Testing weak nulls in matched observational studies

Colin B. Fogarty

Abstract
We develop sensitivity analyses for the sample average treatment effect in matched observational studies while allowing unit-level treatment effects to vary. The methods may be applied to studies using any optimal without-replacement matching algorithm. In contrast to randomized experiments and to paired observational studies, we show for general matched designs that over a large class of test statistics, any procedure bounding the worst-case expectation while allowing for arbitrary effect heterogeneity must be unnecessarily conservative if treatment effects are actually constant across individuals. We present a sensitivity analysis which bounds the worst-case expectation while allowing for effect heterogeneity, and illustrate why it is generally conservative if effects are constant. An alternative procedure is presented that is asymptotically sharp if treatment effects are constant, and that is valid for testing the sample average effect under additional restrictions which may be deemed benign by practitioners. Simulations demonstrate that this alternative procedure results in a valid sensitivity analysis for the weak null hypothesis under a host of reasonable data-generating processes. The procedures allow practitioners to assess robustness of estimated sample average treatment effects to hidden bias while allowing for effect heterogeneity in matched observational studies.

Keywords
additivity, effect heterogeneity, randomization inference, sensitivity analysis, unmeasured confounding

1 | INTRODUCTION

Matching provides an appealing framework for inferring treatment effects in observational studies. Through the solution to an optimization problem, matching partitions individuals who have self-selected into treatment or control into matched sets on the basis of observed covariate information. In so doing, one tries to create a fair comparison between treatment and control individuals, with the hope that after matching observed discrepancies in outcomes may be attributed to differences in treatment status rather than differences in confounding factors. Should the observational study be free of hidden bias, a successful application of matching would justify modes of inference valid for finely stratified experiments (Fogarty, 2018). Matching throws the fundamental differences between randomized experiments and observational studies into stark relief: in randomized experiments such modes of inference are valid by design, whereas in observational studies the same procedures require an assumption of no hidden bias that is both untestable and untenable.
The simplest and most prevalent form of matching is pair matching, where each matched set contains exactly one treated and one control individual who are similar on the basis of the observed covariates. Pair matching is but one of the matching algorithms available to practitioners, and is the least flexible of all matching algorithms. Alternatives include fixed ratio matching, variable ratio matching, and full matching; see Stuart (2010) and Rosenbaum (2020) for a comprehensive overview of and worked examples using modern matching algorithms. These methods produce poststratifications with a common structure: within each matched set, there is either one treated individual and many controls, or one control individual and many treated individuals, and each individual appears in at most one matched set. As described in Rosenbaum (1991), any candidate matching not of this structure could be reduced to a match of this structure without increasing aggregate covariate discrepancy. Of these methods, full matching makes use of the largest percentage of data while maintaining balance on observed covariates (Hansen, 2004). Rosenbaum (1991) further illustrates that full matching enjoys desirable optimality properties in terms of minimizing aggregate covariate distance without discarding individuals. See Kang et al. (2016) for a worked example using full matching. Pair matching can be unnecessarily wasteful of the data if there is a large reservoir of control individuals who would be high-quality matches for each treated individual. That said, if there is not a large reservoir of controls, moving beyond pair matching may instead increase the bias in the treatment effect estimator as the typical match quality will degrade.

In a sensitivity analysis, the practitioner assesses the magnitude of hidden bias that would be required to overturn an observational study’s finding of a treatment effect. Methods for sensitivity analysis in matched observational studies have traditionally focused on tests of sharp null hypotheses, that is, hypotheses which impute the missing values for the potential outcomes. One common expedient is the assumption that treatment effects are constant across individuals. The restrictiveness of this assumption has been a point of contention in the analysis of randomized experiments and observational studies alike, with researchers in many domains instead desiring inference for Neyman’s weak null pertaining to the sample average of the treatment effects (Neyman, 1935). See Sabbaghi and Rubin (2014) and Wu and Ding (2020) for more on arguments surrounding tests of sharp versus weak nulls.

The observed treated-minus-control difference in means centered by the sample average treatment effect has an expectation equal to zero under both constant and heterogeneous effects in a completely randomized design; however, the variance for the difference in means computed under the assumption of constant effects may be too large or too small if instead effects are heterogeneous (Cohen & Fogarty, 2022; Ding, 2017; Loh et al., 2017). A single mode of inference that is both exact under constant effects and asymptotically correct for the weak null is attained by instead employing a studentized randomization distribution, where one simply permutes the centered difference in means divided by a suitable standard error estimator; see Bai et al. (2021) for developments in adaptively paired experiments, and see Janssen (1997) and Chung and Romano (2013) for related developments in robust permutation tests. In an observational study however, the centered difference in means would not have expectation zero under either the sharp or weak null in the presence of hidden bias. When conducting a sensitivity analysis under constant effects, one calculates the worst-case expectation for the test statistic being employed through the algorithm of Gastwirth et al. (2000). This calculation makes explicit use of the constant effects assumption, and one may be concerned that the worst-case expectation when allowing for heterogeneous effects would be materially larger than that under the sharp null.

This work develops sensitivity analyses for flexible matched designs that are valid under effect heterogeneity. The first procedure, developed in Section 3, tightly bounds the expectation under constant effects and retains validity for the weak null under an additional restriction on the potential outcomes. A superpopulation generative model is presented such that this restriction holds almost surely. The second approach, explored in Section 4 and aligning with the extant literature on sensitivity analysis under effect heterogeneity outside of matching, creates a sensitivity analysis that is always valid for the weak null but is conservative for the sharp null. In contrast to the paired case analyzed in Fogarty (2020), we show in Section 5 that over a large class of test statistics, no method of sensitivity analysis for general matched designs that tightly bounds the expectation assuming constant effects can also bound the worst-case expectation over the weak null. In matched observational studies, any sensitivity analysis that is valid for the weak null must be unduly conservative under constant effects, and any sensitivity analysis sharply bounding the expectation under constant effects must only be valid over a subset of the weak null, with the researcher left to decide how restrictive the subset is for any particular test statistic. Simulation studies in Section 6 indicate that the first approach, valid under an additional restriction on the potential outcomes, maintains validity for many reasonable data-generating mechanisms, while a data example in Section 7 highlights the practical benefits of this method. Our developments enable practitioners using observational data to assess the robustness of their findings to hidden bias in any matched design even if
effects are heterogeneous, providing methods for sensitivity analysis when practitioners are interested in average treatment effects.

2 | NOTATION AND REVIEW

2.1 | Notation for finely stratified designs

There are $B$ independent matched sets formed on the basis of observed pretreatment covariates, the $i$th of which contains $n_i$ individuals. There are $N = \sum_{i=1}^{B} n_i$ total individuals in the study. Each matched set contains one treated unit and $n_i - 1$ control units. The developments in this article extend in a straightforward way to full matching; see Rosenbaum (2002, Section 4, Exercise 12) for further details. Let $Z_{ij}$ be an indicator of whether or not the $j$th individual in block $i$ received the treatment, such that $\sum_{i=1}^{B} Z_{ij} = 1$ for all matched sets $i = 1, \ldots, B$. Along with a vector of measured covariates $x_{ij}$, each individual also has an unobserved covariate $0 \leq u_{ij} \leq 1$.

Let $r_{Tij}$ and $r_{Cij}$ denote the potential outcomes under treatment and control, respectively, for individual $ij$. The observed response is $R_{ij} = r_{Tij}Z_{ij} + r_{Cij}(1-Z_{ij})$, and the individual-level treatment effect $\tau_{ij} = r_{Tij} - r_{Cij}$ is not observable for any individual. Collect the potential outcomes, observed covariates, and unobserved covariate for each individual into the set $\mathbf{R} = \{r_{Cij}, r_{Tij}, x_{ij}, u_{ij} : i = 1, \ldots, B; j = 1, \ldots, n_i\}$, the contents of which will be conditioned upon as fixed properties of the observational study at hand. Naturally, the contents of $\mathbf{R}$ will vary with $N$, but this dependence is suppressed in what follows. We will write $\mathbf{R} = (R_{11}, \ldots, R_{Bn_B})^T$ to be the lexicographically ordered vector containing all of the observed responses, $R_i = (R_{i1}, \ldots, R_{in_i})^T$ for the observed responses in stratum $i$, and we will let the analogous notation hold for other vector quantities such as $\mathbf{u}$ and $\mathbf{r}_{C}$.

2.2 | Treatment assignments in experiments and observational studies

Let $\Omega = \{z : \sum_{i=1}^{n_i} z_{ij} = 1, i = 1, \ldots, B\}$ be the set of $\prod_{i=1}^{B} n_i$ possible values of $Z = (Z_{11}, Z_{12}, \ldots, Z_{Bn_B})^T$ under a finely stratified design where one individual receives the treatment and $n_i - 1$ receive the control, and let $\mathcal{Z}$ denote the event $\{Z \in \Omega\}$. In a finely stratified experiment (Fogarty, 2018; Pashley & Miratrix, 2021), $\Pr(Z = z \mid \mathcal{F}, \mathcal{Z}) = \Pr(Z = z \mid \mathcal{Z}) = |\Omega|^{-1}$, and $\Pr(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}) = \Pr(Z_{ij} = 1 \mid \mathcal{Z}) = 1/n_i$, where the notation $|A|$ denotes the cardinality of the set $A$. Before matching, individuals are assigned to treatment independently with unknown probabilities $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$. While one may hope that $\pi_{ij} \approx \pi_{ij'}$ after matching, proceeding as such may produce misleading inference due to the potential presence of unmeasured confounding. The model of Rosenbaum (2002, Chap. 4) states that individuals in the same matched set may differ in their odds of assignment to treatment by at most $\Gamma$.

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1-\pi_{ij'})}{\pi_{ij'}(1-\pi_{ij})} \leq \Gamma, (i = 1, \ldots, B; j, j' = 1, \ldots, n_i).$$ (1)

The parameter $\Gamma$ controls the degree to which unmeasured confounding may have impacted the treatment assignment process. The value $\Gamma = 1$ returns a finely stratified experiment, while $\Gamma > 1$ allows the unobserved covariates to influence the randomization distribution to a degree controlled by $\Gamma$. Returning attention to the matched structure by conditioning on $\mathcal{Z}$, Rosenbaum (2002, Section 4) shows that (1) may be equivalently stated as

$$\Pr(Z = z \mid \mathcal{F}, \mathcal{Z}) = \frac{\exp(\gamma z^T \mathbf{u})}{\sum_{\mathbf{r}_{C}} \exp(\gamma \mathbf{r}_{C}^T \mathbf{u})} = \prod_{i=1}^{B} \frac{\exp(\gamma \sum_{j=1}^{n_i} z_{ij} u_{ij})}{\sum_{j=1}^{n_i} \exp(\gamma u_{ij})},$$ (2)

where $\gamma = \log(\Gamma)$ and $\mathbf{u}$ lies in $U'$, the $N$-dimensional unit cube. At times, we will write $\Pr(\cdot \mid \mathcal{F}, \mathcal{Z})$ and $E(\cdot \mid \mathcal{F}, \mathcal{Z})$ as $\Pr(\cdot \mid \mathcal{F}, \mathcal{Z})$ and $E(\cdot \mid \mathcal{F}, \mathcal{Z})$ to reflect dependence on the hidden bias $\mathbf{u}$.

2.3 | The sample average treatment effect and its usual estimator

The sample average treatment effect is the average of the individual-level treatment effects for the $N$ individuals in the study population, $\bar{\tau} = N^{-1} \sum_{i=1}^{B} \sum_{j=1}^{n_i} \tau_{ij} = \sum_{i=1}^{B} (n_i/N) \bar{\tau}_i$. The value of $\bar{\tau}$ will change as $N$ varies, but this dependence is suppressed in what follows. At $\Gamma = 1$, the conventional unbiased estimator for $\bar{\tau}_i$, the average treatment effect for individuals in block $i$, is simply the observed difference in means between the treated and control individuals in block $i$, $\hat{\tau}_i = \sum_{j=1}^{n_i} (Z_{ij} r_{Tij} - (1 - Z_{ij}) r_{Cij})/(n_i - 1)$. An unbiased estimator for the sample average treatment effect in a finely stratified experiment is $\hat{\tau} = \sum_{i=1}^{B} (n_i/N) \hat{\tau}_i$; however, this estimator may be substantially biased in the presence of unmeasured confounding. In what follows, it will be useful to define an additional quantity $\delta_{ij} = r_{Tij} - \sum_{j' \neq j} r_{Cij'}/(n_i - 1)$ representing what the treated-minus-control difference in means would have been in stratum $i$ had individual $j$ received the treatment. We then have that $\hat{\tau}_i = \delta_i$ and $\hat{\tau}_i = \sum_{j=1}^{n_i} Z_{ij} \delta_{ij}$. 
2.4 | Sensitivity analysis: Sharp nulls, weak nulls, and alternative frameworks

Consider testing a null hypothesis \( H_0 \) pertaining to the treatment effects for the individuals in the observed study population. Let \( \varphi(\alpha, \Gamma) \) be a candidate sensitivity analysis for testing \( H_0 \) while maintaining the Type I error rate at or below \( \alpha \) when (1) is assumed to hold at \( \Gamma \), with \( \varphi \) taking the value 1 if test rejects and 0 otherwise. A sensitivity analysis is called (pointwise) asymptotically level \( \alpha \) if the potential outcomes satisfy \( H_0 \) and (1) holds at \( \Gamma \) for any \( \Gamma \). In what follows, we will develop sensitivity analyses for the weak null hypothesis that the sample average treatment effect equals a particular value \( \tau_0 \),

\[
H_N^{(\tau_0)} : \bar{\tau} = \tau_0.
\]

We will also assess performance under the sharp null hypothesis that the treatment effects \( \tau_{ij} \) are constant at \( \tau_0 \) for all individuals \( i \) in the study population,

\[
H_F^{(\tau_0)} : r_{ij} = r_{ij0} \Leftrightarrow \tau_{ij} = \tau_0 \quad (i = 1, \ldots, B; j = 1, \ldots, n_i).
\]

Through inverting the resulting procedures by enumerating the set of values \( \tau_0 \) for which the methods fail to reject the null at any given value of \( \Gamma \) in (1), we can additionally provide sensitivity intervals for the treatment effect \( \tau_0 \) at any \( \Gamma \).

Alternative approaches for sensitivity analysis exist which have been primarily developed outside the context of matched observational studies. See, among several, Cornfield et al. (1959), Robins et al. (2000), Imbens (2003), Yu and Gastwirth (2005), Egleston et al. (2009), Hosman et al. (2010), Liu et al. (2013), VanderWeele and Ding (2017), and Franks et al. (2020). Of particular relevance to the model (1) is a recent literature developing methods for sensitivity analysis under the marginal sensitivity model, proposed by Tan (2010) and further developed by Zhao et al. (2019), which instead bounds the odds ratio between \( \pi_{ij} \) and the propensity score \( \text{pr}(Z_{ij} = 1 \mid X = x_{ij}) \) between values \( \Lambda^{-1} \) and \( \Lambda \) for some \( \Lambda \geq 1 \) for each \( i,j \). If individuals in the same matched set are perfectly matched on their propensity scores, then the marginal sensitivity model holding at \( \Lambda = \sqrt{1} \) implies that (1) holds at \( \Gamma \) (Zhao et al., 2019, Proposition 7.1). That is, the methods presented herein under (1) are also applicable should one prefer the marginal sensitivity model. Alternative approaches derived under that model are typically based upon inverse probability weighted estimators, assume a superpopula-

tion model, and prioritize performance for inference on the average treatment effect, de-emphasizing performance under the sharp null hypothesis in a departure from the extant literature developed under (1) for matched designs.

3 | A SHARP SENSITIVITY ANALYSIS UNDER CONSTANT EFFECTS AND ITS PERFORMANCE UNDER THE WEAK NULL

3.1 | A statistic with known worst-case expectation under constant effects

Suppose interest lies in the null hypothesis \( H_N^{(\tau_0)} \) with (1) holding at \( \Gamma \), and consider

\[
D_{\Gamma}^{(\tau_0)} = \bar{\tau}_{\Gamma} - \tau_0 - \left( \frac{\Gamma - 1}{1 + \Gamma} \right) \left| \bar{\tau}_{\Gamma} - \tau_0 \right| - \sum_{i=1}^{n_\Gamma} \left( \frac{\Gamma - 1}{1 + \Gamma} \right) \left| \sum_{j=1}^{n_i} Z_{ij} (\delta_{ij} - \tau_0) \right|,
\]

with \( \delta_{ij} \) defined as in Section 2.3. \( D_{\Gamma}^{(\tau_0)} \) is precisely the treated-minus-control mean difference in stratum \( i \), subtracted by the worst-case expectation under a paired design for \( H_F^{(\tau_0)} \) when (1) holds at \( \Gamma \). Fogarty (2020) demonstrated that \( D_{\Gamma}^{(\tau_0)} \) can be used to construct an asymptotically valid sensitivity analysis for \( H_N^{(\tau_0)} \) in paired observational studies. As will be demonstrated, this form continues to prove useful in general matched designs.

Define the weighted average

\[
D_{\Gamma}^{(\tau_0)} = \sum_{i=1}^{B} \left( \frac{n_i}{N} D_{\Gamma}^{(\tau_0)} \right),
\]

which weights each set’s contribution as is typical in finely stratified experiments for inference on the sample average effect. To estimate the variance of \( D_{\Gamma}^{(\tau_0)} \), we follow the construction in Fogarty (2018). Let \( Q = (Q_1, \ldots, Q_B)^T \) with \( Q_i = B(n_i/N) \) for \( i = 1, \ldots, B \), and let \( H_Q = Q (Q^T Q)^{-1} Q^T \) be the hat matrix corresponding to \( Q \). Let \( Y_{\Gamma} = B(n_i/N) D_{\Gamma}^{(\tau_0)} / \sqrt{1 - h_{Qii}} \) where \( h_{Qii} \) is the \( \{i,i\} \) element of \( H_Q \), and let \( Y_{\Gamma} = (Y_{\Gamma 1}, \ldots, Y_{\Gamma B})^T \). Define the squared standard error

\[
\text{se}^2(D_{\Gamma}^{(\tau_0)}) = Y_{\Gamma}^T (I - H_Q) Y_{\Gamma} / B^2.
\]
Consider the candidate level-\(\alpha\) sensitivity analysis when (1) is assumed to hold at \(\Gamma\)

\[
\varphi^{(\tau_0)}(\alpha, \Gamma) = 1 \left\{ \frac{D_\Gamma^{(\tau_0)}}{\text{se}(D_\Gamma^{(\tau_0)})} \geq \Phi^{-1}(1-\alpha) \right\},
\]

where \(\Phi(\cdot)\) is the standard normal cumulative distribution function. The following theorem describes the performance of \(\varphi^{(\tau_0)}\) under the sharp and weak null.

**Theorem 1.** Suppose (1) holds at \(\Gamma\), and consider any \(\alpha \leq 0.5\). If \(H_F^{(\tau_0)}\) holds, then under suitable regularity conditions ensuring asymptotic normality for \(D_\Gamma^{(\tau_0)}\) and the existence of a probability limit for \(\text{se}(D_\Gamma^{(\tau_0)})\), \(\lim_{B \to \infty} \sup_{\Gamma} E[\varphi^{(\tau_0)}(\alpha, \Gamma) \mid F, Z] \leq \alpha\) with equality possible in any matched design. If instead only \(H_N^{(\tau_0)}\) holds, then a sufficient condition for \(\lim_{B \to \infty} \sup_{\Gamma} E[\varphi^{(\tau_0)}(\alpha, \Gamma) \mid F, Z] \leq \alpha\) is that, in addition to regularity conditions, the following holds:

**Condition 1.** Let (1) hold at \(\Gamma\) and let \(u\) be the true vector of unmeasured confounders. There exists a vector of hidden bias \(u'\) such that the following inequalities both hold:

\[
E_u(D_\Gamma^{(\tau_0)}) | F, Z \leq E_u(D_\Gamma^{(\tau_0)}) | F, Z;
\]

\[
\frac{\Gamma - 1}{1 + \Gamma} \sum_{i = 1}^b (n_i/N)E_u[|\tau_i - \hat{\tau}_i| | F, Z] \leq \frac{\Gamma - 1}{1 + \Gamma} \sum_{i = 1}^b (n_i/N)E_u[|\tau_i - \hat{\tau}_i| | F, Z].
\]

(4)

Observe that by taking \(u'\) equal to the vector \(u\) that maximizes \(E_u\left(D_\Gamma^{(\tau_0)} | F, Z\right)\), Condition 1 can be shown to always hold in matched pair designs under \(H_N^{(\tau_0)}\), reflecting results in Fogarty (2020) for paired designs. For other matched designs, potential outcome allocations can be constructed such that Condition 1 is violated. In short, the potential outcomes must be such that, in aggregate, \(\tau_0\) better approximates the median of \(\hat{\tau}_i\) than does \(\hat{\tau}_i\). Importantly, this would not be expected to hold in cases of extreme effect heterogeneity when \(\hat{\tau}_i\) and \(\hat{\tau}_i\) might be drastically different due to the baseline characteristics for individuals in matched set \(i\). Rather, a certain relationship between the skewness of \(\hat{\tau}_i\) and the value of \(\hat{\tau}_i\) must be present. In Section 3.2, we present further insight into this condition through a superpopulation generative model which ensures that it holds.

In words, Theorem 1 states that the proposed sensitivity analysis \(\varphi^{(\tau_0)}(\alpha, \Gamma)\) is asymptotically sharp under the constant effect model \(H_F^{(\tau_0)}\): under this null, if (1) holds at \(\Gamma\) and the sensitivity analysis is conducted at \(\Gamma\), there exists a vector of unmeasured confounding \(u\) such that the limiting size of the procedure is exactly \(\alpha\). Furthermore, the theorem states that Condition 1 implies (in concert with asymptotic normality and the existence of a probability limit for the standard error estimate) that \(\varphi^{(\tau_0)}(\alpha, \Gamma)\) also provides a valid sensitivity analysis for the weak null \(H_N^{(\tau_0)}\). When interpreted as a test of the sharp null \(H_F^{(\tau_0)}\), the procedure \(\varphi^{(\tau_0)}\) delivers an asymptotically valid sensitivity analysis for any \(\Gamma\) in (1): for any \(\Gamma\), a rejection using this method allows us to say that even if hidden bias existed to the degree \(\Gamma\) we still find evidence against the treatment effect being constant at \(\tau_0\), hence ruling out the possibility that evidence for a treatment effect might be explained away by hidden bias. When describing the nature of the detected effect, a rejection of \(\varphi^{(\tau_0)}\) for a particular \(\Gamma\) in (1) indicates either that \(\hat{\tau} > \tau_0\), or that \(\hat{\tau} \leq \tau_0\) but that \(\sum_{i = 1}^b (n_i/N)E(u_i | \tau_i - \tau_0 | F, Z) < E_u(\hat{\tau} - \tau_0 | F, Z)\), which can only occur with heterogeneous effects by Lemma 1. Condition 1 rules out the latter possibility.

The proof is deferred to the web-based supporting information, but relies heavily upon the following two lemmas:

**Lemma 1.** Consider \(A_{\tau_i} = \sum_{i = 1}^{n_i} Z_{ij}q_{ij} - |q_{ij}|(\Gamma - 1)/(1 + \Gamma)\), where \(q_{ij}\) are any constants such that \(\sum_{i = 1}^{n_i} q_{ij} = 0\). Suppose (1) holds at \(\Gamma\). Then, \(E_u(A_{\tau_i} | F, Z) \leq 0\), and \(E_u(A_{\tau_i} | F, Z) = 0\) if \(u_{ij} = 1(1(\hat{q}_{ij} \geq 0))\).

**Lemma 2.**

\[
E[\text{se}^2(D_\Gamma^{(\tau_0)}) | F, Z] = \frac{1}{B^2} E[Y_i | F, Z]^2 (1 - H_0) E[Y_i | F, Z] \geq 0.
\]

Furthermore, under suitable regularity conditions, \(\text{se}^2(D_\Gamma^{(\tau_0)})/\{\text{var}(D_\Gamma^{(\tau_0)}) | F, Z\}\) converges in probability to a value greater than or equal to one.

Observe that Lemma 1 applies to \(D_\Gamma^{(\tau_0)}\). If (1) is assumed to hold at \(\Gamma\), this implies that the worst-case expectation for \(D_\Gamma^{(\tau_0)}\) under \(H_F^{(\tau_0)}\) equals zero as \(\hat{\tau}_i = \tau_0\). Moreover, under Condition 1, the worst-case expectation for \(D_\Gamma^{(\tau_0)}\) under \(H_N^{(\tau_0)}\) is also bounded above by zero if (1) holds at \(\Gamma\). Lemma 2 follows from Proposition 1 and Theorem 2 of Fogarty (2018), and the proof is omitted. The potential for conservativeness of the standard errors is not a deficiency of the estimators themselves, but rather a fundamental limitation of inference under the finite population model (Ding, 2017). Regularity conditions on \(F\) are discussed in the Supporting Information. They are needed to ensure that a central limit theorem holds for \(\hat{D}_\Gamma^{(\tau_0)}\), and that \(\text{se}(\hat{D}_\Gamma^{(\tau_0)})\) has a limits in probability. In the Supporting Information, we also describe a modification of \(\varphi^{(\tau_0)}(\alpha, \Gamma)\).
which replaces critical values from a standard normal with critical values from a biased randomization distribution. At \( \Gamma = 1 \), this provides a test that is both exact for \( H_F^{(\tau_0)} \) and asymptotically correct for \( H_N^{(\tau_0)} \).

3.2 | A superstrobolar model yielding asymptotic validity

To provide additional intuition, we now present conditions within a superstrobolar formulation which are sufficient for Condition 1 to hold almost surely. Suppose that for \( i = 1, \ldots, B \), stratum sizes \( n_i \) are drawn iid from some distribution \( F \). For each individual \( j \) in matched set \( i \), measured and unmeasured covariates are drawn independently from an \( n_i \)-dependent distribution \( G(x_{ij}, u_{ij} \mid n_i) \). Finally, given \( (x_{ij}, u_{ij}) \) potential outcomes for each individual are drawn independently from a distribution \( H(r_{Tij}, r_{Cij} \mid x_{ij}, u_{ij}) \). Suppose that \( (r_{Tij}, r_{Cij}, x_{ij}, u_{ij}) \) and \( n_i \) all have finite first moments. Treatment is assumed strongly significantly given \( x_{ij}, n_{ij} \), and conditional treatment assignments given \( Z, F \) are drawn based upon the model (1) for some \( \Gamma \).

For each \( i \), the treatment effect estimator \( \hat{\tau}_i \) is formed as \( \sum_{j=1}^{n_i} Z_{ij} r_{Tij} - (n_i - 1)^{-1} \sum_{j=1}^{n_i} (1 - Z_{ij}) r_{Cij} \) as before.

Proposition 1. Under the outlined superstrobolar model, suppose that

\[
pr\{\hat{\tau}_i \leq \bar{\tau}_i \mid n_i, \bar{\tau}_i, Z\} = pr\{\bar{\tau}_i \leq \hat{\tau}_i \mid Z\}.
\]

Then, evaluating with \( u' = u \), almost surely

\[
\limsup_{\beta \to \infty} \sum_{i=1}^{B} \frac{(n_i/N)E_u[\|\hat{\tau}_i - \bar{\tau}_i\| \mid \bar{\tau}_i, F, Z] - \sum_{i=1}^{B} (n_i/N)E_u[\|\hat{\tau}_i - \bar{\tau}_i\| \mid F, Z]}{\beta} \leq 0,
\]

such that Condition 1 holds almost surely

The condition outlined in Proposition 1 imposes a relationship between \( (n_i, \bar{\tau}_i) \), and the probability that \( \hat{\tau}_i - \bar{\tau}_i \) is below zero. For it to be violated, the distribution of \( \hat{\tau}_i - \bar{\tau}_i \) must depend upon \( n_i \) and/or \( \bar{\tau}_i \). Importantly, treatment effects may continue to vary across matched sets under this restriction, which instead targets the distribution of residuals \( \hat{\tau}_i - \bar{\tau}_i \). This condition is sufficient, but not necessary, and weaker conditions may be developed using the proof of Proposition 1. For instance, for a violation of Condition 1 to occur, it must actually be the case that \( pr[\hat{\tau}_i \leq \bar{\tau}_i \mid n_i, \bar{\tau}_i, Z] \) is positively correlated with \( n_i(\bar{\tau}_i - \bar{\tau}) \), suggesting that the distribution of \( \hat{\tau}_i - \bar{\tau}_i \) is more left-skewed for larger values of \( n_i \bar{\tau}_i \) than it is for smaller values. See the web-based Supporting Information for a proof of Proposition 1, where more details are provided.

4 | A VALID SENSITIVITY ANALYSIS FOR THE WEAK NULL WHICH IS CONSERVATIVE UNDER CONSTANT EFFECTS

4.1 | Exploiting interval restrictions on stratumwise assignment probabilities

We now develop a sensitivity analysis which controls the worst-case expectation for all allocations of potential outcomes within the weak null \( H_N^{(\tau_0)} \), in turn facilitating a sensitivity analysis for the weak null that does not require a condition such as Condition 1. The approach involves modifying the form of \( D_N^{(\tau_0)} \) in (3) by penalizing the contributions of matched sets to the test statistic depending on the number of individuals in each matched set. As will be shown, such a penalty ensures that the worst-case expectation of the resulting test statistic is bounded above by zero under \( H_N^{(\tau_0)} \) when (1) holds at \( \Gamma \), with equality possible. This will come at the expense of performance when effects are actually constant at \( \tau_0 \), that is, when \( H_F^{(\tau_0)} \) holds. Under constant effects, the procedure will always attain a size strictly smaller than \( \alpha \) unless one has a paired observational study.

Let \( \varphi_{ij} = pr(Z_{ij} = 1 \mid F, Z) \) be the conditional assignment probability for individual \( ij \). For a given stratum size \( n_i \), if (1) holds at \( \Gamma \) the resulting conditional assignment probabilities are confined to

\[
\frac{1}{\Gamma(n_i - 1) + 1} \leq \varphi_{ij} \leq \frac{\Gamma}{(n_i - 1) + \Gamma}.
\]

Observe that the denominators in the bounds above are equal for paired designs, but are unequal for \( n_i \geq 2 \) unless \( \Gamma = 1 \). Let \( \tilde{\varphi}_{n_i} = \Gamma(n_i - 1) + 1 \) be the denominator in the lower bound, and reexpress (5) as

\[
\tilde{\varphi}_{n_i}^{-1} \leq \varphi_{ij} \leq \Gamma n_i \tilde{\varphi}_{n_i}^{-1} ; \quad \Gamma n_i = \Gamma \left\{ \frac{\Gamma(n_i - 1) + 1}{(n_i - 1) + \Gamma} \right\}.
\]

For matched pairs, we have that \( \Gamma n_i = \Gamma \); however, for \( n_i > 2 \), \( \Gamma n_i > \Gamma \) if \( \Gamma > 1 \), and in fact \( \Gamma n_i \to \Gamma^2 \) as \( n_i \to \infty \). Consider the random variable

\[
\tilde{D}_F^{(\tau_0)} = \hat{\tau} - \tau_0 - \left( \frac{\Gamma n_i - 1}{1 + \Gamma n_i} \right) |\hat{\tau} - \tau_0|,
\]

and define the weighted average

\[
\tilde{K}_F^{(\tau_0)} = \sum_{i=1}^{B} \frac{1}{\Gamma n_i} \tilde{\varphi}_{n_i} \tilde{D}_F^{(\tau_0)}.
\]
It can be shown that $R_{Γ}^{(r)}$ is less than or equal to a constant multiple of $R(Γ)$, as matched sets where $n_i > 2$ contribute less to $R_{Γ}^{(r)}$ than they do to $D_{Γ}^{(r)}$. This will facilitate Type I error control under the weak null $H_{N}^{(r)}$, at the expense of performance under the sharp null of constant effects $H_{F}^{(r)}$.

To estimate the variance of $R_{Γ}^{(r)}$, let $Q_{i} = (Q_{i1},...,Q_{iB})^T$ with $Q_{i1} = (B/N)(1 + Γ_{ni})/Γ_{ni}$ for $i = 1,...,B$, and let $H_{OQ} = Q_{Γi}Q_{Γi}^T - 1$ be the hat matrix corresponding to $Q$. Let $Y_{Γi} = (Y_{Γ1},...,Y_{ΓB})^T$, and define the squared standard error $se(Φ_n, Γ) = Y_{Γi}^T(I - H_{OQ})Y_{Γi}/B^2$.

Consider the candidate level-α sensitivity analysis when (1) is assumed to hold at $Γ$.

$$\phi_n^{(r)}(Γ, α, Γ) = 1 \left\{ \frac{\hat{K}_{Γ}^{(r)}}{se(Φ_n, Γ)} \geq Φ^{-1}(1 - α) \right\},$$

where $Φ(·)$ is the standard normal cumulative distribution function.

**Theorem 2.** Suppose (1) holds at $Γ$, and consider any $α \leq 0.5$. If the weak null $H_{N}^{(r)}$ is true, then under suitable regularity conditions $\limsup_{B→∞}E[Φ_n^{(r)}(Γ, α) \mid F, Z] ≤ α$, and equality is possible. If $H_{F}^{(r)}$ also holds, $Γ > 1$, and we do not have a paired design, $\limsup_{B→∞}E[Φ_n^{(r)}(Γ, α) \mid F, Z] < α$, the inequality being strict.

For paired observational studies or if $Γ = 1$, $R_{Γ}^{(r)}$ and $D_{Γ}^{(r)}$ are equivalent, implying no divergence between these modes of inference. For general matched designs with $Γ > 1$, the method $\phi_n^{(r)}(Γ, α, Γ)$ based upon $R_{Γ}^{(r)}$ is valid with heterogeneous effects but is unduly conservative for constant effects. Sensitivity analyses using $Φ_n^{(r)}(Γ, α, Γ)$, based upon $D_{Γ}^{(r)}$, are valid and asymptotically sharp if treatment effects are constant but require additional assumptions to guarantee validity if treatment effects are instead heterogeneous.

The proof of Theorem 3 involves the following additional lemma:

**Lemma 3.** Suppose (1) holds at $Γ$. Then,

$$E(D_{Γ}^{(r)} \mid F, Z) \leq \left(\frac{n_i}{k_{Γi}}\right) \left(\frac{2Γ_{ni}}{1 + Γ_{ni}}\right)(τ - τ_0).$$

We see through the form of $R_{Γ}^{(r)}$ that its expectation is upper bounded by zero when $\bar{r} = τ_0$. The analog of Lemma 2 provides a probability limit for our variance estimator, such that validity under the weak null follows so long as a central limit theorem holds. Strict conservativeness under $H_{F}^{(r)}$ follows from Lemma 1; see the Supporting Information for details.

**5 | CHALLENGES FACING SENSITIVITY ANALYSIS UNDER EFFECT HETEROGENEITY**

**5.1 | An impossibility result for general matched designs**

Theorems 1 and 2 present two methods for inference with diverging properties under the sharp and weak null outside of paired designs. The procedure $Φ_n^{(r)}(Γ, α, Γ)$, described in Section 3, provides a sensitivity analysis that is sharp under $H_{F}^{(r)}$, but requires Condition 1 to ensure validity under the weak null $H_{N}^{(r)}$. The procedure $\hat{ϕ}_n^{(r)}(Γ, α, Γ)$, presented in Section 4, provides a sensitivity analysis that is valid for $H_{F}^{(r)}$ so long as a central limit theorem is satisfied, but is unnecessarily conservative under the sharp null $H_{F}^{(r)}$. In this section, we assess whether or not these trade-offs are unavoidable. That is, can we instead create a sensitivity analysis that is simultaneously sharp over the weak null $H_{N}^{(r)}$ and under constant effects $H_{F}^{(r)}$ without imposing a further restriction on the weak null such as Condition 1?

For $k$ scalar valued, let $h_{Γni}(k)$ be a monotone nondecreasing, nonconstant function allowed to depend upon both the stratum size $n_i$ and the value of $Γ$ at which the sensitivity analysis is being conducted. In the $i$th stratum, let $μ_{Γi}$ be the worst-case expectation for $h_{Γni}(r - τ_0)$ under the sharp null $H_{F}^{(r)}$ when (1) is assumed to hold at $Γ$; unlike under the weak null, this worst-case expectation can be explicitly calculated using the observed data by following the procedure in Gastwirth et al. (2000). If (1) holds at $Γ$, by construction we have that $B^{-1} \sum_{i=1}^{N} E[h_{Γni}(r - τ_0) - μ_{Γi} \mid F, Z] ≤ 0$ when $H_{F}^{(r)}$ actually holds. If $H_{F}^{(r)}$ is instead false but $H_{N}^{(r)}$ is true, $μ_{Γi}$ varies over $z \in Ω$. The sensitivity analysis assuming constant effects would also bound the worst-case expectation under $H_{F}^{(r)}$ for a given collection of functions $\{h_{Γni}\}_{i=2}^{N}$ if $B^{-1} \sum_{i=1}^{N} E[h_{Γni}(r - τ_0) - μ_{Γi} \mid F, Z] ≤ 0$ under $H_{N}^{(r)}$ when (1) holds at $Γ$ for any $Γ$. The following result shows that this cannot be guaranteed over all elements of the composite null $H_{N}^{(r)}$ for general matched designs.

**Theorem 3.** For any collection of nondecreasing, nonconstant functions $\{h_{Γni}\}_{i=2}^{N}$, there exist combinations of
stratum sizes $n_i$ ($i = 1, ..., B$), degrees of hidden bias $\Gamma$, and values for potential outcomes satisfying $H_N^{(c_0)}$ such that if (1) holds at $\Gamma$,  
\[
\max_{u \in L'} \frac{1}{B} \sum_{i=1}^{B} E_u \{ h_{\Gamma_0} (\tilde{r}_i - \tau_0) - \mu_{\Gamma} | F, Z \} > 0.
\]

That is, no sensitivity analysis using the worst-case expectation under $H_F^{(c_0)}$ can bound the worst-case expectation under $H_N^{(c_0)}$ when effects are heterogeneous for all possible combinations of stratum sizes, degrees of hidden bias, and for all elements of the weak null.

The theorem, proven in the Supporting Information, shows that for a large class of test statistics, including any weighted average of treated-minus-control differences in means across strata, $D^{(c_0)}_\Gamma$ and $K^{(c_0)}_\Gamma$, it is in general impossible to devise an upper bound for $B^{-1} \sum_{i=1}^{B} E_u \{ h_{\Gamma_0} (\tilde{r}_i - \tau_0) | F, Z \}$ that is simultaneously tight under both the sharp null and the weak null outside the confines of pair matching. If the bound is tight when effects are assumed constant at $\tau_0$, it must generally not bound the expectation when instead only $\tilde{r} = \tau_0$. The properties of $\varphi^{(c_0)}(\alpha, \Gamma)$ and the requirement of Condition 1 for its validity under the weak null reflect this: $E(D^{(c_0)}_\Gamma | F, Z)$ can equal zero under the sharp null, thus implying $E(D^{(c_0)}_\Gamma | F, Z)$ cannot be bounded above by zero for all potential outcome allocations within $H_N^{(c_0)}$. It must only be bounded above by zero for some subset of $H_N^{(c_0)}$, with the researcher left to judge the stringency of the restriction. On the other hand, if the bound is tight under the weak null $H_N^{(c_0)}$, it must be a strict upper bound when effects are actually constant at $\tau_0$. The behavior of $\varphi^{(c_0)}(\alpha, \Gamma)$ reflects this: $E(K^{(c_0)}_\Gamma | F, Z)$ cannot exceed zero under the weak null, but is strictly less than zero under the sharp null outside of paired designs. Moreover, one could find a less conservative bound for $E(K^{(c_0)}_\Gamma | F, Z)$ than zero under $H_F^{(c_0)}$ by applying the method of Gastwirth et al. (2000), which explicitly calculates the worst-case expectation under a sharp null hypothesis.

### 6  SIMULATIONS WITH WORST-CASE HIDDEN BIAS

#### 6.1  A generative model for inference on the sample average treatment effect

Through a simulation study, we now compare the sensitivity analysis based on $D^{(c_0)}_\Gamma$, $\varphi^{(c_0)}(\alpha, \Gamma)$ to that using $K^{(c_0)}_\Gamma$, $\varphi^{(c_0)}(\alpha, \Gamma)$. There are $B = 500$ matched sets in the $m$th iteration, potential outcomes are drawn as  
\[
r_{C\mid j} = \varepsilon_{C\mid j}; \quad r_{T\mid j} = r_{C\mid j} + \varepsilon_{T\mid j} + \beta_i,
\]

where $\beta_i$ are independent and identically distributed according to some $F_{\beta}()$, $\varepsilon_{C\mid j}$ are independently distributed according to $F_{\varepsilon_{C\mid j}}(\cdot; \beta_i)$, $\varepsilon_{T\mid j}$ are independent according to $F_{\varepsilon_{T\mid j}}(\cdot; \beta_i)$, $\varepsilon_{C\mid j}$ and $\varepsilon_{T\mid j}$ are independent within and across individuals, and $E(\beta_i) = E(\varepsilon_{C\mid j}) = E(\varepsilon_{T\mid j}) = 0$. The variance of $\beta_i$ affects the across-set effect heterogeneity (potentially reflecting the impact of effect modifiers $x_i$), while $\varepsilon_{T\mid j}$ affects the within-set effect heterogeneity. Note the potential dependence of $F_{\varepsilon_{C\mid j}}(\cdot; \beta_i)$ and $F_{\varepsilon_{T\mid j}}(\cdot; \beta_i)$ on $\beta_i$.

Table 1 provides the 11 combinations of distributions for $\beta_i$, $\varepsilon_{C\mid j}$, and $\varepsilon_{T\mid j}$ used in the simulation. Under choice (a), $H_F(0)$ holds in all simulations. For all other settings, the sharp null does not hold. Furthermore, the marginal distributions for potential outcomes under treatment and control in each set $i$ are not equal. One should keep in mind in what follows that a rejection of the null would be the correct conclusion by either of those definitions of no effect, and statements of validity or invalidity in this simulation study are in reference to performance under the weak null hypothesis. Choices (b) and (c) lead to a symmetric distribution for $\delta_j$ within each matched set, with more across-set treatment effect heterogeneity in (c) than in (b). Choices (d) and (e) lead to right-skewed distributions for $\delta_j$ with more heterogeneity of effects in (e), and (f) and (g) lead to left-skewed distributions for $\delta_j$ with more heterogeneity in (g). Choices (h) and (i) result in distributions for $\delta_j$ that are right-skewed when $\beta_i \geq 0$ and left-skewed otherwise, while for (j) and (k) it is reversed. Choice (i) has more across-set effect heterogeneity than (h), and (k) has more across-set heterogeneity than (j).

To assess performance of these methods as a test for the sample average treatment effect, in the $m$th iteration we compute the sample average treatment effect for the $m$th finite population, $\bar{y}(m)$. We then construct the two vectors of unmeasured confounders resulting in the worst-case expectations for $K^{(c_0)}_\Gamma$ and $D^{(c_0)}_\Gamma$ when (1) holds at $\Gamma = 5$ and with $\tau_0 = \bar{y}(m)$. For each test, we generate a single treatment assignment vector through (2) using the worst-case vector of unmeasured confounders for that test with $\Gamma = 5$. We finally conduct the sensitivity analysis at $\Gamma$ for the null hypothesis $\tilde{r} = \bar{y}(m)$. Alternatively, the simulation study could have constructed a single finite population
for each setting (a)–(j), and could have used the resulting sample average treatment effect for inference for all iterations within each setting. Generating a new finite population in each iteration facilitates reproducibility and makes the results less susceptible to the idiosyncrasies of any particular random draw.

### 6.2 Results for testing the sample average treatment effect

Table 2 contains the results of the simulation study. For each setting, we report the estimated Type I error rate. We further include the expected value of the difference between the test statistic employed and its candidate worst-case expectation, scaled by the test statistic’s standard deviation across simulations. The first set of columns summarize the findings for \( D_\Gamma \). As Theorem 1 predicts, under the sharp null the size was controlled at 0.10. With heterogeneous effects, we see that the procedure was conservative for all settings except (h). That the method failed for simulation (h) further highlights Theorem 3: because the procedure is sharp under constant effects, there must be situations where the method is invalid under the weak null. Settings (h) was constructed precisely such that the sufficient condition (4) failed, with the skewness of residuals changing as a function of the treatment effects. The second set of columns show that while \( K_\Gamma^{(\tau_0)} \) was valid in all settings as predicted by Theorem 2, the procedure was extremely conservative. This conservativeness was present even in setting (a) where the sharp null holds, reflecting the consequences of Theorem 3.

A pattern observed throughout the simulation study is that all candidate modes of sensitivity analysis rejected the null less frequently in the presence of strong across-set effect heterogeneity (captured by the variance of \( \hat{\beta}_i \)) than in the presence of mild across-set heterogeneity. This can be understood through (4). When effects are severely heterogeneous across matched sets but \( \tau = \tau_0 \), the overall average of the treatment effects \( \tau_0 \) may be quite far from the stratum-specific treatment effect \( \tau_i \), encouraging a more negative difference between \( E_{\alpha}(|\bar{\tau}_i - \tau_i| \mid F, \mathcal{Z}) \) and \( E_{\alpha}(|\bar{\tau} - \tau| \mid F, \mathcal{Z}) \). Extreme effect modification promotes conservativeness for the method \( \phi^{(\tau_0)}(\alpha, \Gamma) \), and perhaps counterintuitively the method is more likely to be anticonservative when across-set effect modification is present but limited.

### 6.3 Testing equality of marginal distributions under a superpopulation model

Under a superpopulation framework, an alternative notion of “no treatment effect” for which inference may be desired is that of equality of marginal distributions for the potential outcomes under treatment and control for each set \( i \),

\[
H_0 : F_{Ti}(\cdot) = F_{Ci}(\cdot) \quad (i = 1, \ldots, B).
\]

While the theoretical results in this manuscript have not targeted this definition of no effect, focusing instead on...
inference in the finite population model for the sample average treatment effect, through simulation we assess the performance of the tests presented herein under this null. In the \( m \)th of \( M = 5000 \) iterations, generate stratum sizes \( n_i \) \( i.i.d. \) \( \sim 2 + \text{Poisson}(2) \) for \( i = 1, \ldots, 500 \). For each \( i \), let \( (r_{ij}, \varepsilon_{ij}) \) be \( i.i.d. \) multivariate normal with common variance \( \sigma^2_i \) and correlation \( \rho \) for all \( j = 1, \ldots, n_i \). In the first group of simulations, called equal distributions, we set \( \sigma^2_i = 1 \) for all \( i \). In the second group, called varying distributions, in each iteration we instead draw \( \sigma^2_i \) \( i.i.d. \Gamma(\alpha = 2, \beta = 2) \) with \( \rho = -0.5, 0, 0.5 \), respectively. In settings (iv)--(vi), we set \( r_{ij}, r_{ij} = (\exp(\varepsilon_{ij}), \exp(\varepsilon_{ij})) \) with \( \rho = -0.5, 0, 0.5 \), respectively, and in settings (vii)--(ix) we set \( r_{ij}, r_{ij} = -(\exp(\varepsilon_{ij}), \exp(\varepsilon_{ij})) \) with \( \rho = -0.5, 0, 0.5 \), respectively. After drawing the potential outcomes, the worst-case allocation of hidden bias for inference using \( \psi^{(0)}(\alpha, \Gamma) \) and \( \hat{\phi}^{(0)}(\alpha, \Gamma) \) are computed. For each test, we generate a single treatment assignment vector through (2) using the worst-case vector of unmeasured confounders for that test. We finally conduct the sensitivity analysis at \( \Gamma \) with \( \tau_0 = 0 \).

The results are shown in Table 3. We see that in each of the simulation settings considered, the average bias (in units of the statistic’s standard deviation) of both \( \hat{\psi}^{(0)}(\alpha, \Gamma) \) and \( \hat{\phi}^{(0)}(\alpha, \Gamma) \) fall below zero, indicating that the reference distributions deployed by both \( \psi^{(0)}(\alpha, \Gamma) \) and \( \phi^{(0)}(\alpha, \Gamma) \) have expectations that upper bound the actual worst-case expectation under the null of equality of marginal distributions between potential outcomes under treatment and control. This translates into asymptotic Type I error control below the desired level, \( \alpha = 0.1 \), as reflected in Table 3. Once again, we see that \( \phi^{(0)} \) delivers less conservative inference than \( \psi^{(0)} \).

### 7 | DATA EXAMPLE: SMOKING AND LEAD

Using data from the 2007–2008 National Health and Nutrition Examination Survey (NHANES), Rosenbaum (2013) constructed two matched data sets comparing lead levels (\( \mu g/dl \)) in the blood for daily smokers and non-smokers. The first matched comparison deployed optimal pair matching and contains 250 smoker–nonsmoker pairs, while the second comparison deployed fixed ratio matching at a ratio of 1:5 and consists of 150 matched set, each containing one smoker and five matched nonsmokers. Individuals were matched on gender, age, race, education level, and household income level; see Rosenbaum (2013) for additional details.

Table 4 contains the results of the sensitivity analyses performed at \( \alpha = 0.05 \) on both data sets using the \( \hat{\phi}^{(0)} \) and \( \phi^{(0)} \) for various values of \( \Gamma \). All tests are conducted using a normal approximation. As can be seen, for the paired data set \( \phi^{(0)} \) and \( \phi^{(0)} \) yield identical conclusions for all values of \( \Gamma \). This is because when \( n_i = 2, \tau_{ni} = \Gamma + 1 \), and \( \Gamma_{ni} = \Gamma \) in \( (6) \), such that \( \hat{\psi}_{\Gamma}^{(0)} \) and \( \hat{\psi}_{\Gamma}^{(0)} \) are equal up to a constant multiplier. By Theorem 1 in concert with having a paired design, \( \hat{\phi}^{(0)} = \hat{\phi}^{(0)} \) yields inference that is generally asymptotically conservative, but limiting size equal to \( \alpha \) is possible; see also Theorem 1 of Fogarty (2020). The largest values of \( \Gamma \) for which the tests rejected the null hypothesis was 2.01 for \( \phi^{(0)} = \phi^{(0)} \).

For the 1:5 match, \( \phi^{(0)} \) and \( \phi^{(0)} \) are no longer equivalent beyond \( \Gamma = 1 \). We see in Table 4 that this divergence can have a large impact on the performance of the sensitivity analysis, as the worst-case \( p \)-values for \( \phi^{(0)} \) are markedly larger than those of \( \phi^{(0)} \). This is also reflected in the corresponding changepoint \( \Gamma \) values for inference at
TABLE 4 Worst-case p-values for smoking and lead. For each matched data set, sensitivity analyses are conducted using \( \varphi^{(0)} \) and \( \tilde{\varphi}^{(0)} \) for different values of \( \Gamma \). For each method, candidate worst-case p-values are reported.

| \( \Gamma \) | 1:1 matching, \( B = 250 \) | 1:5 matching, \( B = 150 \) |
|-------|-----------------|-----------------|
|       | \( \varphi^{(0)} \) | \( \tilde{\varphi}^{(0)} \) | \( \varphi^{(0)} \) | \( \tilde{\varphi}^{(0)} \) |
| 1.00  | 0.0000          | 0.0000          | 0.0002          | 0.0002 |
| 1.25  | 0.0002          | 0.0002          | 0.0134          | 0.0370 |
| 1.50  | 0.0023          | 0.0023          | 0.0460          | 0.1998 |
| 1.75  | 0.0131          | 0.0131          | 0.1136          | 0.4919 |
| 2.00  | 0.0478          | 0.0478          | 0.2178          | 0.7467 |

\( \alpha = 0.05 \), which are 1.29 and 1.52 for \( \varphi^{(0)} \) and \( \tilde{\varphi}^{(0)} \), respectively. While \( \tilde{\varphi}^{(0)} \) provides asymptotically valid inference over the entirety of the weak null, it does so at a steep price in terms of reported insensitivity to hidden bias. The procedure \( \varphi^{(0)} \) is not valid over all elements of the weak null, but the discussion in Section 3 and simulations in Section 6 suggest that the elements of the weak null and patterns of hidden bias giving rise to anticonservative sensitivity analyses through \( \varphi^{(0)} \) may be of limited practical concern. Only setting (h) in the simulation study produced anticonservative inference, whereas those settings reflecting more conventional generative models failed to break \( \varphi^{(0)} \).

8 DISCUSSION AND RECOMMENDATIONS

For pair matching, the choice of sensitivity analysis for testing the weak null is straightforward: \( K_{\Gamma}^{(r_0)} \) and \( D_{\Gamma}^{(r_0)} \) are equivalent, and the studentized sensitivity analysis described in detail in Fogarty (2020) using these statistics provides a sensitivity analysis for the weak null that remains sharp if effects are constant. In more flexible matched designs, the researcher must instead weigh the benefits and downsides of a few competing methods. In the web-based Supporting Information, we investigate sensitivity analysis using the permutational t-statistic, \( \tilde{\tau} - \tau_0 \), and show that existing methods for sensitivity analysis derived for the sharp null can be anticonservative when erroneously used as tests for the weak null.

If bounding the worst-case expectation over the entirety of the weak null and for all patterns of hidden bias is deemed essential and nonnegotiable, the sensitivity analysis using \( K_{\Gamma}^{(r_0)} \) and \( \varphi^{(r_0)}(\alpha, \Gamma) \), is appropriate. The sensitivity analysis can be very conservative under reasonable data-generating processes such as those described in Section 3.2 or used in Section 6, and can yield markedly lower reported insensitivities to hidden biases as the data example in Section 7 highlighted. The statistic \( D_{\Gamma}^{(r_0)} \) employs an upper bound on the worst-case expectation that is tight under constant effects. The discussion in Section 3.2 gives a sense of what must go wrong in order for \( D_{\Gamma}^{(r_0)} \) to fail under \( H_N^{(r_0)} \).

Those considerations were exploited in simulation setting (h), but it may be difficult to imagine a pattern of hidden bias affecting an observational study in this way being of scientific interest. That the necessary conditions for failure do not hold almost surely under the generative model described in Section 3.2 with adversarial patterns of hidden bias should provide additional comfort in proceeding with a sensitivity analysis based upon \( D_{\Gamma}^{(r_0)} \). Given the nature of the effect heterogeneity which invalidates \( \varphi^{(r_0)}(\alpha, \Gamma) \) and the fact that such patterns cannot arise under many reasonable data-generating models, our view is that the benefits of proceeding using \( \varphi^{(r_0)}(\alpha, \Gamma) \) will generally outweigh the risk of anticonservativeness, and we would recommend the use of \( \varphi^{(r_0)}(\alpha, \Gamma) \) for sensitivity analyses for the sample average treatment effect while assuming Condition 1. Through actively considering performance under constant effects, a markedly more powerful sensitivity analysis can be constructed in exchange for sacrificing validity over a seemingly peculiar collection of heterogeneous effects.

ACKNOWLEDGMENT

The author thanks the editor, the associate editor, and the referee for their constructive comments and suggestions, which substantially improved the quality of the manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this paper are openly available in CRAN within the package sensitivitymv, available at https://cran.r-project.org/web/packages/sensitivitymv/index.html.

ORCID

Colin B. Fogarty @ https://orcid.org/0000-0002-5520-4639

REFERENCES

Bai, Y., Shaikh, A. & Romano, J.P. (2021) Inference in experiments with matched pairs. Journal of the American Statistical Association. Available from: https://doi.org/10.1080/01621459.2021.1883437.
Chung, E. & Romano, J.P. (2013) Exact and asymptotically robust permutation tests. *Annals of Statistics*, 41(2), 484–507.

Cohen, P.L. & Fogarty, C.B. (2022) Gaussian prepivoting for finite population causal inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2), 295–320.

Cornfield, J., Haenszel, W., Hammond, E.C., Lilienfeld, A.M., Shimkin, M.B. & Wynder, E.L. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22, 173–203.

Ding, P. (2017) A paradox from randomization-based causal inference. *Statistical Science*, 32(3), 331–345.

Egleston, B.L., Scharfstein, D.O. & MacKenzie, E. (2009) On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics*, 65(2), 497–504.

Fogarty, C.B. (2018) On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 1035–1056.

Fogarty, C.B. (2020) Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *Journal of the American Statistical Association*, 115(531), 1518–1530.

Franks, A.M., D’Amour, A. & Feller, A. (2020) Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 115(532), 1730–1746.

Gastwirth, J.L., Krieger, A.M. & Rosenbaum, P.R. (2000) Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3), 545–555.

Hansen, B.B. (2004) Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467), 609–618.

Hosman, C.A., Hansen, B.B. & Holland, P.W. (2010) The sensitivity of linear regression coefficients’ confidence limits to the omission of a confounder. *Annals of Applied Statistics*, 4(2), 849–870.

Imbens, G.W. (2003) Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2), 126–132.

Janssen, A. (1997) Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Statistics and Probability Letters*, 36(1), 9–21.

Kang, H., Kreuels, B., May, J. & Small, D.S. (2016) Full matching approach to instrumental variables estimation with application to the effect of malaria on stunting. *Annals of Applied Statistics*, 10(1), 335–364.

Liu, W., Kuramoto, S.J. & Stuart, E.A. (2013) An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science*, 14(6), 570–580.

Loh, W.W., Richardson, T.S. & Robins, J.M. (2017) An apparent paradox explained. *Statistical Science*, 32(3), 356–361.

Neyman, J. (1935) Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, 2(2), 107–180.

Pashley, N.E. & Miratrix, L.W. (2021) Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*, 46(3), 271–296.

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

Rosenbaum, P.R. (1991) A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3), 597–610.

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

Rosenbaum, P.R. (2013) Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics*, 69(1), 118–127.

Rosenbaum, P.R. (2020) Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, 7(1), 143–176.

Sabbaghi, A. & Rubin, D.B. (2014) Comments on the Neyman-Fisher controversy and its consequences. *Statistical Science*, 29(2), 267–284.

Stuart, E.A. (2010) Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25(1), 1–21.

Tan, Z. (2010) Boundend, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3), 661–682.

VanderWeele, T.J. & Ding, P. (2017) Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 167(4), 268–274.

Wu, J. & Ding, P. (2020) Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, 1–16.

Yu, B. & Gastwirth, J.L. (2005) Sensitivity analysis for trend tests: application to the risk of radiation exposure. *Biostatistics*, 6(2), 201–209.

Zhao, Q., Small, D.S. & Bhattacharya, B.B. (2019) Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4), 735–761.

**SUPPORTING INFORMATION**

Web Appendices, Tables, and Figures referenced in §§3.1, 3.2, 4.1, 5.1 and 8 are available with this paper at the Biometrics website on Wiley Online Library. Also available are R scripts to reproduce the simulations in §6 and the data analysis in §7.

**How to cite this article:** Fogarty, C.B. (2023) Testing weak nulls in matched observational studies. *Biometrics*, 79, 2196–2207. [https://doi.org/10.1111/biom.13741](https://doi.org/10.1111/biom.13741)