Statistical Science
2017, Vol. 32, No. 4, 514–530
DOI: 10.1214/17-STS621

The General Structure of Evidence Factors in Observational Studies

Paul R. Rosenbaum

Abstract. The general structure of evidence factors is examined in terms of the knit product of two permutation groups. An observational or nonrandomized study of treatment effects has two evidence factors if it permits two (nearly) independent tests of the null hypothesis of no treatment effect and two (nearly) independent sensitivity analyses for those tests. Either of the two tests may be biased by nonrandom treatment assignment, but certain biases that would invalidate one test would have no impact on the other, so if the two tests concur, then some aspects of biased treatment assignment have been partially addressed. Expressed in terms of the knit product of two permutation groups, the structure of evidence factors is simpler and less cluttered, but at the same time more general and easier to apply in a new context. The issues are exemplified by an observational study of cigarette smoking as a cause of periodontal disease.

Key words and phrases: Evidence factor, knit product, permutation group, permutation inference, randomization inference, semidirect product, sensitivity analysis, wreath product, Zappa–Szep product.

1. INTRODUCTION: MOTIVATION FOR EVIDENCE FACTORS

> [We should] trust rather to the multitude and variety of . . . arguments than to the conclusiveness of any one. [Our] reasoning should not form a chain which is no stronger than its weakest link, but a cable whose fibers may be ever so slender, provided they are sufficiently numerous and intimately connected.
>
> Charles Sanders Peirce (1868)

1.1 Seeking Concurrence of Several Sources of Evidence, Each Susceptible to Different Biases

In an experiment, biased comparisons of treatments are avoided by randomly assigning individuals to treatments, so experimental design focuses on reducing the standard error of consistent or unbiased estimates and reducing the cost of the experiment; see Fisher (1935) and Cox and Reid (2000). When ethical or practical barriers prevent random assignment, observational studies of treatment effects may yield biased inferences about treatment effects by virtue of comparing people who are not comparable, even if they appear to be comparable in terms of measured covariates; see Cochran (1965).

Biases due to nonrandom treatment assignment do not diminish with increasing sample size, so they quickly come to dominate the mean squared error of an estimated effect, and they cannot be addressed simply by acquiring more data of the same kind, more data subject to the same bias; see Rosenbaum (2001). As a result, investigators often examine a variety of sources of evidence, seeking concurrence among several sources of evidence that are likely to be biased in different ways. For instance, Mervyn Susser (Susser, 1973, page 148, and Susser, 1987, page 88) wrote:

> The epidemiologist [. . . seeks. . .] consistency of results in a variety of repeated tests. . . Consistency is present if the result is not dislodged in the face of diversity in times, places, circumstances, and people, as

Paul R. Rosenbaum is Robert G. Putzel Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: rosenbaum@wharton.upenn.edu).

> well as of research design. . . . The strength of the argument rests on the fact that diverse approaches produce similar results.

Similarly, William Cochran [(1965), pages 252–253] wrote:

> The combined evidence on a question that has to be decided mainly from observational studies will usually consist of a heterogeneous collection of results of varying quality, each bearing on some consequence of the causal hypothesis . . . [The investigator] cannot avoid an attempt to weigh the evidence for and against, since some results are so vulnerable to bias that they should be given low weight even if supported by routine tests of significance.

Use of two evidence factors in an observational study is an attempt to employ this strategy inside a single observational study. Obviously, there is little point in reporting at length two highly correlated analyses of the same data that depend upon the same assumptions about unmeasured biases, two analyses that would be invalidated by the same biases. Two evidence factors are two tests of the null hypothesis of no treatment effect that would be nearly independent under the null hypothesis and that are likely to be affected by different biases. Use of evidence factors is an attempt to formalize the sound but informal considerations raised by Susser and Cochran.

1.2 Motivating Example: Smoking and Periodontal Disease

Smoking is widely believed to cause periodontal disease; see Tomar and Asma (2000) and the Centers for Disease Control (2016). Figure 1 depicts 441 matched pairs of a daily cigarette smoker and a never-smoker from the 2011–2012 National Health and Nutrition Examination Survey, as described in greater detail in Rosenbaum (2016a). Smokers smoked every day for the past 30 days, whereas never-smokers smoked fewer than 100 cigarettes in their lives, do not smoke now, and had no tobacco use in the previous five days. Daily smokers began smoking 30 years ago, on average, and 90% began smoking more than 14.9 years ago. The pairs were matched for age, gender, five categories of education, income and black race using the algorithm in Yang et al. (2012); see Table 1 and Figure 1 in Rosenbaum (2016a) for a demonstration that the matching balanced these measured covariates.

Periodontal disease is present on a tooth if the gums and tooth exhibit separation. Following Tomar and Asma (2000), the measure of periodontal disease examines 28 teeth, if present, not including wisdom teeth, at six locations on each tooth, judging a location to exhibit periodontal disease if there is either a loss of attachment $\geq 4mm$ or a pocket depth of $\geq 4mm$. A person is scored by the percent of completed measurements exhibiting periodontal disease in this sense, and Figure 1 plots the smoker-minus-control difference in these percents, which may range from -100% to 100%. (The measure is slightly different from those used in Rosenbaum, 2016a, where upper and lower teeth were considered separately.)

The boxplot on the left in Figure 1(i) displays 441 smoker–control pair differences, which tend to be positive indicating greater periodontal disease among smokers. The asymmetry in Figure 1(i) is significant at <0.0001 by Wilcoxon's signed rank test, with a 95% confidence for a shift of $[10.2, 17.6]$. The scatterplot on the right in Figure 1(ii) plots the 441 smoker–control pair differences against the number of cigarettes smoked per day by the smoker, and there is a weak increasing trend, more cigarettes predicting greater periodontal disease. The trend in Figure 1(ii) is significant at 0.013 with Kendall's rank correlation of 0.084. As discussed later, neither Wilcoxon's test nor Kendall's test is the best test in an observational study.

Figure 1 emphasizes points in the upper and lower quintile of each variable, although all points are plotted. The cross-cut statistic uses the 2×2 table of counts beyond the quintiles, and it yields a two-sided P-value of 0.0044 in a randomization test. The cross-cut statistic has attractive properties, specifically attractive power and design sensitivity, when used in a sensitivity analysis in an observational study; see Rosenbaum (2016b).

The decision to smoke rather than not smoke may be biased in a different way than the decision to smoke more or fewer cigarettes. A well-informed, disciplined person concerned with health is likely to avoid smoking altogether. Some people smoke a few cigarettes as an appetite suppressant, smoking to maintain a moderate weight and attractive appearance. Indeed, the median BMI for smokers in Figure 1 is 27.1, and is 29.0 for matched nonsmokers. Others people compulsively smoke many cigarettes. The biases that affect a comparison of smokers and nonsmokers may be different from the biases that affect a comparison of heavy smokers and light smokers, though both comparisons could easily be biased. We might expect that an effect

of smoking on periodontal disease would reveal greater periodontal disease among smokers than nonsmokers, and greater disease among heavy smokers than among light smokers. Perhaps there is a sense in which the evidence of an effect is stronger if the data concur with both predictions. To what extent are these two predictions redundant and to what extent are they independent pieces of evidence?

The permutation group relevant to Figure 1 is rather large, albeit simple in structure. To illustrate using a small permutation group that permits close inspection, Table 1 describes four individuals in two matched pairs. The first pair contains two white men in their early 50s with some college education, essentially an associate's degree, and solidly middle class incomes. One is a daily smoker who smokes 40 cigarettes per day, the other is a never-smoker. The smoker in this pair has extensive periodontal disease, with 63.04% of measurements indicating separation of gums from teeth. The second pair consists of two black women in their early 60s with a high-school degree or equivalent, and an income below the poverty line. The smoker in this pair smoked 8 cigarettes per day, and had somewhat greater periodontal disease than her matched control. We will be interested in a permutation group that permutes people among treatments while keeping the matched pairs intact.

1.3 Outline: When Are Two Sensitivity Analyses (Essentially) Independent?

Section 2 is background: it introduces notation and concepts for an observational study without evidence factors, and in particular it conducts two sensitivity analyses for Figure 1, one for Figure 1(i) and one for Figure 1(ii). To what extent and in what sense do Figure 1(i) and Figure 1(ii) provide distinct pieces of information about the effects of smoking on periodontal disease? Section 3 answers this question in quite general terms using the knit product of permutation groups. It will be seen that each factor in the knit product yields a permutation test that may be valid despite enormous biases affecting the other factor, and the two factors may be combined into a single sensitivity analysis using methods for combining independent P-values, such as Fisher's method. The main result of Section 3 is Proposition 1: it gives conditions such that the two upper bounds on the two P-values for two evidence factors are stochastically larger than the uniform distribution on the unit square, thereby permitting them to be used as if they were independent P-values.

In early sections, attention focuses on one example of a knit product of permutation groups, namely the one relevant to Figure 1. In Section 4, other examples of knit products are discussed. Observational studies often exhibit less symmetry than designed experiments, and this is apparent in Section 4, where there are knit products that are not wreath products. Additionally, the knit product for Figure 1 acts within and between matched pairs, but in Section 4 there is a knit product that acts in a slightly more complex way within matched sets.

Evidence factors that are strictly independent were proposed for certain rank statistics in Rosenbaum (2010a). These results are fine so far at they go, but they unnaturally restrict the scope of the method, requiring specific statistics that are only available in specific study designs. In Rosenbaum (2011), the goal of strict independence was replaced by the goal of P-value bounds that may be dependent but are stochastically larger than the uniform distribution on the unit square. The argument in Rosenbaum (2011) eliminates the need for rank tests, but is quite limited in the permutation distributions and associated study designs to which it can be applied. In contrast, the formulation in terms of the knit product of permutation groups seems to capture what is essential to evidence factors: it is simpler yet far more general.

For two applications that used particular evidence factors, see Zhang et al. (2011) and Zubizarreta et al. (2012). For an informal discussion of evidence factors, see Rosenbaum (2015a, 2015b, 2017).

2. BACKGROUND FOR STUDIES WITHOUT EVIDENCE FACTORS

2.1 Permuting Units Among Treatment Positions in Randomized Experiments

A randomized experimental design involving N units may be viewed as: (i) N fixed treatment positions, (ii) N units, represented by the vector $\mathbf{i} = (1, 2, \ldots, N)^T$, (iii) a known random process assigning units to treatment positions. If each unit was placed in a different treatment position, there would be a plan for treating everyone in the experiment; that is the meaning of a treatment position. The N treatment positions may be N distinct treatments, but there is no requirement that this be true. In Figure 1, there are $N = 882 = 2 \times 441$ units, while in Table 1 there are $N = 4$ units. The four treatment positions in Table 1 are smoker of 40 cigarettes per day, control paired with that smoker, smoker of 8 cigarettes per day, control

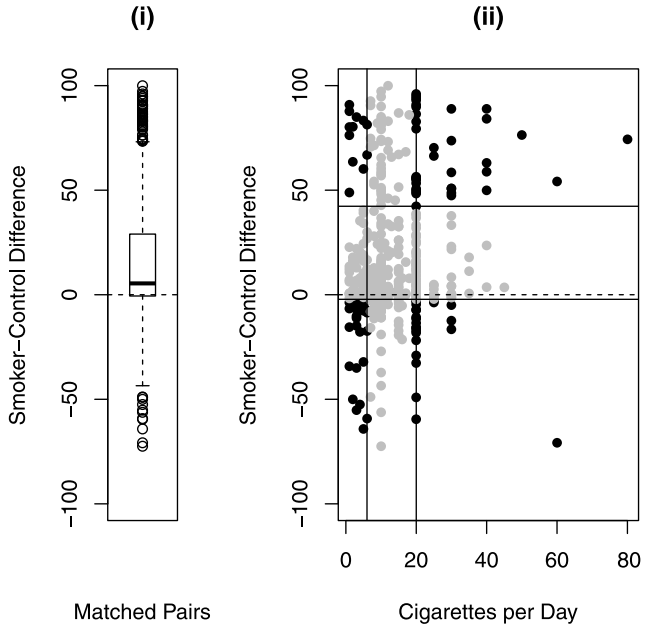

FIG. 1. *Matched pair differences in periodontal disease for* 441 *pairs of a daily smoker and a never smoker, matched for age, gender, education, income and black race. The measure of periodontal disease is the proportion of tooth locations exhibiting separation of tooth and gum. In* (ii), *the smoker-minus-control difference is plotted against the amount smoked by the smoker, with points in the outer quintiles in black, the quintiles being indicated by solid lines. The dashed line is at zero difference.*

paired with that smoker; these positions do not change. However, we could run an experiment with these positions in 8 ways in Table 1 by permuting $\mathbf{i} = (1, 2, 3, 4)^T$ while keeping the pairs intact.

It is convenient to represent the possible treatment assignments by a finite group $\mathfrak{G}$ of $N \times N$ permutation matrices $\mathbf{g} \in \mathfrak{G}$ under the operation of matrix multiplication. By Cayley's theorem, there is no loss of generality in representing a finite group by a group of permutation matrices; see Rotman (1995), Theorem 3.12, page 52. Here, if $\mathbf{g} \in \mathfrak{G}$, then $\mathbf{gi}$ is one possible assignment of units to treatments. If $\mathbf{g} \in \mathfrak{G}$ and $\mathbf{g}^* \in \mathfrak{G}$, then $\mathbf{gg}^* \in \mathfrak{G}$ is the assignment that permutes $\mathbf{i}$ according to $\mathbf{g}^*$, then permutes the result according to $\mathbf{g}$, to obtain the assignment $\mathbf{gg}^*\mathbf{i}$. Generally, $\mathfrak{G}$ is a subgroup of the symmetric group consisting of all $N!$ permutation matrices. Write $|\mathcal{S}|$ for the number of elements in a finite set $\mathcal{S}$, so $|\mathfrak{G}| \leq N!$. The group $\mathfrak{G}$ for Table 1 has $|\mathfrak{G}| = 8 \leq 4! = 24$, with 2 ways to assign a pair to 40 cigarettes or 8 cigarettes, 2×2 ways to assign one person in each pair to smoking or control, making $|\mathfrak{G}| = 8 = 2 \times 2 \times 2$ assignments in total. For a design with $N/2$ pairs comparing a dose of treatment to no treatment in each pair, as in both Table 1 and Figure 1, there are $(N/2)!$ ways to assign the pairs to dose positions, and $2^{N/2}$ ways to pick one treated person in a pair, so $|\mathfrak{G}| = (N/2)! \times 2^{N/2}$. This particular group is discussed by Bell and Haller (1969); moreover, it isomorphic to the group of $(N/2) \times (N/2)$ matrices of permutations and coordinate sign changes, a reflection group that is much discussed in the statistical literature; see Conlon et al. (1977), Eaton and Perlman (1977) and Eaton (1982). In Section 4, there will be groups that are not widely discussed, although they arise naturally in observational studies when matching treated units to several controls, perhaps a variable number of controls.

The group $\mathfrak{G}$ is a finite set of permutation matrices $\mathbf{g}$, and a probability distribution on $\mathfrak{G}$ assigns a probability $p(\mathbf{g})$ to each permutation matrix $\mathbf{g} \in \mathfrak{G}$. More precisely, a probability distribution p on $\mathfrak{G}$ has $p(\mathbf{g}) \geq 0$

TABLE 1

Two of 441 *matched pairs of a daily-smoker and a never smoker, matched for gender* (*female* = 1), *age in years*, *race* (*black* = 1), *education in five categories*, *income measured as the ratio of income to the poverty level and capped at* 5 *times poverty. The outcome is the percent of measurements indicative of periodontal disease*, 0–100%. *The treatment position is defined by the pairing*, *smoking-or-not and cigarettes-smoked-per-day* (*Cigarettes*) *for the smoker. A different treatment assignment would permute the four units without changing who is paired with whom, while leaving the four treatment positions as they are. The smoker-minus-control difference in outcomes is* 63.04 *in pair* 1 *and* 11.08 *in pair* 2

Unit		**Attributes of units**						**Treatment position**	
ID		**Matched covariates**					**Outcome**	**Treatment**	
i	**Pair**	**Female**	**Age**	**Black**	**Education**	**Income**	**Percent diseased**	**Smoker**	**Cigarettes**
1	1	0	54	0	SomeCol	5.00	63.04	1	40
2	1	0	53	0	SomeCol	5.00	00.00	0	0
3	2	1	61	1	HS/GED	0.67	21.50	1	8
4	2	1	64	1	HS/GED	0.72	10.42	0	0

for $\mathbf{g} \in \mathfrak{G}$, and $1 = \sum_{\mathbf{g}\in\mathfrak{G}} p(\mathbf{g})$ and a random $\mathbf{G}$ sampled from this distribution has $\Pr(\mathbf{G} = \mathbf{g}) = p(\mathbf{g})$. Because $|\mathfrak{G}|$ is finite, a probability distribution, p, is a vector of dimension $|\mathfrak{G}|$ indexed by $\mathfrak{G}$, that is, $p = (p_{\mathbf{g}_1}, p_{\mathbf{g}_2}, \ldots, p_{\mathbf{g}_{|\mathfrak{G}|}})$ where the jth coordinate $p_{\mathbf{g}_j}$ is the probability $p(\mathbf{g}_j)$ of permutation matrix $\mathbf{g}_j$. Later, we will speak of a set $\mathcal{P}$ of probability distributions on $\mathbf{g} \in \mathfrak{G}$, so $\mathcal{P}$ will be a subset of $|\mathfrak{G}|$-dimensional Euclidean space. In the current discussion, by definition, an experiment is randomized if the experimenter picks $\mathbf{g} \in \mathfrak{G}$ with equal probabilities, that is, according to the distribution $\overline{p}$ defined by $\overline{p}(\mathbf{g}) = |\mathfrak{G}|^{-1}$ for all $\mathbf{g} \in \mathfrak{G}$. There are other ways to randomize with unequal probabilities, but they will not be part of the discussion of randomized experiments in this paper; see, for instance, Efron (1971). In observational studies, we do not know the true distribution of treatment assignments, and a set $\mathcal{P}$ of distributions p will play a role in calibrating ignorance about the true distribution of treatment assignments.

The current paper will discuss testing Fisher's null hypothesis H_0 of no treatment effect, which says that the responses of each person are not altered by changing the treatment assignments. If changing the treatment a person receives changes the person's responses, then the treatment has some effect, perhaps an effect that is too small or erratic to be interesting, but some effect nonetheless. In Table 1, Fisher's H_0 says that person $i = 1$ would have 63.04% periodontal disease under all $|\mathfrak{G}| = 8$ treatment assignments, and the same is true of persons $i = 2, 3, 4$. Once you can test Fisher's null hypothesis, there are many ways of inverting the test to build confidence intervals or point estimates for the magnitude of the effect; see Lehmann and Romano (2005), Chapter 5, and Rosenbaum (2002), Chapter 5. No new issues arise in inverting the test, so it saves quite a bit of otherwise unneeded notation if we focus on testing the hypothesis of no effect.

A statistic testing Fisher's hypothesis H_0 of no treatment effect is a real-valued random variable, $T = t(\mathbf{G})$, where $t : \mathfrak{G} \to \mathbb{R}$. Of course, T will depend upon measurements describing units besides their treatment positions, but under H_0 these measurements are unchanged by changing the treatment assignment, $\mathbf{G}$, so it simplifies notation to leave this dependence implicit in $T = t(\mathbf{G})$. In other words, under H_0, the test statistic T in a randomized experiment is a random variable with a known distribution because the treatment assignment $\mathbf{G}$ is a random variable with a known distribution, and it is in this sense that randomization forms the "reasoned basis for inference" in randomized experiments, to use Fisher's (1935) phrase.

In the example in Section 1.2, the randomization distribution $\overline{p}(\mathbf{g}) = |\mathfrak{G}|^{-1}$ is the basis for both the Wilcoxon test and the Kendall test, in the following way. The Wilcoxon test ignores the number of cigarettes smoked—its statistic is invariant to the $(N/2)!$ permutations of whole pairs—and it employs the $2^{N/2}$ permutations of smoking status within pairs. The Kendall statistic conditions on the smoker–nonsmoker assignments within pairs, fixing upon one of the $2^{N/2}$ permutations of smoking status within pairs, and it employs the $(N/2)!$ permutations of whole pairs among the doses of cigarettes smoked per day. As noted in Rosenbaum (2010a), these two rank tests are statistically independent under H_0 if treatment assignments are determined by $\overline{p}(\mathbf{g})$. This strict independence turns out to be tied to a narrow choice of rank tests, but one can obtain something just as useful, and much more, for large classes of tests, as is discussed Section 3, specifically in Proposition 1.

There is a small, elegant literature concerned with using one set of data twice to obtain independent rank tests; see, for instance, Alam (1974), Dwass (1960), Marden (1992), and the many references given there. Several authors have noted that strict independence is lost without rank tests, but very strong forms of unrelatedness persist; see, for instance, Randles and Hogg (1971) and Wolfe (1973).

The knit product of permutation groups will provide a general result along these lines, Proposition 1, not only for randomization distributions from $\overline{p}(\mathbf{g})$, but more importantly for sensitivity analyses in observational studies. There is an intuitive sense in which Wilcoxon's signed rank test is affected by biases that determine who smokes in a smoker-control pair, but is unaffected by biases that determine which smokers smoke more and which smoke less; whereas the opposite is true for Kendall's correlation. Proposition 1 will both formalize this intuition and give a basis for constructing two sensitivity analyses that do not affect each other and can, therefore, provide mutual support.

2.2 Sensitivity Analysis in Observational Studies

In an observational study, the distribution of p of the treatment assignment $\mathbf{G}$ is neither known nor identified from the observable data. Beginning with Cornfield et al. (1959), investigators have asked: How far would p have to depart from a randomization distribution to qualitatively alter conclusions reached by acting as if treatment assignments were randomized? They were writing at a time when there was active debate

about the effects of smoking cigarettes on lung cancer. Cornfield et al. (1959) observed that, to explain the strong association between heavy smoking and lung cancer, the departure from random assignment would need to be enormous, the failure to match for an unmeasured covariate strongly predictive of treatment assignment and a near perfect predictor of lung cancer. That is, they concluded that the ostensible effects of smoking on lung cancer are highly insensitive to unmeasured bias, because only enormous biases could produce the observed association in the absence of a causal effect. This objective finding of insensitivity does not eliminate the possibility that the association is produced by unmeasured bias, but it constrains the debate about such claims. Smokers drink more alcohol than nonsmokers, and behave differently in many ways, but none of these commonly noted differences fit the description of the bias that Cornfield et al. (1959) found was needed to explain the association between smoking and lung cancer. Cornfield et al. (1959) replaced the qualitative but largely uninformative statement that "association does not imply causation, sufficiently large biases can explain any association" by a quantitative statement informed by the data: "to explain the observed association, the unmeasured biases in treatment assignment would have to be of at least a particular magnitude".

Although an important conceptual advance, the specific method, an inequality for probabilities, that Cornfield et al. (1959) used is limited in several ways: it is restricted to a binary outcome, ignores sampling variability by equating sample and population quantities, and ignores adjustments for observed covariates. A method that removes these limitations but is otherwise similar in spirit and content to the method of Cornfield et al. (1959) was proposed in Rosenbaum (1987) and is developed in detail in Rosenbaum (2002), Chapter 4; Rosenbaum (2007). In particular, the method is applied in Rosenbaum (2002), Section 4.3.2 to Hammond's (1964) study of smoking and lung cancer, reaching a conclusion similar to that of Cornfield et al. (1959). The method introduces a sensitivity parameter $\Gamma \geq 1$ that measures the magnitude, but not the specific form, of the departure of the distribution p of treatment assignments from the randomization distribution $\overline{p}(\mathbf{g}) = |\mathfrak{G}|^{-1}$ for all $\mathbf{g} \in \mathfrak{G}$. Stated informally, two individuals with the same observed covariates may differ in their odds of receiving one treatment rather than another by a factor of Γ, so $\Gamma = 1$ yields the randomization distribution, $\overline{p}$, while letting $\Gamma \to \infty$ permits large departures from random assignment approaching deterministic assignments. The practical question is: How large would Γ have to be to alter the conclusions of the commonplace but naive analysis of an observational study that pretends that treatment assignment is randomized within matched pairs? For instance, how large would Γ have to be to accept a null hypothesis of no treatment effect rejected under the randomization distribution, $\overline{p}$? An aid to interpreting Γ was given in Rosenbaum and Silber (2009). For details of this method of sensitivity analysis, see the cited references and associated `R` packages.

The presentation of sensitivity analysis below is slightly abstract. If the material is either unfamiliar or uncomfortably abstract, it can be made concrete with minimal effort by installing and loading the `sensitivitymw` package in `R` and executing the `erpcp` example in the help files for `senmw` and `senmwCI`. This example reproduces a sensitivity analysis for an interesting matched observational study by Werfel et al. (1998) of DNA damage caused by electric arc welding, as discussed in Rosenbaum (2007), Chapter 3.3. The `R` package is described in detail in Rosenbaum (2015a).

The notation that follows presumes there are $L \geq 1$ evidence factors, although commonly and in the examples there are $L = 2$ factors. We consider, for each $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_L) \geq (1, \ldots, 1) = \mathbf{1}$, with $L \geq 1$, a set $\mathcal{P}_{\mathbf{\Gamma}}$ of probability distributions, $p \in \mathcal{P}_{\mathbf{\Gamma}}$, on $\mathfrak{G}$, with the following properties: (i) $\mathbf{\Gamma} = \mathbf{1}$ corresponds with the randomization distribution, $\mathcal{P}_{\mathbf{1}} = \{\overline{p}\}$; (ii) the sets grow with increasing $\mathbf{\Gamma}$ so that $\mathcal{P}_{\mathbf{\Gamma}} \subseteq \mathcal{P}_{\mathbf{\Gamma}'}$ for $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_L) \leq (\Gamma_1', \ldots, \Gamma_L') = \mathbf{\Gamma}'$; (iii) for each finite $\mathbf{\Gamma}$, the set $\mathcal{P}_{\mathbf{\Gamma}}$ is a compact set of $|\mathfrak{G}|$-dimensional vectors $p = (p_{\mathbf{g}_1}, p_{\mathbf{g}_2}, \ldots, p_{\mathbf{g}_{|\mathfrak{G}|}})$ all of whose coordinates are strictly positive, $p_{\mathbf{g}_j} > 0$, with $1 = \sum_{j=1}^{|\mathfrak{G}|} p_{\mathbf{g}_j}$. In other words, conditions (i) and (ii) say that, as $\mathbf{\Gamma}$ increases, $\mathcal{P}_{\mathbf{\Gamma}}$ allows for larger departures from randomized treatment assignment. Among other things, condition (iii) says that $\mathfrak{G}$ defines the set of possible treatment assignments, that is, those with positive probability. There is a division of labor between $\mathfrak{G}$ and $\mathcal{P}_{\mathbf{\Gamma}}$, in that $\mathfrak{G}$ defines what is possible and $\mathcal{P}_{\mathbf{\Gamma}}$ defines what is probable.

Unlike a randomized experiment, in an observational study we do not know the true distribution of treatment assignments, $\Pr(\mathbf{G} = \mathbf{g})$, so we cannot compute the tail probability $\Pr\{t(\mathbf{G}) \geq a\}$ of a test statistic under the null hypothesis of no treatment effect. A sensitivity analysis computes bounds on this unknown tail probability $\Pr\{t(\mathbf{G}) \geq a\}$ when the bias in treatment assignment is at most $\mathbf{\Gamma}$, that is, when $p \in \mathcal{P}_{\mathbf{\Gamma}}$. A computation of this form may permit us to say that a bias

of magnitude $\mathbf{\Gamma}$ or smaller would be insufficient to accept the null hypothesis of no treatment effect in a level α test. Tentatively assuming $\Pr(\mathbf{G}=\mathbf{g})=p(\mathbf{g})$ for some unknown $p(\cdot)\in\mathcal{P}_{\mathbf{\Gamma}}$, a sensitivity analysis requires the computation of bounds, $\underline{b}_{\mathbf{\Gamma}}(k)\le\overline{b}_{\mathbf{\Gamma}}(k)$, on the unknown distribution $\Pr\{t(\mathbf{G})\ge a\}$ of T,

$$\begin{aligned}\underline{b}_{\mathbf{\Gamma}}(a)&=\min_{p\in\mathcal{P}_{\mathbf{\Gamma}}}\sum_{\mathbf{g}\in\mathfrak{G}}\chi\{t(\mathbf{g})\ge a\}p(\mathbf{g})\\&\le\Pr\{t(\mathbf{G})\ge a\}\\&\le\max_{p\in\mathcal{P}_{\mathbf{\Gamma}}}\sum_{\mathbf{g}\in\mathfrak{G}}\chi\{t(\mathbf{g})\ge a\}p(\mathbf{g})\\&=\overline{\underline{b}}_{\mathbf{\Gamma}}(a),\end{aligned}\tag{1}$$

where $\chi(E)=1$ if event E occurs and $\chi(E)=0$ if E does not occur; so, $\underline{b}_{\mathbf{\Gamma}}(a)$ and $\overline{\underline{b}}_{\mathbf{\Gamma}}(a)$ provide known bounds on the unknown probability $\Pr\{t(\mathbf{G})\ge a\}$. Then $\mathbf{\Gamma}$ is varied to display the sensitivity of conclusions to biases of different magnitudes measured by $\mathbf{\Gamma}$. Because $\mathcal{P}_{\mathbf{\Gamma}}$ is a compact set of vectors $p=(p_{\mathbf{g}_1},p_{\mathbf{g}_2},\ldots,p_{\mathbf{g}_{|\mathfrak{G}|}})$ and $\sum_{\mathbf{g}\in\mathfrak{G}}\chi\{t(\mathbf{g})\ge a\}p(\mathbf{g})$ is the sum of certain coordinates of p and hence a continuous function of p, it follows that the min and max are attained in (1). If the realized value of the test statistic, $T=t(\mathbf{G})$, satisfied $\overline{\underline{b}}_{\mathbf{\Gamma}}(T)\le\alpha$, then a bias of magnitude $\mathbf{\Gamma}$ is too small to lead us to accept at level α the null hypothesis of no treatment effect. The random variable $\overline{\underline{b}}_{\mathbf{\Gamma}}\{t(\mathbf{G})\}$ is an upper bound on the P-value testing Fisher's null hypothesis, H_0, in the presence of a bias of magnitude at most $\mathbf{\Gamma}$.

In practice, the computation of $\underline{b}_{\mathbf{\Gamma}}(a)$ or $\overline{\underline{b}}_{\mathbf{\Gamma}}(a)$ requires some attention to details not discussed in the current paper. Because $|\mathfrak{G}|=(N/2)!\times 2^{N/2}$ in Figure 1, direct numerical optimization of (1) is not practical. Often, neither $\underline{b}_{\mathbf{\Gamma}}(a)$ nor $\overline{\underline{b}}_{\mathbf{\Gamma}}(a)$ defines a probability distribution, a harmless inconvenience; however, there are many useful cases in which they are probability distributions, a simplifying convenience. For one example of each of these two situations, see Rosenbaum (2007) where the paired case in its Section 3 yields bounds (1) that are probability distributions, while the case of matching with multiple controls in its Section 4 yields bounds (1) that are perfectly serviceable but are not probability distributions.

For various approaches to sensitivity analysis in observational studies, see: Cornfield et al. (1959), Gastwirth (1992), Hosman, Hansen and Holland (2010), Imbens (2003), Liu, Kuramoto and Stuart (2013), McCandless, Gustafson and Levy (2007), Shepherd et al. (2006) and Yu and Gastwirth (2005).

2.3 Sensitivity Analysis in the Example

To illustrate, consider again the application of Wilcoxon's signed rank test to the example in Section 1.2 and Figure 1. If $\mathfrak{K}$ were the group of $N\times N$ permutation matrices that permute the two units in a pair to change their roles as smoker or nonsmoker, then $|\mathfrak{K}|$ would be $2^{N/2}$. One set of probability distributions on this group $\mathfrak{K}$ assigns smoking or control independently in distinct pairs and uses a different biased coin in each pair, such that the $N/2$ biased coins have probabilities of a head between $1/(1+\Gamma_1)$ and $\Gamma_1/(1+\Gamma_1)$; then, for $\Gamma_1=1$ there is random assignment within pairs, and for $\Gamma_1\ge 1$ there is a compact set of probability distributions $p(\mathbf{k})=\Pr(\mathbf{K}=\mathbf{k})$, $\mathbf{k}\in\mathfrak{K}$; see Rosenbaum (1993, 2007) and Rosenbaum (2002), Chapter 4. In this case, using Wilcoxon's statistic, the upper bound $\overline{\underline{b}}_{\mathbf{\Gamma}}\{t(\mathbf{G})\}$ on the one-sided P-value testing Fisher's H_0 is 0.0493 for $\Gamma_1=2.76$ and is 0.0521 for $\Gamma_1=2.77$, so a bias in treatment assignment of $\Gamma_1=2.76$ is just a bit too small to lead to acceptance of H_0 at the conventional 0.05 standard, and a bias of $\Gamma_1=2.77$ is just barely large enough to lead to acceptance. A bias of $\Gamma_1=2.75$ means the treatment assignment probabilities in individual pairs might not be 1/2, but might be anything in the interval $1/(1+\Gamma_1)=0.2667$ and $\Gamma_1/(1+\Gamma_1)=0.7333$. As noted by Rosenbaum and Silber (2009), a bias of $\Gamma_1=2.75$ corresponds with an unmeasured covariate that increases the odds of smoking by a factor of 4 and increases the odds of a positive pair difference in periodontal disease by a factor of 8, not at all an inconsequential covariate; however, failure to match for such a covariate would not suffice to explain rejection of H_0 at the 0.05 level. As mentioned earlier, it is easy to invert this argument to discuss point estimates or confidence intervals instead of P-values; see Rosenbaum (1993, 2007).

It turns out that Wilcoxon's statistic is poor choice, as it exaggerates the harm that a bias of $\Gamma_1=2.77$ can do; see Rosenbaum (2010b). In particular, certain M-estimates report less sensitivity to the same magnitude of bias with the same data; see Huber (1981) for a general discussion of M-estimates, and see Maritz (1979) for their use in randomization tests. Using an M-estimate and M-test designed for sensitivity analyses (ψ_{in} in Rosenbaum, 2013, or `method="p"` in the `sensitivitymw` package in `R`), one obtains an upper bound $\overline{\underline{b}}_{\mathbf{\Gamma}}\{t(\mathbf{K})\}$ on the P-value of 0.0012 at $\Gamma_1=2.77$, and a bound of 0.049 at $\Gamma_1=3.5$. A bias of $\Gamma_1=3.5$ corresponds with an unmeasured covariate that increases the odds of smoking by a factor of

5 and increases the odds of a positive pair difference in periodontal disease by a factor of 11. To attribute rejection of H_0 at the 0.05 level to bias rather than a causal effect is to say that the matching failed to control a covariate strongly predictive of both smoking and periodontal disease.

Is the evidence in Figure 1 stronger than this analysis indicates? Indeed, it is stronger, because this analysis ignores information about the amount smoked. If $\mathfrak{H}$ were the group of $N \times N$ permutation matrices that permutes the pairs among the doses of cigarette smoking, then $|\mathfrak{H}|$ would be $(N/2)!$. Consider the conditional distribution of the number of cigarettes smoked—the dose of smoking—given that a person is a smoker. One set of such conditional probability distributions for doses was considered in Rosenbaum (2016b): it says that two smokers may differ in their odds of smoking a rather than b cigarettes by at most a factor of $\Gamma_2 \geq 1$. If $\Gamma_2 = 1$, then every smoker has the same chance of smoking a cigarettes for each a, yielding the randomization distribution underlying common rank correlation tests, such as Kendall's correlation. If a statistic is not a rank statistic, then the relevant randomization distribution entails conditioning on the order statistic of the number of cigarettes smoked by smokers, or on $(8, 40)$ in Table 1.

The cross-cut statistic is depicted in Figure 1 for a cut at the quintiles: it counts the frequencies in the outer corners, yielding an odds ratio of 3.6, with heavier smoking associated with more extensive periodontal disease. The cross-cut statistic has good power and design sensitivity when used in sensitivity analyses; see Rosenbaum (2016b). The sensitivity bounds (1) for the cross-cut statistic are obtained from the extended hypergeometric distribution with parameter Γ_2, and at $\Gamma_2 = 1.6$, the maximum P-value testing no effect is 0.044. So there is a second aspect to the evidence in Figure 1.

In the two analyses above, Figure 1(i) and Figure 1(ii) were viewed as separate analyses, using (1) with $\mathfrak{G}$ replaced by $\mathfrak{K}$ in the analysis of Figure 1(i), and with $\mathfrak{G}$ replaced by $\mathfrak{H}$ in the analysis of Figure 1(ii), and with different sets of distributions $\mathcal{P}_{\mathbf{\Gamma}}$ over $\mathfrak{K}$ or $\mathfrak{H}$ in these two separate analyses. The next section will illustrate the general concept of evidence factors using the example of the group $\mathfrak{G}$ generated by $\mathfrak{H}$ and $\mathfrak{K}$ as defined above, together with a set of joint distributions $\mathcal{P}_{\mathbf{\Gamma}}$ over $\mathfrak{G}$. In this example, the distributions in $\mathcal{P}_{\mathbf{\Gamma}}$ with $\mathbf{\Gamma} = (\Gamma_1, \Gamma_2)$ will govern who smokes in the $N/2$ pairs, and will govern the conditional distribution of the amount smoked by the smoker in each pair, conditionally given the identity of smokers and the order statistics of the amount smoked. In particular, $\mathcal{P}_{\mathbf{\Gamma}}$ consists of each of the marginal distributions above for $\mathfrak{K}$ multiplied by each of the conditional distributions above for $\mathfrak{H}$ given $\mathfrak{K}$. In Table 1, this means that the identity of the smoker in one pair is determined by two independent coin flips with (possibly different) probabilities in the interval $[\frac{1}{1+\Gamma_1}, \frac{\Gamma_1}{1+\Gamma_1}]$, and then the smoker in one pair is assigned to 40 cigarettes per day, the other smoker to 8 cigarettes per day, by the flip of another biased coin with probability in the interval $[\frac{1}{1+\Gamma_2}, \frac{\Gamma_2}{1+\Gamma_2}]$. The bias in the third coin refers to its conditional distribution given the flips of the first two coins. Although the bias of this conditional distribution of the third coin must be in the interval $[\frac{1}{1+\Gamma_2}, \frac{\Gamma_2}{1+\Gamma_2}]$, the bias may change depending upon the outcome of the flips of the first two coins. If $\Gamma_1 = 1$, the assignment to smoker or control is a fair coin flip, but the amount smoked may be biased. If $\Gamma_2 = 1$, the amount smoked by a smoker is effectively randomized, but the identity of the smoker in a pair may be biased. If $\Gamma_1 > 1$ and $\Gamma_2 > 1$, then both treatment assignments may be biased.

3. KNIT PRODUCTS OF PERMUTATION GROUPS AND EVIDENCE FACTORS

3.1 Definition and Example

A finite group $\mathfrak{G}$ is the knit product or Zappa–Szep product of two of its subgroups $\mathfrak{H}$ and $\mathfrak{K}$ if $\mathfrak{G} = \{\mathbf{hk} : \mathbf{h} \in \mathfrak{H}, \mathbf{k} \in \mathfrak{K}\}$ where each element $\mathbf{g} \in \mathfrak{G}$ can be written in precisely one way as $\mathbf{g} = \mathbf{hk}$ with $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{K}$; see Szep (1950), Gilbert and Wazzan (2008) or Ates and Cevik (2009). Because the representation $\mathbf{g} = \mathbf{hk}$ is unique: (i) $|\mathfrak{G}| = |\mathfrak{K}| \times |\mathfrak{H}|$, (ii) $\mathfrak{K} \cap \mathfrak{H} = \{\mathbf{I}\}$ and (iii) we may determine $\mathbf{h}$ and $\mathbf{k}$ from $\mathbf{g}$. Roman (2012), pages 33, 151, calls this an essentially disjoint product rather than a knit product. A reader familiar with the use of abstract algebra in the construction of experimental designs will sense that the knit product is connected with the concept of orthogonality or balance for nominal factors in such designs, and this is indeed the case.

If $\mathfrak{G}$ is the knit product of two of its subgroups $\mathfrak{H}$ and $\mathfrak{K}$, then it is also the knit product of $\mathfrak{K}$ and $\mathfrak{H}$, the ordering of the two subgroups being unimportant; see, for instance, Lemma 2.18 in Isaacs (2009), page 22. That is, if $\mathfrak{G}$ is the knit product of two of its subgroups $\mathfrak{H}$ and $\mathfrak{K}$, then every element $\mathbf{g} \in \mathfrak{G}$ can be written in precisely one way as $\mathbf{g} = \mathbf{hk}$ with $\mathbf{h} \in \mathfrak{H}$ and $\mathbf{k} \in \mathfrak{K}$,

but also in precisely one way as $\mathbf{g} = \widetilde{\mathbf{k}}\widetilde{\mathbf{h}}$ with $\widetilde{\mathbf{h}} \in \mathfrak{H}$ and $\widetilde{\mathbf{k}} \in \mathfrak{K}$, so $\mathbf{g} = \mathbf{hk} = \widetilde{\mathbf{k}}\widetilde{\mathbf{h}}$, even though, in general, $\widetilde{\mathbf{h}} \neq \mathbf{h}$ and $\widetilde{\mathbf{k}} \neq \mathbf{k}$. In particular, $\mathfrak{G} = \{\mathbf{hk} : \mathbf{h} \in \mathfrak{H}, \mathbf{k} \in \mathfrak{K}\} = \{\mathbf{kh} : \mathbf{h} \in \mathfrak{H}, \mathbf{k} \in \mathfrak{K}\}$ even though, in general, $\mathbf{hk} \neq \mathbf{kh}$.

Group products typically have an "internal" and an "external" definition, yielding isomorphic groups. The internal definition starts with a group $\mathfrak{G}$ and two of its subgroups, $\mathfrak{H}$ and $\mathfrak{K}$, as above. The external form starts with two groups, $\mathfrak{H}$ and $\mathfrak{K}$, and builds a product group $\mathfrak{G}$ from them. For the discussion here, the internal definition suffices.

The knit product has two familiar special cases. If one of the subgroups is a normal subgroup, then the knit product becomes the semidirect product; whereas if both subgroups are normal subgroups, then both the knit product and semidirect product become the direct product. The knit product is more general, because neither subgroup needs to be a normal subgroup, and simpler because there are fewer conditions that the statistician needs to check. Wreath products are semidirect products with a repetitive symmetric structure, and they appear frequently in statistics: see Bell and Haller (1969) for an application to nonparametric inference; Dawid (1985) for an application to multivariate analysis; and Bailey et al. (1983), Brien and Bailey (2006) and Dawid (1988) for applications to experimental design.

Here is a simple, nontrivial example; indeed, it is the example relevant to Table 1. Define $\mathfrak{H}$ and $\mathfrak{K}$ as follows, and define $\mathfrak{G}$ to be the group generated by $\mathfrak{H}$ and $\mathfrak{K}$; that is, $\mathfrak{G}$ is the smallest group containing both $\mathfrak{H}$ and $\mathfrak{K}$.

$$\mathfrak{H} = \left\{ \mathbf{h}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{h}_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \right\},$$

$$\mathfrak{K} = \left\{ \mathbf{k}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{k}_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \right. \tag{2}$$

$$\left. \mathbf{k}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{k}_4 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \right\}.$$

In Table 1, the group $\mathfrak{K}$ assigns one person in each pair to daily smoking, the other to nonsmoking. The group $\mathfrak{H}$ assigns one pair to 40 cigarettes per day and the other pair to 8 cigarettes per day. Wilcoxon's statistic $t(\mathbf{g})$ reacts to the actions of $\mathfrak{K}$ but ignores the actions in $\mathfrak{H}$ in the sense that $t(\mathbf{g}) = t(\mathbf{hk}) = t(\mathbf{k})$; however, it is not generally true that $t(\mathbf{kh}) = t(\mathbf{k})$. Kendall's statistic fixes the action of $\mathfrak{K}$, and reacts to the action of $\mathfrak{H}$. The notions of reacting, ignoring and fixing will be formalized in terms of invariance and conditioning. The groups for Figure 1 are analogous but much larger.

Note that $\mathbf{h}_2\mathbf{k}_3 = \mathbf{k}_2\mathbf{h}_2$: first permuting treatments in the second pair ($\mathbf{k}_3$) and then swapping the two pairs ($\mathbf{h}_2$) is the same as swapping the two pairs ($\mathbf{h}_2$) and then permuting treatments in the first pair ($\mathbf{k}_2$). You can easily check that the smallest group $\mathfrak{G}$ that contains both $\mathfrak{H}$ and $\mathfrak{K}$ consists of exactly eight permutation matrices formed by multiplying each $\mathbf{h} \in \mathfrak{H}$ by each $\mathbf{k} \in \mathfrak{K}$, so $\mathfrak{G}$ is the knit product of $\mathfrak{H}$ and $\mathfrak{K}$. You can easily check that $\mathbf{g}^{-1}\mathbf{kg} \in \mathfrak{K}$ for each $\mathbf{g} \in \mathfrak{G}$ and $\mathbf{k} \in \mathfrak{K}$, so $\mathfrak{K}$ is a normal subgroup and therefore $\mathfrak{G}$ is also the semidirect product of $\mathfrak{H}$ and $\mathfrak{K}$; however, we will not need this additional fact later on.

The first three columns of Table 2 list three times the elements $\mathbf{g}$ of the smallest group $\mathfrak{G}$ that contains both $\mathfrak{H}$ and $\mathfrak{K}$ in (2), giving their two representations as $\mathfrak{G} = \{\mathbf{hk} : \mathbf{h} \in \mathfrak{H}, \mathbf{k} \in \mathfrak{K}\}$ and $\mathfrak{G} = \{\mathbf{kh} : \mathbf{h} \in \mathfrak{H}, \mathbf{k} \in \mathfrak{K}\}$, these being slightly different because $\mathbf{h}_2$ does not commute with $\mathbf{k}_2$ and $\mathbf{k}_3$. The last three columns of Table 2 list the smoker-minus-control pair difference in dental outcomes for the pairs assigned to 40 cigarettes and to 8 cigarettes, together with Wilcoxon's signed rank statistic computed from these two pairs. Notice that Wilcoxon's statistic is unaffected by the action of $\mathbf{h} \in \mathfrak{H}$ in the sense that $t(\mathbf{hk}) = t(\mathbf{k})$; for instance, $t(\mathbf{k}_2) = t(\mathbf{h}_1\mathbf{k}_2) = 1$ in row 2, and $t(\mathbf{h}_2\mathbf{k}_2) = 1$ in row 7 of Table 2.

3.2 Invariant Functions and Invariant Sets of Functions

Let $\mathfrak{F}$ be a subgroup of $\mathfrak{G}$. A function $t : \mathfrak{G} \to \mathbb{R}$ is invariant with respect to $\mathfrak{F}$ if $t(\mathbf{fg}) = t(\mathbf{g})$ for each

TABLE 2

Five probability distributions on the eight element knit product group $\mathfrak{G}$ generated by (2). Each element $\mathbf{g} \in \mathfrak{G}$ is represented as both $\mathfrak{G} = \{\mathbf{kh} : \mathbf{h} \in \mathfrak{H}, \mathbf{k} \in \mathfrak{K}\}$ and $\mathfrak{G} = \{\mathbf{hk} : \mathbf{h} \in \mathfrak{H}, \mathbf{k} \in \mathfrak{K}\}$. Also shown are the 2 smoker-minus-control pair differences in outcomes from Table 1 for treatment positions 40 cigarettes and 8 cigarettes. The final column is Wilcoxon's signed rank statistic

Elements $\mathbf{g} \in \mathfrak{G}$			Five probability distributions p					Differences		Wilcoxon's
$\mathfrak{G}$	kh	hk	(i)	(ii)	(iii)	(iv)	(v)	40 cigs	8 cigs	statistic
I	$\mathbf{k}_1\mathbf{h}_1$	$\mathbf{h}_1\mathbf{k}_1$	0.125	0.320	0.200	0.512	0.128	63.04	11.08	3
$\mathbf{k}_2$	$\mathbf{k}_2\mathbf{h}_1$	$\mathbf{h}_1\mathbf{k}_2$	0.125	0.080	0.200	0.128	0.032	−63.04	11.08	1
$\mathbf{k}_3$	$\mathbf{k}_3\mathbf{h}_1$	$\mathbf{h}_1\mathbf{k}_3$	0.125	0.080	0.200	0.128	0.032	63.04	−11.08	2
$\mathbf{k}_4$	$\mathbf{k}_4\mathbf{h}_1$	$\mathbf{h}_1\mathbf{k}_4$	0.125	0.020	0.200	0.032	0.008	−63.04	−11.08	0
$\mathbf{h}_2$	$\mathbf{k}_1\mathbf{h}_2$	$\mathbf{h}_2\mathbf{k}_1$	0.125	0.320	0.050	0.128	0.512	11.08	63.04	3
$\mathbf{k}_2\mathbf{h}_2$	$\mathbf{k}_2\mathbf{h}_2$	$\mathbf{h}_2\mathbf{k}_3$	0.125	0.080	0.050	0.032	0.128	−11.08	63.04	2
$\mathbf{k}_3\mathbf{h}_2$	$\mathbf{k}_3\mathbf{h}_2$	$\mathbf{h}_2\mathbf{k}_2$	0.125	0.080	0.050	0.032	0.128	11.08	−63.04	1
$\mathbf{k}_4\mathbf{h}_2$	$\mathbf{k}_4\mathbf{h}_2$	$\mathbf{h}_2\mathbf{k}_4$	0.125	0.020	0.050	0.008	0.032	−11.08	−63.04	0

$\mathbf{f} \in \mathfrak{F}$. In Table 1, Wilcoxon's statistic is invariant with respect to $\mathfrak{H}$ in (2), because it does not care that the first treatment position involves smoking 40 cigarettes and the third treatment position involves smoking 8 cigarettes.

In particular, a single probability distribution $p(\cdot)$ on $\mathfrak{G}$ is invariant with respect to $\mathfrak{F}$ if the function $p(\cdot)$ is invariant with respect to $\mathfrak{F}$, so $p(\mathbf{fg}) = p(\mathbf{g})$ for all $\mathbf{f} \in \mathfrak{F}$. For instance, the uniform distribution on $\mathfrak{G}$, namely $\overline{p}(\mathbf{g}) = |\mathfrak{G}|^{-1}$ for all $\mathbf{g} \in \mathfrak{G}$, is invariant with respect to $\mathfrak{F}$ for every subgroup $\mathfrak{F}$ of $\mathfrak{G}$.

Table 2 exhibits five probability distribution on $\mathfrak{G}$ defined by the knit product of the two subgroups in (2). Distribution (i) is $\overline{p}(\cdot)$, randomizing both smoking-control and amount-smoked in Table 1, and it is invariant with respect to all of $\mathfrak{G}$. Distribution (ii) is randomized with respect to the amount smoked by the smoker with $\Gamma_2 = 1$, but is biased with respect to who smokes with $\Gamma_1 = 4$, and it is invariant with respect to $\mathfrak{H}$ in the sense that $p(\mathbf{h}_1\mathbf{k}_j) = p(\mathbf{h}_2\mathbf{k}_j)$ for $j = 1, 2, 3, 4$. Distribution (iii) is randomized with respect to who smokes with $\Gamma_1 = 1$, but biased with respect to the amount smoked with $\Gamma_2 = 4$, and it is invariant with respect to $\mathfrak{K}$ in the sense that $p(\mathbf{k}_j\mathbf{h}) = p(\mathbf{k}_{j'}\mathbf{h})$ for $1 \le j \le j' \le 4$ and for $\mathbf{h} = \mathbf{h}_1$ and $\mathbf{h} = \mathbf{h}_2$. Neither distribution (iv) nor (v) is invariant with respect to $\mathfrak{H}$ or $\mathfrak{K}$ or $\mathfrak{G}$ with $\mathbf{\Gamma} = (\Gamma_1, \Gamma_2) = (4, 4)$. For example, distribution (iv) gives probability 0.512 to the observed treatment assignment in Table 1, with the remaining seven treatment assignments unevenly sharing the remaining 0.488 probability.

A different concept is an invariant set $\mathcal{P}$ of probability distributions $p(\cdot)$ on $\mathfrak{G}$. As will be seen in a moment, to say that a set $\mathcal{P}$ is invariant is not to say that each element $p(\cdot)$ is invariant, but rather to say that the elements may change while leaving the set as a whole unchanged. First, recall from Section 2.1 that a probability distribution $p(\cdot)$ on $\mathfrak{G}$ is both a function $p : \mathfrak{G} \to [0, 1]$ and a $|\mathfrak{G}|$-dimensional vector, $p = (p_{\mathbf{g}_1}, p_{\mathbf{g}_2}, \ldots, p_{\mathbf{g}_{|\mathfrak{G}|}})$, so a set of probability distributions $\mathcal{P}$ is a subset of a $|\mathfrak{G}|$-dimensional Euclidean space. A set $\mathcal{P}$ of distributions $p : \mathfrak{G} \to [0, 1]$ is invariant with respect to a subgroup $\mathfrak{F}$ if for each $p \in \mathcal{P}$ and each $\mathbf{f} \in \mathfrak{F}$ there exist a function $p^* \in \mathcal{P}$ such that $p(\mathbf{gf}) = p^*(\mathbf{g})$ for all $\mathbf{g} \in \mathfrak{G}$. In words, the action $\mathbf{f}$ affects $p(\cdot) \in \mathcal{P}$, but only in the sense of replacing $p(\cdot)$ by another distribution $p^*(\cdot) \in \mathcal{P}$.

There is a slight asymmetry in the definitions of an invariant function and an invariant set of distributions, with $\mathbf{f}$ on the left in the first definition and on the right in the second. That is, $t : \mathfrak{G} \to \mathbb{R}$ is invariant if $t(\mathbf{g}) = t(\mathbf{fg})$ for each $\mathbf{f} \in \mathfrak{F}$, but $\mathcal{P}$ is invariant if for each $p \in \mathcal{P}$ and each $\mathbf{f} \in \mathfrak{F}$ there exist a function $p^* \in \mathcal{P}$, where p^* depends upon p and $\mathbf{f}$, such that $p(\mathbf{gf}) = p^*(\mathbf{g})$ for all $\mathbf{g} \in \mathfrak{G}$.

In a trivial sense, the set consisting of the randomization distribution, $\mathcal{P}_1 = \{\overline{p}\}$, is invariant with respect to every subgroup $\mathfrak{F}$ because $\overline{p}(\mathbf{g}) = |\mathfrak{G}|^{-1}$ for all $\mathbf{g} \in \mathfrak{G}$ so $\overline{p}(\mathbf{g}) = \overline{p}(\mathbf{gf})$ for every $\mathbf{f} \in \mathfrak{G}$. Consider a small but nontrivial example. In Table 2, the set $\mathcal{P}$ consisting of the two distributions (iv) and (v) is invariant with respect to $\mathfrak{H}$; that is, distribution (iv) is transformed into distribution (v) by multiplying on the right by $\mathbf{h}_2$ in column 2 of Table 2.

The set $\mathcal{P}_{\mathbf{\Gamma}}$ of probability distributions on $\mathfrak{G}$ in Section 2.3 is invariant with respect to the subgroup $\mathfrak{K}$ that permutes treatment assignments within pairs. For $\mathbf{\Gamma} \ge \mathbf{1}$, an individual distribution $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$ is not typ-

ically invariant—it pushes some people toward smoking, and some smokers to smoke more—but for any permutation of people within pairs, $\mathbf{k} \in \mathfrak{K}$, there is another distribution $p^*(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$ so that the effect of $\mathbf{k}$ is undone. That is, many of the individual distributions $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$ are biased, not symmetric, but the set of biases $\mathcal{P}_{\mathbf{\Gamma}}$ under consideration is symmetric.

3.3 Sampling a Permutation from a Knit Product

Suppose that $\mathfrak{G}$ is the knit product of two of its subgroups $\mathfrak{H}$ and $\mathfrak{K}$. If we sample a random $\mathbf{G} \in \mathfrak{G}$ such that $\Pr(\mathbf{G} = \mathbf{g}) = p(\mathbf{g})$ for $\mathbf{g} \in \mathfrak{G}$ where $p(\cdot)$ is a particular distribution in $\mathcal{P}_{\mathbf{\Gamma}}$, then we have implicitly sampled an $\mathbf{H} \in \mathfrak{H}$ and a $\mathbf{K} \in \mathfrak{K}$ such that $\mathbf{G} = \mathbf{HK}$. For this particular $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$, the marginal distribution of $\mathbf{K}$ is $\Pr(\mathbf{K} = \mathbf{k}) = p_{\mathbf{K}}(\mathbf{k}) = \sum_{\mathbf{h}\in\mathfrak{H}} p(\mathbf{hk})$, and the conditional distribution of $\mathbf{H}$ given $\mathbf{K} = \mathbf{k}$ is

$$\Pr(\mathbf{H} = \mathbf{h} | \mathbf{K} = \mathbf{k}) = p_{\mathbf{H}|\mathbf{k}}(\mathbf{h}) = \frac{p(\mathbf{hk})}{p_{\mathbf{K}}(\mathbf{k})}.$$

If $\mathcal{P}_{\mathbf{\Gamma}}$ is a set of distributions of $\mathbf{G}$ on $\mathfrak{G}$, write $\mathcal{P}_{\mathbf{\Gamma},\mathbf{K}}$ for the corresponding set of marginal distributions on $\mathfrak{K}$, so $p_{\mathbf{K}}(\cdot) \in \mathcal{P}_{\mathbf{\Gamma},\mathbf{K}}$ if and only if there exists a $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$ such that $p_{\mathbf{K}}(\mathbf{k}) = \sum_{\mathbf{h}\in\mathfrak{H}} p(\mathbf{hk})$.

In parallel, for each $\mathbf{k}$, write $\mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$ for the set of conditional distributions $p_{\mathbf{H}|\mathbf{k}}(\cdot)$ of $\mathbf{H}$ given $\mathbf{K} = \mathbf{k}$ corresponding to $\mathcal{P}_{\mathbf{\Gamma}}$, so $q(\cdot) \in \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$ if and only if there is a $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$ such that $q(\mathbf{h}) = p(\mathbf{hk}) / \sum_{\mathbf{h}\in\mathfrak{H}} p(\mathbf{hk})$. Write $\mathcal{C}_{\mathbf{\Gamma}} = \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{I}}$ for the set of conditional distributions $p_{\mathbf{H}|\mathbf{I}}(\cdot)$ of $\mathbf{H}$ given $\mathbf{K} = \mathbf{I}$. The set $\mathcal{C}_{\mathbf{\Gamma}}$ will play a distinctive role.

Distributions $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$ have support $\mathfrak{G}$. Distributions $p_{\mathbf{K}}(\cdot) \in \mathcal{P}_{\mathbf{\Gamma},\mathbf{K}}$ have support $\mathfrak{K}$. Distributions $p_{\mathbf{H}|\mathbf{k}}(\cdot) \in \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$ have support $\mathfrak{H}$ for each $\mathbf{k} \in \mathfrak{K}$.

3.4 Joint Bounds on P-Values

A bivariate random variable (P_1, P_2) is stochastically larger than uniform if

$$\Pr(P_1 \le \alpha_1, P_2 \le \alpha_1) \le \alpha_1\alpha_2 \quad \text{for all } \alpha_1, \alpha_2 \in [0, 1];$$

see Brannath, Posch and Bauer (2002). Here, (P_1, P_2) may be dependent.

In Section 2.3, one sensitivity analysis applied an M-statistic to the boxplot in Figure 1(i), and another sensitivity analysis used the cross-cut statistic to look for dependence in Figure 1(ii). When can the upper bounds on P-values from two such sensitivity analyses be viewed as providing two (essentially) independent tests of the one null hypothesis of no effect of smoking on periodontal disease? Proposition 1 provides a general answer.

Let $t_1 : \mathfrak{G} \to \mathbb{R}$ and $t_2 : \mathfrak{G} \to \mathbb{R}$ be two test statistics, where $t_1(\cdot)$ is $\mathfrak{H}$-invariant. Define

$$\overline{\overline{b}}_{\mathbf{\Gamma}1}(a) = \max_{q(\cdot)\in\mathcal{P}_{\mathbf{\Gamma},\mathbf{K}}} \sum_{\mathbf{k}\in\mathfrak{K}} \chi\{t_1(\mathbf{k}) \ge a\} q(\mathbf{k}) \quad \text{and}$$
$$\overline{\overline{P}}_1 = \overline{\overline{b}}_{\mathbf{\Gamma}1}\{t_1(\mathbf{K})\}, \tag{3}$$

so $\overline{\overline{P}}_1$ is the upper bound on the P-value using $t_1(\mathbf{K})$ alone and using the marginal distribution of $\mathbf{K}$. Note that, in principle, $\overline{\overline{b}}_{\mathbf{\Gamma}1}(a)$ in (3) differs from $\overline{\overline{b}}_{\mathbf{\Gamma}}(a)$ in (1) because (3) is a maximum over the possible marginal distributions of $\mathbf{K} \in \mathfrak{K}$, whereas (1) is a maximum over the possible joint distributions of $(\mathbf{H}, \mathbf{K})$ with $\mathbf{HK} = \mathbf{G} \in \mathfrak{G}$.

Define

$$\overline{\overline{b}}_{\mathbf{\Gamma}2}(a) = \max_{q(\cdot)\in\mathcal{C}_{\mathbf{\Gamma}}} \sum_{\mathbf{h}\in\mathfrak{H}} \chi\{t_2(\mathbf{hk}) \ge a\} q(\mathbf{h}) \quad \text{and}$$
$$\overline{\overline{P}}_2 = \overline{\overline{b}}_{\mathbf{\Gamma}2}\{t_2(\mathbf{HK})\},$$

where $\mathcal{C}_{\mathbf{\Gamma}} = \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{I}}$ was defined in Section 3.3. In Section 6.2, Lemma 4 shows that

$$\overline{\overline{b}}_{\mathbf{\Gamma}2}(a) = \max_{p_{\mathbf{H}|\mathbf{k}}(\cdot)\in\mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}} \sum_{\mathbf{h}\in\mathfrak{H}} \chi\{t_2(\mathbf{hk}) \ge a\} p_{\mathbf{H}|\mathbf{k}}(\mathbf{h}) \tag{4}$$

implying that $\overline{\overline{P}}_2$ is the upper bound on the P-value from $t_2(\mathbf{HK})$ conditionally given $\mathbf{K} = \mathbf{k}$. Note that, in principle, $\overline{\overline{b}}_{\mathbf{\Gamma}2}(a)$ in (4) differs from $\overline{\overline{b}}_{\mathbf{\Gamma}}(a)$ in (1) because (4) is a maximum over the possible conditional distributions of $\mathbf{H} \in \mathfrak{H}$ given $\mathbf{K} = \mathbf{k}$, whereas (1) is a maximum over the possible joint distributions of $(\mathbf{H}, \mathbf{K})$ with $\mathbf{HK} = \mathbf{G} \in \mathfrak{G}$. Under certain conditions, these two maxima turn out to be equal.

In Section 2.3, the statistic $t_1(\cdot)$ was either Wilcoxon's signed rank test or a one-sample M-test comparing smokers and controls within pairs, and the bound $\overline{\overline{b}}_{\mathbf{\Gamma}1}\{t_1(\mathbf{K})\}$ in (3) was a standard sensitivity bound for matched pairs for the group $\mathfrak{K}$ of permutations within pairs. In Section 2.3, the statistic $t_2(\cdot)$ was the cross-cut statistic relating dose and response, that is, the quantity smoked and extent of periodontal disease. The sensitivity bound $\overline{\overline{b}}_{\mathbf{\Gamma}2}(a)$ in (4) ignored the process that made some people into smokers and others into nonsmokers, acting as if that were fixed, and was a standard sensitivity bound for the group $\mathfrak{H}$ of permutations of doses among the smokers. Can we safely act as if these two sensitivity analysis were from two independent studies by different investigators despite the fact that they were computed from the same data?

PROPOSITION 1. *Let $\mathfrak{G}$ be the knit product of its subgroups $\mathfrak{H}$ and $\mathfrak{K}$, and sample $\mathbf{G} = \mathbf{HK}$ with unknown distribution $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$. Suppose that $t_1(\cdot)$ is $\mathfrak{H}$-invariant and $\mathcal{P}_{\mathbf{\Gamma}}$ is $\mathfrak{K}$-invariant. Then, under H_0, the pair of sensitivity bounds $(\overline{\overline{P}}_1, \overline{\overline{P}}_2)$ is stochastically larger than uniform for every $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$.*

Proposition 1 is applied in Section 3.5 and proved in Section 6.

3.5 Application to the Periodontal Data

In the periodontal data in Section 2.3, $\overline{\overline{P}}_1$ is from the M-test and $\overline{\overline{P}}_2$ is from the crosscut test. If the null hypothesis H_0 of no treatment effect is true, and if the bias in treatment assignment is at most $\mathbf{\Gamma} = (\Gamma_1, \Gamma_2)$, then Proposition 1 implies that the P-value bound, $(\overline{\overline{P}}_1, \overline{\overline{P}}_2)$, is stochastically larger than the uniform distribution on the unit square. It follows that $\overline{\overline{P}}_1$ and $\overline{\overline{P}}_2$ can be combined by methods for combining independent P-values, such as Fisher's method; see Rosenbaum (2011), Lemma 1. Sensitivity analyses often produce P-value bounds that are much larger than uniform, with the consequence that Fisher's method is not the best method for combining them. Hsu, Small and Rosenbaum (2013) found that the truncated product method of Zaykin et al. (2002) often has better power than Fisher's method when used in sensitivity analyses. Where Fisher's method uses the product of the P-values, Zaykin et al. (2002) take the product of only those P-values less than or equal to a truncation point, τ.

TABLE 3

Sensitivity analysis combining an M-test for symmetry of matched pairs with sensitivity parameter Γ_1 and a cross-cut test for dependence upon the amount smoked with sensitivity parameter Γ_2. The table gives the upper bound on the one-sided P-value testing no treatment effect, combining the two separate P-value bounds using Zaykin et al.'s (2002) *truncated product of P-values with truncation $\tau = 0.05$. The largest P-values ≤ 0.05 in a row or column are in* ***bold***

		Γ_2					
		1	**1.4**	**1.5**	**1.6**	**2**	∞
Γ_1	1	0.000	0.000	0.000	0.000	0.000	0.000
	2	0.000	0.000	0.000	0.000	0.000	0.000
	3	0.000	0.001	0.001	0.001	0.013	0.013
	3.3	0.000	0.002	0.003	0.004	**0.046**	**0.046**
	3.5	0.001	0.004	**0.005**	**0.007**	0.095	0.095
	3.6	0.007	**0.043**	0.062	0.087	1.000	1.000
	4	0.007	**0.043**	0.062	0.087	1.000	1.000
	∞	0.007	**0.043**	0.062	0.087	1.000	1.000

For the periodontal data, Table 3 combines the P-value bounds $(\overline{\overline{P}}_1, \overline{\overline{P}}_2)$ for the M-statistic and crosscut statistic, using a truncation of $\tau = 0.05$. If either factor is not biased—if either $\Gamma_1 = 1$ or $\Gamma_2 = 1$—then even infinite biases affecting the other factor are insufficient to lead to acceptance of H_0. Indeed, neither a bias of $(\Gamma_1, \Gamma_2) \leq (3.3, \infty)$ nor a bias of $(\Gamma_1, \Gamma_2) \leq (\infty, 1.4)$ would lead to acceptance of H_0 at the 0.05 level. The upper bound on the pooled P-value is ≤ 0.007 for $(\Gamma_1, \Gamma_2) \leq (3.5, 1.6)$.

In Table 3, the associations in Figures 1(i) and 1(ii) concur, providing mutually supporting evidence against the null hypothesis of no treatment effect. Table 3 is statistical evidence analogous to Peirce's cable of several fibers.

4. OTHER KNIT PRODUCTS IN OBSERVATIONAL STUDIES

The group in (2) and its larger version for Figure 1 are both wreath products, or highly structured, highly symmetrical semidirect products. Many observational studies are not that symmetrical, and so are not well described by wreath products, but nonetheless they are described by semidirect or knit products. This section briefly mentions a few examples.

In Figure 1, the group $\mathfrak{H}$ permutes $N/2$ pairs in all $(N/2)!$ ways. We might wish to put similar pairs in the same stratum, and permute pairs within strata but not across strata. The two people in the same pair are the same or close in terms of observed covariates, but different pairs may differ substantially in terms of observed covariates. Perhaps permutations of whole pairs should be restricted so as to permute pairs with similar values of observed covariates. For instance, periodontal disease increases markedly with age. In Figure 1, there are 78 pairs in which both individuals are more than 60 years old, and 362 pairs in which both individuals are at most 60 years old. (There is one pair in which the smoker is under 60 and the control is over 60, and it is natural to exclude that pair from the following computation.) The group $\mathfrak{H}'$ that permutes pairs within these two age strata has $78! \times 362!$ permutations, rather than the $441!$ permutations in $\mathfrak{H}$, and $\mathfrak{H}'$ is isomorphic to the direct product of two symmetric groups. Permuting pairs within two age strata has only a slight effect on the crosscut test and its sensitivity analysis in Figure 1, perhaps because matched pair, smoker-minus-control differences were permuted, for smokers and controls of similar age. In principle, Proposition 1 permits the responses of smokers to be permuted among

pairs without differencing, and in this case stratification on covariates might be more important. Stratified and adaptive use of the crosscut statistics is discussed in Rosenbaum and Small (2017).

In the NHANES data, cigarette smoking is uncommon among people with at least a BA degree, but is more common among people without a BA. Before matching, there were 44 smokers and 574 nonsmokers with a BA degree, a ratio of more than 13-to-1, but they became just 44 matched pairs in Figure 1. If, instead, these 44 smokers are each matched to 5 nonsmoking controls, then the study has $397 = 441 - 44$ matched pairs and 44 matched sets with a smoker and 5 controls. In this enlarged study, the group $\mathfrak{K}$ permutes individuals within a matched pair or set, while the group $\mathfrak{H}$ permutes matched sets of the same size among the doses of smoking observed for sets of that size. Note that this $\mathfrak{H}$ is isomorphic to a subgroup of the group used for Figure 1, because permutations of doses are restricted to matched sets of the same size. This larger study is not much different from Figure 1, perhaps because there were only 44 smokers with a BA degree. If the M-test in Section 2.3 is used, the upper bound on the P-value is 0.045 at $\Gamma = 2.85$, only slightly less sensitive than the pairs in Figure 1. For a discussion of matching with variable controls, see Pimentel, Yoon and Keele (2015) and the references given there.

An alternative form of matching, called full matching, can use all available nonsmokers in NHANES by permitting a matched set to have one smoker and one or more controls, or else one control and one or more smokers; see Rosenbaum (1991), Hansen and Klopfer (2006), Stuart and Green (2008), Austin and Stuart (2015). Here, again, $\mathfrak{K}$ permutes individuals within a matched set while $\mathfrak{H}$ permutes matched sets of the same structure.

Sometimes treatment is denied to people by several processes acting in sequence, and the biases that affect these different processes are likely to be different. For instance, a person might not receive a stent following a myocardial infarction because the person was rushed to a hospital that lacks the facilities to implant a stent (treatment α), or because the person was treated at a hospital that could implant a stent but this particular person was judged to be a poor candidate for this treatment (treatment β); alternatively, the person might receive a stent in a hospital that can implant one (treatment γ). This two-step process can yield matched sets with two evidence factors permuting individuals within matched sets, one factor examining the effects of treatment at a hospital that cannot implant stents (α versus $\{\beta, \gamma\}$), the other factor examining the effects of not receiving a stent at such a hospital (β versus γ). In a matched triple with one person receiving each of $\{\alpha, \beta, \gamma\}$, there are $3! = 6$ possible treatment assignments; that is, $\mathfrak{G}$ for one matched set is the symmetric group on three letters. Moreover, $\mathfrak{G}$ is the knit product of $\mathfrak{H}$ and $\mathfrak{K}$ below:

$$\mathfrak{H} = \left\{ \mathbf{h}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{h}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \right\},$$

$$\mathfrak{K} = \left\{ \mathbf{k}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{k}_2 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \right. \tag{5}$$

$$\left. \mathbf{k}_3 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \right\}.$$

Here, $\mathfrak{H}$ permutes assignment to β or γ, while $\mathfrak{K}$ picks one person for α. A statistic that is invariant with respect to $\mathfrak{H}$ compares α versus $\{\beta, \gamma\}$, ignoring the distinction between β and γ. A statistic that conditions on $\mathbf{K} \in \mathfrak{K}$ fixes the identity of the person receiving α, so that $\mathbf{H} \in \mathfrak{H}$ permutes the other two people among treatments β and γ. This situation here is similar to situations discussed by Dwass (1960) and Marden (1992), but it does not require the use of rank tests, and it provides for a sensitivity analysis for the two component tests. With $N/3$ matched triples of this form, the relevant group is the direct product of $N/3$ groups each of which is a knit product of $\mathfrak{H}$ and $\mathfrak{K}$ in (5). Although this illustration has one α, one β and one γ in each matched set, other designs are possible. See Rosenbaum (2011) for a discussion of evidence factors of this form with an example.

5. DISCUSSION

In some settings, for instance in slightly idealized randomized trials, the only source of uncertainty comes from a finite sample size. All uncertainty would be resolved by a sufficient increase in sample size. In these settings, we naturally focus on making the most efficient use of a finite sample.

In other settings, for instance in observational studies of treatment effects, increasing the sample size reduces sampling uncertainty, but leaves in place other sources of uncertainty. Many large observational studies reach conclusions that remain controversial and doubtful, and the doubts would not be resolved by increasing the sample size or by exactly replicating the study. In these settings, increasing the sample size from

one source of evidence confers limited benefits, and what is needed is several sources of evidence with different limitations. Most valuable are several sources of evidence that could concur, but would not naturally concur in the absence of a treatment effect. Evidence factors arise in this context.

Proposition 1 offers a general description of the structure of evidence factors in terms of the knit product of permutation groups. Proposition 1 is simpler, yet more general than earlier results, in the following senses: (i) it is not tied to rank tests, (ii) it applies to both a pair of randomization tests and to a pair of sensitivity analyses, (iii) it is not confined to a few limited cases of permutation distributions. The knit product construction resembles the orthogonality of nominal factors that commonly arises in the theory of structured experimental designs.

6. PROOFS

The proof of Proposition 1 in Section 6.4 is a straightforward consequence of several lemmas. Throughout the proofs in this section, $\mathfrak{G}$ is the knit product of $\mathfrak{H}$ and $\mathfrak{K}$.

6.1 Ignoring One Factor in a Knit Product: Lemma 2

In Section 2.3, the Wilcoxon signed rank statistic and the M-statistic ignored the action of $\mathfrak{H}$ that permutes the number of cigarettes smoked. In distribution (iv) of Table 2, $p(\mathbf{hk})$ varies with both $\mathbf{k}$ and $\mathbf{h}$, so that person $i = 1$ in Table 1 is most likely to smoke and likely to smoke 40 cigarettes per day. How does a sensitivity analysis that uses $\mathbf{K}$ alone, ignoring $\mathbf{H}$, compare to a sensitivity analysis that uses $(\mathbf{H}, \mathbf{K})$ or $\mathbf{G} = \mathbf{HK}$ jointly? Lemma 2 says these two sensitivity analyses are the same if the test statistic is invariant with respect to $\mathfrak{H}$, as is true of the Wilcoxon and M-statistics. In the statement of Lemma 2, the two maxima are over different sets of distributions, the joint distributions in $\mathcal{P}_{\mathbf{\Gamma}}$ and the marginal distributions in $\mathcal{P}_{\mathbf{\Gamma},\mathbf{K}}$.

LEMMA 2. *If the function $t : \mathfrak{G} \to \mathbb{R}$ is invariant with respect to $\mathfrak{H}$, then*

$$\max_{p\in\mathcal{P}_{\mathbf{\Gamma}}} \sum_{\mathbf{h}\in\mathfrak{H}} \sum_{\mathbf{k}\in\mathfrak{K}} \chi\{t(\mathbf{hk}) \geq a\} p(\mathbf{hk}) = \max_{p_{\mathbf{K}}(\cdot)\in\mathcal{P}_{\mathbf{\Gamma},\mathbf{K}}} \sum_{\mathbf{k}\in\mathfrak{K}} \chi\{t(\mathbf{k}) \geq a\} p_{\mathbf{K}}(\mathbf{k}).$$

PROOF. Because $t : \mathfrak{G} \to \mathbb{R}$ is invariant with respect to $\mathfrak{H}$, it follows that $t(\mathbf{hk}) = t(\mathbf{k})$ for each $\mathbf{h} \in \mathfrak{H}$. So,

$$\begin{aligned} &\max_{p\in\mathcal{P}_{\mathbf{\Gamma}}} \sum_{\mathbf{h}\in\mathfrak{H}} \sum_{\mathbf{k}\in\mathfrak{K}} \chi\{t(\mathbf{hk}) \geq a\} p(\mathbf{hk}) \\ &\quad = \max_{p\in\mathcal{P}_{\mathbf{\Gamma}}} \sum_{\mathbf{k}\in\mathfrak{K}} \chi\{t(\mathbf{k}) \geq a\} \sum_{\mathbf{h}\in\mathfrak{H}} p(\mathbf{hk}) \\ &\quad = \max_{p_{\mathbf{K}}(\cdot)\in\mathcal{P}_{\mathbf{\Gamma},\mathbf{K}}} \sum_{\mathbf{k}\in\mathfrak{K}} \chi\{t(\mathbf{k}) \geq a\} p_{\mathbf{K}}(\mathbf{k}). \end{aligned} \tag{6}$$

□

6.2 Fixing a Factor in a Knit Product: Lemma 3

If $\mathbf{G} = \mathbf{HK}$ is sampled from the knit product $\mathfrak{G}$ of $\mathfrak{H}$ and $\mathfrak{K}$ according to $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$, then $\mathbf{H}$ and $\mathbf{K}$ may be statistically dependent, so that, in general, $p_{\mathbf{H}|\mathbf{k}}(\cdot) \neq p_{\mathbf{H}|\mathbf{k}'}(\cdot)$ for $\mathbf{k} \neq \mathbf{k}'$. Lemma 3 gives a condition such that the set $\mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$ of conditional distributions does not depend upon $\mathbf{k}$, even though individual distributions $p_{\mathbf{H}|\mathbf{k}}(\cdot) \in \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$ do depend upon $\mathbf{k}$. This condition is that the set $\mathcal{P}_{\mathbf{\Gamma}}$ of probability distributions is invariant with respect to $\mathfrak{K}$. Stated informally, if you knew that $\Pr(\mathbf{G} = \mathbf{g}) = p(\mathbf{g})$ for one specific $p(\cdot)$, then observing $\mathbf{K} = \mathbf{k}$ might change your opinion about the likely value of $\mathbf{H}$ and hence also of $\mathbf{G} = \mathbf{HK}$; however, if you knew only that $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$ where $\mathcal{P}_{\mathbf{\Gamma}}$ is invariant with respect to $\mathfrak{K}$, then observing $\mathbf{K} = \mathbf{k}$ would not change your opinion about the likely value of $\mathbf{H}$ or $\mathbf{G}$. Recall that $\mathcal{C}_{\mathbf{\Gamma}} = \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{I}}$ was defined in Section 3.3.

LEMMA 3. *If the set $\mathcal{P}_{\mathbf{\Gamma}}$ of probability distributions is invariant with respect to $\mathfrak{K}$, then the corresponding set $\mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$ of conditional distributions of $p_{\mathbf{H}|\mathbf{k}}(\cdot)$ of $\mathbf{H}$ given $\mathbf{K} = \mathbf{k}$ is the same for all $\mathbf{k} \in \mathfrak{K}$; that is, $\mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}} = \mathcal{C}_{\mathbf{\Gamma}}$ all $\mathbf{k} \in \mathfrak{K}$.*

PROOF. The proof consists of showing first that $\mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}} \subseteq \mathcal{C}_{\mathbf{\Gamma}}$ for each $\mathbf{k} \in \mathfrak{K}$, and second that $\mathcal{C}_{\mathbf{\Gamma}} \subseteq \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$ for each $\mathbf{k} \in \mathfrak{K}$. Fix one $\mathbf{k} \in \mathfrak{K}$. Suppose $q(\cdot) \in \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$ so $q(\cdot)$ has support $\mathfrak{H}$. Then, by the definition of $\mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$, there is a joint distribution $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$ such that $q(\mathbf{h}) = p(\mathbf{hk}) / \sum_{\mathbf{h}\in\mathfrak{H}} p(\mathbf{hk})$. Because $\mathcal{P}_{\mathbf{\Gamma}}$ is invariant with respect to $\mathfrak{K}$, there exists a $p^*(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$ such that $p(\mathbf{gk}) = p^*(\mathbf{g})$ for all $\mathbf{g} \in \mathfrak{G}$. Hence, $q(\mathbf{h}) = p^*(\mathbf{h}) / \sum_{\mathbf{h}\in\mathfrak{H}} p^*(\mathbf{h}) \in \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{I}} = \mathcal{C}_{\mathbf{\Gamma}}$, so $\mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}} \subseteq \mathcal{C}_{\mathbf{\Gamma}}$. Conversely, suppose $q(\cdot) \in \mathcal{C}_{\mathbf{\Gamma}} = \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{I}}$. Then there exists a $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$ such that $q(\mathbf{h}) = p(\mathbf{h}) / \sum_{\mathbf{h}\in\mathfrak{H}} p(\mathbf{h})$. Because $\mathcal{P}_{\mathbf{\Gamma}}$ is invariant with respect to $\mathfrak{K}$, there exists a $p^*(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$ such that $p(\mathbf{gk}^{-1}) = p^*(\mathbf{g})$ for all $\mathbf{g} \in \mathfrak{G}$. Hence, $p(\mathbf{h}) = p(\mathbf{hkk}^{-1}) = p^*(\mathbf{hk})$ for all $\mathbf{g} = \mathbf{hk} \in \mathfrak{G}$, so $q(\mathbf{h}) = p^*(\mathbf{hk}) / \sum_{\mathbf{h}\in\mathfrak{H}} p^*(\mathbf{hk})$, so $q(\cdot) \in \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$. Therefore $\mathcal{C}_{\mathbf{\Gamma}} \subseteq \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$. □

Lemma 4 says that acting as if $\mathbf{K}$ were fixed, as we did with the cross-cut analysis in Section 2.3, is the

same as conditioning on the observed value of $\mathbf{K}$ providing the set $\mathcal{P}_{\mathbf{\Gamma}}$ of probability distributions is invariant with respect to $\mathfrak{K}$.

LEMMA 4. *If the set $\mathcal{P}_{\mathbf{\Gamma}}$ of probability distributions is invariant with respect to $\mathfrak{K}$, then for each fixed $\mathbf{k} \in \mathfrak{K}$,*

$$\begin{aligned}
&\max_{p\in\mathcal{P}_{\mathbf{\Gamma}}} \frac{\sum_{\mathbf{h}\in\mathfrak{H}} \chi\{t(\mathbf{hk}) \geq a\} p(\mathbf{hk})}{\sum_{\mathbf{h}\in\mathfrak{H}} p(\mathbf{hk})} \\
&\quad = \max_{p_{\mathbf{H}|\mathbf{k}}(\cdot)\in\mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}} \sum_{\mathbf{h}\in\mathfrak{H}} \chi\{t(\mathbf{hk}) \geq a\} p_{\mathbf{H}|\mathbf{k}}(\mathbf{h}) \qquad (7) \\
&\quad = \max_{q(\cdot)\in\mathcal{C}_{\mathbf{\Gamma}}} \sum_{\mathbf{h}\in\mathfrak{H}} \chi\{t(\mathbf{hk}) \geq a\} q(\mathbf{h}).
\end{aligned}$$

PROOF. By definition, $p_{\mathbf{H}|\mathbf{k}}(\cdot) \in \mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}}$ if and only if there exists a $p \in \mathcal{P}_{\mathbf{\Gamma}}$ such that $p_{\mathbf{H}|\mathbf{k}}(\mathbf{h}) = p(\mathbf{hk})/\sum_{\mathbf{h}'\in\mathfrak{H}} p(\mathbf{hk}')$, proving the first equality in (7). By Lemma 3, $\mathcal{P}_{\mathbf{\Gamma},\mathbf{H}|\mathbf{k}} = \mathcal{C}_{\mathbf{\Gamma}}$, proving the second equality. □

6.3 Joint Bounds on P-Values: Lemma 5

Lemma 5 is elementary and is a special case of Lemma 3 in Rosenbaum (2011).

LEMMA 5. *Suppose P_1 is a function of $\mathbf{K}$, and P_2 is a function of $(\mathbf{H}, \mathbf{K})$ such that $\Pr(P_1 \leq \alpha_1) \leq \alpha_1$ and $\Pr(P_2 \leq \alpha_2 | \mathbf{K} = \mathbf{k}) \leq \alpha_2$ for each $\mathbf{k}$, for all $0 \leq \alpha_1 \leq 1$ and $0 \leq \alpha_2 \leq 1$. Then (P_1, P_2) is stochastically larger than uniform.*

PROOF.

$$\begin{aligned}
&\Pr(P_1 \leq \alpha_1, P_2 \leq \alpha_1) \\
&\quad = \mathrm{E}\big[\mathrm{E}\big\{\chi(P_1 \leq \alpha_1)\chi(P_2 \leq \alpha_2)|\mathbf{K}\big\}\big], \\
&\mathrm{E}\big[\chi(P_1 \leq \alpha_1)\mathrm{E}\big\{\chi(P_2 \leq \alpha_2)|\mathbf{K}\big\}\big] \\
&\quad \leq \alpha_2 \mathrm{E}\big[\chi(P_1 \leq \alpha_1)\big] \leq \alpha_1\alpha_2.
\end{aligned}$$ □

6.4 Proof of Proposition 1

PROOF. By assumption, $\mathbf{G} = \mathbf{HK}$ has been sampled with unknown true distribution $p(\cdot) \in \mathcal{P}_{\mathbf{\Gamma}}$. Define the unknown random variables

$$P_1 = \sum_{\mathbf{k}\in\mathfrak{K}} \chi\{t_1(\mathbf{k}) \geq t_1(\mathbf{K})\} \sum_{\mathbf{h}\in\mathfrak{H}} p(\mathbf{hk})$$

and

$$P_{2,\mathbf{k}} = \sum_{\mathbf{h}\in\mathfrak{H}} \chi\{t_2(\mathbf{hk}) \geq t_2(\mathbf{kH})\} \frac{p(\mathbf{hk})}{\sum_{\mathbf{h}'\in\mathfrak{H}} p(\mathbf{h}'\mathbf{k})}$$

and

$$P_2 = P_{2,\mathbf{K}}.$$

Then (P_1, P_2) is stochastically larger than uniform by Lemma 5. By Lemma 2, $P_1 \leq \overline{\overline{P}}_1$. By Lemma 4, $P_2 \leq \overline{\overline{P}}_2$. So the bound we can calculate from data, $(\overline{\overline{P}}_1, \overline{\overline{P}}_2)$, is stochastically larger than uniform. □

REFERENCES

ALAM, K. (1974). Some nonparametric tests of randomness. *J. Amer. Statist. Assoc.* **69** 738–739. MR0370907

ATEŞ, F. and ÇEVIK, A. S. (2009). Knit products of some groups and their applications. *Rend. Semin. Mat. Univ. Padova* **121** 1–11. MR2542131

AUSTIN, P. C. and STUART, E. A. (2015). Optimal full matching for survival outcomes: A method that merits more widespread use. *Stat. Med.* **34** 3949–3967. MR3431315

BAILEY, R. A., PRAEGER, C. E., ROWLEY, C. A. and SPEED, T. P. (1983). Generalized wreath products of permutation groups. *Proc. Lond. Math. Soc.* (3) **47** 69–82. MR0698928

BELL, C. B. and HALLER, H. S. (1969). Bivariate symmetry tests: Parametric and nonparametric. *Ann. Math. Stat.* **40** 259–269. MR0235658

BRANNATH, W., POSCH, M. and BAUER, P. (2002). Recursive combination tests. *J. Amer. Statist. Assoc.* **97** 236–244. MR1947283

BRIEN, C. J. and BAILEY, R. A. (2006). Multiple randomizations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 571–609. MR2301010

CENTERS FOR DISEASE CONTROL (2016). Smoking, gum disease, and tooth loss. Available at https://www.cdc.gov/tobacco/campaign/tips/diseases/periodontal-gum-disease.html.

COCHRAN, W. G. (1965). The planning of observational studies of human populations (with discussion). *J. Roy. Statist. Soc. Ser. A* **128** 234–266.

CONLON, J. C., LEON, R., PROSCHAN, F. and SETHURAMAN, J. (1977). G-Ordered functions, with applications in statistics. I, II. Technical Report M432, M433, Dept. Statistics, Florida State Univ. Tallahassee, FL. Available at http://www.dtic.mil/get-tr-doc/pdf?AD=ADA049316, http://www.dtic.mil/get-tr-doc/pdf?AD=ADA046584.

CORNFIELD, J., HAENSZEL, W., HAMMOND, E., LILIENFELD, A., SHIMKIN, M. and WYNDER, E. (1959). Smoking and lung cancer. *J. Nat. Cancer Inst.* **22** 173–203. Reprinted in *Internat. J. Epidemiol.* **38** (2009) 1175–1201. With discussion by D. R. Cox, J. Vandenbroucke, M. Zwahlen and J. B. Greenhouse.

COX, D. R. and REID, N. (2000). *The Theory of the Design of Experiments*. Chapman and Hall/CRC Press, London. DOI:10.1002/sim.1089.

DAWID, A. P. (1985). Invariance and independence in multivariate distribution theory. *J. Multivariate Anal.* **17** 304–315. MR0813238

DAWID, A. P. (1988). Symmetry models and hypotheses for structured data layouts. *J. Roy. Statist. Soc. Ser. B* **50** 1–34. MR0954729

DWASS, M. (1960). Some k-sample rank-order tests. In *Contributions to Probability and Statistics* 198–202. Stanford Univ. Press, Stanford, CA. MR0120705

EATON, M. L. (1982). A review of selected topics in multivariate probability inequalities. *Ann. Statist.* **10** 11–43. MR0642717

EATON, M. L. and PERLMAN, M. D. (1977). Reflection groups, generalized Schur functions, and the geometry of majorization. *Ann. Probab.* **5** 829–860. MR0444864

EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 403–417. MR0312660

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

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, N. D. and WAZZAN, S. (2008). Zappa–Szép products of bands and groups. *Semigroup Forum* **77** 438–455. MR2457329

HAMMOND, E. C. (1964). Smoking in relation to mortality and morbidity: Findings in first thirty-four months of follow-up in a prospective study started in 1959. *J. Natl. Cancer Inst.* **32** 1161–1188.

HANSEN, B. B. and KLOPFER, S. O. (2006). Optimal full matching and related designs via network flows. *J. Comput. Graph. Statist.* **15** 609–627. MR2280151

HOSMAN, C. A., HANSEN, B. B. and HOLLAND, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Stat.* **4** 849–870. MR2758424

HSU, J. Y., SMALL, D. S. and ROSENBAUM, P. R. (2013). Effect modification and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **108** 135–148. MR3174608

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York. MR0606374

IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93** 126–132.

ISAACS, I. M. (2009). *Algebra*: *A Graduate Course*. *Graduate Studies in Mathematics* **100**. Amer. Math. Soc., Providence, RI. Reprint of the 1994 original. MR2472787

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

LIU, W., KURAMOTO, S. J. and STUART, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prev. Sci.* **14** 570–580.

MARDEN, J. I. (1992). Use of nested orthogonal contrasts in analyzing rank data. *J. Amer. Statist. Assoc.* **87** 307–318. MR1173802

MARITZ, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66** 163–166. MR0529161

MCCANDLESS, L. C., GUSTAFSON, P. and LEVY, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Stat. Med.* **26** 2331–2347. MR2368419

PEIRCE, C. S. (1868). Some consequences of four incapacities. *J. Specul. Philos.* **2** 140–157. Reprinted in R. B. Talisse and S. F. Aikin, eds. (2011). *The Pragmatism Reader: From Peirce through the Present*. Harvard Univ. Press, Cambridge, MA.

PIMENTEL, S. D., YOON, F. and KEELE, L. (2015). Variable-ratio matching with fine balance in a study of the Peer Health Exchange. *Stat. Med.* **34** 4070–4082. MR3431322

RANDLES, R. H. and HOGG, R. V. (1971). Certain uncorrelated statistics and independent rank statistics. *J. Amer. Statist. Assoc.* **66** 569–574.

ROMAN, S. (2012). *Fundamentals of Group Theory. An Advanced Approach*. Birkhäuser/Springer, New York. MR2866265

ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915

ROSENBAUM, P. R. (1991). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. Ser. B* **53** 597–610. MR1125717

ROSENBAUM, P. R. (1993). Hodges–Lehmann point estimates of treatment effect in observational studies. *J. Amer. Statist. Assoc.* **88** 1250–1253. MR1245357

ROSENBAUM, P. R. (2001). Replicating effects and biases. *Amer. Statist.* **55** 223–227. MR1963397

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

ROSENBAUM, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics* **63** 456–464. MR2370804

ROSENBAUM, P. R. (2010a). Evidence factors in observational studies. *Biometrika* **97** 333–345. MR2650742

ROSENBAUM, P. R. (2010b). Design sensitivity and efficiency in observational studies. *J. Amer. Statist. Assoc.* **105** 692–702. MR2724853

ROSENBAUM, P. R. (2011). Some approximate evidence factors in observational studies. *J. Amer. Statist. Assoc.* **106** 285–295. MR2816721

ROSENBAUM, P. R. (2013). Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics* **69** 118–127. MR3058058

ROSENBAUM, P. R. (2015a). How to see more in observational studies: Some new quasi-experimental devices. *Ann. Rev. Statist. App.* **2** 21–48.

ROSENBAUM, P. R. (2015b). Two R packages for sensitivity analysis in observational studies. *Observ. Stud.* **1** 1–17.

ROSENBAUM, P. R. (2016a). Using Scheffé projections for multiple outcomes in an observational study of smoking and periodontal disease. *Ann. Appl. Stat.* **10** 1447–1471. MR3553231

ROSENBAUM, P. R. (2016b). The cross-cut statistic and its sensitivity to bias in observational studies with ordered doses of treatment. *Biometrics* **72** 175–183. MR3500586

ROSENBAUM, P. R. (2017). *Observation and Experiment*. Harvard Univ. Press, Cambridge, MA.

ROSENBAUM, P. R. and SILBER, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist. Assoc.* **104** 1398–1405. MR2750570

ROSENBAUM, P. R. and SMALL, D. S. (2017). An adaptive Mantel–Haenszel test for sensitivity analysis in observational studies. *Biometrics* **73** 422–430. DOI:10.1111/biom.12591.

ROTMAN, J. J. (1995). *An Introduction to the Theory of Groups*, 4th ed. *Graduate Texts in Mathematics* **148**. Springer, New York. MR1307623

SHEPHERD, B. E., GILBERT, P. B., JEMIAI, Y. and ROTNITZKY, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics* **62** 332–342. MR2236845

STUART, E. A. and GREEN, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Dev. Psychol.* **44** 395–406.

SUSSER, M. (1973). *Causal Thinking in the Health Sciences*: *Concepts and Strategies in Epidemiology*. Oxford Univ. Press, New York.

SUSSER, M. (1987). Falsification, verification and causal inference in epidemiology: Reconsideration in the light of Sir Karl Popper's philosophy. In *Epidemiology, Health and Society: Selected Papers* (M. Susser, ed.) 82–93. Oxford Univ. Press, New York.

SZÉP, J. (1950). On the structure of groups which can be represented as the product of two subgroups. *Acta Sci. Math.* (*Szeged*) **12** 57–61. MR0037296

TOMAR, S. L. and ASMA, S. (2000). Smoking-attributable periodontitis in the United States: Findings from NHANES III. *J. Periodont.* **71** 743–751.

WERFEL, U., LANGEN, V., EICKHOFF, I., SCHOONBROOD, J., VAHRENHOLZ, C., BRAUKSIEPE, A., POPP, W. and NORPOTH, K. (1998). Elevated DNA strand breakage frequencies in lymphocytes of welders exposed to chromium and nickel. *Carcinogenesis* **19** 413–418.

WOLFE, D. A. (1973). Some general results about uncorrelated statistics. *J. Amer. Statist. Assoc.* **68** 1013–1018. MR0418333

YANG, D., SMALL, D. S., SILBER, J. H. and ROSENBAUM, P. R. (2012). Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics* **68** 628–636. MR2959630

YU, B. B. and 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). Truncated product method for combining *P*-values. *Genet. Epidemiol.* **22** 170–185. DOI:10.1002/gepi.0042.

ZHANG, K., SMALL, D. S., LORCH, S., SRINIVAS, S. and ROSENBAUM, P. R. (2011). Using split samples and evidence factors in an observational study of neonatal outcomes. *J. Amer. Statist. Assoc.* **106** 511–524. MR2847966

ZUBIZARRETA, J. R., NEUMAN, M., SILBER, J. H. and ROSENBAUM, P. R. (2012). Contrasting evidence within and between institutions that provide treatment in an observational study of alternative forms of anesthesia. *J. Amer. Statist. Assoc.* **107** 901–915. MR3010879