially we want to translate the data provenance, like Figure 3.2(b), into a statistical framework, and we will make decisions about how to structure this framework based on domain knowledge. For each sub-mechanism, define $R_{k,i} = \{0,1\}$ as an indicator for a 'positive' state required for data to be complete. In this setting, an individual has complete data if and only if all $R_{k,i} = 1$, so we would then consider $\pi_i = P(R_{1,i} = 1, ..., R_{K,i} = 1)$.

There are many ways we could structure the K sub-mechanisms, and the interplay of these sub-mechanisms is vital. If they are independent, we can model each one separately. However, if they are nested, we would need to model them sequentially. In this paper, we focus on the situation where the decomposition of the data provenance is monotone, which implies $P\left(R_i^{(k)}\right) > 0$ only if $R_i^{(1)} = \ldots = R_i^{(k-1)} = 1$ for $k = 2, \ldots, K$. Thus, we can write

$$\pi_i = \prod_{k=1}^{K} \pi_{k,i},$$

where $\pi_{k,i} = P(R_{k,i} = 1 |\, R_{1,i} = \ldots = R_{k-1,i} = 1)$. This aligns directly with the data provenance in Figure 3.2(b) where, for example, a BMI measurement at 5 years $(R_{3,i})$ is only relevant if a subject is still enrolled at 5 years $(R_{2,i} = 1)$.

In addition to the specification of the structure of these sub-mechanisms, it is important to decide how to model each particular sub-mechanism. For instance, a generalized linear model (GLM) may be most appropriate to model the binary outcome of a positive or negative state. More specifically, one could specify a logistic or probit regression model for $\pi_i^{(k)}$:

$$\pi_{k,i} = g^{-1}{}_{,k}\left(Z_{k,i}\alpha_k\right),$$

where $Z_k \subset Z$ is is a vector of covariates relevant to whether or not complete data is observed and $g_{\ ,k}(\cdot)$ is the chosen link function for sub-mechanism k.

Some mechanisms fall more naturally into a survival model, especially if folks are censored. For example, in Figure 3.2(b), patients need to remain enrolled intheir healthcare plan until 5 years. We may instead wish to model time to disenrollment for a mechanism like this, in which case $g_{\ ,k}$ is defined as $P(T_i > t_k)$, where t_k is the pre-specified enrollment requirement. This can be done using a fully parametric model, such as an exponential model, where

$$\pi_{k,i} = g^{-1}{}_{,k}\left(Z_{k,i}\alpha_k\right) = \exp\left\{-t_k/Z_{k,i}^T\alpha_k\right\}.$$

Alternatively, we could take a semi-parametric approach, such as estimating a Cox proportional hazards model, then using a Breslow estimator for the baseline hazard[68]. In this case,

$$\pi_{k,i} = g^{-1}{}_{,k}\left(Z_{k,i}\alpha_k\right) = S_{0,k}(t_k)^{\exp\left\{Z_{k,i}^T \quad {}_k\right\}},$$

where $S_{0,k}(t_k)$ is the unknown baseline survival function for mechanism k, evaluated at the pre-specified time t_k. For notational convenience, we will denote $\vartheta_k = \alpha_k$ if the model for sub-mechanism k is fully parametric and $\vartheta_k = (\alpha_k, S_{0,k}(t_k))$ if the model for sub-mechanism k is a semi-parametric Cox proportional hazards model.

Estimation

Based off the specified data provenance and models for the sub-mechanisms, we can directly substitute estimates for $\vartheta^{(k)}$ to obtain

$$\hat{\pi}_{k,i} = g^{-1}{}_{,k}\left(Z_{k,i}; \hat{\vartheta}_k\right).$$

Substituting estimates of all relevant $\vartheta_k \in \vartheta$ yields

$$\sum_{i=1}^{N} R_i \left[\prod_{k=1}^{K} \hat{\pi}_{k,i}\right]^{-1} U_i^{\star}(\vartheta, \beta) = 0, \tag{3.3}$$

which can be solved to obtain $\hat{\beta}$, an estimate of β.

Asymptotics

As shown by van der Vaart[69], under typical regularity conditions, solutions of unbiased estimating equations are asymptotically linear, have a unique influence function, and have a normal limiting distribution. If our models for $\pi_{k,i}$ are correctly specified, the estimating equation (3) is unbiased, and therefore $\hat{\beta}$ is consistent for β with a normal limiting distribution.

More specifically, if $\beta^{\star}$ is the true value of β and $\vartheta_k^{\star}$ is the true value of ϑ_k, we have

$$\sqrt{N}\left(\hat{\beta} - \beta^{\star}\right) \rightarrow \mathit{Normal}(0, \Omega),$$

Hence, we can write

$$\sqrt{N}\left(\hat{\beta} - \beta^{\star}\right) \; \rightarrow \; \text{Normal}(0, \Omega)$$

where $\Omega \; = J^{-1}\Gamma J^{-1}$ and $\Gamma \; = \; \mathit{Var}\left[U^{\star}(\vartheta^{\star}, \beta^{\star}) - Q_1 I_1^{-1} M_1(\vartheta_1^{\star}) - \ldots - Q_K I_K^{-1} M_K(\vartheta_K^{\star})\right]$, the components of which are defined as follows:

$$
\begin{aligned}
J &= \frac{\partial}{\partial \beta} E\left[U_i^{\star}\left(\vartheta^{\star}, \beta\right)\right]\Bigg|_{\quad = \quad {}^{\star}} \\
Q_k &= \frac{\partial}{\partial \vartheta_k} E\left[U_i^{\star}\left(\vartheta, \beta^{\star}\right)\right]\Bigg|_{\quad = \quad {}^{\star}} \\
I_k &= \frac{\partial}{\partial \vartheta_k} E\left[M_{k,i}\left(\vartheta_k\right)\right]\Bigg|_{\quad k= \quad {}^{\star}_{k}},
\end{aligned}
$$

and $M_{k,i}$ is the i^{th} individual's contribution to the score equation for ϑ_k. The full Taylor series expansion to derive these equations can be found in the appendix.

Inference in Practice

In practice, we will need to estimate the components defined in the asymptotic distribution of $\hat{\beta}$. We can rely on Slutsky's theorem[69] and use plug-in estimators for these quantities to obtain estimates of the asymptotic variance-covariance matrix of $\hat{\beta}$. After performing regressions to find all appropriate estimates $\hat{\vartheta}_k$ for $k = 1, \ldots, K$, and $\hat{\beta}$, we can estimate $\hat{\Omega} = \hat{J}^{-1}\hat{\Gamma}\hat{J}^{-1}$, where

$$\begin{aligned}
\hat{J} &= \frac{1}{\sum_i^N R_i} \sum_{i=1}^{N} \frac{\partial}{\partial \beta} U_i\left(\vartheta_k, \beta\right)\Bigg|_{\;=\,\hat{}} \\
\hat{Q}_k &= \frac{1}{\sum_i^N R_i} \sum_{i=1}^{N} \frac{\partial}{\partial \beta} U_i\left(\vartheta_k, \beta\right)\Bigg|_{\;k=\,\hat{}\;\;k} \\
\hat{I}_k &= \frac{1}{\sum_i^N R_i} \sum_{i=1}^{N} \left\{ \frac{\partial}{\partial \vartheta_k} M_{k,i}\left(\vartheta_k\right)\Bigg|_{\;k=\,\hat{}\;\;k} \right\} \\
\hat{\Gamma} &= \frac{1}{\sum_i^N R_i} \left[U_i\left(\hat{\vartheta}_k, \hat{\beta}\right) - \sum_{k=1}^{K} \hat{Q}_k \hat{I}_k^{-1} M_k\left(\hat{\vartheta}_k\right) \right]
\end{aligned}$$

Explicit equations for these are provided in the appendix for logistic, exponential survival, and Cox proportional hazards sub-mechanisms.

3.3.3 Saturated Models

One may also attempt to remove selection bias in the main outcome model by adjusting for all covariates associated with selection[59]. This can be done by fitting a saturated main outcome model that includes all covariates of interest (either exposures, potential confounders, or effect modifiers) as well as covariates that affect selection. Rather than fitting a model for equation (1) above, investigators would instead fit a model for

$$\mu_i = E\left[Y_i | X_{1i}, X_{2i}, X_{3i}, X_{4i}; \beta\right] = g^{-1}\left(X_{1i}, X_{2i}, X_{3i}, , X_{4i}; \beta\right). \tag{3.4}$$

When g is the identity link function, this approach can provide unbiased point estimates for regression coefficients without the need for inverse probability weighting. However, this is not guaranteed for other link functions. For example, when the main outcome model requires a logistic link function, estimates for β from fitting a model for (4) do not necessarily converge to the estimates we would have received by fitting a model for (1) when there are no missing data. This is a result of the non-collapsibility of the odds ratio. Ultimately, fitting saturated outcome models can remove selection bias in some scenarios, but is flawed for generalized linear models.

3.4 Simulations

In most EHR-based settings, $\hat{\beta}$, the regression coefficient of interest when estimated using a modularized IPW, will be expected to exhibit less bias than $\tilde{\beta}$, the regression coefficient of interest when estimated using a modularized IPW. However, given the extra models required for estimation of $\hat{\beta}$, one would also expect its (asymptotic) variance to be larger. As such, researchers will likely contend with a bias-variance trade-ff. To illustrate this, we have conducted a simulation study examining the performance of these estimators, as well as other commonly used approaches, in a number of settings where we may expect to see bias.

3.4.1 Data generation

Figure 3.3 shows three different true data structures we consider in our simulation.

(a) No Effect Modification

(b) Effect Modification of X_3 on $X_1 \rightarrow Y$

(c) Effect Modification of X_4 on $X_1 \rightarrow Y$

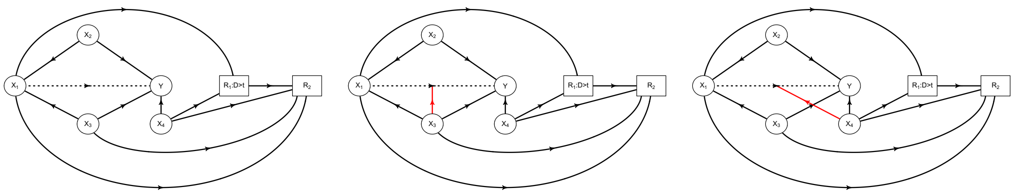


Figure 3.3: DAGs showing data structure for hypothetical study

Consistent with the DURABLE study, X_1 is a binary exposure of interest, Y is a continuous outcome of interest, X_2 is a confounder of the association between X_1 and Y that is not associated with selection, X_3 is a confounder of the association between X_1 and Y that is associated with selection, and X_4 is a covariate associated with selection that is not independently associated with X_1. We consider two selection mechanisms: S_1 requires survival to a certain time t, and S_2 requires observance, given S_1. Figure 3.3(a) shows precisely this data structure in a directed acyclic graph (DAG). In 3(b) we extend this by allowing X_3 to modify the association between X_1 and Y, and in 3(c) we extend this by allowing X_4 to modify the association between X_1 and Y.

Throughout, we considered $N = 10,000$ to examine only large sample properties, with Y, S_1 and S_2 simulated using the following data generating mechanisms:

(*a*) DGM-1: No Effect Modification

$$Y_i \sim N\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i},\ \sigma_y^2\right)$$

$$S_{1,i} = I(D_i > t),\ \ D \sim \mathit{Exp}\left(\alpha_{1,0} + \alpha_{1,1} X_{1i} + \alpha_{1,4} X_{4i}\right)$$

$$S_{2,i} \sim \mathit{Bern}\left(\pi = \mathit{expit}\left\{\alpha_{2,0} + \alpha_{2,1} X_{1i} + \alpha_{2,3} X_{3i} + \alpha_{2,4} X_{4i}\right\}\right)$$

(*b*) DGM-2: Effect Modification of X_3 on $X_1 \rightarrow Y$

$$Y_i \sim N\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_{1,3} X_{1i} X_{3i},\ \sigma_y^2\right)$$

$$S_{1,i} = I(D_i > t),\ \ D \sim \mathit{Exp}\left(\alpha_{1,0} + \alpha_{1,1} X_{1i} + \alpha_{1,4} X_{4i}\right)$$

$$S_{2,i} \sim \mathit{Bern}\left(\pi = \mathit{expit}\left\{\alpha_{2,0} + \alpha_{2,1} X_{1i} + \alpha_{2,3} X_{3i} + \alpha_{2,4} X_{4i}\right\}\right)$$

(*c*) DGM-3: Effect Modification of X_4 on $X_1 \rightarrow Y$

$$Y_i \sim N\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_{1,4} X_{1i} X_{4i},\ \sigma_y^2\right)$$

$$S_{1,i} = I(D_i > t),\ \ D \sim \mathit{Exp}\left(\alpha_{1,0} + \alpha_{1,1} X_{1i} + \alpha_{1,4} X_{4i}\right)$$

$$S_{2,i} \sim \mathit{Bern}\left(\pi = \mathit{expit}\left\{\alpha_{2,0} + \alpha_{2,1} X_{1i} + \alpha_{2,3} X_{3i} + \alpha_{2,4} X_{4i}\right\}\right)$$

(*d*) DGM-4: No Effect Modification and logistic link function

$$Y_i \sim \mathit{Bern}\left(\pi = \mathit{expit}\left\{\beta_0^\dagger + \beta_1^\dagger X_{1i} + \beta_2^\dagger X_{2i} + \beta_3^\dagger X_{3i} + \beta_4^\dagger X_{4i}\right\}\right)$$

$$S_{1,i} = I(D_i > t),\ \ D \sim \mathit{Exp}\left(\alpha_{1,0} + \alpha_{1,1} X_{1i} + \alpha_{1,4} X_{4i}\right)$$

$$S_{2,i} \sim \mathit{Bern}\left(\pi = \mathit{expit}\left\{\alpha_{2,0} + \alpha_{2,1} X_{1i} + \alpha_{2,3} X_{3i} + \alpha_{2,4} X_{4i}\right\}\right)$$

X_1 and X_2 are binary variables generated with probability 0.5, and X_3 and X_4 are generated from zero-mean normal distributions. In all settings, we set $\sigma_y^2 = 10, \beta : \left(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_{13}, \beta_{14}\right) = (0, 2, 1, -1, -1, -2, -2), \alpha_1 = (\alpha_{1,0}, \alpha_{1,1}, \alpha_{1,4}) = (log(30), -1.5, 0.5), \alpha_2 = (\alpha_{2,0}, \alpha_{2,1}, \alpha_{2,3}, \alpha_{2,4}) = (1.5, 2, -1, -1), t = 3$, and $\beta^\dagger : \left(\beta_0^\dagger, \beta_a^\dagger, \beta_1^\dagger, \beta_2^\dagger, \beta_3^\dagger\right) = (0, 1, .5, -.5, -1)$. Using these covariates, 59% of all observations are considered complete, and 41% are excluded due to incomplete data.

3.4.2 Outcome models

In a research study, investigators may be interested in estimating different mean models. In some cases, they will intend to fit the a regression model that accounts for any effect modification main association by including interaction terms. In other cases, however, they may be more interested in estimating the marginal association between the primary exposure, X_1, and the outcome, Y. Although this is, in some sense, a misspecfication of the mean model, investigators may be interested in the overall effect in the population and therefore still be interested in fitting a marginal model. Thus, For each true data setting, we different potential mean models that may be of interest to a researcher:

$$E[Y|X_1, X_2, X_3] = \beta_0 + \beta_1^{(M1)} X_1 + \beta_2 X_2 + \beta_3 X_3 \tag{$M1$}$$

$$E[Y|X_1, X_2, X_3] = \beta_0 + \beta_1^{(M2)} X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{13} X_1 X_3 \tag{$M2$}$$

$$E[Y|X_1, X_2, X_3, X_4] = \beta_0 + \beta_1^{(M3)} X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{14} X_1 X_4 \tag{$M3$}$$

$$E[Y|X_1, X_2, X_3] = expit\left\{\beta_0 + \beta_1^{(M4)\dagger} X_1 + \beta_2^\dagger X_2 + \beta_3^\dagger X_3 + \beta_3^\dagger X_3\right\} \tag{$M4$}$$

When there is no effect modification in the true data generation (i.e. Figure 3.3(a)), only the marginal model $\mathcal{M}1$ is fit. When X_3 is an effect modifier of the association between X_1 and Y (i.e. Figure 3.3(b)), both the marginal model $\mathcal{M}1$ and the interaction model $\mathcal{M}3$ are fit. When X_4 is an effect modifier of the association between X_1 and Y (i.e. Figure 3.3(c)), both the marginal model $\mathcal{M}1$ and the interaction model $\mathcal{M}3$ are fit. When the main outcome model is logistic (DGM-4), only the marginal model $\mathcal{M}4$ is fit.

3.4.3 Accounting for incomplete data

For each of the settings and outcome models outlined above, we consider a series of approaches to account to incomplete data. For each simulation, we generated exposure and outcome data as outlined above, then generated missing indicators based on the specification for α_1 and α_2. We fit the appropriate outcome models first using the full data (i.e. had we observed complete data for all subjects), then compared the estimated coefficients to those resulting from the same regression models using the following approaches using the incomplete data:

- Complete case - include only subjects with complete data
- Complete case with a saturated model - include only subjects with complete data, but additionally include all covariates associated with selection in the main outcome model
- Standard IPW - estimate IP weights using a single missingness mechanism (logistic model for S)
- Modularized IPW - estimate IP weights using two missingness mechanisms (exponential model for S_1 and logistic model for S_2)
- Modularized IPW (Cox) - estimate IP weights using two missingness mechanisms (Cox proportional hazards model for S_1 and logistic model for S_2)

- Modularized IPW (logistic) - estimate IP weights using two missingness mechanisms (logistic model for S_1 and logistic model for S_2)

3.4.4 Simulation Results

Estimation

Percent bias and standard errors for the estimated regression coefficients for the association of interest are presented in Table 3.2. Results for all three data generating mechanisms are provided, including the mean estimated coefficient, as well as the percent bias observed using each approach to account for incomplete data, relative to the estimate using the full data. We see that under all data generation settings, the correctly specified modularized IPW using exponential model for S_1 is unbiased, but the complete case, complete case with a saturated model, standard IPW, and modularized IPW with a misspecified logistic link are prone to bias. The modularized IPW with a Cox PH model for S_1 is also unbiased in all settings, which is expected as the true survival model for S_1 is exponential, which is a proportional hazards model. When there is no effect modification, the complete case analysis with a saturated model is unbiased; however, this is insufficient to adjust for any selection bias in the marginal model when there is effect modification, as seen for DGM-2 (M1) and DGM-3 (M11). We see that for all approaches, the bias in the estimated regression coefficient for the interaction term (β_{13} or β_{14}) is either minimal or zero. When the main outcome model is logistic, DGM-4, the resulting biases reflect the same pattern as those for DGM-1 (M1), with the exception of the complete case with a saturated model, which no longer provides an unbiased coefficient. This is due to the non-collapsibility of the odds ratio.

Across all data generating mechanisms and models, we see the modularized IPW approach is less efficient than both the standard IPW approach and either of the complete case approaches. This is to be expected, as we are fitting a larger number of models and we expect additional variability. However, the modularized IPW is the only approach that is consistently unbiased across all data generating mechanism and models. Thus, we observe a bias-variance trade-off in appropriately adjusting for selection in this setting.

Inference

To evaluate the validity of the inference on these approaches, we calculated the coverage probabilities for 95% confidence intervals for the coefficients of interest using each approach to adjust for selection. Overall, the modularized IPW using either an exponential model or a Cox model achieved the nominal coverage probability in all scenarios. The complete case, standard IPW, and modularized IPW with a logistic model approaches only achieved the nominal coverage probability when the estimate was unbiased. Full results can be found in the supplement.

Overall, we see that both the complete case analysis standard IPW analysis are biased. The modularized IPW using either an exponential model or a Cox PH for S_1 is always unbiased, however using a Cox PH model for S_1 seems to have slightly lower confidence interval coverage. The complete case analysis with a saturated model is unbiased with appropriate coverage for the marginal mean model (M1) when there is no effect modification and for the interaction models (M2) and (M3), but is biased with poor coverage for the marginal mean model (M1) when there is effect modification and is biased for the marginal mean model (M4) when the main outcome model is logistic. Here we see an example of a scenario in which the standard framework fails to capture the complexity of the data provenance and subsequently does not adequately correct for selection bias in the final model.

Table 3.2. Estimated coefficients of X_1

Model	Coefficient	Approach	% Bias	Relative Std Error
DGM-1 (M1)	$\beta_1^{(M1)}$	Complete case	-28	0.79
		Complete case with saturated model	0	0.74
		Standard IPW	-29	0.80
		Modularized IPW with exponential S_1	0	1.00
		Modularized IPW with Cox PH S_1	0	1.00
		Modularized IPW with logistic S_1	-4	0.92
DGM-2 (M1)	$\beta_1^{(M1)}$	Complete case	-12	0.79
		Complete case with saturated model	17	0.75
		Standard IPW	-28	0.81
		Modularized IPW with exponential S_1	0	1.00
		Modularized IPW with Cox PH S_1	0	1.00
		Modularized IPW with logistic S_1	-4	0.93
DGM-2 (M2)	$\beta_1^{(M2)}$	Complete case	-29	0.72
		Complete case with saturated model	0	0.68
		Standard IPW	-29	0.74
		Modularized IPW with exponential S_1	0	1.00
		Modularized IPW with Cox PH S_1	0	1.00
		Modularized IPW with logistic S_1	-4	0.94
DGM-2 (M2)	β_{13}	Complete case	-4	0.72
		Complete case with saturated model	0	0.68
		Standard IPW	-4	0.74
		Modularized IPW with exponential S_1	0	1.00
		Modularized IPW with Cox PH S_1	0	1.00
		Modularized IPW with logistic S_1	-4	0.94

Table 3.2 (continued). Estimated coefficients of X_1

Model	Coefficient	Approach	% Bias	Relative Std Error
DGM-3 (M1)	$\beta_1^{(M1)}$	Complete case	-59	0.62
		Complete case with saturated model	-5	0.52
		Standard IPW	-58	0.65
		Modularized IPW with exponential S_1	0	1.00
		Modularized IPW with Cox PH S_1	0	1.00
		Modularized IPW with logistic S_1	-11	0.80
DGM-3 (M3)	$\beta_1^{(M3)}$	Complete case	0	0.89
		Complete case with saturated model	0	0.89
		Standard IPW	0	0.91
		Modularized IPW with exponential S_1	0	1.00
		Modularized IPW with Cox PH S_1	0	1.00
		Modularized IPW with logistic S_1	0	0.97
DGM-3 (M3)	β_{14}	Complete case	0	0.72
		Complete case with saturated model	0	0.72
		Standard IPW	0	0.73
		Modularized IPW with exponential S_1	0	1.00
		Modularized IPW with Cox PH S_1	0	1.00
		Modularized IPW with logistic S_1	0	0.89
DGM-4 (M4)	$\beta_1^{(M4)}$	Complete case	-55	0.88
		Complete case with saturated model	-35	1.00
		Standard IPW	-55	1.02
		Modularized IPW with exponential S_1	0	1.00
		Modularized IPW with Cox PH S_1	0	0.97
		Modularized IPW with logistic S_1	-6	0.90

3.5 Results for DURABLE Study

Characteristics of N=16,282 patients are identified in DURABLE as having undergone bariatric surgery between 1997 and 2010 are shown in Table 3.1. Patients who have complete baseline covariates (BMI at both 6 months prior to surgery and at the time of surgery), are enrolled at five years, and have a BMI measurement at five years are considered to have 'complete' data. Characteristics of these n=5,636 patients are also shown in Table 3.1. As we have seen through extensive simulation, naive application of standard missing data methods/strategies will generally result in inadequate adjustment for selection bias. To evaluate the impact of choosing different methods to account for selection in these data, we proceed by implementing three approaches to estimate the effect of bariatric surgery type on BMI 5 years after surgery: (1) complete case analysis, (2) standard IPW with a single logistic model for complete data, and (3) modularized IPW with a logistic model for complete baseline covariates, a Cox proportional hazards model for enrollment at 5 years given complete baseline covariates, and a logistic model for BMI information at 5 years given complete baseline covariates and enrollment at 5 years. Results for the selection models are shown in Table 3.3 for both a single-mechanism (standard IPW) approach where the probability of having complete data is estimated using a single logistic model and the modularized IPW described above.

Table 3.3. Estimated regression coefficients for selection models

	Coefficient α (p-value)							
	Standard IPW		Modularized IPW					
	S logistic		S_1 logistic		S_2 coxPH		S_3 logistic	
Covariate	N = 16,282		N = 16,282		N = 10,661		N = 8,332	
VSG vs RYGB	-0.43	(0.53)	-0.74	(0.29)	0.92	(0.12)	0.67	(0.56)
Standardized age at surgery	0.08	(<0.01)	0.04	(<0.01)	-0.08	(<0.01)	0.05	(<0.01)
Year of surgery, years	0.32	(<0.01)	0.86	(<0.01)	0.01	(0.064)	-0.13	(<0.01)
Prior enrollment years	0.05	(<0.01)	0.002	(0.69)	-0.08	(<0.01)	0.02	(<0.01)
Male	-0.28	(<0.01)	-0.13	(0.04)	0.22	(<0.01)	-0.22	(<0.01)
Site								
KNC vs KW	-0.06	(0.36)	-0.06	(0.49)	0.05	(0.56)	-0.01	(0.96)
KSC vs KW	0.43	(<0.01)	0.69	(<0.01)	0.01	(0.87)	0.42	(<0.01)
Standardized baseline BMI	–		–		0.03	(0.05)	-0.04	(0.03)
BMI pre slope	–		–		0.01	(0.35)	-0.02	(0.28)
Type * Site								
VSG & KNC	0.80	(0.28)	0.38	(0.63)	-1.23	(0.07)	-0.13	(0.91)
VSG & KSC	0.28	(0.69)	0.90	(0.21)	-1.17	(0.05)	-0.75	(0.52)

We see that several potential predictors of complete data have differential effects when we compare the standard IPW approach to the modularized approach. For instance, men are less likely to have complete data according to the standard IPW model (p<0.01), but are less likely to have complete baseline data (p<0.01), more likely to be enrolled at five years given complete baseline data (p<0.01), and less likely to have a BMI measurement at 5 years given complete baseline data and enrollment (p<0.01). The simplified standard IPW fails to capture the nuanced association between gender and complete data that is highlighted in the modularized IPW. We can further see the impact of using the standard IPW by examining the predicted probabilities of complete data among those with complete data (Figure 3.4).

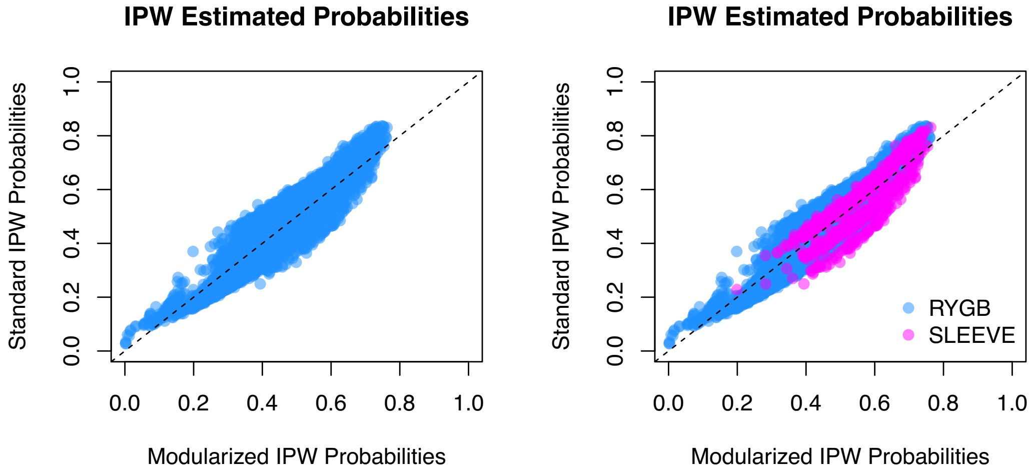


Figure 3.4: Comparison of estimated probabilities of complete data

Here we see that while the probabilities predicted using the standard IPW approach and the modularized IPW approach are highly correlated, they are not identical. The standard IPW approach is more likely to underestimate the probabilities of complete data relative to the modularized approach for modularized IPW probabilities close to 0.5, and it is more likely to overestimate probabilities closer to 0 and 1.

We consider three main outcome models relating bariatric surgery type to BMI at 5 years: (1) BMI slope (defined as BMI at five years - BMI at baseline), (2) BMI slope adjusted for potential confounders, and (3) BMI slope adjusted for potential confounders and an interaction with gender. Results for these outcome regression models examining the association between surgery type and BMI at 5 years are shown in Table 3.4.

Table 3.4. Estimated regression coefficient (95% confidence interval) for VSG vs RYGB

Outcome Model	Complete Case		Standard IPW		Modularized IPW	
BMI 5 year slope (unadjusted)	2.94	(2.52, 3.36)	2.85	(2.59, 3.11)	3.26	(2.79, 3.74)
BMI 5 year slope (adjusted*)	3.17	(2.73, 3.60)	2.98	(2.68, 3.27)	3.43	(2.94, 3.91)
Male	0.78	(0.38, 1.19)	0.70	(0.41, 0.99)	0.61	(0.28, 0.93)
KNC	3.26	(2.53, 3.78)	2.91	(2.43, 3.39)	2.52	(1.61, 3.44)
KSC	2.11	(1.53, 2.70)	2.07	(1.63, 2.51)	1.73	(0.92, 2.53)
Year of surgery	0.15	(0.05, 0.26)	0.15	(0.05, 0.26)	-0.04	(-0.30, 0.23)
Age (centered)	0.34	(0.27, 0.41)	0.35	(0.30, 0.40)	0.31	(0.19, 0.43)
BMI pre slope	-0.42	(-0.48, -0.36)	-0.43	(-0.49, -0.36)	-0.42	(-0.49, -0.35)
BMI 5 year slope (adjusted**)	3.37	(2.89, 3.84)	3.20	(2.89, 3.51)	3.62	(3.14, 4.10)
BMI 5 year slope * Male	-1.08	(-2.11, -0.06)	-1.06	(-1.71, -0.41)	-0.91	(-1.55, -0.26)
Male	0.99	(0.54, 1.43)	0.85	(0.53, 1.17)	0.73	(0.37, 1.09)
KNC	3.18	(2.54, 3.79)	2.92	(2.44, 3.40)	2.53	(1.61, 3.44)
KSC	2.12	(1.53, 2.70)	2.07	(1.63, 2.51)	1.72	(0.91, 2.53)
Year of surgery	0.15	(0.05, 0.26)	0.15	(0.05, 0.25)	-0.04	(-0.30,0.22)
Age (centered)	0.34	(0.27, 0.41)	0.35	(0.30, 0.40)	0.31	(0.19, 0.43)
BMI pre slope	-0.42	(-0.48, -0.36)	-0.43	(-0.49, -0.36)	-0.42	(-0.49, -0.35)

*Adjusted for gender, site, year of surgery, age at surgery, and prior BMI slope

**Additionally including a gender-BMI 5 year slope interaction

We see that in all three models, the standard IPW estimate is in fact in the opposite direction of the complete case estimate than the modularized IPW estimate is. While most of the potential confounders are in the same direction across all three approaches, the magnitude and the confidence intervals vary. This suggests that the standard IPW approach may be incorrectly estimating the probabilities of complete information and therefore insufficiently correcting for selection bias. Note that in this data provenance, we modeled enrollment using a Cox proportional hazards model. We also considered an exponential survival model and a logistic model to estimate the probability of enrollment at 5 years, and the resulting modularized IPW estimates we very similar. Results using the exponential survival model and the logistic model can be found in the supplement.

3.6 Discussion

In EHR-based studies, data provenance relies on a complex structure determined by many clinical decisions made by patients, health care providers, and the healthcare system. As seen in simulation, the standard IPW approach may fail to capture this in EHR-based research, thus outcome models using an oversimplified data provenance do not adequately correct for selection bias.

This result is particularly noteworthy in a number of scenarios, specifically when the true data provenance is multi-stage with large effect sizes in opposite directions, when researchers are specifically interested in a marginal mean model, but the main association is modified by a covariate that is associated with selection, and when the functional form of at least one of the data provenance mechanisms is misspecified. We have shown through simulation that in a scenario such as this, we find biased coefficient estimates and invalid inference with reduced coverage for confidence intervals. It is important to not that this will not be the case in all scenarios, however. When true data provenance is multi-stage with smaller effect sizes in the same direction across mechanisms, the standard approach will not necessarily lead to substantial residual selection bias. It may, however, lead to more efficient regression estimators. It is for this reason we urge researchers to illicit guidance from subject-matter experts and professionals in the healthcare system when performing EHR-based research where data are incomplete.

In this paper, we focus on a small set of possible research questions, but there are plenty to consider. Natural extensions of this methodology could be made to longitudinal analyses in which we are interested in the trajectory of an outcome over time. Similarly, this paper only considers monotone missingness mechanisms. In longitudinal data, however, non-monotone missingness is much more common and should be accounted for. Finally, there are other approaches to estimation and inference in the presence of missingness that this paper does not consider, namely multiple imputation or blended strategies. Further work should be done to extend the general framework outlined in this paper to encompass a broader range of scientific questions and approaches to handling missing data.

While the advantages of performing a modularized approach are many; namely the flexibility in specification of the sub-mechanisms and the ability to more closely reflect the true data provenance, there may be a small sacrifice in efficiency relative to a standard approach. All of this must be considered when analyzing EHR data.

4

Supplement

4.1 Chapter 1 Supplement

4.1.1 S.1 $\widetilde{\varrho}_{ab} = 1 \iff a = b$ for compound symmetric covariance matrices in the discrete case

Here we note that $\widetilde{\varrho}_{ab} \to \varrho_{ab}$ as $n \to \infty$, where ϱ_{ab} is the theoretical correlation between $h_i^\star(a, t_i)$ and $h_i^\star(b, t_i)$. We prove that the theoretical correlation between $h_i^\star(a, t_i)$ and $h_i^\star(b, t_i)$ will be 1 if and only if $a = b$, under the assumption that the covariance matrix for $\tilde{Z}_i$ is compound symmetric, which implies the same will be true for the asymptotic empirical correlation $\widetilde{\varrho}_{ab}$. Under a compound symmetric covariance matrix, $Cov\left[Z_i(s_1), Z_i(s_2)\right] = \sigma^2$ when $s_1 = s_2$ and $Cov\left[Z_i(s_1), Z_i(s_2)\right] = \varrho\sigma^2$ when $s_1 \neq s_2$, where $\varrho < 1$.

$$Cov\left[h_i^{\star}(a,t_i),h_i^{\star}(b,t_i)\right]=\varrho_{ab}\sqrt{Var\left[h_i^{\star}(a,t_i)\right]Var\left[h_i^{\star}(b,t_i)\right]}$$

$$\begin{aligned}
Var\left[h_i^{\star}(a,t_i)\right] &= Var\left[\frac{1}{a+1}\sum_{s=t_i-a}^{t_i} Z_i(s)\right]\\
&= \left(\frac{1}{a+1}\right)^2 Var\left[\sum_{s=t_i-a}^{t_i} Z_i(s)\right]\\
&= \left(\frac{1}{a+1}\right)^2 \sum_{s_1=t_i-a}^{t_i}\sum_{s_1=t_i-a}^{t_i} Cov\left[Z_i(s_1),Z_i(s_2)\right]\\
&= \left(\frac{1}{a+1}\right)^2 \left[(a+1)\sigma^2+\left\{(a+1)^2-(a+1)\right\}\sigma^2\varrho\right]\\
&= \left(\frac{\sigma^2}{a+1}\right)\left[1+a\varrho\right]\\
Var\left[h_i^{\star}(b,t_i)\right] &= \left(\frac{\sigma^2}{b+1}\right)\left[1+b\varrho\right]
\end{aligned}$$

$$\begin{aligned}
\sqrt{Var\left[h_i^{\star}(a,t_i)\right]Var\left[h_i^{\star}(b,t_i)\right]} &= \sqrt{\left(\frac{\sigma^2}{a+1}\right)\left[1+a\varrho\right]\left(\frac{\sigma^2}{b+1}\right)\left[1+b\varrho\right]}\\
&= \sigma^2\sqrt{\left(\frac{1+a\varrho}{a+1}\right)\left(\frac{1+b\varrho}{b+1}\right)} \qquad (S.1.1)
\end{aligned}$$

$$\begin{aligned}
Cov\left[h_i^{\star}(a,t_i),h_i^{\star}(b,t_i)\right] &= Cov\left[\frac{1}{a+1}\sum_{s=t_i-a}^{t_i} Z_i(s),\frac{1}{b+1}\sum_{s=t_i-b}^{t_i} Z_i(s)\right]\\
&= \frac{1}{(a+1)(b+1)}Cov\left[\sum_{s=t_i-a}^{t_i} Z_i(s),\sum_{s=t_i-b}^{t_i} Z_i(s)\right]\\
&= \frac{1}{(a+1)(b+1)}\sum_{s_1=t_i-a}^{t_i}\sum_{s_2=t_i-b}^{t_i} Cov\left[Z_i(s_1),Z_i(s_2)\right]\\
&= \frac{1}{(a+1)(b+1)}\left[(min\{a,b\}+1)\sigma^2+\{(a+1)(b+1)\right.\\
&\left.-(min\{a,b\}+1)\}\varrho\sigma^2\right]\\
&= \sigma^2\left(\frac{1+max\{a,b\}\varrho}{max\{a,b\}+1}\right) \qquad (S.1.2)
\end{aligned}$$

Given that $Cov\left[h_i^\star(a,t_i), h_i^\star(b,t_i)\right] = \varrho_{ab}\sqrt{Var\left[h_i^\star(a,t_i)\right] Var\left[h_i^\star(b,t_i)\right]}$, we can compare equations S.1.1 and S.1.2 above:

$$\begin{aligned}
Cov\left[h_i^\star(a,t_i), h_i^\star(b,t_i)\right] - \sqrt{Var\left[h_i^\star(a,t_i)\right] Var\left[h_i^\star(b,t_i)\right]} &= \sigma^2\left(\frac{1+max\{a,b\}\varrho}{max\{a,b\}+1}\right) - \sigma^2\sqrt{\left(\frac{1+a\varrho}{a+1}\right)\left(\frac{1+b\varrho}{b+1}\right)} \\
&= \sigma^2\left[\left(\frac{1+max\{a,b\}\varrho}{max\{a,b\}+1}\right) - \sqrt{\left(\frac{1+a\varrho}{a+1}\right)\left(\frac{1+b\varrho}{b+1}\right)}\right]
\end{aligned}$$

If $a = b$, $\left(\frac{1+max\{a,b\}}{max\{a,b\}+1}\right) - \sqrt{\left(\frac{1+a}{a+1}\right)\left(\frac{1+b}{b+1}\right)} = 0$, which implies $\varrho_{ab} = 1$. However, if $a \neq b$, $\left(\frac{1+max\{a,b\}}{max\{a,b\}+1}\right) - \sqrt{\left(\frac{1+a}{a+1}\right)\left(\frac{1+b}{b+1}\right)} \neq 0$, which implies $\varrho_{ab} \neq 1$.

S.2 $\widetilde{\varrho}_{ab} = 1 \iff a = b$ for AR(1) covariances matrices in the discrete and continuous cases

Under an AR(1) covariance matrix, $Cov\left[Z_i(s_1), Z_i(s_2)\right] = \varrho^{|s_1 - s_2|}\sigma^2$ where $\varrho < 1$ and $\vartheta \in [0,1]$. We also note that if $\varrho_{ab} = 1 \iff a = b$ holds in the continuous case, it must also hold for discrete exposure measures. Given that $Cov\left[h_i^\star(a,t_i), h_i^\star(b,t_i)\right] = \varrho_{ab}\sqrt{Var\left[h_i^\star(a,t_i)\right] Var\left[h_i^\star(b,t_i)\right]}$, we argue that $\varrho_{ab} < 1$ if $\sqrt{Var\left[h_i^\star(a,t_i)\right] Var\left[h_i^\star(b,t_i)\right]} < Cov\left[h_i^\star(a,t_i), h_i^\star(b,t_i)\right]$, or equivalently if $Var\left[h_i^\star(a,t_i)\right] Var\left[h_i^\star(b,t_i)\right] < Cov\left[h_i^\star(a,t_i), h_i^\star(b,t_i)\right]^2$. Without loss of generality, we assume $a \leq b$.

$$
\begin{aligned}
Var\left[h_i^{\star}(a,t_i)\right] &= Var\left[\frac{1}{a}\int_{t_i-a}^{t_i} Z_i(s)ds\right] \\
&= \left(\frac{1}{a}\right)^2 Var\left[\frac{1}{a}\int_{t_i-a}^{t_i} Z_i(s)ds\right] \\
&= \left(\frac{1}{a}\right)^2 \int_{t_i-a}^{t_i}\int_{t_i-a}^{t_i} Cov\left[Z_i(s_1), Z_i(s_2)\right] ds_1 ds_2 \\
&= \left(\frac{1}{a}\right)^2 \sigma^2 \int_{t_i-a}^{t_i}\int_{t_i-a}^{t_i} \varrho^{|s_1-s_2|} ds_1 ds_2 \\
Var\left[h_i^{\star}(a,t_i)\right] Var\left[h_i^{\star}(b,t_i)\right] &= \frac{\sigma^4}{a^2b^2}\int_{t_i-a}^{t_i}\int_{t_i-a}^{t_i} \varrho^{|s_1-s_2|} ds_1 ds_2 \times \int_{t_i-b}^{t_i}\int_{t_i-b}^{t_i} \varrho^{|s_1-s_2|} ds_1 ds_2 \\
&= \frac{\sigma^4}{a^2b^2}\int_{t_i-a}^{t_i}\int_{t_i-a}^{t_i} \varrho^{|s_1-s_2|} ds_1 ds_2 \times \left[\int_{t_i-b}^{t_i-a}\int_{t_i-b}^{t_i-a} \varrho^{|s_1-s_2|} ds_1 ds_2 + \right. \\
&\left. + \int_{t_i-a}^{t_i}\int_{t_i-a}^{t_i} \varrho^{|s_1-s_2|} ds_1 ds_2 + 2\int_{t_i-a}^{t_i}\int_{t_i-b}^{t_i-a} \varrho^{|s_1-s_2|} ds_1 ds_2 \right] \qquad (S.2.1)
\end{aligned}
$$

$$
\begin{aligned}
Cov\left[h_i^{\star}(a,t_i), h_i^{\star}(b,t_i)\right] &= Cov\left[\frac{1}{a}\int_{t_i-a}^{t_i} Z_i(s)ds, \frac{1}{b}\int_{t_i-b}^{t_i} Z_i(s)ds\right] \\
&= \frac{1}{ab} Cov\left[\int_{t_i-a}^{t_i} Z_i(s)ds, \int_{t_i-b}^{t_i} Z_i(s)ds\right] \\
&= \frac{1}{ab}\int_{t_i-a}^{t_i}\int_{t_i-b}^{t_i} Cov\left[Z_i(s_1), Z_i(s_2)\right] ds_1 ds_2 \\
&= \frac{\sigma^2}{ab}\int_{t_i-a}^{t_i}\int_{t_i-b}^{t_i} \varrho^{|s_1-s_2|} \quad ds_1 ds_2
\end{aligned}
$$

$$\begin{aligned}
Cov\left[h_i^{\star}(a,t_i),h_i^{\star}(b,t_i)\right]^2 &= \frac{\sigma^4}{a^2b^2}\left[\int_{t_i-a}^{t_i}\int_{t_i-b}^{t_i-a}\varrho^{|s_1-s_2|}ds_1ds_2+\int_{t_i-a}^{t_i}\int_{t_i-a}^{t_i}\varrho^{|s_1-s_2|}ds_1ds_2\right]^2 \\
&= \frac{\sigma^4}{a^2b^2}\left[\left(\int_{t_i-a}^{t_i}\int_{t_i-b}^{t_i-a}\varrho^{|s_1-s_2|}ds_1ds_2\right)^2+\left(\int_{t_i-a}^{t_i}\int_{t_i-a}^{t_i}\varrho^{|s_1-s_2|}ds_1ds_2\right)^2\right. \\
&\quad \left.+2\int_{t_i-a}^{t_i}\int_{t_i-b}^{t_i-a}\varrho^{|s_1-s_2|}ds_1ds_2\int_{t_i-a}^{t_i}\int_{t_i-a}^{t_i}\varrho^{|s_1-s_2|}\right] \qquad (S.2.2)
\end{aligned}$$

4.2 Chapter 3 Supplement

Supplemental Table 1. Coverage Probabilities of 95% Confidence Intervals for Coefficients of X_1

Model	Coefficient	Approach	95% CI
DGM-1 (M1)	$\beta_1^{(M1)}$	Complete case	0.01
		Complete case with saturated model	0.95
		Standard IPW	0.00
		Modularized IPW with exponential S_1	0.96
		Modularized IPW with Cox PH S_1	0.95
		Modularized IPW with logistic S_1	0.80
DGM-2 (M1)	$\beta_1^{(M1)}$	Complete case	0.61
		Complete case with saturated model	0.25
		Standard IPW	0.01
		Modularized IPW with exponential S_1	0.96
		Modularized IPW with Cox PH S_1	0.95
		Modularized IPW with logistic S_1	0.81
DGM-2 (M2)	$\beta_1^{(M2)}$	Complete case	0.01
		Complete case with saturated model	0.95
		Standard IPW	0.00
		Modularized IPW with exponential S_1	0.96
		Modularized IPW with Cox PH S_1	0.95
		Modularized IPW with logistic S_1	0.80
DGM-2 (M2)	β_{13}	Complete case	0.86
		Complete case with saturated model	0.95
		Standard IPW	0.73
		Modularized IPW with exponential S_1	0.95
		Modularized IPW with Cox PH S_1	0.95
		Modularized IPW with logistic S_1	0.86

Supplemental Table 1 (continued). Coverage Probabilities of 95% Confidence Intervals for Coefficients of X_1

Model	Coefficient	Approach	95% CI
DGM-3 (M1)	$\beta_1^{(M1)}$	Complete case	0.00
		Complete case with saturated model	0.88
		Standard IPW	0.00
		Modularized IPW with exponential S_1	0.99
		Modularized IPW with Cox PH S_1	0.95
		Modularized IPW with logistic S_1	0.65
DGM-3 (M3)	$\beta_1^{(M3)}$	Complete case	0.95
		Complete case with saturated model	0.95
		Standard IPW	0.87
		Modularized IPW with exponential S_1	0.94
		Modularized IPW with Cox PH S_1	0.95
		Modularized IPW with logistic S_1	0.86
DGM-3 (M3)	β_{14}	Complete case	0.95
		Complete case with saturated model	0.95
		Standard IPW	0.87
		Modularized IPW with exponential S_1	0.96
		Modularized IPW with Cox PH S_1	0.95
		Modularized IPW with logistic S_1	0.78

Table 8. Estimated regression coefficient (95% confidence interval) for VSG vs RYGB using modularized IPW with different models for enrollment at 5 years

Outcome Model	Exponential		Logistic		Cox PH	
BMI 5 year slope (unadjusted)	3.25	(2.77, 3.73)	3.26	(2.78, 3.73)	3.26	(2.79, 3.74)
BMI 5 year slope (adjusted*)	3.42	(2.93, 3.92)	3.42	(2.93, 3.91)	3.43	(2.94, 3.91)
Male	0.63	(0.30, 0.95)	0.62	(0.29, 0.94)	0.61	(0.28, 0.93)
KNC	2.52	(1.61, 3.43)	2.53	(1.62, 3.44)	2.52	(1.61, 3.44)
KSC	1.72	(0.92, 2.53)	1.73	(0.92, 2.54)	1.73	(0.92, 2.53)
Year of surgery	-0.04	(-0.30, 0.23)	-0.04	(-0.30, 0.23)	-0.04	(-0.30, 0.23)
Age (centered)	0.31	(0.18, 0.43)	0.31	(0.19, 0.43)	0.31	(0.19, 0.43)
BMI pre slope	-0.42	(-0.50, -0.35)	-0.42	(-0.49, -0.36)	-0.42	(-0.49, -0.35)
BMI 5 year slope (adjusted**)	3.61	(3.12, 4.10)	3.62	(3.14, 4.10)	3.62	(3.14, 4.10)
BMI 5 year slope × Male	-0.91	(-1.55, -0.26)	-0.92	(-1.56, -0.27)	-0.91	(-1.55, -0.26)
Male	0.75	(0.38, 1.11)	0.74	(0.38, 1.10)	0.73	(0.37, 1.09)
KNC	2.52	(1.61, 3.44)	2.53	(1.62, 3.44)	2.53	(1.61, 3.44)
KSC	1.72	(0.91, 2.53)	1.73	(0.92, 2.54)	1.72	(0.91, 2.53)
Year of surgery	-0.04	(-0.30, 0.22)	-0.04	(-0.30, 0.22)	-0.04	(-0.30,0.22)
Age (centered)	0.31	(0.18, 0.43)	0.42	(0.19, 0.43)	0.31	(0.19, 0.43)
BMI pre slope	-0.91	(-1.55, -0.26)	-0.92	(-48, -0.35)	-0.42	(-0.49, -0.35)

*Adjusted for gender, site, year of surgery, age at surgery, and prior BMI slope

**Additionally including a gender-BMI 5 year slope interaction

5

References

[1] Caplan, Richard J., Gary M. Marsh, and Philip E. Enterline. (1983) A Generalized Effective Exposure Modeling Program for Assessing Dose-response in Epidemiologic Investigations. *Computers and Biomedical Research* . 16.6: 587-96.

[2] Salvan, Alberto, Leslie Stayner, Kyle Steenland, and Randall Smith. (1995) Selecting an Exposure Lag Period. *Epidemiology* . 6.4: 387-90.

[3] Hauptmann, Michael, Jay H. Lubin, Philip Rosenberg, J□rgen Wellmann, and Lothar Kreienbrock. (2000) The Use of Sliding Time Windows for the Exploratory Analysis of Temporal Effects of Smoking Histories on Lung Cancer Risk. *Statist. Med. Statistics in Medicine* . 19.16: 2185-194.

[4] Hauptmann, M., K. Berhane, B. Langholz, and Jh Lubin. (2001) Using Splines to Analyse Latency in the Colorado Plateau Uranium Miners Cohort. *Journal of Epidemiology and Biostatistics* . 6.6: 417-24.

[5] Richardson, D. B., S. R. Cole, H. Chu, and B. Langholz. (2011) Lagging Exposure Information in Cumulative Exposure-Response Analyses. *American Journal of Epidemiology*. 174.12: 1416-422.

[6] Heaton MJ, Peng RD. (2014) Extending distributed lag models to higher degrees. *Biostatistics*. (2):398-412.

[7] Wang, Molin, Xiaomei Liao, Francine Laden, and Donna Spiegelman. (2016) Quantifying Risk over the Life Course - Latency, Age-related Susceptibility, and Other Time-varying Exposure Metrics. *Statist. Med. Statistics in Medicine*. epub.

[8] Rosner, B., Spiegelman, D., and Willett, W.C. (1989) Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics In Medicine* . 8:9, 1051-69.

[9] Rosner, B., Spiegelman, D., and Willett, W.C. (1990) Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am J Epidemiol.* 132: 734-745.

[10] Rosner B., Spiegelman D., Willett W.C. (1992) Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *Am J Epidemiol* .136:1400-13.

[11] Carroll R.J., Wand M.P. (1991) Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society Series B-Methodological.* 53:573-585.

[12] Kuha J. (1994) Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Stat Med.* 13:1135-48.

[13] Spiegelman, D., McDermott, A. and Rosner, B. (1997) Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *American Journal of Clinical Nutrition.* 65: 1179s-1186s.

[14] Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006) Measurement Error in Nonlinear Models: A Modern Perspective. *London: Chapman & Hall* .

[15] Puett RC, Hart JE, Yanosky JD, Paciorek C, Schwartz J, Suh H, Speizer FE, Laden F. (2009) Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses' Health Study. *Environmental Health Perspectives.* 117:1697-1701.

[16] Laden F, Schwartz J, Speizer FE, Ochery DW. (2006) Reduction in fine particulate air pollution and mortality: extended follow-up of the harvard six cities study. *American Journal of Respiratory and Critical Care Medicine*. 173:667-672.

[17] Loève, M. (1977). *Probability Theory 1* (4th ed.). Springer Verlag.

[18] Jennrich, RI. (1969) Asymptotic Properties of Non-Linear Least Squares Estimators. *Ann. Math. Statist.* 40.2: 633-643.

[19] Hauptmann, Michael, Hermann Pohlabeln, Jay H. Lubin, Karl-Heinz J□ckel, Wolfgang Ahrens, Irene Br□ske-Hohlfeld, and H.-Erich Wichmann.(2002) The Exposure-time-response Relationship between Occupational Asbestos Exposure and Lung Cancer in Two German Case-control Studies. *Am. J. Ind. Med. American Journal of Industrial Medicine*.41.2: 89-97.

[20] Liao, Xiaomei, David M. Zucker, Yi Li, and Donna Spiegelman. (2010) Survival Analysis with Error-Prone Time-Varying Covariates: A Risk Set Calibration Approach. *Biometrics* .67.1: 50-58.

[21] Heaphy CM, Yoon GS, Peskoe SB, Joshu CE, Lee TK, Giovannucci E, Mucci LA, Kenfield SA, Stampfer MJ, Hicks JL, De Marzo AM, Platz EA, Meeker AK. (2013) Prostate cancer cell telomere length variability and stromal cell telomere length as prognostic markers for metastasis and death. *Cancer Discov.* 3: (10): 1130?41.

[22] Jardim TV, Sousa AL, Povoa TI, Barroso WK, Chinem B, Jardim L, Bernardes R, Coca A, Jardim PC. (2015) The natural history of cardiovascular risk factors in health professionals: 20-year follow-up. *BMC Public Healt*.15.1: 1111.

[23] Kutner MH, Nachtsheim CJ, Neter J, Li W. (2005). *Applied Linear Statistical Models.* Homewood, Ill, R.D. Irwin..

[24] Gillies M, Richardson DB, Cardis E, Daniels RD, O'Hagan JA, Haylock R, Laurier D, Leuraud K, Moissonnier M, Schubauer-Berigan MK, Thierry-Chef I, Kesminiene A. (2017) Mortality from Circulatory Diseases and other Non-Cancer Outcomes among Nuclear Workers in France, the United Kingdom and the United States (INWORKS). *Radiation Research.*

[25] Barul, C., Fayosse, A, Carton, M, Pilorget, C, Woronoff, AS, Stucker, I, and Luce, D. (2017) Occupational exposure to chlorinated solvents and risk of head and neck cancer in men: a population-based case-control study in France. *Environmental Health* . 16.1: 77.

[26] Carroll, AJ, Carnethon, MR, Liu, K, Jacobs Jr, DR, Colangelo, LA, Stewart, JC, Carr, JJ, Widome, R, Auer, R, and Hitsman, B. (2017) Interaction between smoking and depressive symptoms with subclinical heart disease in the Coronary Artery Risk Development in Young Adults (CARDIA) study. *Health psychology* . 36.2: 101.

[27] Belenchia AM, Johnson SA, Ellersieck MR, Rosenfeld CS, Peterson CA. (2017) In utero vitamin D deficiency predisposes offspring to long-term adverse adipose tissue effects. *Journal of Endocrinology* . JOE-17.

[28] Lin KJ, Mitchell AA, Yau WP, Louik C, Hern□ndez-Diaz S. (2012) Maternal exposure to amoxicillin and the risk of oral clefts. *Epidemiology (Cambridge, Mass.)* . 23.5:699.

[29] Prentice, R. (1982) Covariate measurement errors and parameter es- timation in a failure time regression model. *Biometrika* .69: 331-342.

[30] Wang CY, Hsu L, Feng ZD, Prentice RL. (1997) Regression calibration in failure time regression. *Biometrics* .53.1: 131-45.

[31] Roger L and Spiegelman S. (2012) The SAS %BLINPLUS Macro [Computer software]. Boston, MA: Channing Laboratory.

[32] Puett RC, Hart JE, Yanosky JD, Spiegelman D, Wang M, Fisher JA, Hong B, and Laden F. (2014) Particulate matter air pollution exposure, distance to road, and incident lung cancer in the Nurses? Health Study cohort. *Environmental Health Perspectives*. 122.9: 926-932.

[33] Chiuve SE, Rexrode KM, Spiegelman D, Logroscino G, Manson JE, Rimm EB. (2008) Primary prevention of stroke by healthy lifestyle. *Circulation*. 118.9: 947-54.

[34] Kioumourtzoglou MA, Spiegelman D, Szpiro AA, Sheppard L, Kaufman JD, Yanosky JD, Williams R, Laden F, Hong B, Suh H. (2014) Exposure measurement error in PM2.5 health effects studies: A pooled analysis of eight personal exposure validation studies. *Environ Health*. 13.1: 2.

[35] Sarnat SE, Coull BA, Schwartz J, Gold DR, Suh HH. (2006) Factors affecting the association between ambient concentrations and personal exposures to particles and gases. *Environ Health Perspect*. 14(5):649-54.

[36] Koutrakis P, Suh HH, Sarnat JA, Brown KW, Coull BA, Schwartz J. (2005) Characterization of particulate and gas exposures of sensitive subpopulations living in Baltimore and Boston. *Res Rep Health Eff Inst*. 131:1?65. discussion 67-75.

[37] Liu LS, Box M, Kalman D, Kaufman J, Koenig J, Larson T, Lumley T, Sheppard L, Wallace L. (2003) Exposure assessment of particulate matter for susceptible populations in Seattle. *Environ Health Perspect*. 111(7):909-18.

[38] Meng QY, Turpin BJ, Korn L, Weisel CP, Morandi M, Colome S, Zhang J, Stock T, Spektor D, Winer A, Zhang L. (2005) Influence of ambient (outdoor) sources on residential indoor and personal PM2.5 concentrations: analyses of RIOPA data. *J Expo Anal Environ Epidemiol*. 15(1):17-28.

[39] Sarnat JA, Koutrakis P, Suh HH. (2000) Assessing the relationship between personal particulate and gaseous exposures of senior citizens living in Baltimore. *MD J Air Waste Manag Assoc*. 50(7):1184-98.

[40] Brown KW, Sarnat JA, Suh HH, Coull BA, Koutrakis P. (2009) Factors influencing relationships between personal and ambient concentrations of gaseous and particulate pollutants. *Sci Total Environ*. 407(12):3754-65.

[41] Brown KW, Sarnat JA, Suh HH, Coull BA, Spengler JD, Koutrakis P. (2008) Ambient site, home outdoor and home indoor particulate concentrations as proxies of personal exposures. *J Environ Monit*. 10(9):1041-51.

[42] Williams R, Suggs J, Rea A, Leovic K, Vette A, Croghan C. (2003) The Research triangle park particulate matter panel study: PM mass concentration relationships. *Atmos Environ*. 37:5349-63.

[43] Williams R, Suggs J, Rea A, Sheldon L, Rhodes C, Thornburg J. (2003) The Research Triangle Park Particulate Matter Panel Study: Modeling Ambient Source Contribution to Personal and Residential PM Mass Concentrations. *Atmos Environ*. 37:5365-78.

[44] Paciorek CJ, Yanosky JD, Puett RC, Laden F, Suh H. (2009) Practical large-scale spatio-temporal modeling of particulate matter concentrations. *Ann Appl Stat*. 3:369-96.

[45] Yanosky JD, Paciorek CJ, Schwartz J, Laden F, Puett RC, Suh H. (2008) Spatio-temporal modeling of chronic PM10 exposure for the nurses? health study. *Atmos Environ*. 42(18):4047-62.

[46] Yanosky JD, Paciorek CJ, Laden F, Hart JE, Puett RC, Liao D, Suh HH. (2014) Spatio- temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors. *Environ Health*. 13(1):63.

[47] Puett RC, Schwartz J, Hart JE, Yanosky JD, Speizer FE, Suh H, Paciorek CJ, Neas LM, Laden F. (2008) Chronic particulate exposure, mortality, and coronary heart disease in the nurses? health study. *Am J Epidemiol.* 168(10):1161-8.

[48] Weuve J, Puett RC, Schwartz J, Yanosky JD, Laden F, Grodstein F. (2012) Exposure to particulate air pollution and cognitive decline in older women. *Arch Intern Med*. 172(3):219-27.

[49] Haneuse S, Daniels M. (2016) A general framework for considering selection bias in EHR-based studies: What data are observed and why?. *eGEMS*. 4(1).

[50] Schneeweiss S, Avorn J. (2005) A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology*. 58(4): 323-337.

[51] Weiner MG, Embi PJ. (2009) Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Annals of Internal Medicine*. 151(5): 359-360.

[52] Weiskopf NG, Weng C. (2013) Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*. 20(1): 144-3151.

[53] Gallego B, Dunn AG, Coiera E. (2013) Role of electronic health records in comparative effectiveness research. *J Comp Eff Res*. 2(6): 529-532.

[54] Institute of Medicine (U.S.). Committee on Comparative Effectiveness Research Prioritization. *Initial national priorities for comparative effectiveness research*. Washington, D.C.: National Academies Press; 2009.

[55] Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. (2009) Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report–Part III. *Value in Health*. 12(8): 1062-1073.

[56] Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. (2013) Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. *Medical Care*. 51(8 Suppl 3): S80-86.

[57] Hersh WR, Weiner MG, Embi PJ, et al. (2013) Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care*. 251(8 Suppl 3): S30-37.

[58] Overhage J OL. (2013) Sensible use of observational clinical data. *Statistical Methods in Medical Research*. 22(1): 7-13.

[59] Hernan MA, Hernandez-Diaz S, Robins JM. (2004) A structural approach to selection bias. *Epidemiology*. 15(5): 615-625.

[60] Little RJ, Rubin DB. (2014) *Statistical analysis with missing data*: John Wiley & Sons.

[61] Robins JM, Rotnizky A, Zhao LP. (1996) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *JASA*. 90(429): 106-121.

[62] Carlin AM, Zeni TM, English WJ, Hawasli AA, Genaw JA, Krause KR, Schram JL, Kole KL, Finks JF, Birkmeyer JD, Share D, Birkmeyer, NJO. (2013) The comparative effectiveness of

sleeve gastrectomy, gastric bypass, and adjustable gastric banding procedures for the treatment of morbid obesity. *Annals of Surgery*. 257(5): 791-797.

[63] Arterburn DE and Courcoulas AP. (2014) Bariatric surgery for obesity and metabolic conditions in adults. *BMJ*. 349:g3961.

[64] Courcoulas AP, Yanovski SZ, Bonds D, Eggerman TL, Horlick M, Staten MA, Arterburn DE. (2014) Long-term Outcomes of Bariatric Surgery: A National Institutes of Health Symposium. *JAMA Surgery*. 149(12): 1323-1329.

[65] Colquitt JL, Pickett K, Loveman E, Frampton GK. (2014) Surgery for weight loss in adults. *The Cochrane database of systematic reviews*. 8.

[66] Schneeweiss S. (2007) Understanding secondary databases: A commentary on ?Sources of bias for health state characteristics in secondary databases?. *Journal of Clinical Epidemiology*. 60(7): 648.

[67] Little RJ. (1993) Pattern-mixture models for multivariate incomplete data. *JASA*. 88(421):125-134.

[68] Breslow NE. (1972) Discussion of the paper by D. R. Cox. *J R Statist Soc B*. 34:216?217.

[69] van der Vaart AW. (2000) *Asymptotic Statistics*: Cambridge University Press.