Using Evidence Factors to Clarify Exposure Biomarkers

Abstract. A study has two evidence factors if it permits two statistically independent inferences about one treatment effect such that each factor is immune to some bias that would invalidate the other factor. Because the two factors are statistically independent, the evidence they provide may be combined using methods associated with meta-analysis for independent studies, despite using the same data twice in different ways. We illustrate evidence factors, applying them in a new way in investigations that have both an exposure biomarker and a coarse external measure of exposure to a treatment. To illustrate, we consider the possible effects of cigarette smoking on homocysteine levels, with self-reported smoking and a cotinine biomarker. We examine joint sensitivity of two factors to bias from confounding, a central aspect of any observational study.

Keywords: Biomarkers; evidence factors; reactive doses; sensitivity to confounding.

Doses of exposure are commonly conceived, along experimental lines, as versions or levels of a treatment that an external environment inflicts upon an individual, but many doses are not that simple. Observing that "exposure biomarkers indicate more than just exposure," Savitz and Wellenius [29] write: "many physiological processes that might affect biomarker levels ... may also influence or be influenced by disease processes related to the health outcome of interest, potentially leading to confounding bias." Perera and Weinstein [16, p. 518] write: "Biomarkers of internal dose take into account individual differences in absorption, metabolism, bioaccumulation or excretion of the compound in question," so they are more complex than a direct measure of the intensity of an environmental assault; see also [27].

A dose that is determined by the external environment is called "nonreactive,"

whereas a dose that may incorporate an individual's reaction to the environment, or the environment's reaction to the individual, is "reactive" [22]. A nonreactive dose is a treatment, a shot in the arm; whereas, a reactive dose is an outcome of treatment, say an immune response to a vaccine, partly indicative of the intensity of treatment, but perhaps incorporating heterogenious reactions of different individuals to the same shot in the arm. The level of a toxin in the air is nonreactive, but two people exposed to the same level of that toxin in the air may have different levels of that toxin's metabolites in their blood because their kidneys differ in their ability to filter the toxin; then, blood levels of the toxin are reactive, that is, an outcome related to the intensity of treatment. Issues of this sort led Weisskopf and Webster [32] to argue for coarser, less personal, nonreactive measures of treatment or dose that are unambiguously part of the external environment.

Evidence of causality would be more compelling if there were two independent studies of different data by different investigators, one using a coarse nonreactive dose, the other using a precise but potentially reactive dose, and if these two studies with different limitations concurred. Susser [31, p. 148] argued replication is not repetition: replication occurs only if "diverse approaches produce similar results."

Statistical independence is a precise technical concept meaning that the outcome of one study cannot be used to predict the outcome of the other, but under common circumstances two studies of different data by different investigators would yield statistical independence. These independent studies might be combined using meta-analytic tools to provide stronger evidence than either study provides on its own, with each study providing a check on the limitations of the other study.

Less obviously, with sufficient knowledge and care in design, a single investigator can analyze the same data twice in such a way that two statistically independent inferences are drawn, a method known as "evidence factors"; see [11,12,20,21,23,25] and [26, pp. 136–141]. Care in design is required to produce the needed statistical independence, because the same data are used twice. For the simplest example, see [23, §3.2]. With insufficient care, two analyses of the same data are far from independent, greatly exaggerating the data's strength, as if the investigator had duplicated the data set in a misguided effort to increase the sample size.

Here, we propose the use of two evidence factors, one based on a nonreactive treatment, the other based on a potentially reactive dose. Each analysis provides a check on the other, again with the possibility of combining statistically independent results using meta-analytic tools.

STUDY DESIGNS WITH EVIDENCE FACTORS

A study design has two evidence factors if treatment assignment splits into two aspects exhibiting certain symmetries. In the current paper, the study design is the dose-control design: it consists of treated-control matched pairs, where the treated subjects in different pairs received different doses of treatment. One factor is the treated-control comparison within pairs, ignoring doses. The other factor relates the treated-minus-control pair differences in outcomes to the variation in doses among pairs. It is possible to show that this design yields two evidence factors when analyzed in certain ways [20]. Here, we focus on a dose-control design in which the treatment/control distinction is known to be nonreactive, but the doses may be

reactive. Do the two factors concur?

Besides the dose-control design, there are many other designs with two or more evidence factors. Stated informally, these designs have factors with certain symmetries, in some ways analogous to the symmetries in the theory of experimental design [3], and the associated statistical procedures react to these symmetries in a specific way; however, a general yet precise description requires some mathematical tools not developed here [25]. Instead, consider a few specific designs.

In occupational epidemiology, it is common to have exposed subjects in, say, a factory that exposes individuals to a potential toxin, and controls from some entirely different place without known exposure to toxins. Among those exposed in the factory, some individuals have occupations that entail direct, intense exposure to the toxin, while others have occupations that are remote from direct exposures. The biases that lead people to work in the factory may differ from the biases that lead to specific occupations inside the factory. In this case, using suitable statistical methods, factory-versus-control constitutes one evidence factor, and within the factory, direct-versus-indirect exposure constitutes a second factor. For the analysis of an example, see [21, Table 1] and the help-files for `mtm` and `truncatedP` in the `sensitivitymv` package in `R`. In parallel, an observational study of prenatal mortality built a first evidence factor from babies born before exposure began, and a second factor from babies born during the period of exposure though geographically separated from exposure [35]. This study also contains a "placebo" test for bias [17], comparing the two geographic regions before exposure began.

Some treatments are given by institutions, say hospitals or prisons. Some insti-

Table 1: Distribution of covariates in matched pairs. Values are percentages.

	Never-smoker	Daily smoker
Female	43	43
Age $\geq$ 50	63	63
Black	25	24
Hispanic	17	15
Neither Black nor Hispanic	58	62

tutions typically give treatment A, others typically give B, others selectively give A or B. One factor compares similar patients receiving A or B in the same hospital, the other factor compares similar patients receiving A or B at different hospitals that strongly prefer one treatment to the other [36].

AN EXAMPLE OF TWO EVIDENCE FACTORS

Bazano et al. [1] asked whether cigarette smoking increases homocysteine levels by studying the association between homocysteine and cotinine, a biomarker for exposure to tobacco. To illustrate evidence factors, we reexamine this using more recent data from the 2003-2004 and 2005-2006 National Health and Nutrition Examination Surveys, the most recent to measure homocysteine levels. Daily smokers are compared to never smokers. In total, 1645 daily smokers were individually matched to 1645 never smokers, matching for sex, age, race and education. Figure 1 and Table 1 show the covariates in 1645 matched pairs. Matching was done in a conventional way, using a covariate distance and propensity score [10], exact matching for sex and ten-year age categories, minimizing the total distance within pairs [6].

Figure 2 exhibits the data that enter into the two evidence factors. One coarse

comparison makes no use of the cotinine biomarker, along the lines suggested by Weisskopf and Webster [32], while the other acts as if the biomarker were a nonreactive dose, as in the original investigation [1]. Do these analyses concur? If they do concur, to what quantitative degree is the evidence strengthened?

To the left in Figure 2(i), there is a boxplot of the smoker-minus-control pair differences in logs of homocysteine levels. Despite taking logs to limit extreme observations, the differences in homocysteine levels are not Normally distributed, so nonparametric methods are used throughout [7]. In Figure 2(iii), there is a boxplot of smoker-control differences in cotinine. A few of the 1645 cotinine differences are negative, indicating substantial exposure to tobacco by individuals who described themselves as never-smokers. Finally, in Figure 2(ii), there is the "crosscut plot" in which pair differences in logs of homocysteine levels are plotted against pair differences in cotinine levels. Each pair difference compares two individuals who are similar on covariates. The crosscut plot cuts a cross from the scatter-plot, with points inside the quartiles appearing in gray, points outside appearing in black.

Our first analysis uses conventional methods that assume, somewhat naively, that there is no unmeasured confounding. Using methods derived from Wilcoxon's signed rank statistic [7, §3.1-§3.3], the median difference in logs of homocysteine levels is estimated to be 0.0963 with 95% confidence interval [0.0775, 0.1151], or a multiplicative effect of $e^{0.0963} = 1.101$ or a 10.1% increase for smokers, with 95% confidence interval [8.1%, 12.2%]. The two-sided P-value testing no effect is $\leq 2.2 \times 10^{-16}$. For several definitions of "no effect" that permutation tests correctly do not distinguish, see [13, §5.8-§5.12].

The crosscut statistic [24] forms the 2×2 table of counts of black points in Figure 2(ii). Like the corner test [15], the crosscut test compares the 2×2 table to a hypergeometric distribution, but it focuses on outer corners with the consequence that it performs well in sensitivity analyses: specifically, it has high design sensitivity, even with modest correlations between dose and response [24]; see [30, 37] and [26, §10] for general discussion of design sensitivity. Intuitively, large effects tend to be insensitive to large biases, and the crosscut statistic focuses on large differences in dose and response. The crosscut odds ratio is $1.89 = (122 \times 117) / (90 \times 84)$ and the two sided P-value testing independence is 0.0019: larger smoker-control pair differences in cotinine predict larger smoker-control differences in homocysteine.

The two tests just performed have several remarkable properties; they are two evidence factors [20]. First, the two tests, the two P-values, are statistically independent when their null hypotheses are true: it is as if they came from two unrelated studies using different data sets. This first property does not hold for most pairs of two statistics, but it does hold for the signed-rank and crosscut statistics and some others [25]. Second, bias, no matter how strong, in who reports smoking does not affect the crosscut test, and bias, no matter how strong, in the potentially reactive nature of the cotinine biomarker does not affect the signed-rank test. This second property will be made clearer in the sensitivity analysis to follow later. The two tests in Figure 2 are not infallible — no scientific data are infallible — but the two tests are two entirely separate pieces of information, fallible in different ways, yet supporting the same conclusion.

Because the P-values are independent when testing their null hypotheses, they

may be combined into a single P-value using techniques for meta-analysis, as if they came from unrelated studies. Again, this is valid only because they are evidence factors. A traditional method for combining independent P-values is Fisher's method, which derives a new P-value from the distribution of the product of the two P-values. One generalization, the truncated product P-value [34] is derived from the distribution of the product of P-values that are at most ι, conventionally $\iota = 0.2$; it is implemented as `truncatedP` in the `sensitivitymv` package in `R`, becoming Fisher's method when $\iota = 1$. The truncated product has higher power than Fisher's method in sensitivity analyses in observational studies [9].

For instance, with $\iota = 0.2$, a pair of independent P-values of 0.05 and 0.1 combine via the truncated product method to a single P-value of 0.023, the pair 0.05 and 0.05 combine to 0.013, but the pair 0.05 and 0.5 combine to 0.12. When the two factors concur, the reported evidence of an effect is strengthened, but when one factor finds nothing the reported evidence is weakened. In Figure 2, each factor has a small P-value, so the P-value derived from the truncated product is vanishingly small.

Closed testing is one method for testing several hypotheses. Using the truncated product, closed testing terminates if the combined P-value is $> \alpha$, conventionally $\alpha = 0.05$; otherwise, if the combined P-value is $\leq \alpha$, the two separate factors are each tested at level α. Alternatively, the two hypotheses may be tested in order, testing the first factor at level α, continuing on to test the second factor at level α only if the first factor's P-value is $\leq \alpha$. Both closed testing and testing in order falsely reject at least one true hypothesis with probability $\leq \alpha$ despite testing several hypotheses at level α [5]. Testing in order is appropriate when the second factor has

no prospect of being credible without support from the first factor. Our analysis uses closed testing, thereby giving equal emphasis to the two factors. In Figure 2, closed testing rejects using the combined P-value, then rejects using each P-value separately, so the two factors concur in finding two independent pieces of information linking smoking with increased homocysteine.

COULD THE ASSOCIATIONS REFLECT CONFOUNDING RATHER THAN A CAUSAL EFFECT?

In a randomized trial, random numbers assign treatments, ensuring the treatment an individual receives is statistically independent of every attribute of that individual, observed or not, thereby justifying the randomization inferences reported in a clinical trial [3, §2]. In an observational study, treatments are not randomly assigned: adjustments or matching may remove confounding due to observed covariates [26, §5], but confounding due to unmeasured covariates is possible [26, §9]. Write θ_i for the probability that the first individual in matched pair i receives treatment, with the second assigned to control, so $\theta_i = \frac{1}{2}$ in a paired randomized trial.

A sensitivity analysis in an observational study asks about the magnitude of bias from unmeasured covariates that would need to be present to alter the qualitative conclusions of an observational study. What magnitude of bias from nonrandomized treatment assignment would need to be present to accept a null hypothesis of no treatment effect that was rejected when assuming there is no unmeasured confounding? Stated differently: What magnitude of bias would need to be present for the confidence interval for the size of the effect to include zero effect? A simple

approach [18, 19, 26] builds upon the familiar Cornfield inequality [2], quantifying the departure from random treatment assignment by a number $\Gamma \geq 1$; for other approaches, see [4, 8, 14, 28, 33]. Here, $\Gamma = 1$ signifies no unmeasured confounding, with each matched pair have probability $\theta_i = \frac{1}{2}$. If $\Gamma = 1.25$, then the treatment assignment probabilities θ_i are unknown, but the magnitude of their departure from randomized assignment is limited to $0.44 \leq \theta_i \leq 0.56$, and such a bias could be produced by an unobserved covariate that doubled the odds of treatment and doubled the odds of a positive pair difference in the absence of an actual treatment effect [26, Table 9.1]. If $\Gamma = 1.5$, then the departure from randomized assignment is limited to $0.4 \leq \theta_i \leq 0.6$, and such a bias could be produced by an unobserved covariate that doubled the odds of treatment and increased the odds of a positive pair difference by four-fold. In general, a bias of Γ in treatment assignment means that two individuals matched for observed covariates may differ in their odds of treatment by a factor of Γ, so that $1/(1+\Gamma) \leq \theta_i \leq \Gamma/(1+\Gamma)$, and this bias may be produced by an unobserved covariate that increases the odds of treatment by a factor of Λ and the odds of a positive pair difference in outcomes by a factor of Δ where $\Gamma = (\Lambda\Delta + 1)/(\Lambda + \Delta)$.

Table 2 displays three sensitivity analyses, one for the analysis of Figure 2(i), another for the analysis of Figure 2(ii), and a third for their combination using the truncated product method. The values in Table 2 are the maximum possible P-value that a bias of Γ can produce in the absence of a treatment effect, so if this value is less than α, conventionally $\alpha = 0.05$, then a bias of Γ is too small to explain away rejection of the null hypothesis of no treatment effect. Considered alone,

Table 2: Sensitivity analysis for Wilcoxon's test alone, for the cross-cut test alone, and for their combination using the truncated product of P-values. The tabulated values are upper bounds on one-sided P-values testing no treatment effect; double these for two-sided P-values.

			Γ for Wilcoxon's Test					
			1	1.25	1.5	1.6	1.75	∞
Wilcoxon P-value $\rightarrow$			0.000000	0.000000	0.002143	0.038751	0.400004	1.000000
Cross-cut P-value $\downarrow$			Combining Wilcoxon and cross-cut P-values					
Γ for	1	0.000942	0.000000	0.000000	0.000025	0.000350	0.005981	0.005981
Cross	1.15	0.008249	0.000000	0.000000	0.000182	0.002375	0.034469	0.034469
Cut	1.25	0.024162	0.000000	0.000000	0.000479	0.005950	0.075001	0.075001
Test	1.35	0.056563	0.000000	0.000000	0.001018	0.012064	0.130500	0.130500
	∞	1.000000	0.000000	0.000000	0.011842	0.101982	1.000000	1.000000

the Wilcoxon test in Figure 1(i) is insensitive to a bias of $\Gamma = 1.5$ with maximum possible P-value 0.00214. Taken alone, the crosscut test in Figure 1(ii) is insensitive to a bias of $\Gamma = 1.25$ with maximum possible P-value of 0.024. Moreover, these two analyses are entirely separate pieces of information, so they provide mutually reinforcing information. Combining the P-value of 0.0388 for Wilcoxon's test at $\Gamma = 1.6$ and the P-value of 0.024 for the crosscut test at $\Gamma = 1.25$ yields a combined P-value of 0.00595. Even $\Gamma = \infty$ for either factor is insufficient to accept the hypothesis of no effect, provided the bias affecting the other factor is not too large, $\Gamma = 1.5$ for Wilcoxon's test, $\Gamma = 1.15$ for the crosscut test.

ALTERNATIVE ANALYSES OF TWO EVIDENCE FACTORS AND CONFIDENCE INTERVALS

An on-line supplement presents alternative analyses. In Figure 2(ii), the crosscut analysis controlled for age and sex indirectly, using the fact that pairs are matched

for age and sex. An alternative analysis stratifies pairs matched for age and sex, adjusting twice for age and sex. Wilcoxon's statistic is known to exaggerate sensitivity to bias, and the supplement finds greater insensitivity when a better statistic is used. As always, confidence intervals are obtained by inverting hypothesis tests; e.g., [13, §3.5] and [7, §3.1-§3.3]. The supplement obtains shorter, more informative confidence intervals in sensitivity analyses by replacing Wilcoxon's statistic.

SIMULATED ILLUSTRATION OF TWO EVIDENCE FACTORS

To illustrate, a simulation was conducted. Let y_i be the ith pair difference in homocysteine levels as depicted in Figure 2(i) and let x_i be the ith pair difference in cotinine levels, as depicted in Figure 2(iii), for $i = 1, \ldots, I$ where I is the number of pairs. The simulation oversimplifies some of the mathematics of about evidence factors, but it illustrates key ideas.

The simulation model is: $y_i = \lambda\tau + (1-\lambda)\, x_i + \varepsilon_i$ where (i) the ε_i are Normal with expectation zero and variance ν, (ii) the x_i are Normal with expectation $\tau \geq 0$ and variance κ, (iii) ε_i and the x_i are independent, (iv) $0 \leq \lambda \leq 1$; so, y_i has expectation τ and variance $(1-\lambda)^2 \kappa + \nu$. Setting $(1-\lambda)^2 \kappa + \nu = 1$ makes τ the average effect of smoking in units of the standard deviation of a pair difference, $\sqrt{\text{var}\,(y_i)}$. The correlation of (y_i, x_i) is $(1-\lambda)\sqrt{\kappa}$.

If $\lambda = 1$, the average effect of smoking is τ but x_i is irrelevant, independent of y_i. If $\lambda = 0$, then the average effect is still τ, with the effect on y_i proportional to its effect on x_i. If $\tau = 0$, there is no effect on y_i, but if $\lambda < 1$, then x_i is misleadingly correlated with y_i. Will the evidence factor analysis help to distinguish these very

different situations? Indeed, it will.

Figure 3 simulates two evidence factors in four situations, plotting the P-value from the crosscut test against the P-value from Wilcoxon's test. The simulation creates 5000 studies, each with $I = 500$ matched differences (y_i, x_i), similar to Figure 2. Each point in Figure 3 is one study. In Figure 3(i), there is no effect of smoking ($\tau = 0$) and cotinine is independent of homocysteine ($\lambda = 1$), and in this case the two P-values are independent and uniform on the unit square. In Figure 3(ii), there is a treatment effect ($\tau = .15$) but the size of the effect does not track the level of cotinine ($\lambda = 1$), so P-values from Wilcoxon's test are small but the P-values from the crosscut test are uniform on $[0, 1]$, and again the P-values from the two tests are independent. In Figure 3(iii), there is no treatment effect ($\tau = 0$) but cotinine and homocysteine are related ($\lambda = .8$) with correlation $(1 - \lambda)\sqrt{\kappa} = 0.141$, so the P-values from Wilcoxon's test are uniform on $[0, 1]$ but the P-values from the crosscut test are small, and the two P-values are independent. In Figure 3(iv), there is an average treatment effect ($\tau = .15$) and it is larger when the difference in cotinine is larger, so both P-values are typically small.

The two factors concur if closed testing terminates with rejection at $\alpha = 0.05$ for both factors. A concurrence is false if the null hypothesis is true for either factor. The concurrence rate, ρ, is the probability of concurrence. We would like ρ to be low when at least one null hypothesis is true and high when both are false. The simulation estimates ρ by $\widehat{\rho}$ in Figure 3. In Figure 3(i)-(iii), $\widehat{\rho}$ is less than $\alpha = 0.05$. In Figure 3(iv), there is an association to be found in both factors, and the estimated concurrence rate is $\widehat{\rho} = 0.74$. So the method is performing as theory says it should,

protecting against a false report of concurrence.

DISCUSSION

Biomarkers of exposure reflect both the dose of an environmental assault and an organism's reaction to that assault, and in that sense a biomarker is unlike a dose of treatment manipulated in a laboratory experiment. A biomarker and an outcome may be associated even when the environmental exposure has no effect [29], leading some investigators to prefer coarser but unambiguously nonreactive doses of treatment [32]. Evidence factors offer an investigator the opportunity to perform two or more statistically independent studies at the same time, one using the potentially reactive biomarker, the other using a nonreactive dose. Because the two studies provide two statistically independent tests of the treatment effect, their P-values may be combined by meta-analytic techniques used to combine independent studies by different investigators. Because the two studies have different limitations, they may resolve the ambiguity introduced by the reactive nature of biomarkers. If the two studies concur in finding a treatment effect, they formally provide stronger evidence than either study would provide on its own.

References

[1] Bazzano LA, He J, Muntner P, Vupputuri S, Whelton PK. Relationship between cigarette smoking and novel risk factors for cardiovascular disease in the United States. *Ann Intern Med.* 2003;138:891-897.

[2] Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *J Nat Cancer Inst.* 1959;22:173-203. Reprinted: *Int J Epidemiol* 2009;38:1175-1191.

[3] Fisher RA. *Design of Experiments.* Edinburgh: Oliver and Boyd; 1935.

[4] Fogarty CB. Studentized sensitivity analysis for sample average treatment effect in paired observational studies. *J Am Statist Assoc.* 2019, to appear.

[5] Goeman JJ, Solari A. The sequential rejection principle of familywise error control. *Ann Statist.* 2010;38:3782-3810.

[6] Hansen BB, Klopfer SO. Optimal full matching and related designs via network flows. *J Comp Graph Statist.* 2006;15:609-27.

[7] Hollander M, Wolfe DA, Chicken E. *Nonparametric Statistical Methods.* NY: Wiley; 2014.

[8] Hosman CA, Hansen BB, Holland PW. The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann Appl Statist.* 2010;4:849-870.

[9] Hsu JY, Small DS, Rosenbaum PR. Effect modification and design sensitivity in observational studies. *J Am Statist Assoc.* 2013;108:135-48.

[10] Joffe MM, Rosenbaum PR. Invited commentary: Propensity scores. *Am J Epidemiol.* 1999;150:327-333.

[11] Karmakar B, French B, Small DS. Integrating the evidence from evidence factors in observational studies. *Biometrika*. 2019;1066:353-367

[12] Karmakar B, Small DS, Rosenbaum PR. Using approximation algorithms to build evidence factors and related designs for observational studies. *J Comp Graph Statist*. to appear.

[13] Lehmann EL, Romano JP. *Testing Statistical Hypotheses*. NY: Springer, 2005.

[14] McCandless LC, Gustafson P, Levy A. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statist Med*. 2007;26:2331-2347.

[15] Olmstead PS, Tukey JW. A corner test for association. Ann Math Statist 1947;18:495-513.

[16] Perera FP, Weinstein IB. Molecular epidemiology: recent advances and future directions. *Carcinogenesis*. 2000;21:517-524.

[17] Rosenbaum PR. From association to causation in observational studies. *J. Am. Statist. Assoc*. 1984;79:41-48.

[18] Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*. 1987;74:13-26.

[19] Rosenbaum PR. Discussing hidden bias in observational studies. *Ann Intern Med*. 1991;115:901-905.

[20] Rosenbaum PR. Evidence factors in observational studies. *Biometrika*. 2010;97:333-345.

[21] Rosenbaum PR. Some approximate evidence factors in observational studies. *J. Am. Statist. Assoc.* 2011;106:285-295.

[22] Rosenbaum PR. Nonreactive and purely reactive doses in observational studies. In: Berzuini C, Dawid AP, Bernardinelli L, eds. *Causality*, C Berzuini, P Dawid, L Bernardinelli, eds. NY: Wiley; 2012;273-289.

[23] Rosenbaum PR. How to see more in observational studies: Some new quasi-experimental devices. *Ann Rev Statist App.* 2015;2:21-48.

[24] Rosenbaum PR. The crosscut statistic and its sensitivity to bias in observational studies with ordered doses of treatment. *Biometrics.* 2016;72:175-183.

[25] Rosenbaum PR. The general structure of evidence factors in observational studies. *Stat Sci.* 2017;32:514-530.

[26] Rosenbaum PR. *Observation and Experiment: An Introduction to Causal Inference.* Cambridge, MA: Harvard, 2017.

[27] Rothman N, Stewart WF, Schulte PA. Incorporating biomarkers into cancer epidemiology. *Cancer Epidem Biomark Prev.* 1995;4:301-311.

[28] Rudolph KE, Stuart EA. Using sensitivity analyses for unobserved confounding to address covariate measurement error in propensity score methods. *Am J Epidemiol.* 2017;187:604-13.

[29] Savitz DA, Wellenius GA. Exposure biomarkers indicate more than just exposure. *Am J Epidem.* 2018;187:803-805.

[30] Stuart EA, Hanna DB. Should epidemiologists be more sensitive to design sensitivity? *Epidemiology*. 2013;24:88-89.

[31] Susser M. *Causal Thinking in the Health Sciences*. NY: Oxford, 1973.

[32] Weisskopf MG, Webster TF. Trade-offs of personal versus more proxy exposure measures in environmental epidemiology. *Epidemiol*. 2017;28:635-643.

[33] Yu B, Gastwirth J. Sensitivity analysis of trend tests. *Biostatistics* 2005;6:201-9.

[34] Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method of combining P-values. *Genet Epidemiol*. 2002;22:170-185.

[35] Zhang K, Small DS, Lorch S, Srinivas S, Rosenbaum PR. Using split samples and evidence factors in an observational study of neonatal outcomes. *J Am Statist Assoc*. 2011;106:511-524.

[36] Zubizarreta JR, Neuman M, Silber JH, Rosenbaum PR. Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia. *J Am Statist Assoc*. 2012;107:901-915.

[37] Zubizarreta JR, Cerdá M, Rosenbaum PR. Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design. *Epidemiology*. 2013;24:79–87.

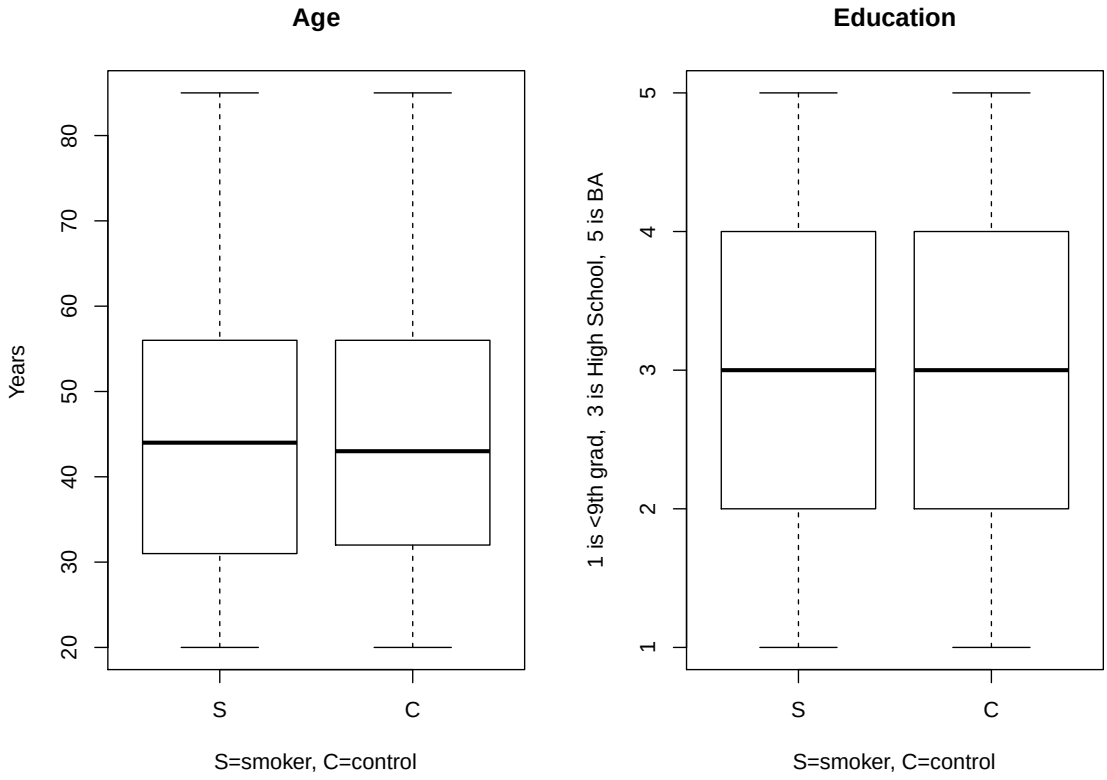

Figure 1: Boxplots of age and education in 1645 matched pairs of one daily smoker (S) and one never-smoking control (C). Education is in five categories, where 1 is $\leq$ 9th grade, 3 is a high school degree or equivalent, and 5 is a BA degree or more.

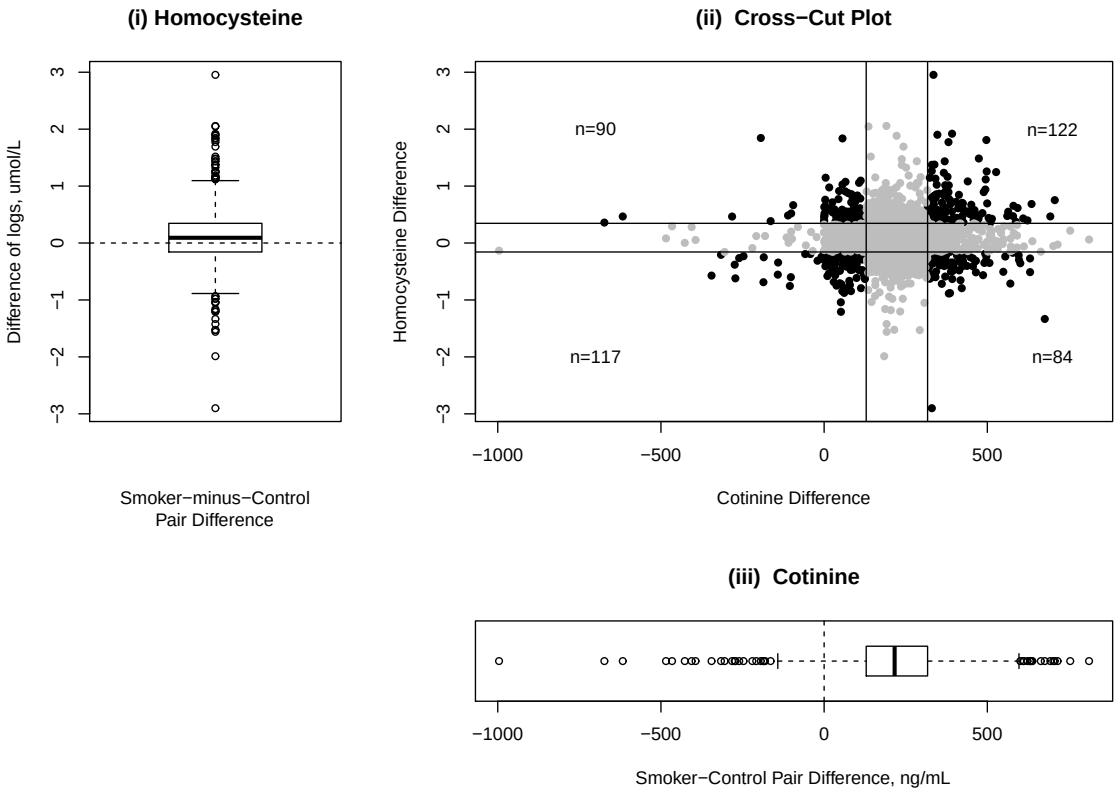

Figure 2: Smoker-minus-control matched pair differences in logs of homocysteine levels, in cotinine levels, and their relationship. Points between the quartiles are in gray, points outside the quartiles are in black. The black points define the cross-cut test.

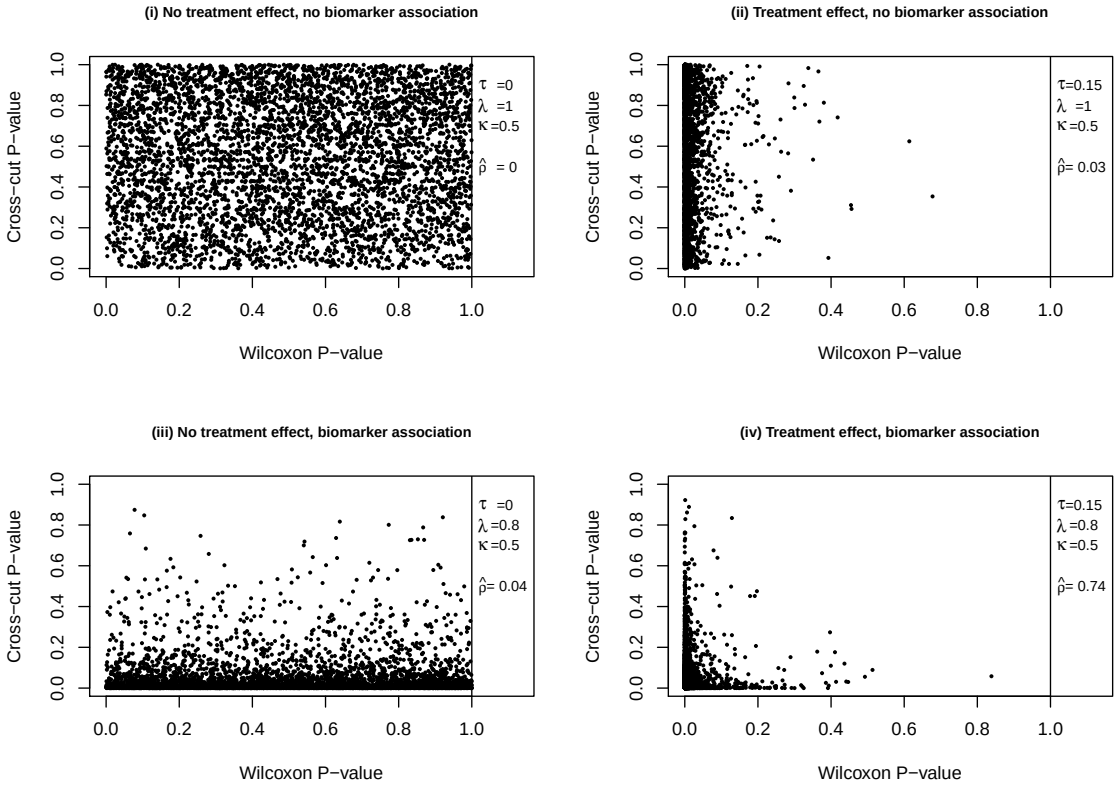

Figure 3: Five thousand simulated pairs of P-values from two evidence factors, one from the Wilcoxon signed rank test, the other from the cross-cut test. The concurrence rate ρ is estimated by the simulation as $\widehat{\rho}$ with standard error ≤ 0.01.