Unobservable Selection and Coefficient Stability: Theory and Validation [*]

Emily Oster

University of Chicago Booth School of Business and NBER

August 9, 2013

Abstract

A common heuristic for evaluating the problem of omitted variable bias in economics is to look at coefficient movements after inclusion of controls. The theory under which this is informative is one in which the selection on observables is proportional to selection on unobservables. However, this connection is rarely made explicit and the underlying assumption is rarely tested. In this paper I first show how, under proportional selection, coefficient movements, *along with* movements in r-squared values, can be used to calculate a measure of omitted variable bias. I discuss practical details of implementation. I then undertake two empirical exercises to explore the performance of this adjustment in the data. First, I relate maternal behavior on child birth weight and IQ. Simple controlled regressions give misleading estimates; estimates adjusted with a proportional selection adjustment do significantly better. Second, I match observational and randomized trial data for 23 relationships in public health. I show that on average bias-adjusted coefficients perform much better than simple controlled coefficients and I suggest that a simple form of this adjustment could dramatically improve inference in many public health contexts.

[*] Ling Zhong, Unika Shrestha, Damian Kozbur and Guillaume Pouliot provided excellent research assistance. I thank David Cesarini, Raj Chetty, Todd Elder, Amy Finkelstein, Larry Katz, Matt Gentzkow, Matt Notowidigdo, Chad Syverson, Manisha Shah, Azeem Shaikh, Jesse Shapiro, Bryce Steinberg, Matt Taddy, Heidi Williams and participants in seminar at University of Chicago Booth School of Business, Wharton and Yale for helpful comments. I gratefully acknowledge financial support from the Neubauer Family.

1 Introduction

Concerns about omitted variable bias are common to most or all non-experimental empirical work in economics, other social sciences and the natural sciences. And although randomized experiments are common in natural sciences and becoming increasingly common within economics, the majority of empirical work in both settings is still not randomized.[1] Within economics, a common heuristic for evaluating the robustness of a result to omitted variable bias concerns is to look at the sensitivity of the treatment effect to added controls. The heuristic suggests that if a coefficient is stable as controls are added, this is a good sign that there is little remaining bias. In a review of non-structural, non-experimental empirical work published in three general interest economics journals[2] in 2012, 75% of papers explored the sensitivity of the results to varying control sets, and a number of these papers were quite explicit about the relationship between coefficient stability and omitted variable bias.[3]

Although it is rarely made explicit, this coefficient stability heuristic relies on the idea that the selection on observable covariates is informative about the selection on *unobservable* covariates, an idea which is formalized in Altonji, Elder and Taber (2005) and Murphy and Topel (1990). I will refer to this as the *proportional selection assumption.* In the context of a linear model, these papers show how this assumption can be used to calculate a causal treatment effect. Neither paper formalizes the link with coefficient movements.

The fact that the link between the proportional selection assumption and coefficient movements is not explicit creates two problems. First, the use of this as a robustness test is rarely done in the most informative way. Second, there has been little or no effort to test whether the proportional selection assumption is better than alternatives (for example, than the alternative that the unobservables are related to the treatment but there is no information provided about that relationship by the link between treatment and observables). The informativeness of robustness tests which rely on this proportional selection assumption rest crucially on whether it is empirically valid.

In this paper I take up both of these issues. I begin by expanding on the theory laid out in Altonji, Elder and Taber (2005) (hence, AET) and connecting the omitted variable bias directly to coefficient movements. I provide some explicit guidance for performing a bias adjustment based on this theory. I then present two validation exercises, both of which take advantage of settings in which I observe a "true" treatment effect matched to possibly biased estimates.

I begin in Section 2 with the formal theory. I consider the following model : $Y = \beta X + W_1 + W_2 + \epsilon$,

[1] For example: in 2012 *JAMA* published 133 major research papers, only 53 of which were randomized. The *American Journal of Public Health* published 128, only 14 of which were randomized. The combination of the *American Economic Review,* the *Quarterly Journal of Economics* and the *Journal of Political Economy* published 69 empirical, non-structural papers, only 11 of which were randomized.

[2] *American Economic Review, Journal of Political Economy* and *Quarterly Journal of Economics.*

[3] For example, Chiappori et al (2012) state: "It is reassuring that the estimates are very similar in the standard and the augmented specifications, indicating that our results are unlikely to be driven by omitted variables bias." Similarly, Lacetera et al (2012) state: "These controls do not change the coefficient estimates meaningfully, and the stability of the estimates from columns 4 through 7 suggests that controlling for the model and age of the car accounts for most of the relevant selection."

where W_1 is observed and W_2 is unobserved and the coefficient of interest is β. Note that β cannot be recovered from regression because of the unobserved elements in the model (this is the standard omitted variable bias issue). I introduce the *proportional selection assumption,* which formally links the relationship between X and the observed variables to the relationship between X and and the unobserved variables. This link invokes a degree of proportionality, which I denote δ. A value of $\delta = 1$ implies that the observed and unobserved variables are equally important in explaining X; $\delta < 1$ implies that observables are more important and $\delta > 1$ that the unobservables are more important.

Under this assumption I show that β can be recovered from: (1) the coefficients on X with and without controls for observed variables; (2) the r-squared values from controlled and uncontrolled regressions; (3) an assumption about the r-squared of a (hypothetical) regression which controlled for X, and both observed and unobserved variables; and (4) a value for the degree of proportionality, δ. The result shows that coefficient movements *do* relate to omitted variable bias, but they must be scaled by movements in the r-squared values.

Section 3 discusses implementation. I show how one can perform the baseline adjustment. I suggest that it would be simple for researchers to calculate a bounding value for δ – namely, the degree of proportionality which would be necessary to produce a treatment effect of zero – and this would be a natural replacement for heuristic statements that coefficient movements are "small".[4] I highlight two extensions. First, I discuss the case where there are additional controls which are unrelated to W_1. Second, I discuss the related heuristic of looking for stability in coefficients as additional controls are added. I then briefly address several estimation choices, including what controls should be included in W_1 and how to evaluate the appropriate r-squared of the hypothetical full regression.

Following the theory, I turn to two applications which explore how this procedure performs empirically. It is not possible to directly test the proportional selection assumption, but I argue I can test the assumption indirectly – and the methodology more generally – by asking whether the proportional selection adjustment improves inference.[5] I do this in two ways. First, by asking whether the more robust relationships (i.e. confirmed in better or randomized data) are those for which a higher value of δ would be required to produce a treatment effect of zero. Second, by asking whether a particular value of δ can match the magnitude of the observational result to true estimates.

In Section 4 I perform the primary validation exercise in the paper. This also provides a model of how this might be used empirically. I consider links between maternal behavior (prenatal and early life), child birth weight and child IQ. Many studies – in economics and elsewhere – have suggested links between maternal behaviors and child outcomes, but most studies are subject to significant concerns about omitted variable bias, notably associated with socioeconomic status. I use data from the National Longitudinal Survey of Youth

[4] A Stata command to perform this calculation (**psacalc**) is available.

[5] Altonji et al (2008) also compare results from their adjustment to randomized results in a single case (catheterization), although they consider only the test of the null hypothesis rather than comparing magnitudes. This general procedure is reminiscent of LaLonde (1986).

(NLSY) and US Natality Detail Files to estimate (1) the impact of breastfeeding, drinking in pregnancy and low birth weight/prematurity on child IQ and (2) the impact of maternal drinking and smoking on child birth weight.

I estimate regressions with and without controls for maternal socioeconomic status, and use the coefficients and r-squared values to perform the proportional selection adjustment. I draw estimates of the full model r-squared from published sibling correlations in IQ and birth weight; this is appropriate under the theory that the omitted variables are measures of family background and, therefore, the maximum variation that family background could hope to explain is captured by within family correlations.

Using this analysis for validation also requires an estimate of the true causal effect – at a minimum a conclusion about the null hypothesis and, for actual estimation of δ, a value for β. I draw on two sources. First, I use external evidence, some of it from randomized trials and others from comprehensive meta-analyses to draw conclusions about the null. Second, I run sibling fixed effects regressions in the NLSY to provide both null estimates and values for β. The assumption is that these are closer to the true causal effects, although I note they are subject to their own concerns. Perhaps comforting, the null conclusions are identical from either source.

In simple observational regressions with controls there are many "false positive" results. The proportional selection adjustment performs well. The relationships which require higher values of δ to produce a treatment effect of zero are more likely to be validated by external evidence and sibling fixed effects regressions. In all seven relationships estimated, there is a positive value of δ for which the adjusted coefficient matches the sibling fixed effects estimate. Further, all of the estimated δ values hover around 1 and I show that performing the proportional selection adjustment using a value of $\delta = 1$, with bootstrapped standard errors, would have led to much improved inference.

A broader application is to ask whether this procedure could be used to organize a set of results within an area of research and, ideally, provide guidance for improving inference when new results appear in that area. I consider this type of application in Section 5 using data on a number of settings in public health which link positive health behaviors to health outcomes. I argue this may be a fruitful area for this type of adjustment given that much of the existing literature relies on simple controlled regressions, and these results often turn out to be biased when better data becomes available. It is also an area of much policy interest.

I argue that a version of this procedure would be especially useful here if one could improve inference using *only* information from the feasible regressions. This would allow external researchers to evaluate results without having to make a detailed analysis of the full model r-squared in each setting. I therefore consider how this procedure performs when I assume that the amount of variation in Y explained by the unobservables is the same as the amount explained by the observables. This is a strong assumption; the goal here is to ask whether even with such a strong assumption we might draw better conclusions with the proportional selection adjustment.

With this additional assumption in place, I then undertake a validation exercise similar to the one in Section 4. I combine NHANES data (for observational correlations) with randomized evidence in two settings: the relationship between exercise and a number of health measures and the relationship between vitamin D/calcium (CaD) supplementation and a similar set of measures. I generate a total of 23 treatment-outcome pairs where I can estimate a relationship in the NHANES and match the point estimate to a treatment effect from a randomized trial.

As in Section 4, standard controlled regressions produce some false positives. I show that, on average, relationships in which a higher value of δ would be necessary to produce an adjusted effect of zero are more likely to be validated in randomized trials. I find that a single value of δ ($\delta = 0.971$) would decrease the overall error rate by around 27%. Many of the estimates which benefit most from this adjustment are ones in which the controlled coefficients overstate the benefit of the intervention and the adjusted coefficients match the truth. I perform several out-of-sample tests and show this performs well. I argue this adjustment may be applicable to a wide swath of the public health literature where the outcome is a health outcome, treatment is a good health behavior and we see only imperfect socioeconomic status controls. This adjustment would be easy for researchers (or research consumers) to perform, and could be helpful in evaluating the plausibility of results.

2 Theory

Consider the regression model

$$Y = \beta X + W_1 + W_2 + \epsilon \tag{1}$$

X represents the treatment and the coefficient of interest is β. W_1 and W_2 represent confounders. Specifically, W_1 is a vector which is a linear combination of observed control variables w_j^o multiplied by their true coefficients: $W_1 = \sum_{j=1}^{J_o} w_j^o \gamma_j^o$. W_2 is a vector which is a linear combination of *unobserved* control variables w_j^u, again multiplied by their true coefficients: $W_2 = \sum_{j=1}^{J_u} w_j^u \gamma_j^u$. I assume that $Cov(W_1, W_2) = 0$ and, without loss of generality, that $Var(X) = 1$. The covariance matrix associated with the vector $[X, W_1, W_2]'$ is positive definite.

Assume that $Cov(W_1, \epsilon) = 0$, $Cov(W_2, \epsilon) = 0$ and $Cov(X, \epsilon) = 0$. Denote the model (1) r-squared as R_{max}. Note that R_{max} may be less than 1 if Y is measured with error or there are components of the variation in Y that are orthogonal to X, W_1 and $W_{2.}$

Define the proportional selection relationship as $\delta \frac{\sigma_{1X}}{\sigma_{11}} = \frac{\sigma_{2X}}{\sigma_{22}}$, where $\sigma_{iX} = Cov(W_i, X)$, $\sigma_{ii} = Var(W_i)$ and δ is the coefficient of proportionality. I assume that $\delta > 0$ and refer to this as the *proportional selection assumption.* This implies that the relationship between X and the vector containing the observables is informative about the relationship between X and the vector containing the unobservables.

Define the coefficient resulting from the short regression of Y on X as $\mathring{\beta}$ and the r-squared from that regression as $\mathring{R}$. Define the coefficient from the intermediate regression of Y on X and W_1 as $\tilde{\beta}$ and the r-squared as $\tilde{R}$. Note these are in-sample values.

The omitted variable bias on $\mathring{\beta}$ and $\tilde{\beta}$ is controlled by the auxiliary regressions of (1) W_1 on X; (2) W_2 on X; and (3) W_2 on X and $W_{1.}$ Denote the in-sample coefficient on X from regressions of W_1 and W_2 on X as $\hat{\lambda}_{w_1|X}$ and $\hat{\lambda}_{W_2|X}$, respectively and the coefficient on X from a regression of W_2 on X and W_1as $\hat{\lambda}_{W_2|X,W_1}$. Denote the population analogs of these values $\lambda_{W_1|X}$, $\lambda_{W_2|X}$ and $\lambda_{W_2|X,W_1}$.

All estimates are implicitly indexed by n. Probability limits are taken as n approaches infinity. All observations are independent and identically distributed according to model (1). By standard omitted variable bias formulas, I can express the probability limits of the short and intermediate regression coefficients in terms of these values:

$$\begin{aligned}\mathring{\beta} &\xrightarrow{p} \beta + \lambda_{w_1|X} + \lambda_{W_2|X}\\ \tilde{\beta} &\xrightarrow{p} \beta + \lambda_{W_2|X,W_1}\end{aligned}$$

Lemma 1 defines the probability limit of the coefficient difference.

Lemma 1. $(\mathring{\beta} - \tilde{\beta}) \xrightarrow{p} \sigma_{1X} \frac{\sigma_{11}^2 - \sigma_{1X}^2(\delta\sigma_{22}+\sigma_{11})}{\sigma_{11}(\sigma_{11}-\sigma_{1X}^2)}$

Proof. This follows directly from the probability limits of the auxiliary regression coefficients under the proportional selection assumption. Proof details are in Appendix A. □

Denote the sample variance of Y as $\hat{\sigma}_{yy}$ and note that $\hat{\sigma}_{yy} \xrightarrow{p} \sigma_{yy}$. Lemma 2 defines probability limits for functions of the r-squared values.

Lemma 2. $(\tilde{R} - \mathring{R})\hat{\sigma}_{yy} \xrightarrow{p} \frac{\left[\sigma_{11}^2 - \sigma_{1X}^2(\sigma_{11}+\delta\sigma_{22})\right]^2}{\sigma_{11}^2(\sigma_{11}-\sigma_{1X}^2)}$ *and* $(R_{max} - \tilde{R})\hat{\sigma}_{yy} \xrightarrow{p} \frac{\sigma_{22}\left[\sigma_{11}^2 - \sigma_{1X}^2(\sigma_{11}+\delta^2\sigma_{22})\right]}{\sigma_{11}(\sigma_{11}-\sigma_{1X}^2)}$.

Proof. This follows directly from the auxiliary regression coefficients and Lemma 1. Proof details are in Appendix A. □

Define the following:

$$\beta^* = \begin{cases} \tilde{\beta} - \delta\left[\mathring{\beta} - \tilde{\beta}\right] \frac{R_{max}-\tilde{R}}{\tilde{R}-\mathring{R}} & \text{if } \delta{=}1\\ \tilde{\beta} - \left[\frac{\sqrt{[\mathring{\beta}-\tilde{\beta}]^2[\Theta^2+\Theta(4\delta(1-\delta)[\mathring{\beta}-\tilde{\beta}]^2[R_{max}-\tilde{R}])}]-\Theta[\mathring{\beta}-\tilde{\beta}]}{2(1-\delta)[\mathring{\beta}-\tilde{\beta}]^2[\tilde{R}-\mathring{R}]}\right] & \text{if } \delta \neq 1, \sigma_{1X} \geq 0\\ \tilde{\beta} - \left[\frac{-\sqrt{[\mathring{\beta}-\tilde{\beta}]^2[\Theta^2+\Theta(4\delta(1-\delta)[\mathring{\beta}-\tilde{\beta}]^2[R_{max}-\tilde{R}])}]-\Theta[\mathring{\beta}-\tilde{\beta}]}{2(1-\delta)[\mathring{\beta}-\tilde{\beta}]^2[\tilde{R}-\mathring{R}]}\right] & \text{if } \delta \neq 1, \sigma_{1X} < 0 \end{cases}$$

where $\Theta = \left(\left[\tilde{R} - \mathring{R}\right]^2 \hat{\sigma}_{yy} + \left[\mathring{\beta} - \tilde{\beta}\right]^2 \left[\tilde{R} - \mathring{R}\right]\right)$.

Proposition 1. $\beta^* \xrightarrow{p} \beta$.

Proof. I outline the proof here, with details in Appendix A. Recall that the bias of interest to calculate is $\hat{\lambda}_{W_2|X,W_1}$ which, under the proportional selection assumption and by Lemma 1, converges in probability to $\frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11}-\sigma_{1X}^2}$. δ is assumed to be known so the unknown variables are σ_{11}, σ_{22} and σ_{1X}.

By Lemmas 1 and 2 we have:

$$\begin{aligned}
\mathring{\beta} - \tilde{\beta} &\xrightarrow{p} \sigma_{1X}\frac{\sigma_{11}^2 - \sigma_{1X}^2(\delta\sigma_{22}+\sigma_{11})}{\sigma_{11}(\sigma_{11}-\sigma_{1X}^2)} \\
\left[\tilde{R} - \mathring{R}\right]\hat{\sigma}_{yy} &\xrightarrow{p} \frac{\left[\sigma_{11}^2 - \sigma_{1X}^2(\sigma_{11}+\delta\sigma_{22})\right]^2}{\sigma_{11}^2(\sigma_{11}-\sigma_{1X}^2)} \\
\left[R_{max} - \tilde{R}\right]\hat{\sigma}_{yy} &\xrightarrow{p} \frac{\sigma_{22}\left[\sigma_{11}^2 - \sigma_{1X}^2(\sigma_{11}+\delta^2\sigma_{22})\right]}{\sigma_{11}(\sigma_{11}-\sigma_{1X}^2)}
\end{aligned}$$

This defines a system of three equations in the three unknowns of interest. This system is identified and solving it completes the proof. □

For values of δ close to 1, the simple expression $\tilde{\beta} - \delta\left[\mathring{\beta} - \tilde{\beta}\right]\frac{R_{max}-\tilde{R}}{\tilde{R}-\mathring{R}}$ will be an approximation for β. The exact value diverges from this as δ gets significantly larger than 1.

I argue in the next section on implementation that a valuable statistic to report as robustness is the value of δ such that $\beta = 0$. This value can be obtained by rearranging the equations above and the formula appears in Appendix A.

It is worth briefly noting how the calculation suggested here differs from that suggested in AET. The proof that the bias is proportional to $\frac{\sigma_{22}\sigma_{1,X}}{\sigma_{11}-\sigma_{1,X}^2}$ is echoed in their work. A primary innovation here is to connect this bias to objects which are observable from regressions. AET suggest a methodology for calculating bounds on δ which relies on using the data directly (effectively, using the auxiliary regression coefficients). Their method recovers δ under the assumption that $\beta^* \to^p \beta$. This will be exact if $\delta = 1$, although only approximate in other cases, as they note. Further discussion is in Appendix B.1.

3 Implementation

The previous section formally establishes the link between coefficient movements with controls and omitted variable bias under the proportional selection assumption. This section discusses implementation. The first subsection details the implementation of the baseline result, and discusses two corollaries. The second discusses the details of several choices about parameters which are necessary for implementation.

3.1 Bias Adjustment Implementation

I am concerned here with the case where the true model is:

$$Y = \alpha + \beta X + W_1 + W_2 + \epsilon$$

andW_2 is unobserved to the researcher. Both W_1 and W_2 are indicies of variables multiplied by their true coefficients. W_1 contains any variables we observe and W_2 includes any variables unobserved to the researcher which are correlated with both Y and X.[6]

Since the elements of W_2 are unobserved one cannot estimate the true model. There are two regressions the researcher can observe in this case, shown in equations 2 and 3 below. The first controls only for X, the variable of interest. The second adds controls for the observed confounders. Each of these produces a coefficient on X.

$$Y = \hat{\alpha} + \mathring{\beta} X + \hat{\epsilon} \tag{2}$$

$$Y = \tilde{\alpha} + \tilde{\beta} X + \Psi W_1 + \tilde{\epsilon} \tag{3}$$

The commonly used coefficient movement heuristic would simply comment on the difference between $\mathring{\beta}$ and $\tilde{\beta}$.

In fact, these two coefficients *are* inputs to the bias calculation. The calculation also requires the r-squared values from these regressions. Referring back to proposition (1) above, equation (2) here recovers $\mathring{\beta}$, which is simply the OLS coefficient from a regression of Y on X alone, and $\mathring{R}$, which is the r-squared from that regression. Similarly, equation (3) recovers $\tilde{\beta}$ and $\tilde{R}$. Completing the bias calculation in proposition 1 additionally requires (a) a value of R_{max} and (b) a value of δ. Neither of these is recoverable directly from regressions, and in Section 3.2 below I will discuss the choice of these values.

To give some intuition for why the r-squared values matter, consider a setup with $\delta = 1$ and $R_{max} = 1$. This implies that (a) the treatment X is equally related to the observed and unobserved variables and (b) were we able to control of the unobservables, all variation in Y would be accounted for. Assume that when Y is regressed on X alone, the coefficient is 0.5 with an r-squared of 0.1 and, when controls are added, the coefficient moves to 0.4. Now consider two polar cases for the controlled r-squared. In Case 1, the r-squared value barely moves when controls are included – say, from 0.1 to 0.15. In this case, the remaining omitted variable bias is huge, because the omitted and included variables are equally related to X, but we expect the omitted variables to move the r-squared all the way from 0.15 to 1, even though the included variables moved

[6] Note that we denote W_1 as a vector here. In practice there may be multiple controls which are observed (and multiple unobserved). The assumption stated at the start of Section 2 is that W_1 is a linear combination of these controls multiplied by their true coefficients. Under this assumption one can operationalize this by including multiple variables as controls.

it only from 0.1 to 0.15.

The opposite case is when the controlled r-squared is 1.0. In this case, there is no bias, since the omitted controls do not explain any of Y. The same coefficient movement can imply wildly different values for the causal effect, depending on the movements in r-squared. Figure 1 shows how the resulting β estimate will vary in this example with the movement in controlled r-squared, even holding constant the coefficient differences.

Putting this intuition and equation into practice, I argue there are two calculations which may be of interest: a direct calculation of the bias, and a bounding argument on δ.

Direct Bias Calculation: Given both a value for R_{max} and δ, one can calculate the coefficient β which would result if one were able to control for the elements of W_2. The full equation for this calculation is given in Proposition 1. If δ is close to 1, then the simple calculation $\beta^* = \tilde{\beta} - \delta\frac{(\tilde{\beta}-\mathring{\beta})(R_{max}-\tilde{R})}{(\tilde{R}-\mathring{R})}$ will be a close approximation to β.

Bounding Argument on δ **:** A second way to use this adjustment – perhaps more akin to the use of the coefficient movement heuristic as robustness – is to adopt a value for R_{max} and calculate the value of δ which would produce $\beta = 0$. This is akin to asking how important the unobservables would need to be relative to the observables to eliminate the estimated effect. The bounding value for δ is calculated by simply setting $\beta = 0$ and solving for δ. The general formula for this is given in Appendix A; for δ close to 1 it is approximated by $\hat{\delta} = \frac{\tilde{\beta}(\tilde{R}-\mathring{R})}{(\tilde{\beta}-\mathring{\beta})(R_{max}-\tilde{R})}$.

Stata code accompanying this paper[7] preform this calculation.

These calculations are appropriate only under the proportional selection assumption – that is, only if the relationship between the observed controls and X is informative about the relationship between the unobserved controls and X. In regression form, the condition states that if an OLS regression of X on the index observed controls W_1 yields a coefficient of 1, then a regression of X on the index of unobserved controls W_2 would yield a coefficient of δ. There may be no *a priori* reason to think this is appropriate and testing it, at least in a limited set of contexts, is the focus on Sections 4 and 5 of the paper.

Section 3.2 below discusses the various choices necessary in doing this estimation – in particular, choices of W_1, R_{max} and, if necessary, δ. Before doing that, however, I briefly discuss two important extensions.

3.1.1 Additional Controls

This procedure recovers the coefficient which the researcher would estimate if the elements of W_2 could be observed. This need not be the causal impact of X on Y, however, and will not be if there are other important controls. Consider the case in which the true model is given by equation 4 below.

$$Y = \alpha + \beta X + W_1 + W_2 + \mathbf{m} + \epsilon \tag{4}$$

[7] The command is **psacalc** and is available through ssc.

and assume that the vector of variables $\mathbf{m}$ is correlated with both X and Y, but orthogonal to W_1 and W_2. If $\mathbf{m}$ is observed, it is possible to recover β using a variation on the procedure described above. In particular, one can estimate equations (2) and (3) including $\mathbf{m}$ in both regressions. The coefficient movement between the two regressions with and without W_1 can be used to recover β.

If $\mathbf{m}$ is unobserved, it is still possible to use the procedure to recover the value $\overline{\beta}$ which would result from estimation of equation (5) below[8], but note that because $\mathbf{m}$ is omitted, $\beta \neq \overline{\beta}$.

$$Y = \overline{\alpha} + \overline{\beta}X + \Psi_1 W_1 + \Psi_2 W_2 + \overline{\epsilon} \tag{5}$$

Appendix B.2 proves this result.

This discussion has two implications. First, this procedure recovers the treatment effect that would be estimated if we could control for the unobservables which are related to the observables. If there is another category of unobserved variables, this will not be the causal effect (although it could be close if $\mathbf{m}$ is relatively unimportant).

The second practical implication is to note that it may be that not all controls should be included as part of W_1. Controls which do not have a corresponding unobserved component should, instead, be included in both observed and unobserved regressions. An example of this would be something like sex: adjusting for sex is likely to matter for many applications, but since it is fully observed it may be inappropriate to assume that resulting coefficient movements reflect what would happen with additional controls. A corollary is that movements in the treatment effect when the components of $\mathbf{m}$ are introduced contain no information. Even very large movements in the treatment effect should not lead one to worry about the robustness of the result. The choice of what variables are in W_1 versus $\mathbf{m}$ is discussed in more detail in Section 3.2.1.

3.1.2 Stabilizing Coefficients

In cases where some controls make a lot of difference in estimates, researchers often invoke a stabilizing coefficient heuristic. This involves showing that although the first controls change the treatment effect a lot, as additional controls are added, the coefficient moves less. The assumption is then that any further controls would not move the coefficient much.

I capture this idea by assuming the true model is given by

$$Y = \alpha + \beta X + W_1^{*} + W_1^{**} + W_2 + \epsilon$$

where both W_1^{*} and W_1^{**} are observed, and W_2 is not. This heuristic involves first, observing that the

[8] Note that the coefficients on W_1 and W_2 are no longer equal to 1 but are biased by the exclusion of $\mathbf{m}$ through the joint correlation with X.

estimated impact of X changes significantly when controls for the elements of W_1^* are included. Then, second, adding controls for the element (or elements) of W_1^{**} and observing the coefficient changes relatively little. Third, concluding that further controls for W_2 would also move the coefficient relatively little.

It is clear, first, from the discussion above that this assumes that the proportional selection assumption holds for the elements of W_1^{**} and W_2 only. That is, it assumes that W_1^* does not have proportionally selected unobservables. Effectively, W_1^* behaves like $\mathbf{m}$ in Section 3.1.1. If, in fact, W_1^* should be considered part of the set which has related unobservables, one cannot learn much from this exercise.[9]

If we are willing to assume that this is appropriate – namely, that the final controls added are representative of the unobservables – then it is still necessary to consider the full bias adjustment. In this case, $\mathring{\beta}$ and $\mathring{R}$ come from the regression with X and W_1^* only and $\tilde{\beta}$ and $\tilde{R}$ come from the regression with controls for both W_1^* and W_1^{**}.

A key is to note that even if the coefficient difference is quite small, it may be scaled up by a vary large number if the r-squared does not move much and is not close to R_{max}. In some cases, the final control added may be relatively unimportant in the regression – moving the r-squared only a bit – and therefore the fact that the coefficient is relatively stable does not imply the adjustment is small.

This suggests that this particular heuristic should be taken with caution. The fact that it is possible to identify some control which, when added, does not move the coefficient much is meaningless on its own. It is informative only if (a) this last control is the one which is proportional to the unobservables and (b) the small coefficient movement is accompanied by a large r-squared movement *or* the r-squared after this control is close to or at R_{max}.

3.2 Parameter and Variable Choice

Calculating the bias adjustment described above requires the researcher make several choices: (1) what elements are in W_1 and what, if any, are in $\mathbf{m}$, (2) what is the value for R_{max} and (3) if one is interested in calculating the true β (rather than providing bounds), a value for δ. These three issues are briefly discussed below.

3.2.1 Choice of Controls

The vector of controls used in W_1 should include observed variables with related unobserved components. A common case empirically (and the one I consider in the applications below) would be one in which the primary omitted variables are components of socioeconomic status. In this case, the elements of W_1 should be whatever measures of socioeconomic status are observed by the researcher – education, income, etc. The

[9] This point relates closely to the discussion of this problem in Murphy and Topel (1990). They suggest there that researchers should consider which observable is "most like" their unobservables, and consider the coefficient movement after that is included.

proportional selection assumption is then that the elements of socioeconomic status the researcher doesn't see – for example, details about education, parental education, etc – relate to the treatment in a way proportional to how the observed elements relate.

In contrast, **m** should contain variables which are potentially important controls in the sense that they are correlated with X and Y but which do not have omitted components. One clear special case would be one of conditional random assignment. The variables on which assignment is conditioned should be included in **m**. It is likely these have large impacts on the estimate of treatment effects, but since assignment is random after they are adjusted for, we should not expect related unobserved controls to matter.

There are also cases in which some demographic elements should be included in **m**. To consider a concrete example: some of the analyses below considers child IQ as an outcome. In that case, because of the way the tests work, older children have higher scores. It's necessary to control for age to generate unbiased estimates. However, once we control for age that source of bias is gone.

A key subtlety arises when variables like age are, in fact, partially markers for the omitted categories. If there are cohort effects in economic circumstances, for example, then age may actually be best thought of as part of W_1 since it captures some element of socioeconomic status. Note that there is no requirement that anything be in **m**.

3.2.2 Choice of R_{max}

Recall that R_{max} is the r-squared from the model:

$$Y = \alpha + \beta X + W_1 + W_2 + \epsilon$$

The choice about R_{max} therefore relates to the variance of ϵ. Assuming β is the causal effect, what is captured in ϵ is one of two things: measurement error, or variables which impact Y but are uncorrelated with X.

Taking into account only the measurement error in Y will provide an upper bound on R_{max} and therefore generate a conservative estimate. Other downward adjustment for R_{max} may result from variation in Y which cannot be related to X. For example, if X is a long term medical treatment for cholesterol, day-to-day variation in cholesterol readings could not be explained by variation in X and the R_{max} should account for that.

3.2.3 Choice of δ

The value of δ defines the proportionality of selection. Recall that a value of $\delta = 1$ indicates equal selection, $\delta < 1$ implies that the unobservables are less important than the observables and visa versa for a value of $\delta > 1$. AET show that $E(\delta) = 1$ is appropriate if W_1 is a randomly drawn sample of the full set $\{W_1, W_2\}$ and argue that probably an assumption of $E(\delta) \leq 1$ is appropriate for most practical settings. Having said that, it is somewhat difficult to have an intuition about the value of δ.

A natural approach, noted above, is to use δ as part of a bounding argument – that is, report the value of δ which would produce a treatment effect of zero. On average, larger values of δ indicate more robust results.

3.3 Summary and Limitations

I argue that a bounding statement on δ would be a reasonable replacement for less precise comments about movements in treatment effects. It may seem that the requirements for this calculation– in particular, the need to make an assumption about R_{max}– are more onerous than simply looking at coefficient movements. However, it should be clear that without this assumption the coefficient movements are not meaningful either.

A brief final note: in a case in which one expects heterogeneous treatment effects along some dimension – age, for example, or sex – this method can be used to recover the treatment effect within each group. To do so, the analysis is run for each group separately, and the adjustment performed. Running a pooled model will, mechanically, recover the β one would get if one incorrectly ignored the heterogeneous treatment effect issue.

4 Validation: Maternal Behavior, Birth Weight and Child IQ

The results above provide a way to recover an estimate of “causal” treatment effects under the assumption that selection on observables and unobservables is proportional. This assumption is fairly strong and not directly testable. Indirectly, I can test the assumption – and the methodology more generally – by asking whether estimates generated by this procedure are closer to the true causal effect. Discussing that requires a setting in which I can match (possibly) biased estimates to some “true” estimate of a treatment effect.

Given such a setting, validation could take several forms. First, I can perform the bounding calculations on δ described above and ask whether relationships which require a higher value of δ to produce $\beta = 0$ are more likely to be true. Second, I can ask what value δ (if any) would match the adjusted effect from the observational regressions to the true treatment effect. Finally, a more constrained test is to ask whether a single value of δ might organize a number of findings. If yes, this would suggest at a minimum that this technique works well in comparing the robustness of multiple findings within a given setting.

In this section I undertake these validation tests in the context of the relationship between maternal behaviors, infant birth weight and child IQ. These relationships are of some interest in economics, and of wider interest in public health and public policy circles. A literature in economics demonstrates that health shocks while children are in the womb can influence early outcomes and later cognitive skills (e.g. Almond and Currie, 2011; Almond and Mazumder, 2011). A second literature, largely in epidemiology and public health, suggests that even much smaller variations in behavior – occasional drinking during pregnancy, not breastfeeding – could impact child IQ and birth weight. These latter studies, however, are subject to significant omitted variable concerns.

The key class of omitted variables relates to socioeconomic status. Women who drink or smoke during pregnancy tend to be of lower socioeconomic status, as are those women who do not breastfeed. Measures of socioeconomic status in standard datasets are useful but incomplete. Broadly, the idea here is to ask whether coefficient movements after inclusion of the observed socioeconomic status controls relate to the movements we would expect to see if we observed very precise information on socioeconomic status.

I consider five relationships in all: the relationship between child IQ and breastfeeding, drinking during pregnancy, low birth weight/prematurity and the relationship between birth weight (as the outcome) and maternal drinking and smoking in pregnancy. Section 4.1 below describes the data, Section 4.2 the empirical strategy and Section 4.3 the results.

4.1 Data

I use data from the National Longitudinal Survey of Youth Children and Young Adult Survey (NLSY) and data from the US Natality Detail Files (from 2001 and 2002).

NLSY

The NLSY is a longitudinal survey of women, and the Children and Young Adult module collects information on the children of NLSY participants. These data contain information on both IQ and birth weight. In the case of IQ, the outcome of interest is PIAT test scores for children aged 4 to 8. The treatments of interest are: months of breastfeeding, any drinking of alcohol in pregnancy and an indicator for being low birth weight and premature (<2500 grams and <37 weeks of gestation). These variables are summarized in the first rows of Panel A of Table 1.

For birth weight, the outcome is simply birth weight in grams. Here, I use all children. The treatments are whether the mother smokes in pregnancy and maternal drinking intensity during pregnancy. These variables are summarized in Panel B of Table 1.

The NLSY data also contain demographic controls. These are summarized in the remainder of Panel A and Panel B of Table 1 (I summarize these twice since the sample differs for the IQ and birth weight analyses). They include: child age and sex, race, maternal age, maternal education, maternal income and maternal marital status.

Natality Detail Files

The US Natality Detail Files contain data on all births in the US. I use data from 2001 and 2002 and focus on birth weight as the outcome. The treatments are, again, whether the mother smokes during pregnancy and maternal drinking intensity. I recode drinking data to match the NLSY definitions. The natality detail files also include demographics: child sex, maternal race, age, education and marital status. These data do not report income.

Panel C of Table 1 reports summary statistics.

4.2 Empirical Strategy

The primary issues in implementation include the choices about control sets (including possible variables in $\mathbf{m}$) and an assumption on the value of R_{max}. In addition, because I am concerned with validation here, it is necessary to have a measure of the true causal effect, which I denote β_{true}.

Since the primary concern here is omitted socioeconomic status, the variables included in W_1 are the standard observed socioeconomic status components: maternal education, income, race, marital status and age. In the vector of $\mathbf{m}$ controls I include child sex and, in the case of IQ, child age. These variables are important controls, but are unlikely to be related to omitted socioeconomic status.

Turning to R_{max} : in theory this should reflect how much of the variation in child IQ and birth weight could be explained if we had full controls for socioeconomic status. This is a figure for which we need to go outside the data. Neither IQ nor birth weight seem likely to have an R_{max} of 1. Even identical twins raised together do not have the same IQ scores or identical birth weight. I suggest that the appropriate figure in either case is the correlation between siblings raised together, which will capture the full effect of family background. For IQ, I use a value of 0.385, based on the average correlations from two studies reported in Scarr and Weinberg (1983).[10] For birth weight, I use a value of 0.5, drawn from Mazumder (2011).

Finally, the estimation requires an estimate of the true causal effect. One natural approach, in the spirit of Lalonde (1986), would be to match the observational analysis with evidence from randomized controlled trials which estimate similar parameters. This is not feasible here. Even in the two cases (breastfeeding and smoking) where I do have some randomized or quasi-random estimates on which to rely, the magnitudes are not comparable.

As an alternative, I undertake two approaches. First, I consider outside evidence in each case on the test of the null of treatment effect or not. Even for outcomes with no randomized trials, it is possible to get a sense from the literature about whether these effects are causal or not. Among the relationships I consider, randomized evidence suggests that breastfeeding is not linked with full-scale IQ (Kramer et al, 2008) and most

[10] This is consistent with other overview studies which suggest values in the range of 0.35 to 0.4 – see, for example, Bouchard and McGue, 2003.

evidence does not suggest an impact of occasional maternal drinking on child IQ (see, for example: Falgreen-Eriksen et al, 2012; O'Callaghan et al, 2007). In contrast, low birth weight and prematurity do seem to be consistently linked to low IQ (Salt and Redshaw, 2006), a link which also has a biological underpinning (de Kieviet et al, 2012). On the birth weight side, occasional maternal drinking is typically not thought to impact birth weight (Henderson, Gray and Brocklehurst, 2007), but there is better evidence that smoking does (e.g. from trials of smoking cessation programs as in Lumley et al, 2009).

With this evidence I can then ask – using the bounding argument described above – whether effects which would require a larger δ to generate $\beta = 0$ are those which are robust based on the outside evidence.

In a second approach, I take advantage of the family structure in the NSLY to estimate sibling fixed effect models and use the estimates of these – which should be purged of much of the family background variation – as the values of $\beta_{true.}$ This approach has the advantage that it can generate actual estimates of β_{true} rather than just a test of the null. On the other hand, sibling fixed effects estimates may also be subject to concerns about causality, and additional concerns about the endogenaity of parental investments. It is for this reason that I pursue both approaches.The sibling fixed effects echo the outside evidence in terms of which impacts are robust.

An issue throughout these analyses is the very likely chance that the treatments – smoking, drinking, breastfeeding – are measured with error. If this error is classical, it will of course not impact the coefficients. However, in this case it may be that much of the error is through under-reporting of bad behaviors (and over-reporting of good ones). This is likely to reinforce the omitted variable bias problem. If high socioeconomic status women are both less likely to smoke and less likely to admit to it then when we estimate the impact of "reported smoking" on child outcomes we will be even more biased than the estimates of actual smoking. Although there is no way for this procedure to address the measurement issue separately, to the extent that it operates like this example and relates to the same omitted variables, the procedure here will also help address this.

4.3 Results

Table 2 reports the results: Panel A shows data on child IQ from the NSLY, Panel B data on birth weight from the NLSY and Panel C data on birth weight from the Natality Detail Files.

The first two columns in each panel show estimated treatment effects and r-squared values with only sex (or age and sex in the case of IQ) as controls. Columns 3 and 4 show similar treatment effects with the full control set. More breastfeeding is associated with higher IQ in these regressions, and low birth weight is associated with lower child IQ. More maternal drinking appears in these data to be associated with *higher* child IQ later. There is no biological reason to think this is the case: it *must* be due to selection. Both samples show smoking and drinking are associated with lower birth weight. All seven analyses reported here show significant effects with the full set of controls. Interpreting these results in a naive way, one would conclude that each has a significant link with child outcomes.

Column 5 reports whether external evidence, summarized above, suggests a causal impact. As noted, low birth weight does seem to be linked to IQ and smoking is linked to low birth weight, but the other relationships do not have broad support. In Column 6 I then combine the regression estimates from the first columns with the estimates of R_{max} (0.385 in the case of IQ, 0.5 in the case of birth weight) and calculate the value of δ for which $\beta = 0$. This is the value I suggested reporting as a summary of the robustness of the results. The evidence in this column provides support for the value of this adjustment: the relationships which appear to be causal based on outside evidence are associated with larger values of δ.

Column 7 shows the sibling fixed effects estimates; in Panel C, I report the estimates drawn from the NLSY for these outcomes, since the natality files do not link mothers across births. The positive impacts of breastfeeding and maternal drinking are eliminated. The impact of low birth-weight and prematurity on IQ remains fairly large – about 0.10 standard deviations – but has a p-value of 0.11. In the case of birth weight, the impact of smoking on child birth weight remains strongly significant in these regressions, but there is no measured impact of maternal drinking. These results – the lack of an impact for breastfeeding and maternal drinking, the possible impact of low birth weight on child IQ and the strong impact of smoking on birth weight – line up well with the conclusions on null hypotheses reported in Column 5.

Column 8 calculates the value of δ which would match the β_{true} estimated from sibling fixed effects regressions. In all seven rows this δ is defined and is positive. That is, these all pass the most basic validation test: the coefficients move toward the truth when the controls are added and there is therefore some value of δ which would match. The values of δ range between 0.5 and 1.5.

Finally, Column 9 asks whether a single value of δ would generate better inference across all these settings. I use a value of $\delta = 1$. This is done for two reasons. First, it seems a natural focal point. Second, looking at the values in Columns 6 and 8, this would appear to fit well. Standard errors in this column are calculated with a bootstrap over individuals, although it is worth keeping in mind that these are sensitive to sample size. The coefficient moves closer to the sibling fixed effect result in all cases. After the adjustment only the impacts of smoking remain significant and sizable.

Coefficient Stability

The above analysis suggests that performing the proportional selection adjustment improves the conclusions. It seems useful to consider whether a similar conclusion could have been reached from using the "coefficient stability" heuristic. To do this, for each treatment I run regressions progressively including controls. I choose the order of controls by ranking the demographics based on the amount of variation in child IQ or birth weight that they explain in the data. I include these controls in the same order for each analysis within outcome (the order differs for IQ and birth weight). Figures 2a-2g show coefficients and r-squared values for the seven analyses.

These figures suggest coefficient stability is not useful distinguishing among these analyses, perhaps not surprising given the discussion in 3.1.2. All of the graphs show a very similar pattern of stabilizing coefficients. Based on these alone it would be quite difficult to identify some of the relationships as more robust than the others. In line with the discussion in 3.1.2, the issue is clear: the r-squared in the fully controlled regressions here is around 0.25 for IQ and less than 0.1 for birth weight, far below the figures of 0.385 or 0.5 that were drawn from existing data. Given this, the fact that the coefficient has stabilized is not fully informative.

Summary

The results in this section – in particular, in Table 2 – are quite supportive of this approach. It passes the most basic validation test by showing that the bounding calculation can identify more versus less robust results. Perhaps more surprisingly, the results show that a single value of δ ($\delta = 1$) performs reasonably well. Returning to the question of applications in economics, this suggests support for the coefficient movement robustness test. However, it also makes clear the importance of taking into account the r-squared movements. If we based our analysis only on the size (say, in percent terms) of the coefficient movements we would conclude the link between drinking and low birth weight is much more robust than the link between low birth weight and IQ since the former moves only 10% and the latter 30%. In fact, the low birth weight and IQ link is more robust – the bias-adjusted coefficient is much larger and is significant at the 11% level – which is reflective of the much larger change in r-squared and lower R_{max}.

5 Application: Health Behaviors and Health Outcomes

The discussion in Section 4 provides a model for performing this adjustment in a single setting with a carefully considered R_{max} and suggests that procedure can improve inference and speak to the robustness of results. A broader application of this is to ask whether this procedure could be used to organize a set of results within an area of research and, ideally, provide guidance for improving inference when new results appear in that area. In this section I use data on several settings in the area of public health.

I first show that a simplified version of this procedure (which, importantly, does not require an R_{max} for each outcome) can organize a number of results and improve inference; I do this by comparing observational data to randomized trials for 23 outcome-treatment relationships.[11] Second, I derive a best-fit *magnitude* value for this adjustment which can be used in this particular area. Finally, I provide several out-of-sample

[11] This is in the spirit of LaLonde (1986).

tests of this for similar types of relationships. The precise magnitude adjustment I supply could be used by researchers and research consumers to evaluate new results in this area. It will not, of course, be portable to very different settings (I would not suggest using this in estimates of returns to schooling, for example), although a similar procedure could be used to derive adjustment magnitudes in those settings.

The area I consider is the relationship between positive health behaviors and health outcomes, a topic of much interest in public health. Do individuals who exercise live longer? Does taking a vitamin supplement lower your blood pressure? Observational studies in this literature suffer from clear omitted variable bias problems, largely stemming from correlations between high socioeconomic status and both positive health behaviors and good health outcomes. Likely due to this issue, when randomized studies are run to look at similar questions the results are often at odds with what was seen in observational data. A classic example is the exploration of the link between diet and health. For years the medical profession recommended a low-fat, high carbohydrate diet as a key to better health. It turned out this was based on biased estimates. When randomized data from a large study was released in 2006, this result was seriously weakened (Prentice et al, 2006; Beresford et al, 2006; Howard et al, 2006).

Given that many of the central issues facing this literature can be boiled down to omitted variable bias, it seems a natural area for which to ask whether this procedure could improve conclusions. A significant barrier to using this, especially as a research consumer, is the need for carefully considering R_{max} in each setting. I suggest here that a general assumption about R_{max} could substitute so adjustment might be performed using *only* results available from regressions. This naturally will cause some loss of information relative to a careful consideration of R_{max} in each setting; whether it still improves inference is an empirical question.

Although $R_{max} = 1$ may seem a natural assumption, it seems unlikely to apply here since many of the outcomes (like lipids, blood pressure, etc) vary within an individual even over the course of a day.[12] Instead, I adopt the assumption that the unobservables explain as much of the variation in the outcome as the observables do. Formally, this means that the increase in r-squared when adding the unobservables would be equal to the increase when the observables were added: $R_{max} = \tilde{R} + (\tilde{R} - \mathring{R})$.

Using this assumption, I then combine observational and randomized estimates of parallel relationships and ask the same questions as in Section 4. First, are estimates which require a higher value of δ to produce $\beta = 0$ more likely to have been confirmed in randomized data? Second, is there a value of δ in each case which allows me to match the adjusted coefficient to the randomized effect. And, finally, is there a value of δ which could be generally applied across all the settings to improve inference on average? This final magnitude conclusion is tested in several out-of-sample relationships.

[12] Demacker et al (1982), for example, show an intra-individual coefficient of variation for triglycerides of 35% within a day.

5.1 Data

This section considers two treatments: exercise and vitamin D+calcium supplementation. In each case I consider the relationship between the treatment and a range of outcomes. This analysis requires two pieces of data: randomized trial results and observational data.

Randomized Trials

Randomized trial results are drawn from existing work.

Exercise Evidence on the impact of exercise is drawn from several papers which are summarized in a Cochrane Review meta-analysis (Shaw et al, 2006). I consider only studies which compared exercise to no exercise (this excluded studies which also used diet). Outcomes considered include weight, blood pressure, cholesterol, blood glucose and triglycerides.

Vitamin D and calcium Evidence on the impact of vitamin D and calcium supplementation comes from the Women's Health Initiative, a large scale study of post-menopausal women which has run a number of important interventions. One trial within the study involved randomizing women into receiving vitamin D and calcium supplements (treatment) or not (control). Outcomes include bone density, lipids, blood pressure, exercise, and weight.

In Appendix Table A.1 I list the citation for each outcome-treatment pair, the treatment and any restrictions on age or gender in the study recruitment.

Observational Data

Exercise Exercise data are drawn from the National Health and Nutrition Examination Survey (NHANES), Wave III. Individuals are asked detailed questions about exercise. I use this to create a treatment measure as close as possible to the treatment in each study. In most cases the study includes some kind of jogging three times a week. Exact populations used are listed in Column 3 of Appendix Table A.1 for each paper, but in general these tend to focus on middle-aged individuals. Exercise data and the outcomes variables considered are summarized in Panel A of Table 3.

Vitamin D and calcium Data on vitamin D and calcium supplementation also comes from the NHANES-III. Individuals are asked about vitamin and mineral supplements, which allows me to create an indicator for taking vitamin D and calcium supplementation. To match the Women's Health Initiative data I use women aged 55 to 85 (recruitment in this study is women 50 to 80, but evaluation is several years later). Summary statistics on share of women using supplements and outcomes variables are in Panel B of Table 3.

Magnitudes

A central issue here is how to compare magnitudes across these settings. The observational coefficients estimate the impact of actually engaging in the behavior. This directly maps to the randomized data only if everyone is compliant, or if we observe average treatment effects and there is no heterogeneity across individuals. In the case of exercise, the three studies used in the data were very intensive with extremely high compliance (they were also fairly small). I therefore use the ITT estimates as the treatment effects. In the case of the vitamin D+calcium supplementation, I use adherence data to generate ATE estimates under the assumption that the control group did not use supplements. There is no way to address heterogeneity here, other than to argue that because the group is exclusively post-menopausal women, this issue is hopefully limited.

It is worth noting that these issues with magnitude comparisons do not spill over to comparisons with the null, so in the case where one is uncomfortable with the use of these magnitude data, the evidence on the null may still be informative.

5.2 Empirical Strategy

As noted, in this section I employ the assumption that $R_{max} = \tilde{R} + (\tilde{R} - \mathring{R})$. In this case, β_{true} is drawn from randomized trial results. As in the analysis above there is an important choice of what is in W_1 and what will be in $\mathbf{m}$. I include in W_1 the standard socioeconomic status measures: education, income, marital status and race. This reflects the observation that the bulk of the omitted variable issue are likely to be socioeconomic status. In $\mathbf{m}$ I include age dummies and sex and, in cases where the outcome is weight in kilograms, measures of height.

The first step below is to calculate, for each relationship, the value of δ which produces $\beta_{adj} = 0$ in each case, and evaluate whether higher values of δ are associated with more robust results. The second step is to estimate a value of δ such that $\beta_{adj} = \beta_{true.}$ This value will be positive as long as controls move the coefficient towards the true β.

In a third step I estimate the single value of δ which provides the best fit across all settings. For outcome-treatment pair i, denote the adjusted coefficient $\beta^i_{adj}(\delta)$ and the true effect β^i_{true}. The trial also produces a standard error, denoted σ^i. I calculate the difference between the bias-adjusted and true coefficient, scaled by the standard error. I sum these over the outcome-treatment pairs and minimize the sum over the choice of δ. Formally, I solve:

$$\hat{\delta} = argmin_{\delta} \sum_i \left(\frac{\beta^i_{adj}(\delta) - \beta^i_{true}}{\sigma^i} \right)^2$$

Given this value it is then possible to explore the performance of this adjustment in several ways. First, I can compare the magnitude of the error under the maximum likelihood value of δ relative to the assumption that $\delta = 0$ (which is the benchmark controlled regression coefficient). Second, I can compare the performance on each outcome-treatment pair, using bootstrapped standard errors, and ask whether I would have drawn more accurate conclusions about the null hypothesis from the adjusted analysis. Finally, I perform several out-of-sample tests using these outcomes and exploring whether the same adjustment would lead to more accurate conclusion in these cases.

5.3 Results

The first five columns of Table 4 show the first step of the results. Column 1 lists the outcome and, in the case of exercise where there are typically multiple studies per outcome, information on the citation. The second and third columns list the uncontrolled and controlled effects, their standard errors and the r-squared values. The effects are significant in many but certainly not all cases, and generally in the expected direction, with exercise and vitamin supplementation linked to improved health outcomes.

Column 4 reports whether randomized evidence rejects the null of no effect; in the case of exercise, this conclusion is drawn from the meta-analysis in Shaw et al (2006). Fewer of the effects are significant than in the observational data. Column 5 reports the value of δ such that $\beta = 0$. On average, higher values of δ are linked with more robust results. The average value of δ for results which have support in randomized data is 1.69, versus 1.84 for those without support. The pattern is certainly less consistent than in Section 4, likely for two reasons. First, I have introduced the assumption on R_{max}, which is at best an approximation. Second, given the small sample sizes here, some of the results which are not significant are still quite large. If the real goal is to match the randomized magnitudes, matching zero may go too far.

Columns 6 and 7 move to matching magnitudes. Column 6 reports the magnitude of the impact from the randomized trial.[13] Column 7 reports the value of δ such that $\beta_{adj} = \beta_{true.}$ In 17 of 23 outcomes there is a positive value of δ such that this holds. The cases in which there is no match – i.e. the coefficients move the wrong way – are all ones where the observational effect is not significant and neither is the randomized effect. These are inherently somewhat noisy, which makes it perhaps less surprising that the coefficient movements are not informative. If I ask the broader question of whether a positive value of δ could generate estimates inside the randomized confidence interval, the answer is yes in all cases.

Turning to the third step, the full estimation procedure described above yields a value of $\hat{\delta} = 0.971$. This suggests very close to equal selection and is surprisingly close to what I suggested as a good fit in Section 4. I can illustrate the overall impact of this bias adjustment. To do so I re-scale each outcome so the 95% confidence interval from the randomized trial ranges from 0 to 1 (and thus the randomized point estimate is close to 0.5); this is necessary for visualization since the scale of the effects varies widely across outcomes. I then convert first the standard controlled coefficient and then the bias-adjusted coefficient onto this scale. Figure 3a shows the interval for the randomized trial (open circles) and the controlled coefficient (filled in circle). Although the controlled and true coefficient are similar in some cases, especially when they are both close to zero, in others the controlled coefficient is wildly outside the confidence interval.

[13] Note that in some cases in exercise the individual effect is not significant even if we report rejection of the null. This is because the null rejection was based on the Cochrane Review meta-analysis (Shaw et al, 2006).

Figure 3b shows the coefficients after the bias adjustment is done with the value of $\delta = 0.971$. The fit is significantly better (note I have retained the large scale for ease of comparison). In a number of cases where the controlled coefficient showed significant errors – for example, the impact of vitamin supplementation on weight and exercise – the adjusted coefficients are within or very close to the confidence interval. The overall error is significantly smaller in the bias adjustment case – a reduction of 27% on average.

The final column of Table 4 describes this result numerically: I perform the bias adjustment with $\delta = 0.971$, and generate standard errors using a bootstrap over individuals. Again, it's worth taking the standard errors with caution since the observational studies here are, in some cases, significantly underpowered to pick up impacts of the size seen in randomized trials. The bias-adjusted impacts are much closer to the estimates from the randomized data, on average.

This table makes clear much of the value in the adjustment comes in cases where the controlled coefficients lead to false positive conclusions, or at least to an overstatement of the magnitude of the impact. For example, the controlled coefficients suggest a large and significant impact of vitamin supplementation on exercise[14], whereas the bias-adjusted coefficient is very close to the small and insignificant impact estimated in randomized trials. At the same time, the bias-adjustment retains significant effects in many of the cases where there are large and significant effects estimated in randomized trials – for example, the impact of exercise on weight, blood pressure and some measures of heart health.

Out of Sample Tests

Within sample, the adjusted coefficients above are closer to the true treatment effect than the controlled coefficients. An important related test is to ask how these perform out of sample.

A simple way to explore this is to perform an out of sample test *within* sample. More specifically, I drop each outcome-treatment pair in turn, re-estimate the best fit $\hat{\delta}$ and then apply the new δ to the dropped relationship. I can then ask whether, on average, the error in that estimate would be diminished. The $\hat{\delta}$ values estimated range from 0.953 to 1.059, all very close to the full sample value. Not surprisingly, then, in line with the evidence in Table 4, the error reduction is 27% on average with the adjustment.

I consider two other out-of-sample tests. In the case of exercise and vitamin D there are several outcomes for which randomized experiments have reached a conclusion about the null but where magnitude comparisons are difficult. This may be due to differences in the timing of follow-up, the fact that randomized effects are reported as odds ratios or because generating an exactly parallel analysis is difficult. However, given the adjustment value estimated above it is possible to return to these outcomes and explore whether the adjustment procedure used here leads to correct conclusions in these cases.

[14] The theory under which this might matter is that calcium and vitamin D increase bone health, which improves ability to exercise.

This is done in Panels A and B of Table 5. This table is structured similarly to Table 4 except that in the third column I simply report the hypothesized direction and significance (or not) of the effect in the randomized trial. In general, the bias-adjustment also performs well here. In the case of exercise, the controlled coefficients show significant impacts on both diabetes and mortality (among individuals with heart disease), and the bias-adjusted coefficients correctly identify only the mortality evidence as robust. In the case of vitamin D the controlled coefficients incorrectly suggest supplementation matters for mortality, a result which is corrected by the bias-adjustment.

A second out-of-sample tests relies on another study, the Physician Health Study (PHS). This work evaluated the impact of beta-carotene, vitamin E and vitamin C on heart disease mortality among men.[15] Published results from the study reject links between mortality and any of these vitamins (Hennekens et al, 1996; Sesso et al, 2008). Because the outcome is mortality and magnitudes are therefore difficult to link, I could not use this study in the estimation, but it is possible to use as an out of sample test. The NHANES-III provides the data, as above.

Panel C of Table 5 shows this evidence. vitamin E and vitamin C are both linked to lower mortality in the controlled regressions but not (at least not significantly) in the bias-adjusted coefficients. This provides further support.

Summary

The final conclusion of this discussion is that, across a range of settings in this area of public health inference might be improved with a very simple adjustment. Since the best fit value of $\hat{\delta}$ is close to 1, the formula for the bias adjustment with $\delta = 1$ will be a very close approximation. This dramatically simplifies the formula. Given a controlled coefficient $\tilde{\beta}$ and uncontrolled coefficient $\mathring{\beta}$, a value of $\beta_{adj} = \tilde{\beta} - 0.971(\mathring{\beta} - \tilde{\beta})$ would be closer, on average, to the true treatment effect.[16] It would be inappropriate, of course, to port this particular value to other areas (for example those closer to economics) but this does provide a model for how an adjustment might be constructed in other settings.

6 Conclusion

The goal of this paper is two-fold. First, I connect the popular robustness heuristic of exploring coefficient sensitivity to controls to the proportional selection assumption formalized in Altonji, Elder and Taber (2005)

[15] This study also evaluated (and supported) the importance of aspirin in preventing heart disease mortality. However, the observational evidence on aspirin is marred by both the omitted variable issue but much more so by a problem of reverse causality. It has long been thought that aspirin was good for heart disease so the kind of people who take it tend to be those with heart disease. This problem crops up in most of the settings I consider but to a much, much lesser extent. When facing this problem a bias adjustment of this type will not address the issue. I therefore do not use this as a test.

[16] This relies only on coefficient movements, and it is worth noting that this is the adjustment that Bellows and Miguel (2009) suggest. It is appropriate under this assumption about R_{max} and the assumption that δ is close to 1.

and Murphy and Topel (1990). I provide some guidance to discipline the use of this coefficient movement heuristic and give a simple form of the adjustment using information on coefficient and r-squared values. I suggest an alternative to commenting on the magnitude of coefficient movements would be to report the degree of proportionality between observables and unobservables which would produce a treatment effect of zero. Second, I explore the performance of this adjustment in the data. I find that, in fact, an adjustment of this form does get closer to causal effects.

In the final section I argue that an even simpler form of this adjustment – one which can therefore be performed using only the coefficient estimates – could improve inference in public health. This suggests a simple way for researchers in parallel settings (i.e. those where the outcome of interest is a health outcome and the treatment is a positive health behavior) to evaluate the plausibility of their results, and for readers of published work to do so, as well.

References

Almond, Douglas and Bhashkar Mazumder, "Health Capital and the Prenatal Environment: The Effect of Ramadan Observance during Pregnancy," *American Economic Journal: Applied Economics*, October 2011, *3* (4), 56–85.

___ **and Janet Currie**, "Killing Me Softly: The Fetal Origins Hypothesis," *Journal of Economic Perspectives*, Summer 2011, *25* (3), 153–72.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 2005, *113* (1), 151–184.

___**, Todd Elder, and Christopher R. Taber**, "Using Selection on Observed Variables to Assess Bias from Unobservables When Evaluating Swan-Ganz Catheterization," *American Economic Review*, 2008, *98* (2), 345–50.

Anderssen, S. A., I. Hjermann, P. Urdal, P. A. Torjesen, and I. Holme, "Improved carbohydrate metabolism after physical training and dietary intervention in individuals with the "atherothrombogenic syndrome'. Oslo Diet and Exercise Study (ODES). A randomized trial," *J. Intern. Med.*, Oct 1996, *240* (4), 203–209.

Bellows, John and Edward Miguel, "War and local collective action in Sierra Leone," *Journal of Public Economics*, December 2009, *93* (11-12), 1144–1157.

Beresford, Shirley et al., "Low-Fat Dietary Pattern and Risk of Colorectal Cancer," *JAMA*, 2006, *295* (6), 643–654.

Bouchard, T. J. and M. McGue, "Genetic and environmental influences on human psychological differences," *J. Neurobiol.*, Jan 2003, *54* (1), 4–45.

Brunner, R. L., B. Cochrane, R. D. Jackson et al., "Calcium, vitamin D supplementation, and physical function in the Women's Health Initiative," *J Am Diet Assoc*, Sep 2008, *108* (9), 1472–1479.

___**, J. Wactawski-Wende, B. J. Caan et al.**, "The effect of calcium plus vitamin D on risk for invasive cancer: results of the Women's Health Initiative (WHI) calcium plus vitamin D randomized clinical trial," *Nutr Cancer*, 2011, *63* (6), 827–841.

Caan, B., M. Neuhouser, A. Aragaki et al., "Calcium plus vitamin D supplementation and the risk of postmenopausal weight gain," *Arch. Intern. Med.*, May 2007, *167* (9), 893–902.

Chiappori, Pierre-Andrei, Sonia Oreffice, and Climent Quintana-Domeque, "Fatter Attraction: Anthropometric and Socioeconomic Matching on the Marriage Market," *Journal of Political Economy*, 2012, *120* (4), 659 – 695.

de Boer, I. H., L. F. Tinker, S. Connelly et al., "Calcium plus vitamin D supplementation and the risk of incident diabetes in the Women's Health Initiative," *Diabetes Care*, Apr 2008, *31* (4), 701–707.

de Kieviet, J. F., L. Zoetebier, R. M. van Elburg, R. J. Vermeulen, and J. Oosterlaan, "Brain development of very preterm and very low-birthweight children in childhood and adolescence: a meta-analysis," *Dev Med Child Neurol*, Apr 2012, *54* (4), 313–323.

Demacker, P. N., R. W. Schade, R. T. Jansen, and A. Van 't Laar, "Intra-individual variation of serum cholesterol, triglycerides and high density lipoprotein cholesterol in normal humans," *Atherosclerosis*, Dec 1982, *45* (3), 259–266.

Falgreen-Eriksen, H. L., E. L. Mortensen, T. Kilburn, M. Underbjerg, J. Bertrand, H. Stavring, T. Wimberley, J. Grove, and U. S. Kesmodel, "The effects of low to moderate prenatal alcohol exposure in early pregnancy on IQ in 5-year-old children," *BJOG*, Sep 2012, *119* (10), 1191–1200.

Henderson, J., R. Gray, and P. Brocklehurst, “Systematic review of effects of low-moderate prenatal alcohol exposure on pregnancy outcome,” *BJOG*, Mar 2007, *114* (3), 243–252.

Hennekens, C. H., J. E. Buring, J. E. Manson, M. Stampfer, B. Rosner, N. R. Cook, C. Belanger, F. LaMotte, J. M. Gaziano, P. M. Ridker, W. Willett, and R. Peto, “Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease,” *N. Engl. J. Med.*, May 1996, *334* (18), 1145–1149.

Heran, B. S., J. M. Chen, S. Ebrahim, T. Moxham, N. Oldridge, K. Rees, D. R. Thompson, and R. S. Taylor, “Exercise-based cardiac rehabilitation for coronary heart disease,” *Cochrane Database Syst Rev*, 2011, (7), CD001800.

Howard, Barbara et al., “Low-Fat Dietary Pattern and Risk of Cardiovascular Disease,” *JAMA*, 2006, *295* (6), 655–666.

Howe, T. E., B. Shea, L. J. Dawson et al., “Exercise for preventing and treating osteoporosis in postmenopausal women,” *Cochrane Database Syst Rev*, 2011, (7), CD000333.

Jackson, R. D., N. C. Wright, T. J. Beck et al., “Calcium plus vitamin D supplementation has limited effects on femoral geometric strength in older postmenopausal women: the Women’s Health Initiative,” *Calcif. Tissue Int.*, Mar 2011, *88* (3), 198–208.

Kramer, M. S., F. Aboud, E. Mironova et al., “Breastfeeding and child cognitive development: new evidence from a large randomized trial,” *Arch. Gen. Psychiatry*, May 2008, *65* (5), 578–584.

Lacetera, Nicola, Devin G. Pope, and Justin R. Sydnor, “Heuristic Thinking and Limited Attention in the Car Market,” *American Economic Review*, August 2012, *102* (5), 2206–36.

LaCroix, A. Z., J. Kotchen, G. Anderson et al., “Calcium plus vitamin D supplementation and mortality in postmenopausal women: the Women’s Health Initiative calcium-vitamin D randomized controlled trial,” *J. Gerontol. A Biol. Sci. Med. Sci.*, May 2009, *64* (5), 559–567.

LaLonde, Robert J, “Evaluating the Econometric Evaluations of Training Programs with Experimental Data,” *American Economic Review*, September 1986, *76* (4), 604–20.

Lumley, J., C. Chamberlain, T. Dowswell, S. Oliver, L. Oakley, and L. Watson, “Interventions for promoting smoking cessation during pregnancy,” *Cochrane Database Syst Rev*, 2009, (3), CD001055.

Margolis, K. L., R. M. Ray, L. Van Horn et al., “Effect of calcium and vitamin D supplementation on blood pressure: the Women’s Health Initiative Randomized Trial,” *Hypertension*, Nov 2008, *52* (5), 847–855.

Mazumder, Bhashkar, “Family and Community Influences on Health and Socioeconomic Status: Sibling Correlations Over the Life Course,” *The B.E. Journal of Economic Analysis & Policy*, 2011, *11* (3), 1.

Murphy, Kevin and Robert Topel, “Efficiency Wages Reconsidered: Theory and Evidence,” in “Advances in the Theory and Measurement of Unemployment” 1990, pp. 204–240.

O’Callaghan, F. V., M. O’Callaghan, J. M. Najman, G. M. Williams, and W. Bor, “Prenatal alcohol exposure and attention, learning and intellectual ability at 14 years: a prospective longitudinal study,” *Early Hum. Dev.*, Feb 2007, *83* (2), 115–123.

Orozco, L. J., A. M. Buchleitner, G. Gimenez-Perez, M. Roque I Figuls, B. Richter, and D. Mauricio, “Exercise or exercise and diet for preventing type 2 diabetes mellitus,” *Cochrane Database Syst Rev*, 2008, (3), CD003054.

Prentice, Ross et al., “Low-Fat Dietary Pattern and Risk of Invasive Breast Cancer,” *JAMA*, 2006, *295* (6), 639–642.

Rajpathak, S. N., X. Xue, S. Wassertheil-Smoller et al., “Effect of 5 y of calcium plus vitamin D supplementation on change in circulating lipids: results from the Women’s Health Initiative,” *Am. J. Clin. Nutr.*, Apr 2010, *91* (4), 894–899.

Rossom, R. C., M. A. Espeland, J. E. Manson et al., "Calcium and vitamin D supplementation and cognitive impairment in the women's health initiative," *J Am Geriatr Soc*, Dec 2012, *60* (12), 2197–2205.

Salt, A. and M. Redshaw, "Neurodevelopmental follow-up after preterm birth: follow up after two years," *Early Hum. Dev.*, Mar 2006, *82* (3), 185–197.

Scarr, Sandra and Richard Weinberg, "The Minnesota Adoption Studies: Genteic Differences and Malleability," *Child Development*, 1983, *54* (2), 260–267.

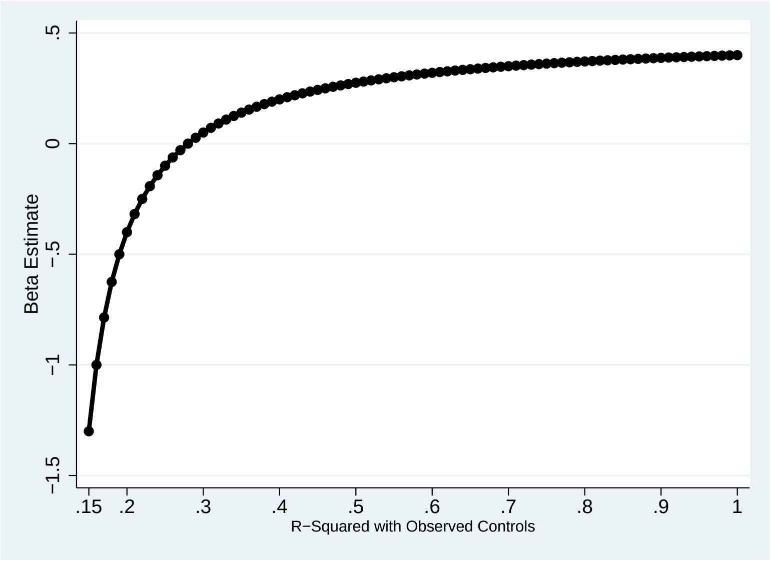
Sesso, H. D., J. E. Buring, W. G. Christen, T. Kurth, C. Belanger, J. MacFadyen, V. Bubes, J. E. Manson, R. J. Glynn, and J. M. Gaziano, "Vitamins E and C in the prevention of cardiovascular disease in men: the Physicians' Health Study II randomized controlled trial," *JAMA*, Nov 2008, *300* (18), 2123–2133.

Shaw, Kelly, Hanni Gennat, Peter ORourke, and Chris Del Mar, "Exercise for overweight or obesity," *Cochrane Database of Systematic Reviews*, 2006, (4).

Stefanick, M. L., S. Mackey, M. Sheehan, N. Ellsworth, W. L. Haskell, and P. D. Wood, "Effects of diet and exercise in men and postmenopausal women with low levels of HDL cholesterol and high levels of LDL cholesterol," *N. Engl. J. Med.*, Jul 1998, *339* (1), 12–20.

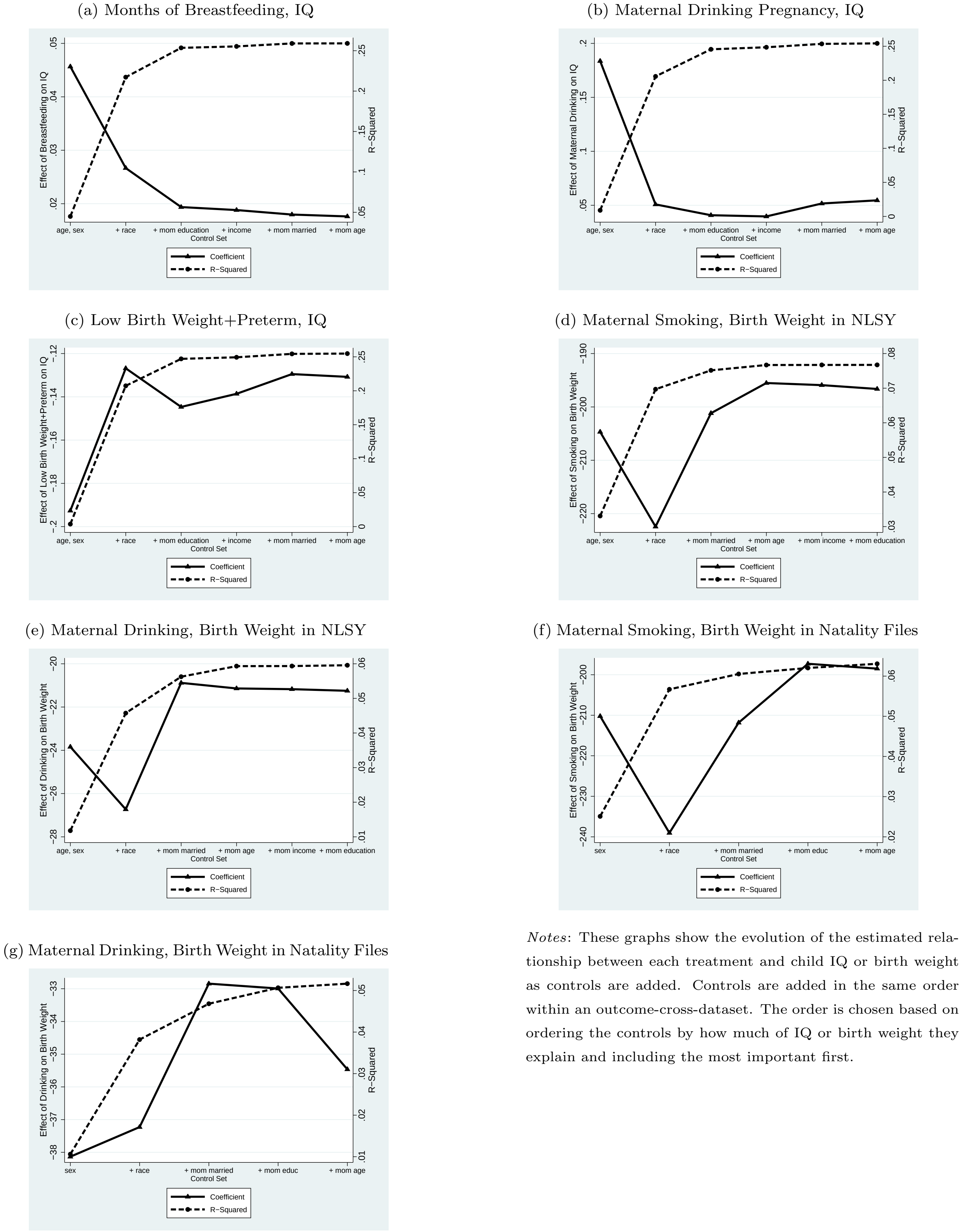
Wood, P. D., M. L. Stefanick, D. M. Dreon et al., "Changes in plasma lipids and lipoproteins in overweight men during weight loss through dieting as compared with exercise," *N. Engl. J. Med.*, Nov 1988, *319* (18), 1173–1179.

Figure 1: **Simulated Effects by Controlled R-Squared**



Notes: This graph shows simulated estimates of treatment effects for a setting with the following assumption: $\mathring{\beta} = .5$, $\tilde{\beta} = .5$, $\mathring{R} = .1$, $R_{max} = 1$ and $\delta = 1$. The X-axis gives values of $\tilde{R}$ and the Y-axis graphs the resulting treatment effects which would be estimated by the proportional selection adjustment.

Figure 2: **Coefficient Stability, Maternal Behavior, Child Birth Weight and IQ**



(a) Months of Breastfeeding, IQ

(b) Maternal Drinking Pregnancy, IQ

(c) Low Birth Weight+Preterm, IQ

(d) Maternal Smoking, Birth Weight in NLSY

(e) Maternal Drinking, Birth Weight in NLSY

(f) Maternal Smoking, Birth Weight in Natality Files

(g) Maternal Drinking, Birth Weight in Natality Files

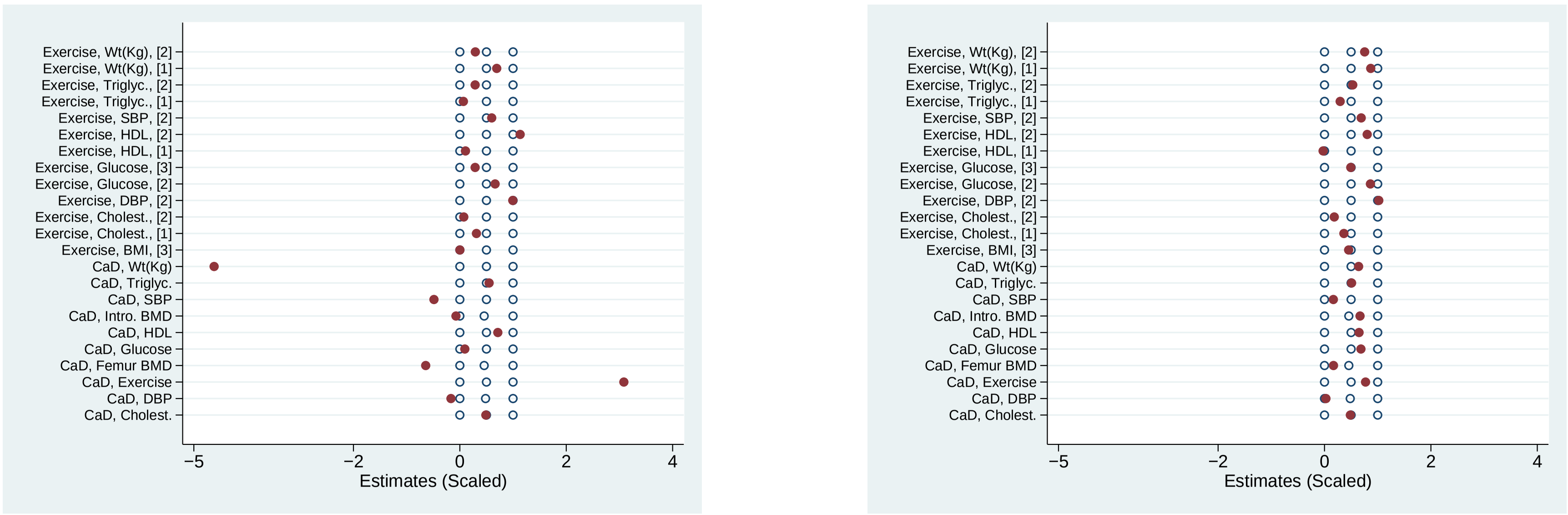
Notes: These graphs show the evolution of the estimated relationship between each treatment and child IQ or birth weight as controls are added. Controls are added in the same order within an outcome-cross-dataset. The order is chosen based on ordering the controls by how much of IQ or birth weight they explain and including the most important first.

Table 1: **Summary Statistics: Early Life and Child IQ**

Panel A: NLSY Data, IQ Analysis			
	Mean	*Standard Deviation*	*Sample Size*
IQ (PIAT Score, Standardized)	0.026	0.991	6613
Breastfeeding Months	2.32	4.51	6184
LBW + Preterm	0.049	0.217	5896
Mom Drink at all in Pregnancy	0.322	0.467	6225
Age	5.57	1.37	6613
Child Female	0.494	0.500	6613
Mother Black	0.284	0.451	6613
Mother Age	25.1	5.42	6613
Mother Education (years)	12.4	3.1	6613
Mother Income	$39,980	$79,069	6613
Mother Married	0.649	0.477	6613
Panel B: NLSY Data, Birth Weight Analysis			
Birth Weight (grams)	3292.8	604.9	7418
Mom Smoke in Pregnancy	0.290	0.453	7418
Drinking Intensity (0-7)	0.634	1.15	7174
Child Female	0.486	0.499	7418
Mother Black	0.277	0.447	7418
Mother Age	24.2	5.42	7418
Mother Education (years)	12.1	3.1	7418
Mother Income	$31,097	$62,975	7418
Mother Married	0.665	0.471	7418
Panel C: Natality Detail Files			
Birth Weight (grams)	3333.8	575.1	5,886,822
Mom Smoke in Pregnancy	0.123	0.328	5,886,822
Drinking Intensity (0-7)	0.023	0.316	5,886,822
Child Female	0.488	0.499	5,886,822
Mother Black	0.167	0.373	5,886,822
Mother Age	27.2	6.13	5,886,822
Mother Education (1-5)	3.51	1.16	5,886,822
Mother Married	0.658	0.474	5,886,822

Notes: This table shows summary statistics for the data used in the analysis in Section 3. Drinking intensity is coded from 0 (never) to 7 (every day). Natality detail files are from 2001 and 2002. NLSY data is from the NLSY Children and Young Adults panel.

Figure 3: **Model Fit With And Without Bias Adjustment**



(a) Controlled Regression

(b) With Bias Adjustment

Notes: These graphs show the randomized effect sizes along with (in Sub-Figure a) the effects estimated in controlled regressions and (in Sub-Figure b) the bias-adjusted coefficients using the best-fit adjustment value of $\delta = 0.971$. Every outcome is scaled so the top and bottom of the 95% confidence interval in the randomized trial take values of 0 and 1 respectively. The mean randomized trial value is typically 0.5, although in some cases it is slightly more or less when the confidence intervals are not symmetric.

Table 2: **Maternal Behavior, Child IQ and Birth Weight**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Child IQ, Standardized (NLSY) (R_{max} = .385)									
Treatment Variable	*Baseline Effect*	*Baseline*	*Effect with*	*Controls*	*Null Reject?*	*δ to match*	*Sibling FE*	*δ to match*	*Bias-Adjusted*
		R^2	*Full Controls*	R^2	*(extrnl. evid.)*	$\beta = 0$	*Estimate*	*Sibling*	*Coefficient,* $\delta = 1$
Breastfeeding (Months)	0.045*** (.003)	.045	0.017***(.002)	.259	No	1.07	-0.008*(.005)	1.59	0.001 (.003)
Drinking in Preg. (Any)	0.183***(.026)	.009	0.054**(.023)	.254	No	0.80	0.024 (.036)	0.43	-0.014 (.025)
LBW + Preterm	-0.192***(.058)	.003	-0.131***(.050)	.254	Yes	3.70	-0.107 (.071)	0.73	-0.098 (.063)
Panel B: Birth Weight in Grams (NLSY) R_{max} = .5									
Treatment Variable	*Baseline Effect*	*Baseline*	*Effect with*	*Controls*	*Null Reject?*	*δ to match*	*Sibling FE*	*δ to match*	*Bias-Adjusted*
		R^2	*Full Controls*	R^2	*(extrnl. evid.)*	$\beta = 0$	*Estimate*	*Sibling*	*Coefficient,* $\delta = 1$
Smoking in Pregnancy	-204.7*** (15.2)	.033	-196.63***(15.6)	.076	Yes	2.52	-74.7***(29.4)	1.56	-118.66*(66.3)
Drinking in Preg. (Intensity)	-23.83***(6.14)	.011	-21.24*** (6.05)	.059	No	0.89	-3.41 (8.74)	0.75	2.62 (17.23)
Panel C: Birth Weight in Grams (Natality Detail Files) R_{max} = .5									
Treatment Variable	*Baseline Effect*	*Baseline*	*Effect with*	*Controls*	*Null Reject?*	*δ to match*	*Sibling FE*	*δ to match*	*Bias-Adjusted*
		R^2	*Full Controls*	R^2	*(extrnl. evid.)*	$\beta = 0$	*Estimate*	*Sibling*	*Coefficient,* $\delta = 1$
Smoking in Pregnancy	-210.2***(.699)	.025	-198.9***(.735)	.064	Yes	1.45	-74.7***(29.4)	1.00	-73.3***(3.70)
Drinking in Preg.(Intensity)	-38.15***(.73)	.010	-34.95***(.72)	.053	No	1.04	-3.41 (8.74)	0.94	-1.45 (1.73)

Notes: This table shows the validation results for the analysis of the impact of maternal behavior on child birth weight and IQ. Baseline effects include only controls for child sex and age dummies in the case of IQ. Full control effects in the NLSY: race, age, education, income, marital status. Full control effects in Natality Detail Files: race, education, marital status and age. Sibling fixed effects estimates come from NLSY in all panels. The value of δ is calculated to match the adjusted β to the sibling fixed effect (Column 6) or to 0 (Column 7). The bias-adjusted effect in Column 8 is generated using the assumption that $\delta = 1$. Standard errors are estimated using a bootstrap over individuals. * significant at 10% level, ** significant at 5% level, *** significant at 1% level.

Table 3: **Summary Statistics: Exercise and Vitamins**

Panel A: Exercise [NHANES-III]			
	Mean	*Standard Deviation*	*Sample Size*
Jogging 3+ Times/Wk	.033	.179	9268
BMI	28.0	6.08	9251
Weight (kg)	78.2	18.4	9252
Diastolic Blood Pressure	76.8	10.3	9197
Systolic Blood Pressure	123.9	17.5	9198
Serum Glucose (mmol/l)	5.61	2.17	8712
Triglycerides (mmol/l)	1.71	1.44	8791
Cholesterol (mmol/l)	5.39	1.13	8811
HDL (mmol/l)	1.31	.41	8740
Panel B: Vitamin D and Calcium Supplements [NHANES-III]			
Took VitD+Calcium	.211	.408	3200
Weight (kg)	69.5	16.3	3180
Diastolic Blood Pressure	73.5	10.1	3003
Systolic Blood Pressure	140.2	20.9	3004
Serum Glucose (mg/dl)	111.9	50.5	2937
Triglycerides (mg/dl)	166.4	111.8	2983
Cholesterol (mg/dl)	232.3	45.6	2988
HDL (mg/dl)	55.7	16.9	2972
Exercise Intensity (METS/wk)	14.3	20.4	3196
Femur BMD	.68	.13	2689
Introchanter BMD	.94	.19	2689

Notes: This table shows summary statistics for the data used in Section 4. NHANES-III : National Longitudinal Health and Nutrition Survey, Wave III. For Exercise, the sample restrictions in the analysis differ slightly depending on which paper I am comparing to. For the summary statistics I consider the most inclusive definition.

Table 4: **Selection Adjustments and Randomized Results**

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Outcome	*Uncontrolled Effect*	*Controlled Effect*	*Null Reject?*	δ *to match*	*Randomized*	δ *to match*	*Bias-Adjusted Effect*
[Citation]	*(Std. Error)*, $[R^2]$	*(Std. Error)*, $[R^2]$	*(extrnl. evid.)*	$\beta = 0$	*Estimate*	*Randomized*	($\delta = .971$)
							(Std. Error)
Panel A: Exercise							
BMI, [3]	-1.93** (0.41) [0.020]	-1.49** (0.41) [0.048]	Yes	1.862	-1.01**	1.056	-1.05**(.34)
Weight in Kg, [1]	-4.56** (1.20) [0.177]	-3.99** (1.21) [0.201]	Yes	2.915	-4.60**	-0.970	-3.43**(1.11)
Weight in Kg, [2]	-2.42** (1.09) [0.210]	-1.53 (1.10) [0.230]	Yes	1.525	-1.15**	0.450	-0.68 (1.11)
Diastolic BP, [2]	-0.09 (0.67) [0.054]	0.004 (0.67) [0.071]	No	-0.039	-1.80	-4.831	0.095 (.74)
Systolic BP, [2]	0.22 (1.00) [0.094]	0.66 (1.01) [0.119]	No	-1.320	0.20	-0.972	1.07 (.98)
Serum Glucose, [2]	-0.21 (0.14) [0.025]	-0.13 (0.14) [0.051]	Yes	1.655	-0.19**	-0.732	-0.059 (.094)
Serum Glucose, [3]	-0.31** (0.12) [0.027]	-0.24** (0.12) [0.048]	Yes	2.178	-0.16	0.998	-0.16** (.068)
Cholesterol, [1]	-0.15* (0.09) [0.033]	-0.12 (0.09) [0.062]	No	3.257	-0.02	2.935	-0.093 (.087)
Cholesterol, [2]	-0.12 (0.08) [0.068]	-0.09 (0.08) [0.095]	No	2.136	0.05	2.796	-0.051 (.089)
Triglycerides, [1]	-0.47** (0.13) [0.034]	-0.36** (0.13) [0.066]	Yes	2.028	-0.16	1.607	-0.25**(.094)
Triglycerides,[2]	-0.37** (0.10) [0.030]	-0.28** (0.10) [0.062]	Yes	1.958	-0.20**	0.857	-0.19** (.083)
HDL, [1]	0.104** (0.03) [0.016]	0.09** (0.03) [0.104]	Yes	4.511	0.13**	-2.491	0.077**(.030)
HDL, [2]	0.11** (0.03) [0.092]	0.08** (0.03) [0.132]	Yes	2.051	0.03	1.638	0.054**(.027)
Panel B: Vitamin D and Calcium							
Weight in Kg	-3.03** (0.77) [0.139]	-1.58** (0.78) [0.184]	Yes	1.077	-0.22**	0.948	-0.18 (.85)
Diastolic BP	-0.26 (0.46) [0.047]	-0.15 (0.48) [0.065]	No	1.478	0.18	3.127	-0.053 (.43)
Systolic BP	-1.12 (0.93) [0.102]	-0.52 (0.96) [0.129]	No	0.864	0.37	1.437	0.066 1.06)
Serum Glucose	-6.92** (2.39) [0.021]	-3.59* (2.44) [0.056]	No	1.066	-1.37	0.681	-0.36 (2.39)
Triglycerides	4.47 (5.39) [0.010]	3.03 (5.45) [0.064]	No	2.095	1.33	1.182	1.63 (6.07)
Cholesterol	0.20 (2.17) [0.012]	0.16 (2.23) [0.030]	No	3.601	0.33	-4.091	0.114 (2.50)
HDL	1.28* (0.80) [0.015]	1.02 (0.82) [0.053]	No	3.612	0.17	3.091	0.76 (.88)
Exercise (METS/wk)	5.27** (0.94) [0.028]	2.88** (0.94) [0.092]	No	1.164	0.30	1.067	0.57 (1.20)
Femur BMD	-0.019** (0.01) [0.175]	-0.006 (0.01) [0.260]	Yes	0.434	0.01**	1.266	0.007 (.007)
Introchanter BMD	-0.020** (0.01) [0.163]	-0.008 (0.01) [0.216]	No	0.662	0.00	0.704	0.004 (.010)

Notes: This table displays the match between the results from observational data and randomized results. Citation Key: [1] Wood et al, 1988; [2] Stefanick et al, 1998; [3] Anderssen et al, 1996. Full citations for randomized data and observational sample restrictions are in Appendix Table A.1. Controls in Panels A and B include : dummies for age and sex (controlled and uncontrolled regressions), dummies for income, dummies for education category, dummies for race, dummies for detailed marital status (controlled regressions only). Throughout the table we assume $R_{max} = \tilde{R} + (\tilde{R} - \mathring{R})$. The bias-adjustment in Column 4 is performed using a value of $\delta = .971$. Standard errors are bootstrapped over individuals. *significant at the 10% level, ** significant at the 5% level.

Table 5: **Selection Adjustments, Out-of-Sample Outcomes**

Panel A: Exercise

Outcome	*Uncontrolled Effect*	*Controlled Effect*	*Randomized Effect*	*Bias-Adjusted Effect*
	(Std. Error)	*(Std. Error)*	*[Possible Direction, Sig.]*	*(Std. Error)*
Ever Diabetes	-0.035**(.009)	-0.019** (.009)	Negative, Not Significant	-0.004 (.010)
Mortality, with heart disease, Men	-0.132**(.041)	-0.115**(.041)	Negative, Significant	-0.010**(.05)
Overall Bone Density, Women	-0.013 (.012)	-0.0003 (.012)	Positive, Not Significant	0.013 (.013)

Panel B: Vitamin D and Calcium Supplementation

Outcome	*Uncontrolled Effect*	*Controlled Effect*	*Randomized Effect*	*Bias-Adjusted Effect*
	(Std. Error)	*(Std. Error)*	*[Possible Direction, Sig.]*	*(Std. Error)*
Ever Diabetes	-0.049**(.015)	-0.023 (.016)	Negative, Not Significant	0.001 (.018)
Mortality	-0.058**(.019)	-0.034*(.020)	Negative, Not Significant	-0.011 (.022)

Panel C: Vitamins and Mortality in Physician Health Study

Outcome	*Uncontrolled Effect*	*Controlled Effect*	*Randomized Effect*	*Bias-Adjusted Effect*
	(Std. Error)	*(Std. Error)*	*[Possible Direction, Sig.]*	*(Std. Error)*
Beta-Carotene Supplements	-0.035*(.019)	-0.022 (.019)	Negative, Not Significant	-0.010 (.020)
Vitamin E Supplements	-0.033***(.012)	-0.026**(.012)	Negative, Not Significant	-0.018 (.013)
Vitamin C Supplements	-0.029** (.011)	-0.021* (.012)	Negative, Not Significant	-0.014 (.013)

Notes: Exercise treatment: total exercise times per month (in units of 100). All adjustments are done using a value of $R_{max} = \tilde{R} + (\tilde{R} - \mathring{R})$ and $\delta = .971$. Citation List: Exercise and (a) diabetes (Orozco et al, 2008); (b) mortality (Heran et al, 2011); (c) bone density (Howe et al, 2011). Vitamin Supplementation and: (a) diabetes (de Boer et al, 2008); (b) mortality (LaCroix et al, 2009); (c) cognitive (Rossom et al, 2012); (d) cancer (Brunner et al, 2011). Physican Health Study: (a) Beta-carotene (Hennekens et al, 1996); vitamins E and C (Sesso et al, 2008).

Appendix A: Details of Proofs

Proof of Lemma 1: Claim: $(\mathring{\beta}-\tilde{\beta}) \xrightarrow{p} \sigma_{1X}\frac{\sigma_{11}^2-\sigma_{1X}^2(\delta\sigma_{22}+\sigma_{11})}{\sigma_{11}(\sigma_{11}-\sigma_{1,X}^2)}$

Proof: Observe that $\hat{\lambda}_{w_1|X}$ converges in probability to $\frac{Cov(W_1,X)}{V(X)}=\sigma_{1,X}$. By a similar logic, $\hat{\lambda}_{W_2|X}$ converges to σ_{2X} and, under proportional selection, to $\frac{\delta\sigma_{1X}\sigma_{22}}{\sigma_{11}}$. $\hat{\lambda}_{W_2|X,W_1}$ converges in probability to $\frac{Cov(W_2,X)}{Var(\tilde{X})}$ where $\tilde{X}$ is the residual from a regression of X on $W_{1.}$ Note that $Var(\tilde{X})$ converges in probability to $1-\frac{\sigma_{1X}^2}{\sigma_{11}}$. Therefore, again invoking proportional selection, $\hat{\lambda}_{W_2|X,W_1}$ converges in probability to $\frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11}-\sigma_{1X}^2}$. Subtracting and simplifying yields the result.

Proof of Lemma 2: Claim: $(\tilde{R}-\mathring{R})\hat{\sigma}_{yy} \xrightarrow{p} \frac{\left[\sigma_{11}^2-\sigma_{1X}^2(\sigma_{11}+\delta\sigma_{22})\right]^2}{\sigma_{11}^2(\sigma_{11}-\sigma_{1X}^2)}$ and $(R_{max}-\tilde{R})\hat{\sigma}_{yy} \xrightarrow{p} \frac{\sigma_{22}\left[\sigma_{11}^2-\sigma_{1X}^2(\sigma_{11}+\delta^2\sigma_{22})\right]}{\sigma_{11}(\sigma_{11}-\sigma_{1X}^2)}$.

Proof: Observe the following definitions. From the short regression coefficient, $\mathring{R}=\frac{(\beta+\hat{\lambda}_{w_1|X}+\hat{\lambda}_{W_2|X})^2}{\hat{\sigma}_{yy}}$. By Lemma 1, this converges in probability to $\frac{(\beta+\frac{\sigma_{1,X}(\delta\sigma_{22}+\sigma_{11})}{\sigma_{11}})^2}{\sigma_{yy}}$. In the intermediate regression the calculation relies on the coefficient on X $(\beta+\hat{\lambda}_{W_2|X,W_1})$ and the coefficient on W_1, which is also biased by the exclusion of W_2 through the joint correlation with X and is equal to1 $-\frac{\sigma_{1X}}{\sigma_{11}}\hat{\lambda}_{W_2|X,W_1}$. Thus, $\tilde{R}=\frac{(\beta+\hat{\lambda}_{W_2|X,W_1})^2+\sigma_{11}(1-\frac{\sigma_{1X}}{\sigma_{11}}\hat{\lambda}_{W_2|X,W_1})^2+2\sigma_{1X}(\beta+\hat{\lambda}_{W_2|X,W_1})(1-\frac{\sigma_{1X}}{\sigma_{11}}\hat{\lambda}_{W_2|X,W_1})}{\hat{\sigma}_{yy}}$. By Lemma 1, $\tilde{R}\xrightarrow{p}\frac{(\beta+\frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11}-\sigma_{1X}^2})^2+(1-\frac{\sigma_{1X}}{\sigma_{11}}\frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11}-\sigma_{1X}^2})^2\sigma_{11}+2(\beta+\frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11}-\sigma_{1X}^2})(1-\frac{\sigma_{1X}}{\sigma_{11}}\frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11}-\sigma_{1X}^2})\sigma_{1X}}{\sigma_{yy}}$. Finally, observe that $R_{max}=\frac{\beta^2+\sigma_{11}+\sigma_{22}+2\beta\sigma_{1,X}+2\beta\frac{\delta\sigma_{1X}\sigma_{22}}{\sigma_{11}}}{\sigma_{yy}}$. Differencing these expressions appropriately yields the result.

Proof of Proposition 1. Claim : Define:

$$\beta^*=\begin{cases}\tilde{\beta}-\delta\left[\mathring{\beta}-\tilde{\beta}\right]\frac{R_{max}-\tilde{R}}{\tilde{R}-\mathring{R}} & \text{if } \delta{=}1\\ \tilde{\beta}-\left[\frac{\sqrt{[\mathring{\beta}-\tilde{\beta}]^2[\Theta^2+\Theta\left(4\delta(1-\delta)[\mathring{\beta}-\tilde{\beta}]^2[R_{max}-\tilde{R}]\right)}]-\Theta[\mathring{\beta}-\tilde{\beta}]}{2(1-\delta)\left[\mathring{\beta}-\tilde{\beta}\right]^2\left[\tilde{R}-\mathring{R}\right]}\right] & \text{if } \delta\neq 1,\sigma_{1X}\geq 0\\ \tilde{\beta}-\left[\frac{-\sqrt{[\mathring{\beta}-\tilde{\beta}]^2[\Theta^2+\Theta\left(4\delta(1-\delta)[\mathring{\beta}-\tilde{\beta}]^2[R_{max}-\tilde{R}]\right)}]-\Theta[\mathring{\beta}-\tilde{\beta}]}{2(1-\delta)\left[\mathring{\beta}-\tilde{\beta}\right]^2\left[\tilde{R}-\mathring{R}\right]}\right] & \text{if } \delta\neq 1,\sigma_{1X}<0\end{cases}$$

where $\Theta=\left(\left[\tilde{R}-\mathring{R}\right]{}^2\hat{\sigma}_{yy}+\left[\mathring{\beta}-\tilde{\beta}\right]{}^2\left[\tilde{R}-\mathring{R}\right]\right)$. $\beta^*\xrightarrow{p}\beta$.

Proof: Recall that the object of interest – the bias – is $\frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11}-\sigma_{1X}^2}$. There are three unknowns here: σ_{11},σ_{22} and σ_{1X}. Note that none of these can be calculated directly from the data. Lemmas 1 and 2 provide a system of three equations in these variables. Lemmas are stated in probability limits; for the proof I will write these as equalities to simplify notation, and return to the probability limit notation at the end. In addition, again to simplify notation in the algebra, I will adopt single letter notation for each of the differences.

$$\begin{aligned}A=\mathring{\beta}-\tilde{\beta} &= \sigma_{1X}\frac{\sigma_{11}^2-\sigma_{1X}^2(\delta\sigma_{22}+\sigma_{11})}{\sigma_{11}(\sigma_{11}-\sigma_{1X}^2)}\\ B=\left[\tilde{R}-\mathring{R}\right]\sigma_{yy} &= \frac{\left[\sigma_{11}^2-\sigma_{1X}^2(\sigma_{11}+\delta\sigma_{22})\right]^2}{\sigma_{11}^2(\sigma_{11}-\sigma_{1X}^2)}\\ C=\left[R_{max}-\tilde{R}\right]\sigma_{yy} &= \frac{\sigma_{22}\left[\sigma_{11}^2-\sigma_{1X}^2(\sigma_{11}+\delta^2\sigma_{22})\right]}{\sigma_{11}(\sigma_{11}-\sigma_{1X}^2)}\end{aligned}$$

The algebra differs slightly for the case of $\delta=1$ and the case of $\delta\neq 1$ but only in a later step of the proof. I will note when the cases diverge below. The method of proof is simply to solve the system of simultaneous equations. Some algebraic steps are suppressed.

Solve Equation (1) for σ_{22}:

$$\begin{aligned} A &= \sigma_{1X}\frac{\sigma_{11}^2 - \sigma_{1X}^2(\delta\sigma_{22} + \sigma_{11})}{\sigma_{11}(\sigma_{11} - \sigma_{1X}^2)} \\ \sigma_{22} &= \frac{1}{\delta}\left[\frac{\sigma_{11}^2\sigma_{1X} - A\sigma_{11}(\sigma_{11} - \sigma_{1,X}^2) - \sigma_{11}\sigma_{1X}^3}{\sigma_{1X}^3}\right] \end{aligned}$$

Solve Equation (2) for σ_{11} and σ_{22} in terms of σ_{1X} :

$$\begin{aligned} B &= \frac{\left[\sigma_{11}^2 - \sigma_{1X}^2(\sigma_{11} + \delta\sigma_{22})\right]^2}{\sigma_{11}^2(\sigma_{11} - \sigma_{1X}^2)\sigma_y} = \frac{\left[\sigma_{11}^2 - \sigma_{1X}^2(\sigma_{11} + \frac{\sigma_{11}^2\sigma_{1X} - A\sigma_{11}(\sigma_{11} - \sigma_{1,X}^2) - \sigma_{11}\sigma_{1X}^3}{\sigma_{1X}^3})\right]^2}{\sigma_{11}^2(\sigma_{11} - \sigma_{1X}^2)} \\ \sigma_{11} &= \left[\frac{\sigma_{1X}^2(A^2 + B)}{A^2}\right] \\ \sigma_{22} &= \frac{1}{\delta}\left[\frac{\sigma_{1X}(B^2 + A^2B)\left[\sigma_{1X} - A\right]}{A^4}\right] \end{aligned}$$

Note that for the bias calculation we do not require σ_{11} alone but only $\sigma_{11} - \sigma_{1X}^2$ which, given values above, equals $\frac{\sigma_{1X}^2 B}{A^2}$ and allows us to collapse the bias calculation to $\frac{\delta\sigma_{22}A^2}{\sigma_{1X}B}$.

Case 1: $\delta = 1$**.** *Solve Equation (3) for σ_{1X} :*

$$C = \frac{\sigma_{22}\left[\sigma_{11}^2 - \sigma_{1X}^2(\sigma_{11} + \sigma_{22})\right]}{\sigma_{11}(\sigma_{11} - \sigma_{1X}^2)} = \frac{\left[\frac{\sigma_{1X}(B^2+A^2B)[\sigma_{1X}-A]}{A^4}\right]\left[\frac{\sigma_{1X}^2(A^2+B)}{A^2} - \sigma_{1X}^2\left(\frac{\sigma_{1X}^2(A^2+B)}{A^2} + \left[\frac{\sigma_{1X}(B^2+A^2B)[\sigma_{1X}-A]}{A^4}\right]\right)\right]}{\frac{\sigma_{1X}^2(A^2+B)}{A^2}\left[\frac{\sigma_{1X}^2(A^2+B)}{A^2} - \sigma_{1X}^2\right]}$$

$$\begin{aligned} \sigma_{1X} &= \frac{CA^3 + A(B^2 + A^2B)}{(B^2 + A^2B)} \\ \sigma_{22} &= \left[CA^3 + A(B^2 + A^2B)\right]\left[\frac{C}{A(B^2 + A^2B)}\right] \end{aligned}$$

Applying these values to the formula above, we have:

$$\frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11} - \sigma_{1X}^2} = \frac{AC}{B} = \delta\left[\hat{\beta} - \tilde{\beta}\right]\frac{R_{max} - \tilde{R}}{\tilde{R} - \hat{R}}$$

which leads us to the $\delta = 1$ result.

Case 2: $\delta \neq 1$. *Solve Equation (3) for σ_{1X} :*

$$\begin{aligned} C &= \frac{\sigma_{22}\left[\sigma_{11}^2 - \sigma_{1X}^2(\sigma_{11} + \delta^2\sigma_{22})\right]}{\sigma_{11}(\sigma_{11} - \sigma_{1X}^2)} \\ C &= \frac{1}{\delta}\frac{\frac{\sigma_{1X}(B^2+A^2B)[\sigma_{1X}-A]}{A^4}\left[\left[\frac{\sigma_{1X}^2(A^2+B)}{A^2}\right]^2 - \sigma_{1X}^2(\frac{\sigma_{1X}^2(A^2+B)}{A^2} + \delta\frac{\sigma_{1X}(B^2+A^2B)[\sigma_{1X}-A]}{A^4})\right]}{\frac{\sigma_{1X}^2(A^2+B)}{A^2}(\frac{\sigma_{1X}^2(A^2+B)}{A^2} - \sigma_{1X}^2)} \end{aligned}$$

This does not simplify to the extent that the $\delta = 1$ case does, and solving for σ_{1X} requires the quadratic formula. Applying this, we find:

$$\sigma_{1X} = \frac{(A(B^2 + A^2B)(1-2\delta)) \pm \sqrt{(A^2(B^2 + A^2B)^2 + 4(B^2 + A^2B)(1-\delta)\delta CA^4)}}{2(B^2 + A^2B)(1-\delta)}$$

Note this has two roots. The positive root corresponds to the case where $\sigma_{1X} \geq 0$; the negative root to the case where $\sigma_{1X} < 0$.

Given this and the resulting formula for σ_{22} we can complete the solution. If $\sigma_{1X} \geq 0$, we have:

$$\frac{\delta\sigma_{22}A^2}{\sigma_{1X}B} = \left[\frac{(-A(B^2 + A^2B) + \sqrt{(A^2(B^2 + A^2B)\left[B^2 + A^2B + 4\delta(1-\delta)CA^2\right]}}{2(1-\delta)BA^2}\right]$$

If $\sigma_{1X} < 0$ we have:

$$\frac{\delta\sigma_{22}A^2}{\sigma_{1X}B} = \left[\frac{(-A(B^2 + A^2B) - \sqrt{(A^2(B^2 + A^2B)\left[B^2 + A^2B + 4\delta(1-\delta)CA^2\right]}}{2(1-\delta)BA^2}\right]$$

Substituting in the difference values for A, B and C yields the result.

Delta Value for $\beta = 0$ In implementation I argue that a valuable statistic to report is the value of δ such that $\beta = 0$. The exact formula is:

$$\hat{\delta} = \frac{\tilde{\beta}^2\left[(\tilde{R} - \mathring{R})\right]^2\left[\mathring{\beta} - \tilde{\beta}\right] + \tilde{\beta}\left[(\tilde{R} - \mathring{R})\right]\left[\hat{\sigma}_{yy}\left[(\tilde{R} - \mathring{R})\right]^2 + \left[\mathring{\beta} - \tilde{\beta}\right]^2\left[(\tilde{R} - \mathring{R})\right]\right]}{\tilde{\beta}^2\left[(\tilde{R} - \mathring{R})\right]^2\left[\mathring{\beta} - \tilde{\beta}\right] + \left[R_{max} - \tilde{R}\right]\left[\mathring{\beta} - \tilde{\beta}\right]\left[\hat{\sigma}_{yy}\left[(\tilde{R} - \mathring{R})\right]^2 + \left[\mathring{\beta} - \tilde{\beta}\right]^2\left[(\tilde{R} - \mathring{R})\right]\right]}$$

Note that this is invariant to the sign of σ_{1X}.

Appendix B: Further Theoretical Results

This appendix discusses two additional issues related to the theory. Subsection A.1 below briefly contrasts the calculation of bias based on the coefficients to the calculation directly from the data suggested by Altonji, Elder and Taber (2005). Subsection A.2 discusses details of the case with **m**.

A.1. Altonji, Elder and Taber (2005) Calculation

Recall the model:

$$Y = \alpha + \beta X + W_1 + W_2 + \epsilon$$

For simplicity, assume that $\epsilon = 0$; and $R_{max} = 1$. Lemma 1 in the text demonstrates that, under the proportional selection relationship the bias on the intermediate regression coefficient $\tilde{\beta}$ is $\frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11} - \sigma_{1X}^2}$.

Altonji, Elder and Taber (2005) suggest that this bias might be calculated directly from the data. In particular, they propose:

1. Run the intermediate regression, which we will denote $Y = \tilde{\beta}X + \Psi W_1 + \tilde{\epsilon}$.

2. Calculate ΨW_1 and denote the variance of the residual $V_{\tilde{\epsilon}}$.\psi=1.01

3. Regress X on $\hat{\Psi}W_1$. Denote the coefficient on $\hat{\Psi}W_1$ as Γ, and the variance of the residual $V_{\tilde{X}}$.

4. Calculate the bias as $\frac{\delta\Gamma V_{\tilde{\epsilon}}}{V_{\tilde{X}}}$

Recall that $Var(X) = 1$. Consider each of the elements of this in turn:

1. $V_{\tilde{X}}$

$$V_{\tilde{X}} \xrightarrow{p} 1 - \frac{\sigma_{1X}^2}{\sigma_{11}}$$

2. Γ. Note first that $\hat{\Psi} \rightarrow^p 1 - \frac{\sigma_{1X}}{\sigma_{11}} \frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11}(1-\sigma_{1X}^2)}$.

$$\begin{aligned} \Gamma &= \frac{Cov(\hat{\Psi}W_1, X)}{Var(\hat{\Psi}W_1)} = \frac{Cov(W_1, X)}{\hat{\Psi}Var(W_1)} \\ \Gamma &\xrightarrow{p} \frac{\sigma_{1X}}{\left[1 - \frac{\sigma_{1X}}{\sigma_{11}} \frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11}-\sigma_{1X}^2)}\right]\sigma_{11}} = \frac{\sigma_{1X}(\sigma_{11} - \sigma_{1X}^2)}{\sigma_{11}(\sigma_{11} - \sigma_{1X}^2) - \delta\sigma_{22}\sigma_{1X}^2} \end{aligned}$$

3. $V_{\tilde{\epsilon}}$.

$$V_{\tilde{\epsilon}} \xrightarrow{p} \sigma_{22} - \frac{(\delta\sigma_{22}\sigma_{1X})^2}{\sigma_{11}(\sigma_{11} - \sigma_{1X}^2)}$$

Combining these, we find:

$$\frac{\delta\Gamma V_{\tilde{\epsilon}}}{V_{\tilde{X}}} \xrightarrow{p} \frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11} - \sigma_{1X}^2}\left[\frac{\sigma_{11}(\sigma_{11} - \sigma_{1X}^2) - \delta^2\sigma_{22}\sigma_{1X}^2}{\sigma_{11}(\sigma_{11} - \sigma_{1X}^2) - \delta\sigma_{22}\sigma_{1X}^2}\right]$$

If $\delta = 1$ the second term cancels, but in cases where $\delta \neq 1$ it does not and this calculation is a close approximation to the bias.

A.2. Additional Category Controls

Section 3 discusses extending the model to a case where there is an additional, orthogonal, category of controls, so the true model is

$$Y = \alpha + \beta X + W_1 + W_2 + \mathbf{m} + \epsilon$$

I suggest in Section 3 that the appropriate procedure for recovering β if $\mathbf{m}$ is observed is to include $\mathbf{m}$ in both the short and intermediate regressions and preform the same procedure. It is trivial to see why this works. I have assumed that $\mathbf{m}$ is orthogonal to W_1 and $W_{2.}$ The only correlations are between $\mathbf{m}$ and X and Y. Consider regressing Y on $\mathbf{m}$ and taking residuals and doing the same for X. We can then run our original procedure on the residuals of X and Y to recover β. Including the $\mathbf{m}$ control in both regressions is equivalent to this exercise.

In the case were $\mathbf{m}$ is not observed, I suggest that it is still possible to use this procedure to recover $\overline{\beta}$ from this regression:

$$Y = \overline{\alpha} + \overline{\beta}X + \Psi W_1 + \Psi W_2 + \overline{\epsilon}$$

Although this will not be the causal effect, since $\mathbf{m}$ is omitted, it will be closer to the causal effect since it adjusts for the influence of $W_{2.}$The procedure for recovering $\overline{\beta}$ differs from the main text only in that $\Psi \neq 1$. [17]

To prove this, we therefore work through a modified version of the proof in Section 2.

Short and intermediate regression coefficients are given below.

$$\begin{aligned} \mathring{\beta} &= \overline{\beta} + \hat{\Psi}\hat{\lambda}_{w_1|X} + \hat{\Psi}\hat{\lambda}_{W_2|X} \\ \tilde{\beta} &= \overline{\beta} + \hat{\Psi}\hat{\lambda}_{W_2|X,W_1} \end{aligned}$$

By the same logic as Lemma 1 in the text, and the observation that $\hat{\Psi} \rightarrow^p \Psi$ we observe that

$$\mathring{\beta} - \tilde{\beta} \xrightarrow{p} \Psi\sigma_{1X}\frac{\sigma_{11}^2 - \sigma_{1X}^2(\delta\sigma_{22} + \sigma_{11})}{\sigma_{11}(\sigma_{11} - \sigma_{1X}^2)}$$

Turning to the r-squared values, we observe $\mathring{R} \xrightarrow{p} \frac{(\overline{\beta} + \frac{\Psi\sigma_{1X}(\delta\sigma_{22}+\sigma_{11})}{\sigma_{11}})^2}{\sigma_{yy}}$,

[17] As mentioned in the text this is because they are biased by the exclusion of $\mathbf{m}$ through the joint correlation with X.

$\tilde{R} \xrightarrow{p} \frac{(\overline{\beta}+\frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11}-\sigma_{1X}^2})^2+(\Psi-\frac{\sigma_{1X}}{\sigma_{11}}\frac{\delta\sigma_{22}\sigma_{1,X}}{\sigma_{11}-\sigma_{1X}^2})^2\sigma_{11}+2(\overline{\beta}+\frac{\delta\sigma_{22}\sigma_{1,X}}{\sigma_{11}-\sigma_{1X}^2})(\Psi-\frac{\sigma_{1X}}{\sigma_{11}}\frac{\delta\sigma_{22}\sigma_{1X}}{\sigma_{11}-\sigma_{1X}^2})\sigma_{1X}}{\sigma_{yy}}$ and $R_{max} = \frac{\overline{\beta}^2+\Psi^2\sigma_{11}+\Psi^2\sigma_{22}+2\overline{\beta}\Psi\sigma_{1X}+2\overline{\beta}\Psi\frac{\delta\sigma_{1X}\sigma_{22}}{\sigma_{11}}}{\sigma_{yy}}$. Algebraic simplification then yields:

$$\left[\tilde{R} - \mathring{R}\right]\sigma_{yy} \xrightarrow{p} \Psi^2 \frac{\left[\sigma_{11}^2 - \sigma_{1X}^2(\sigma_{11} + \delta\sigma_{22})\right]^2}{\sigma_{11}^2(\sigma_{11} - \sigma_{1X}^2)}$$

$$\left[R_{max} - \tilde{R}\right]\sigma_{yy} \xrightarrow{p} \Psi \frac{\sigma_{22}\left[\sigma_{11}^2 - \sigma_{1X}^2(\sigma_{11} + \delta^2\sigma_{22})\right]}{\sigma_{11}(\sigma_{11} - \sigma_{1X}^2)}$$

Combining, we replicate the results from Section 2.

Appendix Tables

Table A1: Citation for Randomized Outcomes

Outcome	*Citation*	*Sample Restrictions (if any)*
Exercise, BMI, [3]	Anderssen et al, 1996	Age 30-50
Exercise, Wt(Kg), [1]	Wood et al, 1988	Female, 30-59
Exercise, Wt(Kg), [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, DBP, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, SBP, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, Glucose, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, Glucose, [3]	Anderssen et al, 1996	Age 30-50
Exercise, Triglyc, [1]	Wood et al, 1988	Female, 30-59
Exercise, Triglyc, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, Cholest, [1]	Wood et al, 1988	Female, 30-59
Exercise, Cholest, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, HDL, [1]	Wood et al, 1988	Female, 30-59
Exercise, HDL, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
CaD, Wt(Kg)	Caan et al, 2007	Women, 55-85
CaD, DBP	Margolis et al, 2008	Women, 55-85
CaD, SBP	Margolis et al, 2008	Women, 55-85
CaD, Glucose	de Boer et al, 2008	Women, 55-85
CaD, Triglyc	Rajpathak et al, 2010	Women, 55-85
CaD, Cholest	Rajpathak et al, 2010	Women, 55-85
CaD, HDL	Rajpathak et al, 2010	Women, 55-85
CaD, Exercise	Brunner et al, 2008	Women, 55-85
CaD, Femur BMD	Jackson et al, 2011	Women, 55-85
CaD, Intro. BMD	Jackson et al, 2011	Women, 55-85

Notes: This table shows the source of the randomized estimates. The text of the outcome matches the form of citation in Figure 2.