Published in the *Journal of the American Statistical Association* 2015; 110: 1389-1398.

Some counterclaims undermine themselves in observational studies

Paul R. Rosenbaum[1]

Abstract: Claims based on observational studies that a treatment has certain effects are often met with counterclaims asserting that the treatment is without effect, that associations are produced by biased treatment assignment. Some counterclaims undermine themselves in the following specific sense: presuming the counterclaim to be true may strengthen the support that the original data provide for the original claim, so that the counterclaim fails in its role as a critique of the original claim. In mathematics, a proof by contradiction supposes a proposition to be true en route to proving that the proposition is false. Analogously, the supposition that a particular counterclaim is true may justify an otherwise unjustified statistical analysis, and this added analysis may interpret the original data as providing even stronger support for the original claim. More precisely, the original study is sensitive to unmeasured biases of a particular magnitude, but an analysis that supposes the counterclaim to be true may be insensitive to much larger unmeasured biases. The issues are illustrated using data from the US Fatality Analysis Reporting System.

1 Motivating Example: an observational study and a counterclaim

1.1 Data from the Fatal Accident Reporting System 2010-2011

Do safety belts reduce injury and death in motor vehicle accidents? The example is built to resemble a compelling study by Evans (1986), but with more recent, detailed data from

[1] *Address for correspondence*: Paul R. Rosenbaum, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340 US, rosenbaum@wharton.upenn.edu, 23 April 2015. Supported by a grant from the MMS Program of the US National Science Foundation. This manuscript was published as https://www.tandfonline.com/doi/full/10.1080/01621459.2015.1054489: Rosenbaum, P. R. (2015). Some counterclaims undermine themselves in observational studies. *Journal of the American Statistical Association*, 110(512), 1389-1398.

the US Fatality Analysis Recording System (FARS) in 2010 and 2011. In the example, a counterclaim will undermine itself. Specifically, the counterclaim says that safety belts are without effect, that the association between injury and use of safety belts is produced by selection bias, that is, by the type of person who wears safety belts. Tentatively taking the counterclaim to be true justifies an analysis that would only strengthen the support that the original data provide for the original claim. Counterclaims are "plausible rival hypotheses"; see Rindskopf (2000). To say that a counterclaim undermines itself is to say it fails in its role as a plausible rival hypothesis.

Information about vehicle accidents with a fatality is recorded in FARS, including use of safety belts and a coarse measure of severity of injury or death. Less is recorded about events leading up to the crash, the speeds involved, the proximity of the cars before braking began, so less is known about the forces involved in the crash. Selection bias is possible. People who wear safety belts may drive more slowly, more cautiously than people who do not, may drive at a greater distance from the car ahead, may be sober more of the time, so it is possible that people who wear safety belts tend to be involved in less severe crashes.

Evan's idea repeated here compares a driver and a passenger of one vehicle in the same crash. Because they are in the same vehicle, their speed is the same, braking is the same, road traction is the same, the vehicle is the same, and some other aspects of the crash are similar even if not the same. In some crashes, the driver is belted and the passenger is not. In other crashes, the driver is unbelted and the passenger is belted. In most crashes, both are belted or both are unbelted. Evan's comparison removes many sources of selection bias, but it is still open to the counterclaim that selection biases exist inside a car when one person is belted and the other is not. The current section describes a basic analysis, while §1.3 and §3 will ask whether the counterclaim undermines itself.

The data consist of pairs of individuals, one in the drivers seat, one in the front passenger

seat, both at least 18 years of age from FARS 2010 and 2011. The comparison is of lap-shoulder belts (ls) or no belt restraints (n), so a pair is included only if the driver was ls-or-n and the passenger was ls-or-n, making four possible combinations for (driver.passenger) namely (ls.ls), (ls.n), (n.ls), (n.n). The injury scores are 0 for none, 1 for possible injury, 2 for nonincapacitating injury, 3 for incapacitating injury, and 4 for fatal injury. The driver-minus-passenger pair differences Y_i ranged from -4 to 4 in each restraint group. A value of 4 indicates the driver died and the passenger was uninjured.

1.2 A preliminary analysis of injury severity: four comparisons with sensitivity analyses

Figure 1 shows frequencies of driver-minus-passenger differences in injury scores for a driver and front passenger in the same vehicle in the same crash. Notably in Figure 1, when the driver is unbelted and the passenger is belted, the difference tends to be positive, the driver suffering more severe injuries. When the passenger is unbelted and the driver is belted, the difference tends to be negative, with the passenger suffering more severe injuries. When neither is belted, the distribution in Figure 1 looks symmetrical about zero. When both are belted, the distribution again looks fairly symmetrical, albeit more concentrated at zero, that is, the same injury category for passenger and driver.

To explain Figure 1 as a consequence of selection bias, the biases would have to be large, comparable in size to the bias that would explain away the effects of heavy smoking on lung cancer; see Rosenbaum (2002a, §4.3.2, Table 4.1). This is seen in Table 1. Table 1 gives counts of pairs, mean pair differences, standard errors of the mean, standard deviations of the differences, and two sensitivity analyses for testing the null hypothesis of no effect of safety belts. The sensitivity analysis uses the method in Rosenbaum (2007, 2013, 2014); see §2.3. The sensitivity analysis allows for a bias of $\Gamma \geq 1$ in assignment to treatment group in each type of pair, where $\Gamma = 1$ yields a paired randomization test. There are

two treatments in each type of pair, but the treatments vary from one type to another, e.g., belted driver versus belted passenger in the ls.ls group. The parameter Γ says that the two people in the front seat may not be equally likely to assume the two possible roles in the given type of pair, but rather one may have an odds of treatment that is Γ times greater than the other. If $\Gamma = 1$ then they have the same odds of treatment 1 rather than treatment 2, whereas $\Gamma = 2$ is a moderately large bias in which one person may have twice the odds of treatment 1 because these two people differ in some unmeasured way. Table 1 reports the maximum possible P-value testing no effect when allowance is made for a bias of magnitude Γ, so if this maximum P-value is less than α, conventionally $\alpha = 0.05$, then a bias of magnitude at most Γ could not lead to acceptance of the null hypothesis of no effect in an α-level test. To make Table 1 easier on the eyes, once the maximum possible P-value rounds to 1.0000, white space replaces repetition of 1.0000.

The first analysis in Table 1, labeled "Huber scores," is for the permutation distribution of one of Huber's M-statistics, as developed by Maritz (1979), for which a sensitivity analysis was developed in Rosenbaum (2007). The ψ-function in this test is Huber's ψ-function, namely $\psi(y) = \min\{1, \max(-1, y)\} = \text{sign}(y) \cdot \min(1, |y|)$, and it is applied to the scaled driver-minus-passenger pair difference in injury scores Y_i in vehicle i in Figure 1. The test statistic is $T = \sum_i \psi(Y_i/s)$, where s is 95% point of the $|Y_i|$, in parallel with Maritz (1979), so it is similar to a slightly trimmed mean. Use of the permutation distribution of the mean produces very similar results in this example.

Randomization tests — that is, tests at $\Gamma = 1$ — find driver-versus-passenger differences in all four groups in Table 1, ls.ls, n.n, ls.n, n.ls, so some of these differences are not effects of lap-shoulder belts; however, with a moderate bias of $\Gamma = 1.2$, the possible P-values in Table 1 for ls.ls and n.n range from near 0 to near 1 so these two comparisons may indicate either small biases or small differences between the safety of driver's and passenger's seats.

A quite large bias of $\Gamma = 4$ produces a maximum P-value of 0.0027 in two comparisons with different restraints for driver and passenger, ls.n and n.ls. The general impression is that injuries in the ls.ls and n.n are not the same but quite similar for driver and passenger; however, in the ls.n and n.ls comparisons, the belted individual typically suffers less severe injuries regardless of seat position, and a large selection bias would have to be present to explain away this difference. Moreover, the biases would have to be quite subtle in form: it cannot just be that the driver's seat is safer or that sturdy individuals drive; rather, it has to be that, regardless of seat position, wearing a lap-shoulder belt in either position is strongly associated with unmeasured additional safety not caused by the belt.

The second sensitivity in Table 1 uses a different ψ-function with inner trimming, as evaluated in Rosenbaum (2013). Theory suggests that statistics designed to optimize power against small treatment effects in large randomized experiments tend to exaggerate sensitivity to unmeasured biases in large observational studies. The ψ-function with inner trimming has $\psi(y) = \text{sign}(y) \cdot \max\left\{0, \min(1, |y|) - \frac{1}{4}\right\}$, so it ignores Y_i with small $|Y_i|/s$. As theory anticipates, the sensitivity analysis using a ψ-function with inner trimming reports greater insensitivity to unmeasured biases, insensitivity at $\Gamma = 5$ for both ls.n and n.ls, as opposed to $\Gamma = 4$ without inner trimming. In effect, inner trimming focuses attention on accidents in which one individual is much more severely injured than the other, saying that a large difference in injury scores more closely tracks restraint use than does a typical difference in injury score.

1.3 A counterclaim to the preliminary analysis: it's still just selection bias

A typical counterclaim says that the association between treatment and outcome is not an effect of the treatment but entirely a selection bias from comparing people who are not comparable. In FARS, there is selection bias in who does and who does not wear safety

belts, though much of it is eliminated by Evan's idea of comparing people in the same vehicle. For the people in Figure 1, the make of the car predicts belt use: there were 6522 people in a Ford of whom 29.0% were unbelted, and 2852 people in a Toyota of whom 20.0% were unbelted. People in Volvos and Mercedes are more likely to be belted than people in Fords. In Figure 1, people who did not wear lap-shoulder belts were on average about 9 years younger than those who did, and people between 18 and 30 were twice as likely as older individuals to not wear lap-shoulder belts. The safety of different vehicles is controlled by comparing two people in the same vehicle in Figure 1, but within one vehicle there could be selection bias in the decision to use a safety belt or not. Within the same vehicle, however, the mean age differences in the four groups in Table 1 are small: 0.36 years for ls.ls, .59 years for n.n, $-.98$ years for ls.n, and 1.34 years for n.ls.

The counterclaim says that when two people sit in the front seat, one belted, the other unbelted, there are selection biases in who is belted. It may say that a sick and frail individual, an intoxicated individual, or a severely obese individual is less likely than someone else to wear a lap-shoulder belt, and each of these individuals will suffer more severe injuries than robust individuals. An airbag does less to protect a severely obese individual, the counterclaim continues, and the punch of an airbag may seriously injure a frail individual, belted or not. In brief, the counterclaim says that a difference in injury scores reflects who wears safety belts, not any effect caused by wearing them.

As will be seen in §3, this counterclaim undermines itself. Where analysis of all pairs in Table 1 is insensitive to a bias of $\Gamma = 5$, an analysis that tentatively supposes the counterclaim to be true is insensitive to a bias of $\Gamma = 11$. How does this happen? An experiment or observational study may, and often does, look at effects in subgroups defined by covariates unaffected by the treatment. If instead, it looks at subgroups defined by affected outcomes, then these analyses may find effects where there are none or remove

actual effects (Rosenbaum 1984). The counterclaim says that the treatment is without effect; so, it is denying that certain quantities are affected, hence licensing their use as covariates unaffected by the treatment. In §3, an analysis focuses on injuries in pairs in which exactly one person was ejected from the vehicle. Someone who thought safety belts might have effects would regard that analysis as inappropriate, because wearing a belt might prevent ejection. However, if associations between outcomes and belt use were merely selection bias, then §2 shows this analysis would be appropriate. Were this analysis appropriate, it would be insensitive to a bias of $\Gamma = 11$. Hence, supposing the counterclaim to be true would justify an analysis that would strengthen the support that the original data provided in support of the orignal claim.

Section 2 defines a segment of data whose separate analysis is not usually justified but becomes justified if the counterclaim is tentatively supposed to be true. The segment may be a subset of the matched pairs, or with matched sets with many controls it may consists of some of the members of those sets; however, the segment is defined by outcomes, not by treatment received. Then, §3 reanalyzes the example from this perspective. Finally, §4 examines a simple model for treatment effects, showing that the pattern seen in the FARS example is the expected pattern under this model.

2 Analysis of the counterclaim of pure selection bias

2.1 Notation: treatment assignment, treatment effect, covariates

There are I matched sets, $i \in \{1, \ldots, I\} = \mathcal{I}$, where set $i \in \mathcal{I}$ contains subjects $\mathcal{J}_i = \{1, \ldots, J_i\}$, one treated with $Z_{ij} = 1$, the rest untreated controls with $Z_{ij} = 0$, so $1 = \sum_{j \in \mathcal{J}_i} Z_{ij}$ for each i. Write $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots Z_{IJ_I})^T$ for the vector of dimension $n = \sum_{i \in \mathcal{I}} J_i$ containing the treatment assignments, and let $\mathcal{Z}$ be the set containing the $\prod_{i \in \mathcal{I}} J_i$ possible values of $\mathbf{Z}$, so $\mathbf{z} \in \mathcal{Z}$ if $\mathbf{z}$ is of dimension n with $z_{ij} = 0$ or $z_{ij} = 1$ and $1 = \sum_{j \in \mathcal{J}_i} z_{ij}$

for each i. Denote by $|\mathcal{A}|$ the number of elements in a finite set $\mathcal{A}$ so that, for instance, $|\mathcal{J}_i| = J_i$ and $|\mathcal{Z}| = \prod_{i\in\mathcal{I}} J_i$. In §1.2, there are four types of matched sets, ls.ls, n.n, ls.n, n.ls, and this notation describes any one type, and every set $\mathcal{J}_i$ is a pair, $\mathcal{J}_i = \{1, 2\}$ and $J_i = 2$ for all i. Conditioning on the event $\mathbf{Z} \in \mathcal{Z}$ is abbreviated as conditioning on $\mathcal{Z}$. If $\mathbf{v}$ is a vector, then $\mathbf{v}^T$ is its transpose, and $\mathbf{v}'$ will mean something else defined in §2.4.

Subject ij has a measured covariate $\mathbf{x}_{ij}$ and an unmeasured covariate u_{ij}. Matching controlled $\mathbf{x}_{ij}$, so that $\mathbf{x}_{ij} = \mathbf{x}_{ik} = \mathbf{x}_i$, say, for each i, j, k, but quite possibly $u_{ij} \neq u_{ik}$ for many i, j, k. In §1.2, $\mathbf{x}_i$ indicates the vehicle and the crash.

Subject ij has two potential responses for the outcome of primary interest, r_{Tij} if assigned to treatment or r_{Cij} if assigned to control, so the observed response of ij is $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$ and the effect of the treatment on ij, namely $r_{Tij} - r_{Cij}$ is not observed; see Neyman (1923) and Rubin (1974). Fisher's (1935) sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tij} = r_{Cij}$ for all ij. In any one of the four types of pairings in §1.2, (r_{Tij}, r_{Cij}) records the injury scores that subject ij would suffer under the two possible treatments in that type of pairing, say belted driver or belted passenger for type ls.ls, R_{ij} is the injury ij actually suffered, and Fisher's H_0 says that swapping the treatments in pair i would not alter the injury suffered by individual ij. Write $\mathbf{R}$, $\mathbf{r}_C$, $\mathbf{r}_T$, and $\mathbf{u}$ for the n dimensional vectors. In addition, each subject may have a K-dimensional row vector of secondary outcomes, $\mathbf{s}_{Tij}$ or $\mathbf{s}_{Cij}$, with observed value $\mathbf{S}_{ij} = Z_{ij}\, \mathbf{s}_{Tij} + (1 - Z_{ij})\, \mathbf{s}_{Cij}$, and associated $n \times K$ matrices $\mathbf{S}$, $\mathbf{s}_C$ and $\mathbf{s}_T$ whose rows are in the lexical order, $i1$, $i2$, …, IJ_I. In the FARS example, the secondary outcome will describe aspects of the accident, such as whether an individual was ejected from the vehicle or direction of initial impact. Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{s}_{Tij}, \mathbf{s}_{Cij}, \mathbf{x}_{ij}, u_{ij}),\ i = 1, \ldots, I,\ j = 1, \ldots, J_i\}$. The subscripts ij are unique but noninformative identifiers, perhaps randomly assigned, and all information about individual ij is in observed or unobserved variables that describe ij. A matched pair,

$J_i = 2$, yields a single treated-minus-control pair difference $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$ in outcomes, as discussed in §1.2 and as plotted in Figure 1.

2.2 Randomization inference in randomized experiments

In a randomized experiment, one individual in each set is picked at random for treatment with independent selections in distinct matched sets, so that

$$\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \prod\nolimits_{i\in\mathcal{I}} J_i^{-1} = |\mathcal{Z}|^{-1} \text{ for each } \mathbf{z} \in \mathcal{Z}. \tag{1}$$

If $t(\mathbf{Z}, \mathbf{R})$ is a test statistic, then in a randomized experiment (1), the distribution of $t(\mathbf{Z}, \mathbf{R})$ under Fisher's null hypothesis H_0 of no effect equals its permutation distribution $\Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq k \mid \mathcal{F}, \mathcal{Z}\} = |\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq k\}| / |\mathcal{Z}|$, because $\mathbf{R} = \mathbf{r}_C$ when H_0 is true, $\mathbf{r}_C$ is fixed by conditioning on $\mathcal{F}$, and $\mathbf{Z}$ is uniform on $\mathcal{Z}$ in a randomized experiment. In an observational study, the counterclaim of selection bias says that the treatment is entirely without effect and $t(\mathbf{Z}, \mathbf{R})$ is large because (1) is false.

2.3 Sensitivity analysis in observational studies

A sensitivity analysis asks how large a departure from (1) would have to be present to materially alter the study's conclusions. One model says that, in the population prior to matching, treatment assignments are independent and two subjects with the same observed covariates may differ in their odds of treatment, $Z_{ij} = 1$, by at most a factor of Γ; then, the distribution of $\mathbf{Z}$ is returned to $\mathcal{Z}$ by conditioning on $\mathbf{Z} \in \mathcal{Z}$. This is equivalent to assuming that there is an unobserved covariate u_{ij} with $0 \leq u_{ij} \leq 1$ such that

$$\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \prod_{i\in\mathcal{I}} \frac{\exp\left(\gamma \sum_{j\in\mathcal{J}_i} z_{ij}\, u_{ij}\right)}{\sum_{j\in\mathcal{J}_i} \exp(\gamma\, u_{ij})} = \frac{\exp\left(\gamma \mathbf{z}^T \mathbf{u}\right)}{\sum_{\mathbf{b}\in\mathcal{Z}} \exp\left(\gamma \mathbf{b}^T \mathbf{u}\right)},\ \mathbf{u} \in [0,1]^n, \tag{2}$$

for each $\mathbf{z} \in \mathcal{Z}$, where $\gamma = \log(\Gamma) \geq 0$; see Rosenbaum (2002a, §4.2). For each $\mathbf{u} \in [0,1]^n$, the null distribution of $t(\mathbf{Z}, \mathbf{R})$ under Fisher's H_0 is obtained by summing terms (2) over $\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq k\}$. As $\mathbf{u}$ is allowed to range over $[0,1]^n$, the sensitivity analysis determines bounds on this null distribution, yielding for instance the upper bounds on P-values in Table 1. This method with point estimates and confidence intervals is implemented for M-statistics, including the permutational t-test, in the `sensitivitymv` and `sensitivitymw` packages in `R`, which determine a worst $\mathbf{u} \in [0,1]^n$ and approximate by a Normal distribution the distribution of $t(\mathbf{Z}, \mathbf{R})$ at this $\mathbf{u}$ under H_0; see Rosenbaum (2007, 2013, 2014) for description of this Normal approximation.

Various aspects of sensitivity analyses for observational studies are discussed by Cornfield et al. (1959), Egleston et al. (2009), Gastwirth (1992), Lin et al. (1998), Hernán and Robins (1999), Gilbert et al. (2003), Hosman et al. (2010), Hsu and Small (2013), Liu et al. (2013), Richardson et al. (2014), Robins et al. (2000), Rosenbaum and discussants (2002b), Scharfstein et al. (1999), and Yu and Gastwirth (2005).

2.4 Segments of the data determined by a matrix W

Informally, a segment consists of some of the individuals in the study. Formally, a segment of the data $\{\mathcal{J}_i, i \in \mathcal{I}\}$ in §2.1 is defined to be $\left\{\mathcal{J}_i^{'}, i \in \mathcal{I}\right\}$ where $\mathcal{J}_i^{'} \subseteq \mathcal{J}_i$ for each $i \in \mathcal{I}$. For example, if there are $n = 9$ subjects in matched triples, $\mathcal{J}_1 = \{1,2,3\}$, $\mathcal{J}_2 = \{1,2,3\}$, $\mathcal{J}_3 = \{1,2,3\}$, then one segment is $\mathcal{J}_1^{'} = \{2,3\}$, $\mathcal{J}_2^{'} = \emptyset$, $\mathcal{J}_3^{'} = \{1,2,3\}$. Let $\mathfrak{S}$ be the set whose 2^n elements are the 2^n possible segments.

For a segment $\left\{\mathcal{J}_i^{'}, i \in \mathcal{I}\right\}$, write m_i for the random variable that counts the number of treated subjects in $\mathcal{J}_i^{'}$, so $m_i = 0$ if $\mathcal{J}_i^{'} = \emptyset$ and otherwise $m_i = \sum_{j \in \mathcal{J}_i^{'}} Z_{ij}$, so $m_i = 0$ or $m_i = 1$. Write $\mathbf{m} = (m_1, \ldots, m_I)$. The contribution from $\mathcal{J}_i^{'}$ in segment $\left\{\mathcal{J}_i^{'}, i \in \mathcal{I}\right\}$ will be degenerate and uninteresting unless $m_i = 1 < \left|\mathcal{J}_i^{'}\right|$, that is, unless $\mathcal{J}_i^{'}$ contains

the treated subject and at least one control from matched set $\mathcal{J}_i$. For matched pairs, $|\mathcal{J}_i| = J_i = 2$ for all i as in §1.2, the interesting (or nondegenerate) part of a segment $\left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$ is simply a subset of the matched pairs. For matched sets with $|\mathcal{J}_i| = J_i > 2$, a segment $\left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$ may have nondegenerate parts $\mathcal{J}_i^{'}$ with $m_i = 1 < \left|\mathcal{J}_i^{'}\right| < |\mathcal{J}_i|$ containing the treated subject from $\mathcal{J}_i$ and some but not all of the controls from $\mathcal{J}_i$.

For a segment $\left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$, add a prime to a quantity to denote the value of a quantity confined to the segment. For instance, write $\mathbf{Z}'$ or $\mathbf{R}'$ for the vectors of dimension $n' = \sum_{i\in\mathcal{I}} \left|\mathcal{J}_i^{'}\right|$ containing, in the lexical order, the Z_{ij} or R_{ij} for $j \in \mathcal{J}_i^{'}$, $i \in \mathcal{I}$, and $\mathcal{Z}'_{\mathbf{m}}$ for the set of possible values of $\mathbf{Z}'$, that is, the set of vectors $\mathbf{z}'$ of dimension n' with 1 or 0 coordinates such that $m_i = \sum_{j\in\mathcal{J}_i^{'}} z_{ij}$. In parallel, write $\mathbf{r}_C^{'}$, $\mathbf{S}'$, etc. As before, conditioning on the event $\mathbf{Z}' \in \mathcal{Z}'_{\mathbf{m}}$ is abbreviated as conditioning on $\mathcal{Z}'_{\mathbf{m}}$, and generally the conditioning will be on $(\mathcal{Z},\, \mathcal{Z}'_{\mathbf{m}},\, \mathbf{m})$ jointly.

There is a matrix $\mathbf{W}$ with $M \geq 1$ columns and n rows $\mathbf{w}_{ij}$ in the lexical order describing the n subjects ij. Write $\mathcal{W}$ for the set of possible values for $\mathbf{W}$.

Definition 1 *The phrase "*$\mathbf{W}$ *determines the segment" means that there is a known function* $\mathcal{S}(\mathbf{W})$ *that receives* $\mathbf{W}$ *and returns a segment from* $\mathfrak{S}$*, that is,* $\mathcal{S} : \mathcal{W} \rightarrow \mathfrak{S}$*.*

To illustrate, in §3.1, there is $M = 1$ variable in $\mathbf{W}$ and the segment $\left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$ consists of the subset of matched pairs with different values for that one variable. In §3.2, there are $M = 2$ variables and the segment consists of the subset of matched pairs with different values for the first variable and the same value for the second variable. With matched sets, $J_i > 2$, a segment might use all sets but only some subjects in those sets.

Definition 1 needs to be guarded from a natural misinterpretation. The treated subject in set i is indicated by $Z_{ij} = 1$, that is, is the subject numbered $\sum_{j\in\mathcal{J}_i} j\, Z_{ij}$. Unless $\mathbf{W}$ includes $\mathbf{Z}$, a segment determined by $\mathbf{W}$ cannot make use of the identity of the treated

subject. In §3.1, the segment is the subset of pairs in which exactly one person was ejected from the vehicle, without reference to whether that individual was belted or not.

2.5 Segments and sensitivity analysis

When can we select a segment $\left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$ using $\mathbf{W}$, yet appropriately analyze this segment as if were an unselected data set? Proposition provides a condition on the segment $\mathcal{S}(\mathbf{W}) = \left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$ such that the distribution of treatment assignments in the segment is nothing more than a distribution with the same form as (2) confined to the segment by conditioning on $\mathbf{m}$.

Proposition 2 *If a segment* $\mathcal{S}(\mathbf{W}) = \left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$ *is determined by* $\mathbf{W}$ *in the sense of Definition 1, and if* $\mathbf{W}$ *is fixed by conditioning on* $\mathcal{F}$*, then (2) implies the distribution of treatment assignments* $\mathbf{z'} \in \mathcal{Z}'_{\mathbf{m}}$ *within the segment is given by*

$$\Pr\left(\mathbf{Z'} = \mathbf{z'} \mid \mathcal{F},\, \mathcal{Z},\, \mathcal{Z}'_{\mathbf{m}},\, \mathbf{m}\right) = \prod_{i \in \mathcal{I}:\left|\mathcal{J}_i^{'}\right|>0} \frac{\exp\left(\gamma \sum_{j \in \mathcal{J}_i^{'}} z_{ij}^{'}\, u_{ij}\right)}{\sum_{j \in \mathcal{J}_i^{'}} \exp\left(\gamma\, u_{ij}\right)},\ \mathbf{u'} \in [0,1]^{n'}\,. \tag{3}$$

Proof. The segment $\left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$ is fixed by conditioning on $\mathcal{F}$; moreover, the set $\mathcal{Z}'_{\mathbf{m}}$ is a fixed set as a consequence of conditioning on $\mathcal{Z}$ and $\mathbf{m}$. So the task is to show that conditioning on $\mathbf{m}$ in (2) yields the distribution in (3). It suffices to consider a single set i. If $\left|\mathcal{J}_i^{'}\right| = 0$, then Z_{ij} is not part of $\mathbf{Z'}$ and there is nothing to prove. If $\left|\mathcal{J}_i^{'}\right| = 1$, then the one Z_{ij} with $j \in \mathcal{J}_i^{'}$ is fixed by conditioning on $m_i = \sum_{j \in \mathcal{J}_i^{'}} Z_{ij}$ and the factor in (3) is 1. If $m_i = 0$, so $\mathcal{J}_i^{'}$ contains no treated subjects, then conditioning on $m_i = 0$ fixes Z_{ij} at 0 for each $j \in \mathcal{J}_i^{'}$, and the factor in (3) is 1. So, the only nondegenerate case has $\left|\mathcal{J}_i^{'}\right| \geq 2$ and $m_i = 1$. The final step is notationally cumbersome but mathematically elementary; it uses nothing more than the definition of conditional probability for a discrete random variable. For $\left|\mathcal{J}_i^{'}\right| \geq 2$ and $m_i = 1$, using (2), the conditional probability that $Z_{ij} = z_{ij}^{'}$

for $j \in \mathcal{J}_i^{'}$ given $\mathcal{F}$, $\mathcal{Z}$, $\mathcal{Z}'_{\mathbf{m}}$, $\mathbf{m}$ is the ratio of $\exp\left(\gamma \sum_{j\in\mathcal{J}_i} z_{ij}^{'} u_{ij}\right) / \sum_{j\in\mathcal{J}_i} \exp\left(\gamma\, u_{ij}\right)$ to the sum of similar terms over $j \in \mathcal{J}_i^{'}$, namely

$$\frac{\exp\left(\gamma \sum_{j\in\mathcal{J}_i} z_{ij}^{'} u_{ij}\right) / \sum_{j\in\mathcal{J}_i} \exp\left(\gamma\, u_{ij}\right)}{\sum_{j\in\mathcal{J}_i^{'}} \left\{\exp\left(\gamma\, u_{ij}\right) / \sum_{j\in\mathcal{J}_i} \exp\left(\gamma\, u_{ij}\right)\right\}} = \frac{\exp\left(\gamma \sum_{j\in\mathcal{J}_i^{'}} z_{ij}^{'} u_{ij}\right)}{\sum_{j\in\mathcal{J}_i^{'}} \exp\left(\gamma\, u_{ij}\right)}$$

as in (2). ■

Corollary 3 *If a segment* $\mathcal{S}(\mathbf{S}) = \left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$ *is determined by the observed value of the supplementary responses* $\mathbf{S}$*, and if the supplementary responses are unaffected by the treatment,* $\mathbf{s}_{Tij} = \mathbf{s}_{Cij}$ *for all* ij*, then (2) implies the distribution of treatment assignments within the segment is given by (3).*

Proof. If $(\mathbf{s}_{Tij}, \mathbf{s}_{Cij})$ is unaffected, then $\mathbf{S}_{ij}$ equals $\mathbf{s}_{Cij}$, so $\mathbf{S}$ is fixed by conditioning on $\mathcal{F}$. Apply Proposition 2. ■

2.6 The logic behind the statistical analysis of counterclaims

Suppose that an investigator had tested Fisher's null hypothesis, $H_0 : r_{Tij} = r_{Cij}$ for all ij, and found that rejection of this H_0 is insensitive to a certain magnitude Γ of bias. Suppose that the investigator believes that the treatment affects the outcomes of interest (r_{Tij}, r_{Cij}), and may also affect the supplementary outcomes $(\mathbf{s}_{Tij}, \mathbf{s}_{Cij})$. In this case, the investigator cannot test H_0 again on a segment $\mathcal{S}(\mathbf{S}) = \left\{\mathcal{J}_i^{'},\, i \in \mathcal{I}\right\}$ determined by the observed values $\mathbf{S}$ of the supplemental outcomes. Believing her claim, the investigator cannot do this because if $\mathbf{s}_{Tij} \neq \mathbf{s}_{Cij}$ then $\mathbf{S}_{ij}$ changes as Z_{ij} changes, so acting as if the segment were fixed is not at all the same as merely conditioning on $(\mathbf{m}, \mathcal{Z}'_{\mathbf{m}})$ in (3): change the treatment assignment $\mathbf{Z}$ and the segment $\mathcal{S}(\mathbf{S})$ might change.

Now a critic makes a counterclaim asserting that the treatment has no effect of any kind and associations are all produced by selection bias, the type of person who gets treated.

This counterclaim asserts $r_{Tij} = r_{Cij}$ and $\mathbf{s}_{Tij} = \mathbf{s}_{Cij}$ for all ij. If the counterclaim were true, then Corollary 3 would permit us to focus on a segment $\mathcal{S}(\mathbf{S}) = \left\{\mathcal{J}_i', i \in \mathcal{I}\right\}$ determined by $\mathbf{S}$, and to test Fisher's $H_0 : r_{Tij} = r_{Cij}$ for all ij by confining attention to this segment. Importantly, this is still a test of no effect on the outcome (r_{Tij}, r_{Cij}) of primary interest, injury severity in §1.2. If based on (3) this test of $H_0 : r_{Tij} = r_{Cij}$ for all ij turned out to be insensitive to a much larger bias $\Gamma' > \Gamma$, then the critic's counterclaim has undermined itself. The investigator had acknowledged that her claim was sensitive to biases larger than Γ, but add the critic's counterclaim and the investigator's claim becomes insensitive to a larger bias $\Gamma' > \Gamma$; that is, the counterclaim fails in its role as a critique of the original claim, because presuming the counterclaim to be true would only strengthen the support that the original data provide for the original claim.

The critic might engage in a tactical retreat, saying that the treatment has no effect on (r_{Tij}, r_{Cij}) — that is just selection bias — but perhaps, yes, the treatment does indeed have an effect on $(\mathbf{s}_{Tij}, \mathbf{s}_{Cij})$. In that case, Corollary 3 could not be invoked to perform the analysis confined to a segment $\mathcal{S}(\mathbf{S}) = \left\{\mathcal{J}_i', i \in \mathcal{I}\right\}$. If the supplemental responses $(\mathbf{s}_{Tij}, \mathbf{s}_{Cij})$ are of doubtful relevance, then perhaps the tactical retreat would work, would seem to offer a credible explanation of the association between R_{ij} and Z_{ij}. On the other hand, the critic's concession that the treatment does affect $(\mathbf{s}_{Tij}, \mathbf{s}_{Cij})$ may be quite a large concession. In some contexts, it might be quite implausible that the treatment genuinely and materially affects $(\mathbf{s}_{Tij}, \mathbf{s}_{Cij})$, the treatment Z_{ij} is strongly associated with R_{ij}, yet the treatment has no effect on (r_{Tij}, r_{Cij}). So the critic is forced to admit that either the treatment does affect $(\mathbf{s}_{Tij}, \mathbf{s}_{Cij})$ or the strengthened sensitivity analysis based on the segment is justified, and both sides of this either-or may strengthen the original claim.

An investigator making a claim need not wait for an actual critic to make a counterclaim. That investigator may say: if the following counterclaim were made, the following

analysis would be appropriate, and so a counterclaim of that form would undermine itself.

This same fact may be expressed in different terms, omitting the story of an investigator and a critic. A certain analysis may demonstrate the following, perhaps interesting, fact: a selection bias of magnitude Γ could explain the observed association between treatment Z_{ij} and response R_{ij} as a bias and not a causal effect on (r_{Tij}, r_{Cij}), but only if there is a causal effect on $(\mathbf{s}_{Tij}, \mathbf{s}_{Cij})$; however, if there is no effect on both (r_{Tij}, r_{Cij}) and $(\mathbf{s}_{Tij}, \mathbf{s}_{Cij})$, then the bias Γ' needed to explain the observed association treatment Z_{ij} and response R_{ij} is much larger, $\Gamma' > \Gamma$. Whether this fact strikes us as interesting and consequential will depend on how plausible it seems that the treatment could genuinely affect $(\mathbf{s}_{Tij}, \mathbf{s}_{Cij})$ while not affecting (r_{Tij}, r_{Cij}).

3 Counterclaim analysis in the FARS data

3.1 Ejection from the vehicle

In some crashes, a person is ejected from the vehicle. The first coordinates of $\mathbf{s}_{Tij}$ and $\mathbf{s}_{Cij}$ describe ejection. Write $s_{Tij1} = 1$ if ij would be ejected if in the treated condition in pair i, $s_{Tij1} = 0$ otherwise, and write $s_{Cij1} = 1$ if ij would be ejected if in the control condition in pair ij, $s_{Cij1} = 0$ otherwise, so that S_{ij1} is the observed ejection status, 1 or 0, as in §2.1. If belts affected safety, they might prevent some ejections, $s_{Tij1} \neq s_{Cij1}$ for some ij, thereby possibly reducing some injury scores in the process, so $r_{Tij} \neq r_{Cij}$ for some ij; however, the counterclaim in §1.3 denies all this, saying the association between treatment and outcome is spurious, produced entirely by selection bias, by the type of person who does not wear seatbelts. If the counterclaim were true, then $r_{Tij} = r_{Cij} = R_{ij}$ and $s_{Tij1} = s_{Cij1} = S_{ij1}$ and by Corollary 1, this would justify an analysis focused on a segment $\mathcal{S}(\mathbf{S}) = \left\{\mathcal{J}_i', i \in \mathcal{I}\right\}$ determined by $\mathbf{S}$. Here, the segment is the set of pairs in which exactly one person, driver or passenger, was ejected, so $\mathcal{S}(\mathbf{S}) = \left\{\mathcal{J}_i', i \in \mathcal{I}\right\}$ with

$\mathcal{J}_i^{'} = \{1, 2\}$ if $S_{i11} \neq S_{i21}$ and $\mathcal{J}_i^{'} = \emptyset$ if $S_{i11} = S_{i21}$.

Figure 2 parallels Figure 1, but only for the subset of 2048 pairs with precisely one ejection from the vehicle. The association between injury and belt use looks stronger in Figure 2 than in Figure 1. Table 2 repeats the analysis in Table 1 for the 2048 pairs with precisely one ejection. Despite the dramatic reduction in sample size, Table 2 exhibits far less sensitivity to unmeasured bias than does Table 1. In this sense, the counterclaim of pure selection bias undermines itself. If it were true that ejection from the vehicle was unaffected by treatment category, then the magnitude of unmeasured bias needed to explain the higher injury scores of unbelted individuals in pairs with one ejection is greater than $\Gamma = 11$ in Table 2 for scores using inner trimming.

Importantly, Table 2 continues to refer to the effect of safety belts on injury severity. The contrast between Table 1 and Table 2 says: Injury severity more closely tracks safety belt use in crashes in which exactly one person was ejected.

As in §2.6, an advocate of the counterclaim could retreat to say that lap-shoulder belts do prevent ejections from the vehicle, but preventing ejection is without consequence for injury scores. This retreat would invalidate Table 2, but not Table 1; however, it is not a small retreat. The position that belt use is strongly associated with injury, that it causes a reduction in ejections, but that belt use has no effect on injury is a logical possibility — it is not self-contradictory — but not a particularly plausible possibility.

3.2 Eliminating sources of heterogeneity: the direction of initial impact

In some vehicle crashes, the initial vehicle impact is from the front or rear, while in others it is from a side, perhaps at an angle. Sivak et al. (2006) considered the possibility that some side crashes are caused by a lack of visibility during lane changes. They compared two-door and four-door models of the same make of car when the B-pillar that divides

front and rear seats is further back in two-door models, thereby improving driver visibility out the side windows. They claimed that such two-door models had fewer side crashes. Sivak et al. (2006) view the direction of the crash as an outcome. Could the direction of the crash be an outcome affected safety belts? It is conceivable that some belted drivers experience limitation of side visibility due to restrictions on movement.

The second coordinates, s_{Tij2} and s_{Cij2}, of $\mathbf{s}_{Tij}$ and $\mathbf{s}_{Cij}$ indicate whether the vehicle was hit from the side, with 0 indicating a vehicle with an initial hit known to be from the side, 1 indicating all other vehicles. The direction of initial vehicle impact is the same for two people, $i1$ and $i2$, in the same vehicle, $S_{i12} = S_{i22}$; however, it is possible that $s_{Tij2} \neq s_{Cij2}$. The counterclaim that safety belts are without safety effects entails $\mathbf{s}_{Tij} = \mathbf{s}_{Cij}$, so by Corollary 3 it would justify analyses on segments determined by $\mathbf{S}$.

Perhaps one reason that a belted driver is more severely injured than an unbelted passenger is that the initial impact came from the left, the driver's side. Perhaps one reason that a belted passenger is more severely injured than an unbelted driver is that the initial impact came from the right, the passenger's side. Perhaps part of the heterogeneity in differences Y_i in injury scores comes from impacts on the left or the right. Reduced heterogeneity of pair differences Y_i reduces sensitivity to unmeasured biases (Rosenbaum 2005). This motivates considering crashes not known to be from the side.

If the analysis in Table 1 is repeated excluding the 5783 vehicles that were known to have an initial impact on the side, the results are somewhat less sensitive. For the test statistic with inner trimming, the upper bounds on the P-values for ls.n and n.ls are, respectively, 0.0000 and 0.0126 at $\Gamma = 5.5$, and 0.0002 and 0.0513 at $\Gamma = 6$. Similarly, if the analysis in Table 2 is repeated including only the 1383 pairs with both exactly one ejected individual and without a known initial impact from the side, then the results are much less sensitive to unmeasured biases, with upper bounds on the P-values for ls.n and

n.ls of 0.0129 and 0.0439 at $\Gamma = 15$ when inner trimming is used.

3.3 Use of a segment or its complement

The complement of a segment is the data excluded in forming the segment, and it is another segment. In the safety belt example, the complements are of no obvious interest. In some other context, it might be unclear whether a segment or its complement is of greater interest, so that both would be examined. One way to do this is using the truncated product of P-values introduced by Zaykin et al. (2002) to generalize Fisher's method for combining independent P-values. The truncated product is the product of those P-values that are less than or equal to a certain prespecified level, say $\widetilde{\alpha} = 0.2$ or $\widetilde{\alpha} = 0.1$, and defined to equal 1 if all P-values are greater than $\widetilde{\alpha}$. Fisher's method equals the truncated product with $\widetilde{\alpha}$ set to 1. With a segment and its complement, there are two P-values conditionally independent given $\mathbf{m}$. As shown by Hsu et al. (2013), the truncated product is more powerful than Fisher's method when used in sensitivity analyses because it eliminates without much cost the large P-value bounds often seen in sensitivity analyses.

4 Properties of the counterclaim analysis in a simple case

4.1 An alternative hypothesis: a simple model for a treatment effect without bias

When do we expect a segment to report greater insensitivity to bias? The behavior of a sensitivity analysis for a test of no effect H_0 will be examined in matched pairs when H_0 is false and there is no bias from unmeasured covariates so (1) is true; that is, the critic is entirely mistaken, and the observed association between treatment and outcome is what it naively appears to be, namely a treatment effect. There is one binary supplemental response, $s_{Ti1} = 0$ or 1 and $s_{Cij} = 0$ or 1, that is positively affected by the treatment, so $s_{Tij} \geq s_{Cij}$ and with strict inequality for some ij. Neither the investigator nor the critic

know that these facts, so they continue to disagree.

The treatment effect has $r_{Tij} = r_{Cij} + \tau + \beta\,(s_{Tij} - s_{Cij})$ with $\tau \geq 0$ and $\beta \geq 0$, so the effect is τ if $s_{Tij} = s_{Cij}$ or is $\tau + \beta$ if $s_{Tij} = 1 > 0 = s_{Cij}$. If $\tau = 0$, then there is an effect on (r_{Tij}, r_{Cij}) only if there is an effect on (s_{Tij}, s_{Cij}), so the model satisfies the exclusion restriction for an instrumental variable. If $\beta = 0$, then the effect is constant. This model is a simple form of mediation effect; see, for instance, Imai et al. (2010).

Under this model for effect, $Y_i = (Z_{i1} - Z_{i2})\,(R_{i1} - R_{i2}) = \tau + \beta\delta_i + \varepsilon_i$ where both $\delta_i = Z_{i1}\,(s_{Ti1} - s_{Ci1}) + Z_{i2}\,(s_{Ti2} - s_{Ci2})$ and $\varepsilon_i = (2Z_{i1} - 1)\,(r_{Ci1} - r_{Ci2})$ are not observed. We observe $\Delta_i = (Z_{i1} - Z_{i2})\,(S_{i1} - S_{i2}) = Z_{i1}\,(s_{Ti1} - s_{Ci2}) + Z_{i2}\,(s_{Ti2} - s_{Ci1})$, but δ_i and Δ_i may often differ. The strategy in §3.1 was to limit attention to pairs i with $|\Delta_i| = 1$, that is, to pairs in which exactly one person was ejected.

The alternative hypothesis is further simplified by assuming that the $2I$ values of (s_{Tij}, s_{Cij}) were obtained as $2I$ independent and identically (iid) distributed observations, independent of the rest of $\mathcal{F}$, sampled from a multinomial distribution with $\Pr\,(s_{Tij} = a,\ s_{Cij} = b) = \pi_{ab}$, for $ab = 00,\ 10,\ 11$, where $1 = \pi_{11} + \pi_{10} + \pi_{00}$. This implies (s_{Tij}, s_{Cij}) is independent of ε_i. It follows that $\mathrm{E}\,(Y_i) = \tau + \beta\pi_{10}$ and $\mathrm{var}\,(Y_i) = \beta^2\pi_{10}\,(1 - \pi_{10}) + \mathrm{var}\,(\varepsilon_i)$. The counterclaim analysis in §3.1 looked only at Y_i with $|\Delta_i| = 1$. Then $\mathrm{E}\left(Y_i \;\middle|\; |\Delta_i| = 1\right) = \tau + \lambda\beta$ and $\mathrm{var}\left(Y_i \;\middle|\; |\Delta_i| = 1\right) = \beta^2\,\lambda\,(1 - \lambda) + \mathrm{var}\,(\varepsilon_i)$, where $\lambda = \Pr\left(\delta_i = 1 \;\middle|\; |\Delta_i| = 1\right) = \{\pi_{10}\,(1 - \pi_{11})\}\,/\,\{\pi_{00}\pi_{11} + (1 - \pi_{00})\,(1 - \pi_{11})\}$. Pairs with $|\Delta_i| = 1$ have $\mathrm{E}\,(Y_i \mid |\Delta_i| = 1) \geq \mathrm{E}\,(Y_i)$ if either $\pi_{11} \leq \frac{1}{2}$ or $\pi_{00} = 0$. If $\mathrm{E}\,(Y_i \mid |\Delta_i| = 1) < \mathrm{E}\,(Y_i)$, then one should focus on the complement with $|\Delta_i| = 0$. In uncertain cases, the truncated product calculation in §3.3 uses both the segment and its complement.

4.2 Design sensitivity: limiting sensitivity to bias in large samples

A level-α sensitivity analysis rejects the null hypothesis H_0 of no treatment effect in the presence of a bias of at most Γ if the upper bound on the P-value is at most α when computed with this Γ; see, for instance, Table 1. The power of a level-α sensitivity analysis is the probability of rejection, that is, the probability that the upper bound on the P-value will be less than or equal to α when computed with a particular Γ. The power is computed under an alternative hypothesis, such as the model in §4.1, in which there is a treatment effect but there is no unmeasured bias. For $\Gamma = 1$, the power of a sensitivity analysis is the same as the power of a randomization test in a randomized experiment.

As the number of pairs increases, $I \rightarrow \infty$, there is a limiting sensitivity to unmeasured biases called the design sensitivity, $\widetilde{\Gamma}$. More precisely, as $I \rightarrow \infty$ in a model with a treatment effect but no unobserved biases, such as the model in §4.1, the power of a sensitivity analysis tends to 1 for $\Gamma < \widetilde{\Gamma}$ and to 0 for $\Gamma > \widetilde{\Gamma}$. If the sample size were large enough, the study would be insensitive to all biases smaller than $\widetilde{\Gamma}$ and sensitive to some biases larger than $\widetilde{\Gamma}$. For detailed discussion of design sensitivity, see Rosenbaum (2005, 2013, 2014, 2015), Stuart and Hanna (2013) and Zubizarreta et al. (2013).

A formula for the design sensitivity $\widetilde{\Gamma}$ of M-tests in matched pairs is given in Rosenbaum (2013, Corollary 1). The formula is the ratio of two integrals. The integrals may be determined with negligible error by monte carlo integration. Table 3 reports the results for various models from §4.1 for the two M-tests used in Table 1. Each value in Table 3 is based on sampling ten million Y_i for the relevant sampling situation from §4.1.

In Table 3, τ and β are adjusted so that the expected pair difference in outcomes is $\mathrm{E}(Y_i) = \frac{1}{2}$, that is, $\beta = \left(\frac{1}{2} - \tau\right)/\pi_{10}$. If $\tau = 0$, then a treatment effect occurs only in pairs with $(s_{Tij}, s_{Cij}) = (1, 0)$, whereas if $\tau = \frac{1}{2}$ then the treatment effect is always $\frac{1}{2}$. If $\tau = \frac{1}{4}$, then there is always a nonzero effect, but the effect is larger when $(s_{Tij}, s_{Cij}) = (1, 0)$. The

errors ε_i are Normal or logistic, symmetric about zero, scaled so that $\text{var}(\varepsilon_i) = 1$. There are two patterns for the $(\pi_{11}, \pi_{10}, \pi_{00})$.

Consistent with results in Rosenbaum (2013), the design sensitivities are generally larger with inner trimming, so focus on the last three columns of Table 3 where inner trimming is used. The design sensitivity $\widetilde{\Gamma} = 3.8$ in the first row of Table 3 says that if all I pairs were analyzed in this setting with $\tau = 0$, $\pi_{10} = 1/3$ and Normal errors, then for sufficiently large I the results would be sensitive to a bias $\Gamma > 3.8$ and insensitive to a bias $\Gamma < 3.8$. The segment limits attention to the subset of pairs with $1 = |\Delta_i| = |(Z_{i1} - Z_{i2})(S_{i1} - S_{i2})|$, as in §3.1 where attention focused on vehicles with exactly one person ejected. The design sensitivity for this segment is larger, $\widetilde{\Gamma} = 4.9$. If the sensitivity analysis were performed at $\Gamma = 4$, then as $I \rightarrow \infty$ the power is tending to zero using all pairs but is tending to 1 using the segment. As one would expect, when the treatment effect is constant, not dependent on (s_{Tij}, s_{Cij}), that is when $\tau = \frac{1}{2}$ in Table 3, the design sensitivity is the same in all pairs, in the segment, and in the complement to the segment.

Table 3 refers to the limit as $I \rightarrow \infty$, but for finite I the analysis using all pairs is using more pairs than the analysis confined to the segment. For instance, in the $\tau = 1/2$ rows, the design sensitivity is the same for all pairs and for the segment, but the sample size is larger using all pairs. From §4.1, a focus on the segment can reduce $\text{E}(Y_i \mid |\Delta_i| = 1)$ if $\pi_{11} > \frac{1}{2}$, and in this case we should be looking at the complement of the segment. As shown by Hsu et al. (2013), the truncated product from §3.3 will have design sensitivity equal to the maximum of the design sensitivities for the segment and the complement.

4.3 Simulation: How do the asymptotic comparisons hold up in finite samples?

The simulation has $I = 2000$ pairs with the structure in Table 3. Each situation is replicated 3000 times, so the standard error of an estimated power is at most $\sqrt{.25/3000} \leq$

0.01. Power is evaluated at $\Gamma = 4$ for the M-test with inner trimming. Table 4 reports the average number of pairs in the segment, I_{Seg}, so the complement has $I - I_{Seg}$ pairs on average. The power estimates are the proportions of the 3000 replicates that produced an upper bound on the one-sided P-value less than or equal to 0.05. When Γ in Table 4 exceeds the design sensitivity $\widetilde{\Gamma}$ in Table 3, the simulated power in Table 4 is negligible, consistent with asymptotic theory. In Table 4, focusing on the segment rather than all pairs increases the power of the sensitivity analysis when $\tau = 0$ or $\tau = \frac{1}{4}$ and decreases it when $\tau = \frac{1}{2}$, consistent with Table 3.

As discussed in §3.3, the truncated product of the two P-values uses all pairs by combining two tests. In Table 4, truncation is at $\widetilde{\alpha} = 0.2$ and the truncated product is never much worse than "All", and is often much better. This is consistent with the fact from Hsu et al. (2013) that the truncated product has design sensitivity equal to the maximum of the design sensitivities for segment and complement. If $\tau = \frac{1}{2}$, the effect and design sensitivity are the same in the segment and the complement, and it is marginally better to use all pairs rather than the truncated product, but the difference is not large.

For the simple model in §4.1, both the asymptotic calculation in Table 3 and the simulation in Table 4 yield a pattern similar to that seen in the FARS data.

5 Summary: counterclaims should be subjected to empirical analysis

Some counterclaims deny certain treatment effects, and if these counterclaims were true they would justify certain otherwise unjustified analyses. The additional analyses tentatively presume the counterclaim to be true, acting as if certain outcomes were unaffected. By definition, the counterclaim undermines itself if, in these additional analyses, the investigator's original claim becomes insensitive to a larger bias $\Gamma' > \Gamma$; that is, presuming the counterclaim to be true would only strengthen support that the data provide for the

original claim. A counterclaim that undermines itself fails in role as a counterclaim: presuming the counterclaim to be true would not weaken, but would instead strengthen the support that the original data provide for the original claim.

References

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., Wynder, E. (1959), "Smoking and lung cancer," *Journal of the National Cancer Institute*, 22, 173-203.

Egleston, B. L., Scharfstein, D. O. and MacKenzie, E. (2009), "On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death," *Biometrics*, 65, 497-504.

Evans, L. (1986), "The effectiveness of safety belts in preventing fatalities," *Accident Analysis and Prevention*, 18, 229-241.

Fatality Analysis Recording System, National Highway Traffic Safety Administration.

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

Gastwirth, J. L. (1992), "Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables," *Jurimetrics* **33** 19-34.

Gilbert, P., Bosch, R., Hudgens, M. (2003), "Sensitivity analysis for the assessment of the causal vaccine effects on viral load in HIV vaccine trials," *Biometrics*, 59, 531-41.

Hernán, M. A. and Robins, J. M. (1999), "Letter to the editor," *Biometrics*, 55, 1316-1317.

Hosman, C. A., Hansen, B. B. and Holland, P. W. H. (2010), "The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder," *Annals of Applied Statistics*, 4, 849-870.

Hsu, J. Y. and Small, D. S. (2013), "Calibrating sensitivity analyses to observed covariates in observational studies," *Biometrics*, 69, 803-811.

Hsu, J. Y., Small, D. S. and Rosenbaum, P. R. (2013), "Effect modification and design

sensitivity in observational studies," *Journal of the American Statistical Association*, 108, 135-148.

Imai, K., Keele, L. and Yamamoto, T. (2010), "Identification, inference and sensitivity analysis for causal mediation effects," *Statistical Science*, 25, 51-71.

Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998), "Assessing the sensitivity of regression results to unmeasured confounders in observational studies," *Biometrics*, 54, 948-963.

Liu, W., Kuramoto, J. and Stuart, E. (2013), "Sensitivity analysis for unobserved confounding in nonexperimental prevention research," *Prevention Science*, 14, 570-580.

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

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

Richardson, A., Hudgens, M. G., Gilbert, P. B., and Fine, J. P. (2014), "Nonparametric bounds and sensitivity analysis of treatment effects," *Statistical Science*, 29, 596-618.

Rindskopf, D. (2000), "Plausible rival hypotheses in measurement, design, and scientific theory," in: L. Bickman, ed., *Research Design*, Thousand Oaks, CA: Sage, pp. 1-12.

Robins, J., Rotnitzky, A., Scharfstein, D. (2000), "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models," In: *Statistical models in epidemiology, the environment, and clinical trials*, Springer: NY, 1-94.

Rosenbaum, P. R. (1984), "The consequences 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. (2002a), *Observational Studies* (2^{nd} edition), New York: Springer.

Rosenbaum, P. R. (2002b), "Covariance adjustment in randomized experiments and observational studies (with Discussion)," *Statistical Science*, 17, 286-327.

Rosenbaum, P. R. (2005), "Heterogeneity and causality," *American Statistician*, 59, 147-52.

Rosenbaum, P. R. (2007), "Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies," *Biometrics*, 63, 456-64. (`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. (2014), "Weighted M-statistics with superior design sensitivity in matched observational studies with multiple controls," *Journal of the American Statistical Association*, 109, 1145-1158. (`R` package `sensitivitymw`)

Rosenbaum, P. R. (2015), "Bahadur efficiency of sensitivity analyses in observational studies," *Journal of the American Statistical Association*, 110, 205-217.

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

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999), "Adjusting for non-ignorable drop-out using semiparametric non-response models," *Journal of the American Statistical Association*, 94, 1096-1120.

Sivak, M., Schoettle, B., Reed, M., Flannagan, M. (2006), "Influence of visibility out of the vehicle cabin on lane-change crashes," *Accident Analysis and Prevention*, 38, 969-72.

Stuart, E. A. and Hanna, D. B. (2013), "Should epidemiologists be more sensitive to design sensitivity?" *Epidemiology*, 24, 88-89.

Yu, B. B., Gastwirth, J. L. (2005), "Sensitivity analysis for trend tests: application to the risk of radiation exposure. *Biostatistics*, 6, 201-209.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002), "Trucated product method of combining P-values," *Genetic Epidemiology*, 22, 170-185.

Zubizarreta, J. R., Cerdá, M. and Rosenbaum, P. R. (2013), "Effect of the 2010 Chilean earthquake on posttraumatic stress," *Epidemiology*, 24, 79-87.

Table 1: Analysis of driver-minus-passenger pair differences in injury scores by restraint group. Values are upper bounds on P-values. n = no restraint. ls = lap-shoulder belt.

	Restraint Use: (driver.passenger)			
	Same Use		Different Use	
Restraint Group	ls.ls	n.n	ls.n	n.ls
Number of Pairs	10996	3274	1412	1198
Mean	-0.059	0.061	-1.076	1.000
Standard error	0.013	0.027	0.042	0.044
Standard deviation	1.335	1.571	1.565	1.513
Γ	Huber Scores without Inner Trimming			
1	0.0000	0.0241	0.0000	0.0000
1.2	1.0000	1.0000	0.0000	0.0000
4			0.0000	0.0027
5			0.0211	0.4673
5.5			0.1808	1.0000
Γ	Inner Trimmed Scores			
1	0.0000	0.0374	0.0000	0.0000
1.2	1.0000	1.0000	0.0000	0.0000
5			0.0000	0.0125
6			0.0031	0.2219
6.5			0.0160	0.5058

6 Tables

Table 2: Renalysis using only 2048 pairs in which exactly one person was ejected from the vehicle. Values are upper bounds on P-values.

	Restraint Use: (driver.passenger)			
	Same Use		Different Use	
Restraint Group	ls.ls	n.n	ls.n	n.ls
Number of Pairs	222	782	522	522
Mean	-0.023	0.141	-1.540	1.584
Standard error	0.117	0.069	0.064	0.057
Standard deviation	1.748	1.938	1.455	1.291
Γ	Huber Scores without Inner Trimming			
1	0.7436	0.0428	0.0000	0.0000
1.2	1.0000	1.0000	0.0000	0.0000
9			0.0388	0.0009
11			0.2783	0.0149
Γ	Inner Trimmed Scores			
1	0.9002	0.0764	0.0000	0.0000
1.2	1.0000	0.8737	0.0000	0.0000
9			0.0047	0.0004
11			0.0322	0.0040

Table 3: Design sensitivities using all pairs (All), the segment (Seg), and its complement (Comp), without or with inner trimming. The largest design sensitivities in each row are in **bold**.

Model			No inner trim			With inner trim		
Distribution	τ	$(\pi_{11}, \pi_{10}, \pi_{00})$	All	Seg	Comp	All	Seg	Comp
Normal	0	(1/3, 1/3, 1/3)	2.7	3.3	2.2	3.8	**4.9**	2.8
Normal	1/4	(1/3, 1/3, 1/3)	3.2	3.6	2.8	4.4	**5.1**	3.7
Normal	1/2	(1/3, 1/3, 1/3)	3.4	3.4	3.4	**4.7**	**4.7**	**4.7**
Logistic	0	(1/3, 1/3, 1/3)	2.8	3.3	2.2	3.9	**4.9**	2.9
Logistic	1/4	(1/3, 1/3, 1/3)	3.3	3.7	2.8	4.4	**5.0**	3.7
Logistic	1/2	(1/3, 1/3, 1/3)	3.5	3.5	3.5	**4.7**	**4.7**	**4.7**
Normal	0	(1/4, 1/2, 1/4)	3.0	3.8	2.1	4.0	**5.3**	2.5
Normal	1/4	(1/4, 1/2, 1/4)	3.3	3.8	2.7	4.5	**5.3**	3.5
Normal	1/2	(1/4, 1/2, 1/4)	3.5	3.5	3.5	**4.8**	**4.8**	**4.8**
Logistic	0	(1/4, 1/2, 1/4)	3.0	3.8	2.1	4.0	**5.3**	2.6
Logistic	1/4	(1/4, 1/2, 1/4)	3.4	3.9	2.8	4.5	**5.2**	3.5
Logistic	1/2	(1/4, 1/2, 1/4)	3.5	3.5	3.5	**4.7**	**4.7**	**4.7**

Table 4: Power of a 0.05-level sensitivity analysis at $\Gamma = 4$ under 12 models, using all $I = 2000$ pairs (All), the segment (Seg), its complement (Comp), and the truncated product (Tprod) based on both the segment and its complement, using inner trimming.

Distribution	τ	I_{Seg}	All	Seg	Comp	Tprod
$(\pi_{11}, \pi_{10}, \pi_{00}) = (1/3, 1/3, 1/3)$						
Normal	0	1111	0.01	0.48	0.00	0.22
Normal	1/4	1111	0.24	0.62	0.01	0.38
Normal	1/2	1111	0.61	0.40	0.33	0.55
Logistic	0	1111	0.02	0.50	0.00	0.25
Logistic	1/4	1111	0.22	0.57	0.01	0.34
Logistic	1/2	1111	0.53	0.34	0.29	0.47
$(\pi_{11}, \pi_{10}, \pi_{00}) = (1/4, 1/2, 1/4)$						
Normal	0	1250	0.04	0.82	0.00	0.60
Normal	1/4	1250	0.39	0.80	0.00	0.60
Normal	1/2	1250	0.60	0.43	0.29	0.54
Logistic	0	1250	0.04	0.78	0.00	0.57
Logistic	1/4	1250	0.29	0.73	0.00	0.48
Logistic	1/2	1250	0.51	0.36	0.25	0.46

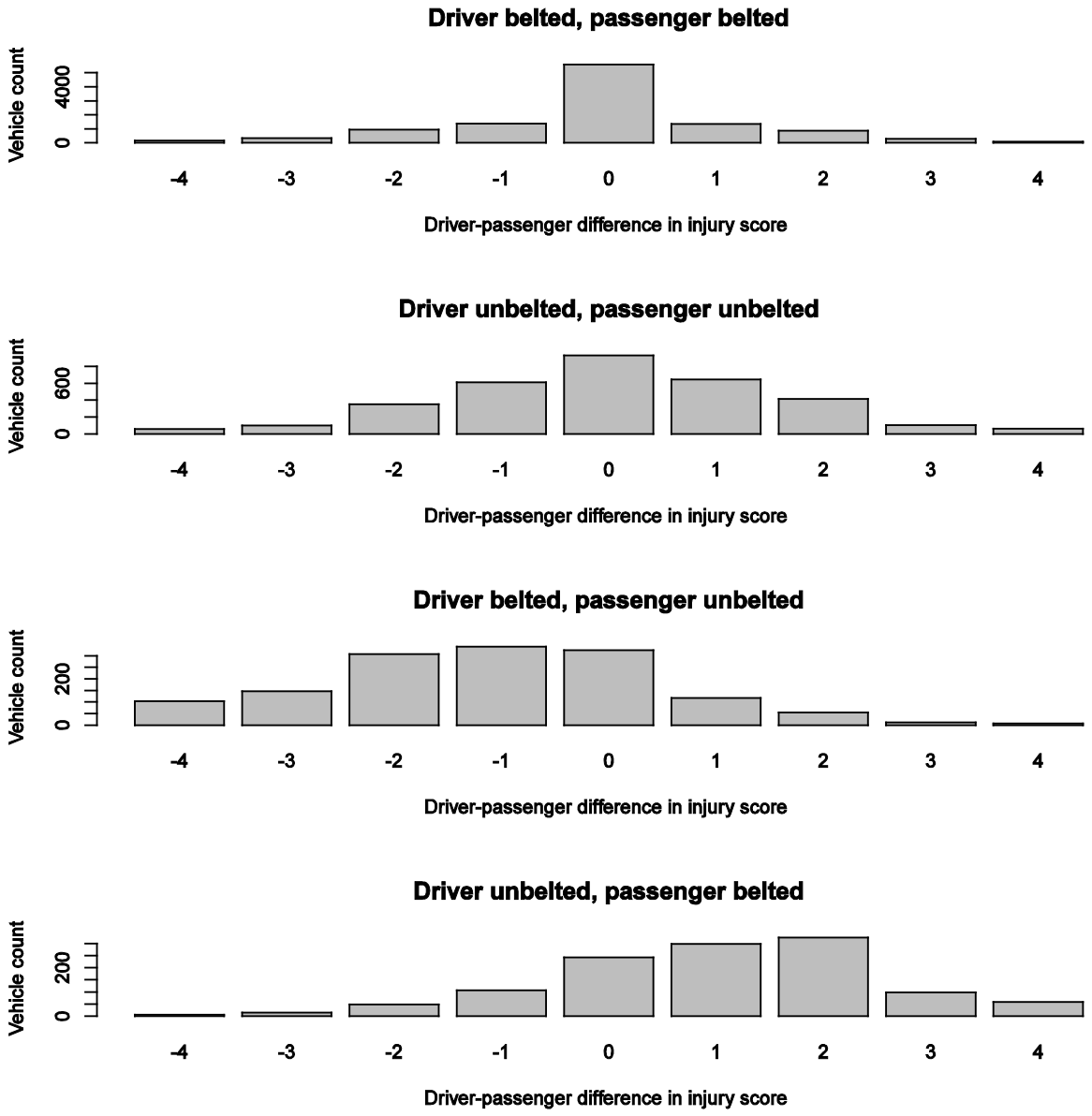

Figure 1: Pair differences in injury scores, driver-minus-passenger, for a driver and a passenger in the same car in FARS 2010-2011, by restraint use. A positive difference indicates the driver suffered more severe injuries than the passenger.

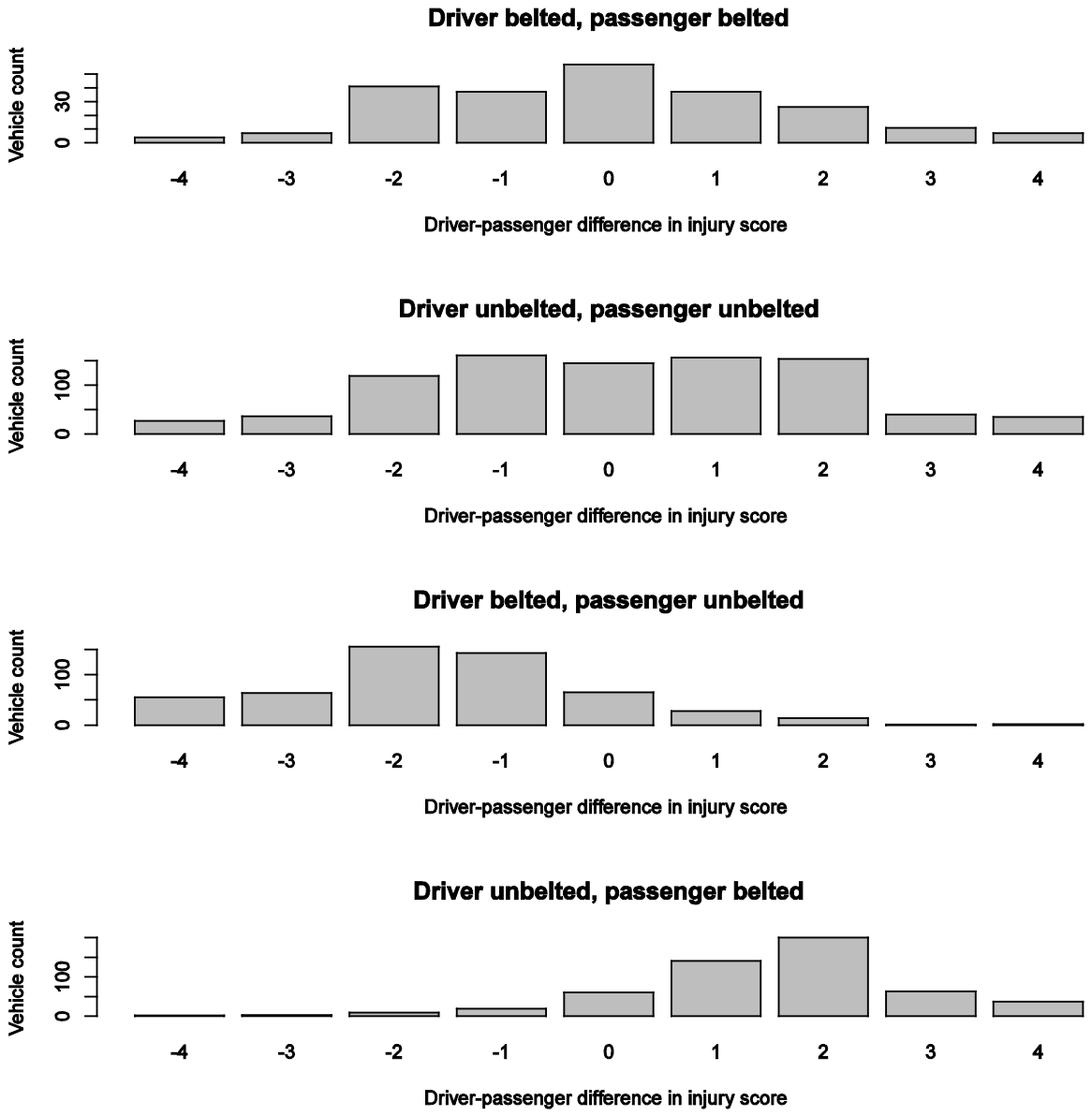

Figure 2: Pair differences in injury scores, driver-minus-passenger, for a driver and a passenger in the same car in FARS 2010-2011, by restraint use, when precisely one individual was ejected from the vehicle, either partially ejected or totally ejected. A positive difference indicates the driver suffered more severe injuries than the passenger.