A new, powerful approach to the study of effect modification in observational studies

Kwonsang Lee, Dylan S. Small, Paul R. Rosenbaum[1]

University of Pennsylvania, Philadelphia

Abstract. Effect modification occurs when the magnitude or stability of a treatment effect varies as a function of an observed covariate. Generally, larger and more stable treatment effects are insensitive to larger biases from unmeasured covariates, so a causal conclusion may be considerably firmer if effect modification is noted when it occurs. We propose a new strategy, called the submax-method, that splits the population into two subpopulations L times, restoring the population after each split, so that successive splits do not make ever smaller subpopulations. For instance, in split 1, the method might distinguish men and women, while in split 2 it might ignore gender and distinguish smokers and nonsmokers. A test statistic is computed from each subpopulation, and the one statistic for the whole population is added to this list, making $2L+1$ test statistics in total. Typically, $2L+1$ is small compared to the number of interaction subpopulations, namely 2^L, and the $2L$ subpopulations are each large compared to the 2^L interaction subpopulations. These $2L+1$ statistics exhibit a fairly high correlation because the same people reappear in many statistics; for instance, many smokers are men. The method corrects for multiple testing using the joint distribution of the $2L+1$ test statistics, but because of the high correlation among statistics, the correction for multiple testing is small compared to a correction using the Bonferroni inequality. The submax-method achieves the highest design sensitivity and the highest Bahadur efficiency of its $2L+1$ component tests. Moreover, the form of the test is sufficiently tractable that its large sample power may be studied analytically. A simulation confirms and elaborates large sample results.

[1]Kwonsang Lee is a PhD student and Dylan Small and Paul Rosenbaum are professors in the Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340 US. 1 February 2017. dsmall@wharton.upenn.edu.

Additionally, the simulation compares the new method to other methods for which power formulas are not available. Strictly speaking, the submax procedure tests a global null hypothesis, but it may be converted into a consonant multiple inference procedure using closed testing. The simulation suggests that the submax method exhibits superior performance when there is effect modification of moderate size. The mathematical formalism of the submax method has applications besides exploiting effect modification, and these other applications are discussed briefly. Using data from the NHANES I Epidemiologic Follow-Up Survey, an observational study of the effects of physical activity on survival is used to illustrate the method.
Keywords: Causal effects; causal inference; design sensitivity; effect modification; epidemiology; observational study; sensitivity analysis; testing twice.

1 Does physical activity prolong life? Equally for everyone?

1.1 A matched comparison of physical inactivity and survival

Davis et al. (1994) used the NHANES I Epidemiologic follow-up study (NHEFS) to ask: Is greater physical activity reported at the time of the NHANES I survey associated with a longer subsequent life? We examine the same data in a similar way, but with new methodology, specifically the subgroup maximum method or submax-method.

A representative national sample was collected in the first NHANES I survey in 1971-1975, and these sampled individuals were followed up for survival until 1992. Data on all variables other than death were collected at baseline (NHANES I). Physical activity was measured in two variables: self-reported nonrecreational activity ("In your usual days, aside from recreation are you physically very active, moderately active or quite inactive?") and self-reported recreational activity ("Do you get much, moderate or little or no exercise in the things you do for recreation?"). We formed a treated group of 470 adults who were quite inactive, both at work and at leisure, and we matched them to a control group of

470 adults who were quite active (very active in physical activity outside of recreation and much or moderate recreational activity). We compare quite inactive to quite active rather than moderately inactive to moderately active individuals because making the treated and control groups sharply differ in dose increases the insensitivity of the study to unobserved biases when there is a treatment effect and no bias (i.e., it increases the design sensitivity, Rosenbaum, 2004). Following Davis et al. (1994), we excluded people who were quite ill at the time of the NHANES I survey. Both of our groups included people aged between 45 and 74 at baseline in the NHANES I study and excluded people who, prior to the NHANES I evaluation, had had heart failure, a heart attack, stroke, diabetes, polio or paralysis, a malignant tumor, or a fracture of the hip or spine.

Table 1 shows the covariates used in matching. Pairs were exactly matched on sex, smoking status (current smoker) and income (cut at two times the Federal poverty level). Other matched variables were age, race (white or other), years of education, employment (employed or not employed outside the home during the previous three months), marital status, alcohol consumption and dietary quality (number of five nutrients – protein, calcium, iron, Vitamin A and Vitamin C – that were consumed at more than two thirds of the recommended dietary allowance). After matching, the groups are fairly similar, whereas before matching, the inactive group was older, more often female, more often nonwhite, more often poor, more often not working in the prior 3 months, more often not married, and less often had an adequate diet.

The top panel of Figure 1 shows the Kaplan-Meier survival curves for the matched active and inactive groups. We ask two interconnected questions: (i) What magnitude of unmeasured bias from nonrandom treatment assignment would need to be present to explain Figure 1 as something other than an effect caused by inactivity? (ii) Is there greater insensitivity to unmeasured bias in some subgroups because the ostensible effect is larger

in those subgroups, or is there similar evidence of effect in all subgroups? We will study sex, smoking and the two categories of income as potential effect modifiers. These three binary covariates are exactly matched.

1.2 A new approach to effect modification in observational studies

If some subgroups experience larger or more stable effects, then the ostensible effect of a treatment may be less sensitive to bias from nonrandomized treatment assignment in these subgroups; see Hsu et al. (2013). Conversely, if a treatment appears to be highly effective in all subgroups, then it is safer to generalize to other populations that may have different proportions of people in the various subgroups.

One approach to studying effect modification in observational studies constructs a few promising subgroups from several measured covariates using an algorithm such as Breiman et al. (1984)'s CART technique, as discussed by Hsu et al. (2013, 2015), and as described in §3.7. A limitation of this approach is that it is hard to study the power and operating characteristics of such a technique except by simulation, because the CART step does not lend itself to such an evaluation. In the current paper we propose a different approach — the submax method — for which a theoretical evaluation is possible. The submax method has a formula for power and design sensitivity, and additionally permits statements about Bahadur efficiency. In particular, the new method achieves the largest — i.e., best — of the design sensitivities for the subgroups, and the highest Bahadur efficiency of the subgroups; moreover, both the power formula and a simulation confirm that the asymptotic results are a reasonable guide to performance in samples of practical size. The simulation in §3.7 also compares the submax and CART methods. An additional limitation of the CART method is that it is defined for matched pairs. In contrast, the submax-method works for matched pairs, for matched sets with multiple controls, variable numbers of controls and

with the full matching method described by Rosenbaum (1991) and Hansen and Klopfer (2012).

The submax-method considers a single combined analysis together with several ways to split the population into subgroups. It does not form the interaction of subgroups, which would quickly become thinly populated with small sample sizes; rather, it considers one split, reassembles the population, then considers another split. If the splits were defined by L binary covariates, then there would be 2^L interaction subgroups, but the submax-method would do only 1 overall test plus $2L$ subgroup tests, making a total of $2L+1$ highly correlated tests, not 2^L independent tests. If the binary covariates each split every subpopulation in half, then each interaction subgroup would contain a fraction 2^{-L} of the population — i.e., not much — but each of our $2L$ subgroup tests would use half the population — i.e., a much larger fraction. The submax-method uses the joint distribution of the $2L + 1$ test statistics, with the consequence that the correction for multiple testing is quite small due to the high correlation among the test statistics. Specifically, the two halves of one binary split are independent because they refer to different people, but each of those test statistics is highly correlated with test statistics for other splits, because all the splits use the same people. In the example, we split the population by gender (male or female), by current cigarette smoking (yes or no), and by two income groups, so we do $2K + 1 = 2 \times 3 + 1 = 7$ correlated tests. Although the test statistics for men and women are independent, the statistics for men and smokers are highly correlated because there are many male smokers.

2 Notation and review of observational studies

2.1 Treatment effects in randomized experiments

There are G groups, $g = 1, \ldots, G$, of matched sets, $i = 1, \ldots, I_g$, with n_{gi} individuals in set i, $j = 1, \ldots, n_{gi}$, one treated individual with $Z_{gij} = 1$ and $n_{gi} - 1$ controls with $Z_{gij} = 0$, so that $1 = \sum_{j=1}^{n_{gi}} Z_{gij}$ for each g, i. Matched sets were formed by matching for an observed covariate x_{gij}, but may fail to control an unobserved covariate u_{gij}, so that $x_{gij} = x_{gik}$ for each g, i, j, k, but possibly $u_{gij} \neq u_{gik}$. In §1.1, the matched sets are pairs, $n_{gi} = 2$, and there are $G = 2^3 = 8$ groups of pairs defined by combinations of $L = 3$ binary covariates, sex, smoking and income group, with $470 = \sum_{g=1}^{8} I_g$ pairs in total.

Individual gij exhibits response r_{Tgij} if treated or response r_{Cgij} if given the control, so this individual exhibits response $R_{gij} = Z_{gij}\, r_{Tgij} + (1 - Z_{gij})\, r_{Cgij}$, and the effect of the treatment, $r_{Tgij} - r_{Cgij}$, is not observed for anyone; see Neyman (1923) and Rubin (1974). Fisher's (1935) null hypothesis of no treatment effect asserts that $H_0 : r_{Tgij} = r_{Cgij}$ for all i, j. Write $\mathcal{F} = \{(r_{Tgij}, r_{Cgij}, x_{gij}, u_{gij}),\, g = 1, \ldots, G,\, i = 1, \ldots, I_g,\, j = 1, \ldots, n_{gi}\}$. Write $|\mathcal{S}|$ for the number of elements in a finite set $\mathcal{S}$.

Write $\mathcal{Z}$ for the set containing the $|\mathcal{Z}| = \prod_{g=1}^{G} \prod_{i=1}^{I_g} n_{gi}$ possible values $\mathbf{z}$ of the treatment assignment $\mathbf{Z} = \left(Z_{111}, Z_{112}, \ldots, Z_{G,I_G,n_{G,I_G}}\right)^T$, so $\mathbf{z} \in \mathcal{Z}$ if $z_{gij} = 0$ or $z_{gij} = 1$ and $1 = \sum_{j=1}^{n_{gi}} z_{gij}$ for each gi. Conditioning on the event $\mathbf{Z} \in \mathcal{Z}$ is abbreviated as conditioning on $\mathcal{Z}$. In an experiment, randomization picks a $\mathbf{Z}$ at random from $\mathcal{Z}$, so that $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = |\mathcal{Z}|^{-1}$ for each $\mathbf{z} \in \mathcal{Z}$. In a randomized experiment, randomization creates the exact null randomization distribution of familiar test statistics, such as Wilcoxon's signed rank statistic or the mean pair difference or Maritz (1979)'s version of Huber M-statistic. In the analysis of the paired censored survival data in §1.1, the test statistic is the Prentice-Wilcoxon test proposed by O'Brien and Fleming (1987). These

test statistics and many others are of the form $T = \sum_{g=1}^{G} \sum_{i=1}^{I_g} \sum_{j=1}^{n_{gi}} Z_{gij}\, q_{gij}$ for suitable scores q_{gij} that are a function of the R_{gij}, n_{gi} and possibly the x_{gij}, so that, under H_0 in a randomized experiment, the conditional distribution $\Pr(T \mid \mathcal{F}, \mathcal{Z})$ of the test statistic T is the distribution of the sum of fixed scores q_{gij} with $Z_{gij} = 1$ selected at random. In a conventional way, randomization tests are inverted to obtain confidence intervals and point estimates for magnitudes of treatment effects; see, for instance, Lehmann (1975), Maritz (1979) and Rosenbaum (2007).

In large sample approximations, the number of groups, G, will remain fixed, and the number of matched sets I_g in each group will increase without bound.

2.2 Sensitivity to unmeasured biases in observational studies

In an observational study, conventional tests of H_0 appropriate in the randomized experiments in §2.1 can falsely reject a true null hypothesis of no effect because treatments are not assigned at random, $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) \neq |\mathcal{Z}|^{-1}$. A simple model for sensitivity analysis in observational studies assumes that, in the population prior to matching for x, treatment assignments are independent and two individuals, gij and $g'i'j'$, with the same observed covariates, $x_{gij} = x_{g'i'j'}$, may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\Pr(Z_{gij} = 1 \mid \mathcal{F})\ \Pr\left(Z_{g'i'j'} = 0 \,\middle|\, \mathcal{F}\right)}{\Pr\left(Z_{g'i'j'} = 1 \,\middle|\, \mathcal{F}\right)\ \Pr(Z_{gij} = 0 \mid \mathcal{F})} \leq \Gamma \text{ whenever } x_{gij} = x_{g'i'j'}; \tag{1}$$

then the distribution of $\mathbf{Z}$ is returned to $\mathcal{Z}$ by conditioning on $\mathbf{Z} \in \mathcal{Z}$.

Under the model (1), one obtains conventional randomization inferences for $\Gamma = 1$, but these are replaced by an interval of P-values or an interval of point estimates or an interval of endpoints for a confidence interval for $\Gamma > 1$. The intervals become longer as Γ increases, the interval of P-values tending to $[0, 1]$ as $\Gamma \to \infty$, reflecting the familiar fact

that association, no matter how strong, does not logically entail causation. At some point, the interval is sufficiently long to be uninformative, for instance including P-values that would both reject and accept the null hypothesis of no effect. The question answered by a sensitivity analysis is: How much bias in treatment assignment, measured by Γ, would need to be present before the study becomes uninformative? For instance, how large would Γ have to be to produce a P-value above α, conventionally $\alpha = 0.05$?

An approximation to the upper bound on the P-value is obtained in the following way; see Gastwirth, Krieger and Rosenbaum (2000) for detailed discussion and see Rosenbaum (2007, §4; 2014) for its application to Huber-Maritz M-tests. Assume H_0 is true for the purpose of testing it, so that $R_{gij} = r_{Cgij}$ and q_{gij} are fixed by conditioning on $\mathcal{F}$. Write $T_g = \sum_{i=1}^{I_g} \sum_{j=1}^{n_{gi}} Z_{gij}\, q_{gij}$, so that $T = \sum_{g=1}^{G} T_g$. Subject to (1) for a given $\Gamma \geq 1$, find the maximum expectation, $\mu_{\Gamma g}$, of T_g. Also, among all treatment assignment probabilities that satisfy (1) and that achieve the maximum expectation $\mu_{\Gamma g}$, find the maximum variance, $\nu_{\Gamma g}$, of T_g. If $T \geq \sum_{g=1}^{G} \mu_{\Gamma g}$, report as the upper bound on the P-value for T,

$$1 - \Phi\left\{\left(\sum_{g=1}^{G} T_g - \mu_{\Gamma g}\right) \Big/ \sqrt{\sum_{g=1}^{G} \nu_{\Gamma g}}\right\}, \tag{2}$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution. The bound is derived as $\min(I_g) \rightarrow \infty$ with some mild conditions to ensure that no one q_{gij} dominates the rest, and that the fixed scores q_{gij} do not become degenerate as $\min(I_g)$ increases. For $\Gamma = 1$, this yields a Normal approximation to a randomization P-value using T as the test statistic. If treatment assignments were governed by the probabilities satisfying (1) that yield $\mu_{\Gamma g}$ and $\nu_{\Gamma g}$, then, under H_0 and mild conditions on the q_{gij}, the joint distribution of

$$\left\{(T_1 - \mu_{\Gamma 1}) / \sqrt{\nu_{\Gamma 1}}, \ldots, (T_G - \mu_{\Gamma G}) / \sqrt{\nu_{\Gamma G}}\right\}^T$$

would converge to a G-dimensional Normal distribution with expectation vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$ as $\min(I_g) \rightarrow \infty$. Simpler methods of proof and formulas apply in simple cases, such as matched pairs; for instance, contrast §3 and §4 of Rosenbaum (2007). These simpler methods of proof bound the distribution of T exactly, then approximate the bounding distribution, whereas the general method is merely a large sample approximation to the upper bound on the P-value when $T \geq \sum_{g=1}^{G} \mu_{\Gamma g}$. Write $\boldsymbol{\mu}_\Gamma = (\mu_{\Gamma 1}, \ldots \mu_{\Gamma G})^T$ and $\mathbf{V}_\Gamma$ for the $G \times G$ diagonal matrix with gth diagonal element $\nu_{\Gamma g}$.

For various methods of sensitivity analysis in observational studies, see Egleston et al. (2009), Gilbert et al. (2003), Hosman et al. (2010), and Liu et al. (2013).

2.3 Design sensitivity and Bahadur efficiency

Suppose that there is a treatment effect and there is no bias from the unobserved covariate u_{gij}, and call this the favorable situation. In an observational study, if an investigator were in the favorable situation, then she would not know it, and the best she could hope to say is that the results are insensitive to small and moderate biases Γ. The power of a sensitivity analysis is the probability that she will be able to say this. More precisely, in the favorable situation, the power of a level α sensitivity analysis at sensitivity parameter Γ is the probability that (2) will be less than or equal to α when computed at the given Γ.

As the sample size increases, there is a value, $\widetilde{\Gamma}$, called the design sensitivity, such that the power tends to 1 if $\Gamma < \widetilde{\Gamma}$ and the power tends to zero if $\Gamma > \widetilde{\Gamma}$, so $\widetilde{\Gamma}$ is the limiting sensitivity to unmeasured bias for a given favorable situation and test statistic; see Rosenbaum (2004; 2010, Part III), Zubizarreta et al. (2013) and Stuart and Hanna (2013). In a particular favorable situation, for a specific Γ, the rate at which (2) declines to zero with increasing sample size yields the Bahadur efficiency of the sensitivity analysis,

and the efficiency drops to zero at $\Gamma = \widetilde{\Gamma}$; see Rosenbaum (2015).

3 Joint bounds for two or more comparisons

3.1 Subgroup comparisons

We are interested in K specified comparisons, $k = 1, \ldots, K$, among the G groups of matched sets. By one comparison we mean a fixed nonzero vector $\mathbf{c}_k = (c_{1k}, \ldots, c_{Gk})^T$ of dimension G with $c_{gk} \geq 0$ for $g = 1, \ldots, G$, and we evaluate a comparison using the statistic $S_k = \sum_{g=1}^{G} c_{gk}\, T_g$. For instance, the comparison $\mathbf{c}_1 = (1, \ldots, 1)^T$ yields the overall test in §2.2. By replacing the scores q_{gij} in §2.2 by scores $q^*_{gij} = c_{gk}\, q_{gij}$, the bound for S_k is obtained in parallel with (2). If groups $1, \ldots, G/2$ are matched sets of men and groups $G/2+1, \ldots, G$ are matched sets of women, then the comparison $\mathbf{c}_2 = (1, \ldots, 1, 0, \ldots, 0)^T$ confines attention to men, while the comparison $\mathbf{c}_3 = (0, \ldots, 0, 1, \ldots, 1)^T$ confines attention to women. Perhaps an additional comparison $\mathbf{c}_4 = (1, \ldots, 1, 0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)^T$ would confine attention to people over the age of 65, and so on.

If the treatment effect for women were larger than the effect for men, the comparison, $\mathbf{c}_3$, restricted to women might be insensitive to larger unmeasured biases than the overall comparison, $\mathbf{c}_1$. Hsu et al. (2013) present an example in which a treatment to prevent malaria is far more effective for children than for adults, so that only very large biases in treatment assignment could explain away the ostensible benefits for children.

3.2 Joint evaluation of subgroup comparisons

Let $\mathbf{C}$ be the $K \times G$ matrix whose K rows are the $\mathbf{c}_k^T = (c_{1k}, \ldots, c_{Gk})$, $k = 1, \ldots, K$. Define $\boldsymbol{\theta}_\Gamma = \mathbf{C}\boldsymbol{\mu}_\Gamma$ and $\boldsymbol{\Sigma}_\Gamma = \mathbf{C}\mathbf{V}_\Gamma\mathbf{C}^T$, noting that $\boldsymbol{\Sigma}_\Gamma$ is not typically diagonal. Write $\theta_{\Gamma k}$ for the kth coordinate of $\boldsymbol{\theta}_\Gamma$ and $\sigma^2_{\Gamma k}$ for the kth diagonal element of $\boldsymbol{\Sigma}_\Gamma$. Define $D_{\Gamma k} = (S_k - \theta_{\Gamma k})/\sigma_{\Gamma k}$ and $\mathbf{D}_\Gamma = (D_{\Gamma 1}, \ldots, D_{\Gamma K})^T$. Finally, write $\boldsymbol{\rho}_\Gamma$ for the $K \times K$

correlation matrix formed by dividing the element of $\mathbf{\Sigma}_\Gamma$ in row k and column k' by $\sigma_{\Gamma k}\,\sigma_{\Gamma k'}$. Subject to (1) under H_0, at the treatment assignment probabilities that yield the $\mu_{\Gamma g}$ and $\nu_{\Gamma g}$, the distribution of $\mathbf{D}_\Gamma$ is converging to a Normal distribution, $N_K\left(\mathbf{0}, \boldsymbol{\rho}_\Gamma\right)$, with expectation $\mathbf{0}$ and covariance matrix $\boldsymbol{\rho}_\Gamma$ as $\min\left(I_g\right) \to \infty$. Using this null distribution, the null hypothesis H_0 is tested using

$$D_{\Gamma\,\max} = \max_{1 \le k \le K} D_{\Gamma k} = \max_{1 \le k \le K} \frac{S_k - \theta_{\Gamma k}}{\sigma_{\Gamma k}}.$$

The α critical value $\kappa_{\Gamma,\alpha}$ for $D_{\Gamma\,\max}$ solves

$$1 - \alpha = \Pr\left(D_{\Gamma\,\max} < \kappa_{\Gamma,\alpha}\right) = \Pr\left(\frac{S_k - \theta_{\Gamma k}}{\sigma_{\Gamma k}} < \kappa_{\Gamma,\alpha},\, k = 1, \ldots, K\right) \tag{3}$$

under H_0. The multivariate Normal approximation to $\kappa_{\Gamma,\alpha}$ is obtained using the `qmvnorm` function in the `mvtnorm` package in R, as applied to the $N_K\left(\mathbf{0}, \boldsymbol{\rho}_\Gamma\right)$ distribution; see Genz and Bretz (2009). Notice that this approximation to $\kappa_{\Gamma,\alpha}$ depends upon Γ only through $\boldsymbol{\rho}_\Gamma$, which in turn depends upon Γ only through $\nu_{\Gamma g}$. The resulting approximate α critical value $\kappa_{\Gamma,\alpha}$ for $D_{\Gamma\,\max}$ is larger than $\Phi^{-1}\left(1 - \alpha\right)$ because the largest of K statistics $D_{\Gamma k}$ has been selected, and it reflects the correlations $\boldsymbol{\rho}_\Gamma$ among the coordinates of $\mathbf{D}_\Gamma$.

3.3 Behavior of the critical constant $\kappa_{\Gamma,\alpha}$ in a simple case

Consider a simple, balanced case under the null hypothesis H_0, in which every matched set is a matched pair, $n_{gi} = 2$ for all g, i, and outcomes are continuously distributed and hence untied with probability one. Additionally, there are L matched binary covariates, such as gender, to be examined as potential effect modifiers making $G = 2^L$ groups of pairs, with the same number of matched pairs in each group, $I_1 = \cdots = I_G = \overline{I}$, say. Suppose that, in each group, T_g is Wilcoxon's signed rank statistic computed from the $\overline{I}$ pairs in that group.

In this case, $\mu_{\Gamma g} = \{\Gamma/(1+\Gamma)\}\,\overline{I}\,(\overline{I}+1)/2$ and $\nu_{\Gamma g} = \left\{\Gamma/(1+\Gamma)^2\right\}\overline{I}\,(\overline{I}+1)\,(2\overline{I}+1)/6$; see Rosenbaum (2002, §4.3.3). In this simple case, by symmetry, the correlation matrix $\boldsymbol{\rho}_\Gamma$ does not depend upon Γ. There are $K = 2L+1$ comparisons, namely $\mathbf{c}_1 = (1,\ldots,1)^T$ in §3.1 using all of the pairs, yielding T as in §2.2, plus two comparisons for each binary covariate for half the pairs at the high and low levels of that covariate, for instance, $\mathbf{c}_2$, $\mathbf{c}_3$ and $\mathbf{c}_4$ in §3.1, making a total of $K = 2L+1$ tests. Because of the symmetry of this situation, the correlation/covariance matrix $\boldsymbol{\rho}_\Gamma$ of $D_{\Gamma k}$ has the simple form in Table 2; that is, $D_{\Gamma 1}$ has correlation $0.707 = 1/\sqrt{2}$ with $D_{\Gamma k}$ for $k \geq 2$, the two consecutive comparisons for the two categories of the same binary variable are uncorrelated, and all other comparisons have correlation 0.5.

In this simple, balanced case, Table 3 shows the critical constant $\kappa_{\Gamma,\alpha}$ for $\alpha = 0.05$ and $L = 0, 1, \ldots, 15$ potential effect modifiers, and $K = 2L+1 = 1, 3, \ldots, 31$ tests. For comparison in Table 3, $\kappa_{\Gamma,\alpha}$ is compared to $\Phi^{-1}(1-\alpha/K)$, the critical constant obtained from the Bonferroni inequality. For instance, the Bonferroni critical constant $\Phi^{-1}(1-\alpha/K)$ for $K = 15$ tests and $L = 7$ is 2.71, which is larger than the submax critical constant of 2.70 for $K = 25$ tests and $L = 12$.

3.4 Application in the NHANES example

Table 4 performs the test in §3.2 for the NHANES data in §1.1 using the Prentice-Wilcoxon statistic T of O'Brien and Fleming (1987). The row of Table 4 for $\Gamma = 1$ consists of Normal approximations to randomization tests, while the rows with $\Gamma > 1$ examine sensitivity to bias from nonrandom treatment assignment. For $\Gamma = 1$, the test statistic $D_{\Gamma\,\max} = 6.09 \geq \kappa_{\Gamma,\alpha} = 2.31$, so Fisher's hypothesis of no treatment effect would be rejected at level α if the data had come from a randomized experiment with $\Gamma = 1$. For $\Gamma = 1$, the maximum statistic is based on all 470 pairs, $D_{\Gamma\,\max} = D_{\Gamma 1}$; however, $D_{\Gamma k} \geq \kappa_{\Gamma,\alpha} = 2.31$ for every

subgroup, $k = 1, \ldots, K = 7$. At $\Gamma = 1.4$, the deviates $D_{\Gamma 2}$ and $D_{\Gamma 6}$ for females ($k = 2$) and the nonpoor ($k = 6$) no longer exceed $\kappa_{\Gamma,\alpha} = 2.31$, and the precise meaning of this is examined in more detail in §4. At $\Gamma = 1.64$, Fisher's hypothesis of no treatment effect is still rejected because the deviate $D_{\Gamma 3}$ for males exceeds $\kappa_{\Gamma,\alpha} = 2.31$. Although there are 275 pairs of women and 195 pairs of men, the strongest evidence, the least sensitive evidence, of an effect of inactivity on survival is for men. The bottom two panels of Figure 1 show the separate survival curves for men and women.

Table 4 is compactly and conveniently indexed by one parameter Γ. It is sometimes helpful to give a two-parameter interpretation of this one parameter. In particular, the longer life of active men in Table 4 is insensitive to an unmeasured bias of $\Gamma = 1.64$. In a matched pair, $\Gamma = 1.64$ corresponds with an unobserved covariate that doubles the odds of a longer life and increases the chance of inactivity by a factor of more than 6-fold; see the amplification of Γ into two equivalent parameters Δ and Λ in Rosenbaum and Silber (2009a), where $1.64 = \Gamma = (\Delta\Lambda + 1)\,/\,(\Delta\Lambda)$ for $\Delta = 2$ and $\Lambda = 6.33$.

In §3.8, an alternative analysis of the NHANES data is presented using Breiman et al. (1984)'s CART regression, as proposed by Hsu et al. (2013, 2015). The CART technique is described in §3.7 where a simulation compares it to the submax method.

3.5 Design sensitivity and Bahadur efficiency

As in Rosenbaum (2012), it is easy to see that under an alternative hypothesis given by a favorable situation — a treatment effect with no unmeasured bias — the design sensitivity of $D_{\Gamma\,\max}$, say $\widetilde{\Gamma}_{\max}$, is equal to the maximum design sensitivity $\widetilde{\Gamma}_k$ of the K component tests, $\widetilde{\Gamma}_{\max} = \max\left(\widetilde{\Gamma}_1, \ldots, \widetilde{\Gamma}_K\right)$. Briefly, by the definition of design sensitivity, if $\Gamma < \widetilde{\Gamma}_k$, then the probability that $D_{\Gamma k} \geq \kappa$ tends to 1 for every κ as $\min(I_g) \rightarrow \infty$, so the probability that $D_{\Gamma\,\max} \geq \kappa_{\Gamma,\alpha}$ tends to 1 because $D_{\Gamma\,\max} \geq D_{\Gamma k}$. Although there is a

price to be paid for multiple testing, that price does not affect the design sensitivity.

Define $\beta_1 = 1$. Berk and Jones (1979) show that, if $D_{\Gamma k}$ has Bahadur efficiency β_k relative to $D_{\Gamma 1}$ for $k = 2, \ldots, K$ under some alternative hypothesis, then $D_{\Gamma\,\max}$ has Bahadur efficiency $\beta_{\max} = \max_{1 \leq k \leq K} \beta_k$. Berk and Jones call this "relative optimality" meaning $D_{\Gamma\,\max}$ is optimal among the fixed set $D_{\Gamma 1}, \ldots, D_{\Gamma K}$. In other words, the correction for multiplicity, $\kappa_{\Gamma,\alpha} > \Phi^{-1}(1-\alpha)$, does reduce finite sample power, but in a limited way, so that the Bahadur efficiency is ultimately unaffected.

3.6 Power calculations and design sensitivity in a simple case

Under an alternative hypothesis, if the T_g are independent and asymptotically Normal with expectation μ_g^* and variance ν_g^*, then straightforward manipulations involving the multivariate Normal distribution yield an asymptotic approximation to the power of tests based on $D_{\Gamma\,\max}$ or $D_{\Gamma k}$ for fixed k.

Specifically, write $\theta_k^* = \sum_{g=1}^{G} c_{gk}\,\mu_g^*$ and σ_k^* for the square root of the kth diagonal element of the $K \times K$ covariance matrix $\mathbf{C}\mathrm{diag}\,(\nu_1^*, \ldots, \nu_K^*)\,\mathbf{C}^T$, so θ_k^* is the expectation and σ_k^* is the standard deviation of S_k under the alternative; moreover, write $\boldsymbol{\rho}^*$ for the $K \times K$ correlation matrix computed from this covariance matrix. The approximate power is the following probability computed under the alternative hypothesis,

$$1 - \Pr\left(D_{\Gamma\,\max} < \kappa_{\Gamma,\alpha}\right) = 1 - \Pr\left(\frac{S_k - \theta_{\Gamma k}}{\sigma_{\Gamma k}} < \kappa_{\Gamma,\alpha},\, k = 1, \ldots, K\right)$$

$$= 1 - \Pr\left(\frac{S_k - \theta_k^*}{\sigma_k^*} < \frac{\theta_{\Gamma k} - \theta_k^* + \kappa_{\Gamma,\alpha}\,\sigma_{\Gamma k}}{\sigma_k^*},\, k = 1, \ldots, K\right). \tag{4}$$

The Normal approximation to the joint distribution of the T_g under the alternative means that the last term in (4) is approximately a particular quadrant probability for the $N_K\left(\mathbf{0}, \boldsymbol{\rho}^*\right)$

distribution, and this may be calculated using the `pmvnorm` function in the `mvtnorm` package in `R`. Under the same assumptions, the power of a test based on one fixed $D_{\Gamma k}$ is approximately

$$1 - \Pr\left\{ \frac{S_k - \theta_k^*}{\sigma_k^*} < \frac{\theta_{\Gamma k} - \theta_k^* + \Phi^{-1}(1-\alpha)\ \sigma_{\Gamma k}}{\sigma_k^*} \right\}, \tag{5}$$

and this may be calculated using the standard Normal distribution.

Moreover, the design sensitivity $\widetilde{\Gamma}_k$ for $S_k = \sum_{g=1}^{G} c_{gk} T_g$ is the limit of values of Γ that solve $1 = \left(\sum_{g=1}^{G} c_{gk}\, \mu_g^*\right) / \left(\sum_{g=1}^{G} c_{gk}\, \mu_{\Gamma g}\right)$. That is, using S_k, as $I \rightarrow \infty$, the power tends to 1 for $\Gamma < \widetilde{\Gamma}_k$ and it tends to 0 for $\Gamma > \widetilde{\Gamma}_k$. This formula emphasizes the importance of effect modification. For instance, with two groups, $G = 2$, say $g = 0$ and $g = 1$, if $\mu_0^* > \mu_1^*$, then the design sensitivity is largest with $c_{0k} = 1$ and $c_{1k} = 0$, so as $I \rightarrow \infty$, there are values of Γ such that the power of the overall test is tending to 0 while the power of a test focused on the first subgroup is tending to 1. This will be quite visible in both theoretical and simulated power calculations.

An oracle would use the one $D_{\Gamma k}$ with the highest power. Lacking such an oracle, it is interesting to compare $D_{\Gamma\,\max}$ to: (i) the oracle, (ii) the one test, $D_{\Gamma 1}$, that uses all of the matched sets, as in §2.2.

To illustrate, consider the simple, balanced case in §3.3, and suppose that there are L binary covariates as potential effect modifiers. We would like to compute power under a favorable alternative, meaning that, unknown to the investigator, the treatment has an effect and there is no unmeasured bias from u_{gij}. Because the investigator cannot know that the data came from the favorable situation, a sensitivity analysis is performed. A simple favorable situation has $I_g = \overline{I}$ independent treated-minus-control pair differences in every group g, where the pair differences are Normal with various expectations and variance 1. Then Wilcoxon's signed rank statistic in group g, namely T_g, is asymptotically Normal under the alternative hypothesis as $\overline{I} \rightarrow \infty$, and simple formulas in Lehmann (1975, §4.2)

give the expectation and variance, μ_g^* and ν_g^*, of T_g, under this alternative. There are $G\overline{I} = 2^L \cdot \overline{I}$ pairs in total. Note that the $K = 2L + 1$ statistics, S_k, are each computed from at least $2^{L-1} \cdot \overline{I}$ pairs, not from $\overline{I}$ pairs, and they are each sums of at least 2^{L-1} signed rank statistics T_g.

Table 5 displays theoretical power for a level $\alpha = 0.05$ test of no effect in several favorable situations, that is, situations with a treatment effect and no bias. In Table 5, "one covariate" refers to $L = 1$ binary covariate, making $G = 2^L = 2$ groups, so that $D_{\Gamma\,\max}$ is the maximum of three statistics, namely the deviates for the signed rank statistics in groups 1 and 2 and for the sum of these two statistics. In Table 5, "five covariates" refers to $L = 5$ binary covariates, making $G = 2^L = 32$ groups, so that $D_{\Gamma\,\max}$ is the maximum of $11 = 2 \times 5 + 1$ statistics, namely the deviates for 10 totals of 16 signed rank statistics at the high and low levels of each covariate, and also for the sum of all 32 signed rank statistics.

The sample size in Table 5 is constant in each group, $I_g = \overline{I}$, with total sample size $2016 = G\overline{I} = 2^L \cdot \overline{I}$, so this is $\overline{I} = 1008$ for $L = 1$ covariate and $\overline{I} = 63$ for $L = 5$ covariates. In both cases, $L = 1$ and $L = 5$, only the first covariate is a potential effect modifier, in the sense that the expected pair difference only changes with the level of the first covariate, being ζ_0 for the 0 level and ζ_1 for the 1 level. When $\zeta_0 \neq \zeta_1$, there is effect modification. With $L = 5$, four of the five covariates are simply a distraction that require $D_{\Gamma\,\max}$ to make a larger correction for multiple testing. The first situation in Table 5 has no treatment effect, $\zeta_0 = \zeta_1 = 0$, so the reported values are the actual size of a level $\alpha = 0.05$ test. The second situation in Table 5 has a constant treatment effect, $\zeta_0 = \zeta_1 = 0.5$, so it is a mistake to look for effect modification because there is none. The third situation in Table 5 has slight effect modification, $\zeta_0 = 0.6 > 0.4 = \zeta_1$, although the average treatment effect is $0.5 = (\zeta_0 + \zeta_1)/2$ as in the second situation. The fourth situation in Table 5 has substantial effect modification, $\zeta_0 = 0.5 > 0 = \zeta_1$, so the average treatment effect is

$0.25 = (\zeta_0 + \zeta_1)/2$. The design sensitivity $\widetilde{\Gamma}_g$ for Wilcoxon's statistic T_g in group g is 3.17 if $\zeta_g = 1/2$ and it is 1 if $\zeta_g = 0$; see Rosenbaum (2010, p. 272) for details of this calculation. For instance, in Table 5 with $\zeta_0 = \zeta_1 = 0.5$, the power of the test is below the level $\alpha = 0.05$ when $\Gamma > \widetilde{\Gamma}_g = 3.17$.

Table 5 compares the power of $D_{\Gamma\,\max}$ to a single combined test $D_{\Gamma 1}$ that uses all pairs and an oracle that performs a single test using all the pairs that have the largest value of ζ_g. Obviously, the oracle is not a statistical procedure because it requires the statistician to know what she does not know, namely which groups have the largest ζ_g. From theory, in the nonnull situations 2, 3 and 4, we know that $D_{\Gamma\,\max}$ has the same design sensitivity as the oracle, whereas the $D_{\Gamma 1}$ has lower design sensitivity than the oracle unless there is no effect modification, $\zeta_0 = \zeta_1$, as in situation 2. In situation 2, all three procedures have design sensitivity $\widetilde{\Gamma} = 3.17$, with negligible power for $\Gamma = 3.2 > 3.17$. In situation 3, $\zeta_0 = 0.6$, and both $D_{\Gamma\,\max}$ and the oracle have design sensitivity $\widetilde{\Gamma} = 4.05$ by focusing on group 0 for covariate 1, and they have nonnegligible power at $\Gamma = 3.4 < 4.05$; however, $D_{\Gamma 1}$ has design sensitivity $\widetilde{\Gamma} = 3.13$ in situation 3, with negligible power at $\Gamma = 3.2$. In situation 4, $\zeta_1 = 0$, and both $D_{\Gamma\,\max}$ and the oracle have design sensitivity $\widetilde{\Gamma} = 3.17$ by focusing on group 0 for covariate 1; however, $D_{\Gamma 1}$ has design sensitivity $\widetilde{\Gamma} = 1.70$ in situation 3, with negligible power at $\Gamma = 2.8$.

In the first situation in Table 5, all tests have the correct size for $\Gamma = 1$, and because there is no actual bias in the favorable situation, they have size below 0.05 for $\Gamma > 1$. In the second situation in Table 5, $D_{\Gamma\,\max}$ pays a price in power in its search for effect modification that is not there. In situations 3 and 4, $D_{\Gamma\,\max}$ has much higher power than the $D_{\Gamma 1}$ statistic, but it is behind the oracle, reflecting the price paid to discover the true pattern of effect modification. For instance, at $\Gamma = 2.8$, with $L = 5$ binary covariates and slight effect modification, $\zeta_0 = 0.6 > 0.4 = \zeta_1$, the statistic $D_{\Gamma\,\max}$ has power .959, the

oracle has power 0.996, and $D_{\Gamma 1}$ has power 0.521.

3.7 Simulated power and a comparison with CART groups

Table 6 describes simulated power for the same situation as the theoretical power in Table 5. Unlike Table 5, the simulation includes the power for a competing method for matched pairs proposed by Hsu et al. (2015), in which groups are built from covariates using the CART procedure of Breiman et al. (1984). There is no known power formula for the CART method, so it cannot be included in Table 5. In this approach, the pairs are initially ungrouped, and so lack a g subscript. However, the pairs have been exactly matched for several covariates that may be effect modifiers. The absolute treated-minus-control pair difference in outcomes in pair i, namely $|Y_i| = |R_{i1} - R_{i2}|$, is regressed on these covariates using CART, and the leaves of the tree define the groups. The P-values with the groups so-defined are combined using the truncated product of P-values proposed by Zaykin et al. (2002). The truncated product is analogous to Fisher's product of P-values, except P-values above a prespecified truncation point, ς, enter the product as 1, so the two methods are the same for $\varsigma = 1$. In Table 5, $\varsigma = 1/10$. Unlike $D_{\Gamma\,\max}$, there is no guarantee that the CART procedure will equal the oracle in terms of design sensitivity. In other words, we expect $D_{\Gamma\,\max}$ to win in sufficiently large samples, tracking the oracle as $\min(I_g) \rightarrow \infty$; however, $D_{\Gamma\,\max}$ may not win in the finite samples in Table 6.

Table 6 provides a check on the theoretical power formulas that yielded Table 5, and in general the two tables are in agreement. The CART procedure has higher power than $D_{\Gamma\,\max}$ when there is no effect modification in situation 2, $\zeta_0 = \zeta_1 = 0.5$, because the CART procedure typically produces a single group in this situation. The CART procedure has lower power than $D_{\Gamma\,\max}$ when there is slight effect modification in situation 3, $\zeta_0 = .6 > .4 = \zeta_1$, perhaps because the CART procedure fails to locate the slight effect

modification. In situation 4, with $\zeta_0 = .5 > 0 = \zeta_1$, the move from $L = 1$ covariate to $L = 5$ covariates reduces the power of both $D_{\Gamma\max}$ and the CART procedure, but it does more harm to $D_{\Gamma\max}$. Presumably, $D_{\Gamma\max}$ pays a higher price for multiple testing with $L = 5$ than with $L = 1$ consistent with Table 3, while the CART procedure has more difficulty finding the right groups with $L = 5$ than with $L = 1$.

There is no uniform winner in Table 6. However, when compared to the CART method, we expect $D_{\Gamma\max}$ to gradually catch up, or to move ahead, or to stay ahead as $\min(I_g) \to \infty$ because it has the best design sensitivity; therefore, relative performance depends upon the sample size.

3.8 Use of CART in the example

As an alternative to the analysis in §3.4, consider using the CART method in §3.7, implemented using the `rpart` package in `R`. In an `rpart` tree, the number of splits is controlled by a complexity parameter that defaults to the value 0.01. Using the default settings in `rpart`, the CART tree is a single group of all 470 pairs. At $\Gamma = 1.64$, the single group test has deviate $D_{\Gamma 1} = 2.29$ and one-sided P-value bound of $1 - \Phi(2.29) = 0.011$. If the complexity parameter in `rpart` is reduced below 0.0062, then the CART tree splits on sex. Hsu et al. suggest combining the P-value bounds from the leaves of the tree using Zaykin et al. (2002)'s truncated product of P-values, an extension of Fisher's method of combining P-values. At $\Gamma = 1.64$, if the two P-value bounds for females and males, $1 - \Phi(0.97) = 0.166$ and $1 - \Phi(2.32) = 0.010$, are combined using the truncated product with truncation 0.1, then the combined P-value bound is 0.028. In this one example, the two analyses give fairly similar impressions.

4 Simultaneous inference and closed testing

Strictly speaking, the statistic $D_{\Gamma\,\max}$ is a test of a global null hypothesis, specifically Fisher's hypothesis H_0 of no treatment effect in the study as a whole. In previous sections, the c_{gk} are either 0 or 1, and the kth comparison defines a subpopulation $\mathcal{S}_k$ as those groups with $c_{gk} = 1$, that is, $\mathcal{S}_k = \{g : c_{gk} = 1\}$, for instance, the subpopulation of men. We are, of course, interested in the hypothesis, say H_k, that asserts there is no effect in subpopulation $\mathcal{S}_k$, say no effect in the subpopulation of men. We would like to test all K hypotheses H_k, $k = 1, \ldots, K$, strongly controlling the family-wise error rate at α in the presence of a bias of at most Γ. We may do this with the closed testing method of Marcus et al. (1976).

Define $H_{\mathcal{I}}$ for $\mathcal{I} \subseteq \{1, \ldots, K\}$ to be the hypothesis that there is no treatment effect in the union of the subpopulations $\mathcal{S}_k$, $k \in \mathcal{I}$. For instance, in Table 4, the hypothesis $H_{\{2,5\}}$ says that there is no effect for females, $k = 2$, and no effect for smokers, $k = 5$. If $H_{\{2,5\}}$ were true, there might nonetheless be an effect for male nonsmokers. If the goal were to test $H_{\mathcal{I}}$ alone at level α in the presence of a bias of at most Γ, then this could be done using $D_{\Gamma\mathcal{I}} = \max_{k\in\mathcal{I}} D_{\Gamma k}$, which is a test of the same form as $D_{\Gamma\,\max}$, whose approximate critical constant from (3), say $\kappa_{\Gamma,\alpha,\mathcal{I}}$, must be recalculated using a $|\mathcal{I}|$-dimensional multivariate Normal distribution. Of course, $D_{\Gamma\mathcal{I}} \geq D_{\Gamma\mathcal{J}}$ whenever $\mathcal{J} \subset \mathcal{I}$, so $\kappa_{\Gamma,\alpha,\mathcal{J}} \leq \kappa_{\Gamma,\alpha,\mathcal{I}}$; that is, the correction for multiple testing is less severe when fewer comparisons are made. In particular, $\kappa_{\Gamma,\alpha,\mathcal{I}} \leq \kappa_{\Gamma,\alpha}$ for all $\mathcal{I} \subseteq \{1, \ldots, K\}$.

The closed testing method of Marcus et al. (1976) rejects $H_{\mathcal{I}}$ at level α in the presence of a bias of at most Γ if it rejects $H_{\mathcal{K}}$ for all $\mathcal{K} \supseteq \mathcal{I}$, that is, if $D_{\Gamma\mathcal{K}} \geq \kappa_{\Gamma,\alpha,\mathcal{K}}$ for all hypotheses $\mathcal{K}$ that contain $\mathcal{I}$. Closed testing has several attractive properties. In general, closed testing strongly controls the family-wise error rate, as demonstrated by Marcus et al. (1976). The extension of this property to sensitivity analyses is straightforward; see Rosenbaum and Silber (2009b, §4.4). That is, no matter which hypotheses are true or

false, the probability that closed testing falsely rejects at least one true $H_{\mathcal{I}}$ is at most α whenever the bias is at most Γ. There is an additional property of closed testing that is specific to sensitivity analyses. Use of the Bonferroni inequality in sensitivity analysis is conservative in a way that closed testing is not conservative; see Rosenbaum and Silber (2009b, §4.4-§4.5) and Fogarty and Small (2016).

There is a short-cut that simplifies closed testing in this context using the inequality $\kappa_{\Gamma,\alpha,\mathcal{I}} \leq \kappa_{\Gamma,\alpha}$ for all $\mathcal{I} \subseteq \{1,\ldots,K\}$, noted above. Specifically, $D_{\Gamma\mathcal{K}} = \max_{k\in\mathcal{K}} D_{\Gamma k} \geq D_{\Gamma k}$ for all $k \in \mathcal{K}$ and yet $\kappa_{\Gamma,\alpha} \geq \kappa_{\Gamma,\alpha,\mathcal{K}}$, so whenever $D_{\Gamma k} \geq \kappa_{\Gamma,\alpha}$ it follows that $D_{\Gamma\mathcal{K}} \geq \kappa_{\Gamma,\alpha,\mathcal{K}}$ for all hypotheses $\mathcal{K}$ with $k \in \mathcal{K}$. This means that closed testing will reject H_k whenever $D_{\Gamma k} \geq \kappa_{\Gamma,\alpha}$, and may reject H_k in other cases as well. For instance, in Table 4, at $\Gamma = 1.5$, we may reject H_3 and H_7 without calculating $\kappa_{\Gamma,\alpha,\mathcal{K}}$ because $2.77 = D_{\Gamma 3} \geq \kappa_{\Gamma,\alpha} = 2.31$ and $2.45 = D_{\Gamma 7} \geq \kappa_{\Gamma,\alpha} = 2.31$. That is, at $\Gamma = 1.5$, closed testing rejects the null hypothesis of no effect on men and the hypothesis of no effect on the poor.

Consider $\Gamma = 1.4$ in Table 4. The short-cut reject in all groups except females ($k = 2$) and nonpoor ($k = 6$), so that, without further computation, $D_{\Gamma\mathcal{K}} \geq \kappa_{\Gamma,\alpha,\mathcal{K}}$ for every nonempty $\mathcal{K}$ except $\{2,6\}$, $\{2\}$, and $\{6\}$. The short-cut does not apply in these cases, so $\kappa_{\Gamma,\alpha,\mathcal{K}}$ must be computed. Using the 2×2 submatrix of $\boldsymbol{\rho}_\Gamma$ for ($D_{\Gamma 2}$, $D_{\Gamma 6}$), we determine $\kappa_{\Gamma,\alpha,\{2,6\}} = 1.92$, and trivially for $\mathcal{K} = \{2\}$ and $\mathcal{K} = \{6\}$ the critical constant is $\kappa_{\Gamma,\alpha,\mathcal{K}} = 1.64$. Because the short-cut has rejected every $H_{\mathcal{I}}$ with $\{2,6\} \subset \mathcal{I}$, we compare $D_{\Gamma\{2,6\}} = 2.07$ to $\kappa_{\Gamma,\alpha,\{2,6\}} = 1.92$ and therefore reject $H_{\{2,6\}}$. Continuing, we compare $D_{\Gamma 2} = 1.86$ and $D_{\Gamma 6} = 2.07$ to $\kappa_{\Gamma,\alpha,\mathcal{K}} = 1.64$, and we reject both H_2 and H_6. So, at $\Gamma = 1.40$, some of the $D_{\Gamma k}$ are below $\kappa_{\Gamma,\alpha} = 2.31$, but nonetheless closed testing rejects all seven hypotheses.

It is possible, in principle, to strengthen closed testing when there are logical implications among the hypotheses, $H_1,\ldots,H_K$, as is true here. Here, strengthening means changing the procedure so that it still controls the family-wise error rate but it may, from

time to time, reject an additional hypothesis not rejected by closed testing. For instance, Holm's (1979) method is the application of closed testing using the Bonferroni inequality, and Shaffer (1986) strengthened Holm's method when applied to the analysis of variance using logical implications among hypotheses. What are the logical implications in Table 4? Recall that the hypotheses assert that no one in certain subpopulations was affected by the treatment. If any of $H_2, \ldots, H_K$ is false, then H_1 is false. Similarly, if H_5 is false, so at least some smokers are affected, then either H_2 or H_3 or both must be false, because every smoker is either male or female. Bergmann and Hommel (1988) discuss the nontrivial general steps required to strengthen a closed testing procedure based on logical implications among hypotheses.

5 Aids to interpreting subgroup comparisons

The analysis in §4 yields indications of a beneficial effect of physical activity on survival in each subpopulation, but these indications are insensitive to larger biases for men than for women. In the second and third panel of Figure 1, the men are matched for observed covariates, so paired men are similar, as are paired women. However, the men may differ from the women; so, it is useful to examine the observed covariates within subgroups, as is done in Table 7. The men and women are of similar age, but the men are more likely than the women to smoke, drink alcohol, be working, be married, and they have somewhat less education.

The deviates, $D_{\Gamma k}$, in Table 4 may be affected by effect modification, but they are also affected by differing sample sizes. For instance, the deviate for the entire population, $D_{\Gamma 1}$, is based on 470 pairs, whereas the deviate for women, $D_{\Gamma 2}$, is based on 275 pairs of women, and the deviate for men, $D_{\Gamma 3}$, is based on 195 pairs of men. If there were an effect but there were no effect modification — that is, if men and women experienced the same effect

of physical inactivity — then we might reasonably expect $D_{\Gamma 1}$ to be larger than $D_{\Gamma 2}$ and $D_{\Gamma 3}$ simply because of the reduced sample size in subpopulations. To separate the sample size and insensitivity to unmeasured bias, a relevant point estimate would be helpful.

It is possible to produce a consistent point estimate of the design sensitivity, $\widetilde{\Gamma}_k$, for the kth statistic. Sample size does not affect the design sensitivity, as it is the limit as the sample size increases without bound. Differing sample sizes alone do not predict an increase or a decrease in the estimated design sensitivity, in contrast with the effect of sample size on the standardized deviates, $D_{\Gamma k}$. This estimate of $\widetilde{\Gamma}_k$ assumes that there is a treatment effect and no unmeasured bias, and then estimates the limiting sensitivity to unmeasured bias as the sample size in this subpopulation increases. In general, $\widetilde{\Gamma}_k$ depends upon the choice of test statistic. In the example, this is the Prentice-Wilcoxon statistic for censored paired survival times, because follow-up ended in 1992 for everyone. Given that the end of follow-up is a fixed date, it is safe to assume that the treatment, physical inactivity, did not affect the length of follow-up. The estimate of $\widetilde{\Gamma}_k$ solves for Γ in the equation $D_{\Gamma k} = 0$ or equivalently in the equation $S_k - \theta_{\Gamma k} = 0$. For all 470 pairs, the estimate of $\widetilde{\Gamma}_1$ is 2.32. For the 275 pairs of women, the estimate of $\widetilde{\Gamma}_2$ is 1.96. For the 195 pairs of men, the estimate of $\widetilde{\Gamma}_3$ is 2.91. In the example, both the deviates $D_{\Gamma k}$ and the estimates of $\widetilde{\Gamma}_k$ suggest there is greater insensitivity to bias for men, and that this is not a consequence of changing sample sizes. In contrast, if the paired survival times for men and for women were independent draws from the same censored bivariate population, then we would expect $D_{\Gamma 2}$ and $D_{\Gamma 3}$ to be smaller than $D_{\Gamma 1}$ because of the reduced sample size, but we would have $\widetilde{\Gamma}_1 = \widetilde{\Gamma}_2 = \widetilde{\Gamma}_3$, so the three point estimates would estimate the same quantity.

6 Pairs or sets that are not exactly matched for some covariates

To avoid confusing a main effect of gender and effect modification involving gender, we search for effect modification in pairs or matched sets that are exactly matched for gender, say in pairs of men, or in pairs of women. In the example in §1.1, all pairs were exactly matched for gender, smoking and the indicator of an income above twice the poverty level. With more potential effect modifiers, it may not be possible to match exactly for every potential effect modifier. What can be done in this case?

If a matched pair were exactly matched for gender, it seems reasonable to use that pair in an analysis that splits on gender, even if the pair is not exactly matched for other potential effect modifiers. Although there may be only a few pairs exactly matched for twenty covariates, it will often be the case that there are many pairs exactly matched for the first covariate, say gender, ignoring the rest, many pairs exactly matched for the second covariate ignoring the rest, and so on. It is straightforward to compare all the pairs of two men, all the pairs of two women, all the pairs of two smokers, etc. It simply requires a small change in the comparison weights, c_{gk}.

Refine the grouping of matched pairs or sets so that there are groups containing only men, groups containing only women, and groups containing matched sets that have both men and women. Then define the comparison weights for men so that a group g of sets containing only men has $c_{gk} = 1$ and all other groups have $c_{gk} = 0$. Define the comparison for women analogously. In this way, there is a comparison for men and a comparison for women, both comparisons use only sets that are exactly matched for gender, some pairs not matched for gender do not get used when analyzing gender, but some of these unused matched sets do get used in other comparisons, say the comparison of smokers.

7 Discussion

7.1 Using the submax method to study effect modification and its consequences

Effect modification is important in observational studies for several reasons.

If there were effect modification, then we expect to report firmer causal conclusions in subpopulations with larger effects. That is, we expect the design sensitivity and the power of the sensitivity analysis to be larger, so we expect to report findings that are insensitive to larger unmeasured biases in these subpopulations. Such a discovery is important in three ways. First, the finding about the affected subpopulation is typically important in its own right as a description of that subpopulation. Second, if there is no evidence of an effect in the complementary subpopulation, then that may be news as well. Third, if a sensitivity analysis convinces us that the treatment does indeed cause effects in one subpopulation, then this fact demonstrates the treatment does sometimes cause effects, and it makes it somewhat more plausible that smaller and more sensitive effects in other subpopulations are causal and not spurious. This is analogous to the situation in which we discover that heavy smoking causes lots of lung cancer, and are then more easily convinced that second-hand smoke causes some lung cancer, even though the latter effect is much smaller and more sensitive to unmeasured bias.

Conversely, it can be useful to discover evidence of a treatment effect of the same sign in every major subpopulation. We often worry whether the findings of an observational study in one population can generalize to second population that was not studied. Will a study done in Georgia generalize to Kansas where no study was done? If the second population were simply a different mixture of the same types of people — e.g., in Table 4, a different mixture of men and women, smokers and nonsmokers, rich and poor — then finding strong evidence of a nontrivial effect of constant sign in all subpopulations gives us

some reason to hope that the direction of effect found in the first population will reappear in the second population.

The simulation contrasted the new submax method with another method using groups formed by CART. One big difference between the two methods is that there is more theory concerning the performance of the submax method, including power, design sensitivity and Bahadur efficiency. The submax method achieves the largest design sensitivity of the subgroups, but there is no similar claim for the CART method. In the simulation, the CART method was cautious about forming groups, so it failed to capitalize on slight effect modification, with a loss of power in situation 3; however, that also meant that CART rarely paid a price for multiple testing when there was no effect modification in situation 2. One might tinker with the settings of the CART procedure or the simulation and produce a different result, but that is part of the attraction of the submax method: it has desirable properties that hold in general, without tinkering. In principle, the CART method might discover complex patterns of effect modification that the submax procedure does not consider. More practically, one could combine the two approaches, using the submax procedure with a combination of groups defined a priori, like gender, and a few groups suggested by CART, say poor, nonsmoking, men; however, we have not studied such a joint procedure, in part because it could only be evaluated by simulation.

7.2 Other uses of the submax method

Although we have discussed the submax method in §3 in the context of effect modification, the same mathematical calculation is useful in other contexts. The method looks at K specified comparisons, $k = 1, \ldots, K$, among the G groups of matched sets using weights $\mathbf{c}_k = (c_{1k}, \ldots, c_{Gk})^T$ of dimension G with $c_{gk} \geq 0$ for $g = 1, \ldots, G$. The $\mathbf{c}_k$ need not pick out subpopulations defined by measured covariates, such as men and women. Two

examples will be described briefly. Essentially, the examples distinguish groups of matched sets, but the groups were not formed using the observed covariates, and effect modification is not the concern.

If the treated condition were recorded in G increasing doses or intensities, then we could group matched sets with multiple controls into G groups of sets based on the dose given to the one treated subject in that matched set. The quality or relevance of the dose information may be uncertain. Consider three statistics defined by the comparisons $\mathbf{c}_1 = (1, 1, \ldots, 1)^T$, $\mathbf{c}_2 = (1, 2, \ldots, G)^T$ and $\mathbf{c}_3 = (0, 0, \ldots, 0, 1, 1, \ldots, 1)^T$. The comparison $\mathbf{c}_1$ uses all the matched sets with equal weights, ignoring the doses. The comparison $\mathbf{c}_2$ gives positive weight to all sets, but gives larger weight to sets with higher doses. The comparison $\mathbf{c}_3$ confines attention to sets that received high doses. See Rosenbaum (2010, §17.3) for calculations of design sensitivities for statistics using doses in different ways. The submax method would use all three statistics, reporting the least sensitive result, adjusting for multiple testing in a manner that reflects the high correlation between three tests that use the same data in different ways.

In an effort to provide information about unmeasured biases, Zubizaretta et al. (2012) produced two types of matched pairs: (i) pairs matched for the hospital providing the treatment in hospitals that used both the treatment and the control, and (ii) pairs with treated and control patients from different hospitals, one hospital that almost invariably used treatment and another hospital that almost invariably used the control. The first type of pair controls unmeasured covariates that are constant within each hospital, say the hospital's nurse-to-bed ratio. However, in the first type of pair, physicians looked at patients, giving treatment to some patients and control to others, so the first type of pair might be affected by selection bias. In the second type of pair, each patient received the treatment that the hospital almost invariably provides, reducing concern about the

selection of individuals for treatment, but the hospitals themselves and the communities they serve may differ in unmeasured ways. In this case, there are $G = 2$ groups of pairs. The comparison $\mathbf{c}_1 = (1,1)^T$ uses all pairs, $\mathbf{c}_2 = (1,0)^T$ uses type (i) pairs, and $\mathbf{c}_3 = (0,1)^T$ uses type (ii) pairs. The submax method would do all three tests with multiple comparisons, as in §4, taking account of the high correlation between comparison $\mathbf{c}_1$ and each of the other comparisons.

References

Bahadur, R. R. (1960), "Stochastic comparison of tests," *Annals of Mathematical Statistics*, 31, 276-295.

Bergmann, B. and Hommel, G. (1988), "Improvements of general multiple test procedures for redundant systems of hypotheses," in *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, New York: Springer, pp. 100-115.

Berk, R. H. and Jones, D. H. (1978), "Relatively optimal combinations of test statistics," *Scandinavian Journal of Statistics*, 5, 158-162.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Davis, M. A., Neuhaus, J. M., Moritz, D. J., Lein, D., Barclay, J. D. and Murphy, S. P. (1994), "Health behaviors and survival among middle aged and older men and women in the NHANES I Epidemiologic Follow-Up Study," *Preventive Medicine*, 23, 369-376.

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

Fogarty, C. B. and Small, D. S. (2016), "Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming,"

Journal of the American Statistical Association, 111, 1820-1830.

Genz, A. and Bretz, F. (2009), *Computation of Multivariate Normal and t Probabilities*, New York: Springer. (`R` package `mvtnorm`)

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

Hansen, B. B. and Klopfer, S. O. (2012), "Optimal full matching and related designs via network flows," *Journal of Computational and Graphical Statistics*, 15, 609-627. (`R` package `optmatch`)

Holm, S. (1979), "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, 6, 65-70.

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

Hsu, J. Y., Small, D. S. and Rosenbaum, P. R. (2013), "Effect modification and design sensitivity in observational studies," *Journal of the American Statistical Association*, 108, 135-148.

Hsu, J. Y., Zubizarreta, J. R., Small, D. S. and Rosenbaum, P. R. (2015), "Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods," *Biometrika*, 102, 767–782.

Lehmann, E. L. (1975), *Nonparametrics*, San Francisco: Holden-Day.

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

Marcus, R., Peritz, E. and Gabriel, K. R. (1976), "On closed testing procedures with special reference to ordered analysis of variance," *Biometrika*, 63, 655–60.

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

166.

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

O'Brien, P. C. and Fleming, T. R. (1987), "A paired Prentice-Wilcoxon test for censored paired data," *Biometrics*, 169-180.

Rosenbaum, P. R. (1991), "A characterization of optimal designs for observational studies," *Journal of the Royal Statistical Society* B, 53, 597-610.

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

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

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

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

Rosenbaum, P. R. and Silber, J. H. (2009b), "Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units," *Journal of the American Statistical Association*, 104, 501-511.

Rosenbaum, P. R. (2010), *Design of Observational Studies*, New York: Springer.

Rosenbaum, P. R. (2012), "Testing one hypothesis twice in observational studies," *Biometrika*, 99, 763-774.

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

Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonran-

domized studies," *Journal of Educational Psychology*, 66, 688-701.

Shaffer, J. P. (1986), "Modified sequentially rejective multiple test procedures," *Journal of the American Statistical Association*, 81, 826-831.

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

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

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

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

Zubizarreta, J. R., Cerdá, M. and Rosenbaum, P. R. (2013), "Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design," *Epidemiology*, 24, 79-87.

Table 1: Covariate balance in 470 matched, treatment-control pairs. The standardized difference (Std. Dif) is the difference in means before and after matching in units of the standard deviation before matching.

	Covariate Mean			Std. Dif.	
Covariate	Treated	Control	P-value	Before	After
Age	61.7	61.7	0.985	0.283	0.001
Male	0.415	0.415	1.000	-0.245	0.000
White	0.789	0.823	0.187	-0.252	-0.093
Poverty	0.460	0.460	1.000	0.377	0.000
Former Smoker	0.170	0.145	0.283	-0.142	0.064
Current Smoker	0.360	0.360	1.000	-0.141	0.000
Working last three months	0.247	0.247	1.000	-0.589	0.000
Married	0.621	0.666	0.153	-0.350	-0.099
Dietary Adequacy	3.254	3.379	0.143	-0.303	-0.098
	Education				
≤ 8	0.494	0.466	0.397	0.309	0.057
9-11	0.183	0.204	0.410	-0.097	-0.053
High School	0.166	0.172	0.794	-0.193	-0.016
Some College	0.066	0.070	0.796	-0.158	-0.015
College	0.085	0.085	1.000	0.038	0.000
Missing	0.006	0.002	0.317	0.004	0.054
	Alcohol Consumption				
Never	0.406	0.432	0.428	0.189	-0.053
<1 time per month	0.198	0.185	0.619	0.016	0.032
1-4 times per month	0.172	0.153	0.427	-0.125	0.048
2+ times per week	0.089	0.089	1.000	-0.069	0.000
Just about everyday/everyday	0.134	0.140	0.776	-0.073	0.000

Table 2: Correlation and covariance matrix $\boldsymbol{\rho}_\Gamma$ under H_0 for $D_{\Gamma k}$ for all $\Gamma \geq 1$ in the balanced situation, using Wilcoxon's statistic, with $L = 3$ potential effect modifiers.

	$D_{\Gamma 1}$	$D_{\Gamma 2}$	$D_{\Gamma 3}$	$D_{\Gamma 4}$	$D_{\Gamma 5}$	$D_{\Gamma 6}$	$D_{\Gamma 7}$
$D_{\Gamma 1}$	1.000	0.707	0.707	0.707	0.707	0.707	0.707
$D_{\Gamma 2}$	0.707	1.000	0.000	0.500	0.500	0.500	0.500
$D_{\Gamma 3}$	0.707	0.000	1.000	0.500	0.500	0.500	0.500
$D_{\Gamma 4}$	0.707	0.500	0.500	1.000	0.000	0.500	0.500
$D_{\Gamma 5}$	0.707	0.500	0.500	0.000	1.000	0.500	0.500
$D_{\Gamma 6}$	0.707	0.500	0.500	0.500	0.500	1.000	0.000
$D_{\Gamma 7}$	0.707	0.500	0.500	0.500	0.500	0.000	1.000

Table 3: The critical constant κ_α for $L = 0, \ldots, 15$ balanced binary effect-modifiers, using Wilcoxon's statistic, yielding $K = 2L + 1$ correlated tests with $\alpha = 0.05$. For comparison, the final column gives the critical constant obtained using the Bonferroni inequality, testing K one-sided hypotheses at family-wise level $\alpha = 0.05$.

L	$K = 2L + 1$	κ_α	Bonferroni
0	1	1.64	1.64
1	3	2.03	2.13
2	5	2.20	2.33
3	7	2.32	2.45
4	9	2.40	2.54
5	11	2.46	2.61
6	13	2.51	2.67
7	15	2.55	2.71
8	17	2.59	2.75
9	19	2.62	2.79
10	21	2.65	2.82
11	23	2.67	2.85
12	25	2.70	2.88
13	27	2.72	2.90
14	29	2.74	2.92
15	31	2.75	2.95

Table 4: Seven standardized deviates from Wilcoxon's test, $D_{\Gamma k}$, $k = 1, \ldots, K = 7$, testing the null hypothesis of no effect and their maximum, $D_{\Gamma \max}$, where the critical value is $d_\alpha = 2.31$ for $\alpha = 0.05$. Deviates larger than $d_\alpha = 2.31$ are in **bold**.

k	1	2	3	4	5	6	7	
Subpopulation	All	Female	Male	Non-smoker	Smoker	$> 2\times$ PL	$\leq 2\times$ PL	Maximum
	$D_{\Gamma 1}$	$D_{\Gamma 2}$	$D_{\Gamma 3}$	$D_{\Gamma 4}$	$D_{\Gamma 5}$	$D_{\Gamma 6}$	$D_{\Gamma 7}$	$D_{\Gamma \max}$
Sample-size	470	275	195	301	169	254	216	
$\Gamma = 1.00$	**6.09**	**3.79**	**4.88**	**4.67**	**3.92**	**3.88**	**4.71**	**6.09**
$\Gamma = 1.20$	**4.66**	**2.73**	**3.91**	**3.52**	**3.06**	**2.89**	**3.68**	**4.66**
$\Gamma = 1.40$	**3.48**	1.86	**3.11**	**2.57**	**2.36**	2.07	**2.83**	**3.48**
$\Gamma = 1.60$	**2.47**	1.11	**2.44**	1.76	1.76	1.37	2.10	**2.47**
$\Gamma = 1.64$	2.29	0.97	**2.32**	1.62	1.65	1.24	1.97	**2.32**
$\Gamma = 1.65$	2.24	0.94	2.29	1.58	1.63	1.21	1.94	2.29

Table 5: Theoretical power for Wilcoxon's signed rank test in subgroup analyses using (i) the maximum statistic $D_{\Gamma\,\max}$, (ii) an oracle that knows a priori which group has the largest effect (Oracle), and (iii) one statistic that sums all Wilcoxon statistics, thereby using all the matched pairs, $D_{\Gamma 1}$.

Situation		One covariate, $L = 1$			Five covariates, $L = 5$		
	Γ	$D_{\Gamma\,\max}$	Oracle	$D_{\Gamma 1}$	$D_{\Gamma\,\max}$	Oracle	$D_{\Gamma 1}$
$(\zeta_0, \zeta_1) = (0, 0)$	1	0.050	0.050	0.050	0.050	0.050	0.050
1. No effect. Values	1.01	0.035	0.033	0.033	0.035	0.033	0.033
are the size test.	1.2	0.000	0.000	0.000	0.000	0.000	0.000
	1.3	0.000	0.000	0.000	0.000	0.000	0.000
$(\zeta_0, \zeta_1) = (0.5, 0.5)$	1	1.000	1.000	1.000	1.000	1.000	1.000
2. Constant effect.	2.8	0.579	0.671	0.671	0.460	0.601	0.601
Every subgroup	3.0	0.177	0.218	0.218	0.126	0.167	0.167
has effect 0.5.	3.2	0.030	0.030	0.030	0.020	0.019	0.019
	3.4	0.004	0.002	0.002	0.002	0.001	0.001
$(\zeta_0, \zeta_1) = (0.6, 0.4)$	1	1.000	1.000	1.000	1.000	1.000	1.000
3. Slight effect	2.8	0.991	0.998	0.593	0.959	0.996	0.521
modification,	3.0	0.928	0.971	0.161	0.791	0.959	0.121
$\zeta_0 > \zeta_1$	3.2	0.733	0.855	0.018	0.492	0.816	0.011
	3.4	0.446	0.615	0.001	0.220	0.554	0.001
$(\zeta_0, \zeta_1) = (0.5, 0)$	1	1.000	1.000	1.000	1.000	1.000	1.000
4. Effect confined	2.8	0.268	0.418	0.000	0.113	0.369	0.000
to a subgroup.	3.0	0.071	0.144	0.000	0.020	0.117	0.000
$\zeta_1 = 0$	3.2	0.013	0.033	0.000	0.002	0.025	0.000
	3.4	0.002	0.006	0.000	0.000	0.004	0.000

Table 6: Simulated power (number of rejections in 10,000 replications) for Wilcoxon's signed rank test in subgroup analyses using (i) the maximum statistic $D_{\Gamma\,\max}$, (ii) groups built by CART, (iii) an oracle that knows a priori which group has the largest effect (Oracle), and (iv) one statistic that sums all of the Wilcoxon statistics, thereby using all matched pairs, $D_{\Gamma 1}$.

		One covariate, $L = 1$				Five covariates, $L = 5$			
$\mu = (\mu_0, \mu_1)$	Γ	$D_{\Gamma\,\max}$	CART	Oracle	$D_{\Gamma 1}$	$D_{\Gamma\,\max}$	CART	Oracle	$D_{\Gamma 1}$
(0,0)	1	540	525	525	525	515	504	503	503
	1.01	382	344	344	344	345	329	328	328
	1.1	7	1	1	1	7	7	7	7
	1.2	0	0	0	0	1	0	0	0
	1.3	0	0	0	0	0	0	0	0
(0.5, 0.5)	1	10000	10000	10000	10000	10000	10000	10000	10000
	2.8	5804	6713	6713	6713	4581	6014	6014	6014
	3.0	1643	2104	2101	2101	1215	1685	1681	1681
	3.2	279	315	313	313	158	187	183	183
	3.4	30	13	12	12	11	10	9	9
(0.6, 0.4)	1	10000	10000	10000	10000	10000	10000	10000	10000
	2.8	9913	7073	9977	6058	9589	6584	9955	5348
	3.0	9264	3788	9701	1657	7975	3471	9588	1242
	3.2	7387	2313	8565	173	5071	2212	8208	121
	3.4	4603	1535	6265	6	2245	1363	5679	8
(0.5, 0)	1	10000	10000	10000	10000	10000	10000	10000	10000
	2.8	2687	1908	4195	0	1105	1626	3686	0
	3.0	678	514	1476	0	174	398	1139	0
	3.2	120	100	320	0	23	67	208	0
	3.4	16	14	47	0	3	10	31	0

Table 7: Covariate means in 275 pairs of women and 195 pairs of men.

	Covariate Mean			
	Female		Male	
	Treated	Control	Treated	Control
Sample size	275	275	195	195
Age	61.2	61.0	62.5	62.7
Male	0.000	0.000	1.000	1.000
White	0.775	0.822	0.810	0.826
Education				
0-8	0.476	0.429	0.518	0.518
9-11	0.211	0.211	0.144	0.195
High School	0.185	0.207	0.138	0.123
Some College	0.069	0.084	0.062	0.051
College	0.051	0.069	0.133	0.108
Missing	0.007	0.000	0.005	0.005
Poverty	0.476	0.476	0.436	0.436
Former Smoker	0.116	0.080	0.246	0.236
Current Smoker	0.273	0.273	0.482	0.482
Working last three months	0.193	0.189	0.323	0.328
Married	0.502	0.553	0.790	0.826
Dietary Adequacy	3.045	3.139	3.549	3.716
Alcohol Consumption				
<1 time per month	0.222	0.222	0.164	0.133
1-4 times per month	0.116	0.135	0.251	0.179
2+ times per week	0.051	0.069	0.144	0.118
Just about everyday/everyday	0.084	0.084	0.205	0.221
Never	0.527	0.491	0.236	0.349

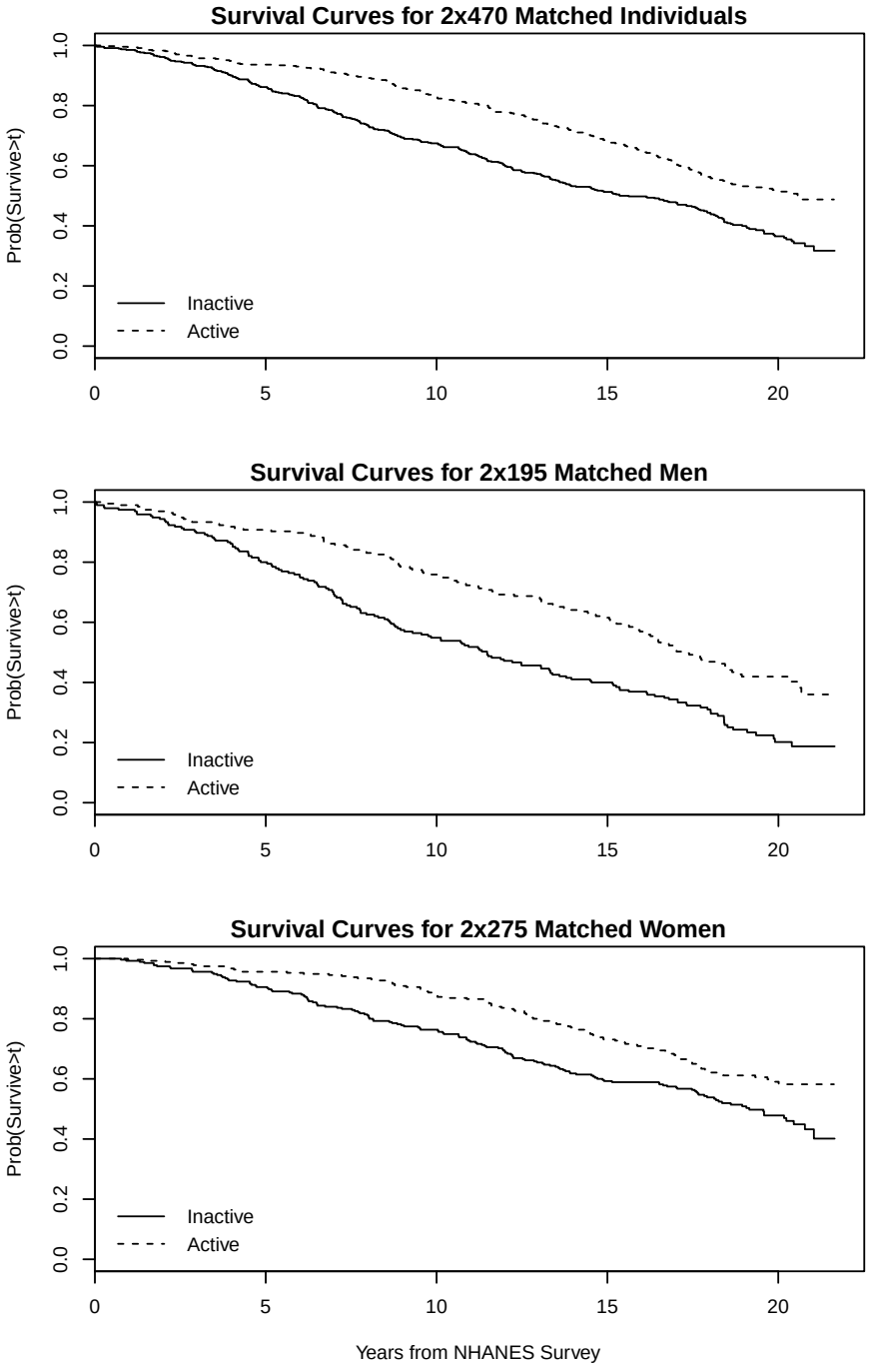

Figure 1: Survival in inactive and matched active groups following the NHANES survey.