Unobservable Selection and Coefficient Stability: Theory and Validation in Public Health *

Emily Oster

University of Chicago Booth School of Business and NBER

February 12, 2013

Abstract

Inferring causal treatment effects in the presence of possible omitted variable bias is as well-known problem. Altonji, Elder and Taber (2005) suggest that the degree of selection on observable variables might be used as a guide to the remaining bias in controlled regressions. I expand on their setup and demonstrate how, with an equal selection assumption, a causal effect can be recovered using coefficients, R-squared values from controlled and uncontrolled regressions and an estimate of the *iid* noise in the outcome. I discuss the relationship between this technique and the heuristic procedure of adding sequential controls until coefficients stabilize. I consider two validation exercises which explore whether coefficients adjusted in this way are closer to the truth. First, I estimate the impact of early life and prenatal influences on child IQ. Simple controlled regressions give misleading estimates; the bias-adjusted estimates are significantly better. However, I show that the progressive adding of controls in this setting is not effective. Second, I match observational and randomized trial data for 31 treatment-outcome pairs in three public health settings. I estimate a best-fit degree of adjustment across all settings (the free parameter is the amount of *iid* noise in the outcome). I show that the bias-adjusted coefficients preform much better than simple controlled coefficients; the total error reduction is 30% and a number of false-positive results are rejected without significant loss in true-positive results. The magnitude of the best-fit adjustment suggests adjusting the controlled coefficient by approximately the same amount as the movement between uncontrolled and controlled coefficients.

*Ling Zhong and Unika Shrestha provided excellent research assistance. I thank David Cesarini, Todd Elder, Matt Gentzkow, Chad Syverson, Azeem Shaikh, Jesse Shapiro, Matt Taddy and participants in a seminar at University of Chicago Booth School for helpful comments. I gratefully acknowledge financial support from the Neubauer Family.

1 Introduction

Inferring causality from observational data with possibly omitted variables is a well known challenge. It has many proposed solutions. These include, but are not limited to: use of instrumental variables, selection models, difference-in-difference analyses and collecting randomized data or identifying naturally randomized settings. These solutions are all effective (assuming the assumptions underlying them are not violated) although all make strong requirements of existing data (a valid instrument, a valid natural experiment) or require collecting new data.

In the absence of meeting any of these requirements, some bias can be addressed by simply controlling for the confounds which the researcher *is* able to see. The assumption under which the resulting treatment effect is causal is very strong – namely, that there are no confounds which remain unobserved – and difficult to test. However, reporting such treatment effects as causal occurs in economics and is especially prevalent in public health. In the latter case, such analyses may have real consequences for policy. For example: 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).

A natural question is whether it is possible to do better in these settings *without* new data or new instruments. In an influential paper, Altonji, Elder and Taber (2005) (hence, AET) suggest that the degree of selection on observable variables might be used to guide assumptions about selection on unobservables. In the context of a linear model they suggest an informal procedure which would calculate bias under this type of assumption; this follows the discussion in Murphy and Topel (1990). This proportional-selection theory also underlies a commonly used heuristic of looking at how the treatment effect moves when controls are added and drawing conclusions about possible movements with unobserved controls. Adjusting coefficients using a version of this assumption provides an alternative benchmark for causal inference in these settings. Whether this benchmark, or a modified version of it, is a good one is ultimately an empirical question. Would we draw more accurate conclusions with a bias adjustment of this type?

It is this question which I take up in this paper. I begin by expanding on the discussion in AET and connecting their bias adjustment directly to coefficient movements. I provide some explicit guidance for preforming a bias adjustment based on this theory and discuss conditions under which the "coefficient movement" heuristic is informative. I then turn to validation with two applications. In the first, I explore a single setting – the impact of prenatal and early life behaviors on child IQ – and demonstrate how this bias adjustment might be carefully applied. In the second, I consider a large range of treatment-outcome pairs with both observational and randomized evidence, and ask whether this bias adjustment might be generally applied to reach better conclusions.

I begin in Section 2 with theory. I consider a simple setup: an outcome Y is fully determined by a treatment variable X, a vector of observable controls W, an (orthogonal) vector of unobserved controls C and some error term which is uncorrelated with X, W and C, denoted γ. I adopt the assumption of equal selection described in AET: $\frac{Cov(W,X)}{Var(W)} = \frac{Cov(C,X)}{Var(C)}$. I demonstrate that the causal effect of X on Y in this setting can be calculated directly from (1) the coefficients on X with and without controls for W; (2) the R-squared values from controlled and uncontrolled regressions and (3) an assumption on the maximum R-squared (i.e. an assumption on the importance of γ). More specifically, denote the true effect β, the uncontrolled coefficient ξ, the coefficient with controls Λ and the two r-squared values as R_1 and R_2. Finally, the maximum R-squared is R_{max}. Under the assumption of equal selection: $\beta = \Lambda - \frac{(\xi-\Lambda)(R_{max}-R_2)}{(R_2-R_1)}$. The intuition is straightforward: if the coefficient moves a lot when the controls are added *and* there is a lot of remaining variation in Y which could be explained by related variables, the bias on Λ is large.[1]

A straightforward corollary of this is that any variables which are correlated with X and Y but do not have any unobserved components should be included in both the "uncontrolled" and controlled regressions and, from there, the bias adjustment is identical. This is simply to say that if the issue is omitted measures of socioeconomic status, but age is also an important control, age should be included in "uncontrolled" regressions and not used to guide the bias adjustment. I also consider the case where the observables have two components which can be included in turn – specifically, I ask whether observing that treatment effect converge as better controls are added should lead one to conclude further controls would not alter the coefficient. Although this is a very common heuristic, I show it only implies that $\Lambda = \beta$ if the R-squared is simultaneously converging to the maximum R-squared.

In Section 3 I turn to the first of two validation exercises. I consider four prenatal and early life behaviors: maternal weight before pregnancy, maternal drinking in pregnancy, low birth weight and months of breastfeeding. All of these have been linked in some observational studies to child IQ, but not all of these links appear robust. All are subject to a concern about unobserved family background characteristics driving the associations.

Using the National Longitudinal Survey of Youth, I first document the results from naive regressions of child IQ on behaviors, child age and sex, and a standard set of socioeconomic status controls (maternal age, education, marital status, etc). I show these regressions lead to misleading conclusions in three of four cases. Maternal weight before pregnancy appears to lower child IQ, a fact which is not supported in the best controlled studies. Maternal drinking in pregnancy appears to *increase* child IQ in these simple regressions. Although there is mixed evidence on whether maternal drinking lowers IQ, virtually no one would suggest it improves it. Both low birth weight and limited breastfeeding seem to decrease IQ; although the link with low

[1]The bias adjustment done this way is identical to what is calculated in the procedure outlined in AET in the case where R_{max} is 1. In particular, AET suggests that selection may not be equal, so $\delta\frac{Cov(W,X)}{Var(W)} = \frac{Cov(C,X)}{Var(C)}$, where $\delta \neq 1$, but their procedure assumes that $\gamma = 0$. I will discuss the comparison in practice when I get to the validation in Section 4.

birth weight is robust, and has good biological underpinnings, randomized evidence on breastfeeding and overall IQ does not support that link.

I then discuss the application of the selection-on-observables bias adjustment to these results and, in particular, ask whether it would pick out the low birth weight link as robust. The key missing piece in doing this is an assumption on the maximum R-squared. I choose this setting precisely because we have, from studies of siblings, some sense of the maximum amount of variation in IQ which could be explained by family background. I use a figure of 0.385, drawn from the sibling correlations reported in Scarr and Weinberg (1983). I show the adjustment preforms well: it rejects the associations between breastfeeding, maternal weight and maternal drinking and child IQ, while confirming the link between child IQ and low birth weight.

It is clear that this is linked with coefficient movements: the low birth weight coefficient moves the least when the controls are introduced, and the impact of mother's drinking gets less positive. However, in this application I also show how the method of continually adding more precise controls can be misleading. In all four cases adding controls in sequence shows convergence of coefficients – in fact, it would be difficult to distinguish the four cases from each other visually. The issue is that in none of the cases does the R-squared converge to what we think is the maximum, which is what would be necessary to suggest these coefficients are converging to the truth.

In Section 4 I consider a broader form of validation. I collect data on several health settings in which (a) I observe both observational and randomized results for the same analysis, (b) the observational data is subject to possible omitted variable bias and (c) the primary omitted variable concern is (broadly) socioeconomic status. The three settings considered are: exercise and adult health, vitamin d+calcium supplementation and women's health and breastfeeding and child health. In each setting I use observational data to run naive regressions which would be standard in the public health literature (i.e. health outcome on treatment with simple socioeconomic status controls). I match the point estimate from the observational data to the corresponding estimate in randomized data.[2]

I then approach this as an estimation with R_{max} as the free parameter. Because outcomes vary wildly in the plausible predictability (due either to measurement error or to correlation with other variables which are unrelated to the treatment), I parametrize R_{max} as a function of R_1 and R_2. I assume that $R_{max} = R_2 + \psi(R_2 - R_1)$ and estimate ψ. This can be interpreted as using the information on how much of the variation in Y is explained by the observables to guess how much would be explained by the unobservables; a value of $\psi = 0$ would suggest the simple controlled coefficient matches the results.

I estimate the value of ψ which would lead the observational point estimate to match the randomized result. I then combine across settings and ask what adjustment value minimizes the distance between the bias-adjusted observational coefficients and the randomized results. Given this, I can then ask whether the

[2] 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.

adjusted coefficients with this common adjustment factor are a better match to the randomized data than the controlled coefficients. It is important to note that this *is* a falsifiable test. Although for any given treatment-outcome pair I will be able to find an adjustment value which matches the two coefficients (as long as the controls cause the coefficient to move in the right direction), it is much less clear that a single adjustment value will be broadly applicable.

I estimate that a value of $\psi = 1.018$ provides the best fit to the data. This value suggests that the unobservables explain about the same amount of variation in Y as the observables. The adjustment with this value of ψ generates coefficients which match the randomized effects significantly better than the controlled coefficients. The reduction in error is 30%. The largest adjustments come in places where the relationship in the observational data significantly overstates the effect in randomized trials. The bias-adjustment is effective at rejecting a number of false-positive results, such as the impact of vitamin D supplementation on exercise and serum glucose levels. At the same time it retains a number of true associations, such as that between exercise and weight.

In its role as validation, this evidence suggests that a version of this adjustment is effective at better matching randomized results. A single value of ψ works well over a large number of settings. In addition, this value of ψ could be used in comparable settings where researchers are considering the relationship between a health outcome and some health behavior and the primary concern is omitted socioeconomic status. This covers a wide variety of studies in public health and epidemiology. It would be simple for researchers to report their coefficient estimates (and, with bootstrapping, their standard errors) under this adjustment value. This could also be helpful in evaluating the plausibility of published results.

2 Theory

2.1 Baseline Result: Bias Calculation Under Equal Selection

Consider a linear model relating an outcome Y to treatment X.

$$Y = \alpha + \beta X + W + Z + \gamma \tag{1}$$

W and Z are indies of control variables which are related to both X and Y. W and Z are orthogonal to each other; the researcher observes W but not Z. The final term, γ, is an *iid* noise term. Without loss of generality I assume the variance of X and W are equal to 1, and the variance of Z is V_z. The key assumption is of equal selection: the relationship between W and X is informative about the relationship between Z and X. Formally, denote the covariance between W and X as C_{wx} and between Z and X as C_{zx}. Equal selection

assumes the following equality holds.

$$C_{wx} = \frac{C_{zx}}{V_z}$$

Were both W and Z observed, it would be possible to recover β from a standard linear regression model. With Z unobserved, the researcher is able to estimate two equations:

$$\begin{aligned} Y &= \hat{\alpha} + \xi X + \wr & (2) \\ Y &= \tilde{\alpha} + \Lambda X + \Psi W + \tau & (3) \end{aligned}$$

ξ is the coefficient on X with no controls, and Λ is the coefficient on X when including all the observed controls. Both ξ and Λ are subject to omitted variable bias. Since the models are linear, the relationship between these coefficients and the true β is straightforward:

$$\begin{aligned} \xi &= \beta + C_{wx} + C_{zx} \\ \Lambda &= \beta + \frac{C_{z\tilde{x}}}{V_{\tilde{x}}} \end{aligned}$$

where $\tilde{X}$ is the residual from a bi-variate regression of X on the observed controls W. The central question I address here is to what extent the difference between these coefficients ξ and Λ can allow us to draw conclusions about the magnitude of the bias on $\Lambda$and, by extension, allow us to calculate β. The results is summarized in Proposition 1.

Proposition 1. *Denote the R-squared in equation (2) as R_1 and the R-squared in equation (3) as R_2. Further, denote the full R-squared from Equation (1) as R_{max}. Under the given assumption of equal selection, $\beta = \Lambda - \frac{(\xi - \Lambda)(R_{max} - R_2)}{(R_2 - R_1)}$.*

Proof. The bias is $\frac{C_{z\tilde{x}}}{V_{\tilde{x}}}$. Since W and Z are orthogonal, $C_{z\tilde{x}} = C_{zx}$;. equal selection implies that $C_{zx} = C_{wx}V_z$. Further, $V_{\tilde{x}} = 1 - C_{xw}^2$. The bias is therefore $\frac{C_{wx}V_z}{1-C_{wx}^2}$. The difference between ξ and Λ is $\xi - \Lambda = C_{wx} + C_{wx}V_z - \frac{C_{wx}V_z}{1-C_{wx}^2}$. Dividing, I can express the relationship between this coefficient difference and the bias:

$$\xi - \Lambda = \left(\frac{C_{wx}(1 - C_{wx}^2 - C_{wx}^2 V_z)}{C_{wx}V_z} \right) \frac{C_{wx}V_z}{1 - C_{xw}^2}$$

Now consider the variances from equations (2) and (3):

$$\begin{aligned} V_\wr &= 1 + V_z - C_{xw}^2[1 + V_z]^2 \\ V_\tau &= V_z - \frac{[C_{wx}V_z]^2}{1 - C_{wx}^2} \end{aligned}$$

Straightforward simplification yields:

$$\frac{V_\tau}{V_\wr - V_\tau} = \frac{C_{wx}V_z}{C_{wx}(1 - C_{wx}^2 - C_{wx}^2 V_z)}$$

Which therefore implies that $\frac{C_{wx}V_z}{1-C_{xw}^2} = (\xi - \Lambda)\frac{V_\tau}{V_\wr - V_\tau}$. The definition of the R-squared in a linear model yields the result. □

The result directly relates coefficient movements to the bias, as well as giving a way to calculate the bias. Calculating this bias requires observing both coefficients and R-squared values from these regressions *and* making an assumption about the maximum R-squared. One assumption (the one adopted by AET) is that this value is 1: that if all of the unobservables were observed, they would explain all variation in Y. This assumption may be too strong in many cases where there is some either random component of Y (measurement error, for example) or some variables which predict Y but do not are orthogonal to X. Below I will discuss how one might develop such an assumption in an empirical context.

A straightforward corollary to this proposition consider the case in which there is another index of observed controls – call these M – which are fully observed, do not have a related unobserved component and are orthogonal to W and Z. In a health context these could be, for example, age or sex – baseline variables which explain some of the variation in Y and related to X but do not generate omitted variable concerns.[3] In this case, Equation (4) below is the full equation, and the two estimable equations are (5) and (6):

$$Y = \alpha + \beta X + W + Z + \Delta M + \gamma \quad (4)$$

$$Y = \hat{\alpha} + \xi X + \Delta M + \wr \quad (5)$$

$$Y = \tilde{\alpha} + \Lambda X + \Psi W + \Delta M + \tau \quad (6)$$

Corollary 1 summarizes the bias calculation in this case.

Corollary 1. *Denote the R-squared in equation (4) as R_{max}, the R-squared in equation (5) as R_1 and the R-squared in equation (6) as R_2. Under the assumption of equal selection, $\beta = \Lambda - \frac{(\xi-\Lambda)(R_{max}-R_2)}{(R_2-R_1)}$.*

Proof. Because M is fully observed, both the coefficient and variance expressions given in the proof to Proposition 1 hold and, therefore, the result goes through as above □

Before moving on, a brief note on the relationship between these results and those in AET. They consider an virtually identical setup, with two changes. First, they assume there is no γ, so $R_{max} = 1$. Second, they assume proportional selection but not (necessarily) equal selection. That is, they allow that $\delta C_{wx} = \frac{C_{zx}}{V_z}$,

[3] In practice, obvious elements of M like age or sex are often correlated with possible omitted variables. This is fine, we simply define the W and C category as the parts of those variables which are orthogonal to M.

where δ may be different than 1. In the results they show how this bias can be calculated using the data (by directly calculating variance and covariance values). Their calculation is identical to the one above, under the assumption that $R_{max} = 1$ and $\delta = 1$. Ultimately, the spirit of the results is the same, with a different free parameter. When I turn to estimation and validation in Section 4 I will show the results with their version as well.

2.2 Bias Results with Added Precision

The proposition in Section 2.1 gives a method for calculating bias using information on the movement of coefficients from the fully uncontrolled fully uncontrolled to the fully controlled regression. It follows simply from that result that if the coefficient on X doesn't change much from the fully uncontrolled to the fully controlled regression, this suggests limited bias. Effectively, this will only occur if C_{wx} is small, which then means the remaining bias is also small. A common, related, heuristic is to look for a slowing in the movement of coefficients as the number of controls increases. Even if there is a large change in the coefficient when some controls are added, if further controls do not change the coefficient very much, the conclusion is that the result is approaching the causal coefficient.

I capture this setup with the assumption that the true model is as follows:

$$Y = \alpha + \beta X + W_1 + W_2 + Z + \gamma$$

In this case, I imagine both W_1 and W_2 are observed, while Z is unobserved. I retain the assumption of equal selection, in this case assuming the variance of W_1 is 1:

$$C_{w_1 x} = \frac{C_{w_2 x}}{V_{w_2}} = \frac{C_{zx}}{V_z}$$

A common procedure in this case is to run the three regressions below in order, and compare the coefficients ξ, Λ_1 to Λ_2.

$$Y = \alpha + \xi X + \wr + \gamma \tag{7}$$

$$Y = \tilde{\alpha} + \Lambda_1 X + \Omega_1 W_1 + \tau + \gamma \tag{8}$$

$$Y = \hat{\alpha} + \Lambda_2 X + \Psi_1 W_1 + \Psi_2 W_2 + \kappa + \gamma \tag{9}$$

All three coefficients are biased, with the exact formulas given below.

$$\xi = \beta + C_{w_1 x}(1 + V_{w_2} + V_z)$$

$$\begin{aligned}\Lambda_1 &= \beta + \frac{C_{w_1x}(V_{w_2} + V_z)}{1 - C^2_{w_1x}} \\ \Lambda_2 &= \beta + \frac{C_{w_1x}V_z}{(1 - C^2_{w_1x}(1 + V^2_{w_2}))}\end{aligned}$$

The common heuristic holds that if Λ_1 and Λ_2 are close, even if ξ and Λ_1 are far apart, then the remaining bias on Λ_2 is small. The proposition below summarizes the condition for this to be the case.

Proposition 2. *A small difference in Λ_1 and Λ_2 relative to the difference between ξ and Λ_1 implies a small remaining bias if and only if a small V_{w_2} implies that V_z is small.*

Proof. The difference between Λ_1 and Λ_2 is $C_{w_1x}\frac{V_{w_2}[1-C^2_{w_1x}(1+V_{w_2}+V_z)]}{(1-C^2_{w_1x})(1-C^2_{w_1x}(1+V_{w_2}))}$. I can express this relative to $\xi - \Lambda_1$:

$$\Lambda_1 - \Lambda_2 = (\xi - \Lambda_1)\frac{V_{w_2}}{(1 - C^2_{w_1x}(1 + V_{w_2}))}$$

If this is small when $\xi - \Lambda_1$ is large, it implies that $\frac{V_{w_2}}{(1-C^2_{w_1x}(1+V_{w_2}))}$ is small. The overall bias on Λ_2 is $\frac{C_{w_1x}V_z}{(1-C^2_{w_1x}(1+V^2_{w_2}))}$. The assumption of a large difference between ξ and Λ_1 rules out the claim that C_{w_1x} is small. Therefore, this will be small only if $\frac{V_z}{(1-C^2_{w_1x}(1+V^2_{w_2}))}$ is small. Which is implied by the small difference between Λ_1 and Λ_2 only if a small V_{w_2} implies a small V_z. □

The mechanics of this claim are very straightforward, but do make it clear that using this heuristic requires two assumptions. The first is the equal selection assumption (or, more generally, that selection on observables guides selection on unobservables). In addition, one must assume that the fact that the second set of observables included are less important that the first set implies that the unobservables are *also* less important. This may be generally plausible, for example if we think that surveys mostly include important confounds and those that are left are of minimal importance. This analysis makes clear, however, that this is also testable, if one takes a view on the maximum R-squared. If V_z is small, then the R-squared from estimating Equation (9) should be close to $R_{max.}$ If it is *not*, then this implies that V_z must be large and, hence, one cannot conclude the remaining bias is small.

It is worth noting that observing that the R^2 stabilizes as further controls are added is also not informative. Under the equal selection assumption the R-squared values will move in proportion with the coefficient movements. Again, only if the R^2 value stabilizes at or close to R_{max} can one conclude that the remaining bias is small.

If one does not want to take a stand on the maximum R-squared, one could combine this pattern in the data with an assumption about proportional decay in the importance of controls. In the language of the above, the researcher could assume that $\frac{1}{V_{w_2}} = \delta\frac{V_{w_2}}{V_z}$ for some $\delta \geq 1$. Using this, if the coefficient stabilize one can automatically conclude that the further control will make only a small difference. This is a stronger and largely untestable assumption. Even if one observes this type of decay in the observables, it is no guarantee

that the unobserved components follow this pattern. In addition, if doing this it is crucial to include controls beginning with the control which explains the most variance in Y and follow with the second-most, etc.

2.3 Summary

The results in this section formalize some commonly used heuristics by expanding on the AET results. Together, they suggest several things.

First, movement in the coefficient of interest when controls are added is informative about remaining bias under the assumption of equal selection, but must be used along with some assumption about the maximum amount of variance explained by the observables and unobservables together.

Second, the relevant movement in the coefficient is that which occurs after inclusion of the set of controls for which we are concerned about omitted components. If a coefficient moves a lot after inclusion of a precise measure of individual age, this is probably not informative about how much further moment would be observed with controls for socioeconomic status. Controls of this type should be included in all regressions.

Third, stability in the coefficient of interest as controls are added is reassuring *only* if the R-squared stabilizes at or close to the maximum R-squared.

Together, this provides guidance in how these heuristics might be better used in practice. But it does not provide evidence on whether this procedure is effective in identifying causal impacts. To learn that, it is necessary to preform some validation. Below, I consider two validation examples. First, I explore a single setting (child IQ and early life influences) in detail. Second, I preform an estimation exercise using a large number of public health settings where we have both randomized and observational data. In both cases I ask whether the bias adjustment described would lead us to correct conclusions based on observational analyses.

3 Results: Impact of Early Life on Child IQ

I begin first with exploring a single application, estimating the impact of early life and prenatal influences on later child IQ. A literature in economics demonstrates that health shocks while children are in the womb can influence later cognitive skills (e.g. Almond and Currie, 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. These latter studies, however, are subject to significant omitted variable concerns. The behaviors which are linked to child IQ tend to also be closely linked to maternal socioeconomic status.

In this section I demonstrate how the adjustment described in Section 2 would work in this setting. I consider possible links between child IQ and four early life influences: breastfeeding, maternal drinking in pregnancy, low birth weight and maternal weight prior to conception. All of these have been linked in some

observational studies to lower child IQ.[4] However, when I look at the best studies – either randomized data or meta-analyses of high-quality observational data – these links are not all confirmed. Low birth weight does seem to contributes to lower IQ (Salt and Redshaw, 2006), a link which also has a biological underpinning (de Kieviet et al, 2012). Randomized data does not show a significant link between breastfeeding and full-scale IQ (Kremer et al, 2008). In the case of maternal drinking, it is known that heavy drinking in pregnancy can use significant disability, but the vast majority of evidence on light or occasional drinking indicates no impact on child IQ (see, for example: Faltreen Eriksen et al, 2012; O'Callaghan et al, 2007). The link between IQ and maternal pre-pregnancy weight has not been tested in a randomized settings but well controlled studies on many cohorts do not suggest a link (Brion et al, 2011) and there is no strong biological underpinning. It therefore seems unlikely this link is causal.

The analysis here represents both a demonstration of the method and a validation test. I show that in simple regressions with controls, there is a significant link between these behaviors and child IQ. I can then ask whether the adjustment procedure in Section 2 would lead to the "correct" conclusions about the validity of these relationships. I will also explore the extent to which the adjusted relationships are sensitive to variation in the control set, and compare this procedure to the heuristic of looking for stability in coefficients as controls are added.

3.1 Data

This section uses data from the National Longitudinal Survey of Youth Children and Young Adult Survey (NLSY). The NLSY is a longitudinal survey of women, and the Children and Young Adult module collects information on the children of NLSY participants. The outcome of interest is IQ for children aged 5 to 9, measured by PIAT test scores. The treatments of interest are: months of breastfeeding, maternal drinks per week during pregnancy, an indicator for low birth weight (<2500 grams) and an indicator for maternal overweight prior to conception. The last of these uses maternal weight as close as possible to the time of conception. These variables are summarized in the first rows of Table 1.

These data also contain demographic controls. These are summarized in the remainder of Table 1. They include: child age and sex, race, maternal age, maternal education, maternal income, maternal marital status and maternal AFQT score.

3.2 Results

Preforming the bias adjustment here requires the researcher to take a stand on two elements. First, what controls are in W and which are in M? Recall that M should contain any elements which may impact the coefficient but do not have omitted counterparts. In this case, it seems appropriate to include child age

[4]See Gabbe, Neibyl and Simpson (2006) for a summary of information on everything other than pre-pregnancy weight. For the latter, see Basatemur et al, 2013.

dummies and child sex in this vector. These are fully observed and we do not think, for example, that there is remaining bias associated with child age after we control for age dummies. These controls will be included in both the "uncontrolled" and controlled regressions.

The vector W includes the variables with omitted counterparts. In the primary results below I include standard maternal demographics: education, income, race, martial status and maternal age. These variables noisily measure socioeconomic status and certainly there are related variables which are not observed. I will also explore the robustness of the adjustment to adding or subtracting some from this W set. In principle, since such a change should move both the R-squared and the coefficient, the results should be relatively stable. This would be an attractive feature of this procedure.

The second important element is the maximum R-squared. How much of the variation in child IQ could be explained if I had full controls for family background? This is a figure for which we need to go outside the data. In this case it seems unlikely that it is 1 – even identical twins raised together do not have the same IQ scores. I suggest that the appropriate figure is the IQ correlation between siblings raised together. This captures the effects of family background which include maternal IQ. Based on the average correlations from the two studies reported in Scarr and Weinberg (1983) this figure is 0.385.[5]

Panel A of Table 2 shows the initial results using the most standard control set: maternal education, income, race, marital status and age. The first column shows the regression of IQ on treatment including only child age and sex and the second column reports the R-squared from this regression. The third and fourth columns show the coefficient and R-squared with controls. More breastfeeding is associated with higher IQ in these regressions, and low birth weight and maternal pre-pregnancy overweight status are associated with lower child IQ. Interestingly, 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; in a sense, this *must* be due to selection.

The final column in Panel A shows the bias-adjusted coefficient. Standard errors are calculated with a bootstrap over individuals. The adjustment preforms quite well. The relationship between low birth weight and IQ remains significant while none of the others do. Put simply, if I relied on the coefficients in Column 3 I would have conclude that breastfeeding increases IQ, as does maternal drinking in pregnancy, whereas maternal weight decreases it. The adjusted coefficients in Column 5 would lead to us to reject those associations, while continuing to accept that low birth weight impacts later child IQ. The bias-adjusted results reflect the best available evidence: breastfeeding doesn't impact IQ, maternal drinking does not affect IQ and low birth weight does. For the more equivocal issue of maternal weight, the evidence is consistent with biological theory in rejecting the claim that it impacts IQ and suggests the existing results may well be due to confounding.

Panel B shows the same analysis but includes maternal AFQT score in the controls; Panel C does a

[5]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.

similar thing but excludes maternal education (and excludes maternal AFQT). The goal here is to explore the sensitivity of these results to changes in the control set. Since the possible set of controls varies across studies, this procedure would appear to be most useful if it delivers similar implications even with variations in controls. Although the exact figures vary across panels, the qualitative conclusions are very similar. In all three cases the low birth weight relationship is confirmed by the bias adjustment and the three others are rejected.

These results suggest that doing this bias adjustment on these simple observational analyses would do a good job in separating true from false associations. 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 that they explain in the data; I include these controls in the same order in each analysis. Figures 1a-1d show coefficients and R-squared values for the four analyses.

These figures are not very useful for distinguishing among these analyses. All four show a very similar pattern of stabilizing coefficients. Based on these alone it would be quite difficult to identify one of the relationships as more robust than the others. In line with the discussion in Section 2.2, the issue is clear: the R-squared in the fully controlled regressions here is around 0.25, far below the figure of 0.385 that was drawn from existing data. Given this, the fact that the coefficient has stabilized is not fully informative.

I argue this section provides significant support for the use of this selection-on-observables adjustment. The regressions run here are very straightforward: the data is easy to obtain and the analysis easy to preform. Standard controlled regressions give clearly biased results – they contradict randomized data and, in one case, show coefficients which are clearly wrong-signed. Although of course more nuanced analyses, perhaps with more data or randomization, would be the optimal way to derive causal conclusions, the evidence here suggests that significant progress could be made using evidence on coefficient movements when controls are introduced. At the same time, this does suggest that the even simpler heuristic of looking for stability in coefficients may be problematic.

This suggests the value of this procedure in a particular context, and in a setting where we have the ability to think carefully about the maximum R-squared. In the next section I ask a broader validation question: in the general context of the link between positive health behaviors and health outcomes, can a version of this procedure help us separate the "true" from "false" associations *without* us having to carefully consider the maximum R-squared for a given setting?

4 Results: Health Behaviors and Health Outcomes

A large literature in epidemiology and public health looks to estimate the relationship between positive heath behaviors and health outcomes. 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.

In this section I combine observational data on a number of relationships estimated in the public health literature with randomized evidence on those relationships. In some cases, randomized trials have confirmed observational links and in others they have not. Very simply, I ask here whether a version of this adjustment would have led us to draw more accurate conclusions based on the observational data – more accurate in the sense that they better match the randomized results. This is similar to what is done in the previous section, with the main difference that here I do not research the likely maximum R-squared in each setting. Instead, for each setting I *estimate* the maximum R-squared that would match the randomized data. I then ask whether a single adjustment value might provide better inference in a number of settings.

This section serves both as validation and, potentially, provides guidance for using this adjustment outside of this paper. From a validation standpoint, this addresses the question of whether this procedure could be used to draw better conclusions. Although I am preforming estimation – effectively, fitting the observational data to the truth – the model *is* falsifiable. This procedure will only work for all the cases together if the coefficients move more with controls in cases where the link is not causal. Put differently, for any given relationship I can likely estimate an value of R_{max} which would lead me to the correct conclusion (as long as the coefficient moves in the correct direction). But there is no guarantee that a similar value will work for many settings.

Going forward, this estimation provides guidance for evaluating the robustness of other work in this area (either existing or new). This section will suggest a precise adjustment, which could be applied elsewhere and used to comment more concretely on the likelihood that an observational link will be confirmed in randomized data. It is important, of course, to be clear on the set of settings for which this is likely to be valid. I am concerned with settings where the left hand side variable is some health outcome and the right hand side variable of interest is some health behavior. Further, these are all settings where the omitted variables are around the issue of socioeconomic status. I would argue that this covers many interesting settings in public health, although of course not all.

4.1 Data

This analysis requires two pieces of data: randomized trial results and observational data. I discuss these in turn.

Randomized Trials

Randomized trial results are drawn from existing work.

Exercise Evidence on the impact of exercise is drawn from a several papers which are summarized in a

Cochrane Review meta-analysis (Shaw et al, 2006). I consider only studies which compared exercise to not exercise; this excludes studies which compared a combination of diet and exercise to diet alone, or a combination of diet and exercise to no treatment. Outcomes considered include weight, blood pressure, cholesterol, blood glucose, 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.

Breastfeeding Evidence on the impact of breastfeeding is drawn from a large randomized study called the PROBIT study, run in Belarus in the 1990s and with follow-up through early childhood (Kramer et al, 2009). This was an encouragement design with much less than full take-up so I scale the impacts to reflect the increase in breastfeeding at 3 months. Outcomes used are child weight and height.

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 is also 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.

Breastfeeding I again use the National Longitudinal Survey of Youth Children and Young Adult Survey (NLSY) for breastfeeding. In this case the outcomes of interest are BMI and height in centimeters. The treatment is breastfeeding at 3 months. These variables are summarized in Panel C of Table 3. The randomized evidence on breastfeeding measured children at age 6.5. The NSLY sample size is too small to limit to only 6 year olds so I use 5, 6 and 7-year-olds for this analysis.

4.2 Empirical Strategy

Recall that the bias-adjusted coefficient is calculated :

$$\beta = \Lambda - \frac{(\xi - \Lambda)(R_{max} - R_2)}{R_2 - R_1}$$

where ξ and R_1 are the coefficient and R-squared from the fully uncontrolled regression (or the regression including only the orthogonal controls like age and sex) and Λ and R_2 are the coefficient and R-squared from the fully controlled regression. The free parameter is R_{max}. Broadly, the empirical strategy is to estimate – for each behavior-outcome pair – a value for R_{max}which would most closely match the randomized conclusions and combine these to estimate a value which would minimize the error across all settings.

A key issue is the parametrization of R_{max}. One option would be to simply estimate a value for R_{max} which would be the same across all settings. This is somewhat unappealing if, as is true in our settings, outcomes differ in their predictability, either due to measurement error, or other noise which is uncorrelated with the treatment behavior.

Instead, I assume that $(R_{max} - R_2) = \psi(R_2 - R_1)$ and estimate ψ. Effectively, this assumes that the amount of Y which is explained by the observables is a guide to how much would be explained by the unobservables. A value of $\psi = 1$ would imply that the unobservables explain as much of the variation in Y as the unobservables; a value larger than 1 implies that the unobservables would explain more, a value less than 1, that they explain less. In addition to having some intuitive appeal, this is a convenient assumption when the goal is to use the conclusions to evaluate existing work. With this assumption, the calculation of the bias-adjusted coefficient collapses to $\beta = \Lambda - \psi(\xi - \Lambda)$ and it is not necessary to observe the R-squared values. Since published papers in public health and epidemiology only very rarely report these values, this makes this procedure significantly more useful.

Armed with this assumption, the estimation is straightforward. For each outcome treatment pair I estimate the uncontrolled regression (which could include some simple controls as described above) and the controlled regression. The selection of controls is addressed below. Given this, and some assumption on ψ, it is possible to calculate a bias-adjusted estimate. For outcome-treatment pair i denote this adjusted coefficient $\beta^i_{adj}(\psi)$. From the randomized trial I obtain a value β^i_{true} which is the measure of the true causal coefficient. 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 ψ. Formally, I solve:

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

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 ψ relative to the assumption that $\psi = 0$ (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, there are a few outcomes for which the trials suggests a conclusion about the null hypothesis but where matching magnitudes is difficult. It is possible to preform an "out-of-sample" test using these outcomes and exploring whether the same adjustment would lead to more accurate conclusion in these cases.

A final issue which needs discussion is the selection of the control set. In general in these settings, omitted variable bias concerns center around socioeconomic status. Individuals who have more education, are wealthier or have more stable home lives are more likely to undertake any given positive health behavior but also are healthier for other reasons. Most studies observe some rough measures of this – education category, income category, marital status – but not detailed data. I therefore proceed as if this is *the* omitted category. I include available socioeconomic controls – typically, education, income, marital status and race – in W. Effectively, this assumes that whatever omitted variables there are in these regressions, they are proxied by these socioeconomic status variables. A related question is what controls should be included in both the "uncontrolled" and controlled regressions. As in Section 3 it seems natural to include age and sex (the latter only in cases where both sexes are included in the trial). In addition, in cases where I estimate the impacts on weight in kilograms I also include a control for height.

4.3 Results

The estimation procedure described above yields a value of $\psi = 1.018$. This suggests that the omitted characteristics explain approximately as much of the variation in outcome as the included characteristics.

I begin by illustrating the impact of the 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 2a 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.

Figure 2b shows the coefficients after the bias adjustment is done with the value of $\hat{\psi} = 1.018$. The fit is significantly better; note the large decrease in scale (the bias-adjusted coefficients on the same scale as the controlled coefficients can be seen in Appendix Figure 1). 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 30% on average.

Table 4 describes the results numerically. Column 1 shows the magnitude of the impacts in randomized trials (in the case of exercise, where there are often multiple studies used, I show a range), and indicates significance.[6] Columns 2 and 3 show the uncontrolled (i.e. with only age and sex) coefficients and the coefficients with socioeconomic status controls. Column 4 shows the bias-adjusted coefficients. Standard errors in the bias-adjusted case are bootstrapped over individuals. It is worth noting that the randomized experiments here use much larger sample sizes than the observational data and are therefore able to detect much smaller impacts. Although I report conclusions on the null hypothesis as well as the sample size, it is therefore worth keeping in mind that in some cases (for example, the impact of vitamin supplementation on weight) the observational data has no where near enough power to detect impacts of the size seen in the randomized data.

The evidence in this table shows bias-adjusted impacts which are much closer to the estimates from the randomized data; this is not surprising given the evidence in Figure 2b. In addition, 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, whereas the bias-adjusted coefficient is very close to the small and insignificant impact estimated in randomized trials. A similar story can be told for the impact of supplementation on serum glucose and the impact of breastfeeding on child weight.

In the case of vitamin supplementation and weight, while the randomized impact is significant it is very small. The simple controlled coefficients suggests an impact of about 1.5 kilograms on weight, where as the randomized impact is only about 0.1 kilograms. The bias-adjusted coefficient is quite close to this in size. Although it is not significant, this reflects the fact that the observational data is simply under-powered to detect significant effects of that magnitude.

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. This bias-adjustment does a good job of identifying true from false associations among those which simple controlled regressions show are significant.

The estimation preformed here takes advantage of outcome-treatment pairs where I can generate comparable magnitudes. In the case of exercise and Vitamin D there are also 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

[6] In the case of exercise this significance is based on estimates from a meta-analysis (Shaw et al, 2006).

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.

This is done in Table 5. This table is structured similarly to Table 4 except that in the first column I simply report the hypothesized direction and significance (or not) of the effect in the randomized trial. In general, the bias-adjustment also preforms 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. Obviously this is a very small list, but it provides some "out-of-sample" evidence on the fit of the adjustment.

The theory here draws heavily on the discussion in Altonji, Elder and Taber (2005), as noted. However, a primary difference is that I assume equal selection and allow the maximum R-squared to vary, whereas they assumed the maximum R-squared was equal to 1 and suggest there may be variation in the degree of proportionality of selection. That is, they suggest that rather than assuming $C_{wx} = \frac{C_{zx}}{V_z}$ one assumes that $\delta C_{wx} = \frac{C_{zx}}{V_z}$, with δ as a "free" parameter, but that $\gamma = 0$. Both approaches allow for one free parameter and the same estimation which I describe above can also be preformed with their version of the adjustment. The analog to Figure 2b, but with the best-fit value of δ used for the adjustment, is shown in Appendix Figure 2. The estimated value of $\delta = 0.055$. Not surprisingly, this adjustment is also a better fit than the simple controlled coefficients, as it relies on the same basic idea that the coefficients move more when the controls are more important.

However, the fit is less good than in Figure 2b, and the reduction in error is only 20%. In addition, many more of the estimates fall outside the randomized confidence interval. The primary issue is that in many of these cases the actual R-squared, even with the full set of controls, is quite small, often under 10%. Assuming that the unobservables would explain all of the additional variation in Y may be less appropriate in these settings.

5 Conclusions

The goal of this paper is two-fold. First, I expand on the framework in Altonji, Elder and Taber (2005) and connect the idea of equal selection explicitly to coefficient movements. I show circumstances under which such movements can be used to generate causal coefficients under the equal selection assumption. I provide some guidance to discipline the use of this coefficient movement heuristic. I provide a simple form of the adjustment using only information on coefficient and R-squared values. In particular, I argue that under the assumption of equal selection, the causal coefficient β can be recovered from the uncontrolled coefficient, ξ, the coefficient

with controls, Λ, the R-squared from the uncontrolled and controlled regressions (R_1 and R_2) and an assumption about the maximum R-squared (R_{max}). The exact calculation is:

$$\beta = \Lambda - \frac{(\xi - \Lambda)(R_{max} - R_2)}{(R_2 - R_1)}$$

Second, I describe two validation exercises. I argue that, regardless of the intuitive appeal of this approach, it has value only if it is effective in drawing better causal conclusions. In two validations I show this approach preforms well. In both validation exercises I consider cases where there exists both observational data which may be biased alongside either randomized data or better observational studies which are more likely to reflect a "true" relationship.

In the case of child IQ and early life influences I show that a carefully applied version of this approach does a good job of separating out true associations (between low birth weight and child IQ) from false ones (positive impacts of breastfeeding and maternal drinking on child IQ, negative impacts of maternal weight). This application also urges caution in using the heuristic of including controls until the coefficient stops moving. I show – both in theory and in practice – that that approach only works if one is confident that the final R-squared is at or close to the maximum possible R-squared.

The second validation exercise takes a number of settings and asks whether I can estimate a general version of the adjustment which would lead to better conclusions. I consider settings where (a) the outcome is a health outcome and the treatment is a health behavior and (b) the primary omitted variable bias comes from socioeconomic status, broadly construed. I argue that this applies to many relationships of interest and is not limited to the ones I consider here. I approach this as an estimation with the assumption that $(R_{max} - R_2) = \psi(R_2 - R_1)$, and ψ as a parameter to be estimated. I find that a value of $\psi = 1.018$ provides a much better fit to randomized results than the simple controlled coefficients. With bootstrapped standard errors it rejects a number of false-positive associations with limited cost in terms of rejecting true-positive ones.

To the extent that one is comfortable porting these results into other contexts, this suggests a simple way for researchers to evaluate the plausibility of their results, and for readers of published work to do so, as well. In particular, the results suggest the following adjustment:

$$\beta = \Lambda - 1.018(\xi - \Lambda)$$

This adjustment can be done without knowing the R-squared values (often not provided in public health papers). The results in this paper suggest not only can this adjustment be done, but doing so would, in many settings, lead one much closer to the true β.

References

Almond, Douglas 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.

Basatemur, E., J. Gardiner, C. Williams, E. Melhuish, J. Barnes, and A. Sutcliffe, "Maternal Prepregnancy BMI and Child Cognition: A Longitudinal Cohort Study," *Pediatrics*, Jan 2013, *131* (1), 56–63.

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.

Brion, M. J., M. Zeegers, V. Jaddoe, F. Verhulst, H. Tiemeier, D. A. Lawlor, and G. D. Smith, "Intrauterine effects of maternal prepregnancy overweight on child cognition and behavior in 2 cohorts," *Pediatrics*, Jan 2011, *127* (1), e202–211.

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.

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.

Eriksen, H. L. Falgreen, 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.

Gabbe, Stephen, Jennifer Niebyl, and Joe Simpson, *Obstetrics: Normal and Problem Pregnancies*, Philadelphia, PA: Churchill Livingstone, 2006.

Hellenius, M. L., U. de Faire, B. Berglund, A. Hamsten, and I. Krakau, "Diet and exercise are equally effective in reducing risk for cardiovascular disease. Results of a randomized controlled study in men with slightly to moderately raised cardiovascular risk factors," *Atherosclerosis*, Oct 1993, *103* (1), 81–91.

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.

___ , L. Matush, I. Vanilovich et al., "A randomized breast-feeding promotion intervention did not reduce child obesity in Belarus," *J. Nutr.*, Feb 2009, *139* (2), 417S–21S.

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.

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.

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.

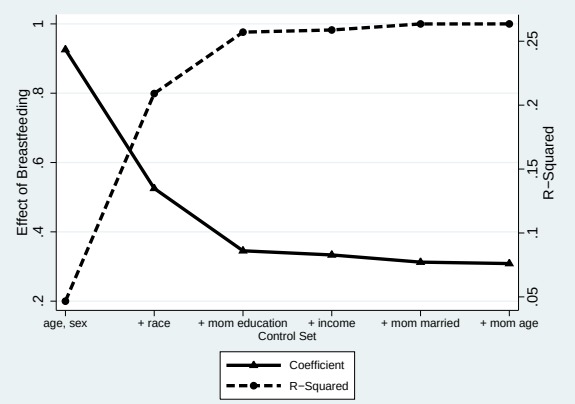
Shaw, Kelly, Hanni Gennat, Peter ORourke, and Chris Del Mar, "Exercise for overweight or obesity," *Cochrane Database of Systematic Reviews*, 2006, (4).

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.

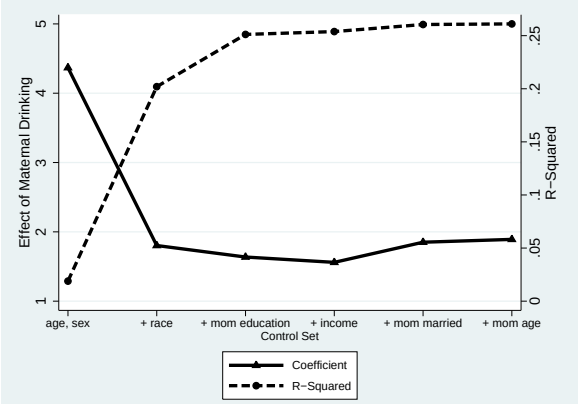
Wood, P. D., M. L. Stefanick, D. M. Dreon et al., "Changes in plasma lipids and lipoproteins in overweight men during weight loss through dieting as compared with exercise," *N. Engl. J. Med.*, Nov 1988, *319* (18), 1173–1179.

Figure 1: **Coefficient Stability, Child IQ and Early Life Influences**

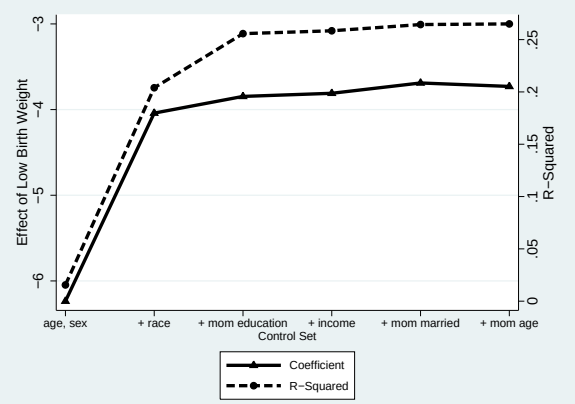
(a) Months of Breastfeeding



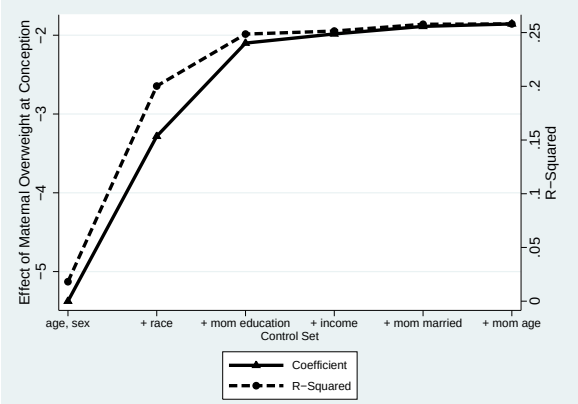
(b) Maternal Drinking Pregnancy



(c) Low Birth Weight (<2500 grams)



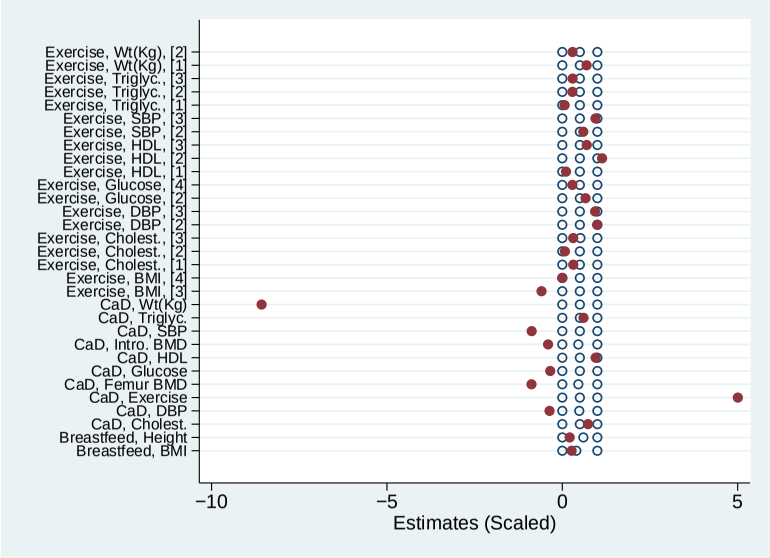
(d) Maternal Overweight before Birth



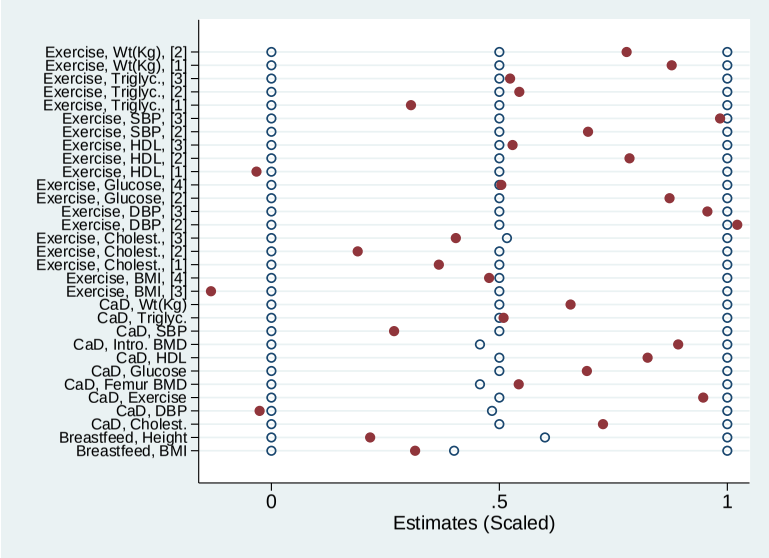
Notes: These graphs show the evolution of the estimated relationship between each treatment and child IQ as controls are added. Controls are added in the same order in each figure. The order is chosen based on ordering the controls by how much of IQ they explain and including the most important first.

Figure 2: **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 $\psi = 1.018$. 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 1: **Summary Statistics: Early Life and Child IQ**

Outcome	*Mean*	*Standard Deviation*	*Sample Size*
IQ (PIAT Score)	89.1	20.5	7531
Breastfeeding Months	2.26	4.45	7018
Mom Drink in Pregnancy	0.325	0.468	7039
Low Birth Weight (<2500 g)	0.083	0.276	6978
Mom Overweight before Pregnancy	0.172	0.377	7488
Age	5.96	1.68	7531
Female	0.494	0.499	7531
Black	0.288	0.453	7531
Mother Age	24.8	5.35	7531
Mother Education	12.3	3.1	7531
Mother Income	$39,622	$79,904	7531
Mother Married	0.644	0.478	7531
Mother AFQT	36.2	27.3	7345

Notes: This table shows summary statistics for the data used in the IQ analysis in Section 3. Data comes from the NLSY and is limited to children aged 5 to 9.

Table 2: **Early Life Behaviors and Child IQ**

Panel A: Baseline Controls (Education, Income, Marital Status, Maternal Age, Race)						
Treatment Variable	*Baseline Effect*	*Baseline* R^2	*Effect with Full Controls*	*Controls* R^2	*Maximum R-Squared*	*Bias-Adjusted Coefficient*
Breastfeeding (Months)	0.925*** (.062)	.047	0.308***(.058)	.263	0.385	-0.037 (.082)
Drinking in Pregnancy	4.368***(.593)	.019	1.891***(.524)	.261	0.385	0.623 (.581)
Low Birth Weight	-6.23***(1.01)	.016	-3.72***(.876)	.265	0.385	-2.52***(.95)
Overweight Before Pregnancy	-5.37***(.706)	.035	-1.85***(.621)	.258	0.385	0.004 (.63)
Panel B: More Controls (Add Maternal AFQT Score)						
Treatment Variable	*Baseline Effect*	*Baseline* R^2	*Effect with Full Controls*	*Controls* R^2	*Maximum R-Squared*	*Bias-Adjusted Coefficient*
Breastfeeding (Months)	0.925*** (.062)	.052	0.167***(.058)	.294	0.385	-0.108 (.084)
Drinking in Pregnancy	4.368***(.593)	.012	0.871* (.514)	.299	0.385	-0.189 (.55)
Low Birth Weight	-6.23***(1.01)	.016	-2.89***(.854)	.304	0.385	-1.95**(.96)
Overweight Before Pregnancy	-5.37***(.706)	.035	-1.55***(.60)	.296	0.385	-0.34 (.61)
Panel B: Fewer Controls (No Income Control)						
Treatment Variable	*Baseline Effect*	*Baseline* R^2	*Effect with Full Controls*	*Controls* R^2	*Maximum R-Squared*	*Bias-Adjusted Coefficient*
Breastfeeding (Months)	0.925*** (.062)	.052	0.406***(.059)	.226	0.385	-0.050 (.084)
Drinking in Pregnancy	4.368***(.593)	.012	2.108***(.536)	.226	0.385	0.382 (.591)
Low Birth Weight	-6.23***(1.01)	.016	-3.91***(.898)	.227	0.385	-2.19**(1.06)
Overweight Before Pregnancy	-5.37***(.706)	.035	-2.61***(.63)	.223	0.385	-0.45 (.68)

Notes: This table shows the validation results for the analysis of the impact of early life and prenatal behaviors on child IQ. Baseline effects include only controls for child age (dummies) and sex. Full control effects include the listed controls. The bias-adjusted effect is generated using the formula derived in Section 2: $\beta = \Lambda - \frac{(R_{max}-R_2)}{(R_2-R_1)}(\xi - \Lambda)$. 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, Vitamins and Breastfeeding**

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
Panel C: Breastfeeding [NLSY]			
Outcome	*Mean*	*Standard Deviation*	*Sample Size*
Breastfed >=3 months	.280	.448	10,085
BMI	16.1	3.05	9917
Height (cm)	116.2	9.93	10,923

Notes: This table shows summary statistics for the data used in the validation exercise in the paper. NLSY = National Longitudinal Survey of Youth; NHANES-III : National Longitudinal Health and Nutrition Survey, Wave III. For Exercise, the sampel restirctions in the analysis differ slightly depending on which paper I am comparing to. For the summary statistics we consider the most inclusive definition.

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

Panel A: Exercise				
Outcome	*Randomized Effect [Range if multiple]*	*Uncontrolled Effect (Std. Error)*	*Controlled Effect (Std. Error)*	*Bias-Adjusted Effect (Std. Error)*
BMI	[-0.6, -1.1]**	-1.99** (.29)	-1.55**(.29)	-1.10**(.32)
Weight in Kg	[-4.6, -1.15]**	-5.60**(.82)	-4.55**(.83)	-3.49**(.89)
Diastolic Blood Pressure	[-3, -1.8]**	-1.36**(.58)	-1.20**(.58)	-1.04*(.60)
Systolic Blood Pressure	[-4, 0.2]	-0.98 (.84)	-0.19 (.84)	0.74 (.90)
Serum Glucose	[-0.19, -0.16]**	-0.31**(.062)	-0.21**(.065)	-0.098 (.076)
Triglycerides	[-0.2,-0.16]**	-0.29**(.062)	-0.21**(.066)	-0.13*(.071)
Cholesterol	[-0.02,0.05]	-0.026 (.063)	-0.0004 (.062)	0.025 (.063)
HDL	[0.03,0.13]**	0.13**(.025)	0.11**(.025)	0.092**(.026)
Panel B: Vitamin D and Calcium Supplementation				
Outcome	*Randomized Effect [Range if multiple]*	*Uncontrolled Effect (Std. Error)*	*Controlled Effect (Std. Error)*	*Bias-Adjusted Effect (Std. Error)*
Weight in Kg	-0.13**	-2.79**(.69)	-1.44**(.69)	-0.06 (.78)
Diastolic Blood Pressure	0.11	-0.255 (.45)	-0.152 (.46)	-0.048 (.52)
Systolic Blood Pressure	0.22	-1.12 (.92)	-0.52 (.94)	0.095 (1.05)
Serum Glucose	-0.82	-6.92**(2.11)	-3.58*(2.22)	-0.19 (2.59)
Triglycerides	1.43	4.47 (5.16)	3.30 (5.39)	1.57 (6.20)
Cholesterol	-1.67	0.199 (2.15)	0.156 (2.24)	0.112 (2.52)
HDL	0.050	1.28 (.86)	1.02 (.85)	0.75 (.96)
Exercise Intensity (METS/wk)	0.18	5.27**(1.03)	2.88**(1.06)	0.44 (1.20)
Femur BMD	0.007**	-0.018**(.006)	-0.006 (.006)	0.007 (.007)
Introchanter BMD	0.0003	-0.020**(.008)	-0.008 (.008)	0.004 (.010)
Panel C: Breastfeeding				
Outcome	*Randomized Effect [Range if multiple]*	*Uncontrolled Effect (Std. Error)*	*Controlled Effect (Std. Error)*	*Bias-Adjusted Effect (Std. Error)*
BMI	0	-0.25** (.078)	-0.18**(.082)	-0.11 (.094)
Height (in cm)	2.43	0.28 (.20)	0.32 (.21)	0.36 (.25)

Notes: This table displays the match between the results from observational data and randomized results. Citations for randomized data nd 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). Controls in Panel C: dummies for age and sex (controlled and uncontrolled regressions), maternal age, dummies for race, income, maternal education, maternal marital status (controlled regressions only). The bias-adjustment is preformed using a valeu of $\psi = 1.018$. 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	*Randomized Effect*	*Uncontrolled Effect*	*Controlled Effect*	*Bias-Adjusted Effect*
	[Possible Direction, Sig.]	*(Std. Error)*	*(Std. Error)*	*(Std. Error)*
Ever Diabetes	Negative, Not Significant	-0.035**(.009)	-0.019** (.009)	-0.003 (.010)
Mortality, with heart disease, Men	Negative, Significant	-0.132**(.041)	-0.115**(.041)	-0.098**(.05)
Overall Bone Density, Women	Positive, Not Significant	-0.013 (.012)	-0.0003 (.012)	0.013 (.014)
Panel B: Vitamin D and Calcium Supplementation				
Outcome	*Randomized Effect*	*Uncontrolled Effect*	*Controlled Effect*	*Bias-Adjusted Effect*
	[Possible Direction, Sig.]	*(Std. Error)*	*(Std. Error)*	*(Std. Error)*
Ever Diabetes	Negative, Not Significant	-0.058**(.017)	-0.023 (.016)	0.002 (.018)
Mortality	Negative, Not Significant	-0.058**(.019)	-0.034*(.020)	-0.010 (.023)

Notes: Exercise treatment: total exercise times per month (in units of 100). 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).

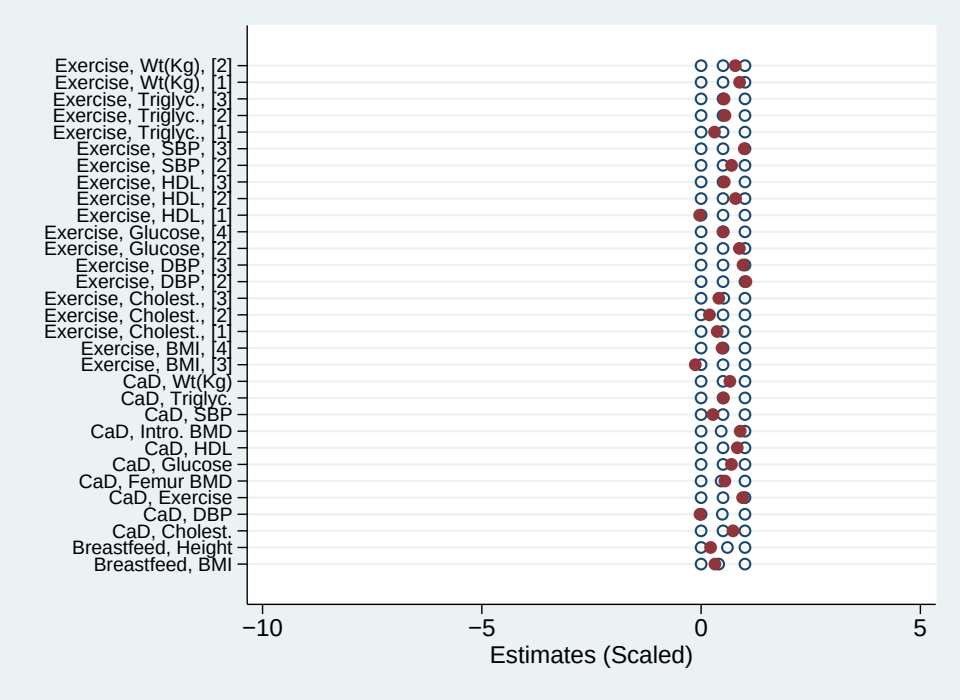
Appendix Tables and Figures

Table A1: Citation for Randomized Outcomes

Outcome	*Citation*	*Sample Restrictions (if any)*
Exercise, BMI, [3]	Hellenius et al, 1993	Men, 35-60, no heart disease
Exercise, BMI, [4]	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, DBP, [3]	Hellenius et al, 1993	Men, 35-60, no heart disease
Exercise, SBP, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, SBP, [3]	Hellenius et al, 1993	Men, 35-60, no heart disease
Exercise, Glucose, [2]	Stefanick et al, 1998	Women 45-64, men 30-64, no heart disease
Exercise, Glucose, [4]	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, Triglyc, [3]	Hellenius et al, 1993	Men, 35-60, 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, Cholest, [3]	Hellenius et al, 1993	Men, 35-60, 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
Exercise, HDL, [3]	Hellenius et al, 1993	Men, 35-60, 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
Breastfeed, BMI	Kramer et al, 2009	Age 6.5
Breastfeed, Height	Kramer et al, 2009	Age 6.5

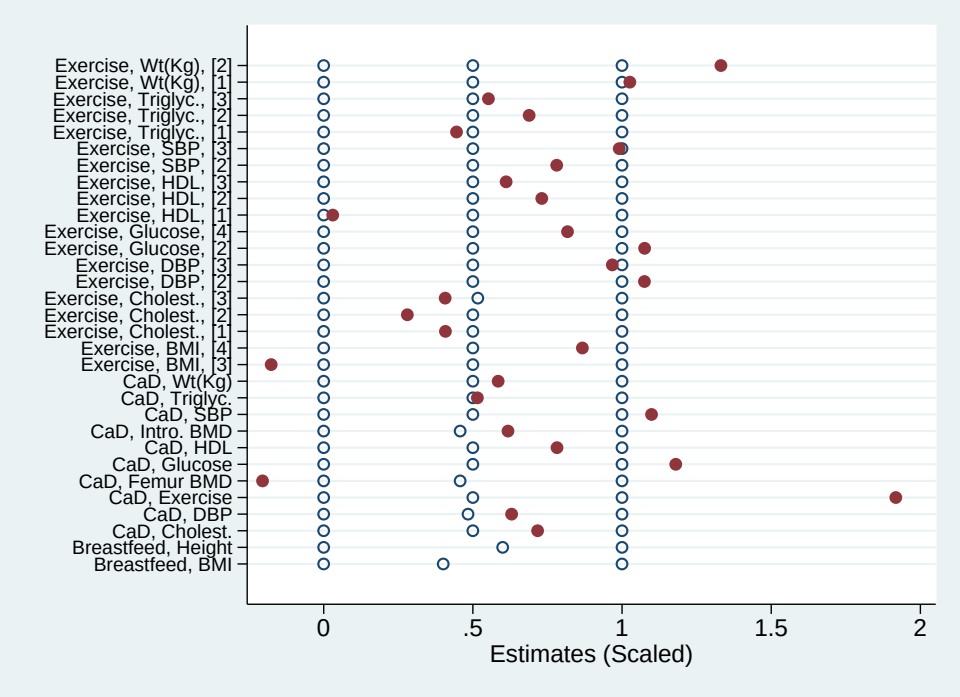
Notes: This table shows the source of the randomized estimates. The text of the outcome matches the form of citation in Figure 2.

Appendix Figure 1: Adjusted Coefficients on Controlled Coefficient Scale



Notes: This table shows Figure 2b graphed on the same scale as Figure 2a.

Appendix Figure 2: Bias-Adjusted Coefficients Assuming No Noise and Proportion (Not Equal) Selection



Notes: This figur shows the bias-adjuted coefficients with the best-fit value of δ, where δ is defined so $\delta C_{wx} = \frac{C_{zx}}{V_z}$. We assume that the maximum R-squared is equal to 1. The best-fit value of δ is 0.055.