Published in final edited form as:

Pharmacoepidemiol Drug Saf. 2015 December ; 24(12): 1286–1296. doi:10.1002/pds.3876.

Evaluating the Treatment Effects Model For Estimation of Cross-sectional Associations Between Risk Factors and Cardiovascular Biomarkers Influenced by Medication Use

Andrew J. Spieker¹, Joseph A.C. Delaney², and Robyn L. McClelland¹

¹University of Washington Department of Biostatistics

²University of Washington Department of Epidemiology

Abstract

Purpose—In cross-sectional observational data, evaluation of biomarker-to-exposure associations is often complicated by nonrandom medication use. Traditional approaches often lead to biased estimates, consistent with known results involving confounding by indication. More sophisticated, yet easy to implement approaches such as inverse probability weighting and censored normal regression can address medication use in certain settings, but have poor performance when medication use depends on off-medication biomarker values. More sophisticated approaches are necessary.

Methods—Heckman's treatment effects model resembles the process which gives rise to cross-sectional data. In this study, we conduct a variety of simulation studies to illustrate why traditional approaches are inappropriate when medication use depends on underlying biomarker values. We illustrate how Heckman's model can accommodate this feature. We also apply the models to data from the Multi-Ethnic Study of Atherosclerosis.

Results—Inverse probability weighting and censored normal regression are sensitive to how strongly medication use is associated with untreated biomarker values (the untreated value acts as an unmeasured predictor of medication use in this context). Heckman's model can often adequately remove bias and is robust to certain forms of model misspecification, but relies on knowing important predictors of medication use, even when they are independent of the biomarker. The advantages of Heckman's model can be negated if the effect of medication on biomarker values is proportionate to the underlying biomarker.

Conclusions—If predictors of medication use are measured, data are cross-sectional, and effects are approximately additive, then Heckman's model is more accurate relative to alternative approaches.

Corresponding Author: Andrew J. Spieker, F-600 Health Sciences Building, Box 357232, University of Washington Seattle, WA 98195.

The authors have no conflicts of interest to declare.

Prior Presentations: Some results from this study were highlighted in a contributed oral presentation at the Eastern North American Region Conference on March 17, 2014 in Baltimore, MD.

Keywords

Cross-sectional; observational study; biomarker; endogenous treatment; confounding by indication

1. INTRODUCTION

In cross-sectional observational data, estimating the association between an exposure of interest and a biomarker is challenging when participants are on medication intended to alter their biomarker values. Consider using cross-sectional observational data to evaluate the association between diabetes (exposure of interest) and low-density lipoprotein (LDL). Depending on the age distribution of the cohort, many participants may be on medication to lower their LDL, so that their underlying, untreated LDL values (which would theoretically have been observed had they not been on medication at the time of measurement) are unobservable. Medication use occurs most frequently in participants who would have had a high LDL off medication. In addition, medication use is often associated with the exposures of interest (diabetes in this example), and other observable factors such as the Framingham Risk Score and insurance. Data on underlying LDL values are missing according to medication use status, which depends on unobservable and/or observable risk factors. That is, medication use is endogenous.

Methods have been proposed to address nonrandom medication use in a longitudinal setting.¹ Commonly, however, only cross-sectional data are available (e.g., when safety challenges or substantial cost/inconvenience preclude the possibility of repeated measurements). In these cases, we must turn to methods which are applicable to cross-sectional data in order to estimate biomarker associations. Traditional methods such as excluding participants on medication from analysis or adjusting for medication use, despite their frequent use^{2–8}, require strong assumptions and often result in invalid estimates.

More sophisticated methods that can be applied or generalized to a cross-sectional setting such as inverse probability-of-treatment weighting (IPTW)^{9–10} and censored normal regression¹¹ have been proposed, but rely on stringent assumptions that are almost assuredly not satisfied in practice. James Heckman devised what he refers to as a hybrid model with structural shift¹², better known as his “treatment effects model,” which can be used to correct the observed biomarker for the effect of medication use by jointly modeling the biomarker with a latent medication use model. Parameters can be estimated through maximum likelihood.¹³

Figure 1 illustrates the data generation mechanism presumed by the Heckman model, which we believe is a fairly realistic representation of cross-sectional data with endogenous medication use. The Heckman model provides us with a means of addressing nonrandom medication use with an alternative set of assumptions to those made by censored normal regression and IPTW. It is important to better understand the performance of these approaches when medication use depends on participants’ off-medication biomarker values. Thus, there are two primary goals to this study: (1) to demonstrate the lack of plausibility of traditional methods, and (2) to better understand to what extent misspecification of the Heckman model negates its advantages under correct specification. In turn, we apply these

models to our motivating example of LDL and diabetes from the Multi-Ethnic Study of Atherosclerosis (MESA).

2. THE FAILURE OF TRADITIONAL APPROACHES

We demonstrate the failure of traditional approaches with a simplistic example. Figure 2 depicts the results of fitting three simple approaches based on a single realization of a hypothetical data generating mechanism with endogenous medication use: (i) fitting ordinary least squares (OLS) on the observed data, or “ignoring the issue”, (ii) excluding on-medication participants from analysis, and (iii) adjusting for medication use. The off-medication participants are represented by black circles (these values represent the underlying biomarker value, which is observed). The on-medication participants are depicted by representing their underlying biomarker value as a gray $\times$ and the observed value as a gray circle. In Figure 2A, we show the true association with a solid black line, which would be estimable if no participants were on medication. The results from these three models fail to capture the true association well. These methods perform poorly because they rely on assumptions about medication use that are not satisfied. Neither ignoring the issue nor excluding participants is reasonable as each approach presumes that medication use is independent of the underlying biomarker. Adjusting for medication use as though a confounder (as in Figure 2D) presumes that medication use is independent of the underlying biomarker, conditional on the exposure of interest. Based on Figure 1, medication use is not exactly a confounder for the relationship of interest, but is rather a cause of missing data on the underlying biomarker.^{14–15} Importantly, the confidence intervals for each of these estimates fail to include the truth.

The adjustment approach is sufficient if medication use is random (in the setting of a randomized trial, for instance). More generally, it is sufficient if medication use is independent of the underlying biomarker value. For example, suppose the biomarker in question is systolic blood pressure (SBP) and the medication under consideration is a statin; traditional methods are sufficient since, despite the fact that statins are known to lower SBP, participants do not traditionally take statins for that purpose.¹⁶

3. METHODS

3.1 Regression and Likelihood Methods

In this section, we describe the assumptions of each of the discussed models, and how they are fit. We focus our attention on simple approaches which are easy to implement. *Ad hoc* methods such as “fixed” or “random” addition¹¹ are not considered in this paper because they rely on external information on study participants or population average treatment effects; such information often does not exist. Let y_i denote the observed biomarker for subject $i = 1, \dots, n$ and let z_i denote the medication use indicator taking on 1 for a subject on medication and 0 otherwise. We denote the underlying, untreated biomarker value by y_i^* , which is unobservable if $z_i = 1$. In untreated participants, y_i and y_i^* are equivalent. Let $\mathbf{X} = (1, X_1, \dots, X_K)$ be the design matrix of the exposures of interest (e.g., age, gender, and race), and $\boldsymbol{\beta} = (\beta_0, \dots, \beta_K)^T$ be a vector of regression coefficients (fitting an intercept). Suppose that

$y_i^* = x_i\beta + \varepsilon_i$, where ε_i are independent, each with mean 0 and variance σ_ε^2 . The exposure-to-biomarker association, β , is our target of inference. Estimates from traditional models are obtained from fitting the following from OLS using the observed y -values: $\mathbf{E}[y_i|x_i] = x_i\beta$ (Ignore), $\mathbf{E}[y_i|x_i, z_i = 0] = x_i\beta$ (Exclude), and $\mathbf{E}[y_i|x_i, z_i] = x_i\beta + z_i\eta$ (Adjust). As mentioned in Section 2, these approaches presume that medication use is either completely independent (Ignore and Exclude) or conditionally independent (Adjust) of the underlying biomarker.

3.1.1 Inverse Probability of Treatment Weighting—Wang et al. describe an inverse probability-of-treatment weighting (IPTW) approach for a longitudinal setting, up-weighting study participants who are less likely to be observed in an attempt to reduce bias (namely, participants on medication with lower observed biomarker values, and participants off medication with higher observed biomarker values).^{9–10}

This approach can be generalized to a cross-sectional setting as well. The probability of being on medication can naïvely be estimated using a (probability of) medication use model from logistic regression on observed $\mathbf{X}$'s: $\text{logit}(P(z_i = 1|x_i)) = x_i\tau$, from which stabilized weights can be generated. An analogue to the adjustment model can then be fit with these weights. This method differs from the approach of Wang et al.⁹, as we do not generate weights from the unobservable biomarker value, but it is still truly an IPTW-type approach. Note that only variables which are associated with both the biomarker and medication use are used to generate the weights, excluding variables which predict medication use only; this has been shown to have better behavior than including covariates which predict medication use only.^{17–18} Of note is that the IPTW medication use model cannot incorporate what is arguably the most important predictor of treatment in this setting: the (unobservable) untreated biomarker value. This is a meaningful departure from the assumptions of IPTW, under which this model would otherwise be expected to perform well. IPTW also has known challenges when positivity assumptions are violated.

Our approach differs from Wang's in the following ways: he classifies each participant's underlying biomarker into two categories (high or low); this category is presumed to be observed, even though the underlying biomarker itself is unobservable. The probability of medication use, in turn, is presumed only to depend on the exposure and the underlying biomarker category (high or low).

3.1.2 Censored Normal Regression—Tobin outlines a censored normal regression approach which attempts to correct the observed biomarker, y_i , in participants on medication by modeling them as right-censored versions of their untreated values.¹¹ The approach is as follows: y_i is modeled according to:

$$y_i^* \begin{cases} = y_i & \text{if } z_i = 0 \\ \geq y_i & \text{if } z_i = 1 \end{cases} \quad (1)$$

We then may modify the normal likelihood for ordinary least squares linear regression, incorporating this right-censoring: let ϕ and Φ denote the standard normal density and

cumulative distribution, respectively, and σ_ε^2 the error variance of y_i^* , we have that the likelihood is given by:

$$\mathcal{L}(\beta, \sigma_\varepsilon^2 | \mathbf{Y}, \mathbf{Z}, \mathbf{X}) = \prod_{i=1}^n \left(\phi \left(\frac{y_i - x_i \beta}{\sigma_\varepsilon} \right) \right)^{1-z_i} \left(1 - \Phi \left(\frac{y_i - x_i \beta}{\sigma_\varepsilon} \right) \right)^{z_i}. \quad (2)$$

To estimate parameters, this likelihood function can be maximized. This model presumes that (i) the medication's effect is non-negative for each participant; (ii) the untreated value follows a normal distribution; and, (iii) censoring is non-informative. Tobin acknowledges that the non-informative censoring assumption is not quite satisfied, but believes in his motivating example of genetic exposures and blood pressure that this approach well-approximates the association of interest.

3.1.3 Extension to Endogenous Medication Use (Heckman's Treatment Effects Model)

—Let $\mathbf{W} = (\mathbf{1}, \mathbf{W}_1, \dots, \mathbf{W}_M)$ denote a design matrix corresponding to observed covariates which predict medication use (they need not all be distinctly different from the covariates which predict y_i^*). Define $z_i^* = \mathbf{w}_i \alpha + \gamma_i$, where $\alpha = (\alpha_0, \dots, \alpha_M)$ and γ_i is a random error with mean zero and variance σ_γ^2 . Here, z_i^* is a continuous latent variable. Although z_i^* is unobserved, medication use, defined by $z_i = 1$ ($z_i^* > 0$) (the indicator that z_i^* is positive) is observed. We refer to this as the “medication use” model.

Finally, let $y_i = y_i^* - \mu_\delta z_i$ denote the observed biomarker value, which is equivalent to y_i^* for participants off medication ($z_i = 0$), but is some unknown μ_δ lower than y_i^* for participants on medication ($z_i = 1$). To estimate parameters, this model was originally proposed as a two-stage model¹², but Maddala proposed a more efficient maximum likelihood approach which jointly models $(\varepsilon_i, \gamma_i)$ as bivariate normal, with correlation parameter ρ .¹³ We consider the latter approach. The parameters σ_γ and ρ are weakly identifiable, and hence it is convenient to set $\sigma_\gamma^2 = 1$ without loss of generality so that all parameters may be estimated (this is a well known property of the probit model).¹⁹ The likelihood can then be written as follows, letting $\theta = (\alpha, \beta, \mu_\delta, \rho, \sigma_\varepsilon)$:

$$\mathcal{L}(\theta | \mathbf{Y}, \mathbf{Z}, \mathbf{Z}, \mathbf{W}) = \prod_{i=1}^n \phi \left(\frac{y_i - x_i \beta + \mu_\delta z_i}{\sigma_\varepsilon} \right) \Phi(-r_i)^{1-z_i} \Phi(r_i)^{z_i}. \quad (3)$$

The quantity r_i is the conditional mean of $z_i^* | y_i^*$, divided by its conditional standard deviation, which can be obtained from standard bivariate normal theory. In this model, our parameters of primary interest (namely, β , σ_ε , and μ_δ) are identifiable. They can be estimated from maximizing this likelihood function.

It turns out that this model very easily accommodates direct dependence of z_i^* on the underlying biomarker y_i^* : if we add a term λy_i^* into the medication use model, it may be expanded as follows: $z_i^* = \mathbf{w}_i \alpha + \lambda x_i \beta + \lambda \varepsilon_i + \gamma_i$. then, if we intentionally include all

covariates of $\mathbf{X}$ in $\mathbf{W}$, the total error on z_i^* is $\lambda \varepsilon_i + \gamma_i$, and the total errors are still bivariate normal, just as before. Hence, the model is still correctly specified and provides consistent estimates of β . Since α is not identifiable, there is no need to estimate λ in this procedure.

Additionally, this model indeed allows for the effect of the medication use to be normally distributed around μ_δ across subjects. This introduces heteroscedasticity, and therefore robust “sandwich” based standard errors should be used for valid inference.²⁰

The “medication use” model is somewhat similar in spirit to that of IPTW, except that its total errors are permitted to be correlated to the errors on the biomarker. This is the mechanism by which endogeneity of medication use is incorporated; the alternative approaches described do not have this advantage. Use of the probit model rather than a logistic model as in IPTW is a convenient way to obtain the conditional distribution of $z_i^* | y_i^*$. Another key difference in the models is that all predictors of medication use are to be placed in the medication use model, as opposed to IPTW in which this is not the case. This difference will be reflected in how we conduct our simulation studies.

3.2 Simulation Study

3.2.1 Heckman-Ideal Scenario—We designed a number of scenarios with no departures from the Heckman model assumptions for simulation purposes (the assumptions of IPTW and the censored normal model are hence not satisfied). The primary goal of this component of the study is to evaluate the advantage over alternative approaches which have been previously described to address medication use (censored normal regression and IPTW). Note: the simulation setup of Section 2 is a simplified version of the one we present in this section.

Let $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ denote a design matrix for all predictors of interest measured in the study. Assume that they are fixed (we generated each X_i independently from a normal distribution with standard deviation 1, and treated them as fixed through all replications of the simulation study). Suppose that the data generating mechanism is given by

$z_i^* = x_{1i} + x_{3i} + \lambda y_i^* + \gamma_i$ so that $\alpha = (0, 1, 0, 1)^T$, and that $y_i^* = x_{1i} + x_{2i} + \varepsilon_i$ so that $\beta = (\beta_0, \beta_1, \beta_2)^T = (0, 1, 1)^T$. Thus, in accordance with Figure 1, X_1 predicts the underlying biomarker and medication use, X_2 predicts the underlying biomarker, but not medication use (except through λ), and X_3 predicts medication use, but not the underlying biomarker. Suppose that the effect of the medication is generated as $\delta_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_\delta, \sigma_\delta^2)$ to allow inter-individual variation in the effect of treatment. We fixed the variance of the error terms ε_i and γ_i to be 4 for all settings. Finally, the observed outcome is generated as follows: $y_i = y_i^* - \delta_i$.

We alter the parameters λ , μ_δ , and ρ to generate four “Heckman-ideal” scenarios, noting that λ is not estimated. In the first, $\lambda = 0.2$, $\mu_\delta = 1$, and $\rho = 0$. The other three scenarios alter one of each of these parameters at a time ($\lambda = 0.8$ in Scenario 2, $\mu_\delta = 1.5$ in Scenario 3, and $\rho = 1/16$ in Scenario 4). These three alterations can be described as “increasing endogeneity through exposures and correlation,” “increasing effect of medication use,” and “increasing endogeneity through random correlation only,” respectively. For all settings, $\sigma_\delta^2 = 1/10$.

Two-thousand replicates were used for each simulation scenario (each based on $n = 2000$ participants, and the nuisance intercept in the generation of z^* was set to yield approximately a 50% prevalence of medication use in each replication). β was estimated from each of the six methods outlined, along with Huber-White sandwich variance estimates to account for the heteroscedasticity.²⁰ We then obtained an estimate of the true standard error by taking the standard deviation of the simulated estimates. We focused our attention on β_1 and β_2 , as the intercept is not of interest in most association studies.

Note: although medication use is dichotomized as described in Section 3.1.3, all participants have a nonzero probability of being on medication due to random error in the medication use model; to verify that we have not induced positivity challenges for IPTW, we report descriptive statistics (mean, minimum, and maximum) on the IPTW weights from the simulations.

3.2.2 Formal Sensitivity Analyses—We performed a number of sensitivity analyses in which we altered the data generating mechanism of “Ideal Scenario 1” so that the Heckman model was no longer correctly specified. (NB: the other models remain incorrectly specified even under these modifications). The goal of this component of the study was to evaluate to what extent the advantages of the Heckman model seen in the previous study are negated when we perturb its assumptions.

The first three sensitivity analyses pertain specifically to misspecification of the IPTW and Heckman medication use models. We first consider estimation of β when X_1 is removed from the medication use models, and then when X_2 is removed. Recall that X_3 is not originally included in the IPTW medication use model, so we may only evaluate the Heckman model when X_3 is not included.

The next three sensitivity analyses pertain to distributional/measurement assumption violations commonly encountered in practice: a departure from the normal errors in the biomarker model; measurement error in X_1 ; and, an effect of medication use that is proportional to the underlying biomarker value, with a mean-variance relationship. Details of the sensitivity analyses are provided in Table 1.

3.3 Application to the Multi-Ethnic Study of Atherosclerosis

We applied these methods to data collected from the Multi-Ethnic Study of Atherosclerosis (MESA). MESA is a prospective cohort study designed to study the progression of subclinical cardiovascular disease. The study includes 6,814 men and women aged 45–84 years who were free of clinical cardiovascular disease at entry. The participants were recruited from six U.S. communities, and included 47% men with an ethnic distribution of 38% white, 28% African-American, 22% Hispanic, and 12% Chinese-American. All subjects gave informed consent. Details of the sampling, recruitment, and data collection have been reported elsewhere.²¹

Specifically, we modeled low-density lipoprotein (LDL) as the biomarker outcome, with diabetes the predictor of interest, and adjusting for: age, gender, race category, smoking & alcohol status, and BMI. The medication use indicator variable was taken to be self-reported

use of any lipid-lowering medication. In assigning the weights for IPTW, these variables were all used in the medication use model. For the Heckman medication use model, the following variables were added: insurance status, education status, and Framingham risk score.

All simulations and analyses were conducted using Stata Version 12 (StatCorp, College Station, TX) or R Version 2.14.2 (The R Foundation for Statistical Computing, <http://www.R-project.org>).^{22–23} Censored normal regression can be performed using Stata's "cnreg" command, and Heckman's treatment effects model can be performed with the "treatreg" command.

4. RESULTS

4.1. Simulation Studies

4.1.1 Heckman-ideal Simulation Results—Table 2 provides a summary of the estimates, simulated standard errors, estimated (Huber-White) standard errors, and coverage probabilities for Scenario 1. The results for the remaining scenarios are illustrated in Figure 3. In each of these "Heckman-ideal" scenarios, the Heckman model provides valid estimates and inference for all coefficients. In general, Huber-White standard errors estimate the true standard errors well. In estimating β_1 , the Heckman model far outperforms the other approaches in coverage probability and bias. The censored normal approach meaningfully overestimates β_1 . Traditional methods tend to perform better in estimating β_2 with only modest bias, which is unsurprising since X_2 does not predict medication use as strongly as X_1 .

When the strength of dependence of medication use on the underlying biomarker increases (Scenarios 2 and 4), all methods but Heckman's model show more bias in comparison to Scenario 1. For IPTW, this result is consistent with previous findings that this method is sensitive to increases in endogeneity due to unobservables.²⁴ The weights from IPTW in a single replication ranged from 0.41 to 7.7 (mean 1.0), suggesting that we do not have positivity violations in the data.

When the mean effect of medication is increased from $\mu_\delta = 1.0$ to $\mu_\delta = 1.5$, the Heckman model still fares well in comparison to other methods. Note that the true mean difference in the underlying biomarker value between treated and untreated participants is approximately 1.38 based on this simulation, so an effect of 1.0 could signify a moderate effect, but an effect of 1.5 is very large. Although the censored normal approach improved marginally when the effect of medication is large, it is clearly sensitive to the level of endogeneity which is not truly estimable from the censored normal approach. Estimates from the exclusion method are trivially equivalent under all possible μ_δ .

We conducted a follow-up simulation study based on these results to evaluate the performance of these methods when medication use was dependent only on the observable covariates (in particular, when $\lambda = 0$, but all other parameters were as described in Scenario 1). All methods except ignoring the issue and censored normal regression provided valid estimates and approximately 95% coverage in this case.

4.1.2 Medication Use Model Misspecification—Table 3 provides a summary of the estimates, their variability, and coverage probabilities from the sensitivity analysis involving medication use model misspecification.

Removal of X_1 from the medication use model has a meaningful effect on the bias of $\hat{\beta}_1$ for both IPTW and the Heckman model. The variability of $\hat{\beta}_1$ and $\hat{\beta}_2$ decrease, with coverage of β_2 still being adequate. Removing X_2 from the medication use model has little impact on the bias of $\hat{\beta}_1$, although for the Heckman model, there is a slight bias induced for $\hat{\beta}_2$ and a reduction of coverage by about 2%. The variability of each of the estimates decreases.

Removal of X_3 from the Heckman medication use model results in a downward bias on $\hat{\beta}_1$ with a loss of coverage and substantially higher variability of the estimates. The Huber-White standard error does not estimate the true standard error well in this case. Of note is that estimation of $\hat{\beta}_2$ is not meaningfully impacted by this change. This appears to be the reverse of IPTW.¹⁸

4.1.3 Distributional and Measurement-Based Assumption Violations—None of the methods were heavily influenced by moderate right-skewness of the errors or slight measurement error on X_1 (Table 4). Additional simulations with higher measurement error resulted in shrinkage of the estimates (that is, a tendency for the estimates to be closer to zero), consistent with the known properties of nondifferential misclassification of the exposure.¹⁷

When the effect of medication use is multiplicative and not additive, censored normal regression provides higher upward bias. Both the Heckman model and IPTW appear to be sensitive to this violation for both $\hat{\beta}_1$ and $\hat{\beta}_2$, as we see shrinkage and a loss of coverage. Under this example of multiplicative effects, the Heckman model does not show much improvement over the traditional approaches.

4.2 Application to MESA biomarkers

The prevalence of lipid-lowering medication use in the sample was 16.2%. Figure 4 depicts the point estimates and confidence intervals for estimating the diabetes-LDL association from each method. In estimating the association between diabetes and LDL, all traditional methods, as well as IPTW, would suggest protective effects of diabetes on LDL. These results would be inconsistent with the literature.²⁵ Censored normal regression and Heckman provide estimates that are consistent with no such association. This mirrors prior results found by a longitudinal imputation approach.¹ Interestingly, the censored normal approach does not provide the highest estimates; it is possible that the effect of lipid lowering medications is large enough in this setting that censored normal is performing reasonably well.

As noted in Section 3.1.3, Heckman's model can also be used to estimate the effect of the medication on the biomarker. In this example, we estimate a treatment effect of 47.7 mg/dL (95% CI: [42.5, 52.9]). The adjustment model provides an estimate of 14.8 (95% CI: [12.9, 16.8]), and IPTW provides an estimate of 13.4 (95% CI: [11.2, 15.6]). The estimates

provided by Heckman's model closely resemble estimates obtained from longitudinal data; estimates from the other two approaches are severely attenuated.¹

5. DISCUSSION

We have demonstrated that in cross-sectional data, traditional methods can perform poorly in estimating an exposure-to-biomarker association when medication use depends on untreated biomarker values. Alternative methods such as censored normal regression and IPTW can show poor performance when used without regard to this dependence; however, the Heckman model shows high advantages over alternatives when correctly specified, as expected. Although cross-sectional data have limitations surrounding understanding temporality of events, there are settings in which they are the only means we have to estimate associations. We believe that the assumptions of the Heckman model are more reasonable than those of the alternatives and hence provide us with estimates that are closer to the truth.

We have confirmed the inappropriateness of IPTW in this setting by noting its sensitivity to the strength of correlation of the biomarker and medication use; when increased, the ability to properly weight participants deteriorates since treated and untreated subjects will be more well-separated, creating challenges with positivity of medication use probability.²⁴ Arguably the most important predictor of medication use (the untreated biomarker), is unobservable and hence not incorporated in the weighting procedure. The Heckman model addresses this by modeling the correlation. The errors (ϵ_i , γ_i) need not themselves be correlated for there to be endogeneity, as the total errors on the medication use variable and the underlying biomarker are correlated if $\lambda \neq 0$. Medication to lower a biomarker can often be assumed to be endogenous, as those with undesirable biomarker values are likely to be on medication.

The Heckman model presumes that data are collected on X_1 (associated with both the biomarker and medication use) as it is included in the biomarker model. Hence, bias induced by excluding X_1 from the medication use model can be addressed by including it. To that end, this component of the study does not truly describe a limitation of the Heckman model, but confirms that variables strongly associated with the biomarker should be measured.

However, a variable associated with medication use, irrespective of its association with the biomarker, impacts the variability of the estimates when not included. Variables which are strongly associated with the biomarker outcome are typically known and hence collected throughout a study; however, data on variables associated with medication use may not be collected. This is a clear shortcoming of the Heckman model when the target of inference corresponds to a covariate associated with medication use. We recognize in our application to MESA that there are likely unmeasured variables associated with medication use, although the level of association and the extent of their impact is not verifiable. This shortcoming does not hold for IPTW as it does not model variables associated with only medication use.^{17–18} In our application to LDL, we have demonstrated that Heckman's model can provide estimates of both the biomarker association and the effect of the medication that are consistent with longitudinal data, whereas IPTW does not.

The Heckman model is robust to departures from several distributional assumptions (skewness of biomarker errors and measurement error, in particular), still showing advantages over alternative approaches. However, non-additive effects can substantially impact bias. This is a meaningful limitation, since in cross-sectional observational data, there is no reliable way to estimate the distribution of the effect of medication use. Although, this is no more of a limitation for the Heckman model than for adjustment and IPTW, which also presume additive effects of medication. Methods such as ignoring and excluding do not show substantial worsening of bias precisely because they fail to model the effect of medication at all.

While the Heckman model remains robust to certain assumptions, unverifiable misspecification of the medication use model can be difficult to overcome in settings where predictors of medication use are not well understood. However, if predictors of medication use are thought to be known and the effect of medication is to change the biomarker by a fixed amount on average, the Heckman model should be considered as a means of correcting bias. Special consideration of endogeneity and the associations of variables with medication use and biomarkers should inform model selection for estimating biomarker-to-exposure associations. Research to develop further extensions of these models is needed in order to correctly estimate the natural association between a predictor of interest and a biomarker in cross-sectional observational data with endogenous medication use.

Acknowledgments

This study was supported by R01 HL 103729-01A1. MESA was supported by contracts N01-HC-95159 through N01-HC-95169 from NHLBI. The authors thank the other investigators, the staff, and MESA participants for their valuable contributions. A full list of participating MESA investigators and institutions can be found at <http://www.mesa-nhlbi.org>.

References

1. McClelland RL, Kronmal RA, Haessler J, et al. Estimation of risk factor associations when the response is influenced by medication use: An imputation approach. *Statistics in Medicine*. 2008; 27(24):5039–5053. [PubMed: 18613245]
2. Brand E, Wang JG, Herrmann SM, et al. An epidemiological study of blood pressure and metabolic phenotypes in relation to the Gbeta3 C825T polymorphism. *Hypertension*. 2003; 21(4):729–737.
3. Iwai N, Baba S, Mannami T, et al. Association of sodium channel gamma-subunit promoter variant with blood pressure. *Hypertension*. 2001; 38(1):86–89. [PubMed: 11463765]
4. Matsubara M, Kikuya M, Ohkubo T, et al. Aldosterone synthase gene (CYP11B2) C-334T polymorphism, ambulatory blood pressure and nocturnal decline in blood pressure in the general Japanese population: the Ohasama Study. *Journal of Hypertension*. 2001; 19(12):2179–2184. [PubMed: 11725161]
5. O'Donnell CJ, Lindpaintner K, Larson MG, et al. Evidence for association and genetic linkage of the angiotensin-converting enzyme locus with hypertension and blood pressure in men but not women in the Framingham heart study. *Circulation*. 1998; 97(18):1766–1772. [PubMed: 9603529]
6. Rice T, Rankinen T, Province MA, et al. Genome-wide linkage analysis of systolic and diastolic blood pressure: the Quebec family study. *Circulation*. 2000; 102(16):1956–1963. [PubMed: 11034945]
7. Schunkert H, Hense HW, Doring A, et al. Association between a polymorphism in the G protein beta3 subunit gene and lower renin and elevated diastolic blood pressure levels. *Hypertension*. 1998; 32(3):510–513. [PubMed: 9740618]

8. Sethi AA, Nordestgaard BG, Tybjaerg-Hansen A. Angiotensinogen gene polymorphism, plasma angiotensinogen, and risk of hypertension and ischemic heart disease: a meta-analysis. *Arteriosclerosis Thrombosis and Vascular Biology*. 2003; 23(7):1269–1275.
9. Wang Y, Fang Y. Adjusting for treatment effect when estimating or testing genetic effect is of main interest. *Journal of Data Science*. 2011; 9:127–138.
10. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression with missing data. *Journal of the American Statistical Association*. 1995; 90(429):122–129.
11. Tobin MD, Sheehan NA, Scurrah KJ, et al. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in Medicine*. 2005; 24(19): 2911–2935. [PubMed: 16152135]
12. Heckman JJ. Dummy endogenous variables in a simultaneous equation system. *Econometrica*. 1978; 46(4):931–959.
13. Maddala, GS. Limited-dependent and qualitative variables in econometrics. Cambridgeshire: Cambridge University Press; 1983.
14. Pearl J. Causal inference in statistics: a review. *Statistical Surveys*. 2009; 3:96–146.
15. Miettinen OS. The need for randomization in the study of intended effects. *Statistics in Medicine*. 1983; 2(2):267–271. [PubMed: 6648141]
16. Statins and high blood pressure: The latest trial of these cholesterol-lowering drugs suggest they may help people with high blood pressure even if they don't have high cholesterol. *Harvard Heart Letter: from Harvard Medical School*. 2003:7.
17. Lefebvre G, Delaney JAC, Platt RW. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics in Medicine*. 2008; 27(18):3629–3642. [PubMed: 18254127]
18. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *American Journal of Epidemiology*. 2006; 163(12):1149–1156. [PubMed: 16624967]
19. Freedman DA, Sekhon JS. Endogeneity in probit response models. *Political Analysis*. 2010; 18(2): 138–150.
20. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980; 48(4):817–838.
21. Bild DE, Bluemke DA, Burke GL, et al. Multi-ethnic study of atherosclerosis: objectives and design. *American Journal of Epidemiology*. 2002; 156(9):871–881. [PubMed: 12397006]
22. StataCorp. Stata Statistical Software: Release 2013. College Station, TX: StataCorp LP; 2013.
23. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2013. URL <http://www.R-project.org/>
24. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*. 2008; 168(6):656–664. [PubMed: 18682488]
25. American Diabetes Association (Position Statement). Dyslipidemia management in adults with diabetes. *Diabetes Care*. 2004; 27:S68S71.

Key points

- Simple, traditional approaches to estimate biomarker associations in cross-sectional observational data tend to be highly biased.
- Common fixes such as inverse probability weighting do not reduce bias in this setting except in rare circumstances.
- Heckman's treatment effects model resembles the process by which cross-sectional data arise.
- The treatment effects model is reasonably robust to a wide range of assumption violations.

Figure 1.

Diagram depicting the presumed data generation mechanism for the Heckman treatment effects model in a cross-sectional setting. All measurements refer to a single point in time. Dashed lines indicate associations of particular interest.

Figure 2.

Observed values for untreated participants are shown as black circles; treated participants are in gray; the observed (o) and treated (x) values are connected with a line. The dark black line signifies the true association of $\beta = 2$; the gray lines show the results from three traditional approaches: ignoring the issue (panel B), excluding treated participants (panel C), and adjusting for treatment (panel D).

Figure 3.

Point estimates are averaged over 2000 replications, denoted as a black dot; vertical lines represent the simulation-standard error, and the diagonal lines denote the estimated Huber-White standard errors. Scenario 2 corresponds to an increase in λ , Scenario 3 corresponds to an increase in μ_δ , and Scenario 4 corresponds to an increase in ρ .

Figure 4. Graphical depiction of the application to diabetes-LDL model in the Multi-Ethnic Study of Atherosclerosis. Estimates are reported in mg/dL.

Table 1

Simulation Setup: Sensitivity Analyses

Violation		Description
1	Treatment Model Misspecification I	X_1 removed from treatment model ^a
2	Treatment Model Misspecification II	X_2 removed from treatment model ^a
3	Treatment Model Misspecification III	X_3 removed from treatment model ^b
4	Non-Normal Errors	$\varepsilon_i \sim \text{Gamma} \left(\alpha = 8, \beta = \sqrt{2} \right) - 8 / \sqrt{2}$
5	Measurement Error	$X_1^{obs} = X_1 + \xi_i; \xi_i \sim \mathcal{N} \left(0, (1/20)^2 \right)$
6	Proportional Treatment Effect	$\delta_i \sim \mathcal{N} \left(Y_i^* / 5, (Y_i^* / 20)^2 \right)$

^a - Only applies to IPTW and Heckman Model

^b - Only applies to Heckman model

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Simulation Results under Heckman-Ideal Scenarios ($\beta_1 = \beta_2 = 1$)

Scenario 1	β_1				β_2				
	Default	Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP	Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP
Ignore		0.81	0.046	0.046	0.02	0.98	0.046	0.045	0.91
	Exclude	0.90	0.069	0.069	0.66	0.99	0.066	0.064	0.94
	Adjust	0.89	0.049	0.049	0.41	0.99	0.046	0.045	0.94
Censored Normal		1.53	0.064	0.06	0.00	1.07	0.060	0.062	0.81
	IPTW	0.88	0.061	0.059	0.49	0.99	0.051	0.050	0.93
	Heckman	1.00	0.068	0.068	0.95	1.00	0.046	0.046	0.95

Table 3

Simulation Results Under Treatment Model Misspecification ($\beta_1 = \beta_2 = 1$)

β_1									
β_2									
X_1 Removed	Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP	Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP	
IPTW	0.89	0.049	0.049	0.41	0.99	0.046	0.045	0.94	0.94
Heckman	0.88	0.049	0.049	0.31	1.00	0.047	0.046	0.95	0.95

β_1									
β_2									
X_2 Removed	Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP	Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP	
IPTW	0.88	0.061	0.059	0.49	0.99	0.050	0.050	0.93	0.93
Heckman	1.00	0.068	0.069	0.95	0.98	0.046	0.045	0.92	0.92

β_1									
β_2									
X_3 Removed	Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP	Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP	
Heckman	0.95	0.22	0.16	0.73	0.99	0.053	0.052	0.94	0.94

Table 4

Simulation Results Under Distribution/Measurement Violations ($\beta_1 = \beta_2 = 1$)

β_1										β_2											
Skewed Errors		Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP	Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP			Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP			Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP
Ignore		0.81	0.046	0.046	0.01	0.98	0.046	0.045	0.93			0.98	0.046	0.045	0.93			0.98	0.046	0.045	0.93
Exclude		0.90	0.067	0.066	0.67	0.99	0.062	0.061	0.94			0.99	0.062	0.061	0.94			0.99	0.062	0.061	0.94
Adjust		0.89	0.050	0.049	0.40	0.99	0.045	0.045	0.94			0.99	0.045	0.045	0.94			0.99	0.045	0.045	0.94
Censored Normal		1.54	0.065	0.065	0.00	1.07	0.060	0.062	0.81			1.07	0.060	0.062	0.81			1.07	0.060	0.062	0.81
IPTW		0.88	0.062	0.059	0.46	0.99	0.051	0.050	0.93			0.99	0.051	0.050	0.93			0.99	0.051	0.050	0.93
Heckman		1.00	0.072	0.071	0.95	1.00	0.046	0.046	0.95			1.00	0.046	0.046	0.95			1.00	0.046	0.046	0.95

β_1										β_2											
Measurement Error		Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP	Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP			Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP			Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP
Ignore		0.81	0.045	0.046	0.02	0.98	0.047	0.045	0.91			0.98	0.047	0.045	0.91			0.98	0.047	0.045	0.91
Exclude		0.89	0.067	0.069	0.66	0.99	0.066	0.064	0.94			0.99	0.066	0.064	0.94			0.99	0.066	0.064	0.94
Adjust		0.89	0.048	0.049	0.39	0.99	0.047	0.045	0.93			0.99	0.047	0.045	0.93			0.99	0.047	0.045	0.93
Censored Normal		1.53	0.062	0.064	0.00	1.07	0.061	0.062	0.81			1.07	0.061	0.062	0.81			1.07	0.061	0.062	0.81
IPTW		0.88	0.060	0.059	0.47	0.99	0.052	0.050	0.93			0.99	0.052	0.050	0.93			0.99	0.052	0.050	0.93
Heckman		1.00	0.068	0.068	0.95	1.00	0.047	0.046	0.95			1.00	0.047	0.046	0.95			1.00	0.047	0.046	0.95

β_1										β_2											
Proportional Treatment		Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP	Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP			Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP			Estimate	SE_{Sim}	$\widehat{SE}_{HW}$	CP
Ignore		0.90	0.042	0.041	0.32	0.90	0.041	0.041	0.30			0.90	0.041	0.041	0.30			0.90	0.041	0.041	0.30
Exclude		0.90	0.069	0.069	0.66	0.99	0.066	0.064	0.94			0.99	0.066	0.064	0.94			0.99	0.066	0.064	0.94
Adjust		0.80	0.044	0.044	0.01	0.89	0.041	0.041	0.20			0.89	0.041	0.041	0.20			0.89	0.041	0.041	0.20
Censored Normal		1.61	0.065	0.065	0.00	1.02	0.062	0.064	0.95			1.02	0.062	0.064	0.95			1.02	0.062	0.064	0.95
IPTW		0.79	0.055	0.053	0.04	0.89	0.046	0.045	0.30			0.89	0.046	0.045	0.30			0.89	0.046	0.045	0.30

β_1				β_2					
Skewed Errors		Estimate	$\widehat{SE}_{\text{Sim}}$	$\widehat{SE}_{\text{HW}}$	CP	Estimate	$\widehat{SE}_{\text{Sim}}$	$\widehat{SE}_{\text{HW}}$	CP
Heckman		0.90	0.062	0.063	0.63	0.90	0.042	0.042	0.31

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript