10-2016

Constructed Second Control Groups and Attenuation of Unmeasured Biases

Samuel D. Pimentel

Dylan S. Small
University of Pennsylvania

Paul R. Rosenbaum
University of Pennsylvania

Recommended Citation

Pimentel, S. D., Small, D. S., & Rosenbaum, P. R. (2016). Constructed Second Control Groups and Attenuation of Unmeasured Biases. *Journal of the American Statistical Association, 111* (515), 1157-1167. http://dx.doi.org/10.1080/01621459.2015.1076342

Constructed Second Control Groups and Attenuation of Unmeasured Biases

Abstract
The informal folklore of observational studies claims that if an irrelevant observed covariate is left uncontrolled, say unmatched, then it will influence treatment assignment in haphazard ways, thereby diminishing the biases from unmeasured covariates. We prove a result along these lines: it is true, in a certain sense, to a limited degree, under certain conditions. Alas, the conditions are neither inconsequential nor easy to check in empirical work; indeed, they are often dubious, more often implausible. We suggest the result is most useful in the computerized construction of a second control group, where the investigator can see more in available data without necessarily believing the required conditions. One of the two control groups controls for the possibly irrelevant observed covariate, the other control group either leaves it uncontrolled or forces separation; therefore, the investigator views one situation from two angles under different assumptions. A pair of sensitivity analyses for the two control groups is coordinated by a weighted Holm or recycling procedure built around the possibility of slight attenuation of bias in one control group. Issues are illustrated using an observational study of the possible effects of cigarette smoking as a cause of increased homocysteine levels, a risk factor for cardiovascular disease. Supplementary materials for this article are available online.

Keywords
attenuation of unmeasured biases, casual inference, observational study, second control group, sensitivity analysis

Disciplines
Business | Business Analytics | Management Sciences and Quantitative Methods | Statistics and Probability

This is an Accepted Manuscript of an article accepted by Taylor & Francis for publication in the Journal of the American Statistical Association and posted online 7 August 2015, available online: http://amstat.tandfonline.com/10.1080/01621459.2015.1076342.

Constructed second control groups and attenuation of unmeasured biases

Samuel D. Pimentel, Dylan S. Small, Paul R. Rosenbaum[1]

University of Pennsylvania, Philadelphia

Abstract. The informal folklore of observational studies claims that if an irrelevant observed covariate is left uncontrolled, say unmatched, then it will influence treatment assignment in haphazard ways, thereby diminishing the biases from unmeasured covariates. We prove a result along these lines: it is true, in a certain sense, to a limited degree, under certain conditions. Alas, the conditions are neither inconsequential nor easy to check in empirical work; indeed, they are often dubious, more often implausible. We suggest the result is most useful in the computerized construction of a second control group, where the investigator can see more in available data without necessarily believing the required conditions. One of the two control groups controls for the possibly irrelevant observed covariate, the other control group either leaves it uncontrolled or forces separation; therefore, the investigator views one situation from two angles under different assumptions. A pair of sensitivity analyses for the two control groups is coordinated by a weighted Holm or recycling procedure built around the possibility of slight attenuation of bias in one control group. Issues are illustrated using an observational study of the possible effects of cigarette smoking as a cause of increased homocysteine levels, a risk factor for cardiovascular disease.

Keywords: Attenuation of unmeasured biases; causal inference; observational study; sensitivity analysis; second control group.

[1]Samuel Pimentel is a doctoral student and Dylan Small and Paul Rosenbaum are professors in the Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340 US. spi@wharton.upenn.edu. Supported by the Measurement, Methodology, and Statistics Program of the National Science Foundation and by Fellowship FA9550-11-C-0028 from the Department of Defense, Army Research Office, National Defense Science and Engineering Graduate (NDSEG) Fellowship Program, 32 CFR 168a4. 11 June 2015.

1 Introduction: background; motivating example

1.1 Is it advantageous to omit adjustments for some measured covariates?

In an observational study of treatment effects, treatments are not randomly assigned to individuals, so treated and control groups are often visibly different in terms of measured pretreatment covariates $\mathbf{x}$, and may differ in terms of unmeasured covariates u. Differing outcomes in treated and control groups after treatment may reflect the lack of comparability of these groups before treatment, rather than an effect caused by the treatment. It is common to adjust for the observed covariates $\mathbf{x}$, perhaps by matching individuals with the same $\mathbf{x}$, and to examine the sensitivity of conclusions to assumptions about unobserved covariates u.

It is sometimes argued informally that parts of $\mathbf{x}$ may be irrelevant, and that there would be less bias from u if adjustments were not made for the parts of $\mathbf{x}$ that are irrelevant; see Brooks and Ohsfeldt (2013) and Sanni Ali et al. (2014) for two general perspectives on this issue, and see Walker (2013) and Zubizarreta et al. (2012) for discussion of a specific situations. The intuitive idea is that it is desirable that something irrelevant decides treatment assignment — that is similar to what happens in a randomized experiment — and if one removes every irrelevant aspect of treatment assignment, one is left with biases from u deciding treatment assignment. Under what circumstances does this line of reasoning have a rigorous basis?

1.2 Motivating example: Does smoking increase homocysteine levels?

To permit a tangible discussion, consider an interesting study by Bazzano et al. (2003) concerned with the possibility that cigarette smoking causes an increase in homocysteine levels, a possible risk factor for cardiovascular disease. Bazzano et al. (2003) compared smokers and nonsmokers in NHANES adjusting for certain covariates, $\overline{\mathbf{x}}$, that might have

a direct biological connection with homocysteine levels, such as age, race and body mass index. They did not adjust for income and education, $\widetilde{\mathbf{x}}$, two covariates strongly related to smoking. In the US today, smoking is much less common among more educated, higher income individuals than among less educated, lower income individuals. Should one adjust for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ jointly or is it better to adjust for $\overline{\mathbf{x}}$ alone? One might argue that income and education have no known direct biological effect on homocysteine levels, so it makes sense to compare poor, less educated smokers to wealthier, better educated nonsmokers, because then something irrelevant has decided whether an individual smokes or not. Conversely, one might argue that one should adjust for all of $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ because education and income are associated with many aspects of daily life that could affect homocysteine levels, from exercise to diet to the quality of health care. Our goal is to shed some light on this decision and related options for study design and analysis.

Figure 1 displays 1536 distinct individuals in $I = 512$ matched triples containing one daily smoker and two nonsmokers from NHANES 2005-2006. Smokers smoked every day for the last 30 days and reported smoking at least 10 cigarettes per day (median = 20). Nonsmokers did not smoke at all in the last 30 days and had smoked fewer than 100 cigarettes in their lives. All controls were matched to smokers for biological covariates, $\overline{\mathbf{x}}$, including age, gender, race (black/other), (Hispanic/other), and body mass index (BMI). Controls labeled M were also matched for two socioeconomic (SES) measures, $\widetilde{\mathbf{x}}$, namely education on a five point scale (with 1 meaning $<$ 9th grade, 3 meaning high school graduate, and 5 meaning at least a BA degree) and income recorded as the ratio of income to the poverty level capped at 5 times poverty. Controls labeled P were pushed apart in terms of $\widetilde{\mathbf{x}}$, that is, they had high levels of education and income. Notably in Figure 1, the three groups are similar in terms of biological covariates, the smokers and M-controls are similar in terms of SES, and the P-controls have higher education and income than the

smokers. There is an obvious sense in which the M-controls are better than the P-controls: they are similar to smokers in terms of SES. Is there any sense in which the P-controls are better than the M-controls?

Section 2 reviews definitions and notation from existing literature. Section 3 considers the possibility that ignoring an irrelevant covariate $\widetilde{\mathbf{x}}$ attenuates bias from an unmeasured covariate u, concluding that it is possible, but the assumptions required are heroic and even then the magnitude of the attenuation is meaningful but not large. Also discussed is the possibility that forcing separation on $\widetilde{\mathbf{x}}$ can produce greater attenuation. Section §3.2 examines the relationship between an irrelevant covariate $\widetilde{\mathbf{x}}$ and an instrumental variable that might be used with the Wald estimator to estimate a complier-average-causal-effect (CACE). The remainder of the paper concerns the construction and analysis of two control groups, one controlling for all of $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, the other controlling for $\overline{\mathbf{x}}$ and allowing or forcing separation on $\widetilde{\mathbf{x}}$. In particular, a form of simultaneous inference is proposed in which two sensitivity analyses are conducted for the two control groups, but the power loss for the controls matched for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ is small, so the second analysis adjusted for $\overline{\mathbf{x}}$ comes at little cost. The example uses data from NHANES 2005-2006 to examine the effects of smoking on homocysteine levels, in parallel with Bazzano et al. (2003) who used data from an earlier NHANES.

2 Review of notation and definitions

2.1 Treatment assignments and treatment effects

There are L individuals $\ell = 1, \ldots, L$ randomly sampled from an infinite population. Individual ℓ is described by $(r_{T\ell}, r_{C\ell}, Z_\ell, \overline{\mathbf{x}}_\ell, \widetilde{\mathbf{x}}_\ell, u_\ell)$, $\ell = 1, \ldots, L$, where $(\overline{\mathbf{x}}_\ell, \widetilde{\mathbf{x}}_\ell)$ are observed covariates, u_ℓ is an unobserved covariate, and individual ℓ exhibits response $r_{T\ell}$ if assigned to treatment, denoted $Z_\ell = 1$, or response $r_{C\ell}$ if assigned to control, denoted $Z_\ell = 0$, so the

observed response from individual ℓ is $R_\ell = Z_\ell\, r_{T\ell} + (1 - Z_\ell)\; r_{C\ell}$, and the effect $r_{T\ell} - r_{C\ell}$ caused by the treatment is not observed for any individual ℓ; see Neyman (1923), Welch (1937) and Rubin (1974). Fisher's (1935) sharp null hypothesis of no treatment effect H_0 asserts that $r_{T\ell} = r_{C\ell}$ for all ℓ. When referring to probability distributions in the population, the subscript ℓ is omitted. Following Dawid (1979), conditional independence of A and B given C is written $A \perp\!\!\!\perp B \,\big|\, C$.

When does it suffice to adjust for covariates $\mathbf{v}$ in causal inference? When may a portion of $\mathbf{v}$ safely be omitted from adjustments? We recall two definitions from the literature.

Definition 1 *(Rosenbaum and Rubin 1983). Treatment assignment Z is said to be strongly ignorable given covariates $\mathbf{v}$ if*

$$(r_T,\, r_C) \;\perp\!\!\!\perp Z \,\big|\, \mathbf{v}, \quad \textit{and } 0 < \Pr\left(Z = 1 \,|\, \mathbf{v}\right) < 1, \textit{ for all } \mathbf{v}. \tag{1}$$

For brevity and without further mention, the word ignorable is used in place of the term "strongly ignorable." If treatment assignment is ignorable given covariate $\mathbf{v}$, and if $\mathbf{v}$ were observed, then one can estimate causal effects such as $\mathrm{E}\left(r_T - r_C\right)$ or $\mathrm{E}\left(r_T - r_C|\mathbf{v}\right)$ or the average effect of the treatment on the treated, namely $\mathrm{E}\left(r_T - r_C|Z = 1\right)$, by adjusting for $\mathbf{v}$, for instance by matching or stratification; see Rosenbaum and Rubin (1983).

Definition 2 *(Heller, Rosenbaum and Small 2010). Covariates $\mathbf{v}_2$ in $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$ are said to be innocuous given $\mathbf{v}_1$ if*

$$(r_T,\, r_C) \;\perp\!\!\!\perp (Z,\, \mathbf{v}_2) \,\big|\, \mathbf{v}_1. \tag{2}$$

It is straightforward to show that if treatment assignment Z is ignorable given $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$ and if $\mathbf{v}_2$ is innocuous, then treatment assignment is also ignorable given $\mathbf{v}_1$

alone. If $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$ were a measured covariate, if treatment assignment Z were ignorable given $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$, and if $\mathbf{v}_2$ were innocuous given $\mathbf{v}_1$, then causal parameters, such as $\mathrm{E}(r_T - r_C)$, could be consistently estimated adjusting for $\mathbf{v}_1$, ignoring $\mathbf{v}_2$.

If (1) and (2) both hold, then causal inference need not include adjustments for $\mathbf{v}_2$. Is there a benefit — not merely absence of harm — from not adjusting for $\mathbf{v}_2$? Claims of benefit in the literature refer to a situation with an unobserved covariate u that cannot be controlled by adjusting for observed covariates, whether $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\overline{\mathbf{x}}$. If treatment assignment were ignorable given $\mathbf{v} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u)$ but not given $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\overline{\mathbf{x}}$, then causal effects could not be estimated by matching for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\overline{\mathbf{x}}$ because u is not controlled. In this case, ask: Is it advantageous to ignore $\widetilde{\mathbf{x}}$ and adjust for $\overline{\mathbf{x}}$ alone? Informal discussions (e.g., Brooks and Ohsfeldt 2013; Sanni Ali et al. 2014) debate the possibility that if an innocuous $\widetilde{\mathbf{x}}$ is left unmatched then it decreases the role that u plays in determining treatment assignment, thereby reducing the bias created by our inability to adjust for an unmeasured covariate u. Is this true in any formal sense?

2.2 Quantifying the impact of an unobserved covariate on treatment assignment

If $\mathbf{x}$ is some observed covariate, perhaps $\mathbf{x} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\mathbf{x} = \overline{\mathbf{x}}$, then one model for sensitivity to unmeasured bias from u is expressed in terms of the potential influence of u on the odds $\Pr(Z = 1 \,|\, \mathbf{x},\, u) \,/\, \{1 - \Pr(Z = 1 \,|\, \mathbf{x},\, u)\}$ of treatment; see Rosenbaum (1987a, 2002, §4; 2007). This model quantifies bias in treatment assignment in terms of how the propensity score might be different if it took account of the unobserved u in addition to the observed $\mathbf{x}$. Consider two subjects with treatment assignments Z and Z' and unobserved covariates u and u' but the same value of the observed covariate, $\mathbf{x} = \mathbf{x}'$, so these two subjects might be matched when matching for $\mathbf{x}$. Then the odds ratio (for Z given $\mathbf{x}$ and u) or density ratio (for u given $\mathbf{x}$ and Z) linking treatment Z and the unobserved covariate u for these

two subjects is:

$$\omega\left(\mathbf{x}, u, u'\right) = \frac{\Pr\left(Z = 1 \,|\, \mathbf{x},\, u\right) \Pr\left(Z' = 0 \,|\, \mathbf{x},\, u'\right)}{\Pr\left(Z = 0 \,|\, \mathbf{x},\, u\right) \Pr\left(Z' = 1 \,|\, \mathbf{x},\, u'\right)} = \frac{\Pr\left(u \,|\, \mathbf{x},\, Z = 1\right) \Pr\left(u' \,|\, \mathbf{x},\, Z' = 0\right)}{\Pr\left(u \,|\, \mathbf{x},\, Z = 0\right) \Pr\left(u' \,|\, \mathbf{x},\, Z' = 1\right)}, \tag{3}$$

where the second equality follows from Bayes theorem. The sensitivity model says that the impact of failing to control u is at most $\Gamma \geq 1$ in the sense that

$$\frac{1}{\Gamma} \leq \omega\left(\mathbf{x}, u, u'\right) \leq \Gamma \text{ for all } \mathbf{x},\, u,\, u'; \tag{4}$$

that is, two subjects with the same $\mathbf{x}$ may differ in their odds of treatment by at most a factor of Γ because they differ in terms of u. Because $\omega\left(\mathbf{x}, u, u'\right) = 1/\omega\left(\mathbf{x}, u', u\right)$, equation (4) is actually redundant, and it is equivalent to write

$$\omega\left(\mathbf{x}, u, u'\right) \leq \Gamma \text{ for all } \mathbf{x},\, u,\, u'. \tag{5}$$

Typically, one would match a treated subject to a control with the same $\mathbf{x}$, so $Z + Z' = 1$, but they might differ in terms of $u \neq u'$. Conditionally given $Z + Z' = 1$, the probability of $(Z, Z') = (1, 0)$ is

$$\frac{\Pr\left(Z = 1 \,|\, \mathbf{x},\, u\right) \Pr\left(Z' = 0 \,|\, \mathbf{x},\, u'\right)}{\Pr\left(Z = 1 \,|\, \mathbf{x},\, u\right) \Pr\left(Z' = 0 \,|\, \mathbf{x},\, u'\right) + \Pr\left(Z = 0 \,|\, \mathbf{x},\, u\right) \Pr\left(Z' = 1 \,|\, \mathbf{x},\, u'\right)} = \frac{\omega\left(\mathbf{x}, u, u'\right)}{\omega\left(\mathbf{x}, u, u'\right) + 1},$$

so that (4) or (5) implies $\varrho\left(\mathbf{x}, u, u'\right) = \Pr\left(Z = 1 \,|\, \mathbf{x},\, u,\, u',\, Z + Z' = 1\right)$ is bounded by

$$\frac{1}{1 + \Gamma} \leq \varrho\left(\mathbf{x}, u, u'\right) \leq \frac{\Gamma}{1 + \Gamma}, \text{ for all } \mathbf{x}, u, u'. \tag{6}$$

The one parameter Γ may be interpreted or amplified into an equivalent formulation in terms of two parameters, Λ and Δ, where Λ controls the relationship between treatment

assignment Z and u, Δ controls the relationship between response under control r_C and u, and one sensitivity analysis at Γ is exactly equivalent to an infinite curve of sensitivity analyses with $\Gamma = (\Lambda\Delta + 1) / (\Lambda + \Delta)$; see Rosenbaum and Silber (2009) for a precise statement using the semiparametric model introduced by Wolfe (1974). For instance, as $1.25 = (2 \times 2 + 1) / (2 + 2)$, it follows that $\Gamma = 1.25$ is equivalent to an unobserved covariate that doubles the odds of treatment ($\Lambda = 2$) and doubles the odds of a positive treated-minus-control response difference ($\Delta = 2$). In other words, one may calculate and report a one-dimensional sensitivity analysis in terms of Γ but have available the interpretations of a two-dimensional sensitivity analysis in terms of (Λ, Δ).

3 When does ignoring an observed covariate attenuate the association between treatment assignment and an unobserved covariate?

3.1 Prods to receive treatment

To prod is to "goad, stimulate [or] prompt," according to the *Oxford English Dictionary.*

Definition 3 *The observed covariates* $\widetilde{\mathbf{x}}$ *are a prod to receive treatment given* $(\overline{\mathbf{x}}, u)$ *if*

$$\widetilde{\mathbf{x}} \underline{\,|\,|\,} u \,\Big|\, \overline{\mathbf{x}}, \quad \textit{and } \operatorname{var}\{\Pr(Z = 1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u) \,|\, \overline{\mathbf{x}}, u\} > 0, \textit{ for all } (\overline{\mathbf{x}}, u). \tag{7}$$

In (7), the condition $\widetilde{\mathbf{x}} \underline{\,|\,|\,} u \,\Big|\, \overline{\mathbf{x}}$ says that, given $\overline{\mathbf{x}}$, there is no information in $\widetilde{\mathbf{x}}$ about u. In other words, trying to remove some bias from the unobserved u by adjusting for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, rather than adjusting for $\overline{\mathbf{x}}$ alone, is not going to work, because $\widetilde{\mathbf{x}}$ is unrelated to u. The requirement in (7) that $\Pr(Z = 1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u)$ varies with $\widetilde{\mathbf{x}}$ for fixed $(\overline{\mathbf{x}}, u)$ says that, although $\widetilde{\mathbf{x}}$ is not informative about u, nonetheless $\widetilde{\mathbf{x}}$ does vary with treatment assignment.

Proposition 4 says that not matching for a prod $\widetilde{\mathbf{x}}$ strictly attenuates the relationship between treatment assignment Z and the unobserved covariate u, or in the notation of §2.2

that $\omega(\overline{\mathbf{x}},\, u,\, u')$ is strictly closer to 1 than is $\omega\{(\overline{\mathbf{x}},\, \widetilde{\mathbf{x}}),\, u, u'\}$.

Proposition 4 *Let $\widetilde{\mathbf{x}}$ be a prod to receive treatment given $(\overline{\mathbf{x}}, u)$. For any fixed $\overline{\mathbf{x}}$, u, u', if*

$$\frac{1}{\Gamma} \leq \omega\left\{(\overline{\mathbf{x}},\, \widetilde{\mathbf{x}}),\, u, u'\right\} \leq \Gamma \text{ for all } \widetilde{\mathbf{x}} \text{ with } \Gamma > 1, \tag{8}$$

then there exists an Υ with $1 \leq \Upsilon < \Gamma$ such that

$$\frac{1}{\Upsilon} \leq \omega\left(\overline{\mathbf{x}},\, u,\, u'\right) \leq \Upsilon. \tag{9}$$

Proof. Following Freedman (2008, §9), define $f : (0,1) \to (0,\infty)$ by $f(p) = p/(1-p)$, so that $f(\cdot)$ is strictly increasing and $f^{-1}(v) = v/(1+v)$, and write $h(p) = f^{-1}\{\Gamma f(p)\}$. Freedman shows that $h(\cdot)$ is strictly concave on its domain, the open interval $(0,1)$. Now the second inequality in (8) implies $f\{\Pr(Z=1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u)\} \leq \Gamma f\{\Pr(Z=1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u')\}$ or equivalently that $\Pr(Z=1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u) \leq h\{\Pr(Z=1 \,|\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u')\}$. Using this and Jensen's inequality (e.g., Lange 2003, Proposition 3.5.1, page 61) for a strictly concave function yields

$$\begin{aligned}
\Pr(Z=1 \,|\, \overline{\mathbf{x}}, u) &= \int \Pr(Z=1 \,|\, \overline{\mathbf{x}},\, \widetilde{\mathbf{x}},\, u)\, \Pr(\widetilde{\mathbf{x}}|\overline{\mathbf{x}})\, d\widetilde{\mathbf{x}} \\
&\leq \int h\left\{\Pr\left(Z=1 \,|\, \overline{\mathbf{x}},\, \widetilde{\mathbf{x}},\, u'\right)\right\} \Pr(\widetilde{\mathbf{x}}|\overline{\mathbf{x}})\, d\widetilde{\mathbf{x}} \\
&< h\left\{\int \Pr\left(Z=1 \,|\, \overline{\mathbf{x}},\, \widetilde{\mathbf{x}},\, u'\right) \Pr(\widetilde{\mathbf{x}}|\overline{\mathbf{x}})\, d\widetilde{\mathbf{x}}\right\} \\
&= h\left\{\Pr\left(Z=1 \,|\, \overline{\mathbf{x}},\, u'\right)\right\}.
\end{aligned} \tag{10}$$

Applying the increasing function $f(\cdot)$ to the first and last term in (10) yields $f\{\Pr(Z=1 \,|\, \overline{\mathbf{x}}, u)\} < \Gamma f\{\Pr(Z=1 \,|\, \overline{\mathbf{x}},\, u')\}$ or equivalently $\omega(\overline{\mathbf{x}},\, u,\, u') < \Gamma$. Using instead the first inequality in (8) and $\omega\{(\overline{\mathbf{x}},\, \widetilde{\mathbf{x}}),\, u',\, u\} = 1/\omega\{(\overline{\mathbf{x}},\, \widetilde{\mathbf{x}}),\, u,\, u'\} \leq \Gamma$, the same argument shows

$\omega(\overline{\mathbf{x}}, u', u) < \Gamma$, and hence that $\omega(\overline{\mathbf{x}}, u, u') = 1/\omega(\overline{\mathbf{x}}, u', u) > 1/\Gamma$. To complete the proof, define $\Upsilon = \max\{\omega(\overline{\mathbf{x}}, u, u'), 1/\omega(\overline{\mathbf{x}}, u, u')\}$. ■

A few technical comments about Proposition 4 follow. First, in the definition of a prod, the requirement that $\text{var}\{\Pr(Z = 1 \mid \overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u) \mid \overline{\mathbf{x}}, u\} > 0$ in (7) is used to obtain the strict inequality in (10) by way of Jensen's inequality (e.g., Lange 2003, Proposition 3.5.1, page 61). Proposition 4 says there is strict attenuation, $\Gamma > \Upsilon$, for each $\overline{\mathbf{x}}$, u, u'; however, the degree of attenuation Υ in (9) generally depends upon $\overline{\mathbf{x}}$, u, u'. As a consequence, if the sensitivity model (4) were true with $\mathbf{x} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, then (8) would hold uniformly in $\overline{\mathbf{x}}$, $\widetilde{\mathbf{x}}$, u, u', but this would not imply that there exists one $\Upsilon < \Gamma$ such that (9) holds uniformly in $\overline{\mathbf{x}}$, u, u'. That is, Proposition 4 shows there is strict attentuation at each $\overline{\mathbf{x}}$, u, u', not that there is uniformly strict attenuation. It is clear that if one focused on the subpopulation with $\widetilde{\mathbf{x}} \in \mathcal{C}$ for some subset $\mathcal{C}$, then essentially the same proof shows there is attenuation in every subpopulation defined by $\widetilde{\mathbf{x}}$.

Proposition 4 is of no use on its own. However, if treatment assignment were ignorable given $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u)$, if $\widetilde{\mathbf{x}}$ were innocuous given $(\overline{\mathbf{x}}, u)$ and if $\widetilde{\mathbf{x}}$ were a prod to receive treatment given $(\overline{\mathbf{x}}, u)$, then: (i) it is suffices to focus attention on $(\overline{\mathbf{x}}, u)$ ignoring $\widetilde{\mathbf{x}}$, because adjustments for $(\overline{\mathbf{x}}, u)$ would permit estimation of causal effects, and (ii) it is also advantageous to focus attention on $(\overline{\mathbf{x}}, u)$ ignoring $\widetilde{\mathbf{x}}$, because the association between treatment assignment Z and u has been attenuated.

The heavy assumptions required to use Proposition 4 are consequential. Failing to adjust for $\widetilde{\mathbf{x}}$ could increase the bias for either or both of two reasons: (i) if treatment assignment were ignorable given $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u)$ but not given $(\overline{\mathbf{x}}, u)$, then adjusting for $\widetilde{\mathbf{x}}$ may reduce bias from $\widetilde{\mathbf{x}}$, (ii) even if $\widetilde{\mathbf{x}}$ itself seems to have no direct relevance, adjusting for $\widetilde{\mathbf{x}}$ might possibly reduce bias from u to the extent that $\widetilde{\mathbf{x}}$ and u are associated and the left side of (7) fails to hold.

Because Proposition 4 is of no use on its own, its actual usefulness is a matter of speculation. The additional assumptions that would make Proposition 4 useful are stringent assumptions about an unobserved covariate, and any investigator who makes these assumptions can expect an argument from skeptics. Rather than argue for or against the additional assumptions that would make Proposition 4 useful, we suggest conducting two analyses, one with and the other without these assumptions. A simple version of this has two control groups, one matched to treated subjects for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, the other matched for $\overline{\mathbf{x}}$ alone. Heller et al. (2010) observe that if treatment assignment were ignorable given $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ and if $\widetilde{\mathbf{x}}$ were innocuous given $\overline{\mathbf{x}}$, then these two comparisons of treated subjects to these two matched control groups would estimate the same parameter, the average effect of the treatment on the treated, so contrasting these two estimates provides a test of these two assumptions. In contrast, Proposition 5 in §6 frames the discussion of these two control groups when they may both be affected by bias from an unmeasured covariate u.

3.2 Is a prod an instrument?

So far, §3 has considered the possibility of comparing outcomes R in treated, $Z = 1$, and control, $Z = 0$, groups without adjustment for a covariate $\widetilde{\mathbf{x}}$ that meets certain additional, fairly speculative, conditions required of a prod. As noted in §1.1, this possibility has been discussed in several recent articles concerned with health outcomes research, including Brooks and Ohsfeldt (2013), Sanni et al. (2014), Walker (2013) and Zubizarreta et al. (2012). The method we propose in §5 takes the analysis adjusting for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ as the primary analysis, then adds at negligible cost in power a secondary analysis adjusting for $\overline{\mathbf{x}}$ but not for $\widetilde{\mathbf{x}}$, while controlling the familywise error rate in these two analyses, and making use of controls who might otherwise have been discarded. Could one, instead, view $\widetilde{\mathbf{x}}$ as an instrument or instrumental variable? Viewing $\widetilde{\mathbf{x}}$ as an instrument might suggest a

different analysis, say the Wald estimator or two-stage least squares, aimed at estimating the so-called "complier-average causal effect" or CACE.

By definition in the Neyman-Rubin framework, a covariate is a variable whose value is determined prior to treatment assignment Z and hence unaffected by which treatment an individual ultimately receives; that is, a covariate has single version that is the same whether or not $Z = 1$ or $Z = 0$, like $\overline{\mathbf{x}}$ or $\widetilde{\mathbf{x}}$ and unlike R or Z. In this framework, an instrument (recorded in an instrumental variable) is a very special kind of treatment that encourages an experimental subject to take a second treatment over which the experimentor lacks direct control, but the encouragement-treatment affects outcomes only to the extent that it alters acceptance of the second treatment; see Angrist, Imbens and Rubin (1996), Hirano et al. (2000) and Holland (1988). The CACE is the average effect of the second treatment on subjects who would respond to the encouragement treatment by changing their adoption of the second treatment, and Angrist et al. (1996) show that the CACE is the estimand of the Wald estimator. For instance, the Vietnam War draft lottery randomly selected people for the draft, a treatment that "encouraged" some people to serve in the military, though many men served without being drafted and others found ways to dodge the draft; see Angrist et al. (1996). For the draft lottery, the CACE is the average effect of military service on the subset of men who would serve in the military only if drafted.

A substantial literature consistent with the Neyman-Rubin framework cautions against adjusting for certain variables that, unlike $\widetilde{\mathbf{x}}$, are not covariates. In particular, Rosenbaum (1984, 2015c) cautions against adjusting for other outcomes of treatment, noting that such an adjustment can create a bias that would otherwise be absent. Several authors wisely advise against adjusting for instruments, such as the draft lottery used as an instrument for military service; see, for instance, Wooldridge (2006), Myers et al. (2011), Pearl (2010, 2011), and Bhattacharya and Vogt (2012).

In §1.2 and §6.2, $\widetilde{\mathbf{x}}$ describes income and education. In the context of NHANES, income and education are plausible covariates for smoking. In particular, we have a clear idea about what it means to be poor and uneducated, and we have no difficulty imagining a person of any fixed income or education choosing to smoke or not smoke. If treatment assignment were ignorable given $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ and if $\widetilde{\mathbf{x}}$ were innocuous given $\overline{\mathbf{x}}$, then the two matched comparisons of the treated group to each of the two control groups would estimate the same parameter, namely the average effect of the treatment on the treated. Although smoking is, in 2015, relatively uncommon among individuals with relatively high income and education, it would be quite a stretch to regard income and education as "treatments" that discourage smoking. For income and education to be instruments, the estimand in instrumental variables estimation, the CACE, would then be the average effect of smoking on people who would change their smoking behavior in response to a substantial change in income and education, a nebulous estimand at best. Within the view of instruments proposed by Angrist et al. (1996), it is not easy to think of income and education as instruments, so within that view, a prod — a type of covariate — is not an instrument — a type of treatment. An older view of instruments defines them in a context-free manner purely in terms of conditional independence or moment conditions. Within this older view, instruments of the type studied by Angrist et al. (1996) and prods might be viewed as two nonoverlapping subsets. A general principle is that an estimand should be clear and intelligible before an investigator sets out to estimate it. Our sense is that the CACE fails that principle for income and education in the NHANES example in §1.2 and §6.2. We do not regard the Wald estimator or two-stage least squares as options in this example.

4 The magnitude of the attenuation: direct calculation under a simple model

Proposition 4 says that not adjusting for a prod $\widetilde{\mathbf{x}}$ attenuates the bias in (8) because the inequality in (9) is strict. How large is this attenuation? For fixed (u, u'), how much closer to 1 is $\omega\left(\overline{\mathbf{x}}, u, u'\right)$ than $\omega\{(\overline{\mathbf{x}}, \widetilde{\mathbf{x}}), u, u'\}$? As in §3, Bayes theorem permits us to think about the answer in terms of the imbalance in u in treated, $Z = 1$, and control, $Z = 0$, groups. Table 1 provides an answer to how large the attenuation is in a simple case in which there is no $\overline{\mathbf{x}}$, $\widetilde{x}$ is a scalar prod with $\widetilde{x} \sim N\left(0, \sigma^2\right)$ for $\sigma = 1/2$ or 1, and treatment assignment probabilities follow a logit model, $\text{logit}\left\{\Pr(Z = 1|\widetilde{x}, u)\right\} = \alpha + \widetilde{x} + \gamma u$, so that for $u = 0$ and $u' = 1$, condition (8) holds with equality as $\Gamma = \exp\left(\gamma\right) = \omega(\widetilde{\mathbf{x}}, u, u')$. Under this model, for fixed u and u', the odds of treatment are $\exp\left(2\sigma\right)$ times greater when $\widetilde{x}$ is one standard deviation above its mean than when it is one standard deviation below its mean, or $\exp\left(2\sigma\right) = 2.71$ for $\sigma = 1/2$ and $\exp\left(2\sigma\right) = 7.39$ for $\sigma = 1$, so for both values of σ the prod $\widetilde{x}$ substantially alters the treatment assignment probabilities. Table 1 displays the attenuated $\omega\left(\overline{\mathbf{x}}, u, u'\right)$ with $u = 0$ and $u' = 1$, obtained by evaluating (10) by numerical integration. For example, for $\alpha = -1$, for $\sigma = 1/2$, a moderate bias of $\Gamma = \exp\left(\gamma\right) = 1.5$ attenuates to 1.47, whereas for $\sigma = 1$ a large bias of $\Gamma = 5$ attenuates to 3.81. The impression from the simple example in Table 1 is that: (i) a prod $\widetilde{x}$ must substantially affect the treatment assignment probabilities to produce substantial attenuation, and (ii) even when there is substantial attenuation, the bias that remains is far from small.

5 Two control groups: controlling for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\overline{\mathbf{x}}$

5.1 Using two control groups

Proposition 4 reaches an attractive conclusion — a reduction in unmeasured biases in (9) — on the basis of heroic assumptions in (2) with $\mathbf{v}_1 = (\overline{\mathbf{x}}, u)$ and $\mathbf{v}_2 = \widetilde{\mathbf{x}}$ and (7) — the

strong influence but total irrelevance of the prod $\widetilde{\mathbf{x}}$. In many applications, investigators will be understandably reluctant to rely on such strong assumptions to achieve the modest level of attenuation seen in Table 1. There is, however, a practical way to use Proposition 4 to see a little more in observational data without committing to the strong assumptions in Proposition 4, that is, a way to have it both ways.

The possibility of using two control groups subject to different biases is much discussed in the literature on observational studies; see, for instance, Campbell (1969), Rosenbaum (1987b, 2015a), Meyer (1995), Shadish et al. (2002), Stuart and Rubin (2008), West et al. (2008), Heller et al. (2010) and Lu et al. (2011). Typically, these two control groups are found rather than constructed; that is, the groups existed as groups before the investigation began.

With varied motivations, several recent studies have used the computer to construct two control groups, one matched for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, the other match only for $\overline{\mathbf{x}}$; see Daniel et al. (2008), Heller et al. (2010), and Silber et al. (2012, 2013). These two control groups may be nonoverlapping, perhaps constructed using the tapered matching algorithm of Daniel et al. (2008), or they may share controls. Matched control groups that share controls may be compared to each other using a device known as the exterior match; see Rosenbaum and Silber (2013). Matching ensures that $\overline{\mathbf{x}}$ has the same distribution in the treated group and both control groups, a helpful fact if the magnitude of the treatment effect varies with $\overline{\mathbf{x}}$; however, at the risk of losing this desirable property, one could alternatively adjust for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ or $\overline{\mathbf{x}}$ using some form of covariance adjustment.

Suppose that two control groups are formed, perhaps overlapping, perhaps not, one matched for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, the other just for $\widetilde{\mathbf{x}}$. In the context of Proposition 4, if there are benefits to not matching for $\widetilde{\mathbf{x}}$, then we see such an analysis, but if the strong assumptions in Definitions 2 and 3 are false or doubtful, then we see an analysis that does not depend

upon these assumptions. Moreover, we are able to compare these two analyses.

Strict use of Proposition 4 would perform two unrelated and therefore typically overlapping matches, one for $\overline{\mathbf{x}}$ alone, the other for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$. In this strict use, each match does not alter the other match: the match for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ does not alter the distribution of $\widetilde{\mathbf{x}}$ in the match for $\overline{\mathbf{x}}$ alone, so Proposition 4 speaks directly to the consequences of leaving $\widetilde{\mathbf{x}}$ unmatched. An alternative approach inspired by Proposition 4 but only informally linked to it would force the two matches to use different controls, thereby typically using more controls, with better matches for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ going to the match that controls $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ and worse matches for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ going to the match for $\overline{\mathbf{x}}$ alone, as happens in tapered matching (Daniel et al. 2008). Because this alternative approach forces the two matched control groups to be nonoverlapping, the two control groups compete for controls, so there is some distortion of the distribution of the unmatched prod $\widetilde{\mathbf{x}}$. Another alternative also inspired by Proposition 4 but even more informally linked to it would force the two matches to use different controls and additionally force the controls matched for $\overline{\mathbf{x}}$ alone to differ from the treated group in terms of $\widetilde{\mathbf{x}}$. The goal in this second alternative is to achieve greater attenuation of bias from u by picking controls precisely because the prod $\widetilde{\mathbf{x}}$ pushed them into the control group; see §5.2.

As noted previously, the attenuation result in Proposition 4 holds whether or not $\widetilde{\mathbf{x}}$ is innocuous, but attenuation is useful in an observational study only if $\widetilde{\mathbf{x}}$ is innocuous given $(\overline{\mathbf{x}}, u)$ in the sense of Definition 2, for otherwise the attenuation of bias from u may be more than offset by bias from failure to control $\widetilde{\mathbf{x}}$. In the current paragraph, assume treatment assignment is ignorable given $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}}', u)$. Were it true that $\widetilde{\mathbf{x}}$ is innocuous given $(\overline{\mathbf{x}}, u)$, then

$$\Pr\left(r_T,\, r_C \mid \overline{\mathbf{x}},\, u\right) = \Pr\left(r_T,\, r_C \mid \overline{\mathbf{x}}, \widetilde{\mathbf{x}},\, u\right)$$

$$= \Pr\left(r_T,\, r_C \mid Z = z,\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}},\, u\right) = \Pr\left(r_T,\, r_C \mid Z = z,\, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}',\, u\right) \text{ for all } \widetilde{\mathbf{x}},\, \widetilde{\mathbf{x}}'. \qquad (11)$$

We observe treated response distributions from treated subjects, say $\Pr(R \mid Z=1, \overline{\mathbf{x}}) = \Pr(r_T \mid Z=1, \overline{\mathbf{x}})$ or $\Pr(R \mid Z=1, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}) = \Pr(r_T \mid Z=1, \overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, and control response distributions from control subjects, say $\Pr(R \mid Z=0, \overline{\mathbf{x}}) = \Pr(r_C \mid Z=0, \overline{\mathbf{x}})$, or $\Pr(R \mid Z=0, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}) = \Pr(r_C \mid Z=0, \overline{\mathbf{x}}, \widetilde{\mathbf{x}})$. Treated response distributions may differ from control response distributions either because of a treatment effect or because of a bias. In contrast, if we compare two control response distributions, say $\Pr(R \mid Z=0, \overline{\mathbf{x}}) = \Pr(r_C \mid Z=0, \overline{\mathbf{x}})$ versus $\Pr(R \mid Z=0, \overline{\mathbf{x}}, \widetilde{\mathbf{x}}) = \Pr(r_C \mid Z=0, \overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ for controls matched to the same treated subject, then these differ when (11) holds only because of bias from the failure to control the unobserved covariate u. This is true of all three matches in the previous paragraph when (11) holds, and forcing $\widetilde{\mathbf{x}}$ to differ in the third match may provide a greater opportunity to check whether or not $\Pr(r_C \mid Z=0, \overline{\mathbf{x}})$ and $\Pr(r_C \mid Z=0, \overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ differ. If $\Pr(r_C \mid Z=0, \overline{\mathbf{x}})$ and $\Pr(r_C \mid Z=0, \overline{\mathbf{x}}, \widetilde{\mathbf{x}})$ do differ, then this can indicate bias from u or it can indicate that $\widetilde{\mathbf{x}}$ is not innocuous given $(\overline{\mathbf{x}}, u)$ or both (so (11) does not hold), but it surely indicates that at least one control group cannot be trusted.

5.2 Attenuation with forced separation

The magnitude of attenuation is now considered under a simple method for forcing separation on a prod, so treated and control groups are further apart on the prod than they would be if the prod were left unmatched. In brief summary, forcing separation increases attentuation when the initial bias is large, but the attenuated bias that remains is still large, even when treated and control groups are widely separated on the prod, as in the example in Figure 1. The logit-model formulation used here is similar to §4 except treated units are matched to controls whose prod $\tilde{x}$ is less than or equal to $c\sigma$ for some cutoff c. The smaller c is for a given σ, the greater the separation on the prod. By analogy with (3), we use Bayes theorem and measure attenuation by comparing odds ratios of $u=1$ versus

$u = 0$ in treated, $Z = 1$, and control, $Z = 0$, groups. Here we consider $\alpha = -1$ so that there are more control units than treated units, which is needed for matching to ensure separation on a prod; the results (not shown) were similar for $\alpha = 0$ and $\alpha = 1$. Table 2 shows the attenuation for different values c. The top half of Table 2 can be compared to the first line of Table 1 and the second half of Table 2 can be compared to the fourth line of Table 1. In Table 2 matching to ensure separation on a prod creates greater attenuation than leaving the prod unmatched. The differences are fairly small for moderate Γ: for $\Gamma = 1.5$ and $\sigma = 1/2$, even for $c = -1$, ensuring separation on the prod only increased the attenuation from 1.47 to 1.39. The differences are more substantial for larger Γ, e.g., for $\Gamma = 10$ and $\sigma = 1/2$, for $c = -1$, ensuring separation on the prod increases the attenuation from 8.88 to 6.42. Table 3 shows how much separation on the prod is created by matching to ensure separation on the prod for different values of c. The table reports the standardized difference on the prod when matching treated units to control units with prod $\leq c\sigma$. For $\sigma = 1/2$, the standardized difference ranges from about $1.8 - 1.9$ (depending on Γ) with $c = -1$ to $0.6 - 0.7$ with $c = 1$.

5.3 An algorithm for matching to ensure separation on a prod

We now introduce an algorithm to create matches that exhibit balance on $\overline{\mathbf{x}}$ and force separation on $\widetilde{\mathbf{x}}$. The algorithm produced the match in Figure 1. This new algorithm slightly extends the balanced optimal matching technique of Pimentel et al. (2015); see Hansen and Klopfer (2006) and Stuart (2010) for other discussions of matching algorithms in observational studies. That approach used penalized network flows to select controls with a covariate distribution as similar as possible to the treated group for large numbers of nominal covariates and their interactions. The extension proposed here selects controls to be similar to treated subjects in some ways and as different as possible in others. The

original algorithm has a target distribution for the covariates in the control group, and the extension simply changes the target distribution. In the example, this means that controls should resemble the treated group in terms of biological quantities, age, gender, BMI, but should be as high as possible in terms of education and income. A precise description is given in the Appendix. To create separation on a prod $\widetilde{\mathbf{x}}$ while balancing $\overline{\mathbf{x}}$, we first define a new covariate

$$\eta(\widetilde{\mathbf{x}}_i) = \begin{cases} 1 & \text{if } \widetilde{\mathbf{x}}_i \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{X}$ is a set of desired values for the prod. The target distribution for controls has the same distribution of $\overline{\mathbf{x}}_i$ as the treated group and has $\eta(\widetilde{\mathbf{x}}_i) = 1$. Running the algorithm for this target group and with balance constraints on $\overline{\mathbf{x}}$ and $\eta(\widetilde{\mathbf{x}})$ selects a control group with a distribution of $\overline{\mathbf{x}}$ very similar to that in the treated population, but also ensures that as many of the controls as possible are chosen with $\widetilde{\mathbf{x}}$ values in the region $\mathcal{X}$, thereby creating separation on the prod.

6 Inference with and without a prod

6.1 Sensitivity analysis with two control groups controlling the familywise error rate

Figure 2 shows homocysteine levels in blood plasma for the $I = 512$ matched triples in Figure 1; see §1.2. The current section is concerned with the simultaneous analysis of prodded and unprodded match sets of the type displayed in Figure 2.

Define the null hypothesis H'_Γ to be the conjunction of (i) Fisher's hypothesis of no effect, H_0, (ii) treatment assignment Z is ignorable given $\mathbf{v} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u)$ and (iii) a bias in treatment assignment from u of at most $\Gamma \geq 1$ in pairs of individuals matched for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, so that (8) holds for all $\overline{\mathbf{x}}$, $\widetilde{\mathbf{x}}$, u, u'. Define the null hypothesis H^*_Υ to be the conjunction of (i) Fisher's hypothesis of no effect, H_0, (ii) treatment assignment Z is ignorable given

$\mathbf{v} = (\overline{\mathbf{x}}, \widetilde{\mathbf{x}}, u)$, (iv) $\widetilde{\mathbf{x}}$ is innocuous given $(\overline{\mathbf{x}}, u)$, (v) a bias in treatment assignment from u of at most $\Upsilon \geq 1$ in pairs of individuals matched for $\overline{\mathbf{x}}$, so that (9) holds for all $\overline{\mathbf{x}}$, u, u'. Obviously, rejecting H'_Γ or H^*_Υ leaves open whether Fisher's H_0 is false or whether the additional assumptions are false. Notably, H'_Γ and H^*_Υ share (i) and (ii) but H'_Γ adds (iii) while H^*_Υ omits (iii) and adds (iv) and (v), although all of assumptions (i)-(v) could be jointly true. The data used to test H'_Γ and H^*_Υ are dependent because the same treated subjects are used in both tests, as in Figure 2, and also if the control groups are allowed to overlap or share some controls, as is not true in Figure 2.

If H'_Γ or H^*_Υ were both true, then Proposition 4 would lead us to anticipate modest attenuation of unmeasured biases. That is, Proposition 4 leads us to be interested in testing pairs $(H'_\Gamma, H^*_\Upsilon)$ with Υ modestly smaller than Γ, perhaps $\Upsilon = \omega\Gamma$ for $\omega = 0.9$, or 10% smaller based on Table 2.

We propose to use a multiple testing procedure to conduct two sensitivity analyses, one for H'_Γ and one for H^*_Υ, correcting for multiple testing using the recycling method of Burman, Sonesson and Guilbaud (2009). The recycling procedure strongly controls the familywise error rate. Let $0 < \alpha' \leq \alpha < 1$ be two fixed numbers, conventionally $\alpha = 0.05$. Fix (Γ, Υ), say $(\Gamma, \Upsilon) = (\Gamma, \omega\Gamma)$, and compute the two upper bounds on P-values, say $p'_{\Gamma,\max}$ and $p^*_{\Upsilon,\max}$, from separate sensitivity analyses for H'_Γ and H^*_Υ, respectively. In the example, the method in Rosenbaum (2007) yields $p'_{\Gamma,\max}$ and $p^*_{\Upsilon,\max}$ using the `R` package `sensitivitymw`. The recycling steps are:

Recycling procedure:

1. Test H'_Γ: Reject H'_Γ at level α in the presence of a bias of at most Γ if $p'_{\Gamma,\max} \leq \alpha'$.

2. Test H^*_Υ: If H'_Γ was rejected in step 1, then reject H^*_Υ at level α in the presence of a bias of at most Υ if $p^*_{\Upsilon,\max} \leq \alpha$. Otherwise, if H'_Γ was not rejected in step 1, then

reject H^*_Υ at level α in the presence of a bias of at most Υ if $p^*_{\Upsilon,\max} \leq \alpha - \alpha'$.

3. Recycle to retest H'_Γ: If H'_Γ was not rejected in step 1 but H^*_Υ was rejected in step 2, then reject H'_Γ at level α in the presence of a bias of at most Γ if $p'_{\Gamma,\max} \leq \alpha$.

For a fixed (Γ, Υ) with $\alpha' = \alpha/2$, then this recycling procedure is easily seen to be equivalent to the standard version of Holm's (1979) procedure, and if $0 < \alpha' < \alpha$, then it is equivalent to Holm's (1979) weighted procedure with $w' = \alpha'/\alpha$ and $w^* = (\alpha - \alpha')/\alpha$. These equivalences are seen by considering the four possible outcomes of steps 1-3. As noted by Benjamini and Hochberg (1997, p. 411), the weighted Holm procedure is superior to another weighting scheme with two hypotheses, as here. Taking $\alpha' = \alpha$ is fixed sequence testing, so rejection of H^*_Υ can occur only if H'_Γ is rejected in step 1, and step 3 is redundant. So in our case with two hypotheses, the recycling procedure reduces to one of two other methods, but is attractive in unifying them. To reject both H'_Γ and H^*_Υ is to have $\max\left(p'_{\Gamma,\max}, p^*_{\Upsilon,\max}\right) \leq \alpha$ as for intersection-union testing (Berger 1982, Laska and Meiser 1989); however, intersection-union testing could reject when recycling does not if $\alpha' < \alpha$, and recycling could reject just one hypothesis, either H'_Γ and H^*_Υ, which intersection-union testing cannot.

Conventionally, $\alpha = 0.05$. How should α' be chosen? If an analysis that controlled $\overline{\mathbf{x}}$ but not $\widetilde{\mathbf{x}}$ would be implausible if it disagreed with an analysis that controls $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, then α' should be close to α, perhaps $\alpha' \in [0.8\alpha, \alpha]$. Arguably this is the case with $\widetilde{\mathbf{x}}$ recording income and education in the smoking example, so we take $\alpha' = 0.04 < \alpha = 0.05$, but taking $\alpha' = \alpha = 0.05$ would be reasonable also. In this way, little power is lost in the analysis that adjusts for $(\overline{\mathbf{x}}, \widetilde{\mathbf{x}})$, yet both analyses are considered with strong control for testing two null hypotheses.

The discussion above considered a single fixed (Γ, Υ). In fact, we consider not a fixed (Γ, Υ) but rather a sequence $(\Gamma, \Upsilon) = \{\Gamma_n, \max(1, \omega\Gamma_n)\}$, $n = 1, 2, \ldots$, with $\Gamma_1 = 1$ and

$\Gamma_n \to \infty$ as $n \to \infty$, where $\omega > 0$ is fixed. In practice, reasonable values of ω are $\omega = 0.9$, hoping for modest attenuation, or $\omega = 1$, preferring to handle the two control groups symmetrically. At step n, a total of $2n$ hypotheses have been tested using the recycling procedure.

Proposition 5 *For fixed $\omega > 0$, apply the recycling procedure to $(\Gamma_n, \Upsilon_n) = \{\Gamma_n, \max(1, \omega\Gamma_n)\}$ for $n = 1, 2, \ldots$. The chance of falsely rejecting at least one true hypothesis, H'_{Γ_n} or $H^*_{\Upsilon_n}$, $n = 1, 2, \ldots$, is at most α.*

Proof. Recall that $p'_{\Gamma,\max}$ is a valid P-value for testing H'_Γ alone and $p'_{\Gamma,\max}$ increases with Γ, whereas $p^*_{\Upsilon,\max}$ is a valid P-value for testing H^*_Υ alone and $p^*_{\Upsilon,\max}$ increases with Υ. Also, the recycling procedure controls the familywise error when testing both H'_Γ and H^*_Υ with any one fixed (Γ, Υ). Let $\overline{\Gamma} = \inf\{\Gamma_n : H'_{\Gamma_n} \text{ is true}\}$ and $\overline{\Upsilon} = \inf\{\Upsilon_n : H^*_{\Upsilon_n} \text{ is true}\}$, where $\overline{\Gamma} = \infty$ and $\overline{\Upsilon} = \infty$ are possible values. To avoid a separate discussion of the infinite cases, define $p'_{\infty,\max} = p^*_{\infty,\max} = 1$. By definition of the hypotheses earlier in this section, H'_{Γ_n} is true for all $\Gamma_n \geq \overline{\Gamma}$ and $H^*_{\Upsilon_n}$ is true for all $\Upsilon_n \geq \overline{\Upsilon}$. Hence, the smallest $p'_{\Gamma_n,\max}$ for a true H'_{Γ_n} is $p'_{\overline{\Gamma},\max}$ and the smallest $p^*_{\Upsilon_n,\max}$ for a true $H^*_{\Upsilon_n}$ is $p^*_{\overline{\Upsilon},\max}$. We consider cases. If $\overline{\Gamma} = \overline{\Upsilon} = \infty$, then there is nothing to prove, because no true hypothesis is tested. If $\overline{\Upsilon} = \omega\overline{\Gamma} < \infty$, then to reject any true hypothesis, one must have $\left(p'_{\overline{\Gamma},\max} \leq \alpha'\right) \vee \left(p^*_{\overline{\Upsilon},\max} \leq \alpha - \alpha'\right)$ and the chance of this is as most α. If $\overline{\Upsilon} < \omega\overline{\Gamma}$, then a false rejection for (Γ_n, Υ_n) with $\Gamma_n < \overline{\Gamma}$ and $\overline{\Upsilon} \leq \Upsilon_n < \omega\overline{\Gamma}$ requires rejection of the true $H^*_{\overline{\Upsilon}}$ with $p^*_{\overline{\Upsilon},\max} \leq \alpha$ which occurs with probability at most α, whereas false rejection for (Γ_n, Υ_n) with $\Gamma \geq \overline{\Gamma}$ and $\Upsilon \geq \omega\overline{\Gamma}$ requires $\left(p'_{\Gamma_n,\max} \leq \alpha'\right) \vee \left(p^*_{\Upsilon_n,\max} \leq \alpha - \alpha'\right)$, which implies $\left(p'_{\overline{\Gamma},\max} \leq \alpha'\right) \vee \left(p^*_{\overline{\Upsilon},\max} \leq \alpha - \alpha'\right)$ which has probability at most α. The case $\overline{\Upsilon} > \omega\overline{\Gamma}$ is analogous. ■

6.2 Example: Using wealthy, educated nonsmokers as a second control group

For the data in §1.2, Figure 2 compares homocysteine levels among smokers to two control groups, one (M) matched to controls for all measured covariates, the other (P) separated from the smokers on the prod $\widetilde{\mathbf{x}}$ of education and income; see, again, Figure 1 for the difference in education and income among these groups. The smokers in Figure 2 appear to have somewhat higher homocysteine levels than both control groups, whereas control groups M and P appear similar. We now conduct a sensitivity analysis using the procedure in §6.

With I treatment-control matched pairs, Maritz (1979) used the null randomization distribution of Huber's one-sample M-statistic $T = \sum_{i=1}^{I} \psi\left(Y_i/s\right)$ to test Fisher's null hypothesis of no effect, where Y_i is a treated-minus-control matched pair difference in responses R, s is the median $|Y_i|$, and $\psi\left(\cdot\right)$ is an odd function, $\psi\left(y\right) = -\psi\left(-y\right)$. Taking $\psi_{\text{t}}\left(y\right) = y$ makes T into a constant multiple of the sample mean, and then Maritz's method is equivalent to the randomization distribution of the mean, that is, the permutational t-test; see Pitman (1937) and Welch (1937). Huber's $\psi_{\text{hu}}\left(\cdot\right)$ has $\psi_{\text{hu}}\left(y\right) = \text{sign}\left(y\right) \cdot \min\left(\left|y\right|, \kappa\right)$ for some $\kappa > 0$, where $\text{sign}\left(y\right) = 1,\ 0,\ -1$ as $y > 0$, $y = 0$, $y < 0$, so $\psi_{\text{hu}}\left(\cdot\right)$ has the same influence function as a trimmed mean. A sensitivity analysis for T when used in observational studies was proposed in Rosenbaum (2007), its power and large sample properties in sensitivity analysis with various choices of $\psi\left(\cdot\right)$ were examined in Rosenbaum (2013), and the method was implemented in the `sensitivitymv` and `sensitivitymw` packages in `R`; see Rosenbaum (2015b). In particular, taking $\psi_{\text{in}}\left(y\right) = \left\{\kappa/\left(\kappa - \iota\right)\right\} \cdot \text{sign}\left(y\right) \cdot \max\left\{0, \min\left(\left|y\right|, \kappa\right) - \iota\right\}$ for some $\kappa > \iota \geq 0$ entails inner trimming and means $\psi_{\text{in}}\left(y\right)$ is zero for $|y| \in [0, \iota]$, is $\text{sign}\left(y\right) \cdot \kappa$ for $|y| \geq \kappa$, and rises linearly from 0 to κ on $[\iota, \kappa]$. For many distributions of Y_i, the M-statistic $T = \sum_{i=1}^{I} \psi_{\text{in}}\left(Y_i/s\right)$ reports greater insensitivity to unmeasured biases than does $\psi_{\text{hu}}\left(\cdot\right)$. Here, we set $\kappa = 2$ and $\iota = 1/2$; see Rosenbaum (2013, Table 3) and `method="p"`

in the `senmw` function of the `sensitivitymw` package in `R`.

Table 4 performs a sensitivity analysis with $(\Gamma, \Upsilon) = (\Gamma,\, 0.9 \times \Gamma)$ for an increasing sequence of values of Γ, as discussed in §6.1, reporting the upper bounds, $p^{'}_{\Gamma,\max}$ and $p^{*}_{\Upsilon,\max}$, on the marginal P-values testing H'_Γ and H^*_Υ, respectively. Table 4 does not control for testing two hypotheses. Using the method in §6.1 to control for testing two hypotheses and testing in a fixed sequence with $\alpha = \alpha' = 0.05$ leads to rejection of H'_Γ and H^*_Υ for $(\Gamma, \Upsilon) = (1.640,\, 1.476)$, and no rejections at $(\Gamma, \Upsilon) = (1.650,\, 1.485)$. Recycling with $\alpha = 0.05$ and $\alpha' = 0.04$ rejects H'_Γ and H^*_Υ for $(\Gamma, \Upsilon) = (1.640,\, 1.476)$, tests H'_Γ at the 0.05 level for $\Gamma = 1.650$ but fails to reject, and barely rejects H^*_Υ for $\Upsilon = 1.575$. To put this in context, $\Gamma = 5/3 = 1.667$ corresponds with an unobserved covariate that triples the odds of treatment and triples the odds of a positive pair difference in outcomes, while $\Gamma = 1.5$ corresponds with an unobserved covariate that doubles the odds of treatment and doubles the odds of positive pair difference in outcomes; see Rosenbaum and Silber (2009) and the `amplify` function in the `sensitivitymv` package in `R`.

If, as may be, it is important to adjust for socioeconomic factors $\widetilde{\mathbf{x}}$, then Table 4 does this, finding that the results are not sensitive to small unmeasured biases. If, as may be, socioeconomic factors introduce a biologically irrelevant source of variation in smoking behavior, so that comparing people differing in $\widetilde{\mathbf{x}}$ attenuates bias from unmeasured covariates u, then Table 4 does this also, finding again that the results are not sensitive to small unmeasured biases. Whether you compare people with the same or different education and income, smokers tend to have higher homocysteine levels than nonsmokers. These two analyses bracket the one analysis in Bazzano et al. (2003), where $\widetilde{\mathbf{x}}$ was neither controlled nor separated.

Arguably, Table 4 allows us to see more in an observational study than we would have seen with either comparison alone, yet it avoids committing us to one or another set of

assumptions about unmeasured covariates, assumptions that are easy enough to state but difficult if not impossible to justify. Moreover, in this example, the M-controls were tested at level 0.05, yet the familywise error rate for two tests was also controlled at $\alpha = 0.05$, so the addition of the P-controls came without cost. The simulation in §6.3 asks whether this pattern is expected in general.

6.3 Simulation: power of the recycling procedure in a sensitivity analysis

Ideally, a sensitivity analysis would reject the null hypothesis of no effect when there is no unmeasured bias and there is a treatment effect, and the power of a sensitivity analysis is the probability that this will happen; see Rosenbaum (2004, 2013). More precisely, the power of an α-level sensitivity analysis allowing for bias Γ is the probability that the upper bound on the P-value leads to rejection when computed with this Γ. The simulation contrasts testing in a fixed sequence, $\alpha = \alpha' = 0.05$, and recycling with $\alpha = 0.05$ and $\alpha' = 0.04$. The simulation also contrasts exploring the (Γ, Υ) sequence along $(\Gamma, \Upsilon) = (\Gamma, \Gamma)$ with equal sensitivity parameters and along $(\Gamma, \Upsilon) = \{\Gamma, \max(1, 0.9 \times \Gamma)\}$. The latter sequence makes sense if the investigator included the prodded controls anticipating moderate attenuation of unmeasured biases.

Table 5 simulates a simple situation in which all treated-minus-control pair differences in both control groups are Normal with expectation τ and variance 1. The correlation between the pair-differences in the two control groups is 1/2 because the same treated subject is matched to two different controls, as in §1.2. The effect size is either $\tau = 1/4$ or $\tau = 1/2$. Of course, the results are less sensitive with a larger effect, and Γ is adjusted accordingly, $\Gamma = 1.5$ for $\tau = 1/4$, $\Gamma = 2.8$ for $\tau = 1/2$. Columns a and b, labeled $\alpha' = 0.05$, refer to testing in a fixed sequence. Columns c and d, labeled $\alpha' = 0.04$, refer to recycling. The final column is for comparison only: column e gives the power if H^*_Υ were

tested at the $\alpha = 0.05$ level with no correction for testing two hypotheses. Although a part of fixed sequence testing, column a for $\alpha' = 0.05$ and $H_\Gamma^{'}$ analogously gives the power when testing $H_\Gamma^{'}$ without correction for multiple testing, because in fixed sequence testing the first hypothesis in the sequence is tested without correction. Table 5 also compares the power when using $\psi_{\text{hu}}(\cdot)$ and $\psi_{\text{in}}(\cdot)$ with $\kappa = 2$ and $\iota = 1/2$ and, as expected from Rosenbaum (2013), the power is greater with $\psi_{\text{in}}(\cdot)$. Each situation is replicated 50,000 times, so the standard error of an estimated power is at most $\sqrt{0.25/50000} = 0.0022$.

With fixed sequence testing, adding a second comparison does not reduce the power of the first comparison, but it affects the power of subsequent comparisons. For instance, in Table 5 with $(\Gamma, \Upsilon) = (1.5, 1.5)$, $\tau = 1/4$, $\psi_{\text{hu}}(\cdot)$, the power is 0.47 for $H_\Gamma^{'}$ alone, for $H_\Gamma^{'}$ as first in sequence, and for H_Υ^{*} alone in the last column, but H_Υ^{*} tested in fixed sequence after testing $H_\Gamma^{'}$ has power of only 0.30. In contrast, with $(\Gamma, \Upsilon) = (\Gamma, \Gamma)$, recycling with $\alpha' = 0.04$ slightly reduces the power for $H_\Gamma^{'}$ and somewhat increases the power for H_Υ^{*}.

There is some attraction to conducting the sensitivity analysis through a sequence of the form $(\Gamma, \Upsilon) = \{\Gamma, \max(1, 0.9 \times \Gamma)\}$, as was done in the example in Table 4. That sequence is interesting because the prod, if it actually works, is intended to attenuate bias, as in Proposition 4, so values of Υ somewhat below Γ are not without interest. At the same time, in the simulated situation, recycling of unused α is much more likely to occur when $\Upsilon = 0.9 \times \Gamma$, so the power loss is smaller. Specifically, when $\Upsilon = 0.9 \times \Gamma$ in Table 4, the power for $H_\Gamma^{'}$ is only slightly lower in column c than in column a, typically about 1% lower, whereas the power for H_Υ^{*} is much higher in column d than in column b. The combination of $\alpha' = 0.04$ and $\Upsilon = 0.9 \times \Gamma$ is, therefore, attractive: despite correction for performing two tests, the M-controls are tested at nearly the power of a single test, while the smaller value of $\Upsilon = 0.9 \times \Gamma$ for the P-controls means the second comparison also has high power.

7 Summary: Prefer additional analyses to additional assumptions

It has been argued in the literature that leaving a measured covariate $\widetilde{\mathbf{x}}$ uncontrolled, say unmatched, may attenuate biases from an unmeasured covariate u. Although this is formally true, the argument requires very strong, typically doubtful, assumptions about both observed and unobserved covariates, and even when those assumptions are true the magnitude of the attenuation is modest. We suggest that one should not conduct a single analysis that presumes these doubtful assumptions are true. Rather, we suggest building two control groups, with two analyses, one that controls for $\widetilde{\mathbf{x}}$ and one that leaves $\widetilde{\mathbf{x}}$ uncontrolled. Often, the second control group uses individuals who would otherwise be excluded from the analysis because they are so different from treated subjects in terms of $\widetilde{\mathbf{x}}$. A second control group entails a second hypothesis test, hence a correction for testing two hypotheses; however, by careful organization of the analyses, there is only a slight loss of power in the primary comparison controlling $\widetilde{\mathbf{x}}$, so the second control group is nearly without cost.

References

Bazzano, L. A., He, J., Muntner, P., Vupputuri, S., and Whelton, P. K. (2003), "Relationship between cigarette smoking and novel risk factors for cardiovascular disease in the United States," *Annals of Internal Medicine*, 138(11), 891-897.

Benjamini, Y. and Hochberg, Y. (1997), "Multiple hypotheses testing with weights," *Scandinavian Journal of Statistics*, 24, 407-418.

Berger, R. L. (1982), "Multiparameter hypothesis testing and acceptance sampling," *Technometrics*, 24, 295-300.

Bhattacharya, J. and Vogt, W. B. (2012), "Do instrumental variables belong in propensity scores?" *International Journal of Statistics & Economics*, 9, 107-127.

Brooks, J. M. and Ohsfeldt, R. L. (2013), "Squeezing the balloon: propensity scores and unmeasured covariate balance," *Health Services Research*, 48, 3078-3094.

Burman, C. F., Sonesson, C., and Guilbaud, O. (2009), "A recycling framework for the construction of Bonferroni-based multiple tests," *Statistics in Medicine*, 28, 739-761.

Campbell, D. T. (1969), "Prospective: artifact and control," In R Rosenthal and R Rosnow, eds, *Artifact in Behavioral Research*, New York: Academic Press, pp. 351-382. Reprinted in Campbell (1988) *Methodology and Epistemology for Social Science: Selected Papers*. Chicago: University of Chicago Press.

Daniel, S., Armstrong, K., Silber, J.H., Rosenbaum, P.R. (2008), "An algorithm for optimal tapered matching, with application to disparities in survival," *Journal of Computational and Graphical Statistics*, 17, 914-924.

Dawid, A. P. (1979), "Conditional independence in statistical theory," *Journal of the Royal Statistical Society*, B 41, 1-31.

Freedman, D. A. (2008), "Randomization does not justify logistic regression," *Statistical Science*, 23, 237-249.

Hansen, B. B. and Klopfer, S. O. (2006), "Optimal full matching and related designs via network flows," *Journal of Computational and Graphical Statistics*, 15, 609-627. (Package `optmatch` in `R`)

Heller, R., Rosenbaum, P. R. and Small, D. S. (2010), "Using the cross-match test to appraise covariate balance in matched pairs," *American Statistian*, 64, 299-309.

Holland, P. W. (1988), "Causal inference, path analysis, and recursive structural equation models," *Sociological Methodology*, 18, 449-484.

Holm, S. (1979), "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 65-70.

Lange, K. (2003), *Applied Probability*, New York: Springer.

Laska, E. M. and Meisner, M. J. (1989), "Testing whether an identified treatment is best," *Biometrics*, 45, 1139-1151.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011), "Optimal nonbipartite matching and its statistical applications," *American Statistician* **65** 21-30. (`R` package `nbpmatching`)

Maritz, J. S. (1979), "Exact robust confidence intervals for location," *Biometrika*, 66, 163-166.

Meyer, B. D. (1995), "Natural and quasi-experiments in economics," *Journal Business and Economic Statistics*, 13, 151-161.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M. and Glynn, R. J. (2011), "Effects of adjusting for instrumental variables on bias and precision of effect estimates," *American Journal of Epidemiology*, 174, 1213-1222.

Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments," *Statistical Science*, 5, 463-480.

Pearl, J. (2010), "On a class of bias-amplifying variables that endanger effect estimates," in P. Grunwald and P. Spirtes, editors, *Proceedings of UAI*, 417-424.

Pearl, J. (2011), "Understanding bias amplification," *American Journal of Epidemiology*, 174, 1223-1227.

Pimentel, S. D., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2015), "Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons," *Journal American Statistical Association*, to appear. (`R` package `rcbalance`)

Pomp, E. R., Van Stralen, K. J., Le Cessie, S., Vandenbroucke, J. P., Rosendaal, F. R., Doggen, C. J. M. (2010), "Experience with multiple control groups in a large population-based case–control study on genetic and environmental risk factors," *European Journal*

Epidemiology, 25, 459-466.

Rosenbaum, P. R. (1984), "The consquences of adjustment for a concomitant variable that has been affected by the treatment," *Journal of the Royal Statistical Society*, A 147, 656-666.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41-55.

Rosenbaum, P. R. (1987a), "Sensitivity analysis for certain permutation inferences in matched observational studies," *Biometrika*, 74, 13-26.

Rosenbaum PR. (1987b), "The role of a second control group in an observational study," *Statistical Science*, 2, 292-316.

Rosenbaum, P. R. (2002), *Observational Studies*, New York: Springer.

Rosenbaum, P. R. (2004), "Design sensitivity in observational studies," *Biometrika*, 91, 153-164.

Rosenbaum, P. R. (2007), "Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies," *Biometrics*, 63, 456-464. (`R` packages `sensitivitymv` and `sensitivitymw`)

Rosenbaum, P. R. and Silber, J. H. (2009), "Amplification of sensitivity analysis in observational studies," *Journal American Statistical Association,* 104, 1398-1405. (`amplify` function in the `R` package `sensitivitymv`)

Rosenbaum, P. R. (2013), "Impact of multiple matched controls on design sensitivity in observational studies," *Biometrics*, 69, 118-127.

Rosenbaum, P. R. and Silber, J. H. (2013), "Using the exterior match to compare two entwined matched control groups," *American Statistian*, 67, 67-75.

Rosenbaum, P. R. (2015a), "How to see more in observational studies: Some new quasi-experimental devices," *Annual Review of Statistics and its Applications*, 2, 21-48.

Rosenbaum, P. R. (2015b), "Two R packages for sensitivity analysis in observational studies," *Observational Studies*, on-line early, http://obsstudies.org/.

Rosenbaum, P. R. (2015c), "Some counterclaims undermine themselves in observational studies," *Journal American Statistical Association,* on-line early.

Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688-701

Sanni Ali, M., Groenwold, R. H. H. and Klungel, O. H. (2014), "Propensity score methods and unobserved covariate balance," *Health Services Research*, 49, 1074-1082.

Shadish, W. R., Cook, T. D., Campbell, D. T. (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Silber, J. H., Rosenbaum, P. R., Ross, R. N., Even-Shoshan, O., Kelz, R. R., Neuman, M. D., Reinke, C. E., David, G., Saynisch, P. A., Kyle, F., Bratzler, D. W., Fleisher, L. A. (2011), "Medical and financial risks associated with surgery in the elderly obese," *Annals of Surgery*, 256, 79-86.

Silber, J. H., Rosenbaum, P. R., Clark, A. S., Giantonio, B. J., Ross, R. N., Teng, Y., Wang, M., Niknam, B. A., Ludwig, J. M., Wang, W., Even-Shoshan, O., Fox, K. R. (2013), "Characteristics associated with differences in survival among black and white women with breast cancer," *Journal American Medical Association*, 310:389-97.

Stuart, E. A. and Rubin, D. B. 2008, "Matching with multiple control groups with adjustment for group differences," *Journal Educational and Behavioral Statistics*, 33, 279-306.

Stuart, E. A. (2010), "Matching methods for causal inference," *Statistical Science*, 25: 1-21.

Welch, B. L. (1937), "On the z-test in randomized blocks and Latin squares," *Biometrika*, 29, 21-52.

Walker, A. M. (2013), "Matching on provider is risky," *Journal of Clinical Epidemiology*, 66, 565-568.

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., Mullen, P. D. (2008), "Alternatives to the randomized controlled trial," *American Journal Public Health* 98, 1359-66.

Wolfe, D. A. (1974), "A characterization of population weighted symmetry and related results," *Journal of the American Statistical Association*, 69, 819-822.

Wooldridge, J. (2006), "Should instrumental variables be used as matching variables?" Unpublished manuscript. East Lansing, MI: Michigan State University.

Zubizarreta, J. R., Neuman, M., Silber, J. H. and Rosenbaum, P. R. (2012), "Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia," *Journal of the American Statistical Association*, 107, 901-915.

Table 1: Degree of attenuation of bias Γ by not matching for a Normally distributed prod $\widetilde{x}$ with expectation 0 and standard deviation σ.

		Γ						
σ	α	1	1.5	2	3	4	5	10
1/2	−1	1.00	1.47	1.93	2.83	3.71	4.59	8.88
1/2	0	1.00	1.47	1.93	2.83	3.73	4.63	9.07
1/2	1	1.00	1.47	1.94	2.88	3.82	4.75	9.41
1	−1	1.00	1.40	1.78	2.49	3.16	3.81	6.82
1	0	1.00	1.40	1.78	2.50	3.19	3.86	7.11
1	1	1.00	1.41	1.81	2.58	3.34	4.08	7.75

Table 2: Degree of attenuation of bias Γ by matching treated units to control units with prod $\leq c\sigma$ for a normally distributed prod with expectation 0 and standard deviation σ, where the treatment assignment probabilities follow a logit model $\log\{\Pr(Z = 1|\tilde{x}, u)/\Pr(Z = 0|\tilde{x}, u)\} = \alpha + \tilde{x} + \gamma u$ with $\alpha = -1$ and $\Gamma = \exp(\gamma)$. The attenuation is measured by the odds ratio linking u and the group.

		Γ						
σ	c	1	1.5	2	3	4	5	10
1/2	-1	1.00	1.39	1.77	2.40	3.07	3.62	6.42
1/2	0	1.00	1.44	1.83	2.61	3.33	4.09	7.58
1/2	1	1.00	1.45	1.88	2.74	3.60	4.43	8.48
1	-1	1.00	1.30	1.54	1.99	2.35	2.67	4.14
1	0	1.00	1.36	1.66	2.20	2.73	3.22	5.42
1	1	1.00	1.38	1.74	2.42	3.04	3.68	6.49

Table 3: Standardized difference on the prod $\tilde{x}$ when matching treated units to control units with prod $\leq c\sigma$ for a normally distributed prod with expectation 0 and standard deviation σ, where the treatment assignment probabilities follow a logit model $\log\{\tilde{\omega}(\bar{x}, \tilde{x}, u)\} = \alpha + \tilde{x} + \gamma u$, with $\alpha = -1$ and $\Gamma = \exp(\gamma)$. The standardized difference is $\frac{E(\tilde{x}|Z=1) - E(\tilde{x}|Z=0, \tilde{x} \leq c\sigma)}{\sqrt{\frac{Var(\tilde{x}|Z=1) + Var(\tilde{x}|Z=0)}{2}}}$.

		Γ						
σ	c	1	1.5	2	3	4	5	10
1/2	-1	1.93	1.91	1.89	1.86	1.83	1.81	1.75
1/2	0	1.21	1.19	1.16	1.14	1.12	1.10	1.05
1/2	1	0.72	0.71	0.70	0.68	0.65	0.63	0.58
1	-1	2.31	2.28	2.24	2.19	2.15	2.11	2.00
1	0	1.56	1.53	1.50	1.45	1.41	1.39	1.30
1	1	1.09	1.07	1.05	1.01	0.98	0.96	0.87

Table 4: Sensitivity analysis in the example with two control groups, one matched (M) for $\widetilde{\mathbf{x}}$ with sensitivity parameter Γ, the other using $\widetilde{\mathbf{x}}$ as a prod (P) with sensitivity parameter $\Upsilon = 0.9 \times \Gamma$. The tabled values are upper bounds on marginal P-values using M-statistics with inner trimming, ψ_{in} with $\iota = 0.5$, $\kappa = 2$.

	Sensitivity parameters						
Γ	1.000	1.250	1.500	1.600	1.640	1.650	1.750
$\Upsilon = 0.9 \times \Gamma$	1.000	1.125	1.350	1.440	1.476	1.485	1.575
	Upper bounds on P-values testing no effect						
M-controls ($p'_{\Gamma,\max}$)	0.000	0.000	0.009	0.031	0.047	0.051	0.119
P-controls ($p^*_{\Upsilon,\max}$)	0.000	0.000	0.000	0.001	0.002	0.002	0.009

Table 5: Simulated power of an $\alpha = 0.05$ level sensitivity analysis with $I = 500$ matched triples, Normal errors, and an additive constant treatment effect that is τ standard deviations of a treated-minus-control matched pair difference. For $\alpha' = 0.05$, there is fixed-sequence testing, whereas for $\alpha' = 0.04$ there is recycling or equivalently a weighted Holm procedure. Either $\Upsilon = 0.9 \times \Gamma$ or $\Upsilon = \Gamma$. Two ψ-functions are compared. Estimated from 50000 independent replicates.

Γ	Υ	$\alpha' = 0.05$		$\alpha' = 0.04$		
Column Label		a	b	c	d	e
		H'_Γ	H^*_Υ	H'_Γ	H^*_Υ	H^*_Υ Alone
		$\tau = 1/4$ with ψ_{hu}				
1.50	1.35	0.47	0.44	0.46	0.66	0.82
1.50	1.50	0.47	0.30	0.44	0.35	0.47
		$\tau = 1/4$ with ψ_{in}				
1.50	1.35	0.66	0.63	0.65	0.80	0.90
1.50	1.50	0.66	0.50	0.63	0.56	0.66
		$\tau = 1/2$ with ψ_{hu}				
2.80	2.52	0.32	0.28	0.31	0.47	0.68
2.80	2.80	0.32	0.17	0.29	0.20	0.32
		$\tau = 1/2$ with ψ_{in}				
2.80	2.52	0.77	0.75	0.77	0.87	0.94
2.80	2.80	0.77	0.65	0.75	0.69	0.77

Supplement to "Constructed second control groups and attenuation of unmeasured biases"

Samuel D. Pimentel, Dylan S. Small, Paul R. Rosenbaum[1]

Abstract. This supplement provides a formal description of the matching algorithm used in the main paper to create balance on certain covariates and separation on others. This algorithm generalizes the large, sparse matching method of Pimentel et al. (2015) to a broader class of target covariate distributions. A method for tuning the algorithm to provide better simultaneous separation and balance is also detailed.

1 Introduction

In Section 5.3 of the main paper we briefly describe an algorithm that can balance treated and control groups closely on certain covariates $\overline{\mathbf{x}}$ while separating them on others $\widetilde{\mathbf{x}}$. This supplement provides a full technical specification of this algorithm. In Section 2 below, the general algorithm is described and its optimality is proven, using concepts and notation from Pimentel et al. (2015). In Section 3, further detail is given about how the general algorithm can be fine-tuned to create better separation on a prod. This section also gives specifics about how the second control group was created in the NHANES example given in the main paper.

[1]Samuel Pimentel is a doctoral student and Dylan Small and Paul Rosenbaum are professors in the Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340 US. spi@wharton.upenn.edu. Supported by the Measurement, Methodology, and Statistics Program of the National Science Foundation and by Fellowship FA9550-11-C-0028 from the Department of Defense, Army Research Office, National Defense Science and Engineering Graduate (NDSEG) Fellowship Program, 32 CFR 168a4. 11 June 2015.

2 Matching to a different target distribution

The large, sparse matching algorithm of Pimentel et al. (2015) requires that balance covariates $\nu_1, \nu_2, \ldots, \nu_K$ (given in decreasing order of importance) be nested within each other, i.e. all categories of ν_j are finer subdivisions of the categories of ν_{j-1}. In practice the covariates ν_i are often interactions of many nominal covariates measured in the dataset. The algorithm computes an optimal match by formulating the task as a network flow problem. Flow constraints in certain edges of the network are set based on the empirical covariate distribution of the treated units, and require the covariate distribution of the controls to be as close as possible to this distribution. In this technical sense the treated group provides the "target distribution" to which the selected control will be made similar. We wish to modify this algorithm so that a different target distribution can be used.

Formally, we transform the algorithm as follows. Here we adopt the notation of Pimentel et al. (2015). In the original algorithm there was a treated group $\mathcal{T}$ and a control group $\mathcal{C}$. Define a third group $\mathcal{T}' = \{\tau'_1, \tau'_2, \ldots, \tau'_T\}$ where $T = |\mathcal{T}|$ and call it the target group. We also extend the domain of each nested covariate ν_k to include $\mathcal{T}'$ so now $\nu_k : \mathcal{T} \cup \mathcal{C} \cup \mathcal{T}' \longrightarrow \mathcal{K}_k$; in other words, the units in the target group take values for each of the nested covariates. We now alter the algorithm by changing the definition of the quantities $d_{k\ell}$ for $\ell = 1, \ldots, L_k$ and $k = 1, \ldots, K$. In the original algorithm, these are defined as:

$$d_{k\ell} = |\{\tau_t \in \mathcal{T} \ : \ \nu_k(\tau_t) = \lambda_{k\ell}\}|$$

In short, $d_{k\ell}$ counts the number of individuals in category ℓ of covariate k in the treated group $\mathcal{T}$. We change the definition so that instead $d_{k\ell}$ is equal to the number of individuals

in category ℓ of covariate k in the target group $\mathcal{T}'$:

$$d_{k\ell} = \left|\left\{\tau'_t \in \mathcal{T}' \ : \ \nu_k(\tau'_t) = \lambda_{k\ell}\right\}\right|$$

This mainly affects the algorithm through the quantities $\beta_{k\ell}$, which give the covariate imbalance at a particular category and are defined as follows:

$$\beta_{k\ell} = m \times d_{k\ell} - \left|\{(\tau_t, \kappa_c) \in \mathcal{M} \ : \ \nu_k(\kappa_c) = \lambda_{k\ell}\}\right|$$

These $\beta_{k\ell}$ terms are used in Definition 2 of Pimentel et al. (2015) to define refined covariate balance. So in changing the $d_{k\ell}$ values we not only transform the algorithm but broaden the definition of refined covariate balance, so that balance is now with respect to a particular target distribution $\mathcal{T}'$. This leads to the following proposition;

Proposition (A1). *Given a target group $\mathcal{T}'$, if we alter the $d_{k\ell}$ values and associated $\beta_{k\ell}$ values as outlined above to obtain a new algorithm and a new definition of refined covariate balance, then the new algorithm produces an optimal match with refined covariate balance with respect to $\mathcal{T}'$.*

Proof. The proof is identical to the optimality proof for the original algorithm, except that we use the new definition for $d_{k\ell}$ and the resulting new definition of $\beta_{k\ell}$. □

Notice that the new version of the proof includes the old version as the special case when $\mathcal{T} = \mathcal{T}'$. However, it also shows the optimality of the modified algorithm for matching under refined covariate balance with respect to any empirical covariate distribution generated by T observations on $\nu_1 \times \nu_2 \times \ldots \times \nu_K$.

3 Creating better separation on a prod

The balance constraints for the large, sparse optimal matching algorithm are described by the decreasingly-important, increasingly-fine nominal covariates $\nu_1, \nu_2, \ldots, \nu_K$. When matching to create separation, these covariates could be formed by relevant functions and interactions of $\overline{\mathbf{x}}$ and $\eta(\widetilde{\mathbf{x}})$ (where η is defined as in Section 5.3 of the main paper). As in the original version of large, sparse matching with refined covariate balance, the best choice of K and of the nested covariates $\nu_1, \ldots, \nu_K$ is highly application- and data-dependent, and researchers may need to experiment with several different configurations to obtain acceptable balance results.

To improve observed separation on $\widetilde{\mathbf{x}}$, the researcher may find it useful to define balance constraints not just in terms of the single function of $\widetilde{\mathbf{x}}$ described by η but in terms of a series of such constraints $\eta_1, \ldots, \eta_J$. For example, one might define a series of J sets $\mathcal{X}'_j \subset \mathcal{X}$ such that $\mathcal{X}'_1 \subset \mathcal{X}'_2 \subset \ldots \subset \mathcal{X}'_J$ where $\mathcal{X}'_1$ is the region from which the researcher would most like controls to be selected and $\mathcal{X}'_2, \ldots, \mathcal{X}'_J$ are regions from which to select the controls if this is not possible, in decreasing order of preference. Then one could define new covariates

$$\eta_j(\widetilde{\mathbf{x}}_i) = \begin{cases} 1 & \text{if } \widetilde{\mathbf{x}}_i \in \mathcal{X}_j \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1, \ldots, J$ and set the values of each $\eta_j(\widetilde{\mathbf{x}}_i)$ to 1 in the target distribution. These covariates and their interactions with $\overline{\mathbf{x}}$ would grant the researcher greater flexibility in defining the balance constraints $\nu_1, \ldots, \nu_K$ and might lead to better combinations of $\widetilde{\mathbf{x}}$-separation and $\overline{\mathbf{x}}$-balance.

In the NHANES example of Section 1.2 of the main paper, we used the 5-level ordinal measure of education and the continuous measure of socioeconomic status to define the

following desirable regions from which to draw controls:

$$\mathcal{X}_1 = \text{income-to-poverty ratio above 2}$$

$$\mathcal{X}_2 = \text{income-to-poverty ratio above 2, high school graduate}$$

$$\mathcal{X}_3 = \text{income-to-poverty ratio above 4, some college}$$

$$\mathcal{X}_4 = \text{income-to-poverty ratio above 4, college graduate}$$

We then enforced balance on a series of interactions of the resulting variables $\eta_1(\widetilde{\mathbf{x}}), \ldots, \eta_4(\widetilde{\mathbf{x}})$ with the balance covariates $\overline{\mathbf{x}}$. We controlled for η_1 at an early, coarse level in the balance hierarchy (to ensure most controls had at least a moderate level of income) and added the other more stringent variables η_j at finer, less-prioritized levels in the hierarchy (to ensure better-educated and wealthier controls were chosen when available).

References

Pimentel, S. D., Kelz, R. R., Silber, J. H. and Rosenbaum, P. R. (2015), "Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons," *Journal American Statistical Association*, to appear. (`R` package `rcbalance`)