INFORMS Journal on Data Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Robust Approach to Quantifying Uncertainty in Matching Problems of Causal Inference

Marco Morucci, Md. Noor-E-Alam, Cynthia Rudin

To cite this article:
Marco Morucci, Md. Noor-E-Alam, Cynthia Rudin (2022) A Robust Approach to Quantifying Uncertainty in Matching Problems of Causal Inference. INFORMS Journal on Data Science 1(2):156-171. https://doi.org/10.1287/ijds.2022.0020

https://pubsonline.informs.org/journal/ijds

INFORMS JOURNAL ON DATA SCIENCE
Vol. 1, No. 2, October–December 2022, pp. 156–171
ISSN 2694-4022 (print), ISSN 2694-4030 (online)

A Robust Approach to Quantifying Uncertainty in Matching Problems of Causal Inference

Marco Morucci,[a] Md. Noor-E-Alam,[b,*] Cynthia Rudin[c]

[a] Center for Data Science, New York University, New York 10012; [b] Department of Mechanical and Industrial Engineering, Northeastern University, Boston, Massachusetts 02115; [c] Department of Computer Science, Duke University, Durham, North Carolina 27708
*Corresponding author

Contact: marco.morucci@duke.edu (MM); mnalam@neu.edu, https://orcid.org/0000-0001-5353-9710 (MN-E-A); cynthia@cs.duke.edu, https://orcid.org/0000-0003-4283-2780 (CR)

Received: January 25, 2022
Revised: August 21, 2022
Accepted: August 30, 2022
Published Online in Articles in Advance: October 27, 2022

https://doi.org/10.1287/ijds.2022.0020

Abstract. Unquantified sources of uncertainty in observational causal analyses can break the integrity of the results. One would never want another analyst to repeat a calculation with the same data set, using a seemingly identical procedure, only to find a different conclusion. However, as we show in this work, there is a typical source of uncertainty that is essentially never considered in observational causal studies: the choice of match assignment for matched groups—that is, which unit is matched to which other unit before a hypothesis test is conducted. The choice of match assignment is anything but innocuous and can have a surprisingly large influence on the causal conclusions. Given that a vast number of causal inference studies test hypotheses on treatment effects after treatment cases are matched with similar control cases, we should find a way to quantify how much this extra source of uncertainty impacts results. What we would really like to be able to report is that *no matter* which match assignment is made, as long as the match is sufficiently good, then the hypothesis test results are still informative. In this paper, we provide methodology based on discrete optimization to create robust tests that explicitly account for this possibility. We formulate robust tests for binary and continuous data based on common test statistics as integer linear programs solvable with common methodologies. We study the finite-sample behavior of our test statistic in the discrete-data case. We apply our methods to simulated and real-world data sets and show that they can produce useful results in practical applied settings.

History: Galit Shmueli served as the senior editor for this article.
Funding: Financial support from the Natural Sciences and Engineering Research Council of Canada and the National Science Foundation [Grant IIS 2147061] is gratefully acknowledged.
Data Ethics & Reproducibility Note: No data ethics considerations are foreseen related to this paper. Code and data to reproduce all the results in this paper are avilable at https://github.com/marcomorucci/robust-tests and in the e-Companion to this article (available at https://doi.org/10.1287/ijds.2022.0020).

Keywords: matching • hypothesis testing • robust optimization • causal inference

1. Introduction

We have a reproducibility crisis in science. Part of the reason for the crisis is sources of uncertainty in the analysis pipeline that are not accounted for. Observational causal studies have serious problems with reproducibility, despite the fact that these studies underlie important policy decisions. In our paper, we focus on a source of uncertainty that can have a large impact on the result, but is almost never mentioned: match assignments of treatment units to control units.

Classically, assignments of treatment and control units to matches are constructed by using a single method aimed at constructing pairs that achieve a good level of some measure of quality predefined by the analyst (e.g., Optimal Matching (Rosenbaum 1989) or Genetic Matching (Diamond and Sekhon 2013)). Choosing a single fixed match ignores a major source of uncertainty, which is the design itself, or, in other words, the uncertainty related to the choice of experimenter. What if there were two possible equally good match assignments, one where the treatment-effect estimate is very strong and one where it is nonexistent? When we report a result on a particular matching assignment, we thus ignore the possibility of the opposite result occurring on an equally good assignment. It is entirely possible that two separate researchers studying the same effect on the same data, using two different, equally good sets of matched groups, would get results that disagree. When researchers follow the classic pipeline of match-then-estimate, their hypothesis tests are conditioned on the match assignment; in other words, only uncertainty *after* matching is considered, and not the uncertainty stemming from the match assignment itself.

Our goal is to create robust matched-pairs hypothesis tests for observational data. These tests implicitly consider *all possible reasonably good assignments* and consider the *range of possible outcomes* for tests on these data. This is a more computationally demanding approach to hypothesis testing than the standard approach, where one considers just a single assignment, but the result is then more robust to the choice of experimenter who chooses the match assignment. It is computationally infeasible (and perhaps not very enlightening) to explicitly compute all possible assignments, but it is possible to look at the range of outcomes associated with them. In particular, our algorithms compute the maximum and minimum of quantities like the test statistic values and their associated p-values, in accordance with the principles of robust optimization. This is a type of *hacking interval* that includes the matching process itself as part of the randomness in the problem (Coker et al. 2021).

After motivation and formalization for our framework in Sections 1.1 and 2, we offer a robust formulation for McNemar's statistic for binary outcomes in Section 3. In Section 4, we formulate an integer linear program (ILP) to compute the robust version of the canonical z-test in a general case. We then study the finite-sample distribution of our robust version of McNemar's test that includes the uncertainty stemming from the match assignment. Finally, we present evidence of the performance of our methods in Section 6 using simulated data where the ground truth is known, and two real-data case studies are discussed in Section 7. We know of no previous approaches to matching in observational causal inference that use robust optimization to handle uncertainty in the match assignment.

1.1. Matching Analyses and Their Sensitivity to Arbitrary Choices

Existing studies that use matching for hypothesis testing all roughly follow the following template: (step 1) Choose a test statistic; (step 2) define criteria that acceptable matches should satisfy (e.g., covariate balance); (step 3) use an algorithm to find matched groups (and corresponding subsamples) that satisfy the criteria; (step 4) implicitly choose the subsample of interest to be defined by the matched data; and (step 5) compute the test statistic and its p-value on the matched data (see, e.g., Rosenbaum 2010).

This procedure does not explicitly incorporate uncertainty arising from steps 3 and 4: Values of the test statistic computed under different, but equally good, matched subsamples could differ. These matches are indistinguishable in terms of quality. Which of the matched sets should be taken as representative of the whole sample? There is no clear answer to this question.

Even a trivial source of randomness, such as the order in which the data appear in the data set, could produce two equivalently good match assignments under a variety of popular matching methods, such as nearest-neighbors. If we apply the same test statistic to each of the two matched samples, we would be implicitly testing two *different* hypotheses: The first one is defined on the subset of the data selected by the first match assignment, and the second hypothesis on the second. Under the assumption that both matches are equally as good, common statistical tests will yield valid results for each of these two hypotheses, yet the hypothesis test conducted on the first matched sample could reject, whereas the test conducted on the second could not. Because both assignments achieve similar quality, there is no clear way to choose one result over the other, and arbitrarily reporting one of the two would hide the fact that the other exists. Such uncertainty would likely be of interest to policy makers and researchers, who would want their results to be reproducible under all arbitrary choices of analyst.

There is no easy fix. Assigning multiple control units to each treatment unit, or using deterministic matching methods that yield a unique result for each data set, ignores the possibility of a slightly different—but still good—match assignment giving opposite results. Asymptotics (e.g., Abadie and Imbens 2011) provide no remedy, because asymptotic results are usually intended to apply to all good matching algorithms asymptotically; the problem we encounter in practice occurs because we always work with finite samples. The problem extends not only to treatment effects, but also to variance estimates and beyond.

1.2. Empirical Evidence of the Problem

The problem we identified is serious in practice. Morucci and Rudin (2020) replicate several social science studies—all published in top journals after 2010—that use matching. They perform the same hypothesis tests after applying several popular matching algorithms. Table 1 reports agreement between any two of six different matching methods on the same hypothesis test—that is, the percentage of the tests replicated that have the same result (reject/fail to reject) under each pair of methods. If our hypothesis were false, and matching at the same level of quality always produced the same results, then we would see high agreement in the table. This is clearly not the case. Additionally, Table 1 also reports a measure of similarity of the balance between the data sets produced by each pair of matching methods. We measure balance as the proportion of covariates whose difference in means between the matched treated and control group is statistically insignificant at the 5% level. The numbers in b of each a/b entry of Table 1 represent the correlation between indicator vectors that keep track of which covariates were balanced by each method according to the criterion just presented. Clearly, although imperfectly, methods tend to "agree" on balance more than they do on rejection: Agreement on

Table 1. Agreement Between Rejection Decisions on the Same Data Between Different Matching Methods

	Rejection agreement/balance agreement					
Method	CEM	Genetic	Optimal	Nearest	Subclass	Optimal
CEM	—	0.15/0.82	0.40/0.68	0.60/0.71	0.41/0.74	0.23/0.62
Genetic		—	0.27/0.63	0.51/0.68	0.68/0.63	0.21/0.83
Optimal			—	0.42/0.99	0.44/0.94	0.40/0.75
Nearest				—	0.64/0.95	0.31/0.78
Subclass					—	0.62/0.83
Optimal						—

Notes. In each a/b entry, the a is the agreement on rejecting the null between methods, and b measures how often covariate balance agrees between the two methods. The bottom rows show some summary statistics of the replication experiment. Results in this table are a summary of results from Morucci and Rudin (2020). This table shows that a seemingly arbitrary choice of matching method can heavily impact the results. Number of hypotheses tested: 68 times six methods to test each hypothesis. Percent of tests with at least one rejection and one nonrejection: 64.7. Percent of tests with at least one positive rejection and one negative rejection: 5.9. Percent agreement between two methods, on average: 40.5.

balance is always greater than agreement on rejection, sometimes by a large margin. This implies that differences in balance produced are not, at least in full, to blame for the problem of different matching methods leading to different conclusions: Even when methods produce similar balance, they still disagree.

On average, two methods alone agree on rejecting versus not rejecting the same null hypothesis *only 40.5% of times*, but achieve similar balance 70.5% of times. These numbers drop dramatically if any group of three methods is considered and are virtually zero if agreement among all six methods at once is considered. As stated before, we believe that this happens because different matching methods choose among matches of similar quality arbitrarily and in different ways. According to these results, seemingly arbitrary choices of match assignment affect conclusions. This is true in practical scenarios that involve real-world data, and addressing it is an important step toward fully robust data analysis.

2. Proposed Approach

Throughout this paper we adopt the potential outcomes framework (see Rubin 1974 and Holland 1986). For each unit $i = 1, \ldots, N$, we have potential outcomes $Y_i(t) \in \mathbb{R}$, where $t \in \{0, 1\}$. As is standard in causal inference settings, there are N^t units that receive the treatment and N^c units that receive the control condition, with $N^c + N^t = N$. We denote the condition administered to each unit with $T_i \in \{0, 1\}$. We never observe realizations of $Y_i(1)$ and $Y_i(0)$ for each unit at the same time, but only of $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$. We also observe a P-dimensional vector of covariates X_i taking value in some set $\mathcal{C}$ for every unit. We will denote the complete set of random variables representing the data with $\mathcal{D} = \{X_i, T_i, Y_i\}_{i=1}^N$, with each element being an independent draw from some probability distribution. Analogously, we will denote the set of observed values for the data with the lowercase equivalents of the notation above: $\{x_i, t_i, y_i\}_{i=1}^N$. We assume that potential outcomes are related to the covariates as follows: $Y_i(t) = \mu_t(x_i) + \nu_i$, with $\mathbb{E}[\nu_i \mid X_i = x_i] = 0$. To conduct our hypothesis test, we will also notate our observations as: $(x_1^t, y_1^t), \ldots, (x_{N^t}^t, y_{N^t}^t)$ and $(x_1^c, y_1^c), \ldots, (x_{N^c}^c, y_{N^c}^c)$, where x_i^t, y_i^t and x_i^c, y_i^c represent observed values of X and Y for the ith treated and control units, respectively. We make the classical assumptions of Conditional Ignorability of treatment assignment and Stable Unit Treatment Value (SUTVA):

Assumption 1 (Strong Conditional Ignorability). *For any unit in the sample $i = 1, \ldots, N$ treatment allocation is independent of potential outcomes conditional on the observed covariates—that is, $T_i \perp\!\!\!\perp (Y_i(1), Y_i(0)) \mid X_i$.*

Assumption 2 (SUTVA). *A unit's potential outcomes depend only on that unit's treatment level—that is, for all units, i: $Y_i(t_1, \ldots, t_N) = Y_i(t_i)$, where $Y_i(t_1, \ldots, t_N)$ is the potential outcome for unit i under all units' treatment assignments.*

Under these two assumption, the Sample Average Treatment Effect (SATE) is our quantity of interest and is defined as:

$$\tau = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{Y_i \mid X_i = x_i}[Y_i(1) - Y_i(0) \mid X_i = x_i]. \tag{1}$$

Under SUTVA and Conditional Ignorability, τ is identified—it can be consistently estimated with the observed data—and hypothesis tests based on it can be conducted in a variety of ways. In general, we will consider testing a null hypothesis of zero average treatment effect in the observed sample, which is canonically defined as follows:

$$\mathbb{H}_0 : \tau = 0 \mid X_1 = x_1, \ldots, X_N = x_N. \tag{2}$$

Note that the defining feature of the SATE is that the contextual covariates, $\mathbf{X}$, are only considered at the observed values, as implied by the definition of $\mathbb{H}_0$. As is done in much of the causal inference literature, we would like to take advantage of matching in order to test $\mathbb{H}_0$. A matching operator determines which

control is assigned to which treatment. In this paper, we focus on one-to-one matching (though our results could be generalized to many-to-one or many-to-many matching); therefore, we define the matching operator as follows.

Definition 1 (Matching Operator). A matching operator $\mathbf{a}: \{1,\dots,N^t\} \to \{1,\dots,N^c,\emptyset\}$ obeys the following: If $i \neq k$ and $\mathbf{a}(i) \neq \emptyset$, then $\mathbf{a}(i) \neq \mathbf{a}(k)$. That is, no two treatment units i and k are assigned to the same control unit.

We define the size of the matching—that is, the number of matched pairs—to be $M = \sum_{i=1}^{N_t} \mathbb{I}(\mathbf{a}(i) \neq \emptyset)$, with $\mathbb{I}(E)$ representing the indicator function for the event E throughout the paper. The set of all matching operators is $\mathcal{A}$: the set of all possible assignments $\{\mathbf{a}\}$. Throughout the rest of the paper, we also use $\mathbf{a}$ to represent a $N^t \times N^c$ matrix of matches such that $a_{ij} = 1$ if i and j are matched together.

How should we make our matches? Choosing a matched set that leads to the most reliable estimates of τ is a canonical problem in observational causal inference, and, indeed, many good methods to find such an assignment have been proposed in the existing literature. One of the contributions of this paper is to not restrict consideration to a single good assignment, but, instead, to consider a set of potentially many equally good match assignments. We define such a set as $\mathcal{A}_{\text{good}} \subset \mathcal{A}$: the set of match assignments that satisfy some user-defined criterion of quality. Many criteria for defining a good match assignment already exist—for example, moment balance in the covariate distributions of treated and control samples after matching, low aggregate distance between units in covariate space, or similarity in propensity scores among matched units. Ultimately, $\mathcal{A}_{\text{good}}$ should be viewed as the set of match assignments that produce the most reliable estimates of τ with the given data. To represent the fact that one assignment is chosen among the many in $\mathcal{A}_{\text{good}}$, we will use the random variable $\mathbf{A}$, having domain equal to $\mathcal{A}_{\text{good}}$. The probability distribution of $\mathbf{A}$ is determined purely by human analysts' choice of matching algorithm.

Consider, for now, testing $\mathbb{H}_0$ against a left-tailed alternative: $\mathbb{H}_1 : \tau < 0$ with a generic test statistic $\Psi_{\mathcal{D}}(\mathbf{a})$, dependent on both the data, $\mathcal{D}$, and the chosen match assignment, $\mathbf{a}$, and denote the observed value of $\Psi_{\mathcal{D}}(\mathbf{a})$ with $\psi_D(\mathbf{a})$. Virtually all hypothesis tests done on data matched by any of the existing methods require the following assumption, which we also maintain:

Assumption 3 (Test Statistic Distribution). *Under Assumption 1 and Assumption 2, we assume that for any assignment* $\mathbf{a} \in \mathcal{A}_{\text{good}}$: $\Pr(\Psi_{\mathcal{D}}(\mathbf{a}) \leq \psi_{\mathcal{D}}(\mathbf{a}) \,|\, \mathbb{H}_0, \mathbf{A} = \mathbf{a}) \approx F(\psi_{\mathcal{D}}(\mathbf{a}))$, *for some known cumulative density function F that does not depend on* $\mathbf{a}$.

This assumption means that, if the matches are good enough, then the cumulative density function (CDF) of $\Psi_{\mathcal{D}}(\mathbf{a})$ under the null can be at least approximated well by some known function, F. This assumption is implicitly made in all studies that perform a hypothesis test on matched data. It is often justified by appealing to asymptotic arguments (Abadie and Imbens 2011) or by making parametric assumptions about the data (Rubin 2007). For example, if a researcher performs a z-test on matched data, they implicitly assume that F is the CDF of the standard normal distribution.

The key problem we wish to highlight in this paper can be stated as follows: Existing matching procedures wishing to test $\mathbb{H}_0$ under Assumption 3 use $\Pr(\Psi_{\mathcal{D}}(\mathbf{a}) \leq \psi_{\mathcal{D}}(\mathbf{a}) \,|\, \mathbb{H}_0, \mathbf{A} = \mathbf{a})$ to compute the p-value for $\Psi_{\mathcal{D}}(\mathbf{a})$ under $\mathbb{H}_0$; however, this p-value is both conditioned on $\mathbb{H}_0$ *and* the event $\mathbf{A} = \mathbf{a}$—that is, conditional on assignment $\mathbf{a}$ being the one chosen among the many in $\mathcal{A}_{\text{good}}$. The hypothesis being tested here is not $\mathbb{H}_0$, but, instead: $\mathbb{H}_0 \,|\, \mathbf{A} = \mathbf{a}$:

$$\mathbb{H}_0 \,|\, \mathbf{A} = \mathbf{a} : \frac{1}{M} \sum_{i=1}^{N} \mathbb{E}[Y(1) - Y(0) \,|\, X = x_i] \mathbb{I}(\mathbf{a}(i) \neq \emptyset) = 0.$$

This is essentially a version of $\mathbb{H}_0$ restricted to consider only those units that do receive a match under assignment $\mathbf{a}$, and, most importantly, there is no guarantee that we will fail to reject $\mathbb{H}_0 \,|\, \mathbf{A} = \mathbf{a}$ *if* $\mathbb{H}_0$ is true. Thus, although testing $\mathbb{H}_0 \,|\, \mathbf{A} = \mathbf{a}$ rather than $\mathbb{H}_0$ seems not to be what the user ultimately desires, it is the calculation that they typically perform. By overlooking the difference between these two quantities, they would fail to ask questions such as: How should we proceed if we find two assignments, $\mathbf{a}_1 \in \mathcal{A}_{\text{good}}$ and $\mathbf{a}_2 \in \mathcal{A}_{\text{good}}$, that successfully reject $\mathbb{H}_0 \,|\, \mathbf{A} = \mathbf{a}_1$, but fail to reject $\mathbb{H}_0 \,|\, \mathbf{A} = \mathbf{a}_2$ at the same significance level? We now articulate our proposed solution to this problem.

Because we have modeled which assignment is selected among those in $\mathcal{A}_{\text{good}}$ as a random variable, it would be natural, but ultimately impossible, to propose computing p-values in expectation over the entirety of $\mathcal{A}_{\text{good}}$ to handle the issue just introduced. This would be possible under Assumption 3 by computing:

$$\begin{aligned}\sum_{\mathbf{a} \in \mathcal{A}_{\text{good}}} \Pr(\Psi_{\mathcal{D}}(\mathbf{a}) \leq \psi_{\mathcal{D}}(\mathbf{a}) \,|\, \mathbb{H}_0, \mathbf{A} = \mathbf{a}) \Pr(\mathbf{A} = \mathbf{a} \,|\, \mathbb{H}_0) \\ = \sum_{\mathbf{a} \in \mathcal{A}_{\text{good}}} F(\psi_{\mathcal{D}}(\mathbf{a})) \Pr(\mathbf{A} = \mathbf{a} \,|\, \mathbb{H}_0) \\ = \mathbb{E}_{\mathbf{A} \,|\, \mathbb{H}_0}[F(\psi_{\mathcal{D}}(\mathbf{A}))].\end{aligned}$$

This quantity cannot be computed due to the presence of $\Pr(\mathbf{A} = \mathbf{a} \,|\, \mathbb{H}_0)$ in its definition: The probability of observing any match assignment among the ones in $\mathcal{A}_{\text{good}}$ cannot be modeled statistically. This is because this probability depends on factors such as which matching method is chosen by the analyst, which hyperparameter values the method is using, and how the method itself may determine how to choose one among many potentially equivalent matched sets. All these are variables that cannot be

analytically studied in a general context and will depend on the specific analyst, data, and method.

Because directly modeling $\Pr(\mathbf{A}=\mathbf{a}\,|\,\mathbb{H}_0)$ is nonsensical, we propose to instead *bound* this quantity using the assumption that all analysts choose reasonably good match assignments. This assumption leads to the following simple relationship:

$$\mathbb{E}_{\mathbf{A}|\mathbb{H}_0}[F(\psi_{\mathcal{D}}(\mathbf{A}))] = \sum_{\mathbf{a}\in\mathcal{A}_{\text{good}}} F(\psi_{\mathcal{D}}(\mathbf{a}))\Pr(\mathbf{A}=\mathbf{a}\,|\,\mathbb{H}_0) \leq \max_{\mathbf{a}\in\mathcal{A}_{\text{good}}} F(\psi_{\mathcal{D}}(\mathbf{a})).$$

In words, we bound the expected p-value over all good assignments, with the largest p-value found among all good assignments. Bounding approaches to handle situations without clear probability distributions are common in robust optimization and robust statistics.

To further reduce the problem, we note that the CDF of the test statistic, $F(\psi_{\mathcal{D}}(\mathbf{a}))$, is monotonically increasing in $\psi_D(\mathbf{a})$ (e.g., the larger the z-statistic, the larger the CDF of the normal), which implies that $\max_{\mathbf{a}\in\mathcal{A}_{\text{good}}} F(\psi_{\mathcal{D}}(\mathbf{a})) = F(\max_{\mathbf{a}\in\mathcal{A}_{\text{good}}} \psi_{\mathcal{D}}(\mathbf{a}))$. Letting $\psi^+ = \max_{\mathbf{a}\in\mathcal{A}_{\text{good}}} \psi_{\mathcal{D}}(\mathbf{a})$ be the largest observed test statistic obtainable with a good match assignment, we have our final proposed bound for the robust p-value:

$$\mathbb{E}_{\mathbf{A}|\mathbb{H}_0}[F(\psi_{\mathcal{D}}(\mathbf{A}))] \leq F(\psi^+). \tag{3}$$

Clearly, if $\mathbb{H}_0$ is rejected at a chosen level of significance by $F(\psi^+)$, then it will also be rejected under $\mathbb{E}_{\mathbf{A}|\mathbb{H}_0}[F(\psi_{\mathcal{D}}(\mathbf{A}))]$. It is in this sense that our proposed tests are *robust*, as they target minimization of the probability of Type-I error: incorrect rejection of $\mathbb{H}_0$. That is, our tests are less likely to reject the null hypothesis when it is true.

Although the bound in Equation (3) is a robust p-value against a left-sided alternative, the same set of arguments shown above can be applied to derive the following upper bound on the p-value for a right-sided alternative: $\mathbb{H}_1 : \tau > 0$:

$$\mathbb{E}_{\mathbf{A}|\mathbb{H}_0}[1-F(\psi_{\mathcal{D}}(\mathbf{A}))] \leq \max_{\mathbf{a}\in\mathcal{A}_{\text{good}}} (1-F(\psi_{\mathcal{D}}(\mathbf{a}))) = 1-F\left(\min_{\mathbf{a}\in\mathcal{A}_{\text{good}}} \psi_{\mathcal{D}}(\mathbf{a})\right) = 1-F(\psi^-). \tag{4}$$

Finally, it follows from the bounds above that a robust p-value for testing $\mathbb{H}_0$ against a two-sided alternative: $\mathbb{H}_1 : \tau \neq 0$ can be found by applying the canonical definition of p-value to the bounds introduced above:

$$2\min\{\mathbb{E}_{\mathbf{A}|\mathbb{H}_0}[F(\psi_{\mathcal{D}}(\mathbf{A}))], \mathbb{E}_{\mathbf{A}|\mathbb{H}_0}[1-F(\psi_{\mathcal{D}}(\mathbf{A}))]\} \leq 2\min\{F(\psi^+), 1-F(\psi^-)\}. \tag{5}$$

The problem we face is now one of finding, for a given data set, the values of ψ^+ and ψ^-. Fortunately, this issue can be solved by adapting modern Mixed-Integer-Programming (MIP) tools to finding the maximal and minimal matches that satisfy the set of constraints that define $\mathcal{A}_{\text{good}}$. In the rest of the paper, we propose MIP formulations and algorithms for computing these values for two of the most widely used test statistics.

2.1. Motivation for Using a Robust Approach

This robust approach is justified as follows:

2.1.1. Why Modeling $Pr(\mathbf{A}=\mathbf{a}\,|\,\mathbb{H}_0)$ Is Not Possible. First, we do not want to consider how likely a certain assignment $\mathbf{a}$ is to appear. This would involve modeling human behavior of the experimenter, or arbitrary and complex choices of different matching algorithms. We do not want to place a distribution over the choice of algorithms or matches that an experimenter would choose. As Morgan and Winship (2007) note, there is no clear guidance on the choice of matching procedure. We do not presuppose a distribution over these procedures. In reality, most researchers tend to use popular and widely cited matching software packages, and they make this choice independently of the specific problem being studied.

2.1.2. Why the Bounding Approach Is Not Too Extreme. It is inaccurate to assume that our tests might lead to results that are "too" extreme. This is because there is no reason why maxima or minima over $\mathcal{A}_{\text{good}}$ should be considered outlying or extreme, even if most of the matches in $\mathcal{A}_{\text{good}}$ are different than those at the extrema of the set. Note also that considering a result output by one of the extrema as an "outlier" in the sense that it is believed that most other results with good matches should be concentrated elsewhere in the space of $\psi_D(\mathbf{a})$ is circular logic: If a set of assignments is included in $\mathcal{A}_{\text{good}}$, then it must be thought to be good enough, according to the analyst's criteria. Excluding this assignment because it produces values that are too extreme is tantamount to excluding observations from the data set because they are not producing the desired result.

2.1.3. It Is Unlikely That There Is Only One Good Assignment. It would be tempting to say that matches chosen are often the unique maximizers of some metric of quality and that, because of this, there never is uncertainty over choice of matches in practice. This is not true in practice, as researchers often start with a predefined level of quality that matches must meet to be considered acceptable and only report results from the assignment that maximizes that quality criterion if the maximum achieved is above that predefined quality threshold. Clearly, all assignments above that threshold should be considered, and that is what we propose in this work.

2.1.4. Asymptotic Theory Does Not Solve the Problem. Our robust method is also necessary because existing asymptotic arguments used to derive valid p-values

and confidence intervals for matched data (see, e.g., Abadie and Imbens 2006, 2011). As argued previously, the defining feature of a "good" match assignment is that it leads to at least well-approximated p-values for the desired test statistic. In the case of asymptotic approximation, it is assumed that $\Pr(\Psi_{\mathcal{D}}(\mathbf{a}) \leq \psi_{\mathcal{D}}(\mathbf{a}) | \mathbb{H}_0, \mathbf{A} = \mathbf{a}) \to F(\psi_{\mathcal{D}}(\mathbf{a}))$ as $N \to \infty$. This is assumed true for all assignments $\mathbf{a} \in \mathcal{A}_{\text{good}}$: Although it might be the case that the p-values produced under all good assignments will converge to the same as the sample size grows, this is demonstrably not necessarily true in finite samples; hence, the need for a robust procedure.

3. Robust McNemar's Test

We consider the problem of hypothesis testing for binary outcomes—that is, $y(t) \in \{0,1\}$. After matching, treated and control units will be paired together, and because the outcome is binary, there can exist four types of pairs: A is the number of pairs such that $y_t = 0, y_c = 0$; B is the number of pairs such that $y_t = 1, y_c = 0$; C is the number of pairs such that $y_t = 0, y_c = 1$; and D is the number of pairs such that $y_t = 1, y_c = 1$. More formally, let $B(\mathbf{a}) = \sum_{i=1}^{N^t} \sum_{j=1}^{N^c} a_{ij} y_i^t (1 - y_j^c)$ be the count of matched pairs in which the treated unit has outcome one and the control unit has outcome zero under assignment $\mathbf{a}$, and let $C(\mathbf{a}) = \sum_{i=1}^{N^t} \sum_{j=1}^{N^c} a_{ij} y_j^c (1 - y_i^t)$ be the number of matched pairs in which the treated unit has outcome zero and the control unit has outcome one. We use the following test statistic for all null hypotheses:

$$\chi = \frac{B(\mathbf{a}) - C(\mathbf{a}) - 1}{\sqrt{B(\mathbf{a}) + C(\mathbf{a})}} = \frac{\sum_{i=1}^{N^t} \sum_{j=1}^{N^c} a_{ij}(y_i^t - y_j^c) - 1}{\sqrt{\sum_{i=1}^{N^t} \sum_{j=1}^{N^c} a_{ij}(y_i^t + y_j^c - 2y_i^t y_j^c)}}. \quad (6)$$

This is the classical formulation of McNemar's statistic with a continuity correction at the numerator (see Tamhane and Dunlop 2000). Our proposed robust test statistic can be defined as the pair:

$$\chi^+ = \max_{\mathbf{a} \in \mathcal{A}_{\text{good}}} \frac{B(\mathbf{a}) - C(\mathbf{a}) - 1}{\sqrt{B(\mathbf{a}) + C(\mathbf{a})}}, \chi^- = \min_{\mathbf{a} \in \mathcal{A}_{\text{good}}} \frac{B(\mathbf{a}) - C(\mathbf{a}) - 1}{\sqrt{B(\mathbf{a}) + C(\mathbf{a})}}.$$

In what follows, we outline a strategy to compute (χ^+, χ^-), subject to general constraints that define $\mathcal{A}_{\text{good}}$. Theorem 1 in Section 3.2 shows that, in a special case where the constraints on $\mathcal{A}_{\text{good}}$ explicitly define strata for the data, the optimal statistic can be computed in linear time. In Section 5, we provide theoretical statements of the finite-sample behavior of the optimized statistics.

3.1. Optimizing McNemar's Statistic with General Constraints

In this section, we give an ILP formulation that optimizes χ with a predefined number of matches.

One can show that the number of pairs where the same outcome is realized for treatment and control (the number of $A + D$ pairs) is irrelevant, so we allow it to be chosen arbitrarily, with no constraints or variables defining it. The total number of pairs is thus also not relevant for this test. Therefore, we are permitted to choose only the total number of untied responses (B and C pairs), denoted m, which must always be greater than zero for the statistic to exist. The problem can be solved either for a specific m or globally for all $m > 0$ by looping over all possible values of m until the problem becomes infeasible. In most practical scenarios, the largest m for which a solution is feasible is chosen.

Formulation 1 (General ILP Formulation for McNemar's Test).

$$\text{Maximize/Minimize}_{\mathbf{a}} \quad \chi(\mathbf{a}) = \left[\frac{B(\mathbf{a}) - C(\mathbf{a}) - 1}{\sqrt{m}}\right]$$

subject to :

$$\sum_{i=1}^{N^t} \sum_{j=1}^{N^c} a_{ij} y_i^t (1 - y_j^c) = B(\mathbf{a})$$

(Total number of first type of discordant pairs), (7)

$$\sum_{i=1}^{N^t} \sum_{j=1}^{N^c} a_{ij} y_j^c (1 - y_i^t) = C(\mathbf{a})$$

(Total number of second type of discordant pairs), (8)

$$B(\mathbf{a}) + C(\mathbf{a}) = m \quad \text{(Total number of discordant pairs)}, \quad (9)$$

$$\sum_{i=1}^{N^t} a_{ij} \leq 1 \quad \forall j \text{ (Match each control unit at most once)}, \quad (10)$$

$$\sum_{j=1}^{N^c} a_{ij} \leq 1 \quad \forall i \text{ (Match each treatment unit at most once)}, \quad (11)$$

$$a_{ij} \in \{0,1\} \quad \forall\, i,j \text{ (Defines binary variable } a_{ij}), \quad (12)$$

(Additional user-defined match quality constraints.) (13)

Equations (7) and (8) are used to define variables B and C. To control the total number of untied responses, we incorporate Equation (9). Equations (10) and (11) confirm that only one treated/control unit will be assigned in a single pair. As it is common practice in matching, it is possible to test a null hypothesis for the Sample Average Treatment Effect on the Treated by adding the constraint $\sum_{i=1}^{N^t} \sum_C a_{ij} = N^t$, only in the case where $N^t \leq N^c$, so that all treatment units are matched.

3.2. Fast Optimization of McNemar's Test Under Special Constraints

The general optimization problem stated in Formulation 1 can also be solved quickly when constraints have the following property: There exists a stratification of the data such that an assignment is a feasible solution if and only if it matches units exclusively within the same stratum. This is stated formally below:

Definition 2 (Exclusively Binning Constraints). The constraints on $\mathcal{A}_{\text{good}}$ are exclusively binning if there exists a partition $\mathcal{S}$ of $\{1,\ldots,N\}$ such that for all $\mathbf{a} \in \mathcal{A}$, we have $\mathbf{a} \in \mathcal{A}_{\text{good}}$ if and only if: $\forall i \in \{1\ldots,N\}$, $i \in S$, and $\mathbf{a}(i) \in S$ for some subset $S \in \mathcal{S}$.

This type of grouping is commonly referred to as blocking (Imai et al. 2008), and there already exist matching methods that openly adopt it as a strategy to construct good-quality matches (Iacus et al. 2012). Blocking can also occur naturally in data; there need only exist natural subcommunities. In practice, several types of constraints on the quality of matches, particularly balance constraints, as defined in Equation (31) in the online appendix, can be implemented by making coarsening choices on the covariates (Iacus et al. 2011), and, as such, these coarsening choices are exclusively binning constraints. Under exclusively binning constraints, solving the optimization problem to find (χ^+, χ^-) becomes much simpler, as shown in the following theorem:

Theorem 1. *Let the constraints on $\mathcal{A}_{\text{good}}$ be Exclusively Binning Constraints. Then, either the max or the min optimization problem in Formulation 1 can be solved in linear time, in the number of units N.*

The proof of the theorem can be found in the online appendix, where we explicitly state these linear-time algorithms, together with a proof of their correctness. These algorithms are not difficult, though stating them and providing the proof requires several pages. The algorithms work by first determining the sign of the optimal test statistic, and then by matching units in each stratum separately to optimize the test statistic once the sign is known. Maximizing (minimizing) the test statistic in each stratum is then accomplished by matching as many treated (control) units as possible with outcome one to control (treated) units with outcome zero. This procedure is intuitively linear in N, the total number of units. Figure 1 gives a simple example of how our linear-time algorithms optimize McNemar's test in each stratum.

4. Robust *z*-Test

In this section we consider the canonical z-test for estimating whether the difference in mean of the treatment and control populations is sufficiently greater than zero, when outcomes are real-valued—that is, $y(t) \in \mathbb{R}$. Again, we will compute extreme values of z that could occur from the set of feasible match assignments.

Again, after computing the extreme values of the z-statistics, p-values for them can be obtained with the canonical Normal asymptotic approximation. At that point, we can determine whether the hypothesis test result is robust to the choice of match assignment.

In what follows, M is the total number of pairs, and $\hat{\sigma}$ is the sample standard deviation of the differences, $y_i^t - y^c_{\mathbf{a}(i)}$. The z-score is:

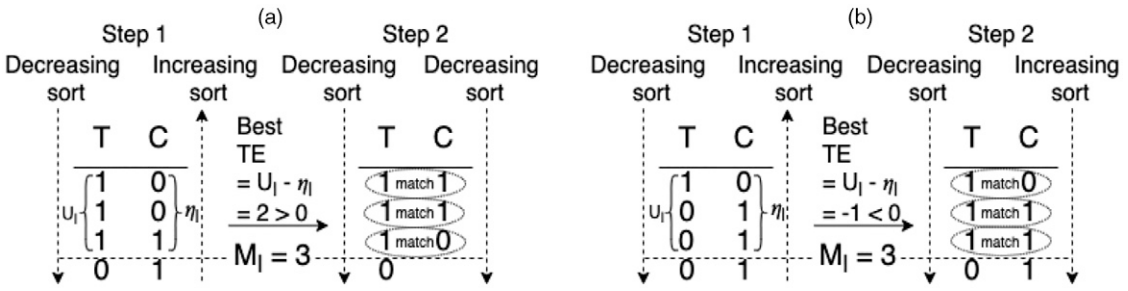
$$z_{\mathbf{y}}(\mathbf{a}) = \frac{\bar{d}_{\mathbf{a}}\sqrt{M}}{\hat{\sigma}_{\mathbf{a}}}, \quad \text{where :}$$
$$\bar{d}_{\mathbf{a}} = \frac{1}{M}\sum_{i=1}^{N^t}\sum_{j=1}^{N^c} a_{ij}(y_i^t - y_j^c), \quad \text{and}$$
$$\hat{\sigma}_{\mathbf{a}} = \sqrt{\frac{1}{M}\sum_{i=1}^{N^t}\sum_{j=1}^{N^c} a_{ij}(y_i^t - y_j^c)^2 - \bar{d}_{\mathbf{a}}^2}. \tag{14}$$

Our robust statistic in this case is defined as the pair:

$$z^+ := \max_{\mathbf{a}\in\mathcal{A}_{\text{good}}} \frac{\bar{d}_{\mathbf{a}}\sqrt{M}}{\hat{\sigma}_{\mathbf{a}}}, z^- := \min_{\mathbf{a}\in\mathcal{A}_{\text{good}}} \frac{\bar{d}_{\mathbf{a}}\sqrt{M}}{\hat{\sigma}_{\mathbf{a}}}. \tag{15}$$

Below, we provide general ILP formulations for computing the robust statistic under (balance and other types of) constraints on $\mathcal{A}_{\text{good}}$ by devising a linearized formulation of the z-statistic optimization problem that allows it to be solved with any ILP solver. ILP formulations that are slightly different from each other

Figure 1. Example of Linear-Time Optimization of McNemar's Test in One Stratum, in Which Exactly M_l Matches Are Made



Notes. (a) Max $\chi > 0$. (a) Max $\chi < 0$. Numbers "0" and "1" in the figure are unit outcome values. In each stratum l, U_l denotes the sum of treated outcomes matched, η_l is the sum of matched control outcomes, and M_l is the number of matches made.

(that we will discuss) can handle testing of $\mathbb{H}_0$ for the Average Treatment Effect and Average Treatment Effect on the Treated.

4.1. Computing (z^+, z^-) Under General Constraints

The z statistic is clearly not linear in the decision variables (the match assignments). If one were to optimize it directly, a solution could be approximated by using a mixed-integer nonlinear programming solver (MINLP), but guarantees on the optimality of the solution might take an incredibly long time. In what follows, we show how this problem can be simplified to be solved by an algorithm that solves several linear integer programming problems instead. This algorithm benefits from the computational speed of ILP solvers, compared with MINLP solvers, and has a guarantee on the optimality of the solution.

To create the ILP formulation, we note that the objective is increasing in the average of the differences (this term appears both in the numerator and denominator), and it is decreasing in the sum of the squared differences (this term is the first term of $\hat{\sigma}$). We then replace the nonlinear objective in (15), expanded in (14), as follows:

$$\text{Maximize/Minimize}_{\mathbf{a}} \quad \sum_{i=1}^{N^t}\sum_{j=1}^{N^c}(y_i^t - y_j^c)a_{ij}, \tag{16}$$

which is now linear in the decision variables. The quantity in (16) is the estimated treatment effect. Simultaneously, we will limit the sum of squared differences term in the denominator by b_l, which is now a parameter rather than a decision variable. Thus, we will optimize treatment effect subject to a bound on the variance. We accomplish this by introducing a new constraint:

$$\sum_{i=1}^{N^t}\sum_{j=1}^{N^c}(y_i^t - y_j^c)^2 a_{ij} \le b_l. \tag{17}$$

Putting this together, the new formulation is an ILP. We simply need to solve it for many values of b_l.

Formulation 2 (ILP Formulation for z-Test).

$$\text{Maximize/Minimize}_{\mathbf{a}} \quad \sum_{i=1}^{N^t}\sum_{j=1}^{N^c}(y_i^t - y_j^c)a_{ij}$$

$$\text{(Treatment effect), subject to,} \tag{18}$$

$$\sum_{i=1}^{N^t}\sum_{j=1}^{N^c}(y_i^t - y_j^c)^2 a_{ij} \le b_l$$

$$\text{(Upper bound on sample variance),} \tag{19}$$

$$\sum_{i=1}^{N^t}\sum_{j=1}^{N^c} a_{ij} = M \qquad \text{(Choose } M \text{ pairs),} \tag{20}$$

$$\sum_{i=1}^{N^t} a_{ij} \le 1 \quad \forall j \quad \text{(Match each control unit at most once),} \tag{21}$$

$$\sum_{j=1}^{N^c} a_{ij} \le 1 \quad \forall i \quad \text{(Match each treatment unit at most once),} \tag{22}$$

$$a_{ij} \in \{0,1\} \quad \forall i,j \quad \text{(Defines binary variable } a_{ij}\text{),} \tag{23}$$

$$\text{(Additional user-defined covariate balance constraints.)} \tag{24}$$

This formulation optimizes treatment effect, subject to the variance of the treatment effect being small. This formulation can be used by itself to find the range of reasonable treatment effects, given a fixed bound b_l on the variance. The problem of testing $\mathbb{H}_0$ for the ATT and under full matching can be formulated by setting $M = N^t$ (where N^t is the number of treatment points) in Formulation 3.

Modification 1. Formulation 3 can also be modified to solve the problem of choosing both the treatment and control populations simultaneously. This is also the setting of Rosenbaum (2012). In that case, the mean is taken over the region of overlap between the control and treatment populations, removing extreme regions. This setting can be handled by looping over increasing values of M until the program becomes infeasible. We would choose the solutions corresponding to the largest values of M.

Modification 2. For testing a sharp null hypothesis of zero treatment effect, a simplified formulation is possible, because the sample variance is fixed and known in this case. We would use a special case of Formulation 3, where the variance constraint in (19) is replaced with an equality constraint, requiring the variance of the solution to be equal to the known sample variance. (The formula for the variance computed under the sharp null is available in textbooks, e.g., Imbens and Rubin 2015.)

Algorithm for Optimizing z-Score. Let us get back to optimizing the z-score. Our algorithm will solve this formulation for many different values of b_l to find the optimal z-scores and p-values. Let us denote the solution of the maximization problem as $\mathbf{a}_l$. Here, $\mathbf{a}_l$ is an optimal match assignment for a specific value of b_l. The indices of the match assignment $\mathbf{a}_l$ are, as usual, ij, which are pairs of treatment and control units. Using $\mathbf{a}_l$, we will then be able to bound the value of z. Shortly, we will use Theorem 2 to prove bounds on the z-score as follows:

$$\max_l \frac{\bar{d}_{\mathbf{a}_l}\sqrt{M}}{\sqrt{\frac{1}{M}b_l - (\bar{d}_{\mathbf{a}_l})^2}} \le \max_{\mathbf{a}} z(\mathbf{a}) \le \max_l \frac{\bar{d}_{\mathbf{a}_l}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l-1} - (\bar{d}_{\mathbf{a}_l})^2}}. \tag{25}$$

That is, when b_{l-1} is close to b_l, we have little uncertainty about $\max_{\mathbf{a}} z(\mathbf{a})$.

Using this bound, we can now formulate an algorithm (Algorithm 6 in Online Appendix D) to choose progressively finer meshes for b_l to maintain the guarantee on the quality of the solution for $\max_{\mathbf{a}} z(\mathbf{a})$, repeatedly solving the ILP formulation. An analogous algorithm (with some signs flipped) will allow us to compute $\min_{\mathbf{a}} z(\mathbf{a})$. Algorithm 6 in Online Appendix D works as follows:

- It first solves the ILP Formulation by relaxing (removing) the first constraint (upper bound on sample standard deviation).
- It then uses the resulting matches to compute the first upper bound on the standard deviation, $b_L^{(0)}$ (line 2 in Algorithm 6 in Online Appendix D).
- The algorithm then creates a coarse mesh $b_1^{(iter)}, \ldots, b_L^{(iter)}$, where $b_1^{(iter)} < b_l^{(iter)} < b_L^{(iter)}$ (Line 3) new refined mesh will be created at each iteration, and we denote iterations by *iter*. We want the interval $[b_1^{(iter)}, b_L^{(iter)}]$ to be wide enough to contain the true value of $f_2(\mathbf{a}^*) := \sum_{i=1}^{N^t}\sum_{j=1}^{N^c}(y_i^t - y_j^c)^2 a_{ij}^*$, where $\mathbf{a}^* \in \arg\max z\ (\mathbf{a})$, which we do not know and are trying to obtain. Determining which procedure to use to create this mesh is left up to the user; the $b_l^{(iter)}$ could be chosen evenly spaced, though they do not need to be. Note that the choice of $(b_1^{(iter)}, \ldots, b_L^{(iter)})$ at each iteration does not affect the optimality of the solution, only the speed at which it is obtained.
- The algorithm then determines the sign of the maximal z-statistic; if negative (line 17), it will try to maximize the denominator (variance) of the z statistic, and, if positive, it will try to minimize it (line 5).
- The algorithm then computes the solution to the ILP Formulation, as well as upper and lower bounds for the solution using (25) (lines 7 and 8 and 21 and 22).
- We then do some reindexing. We take the union of all grid points $\cup_{l,(iter)} b_l^{(iter)}$ created over all iterations, order them, and create an ordered single vector $\mathbf{b} = [b_1, \ldots, b_L]$. This is a single 1D grid. On each point of this grid, we have match assignment $\mathbf{a}_l$ that maximizes z, subject to a constraint b_l on the variance.
- For each l, the algorithm determines whether the interval $[b_{l-1}, b_l]$ can be excluded because it provably does not contain a $f_2(\mathbf{a})$ value corresponding to the maximum value of $z(\mathbf{a})$ (lines 9–12 or 22–25). In particular, we know from the bounds in (25) and from Theorem 2 that if the upper bound on the objective for a particular $b_{l'}$ is lower than all lower bounds for the optimal solution $\mathbf{a}^*$, then l' cannot equal l^*, and the interval $[b_{l'-1}, b_{l'}]$ can be excluded from further exploration. Specifically, we check for each l whether

$$\frac{\bar{d}_{\mathbf{a}_l}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l-1} - (\bar{d}_{\mathbf{a}_l})^2}} < \max_{\ell} \frac{\bar{d}_{\mathbf{a}_\ell}\sqrt{M}}{\sqrt{\frac{1}{M}b_{\ell} - (\bar{d}_{\mathbf{a}_\ell})^2}}.$$

If this holds for some l, it means l cannot equal l^*, and the interval $[b_{l-1}, b_l]$ can be excluded from further exploration.

- The intervals that remain included after this procedure are then refined again at each iteration, thus creating finer and finer meshes, and the process repeats on these finer meshes until most intervals are excluded and the desired tolerance (ϵ) is achieved.
- The output is a binary matrix of match assignments, $\mathbf{a}$, and the optimal value of $z(\mathbf{a})$.

Correctness of the algorithm follows directly from optimality of the solutions at the bounds mentioned earlier and through use of the following theorem:

Theorem 2 (Optimal Solution of ILP with Upper Bound on Variance). *Let the functions f_1 and f_2 be real-valued functions of $x \in X$, and let $F(f_1(x), f_2(x))$ be monotonically increasing in f_1 and monotonically decreasing in f_2. Consider the optimization problem of finding $x^* \in \arg\max_x F(f_1(x), f_2(x))$, and assume we are given $[b_1, b_2, \ldots, b_l, .., b_L]$ that span a wide enough range so that x^* obeys: $b_{l^*-1} \le f_2(x^*) \le b_{l^*}$ forsome $l^* \in \{1, \ldots, L\}$.*

1. *For $x_l \in \arg\max_{x: f_2(x) \le b_l} F(f_1(x), b_l) = \arg\max_{x: f_2(x) \le b_l} f_1(x)$, where the equality follows because F monotonically increases in f_1, we have:*

$$\max_l F(f_1(x_l), b_l) \le \max_x F(f_1(x), f_2(x)) \le F(f_1(x_{l^*}), b_{l^*-1}) \le \max_l F(f_1(x_l), b_{l-1}).$$

2. *For $x_l \in \arg\min_{x: f_2(x) \le b_l} F(f_1(x), b_l) = \arg\min_{x: f_2(x) \le b_l} f_1(x)$, we have:*

$$\min_l F(f_1(x_l), b_l) \ge \min_x F(f_1(x), f_2(x)) \ge F(f_1(x_{l^*}), b_{l^*-1}) \ge \min_l F(f_1(x_l), b_{l-1}).$$

Proof. See Online Appendix B.

This theorem bounds the optimal value of F along the whole regime of x in terms of the values computed at the L grid points. Note that the objective function of Formulation 1 and Formulation 2 in the online appendix is exactly of the form of Theorem 2, where

$$f_1(\mathbf{a}) = \frac{1}{M}\sum_{i=1}^{N^t}\sum_{j=1}^{N^c}(y_i^t - y_j^c)a_{ij},$$

$$f_2(\mathbf{a}) = \sum_{i=1}^{N^t}\sum_{j=1}^{N^c}(y_i^t - y_j^c)^2 a_{ij}, \text{ and}$$

$$F(f_1(\mathbf{a}), f_2(\mathbf{a})) = \frac{f_1(\mathbf{a})\sqrt{M}}{\sqrt{\frac{1}{M}f_2(\mathbf{a}) - (f_1(\mathbf{a}))^2}}.$$

Note that monotonicity of F in f_2 is ensured in the algorithm by the conditions on the sign of the treatment effect at lines 5 and 17 of the algorithm. The extra constraints on $\mathbf{a}$ in the ILP Formulation are compatible with Theorem 2. Thus, the bounds in (25) are

direct results of Theorem 2 applied to the z statistic as an objective function. Finally, minimization of z can be achieved with Algorithm 6 in the online appendix by flipping the treatment indicator—that is, calling control units as treated and treated units as control in the input to the algorithm, then running the procedure as is.

5. Approximate and Exact Distributions of Robust Test Statistics

As introduced in Section 2, our method aims to find an upper bound on the p-value of a desired test statistic over the set of good matches. Although not directly used, a quantity relevant to our robust tests is the distribution of the test statistics under the maximal and minimal match assignments themselves—that is, the distribution of $\Psi^+_{\mathcal{D}} = \max_{\mathbf{a}\in\mathcal{A}_{\text{good}}} \Psi_{\mathcal{D}}(\mathbf{a})$ and $\Psi^-_{\mathcal{D}} = \min_{\mathbf{a}\in\mathcal{A}_{\text{good}}} \Psi_{\mathcal{D}}(\mathbf{a})$, which are random variables denoting a generic test statistic (Ψ) under the match assignment that maximizes and minimizes it, respectively. Although we do not employ these distributions to derive p-values for our tests, we still choose to study them here to better understand the statistical behavior of our robust tests.

Specifically, we will study the distribution of the robust version of McNemar's test, taking into account the randomness due to the choice of match assignment, in order to provide analysts with exact distributions for the robust statistics. Without Exclusively Binning Constraints, as defined in Section 3, the joint distributions can be extremely complex, but with these constraints, the calculation becomes much clearer. As a reminder, these are constraints that uniquely define a division of the units into strata, such that matches between two units in the same stratum always satisfy the quality constraints that define $\mathcal{A}_{\text{good}}$. To maintain exposition, we refrain from stating our results for this section formally (as the statements of these distributions are analytical, but cumbersome), and instead give informal statements of our theorems. We refer to the online appendix for rigorous versions of all our results and their proofs. This computation can be used to derive exact p-values for our robust McNemar's test. Specifically, in the following sections, we show that analytical formulas exist for both null distributions we seek under Exclusively Binning Constraints and that these formulas can be used to generate a lookup table for the distributions of (χ^+, χ^-) in polynomial time.

5.1. Exact Randomization Distribution of (χ^+, χ^-)

Let us compute the finite sample joint distribution of (χ^+, χ^-) under Fisher's sharp null hypothesis of no treatment effect, which is defined as, for all i:

$$\mathbb{H}_0^{sharp} := Y_i(1) = Y_i(0).$$

We compute this distribution for the case in which constraints are exclusively binning and $\mathcal{S}$ can be constructed prior to matching. In this case, potential outcomes for all of the units in our sample are fixed and nonrandom: The only randomness in our calculations will stem from the treatment assignment distribution.

Our approach to constructing this randomization distribution without conditioning on the match assignment is as follows: We assume that any units that could be matched together under a good assignment have the same probability of receiving the treatment and that, after treatment is assigned, matches are made to compute (χ^+, χ^-) with the ILP in Formulation 1. This ILP is modified to include Exclusively Binning Constraints, in addition to those in the formulation, and to make as many matches as possible within each of the strata defined by the constraints. In this setting, we assume that treatment assignment is random within the strata defined by the Exclusively Binning Constraint. This implies that Assumption 1 becomes:

Assumption 4 (Stratified Conditional Ignorability). *All units i in stratum l receive the treatment with fixed probability e_l. That is: For all $l = 1, \ldots, L$: $\Pr(T_i = 1 \mid X = x_i, i \in S_l) = \Pr(T_i = 1 \mid i \in S_l) = e_l$, with $0 < e_l < 1$.*

Let us introduce some notation to represent outcome counts within strata. We use U_l to denote the number of treated units in stratum l with outcome equal to one, V_l to denote the number of treated units with outcome zero, η_l, the number of control units with outcome one, and ν_l the number of control units with outcome zero. Note that, by $\mathbb{H}_0^{sharp}$, we have the number of units with outcome one and zero fixed in each stratum; thus, for stratum l, we use N_l^1 and N_l^0 to denote the number of units with outcome one and zero in stratum l, respectively. Under Assumptions 2–4, these count variables follow the following generating process:

$$U_l \mid \mathbb{H}_0^{sharp} \overset{iid}{\sim} Bin(e_l, N_l^1), \qquad \eta_l = N_l^1 - U_l, \tag{26}$$

$$V_l \mid \mathbb{H}_0^{sharp} \overset{iid}{\sim} Bin(e_l, N_l^0), \qquad \nu_l = N_l^0 - V_l, \tag{27}$$

$$N_l^t = U_l + V_l, \qquad N_l^c = \nu_l + \eta_l. \tag{28}$$

One can view this data-generation process as a result of our assumptions, including those of $\mathbb{H}_0^{sharp}$. The distribution of (χ^+, χ^-) under $\mathbb{H}_0^{sharp}$ has the following properties:

Theorem 3 (Randomization Distribution of (χ^+, χ^-)). *Under Assumptions 1, 2, and 4 and conditionally on $\mathbb{H}_0^{sharp}$:*

- *For a given stratum, l, the joint distribution of B_l and C_l (the counts of optimal matched pairs) can be expressed as a function of the distribution of the count variables U_l, V_l, ν_l, N_l^t, and N_l^c defined in (26)–(28).*
- *Given the signs of (χ^+, χ^-), for any two strata, l and l', B_l and C_l are independent of $B_{l'}$ and $C_{l'}$.*
- *The joint distribution of (χ^+, χ^-) is the convolution of the joint distributions over B_l, C_l with each other.*

A precise formulation for the distribution of test statistics in the theorem above is given in Theorem 7 in Online Appendix E.

5.2. Exact Conditional Distribution of (χ^+, χ^-)

In this section, we seek a formulation for $\Pr(\chi^- = s, \chi^+ = r \mid X)$, the distribution of robust McNemar's statistics under exclusively binning constraints and a fixed number of treated units, N_l^t in each stratum, but without conditioning on a specific match assignment. Potential outcomes are treated as random quantities in this case, and treated and control outcomes are assumed to be equal only on average. This distribution permits testing of hypotheses other than $\mathbb{H}_0^{sharp}$ and is more general than the randomization distribution derived in the previous subsection. Now, denote with N_l^t the number of treated units in the stratum l, and with N_l^c the number of control units in that same stratum. In addition, let $p_l^t = \Pr(Y_i = 1 \mid i \in S_l, T_i = 1)$, and $p_l^c = \Pr(Y_i = 1 \mid i \in S_l, T_i = 0)$. For each stratum, $l = 1, \ldots, L$, the data are generated as follows:

$$U_l \mid N_l^t \overset{iid}{\sim} Bin(p_l^t, N_l^t), \quad V_l = N_l^t - U_l, \tag{29}$$

$$\eta_l \mid N_l^c \overset{iid}{\sim} Bin(p_l^c, N_l^c), \quad \nu_l = N_l^c - \eta_l. \tag{30}$$

As we did before, we can use the distributions of these count variables to arrive at a formulation for the joint distribution of (χ^+, χ^-) under the DGP just outlined.

Theorem 4 (Conditional Distribution of (χ^+, χ^-)). *Under Assumptions 1, 2, and 4 and, for all strata, l, conditionally on N_l^t, N_l^c:*

- *For a given stratum, l, the joint distribution of B_l and C_l (the counts of optimal matched pairs) can be expressed as a function of the distribution of the count variables U_l, V_l, ν_l, N_l^t, and N_l^c defined in (29) and (30).*
- *Given the signs of (χ^+, χ^-), for any two strata, l and l', B_l and C_l are independent of $B_{l'}$ and $C_{l'}$.*
- *The joint distribution of (χ^+, χ^-) is the convolution of the joint distributions over B_l, C_l with each other.*

A precise formulation for the distribution in the theorem above is in Theorem 8 in Online Appendix F. Note that this distribution can be used to test the null hypothesis of no effect in each stratum by requiring $p_l^t = p_l^c$ for all l in the general formulation we provide in the theorem.

5.3. Fast Computation of Lookup Tables for (χ^+, χ^-)

Another important advantage of the theorems given in the last two sections is that they permit us to generate lookup tables for the finite-sample null distributions of interest in polynomial time. Let $M = \sum_{l=1}^{L} M_l$ be the total number of matches made. A naïve procedure for computing a lookup table for the distribution in Theorems 3 and 4 for a given data set based on brute-force enumeration has computational complexity $O(2^{N^M + (N^M)^2})$, because naively computing the domain of these distributions is a version of the subset-sum problem.

Fortunately, the complexity of computing the distribution of (χ^+, χ^-) under both $\mathbb{H}_0$ and $\mathbb{H}_0^{sharp}$ can be significantly reduced because by Theorems 3 and 4, these distributions are convolutions of simpler distributions. We take advantage of this fact and of existing fast convolution algorithms to establish the following result:

Theorem 5. *There exists an algorithm that creates a probability table for the null distribution of (χ^+, χ^-) given in Theorems 3 and 4 in $O(4MN^{4M}\log N)$.*

We give proof of this theorem and outline the algorithm in question in the online appendix. Note that the theorem directly implies that the worst-case time of computing the exact distributions of (χ^+, χ^-) is polynomial in N. Figure 10 in Online Appendix H shows marginal distributions of (χ^+, χ^-) computed with the algorithm in Theorem 5 and shows that there is a location and scale difference between the distribution of (χ^+, χ^-) when the null is true and when it is not. This demonstrates our tests' capacity to detect full-sample treatment effects.

6. Simulations

We present a set of experiments on simulated data sets that demonstrated the usefulness and performance of our method in a series of settings. We compare the performance of both our robust tests against several popular matching methods and find that our proposed tests tend to display better performance (lower Type-I error) in several cases, without sacrificing statistical power.

In all our simulations, we prespecify the number of units N, number of treated units N^t, and number of covariates P. We will vary N throughout; however, we will keep N^t at 20% of the whole sample and $p = 20$, unless otherwise specified (simulations with varying P and constant N are available in the online appendix). We then generate data as follows, for $i = 1, \ldots, N$:

$$x_{i1}, \ldots, x_{iP} \sim \text{Uniform}(-1, 1), \quad \epsilon_i \sim \text{Normal}(0, 1).$$

Propensity scores are then simulated according to:

$$e_i = \frac{1}{1 + \exp(-x_i \beta_0)},$$

where β_0 is a P-dimensional vector of coefficients, which are prespecified before data generation. Treatment t_i is then set to one for exactly N_t units, such that each unit has $t_i = 1$ with probability e_i. We then introduce two different data-generating processes for our outcome variable:

$$\begin{aligned} \text{Linear}: \quad & y_i^* = t_i\tau + x_i\beta + \epsilon_i \\ \text{Complex}: \quad & y_i^* = t_i\tau + x_i\beta + x_i^2\beta_2 + \sin(x_i)\beta_3 \\ & \quad + \mathbb{I}[x_{i1} > 0]\beta_4 + \epsilon_i. \end{aligned}$$

To simulate binary data for McNemar's test, we then set $y_i = \mathbb{I}[y_i^* > 0]$, and to simulate continuous data for the z-test, we set $y_i = y_i^*$. The coefficient vectors $\beta_0, \ldots, \beta_4$, as

Table 2. Methods Used in the Simulation Studies

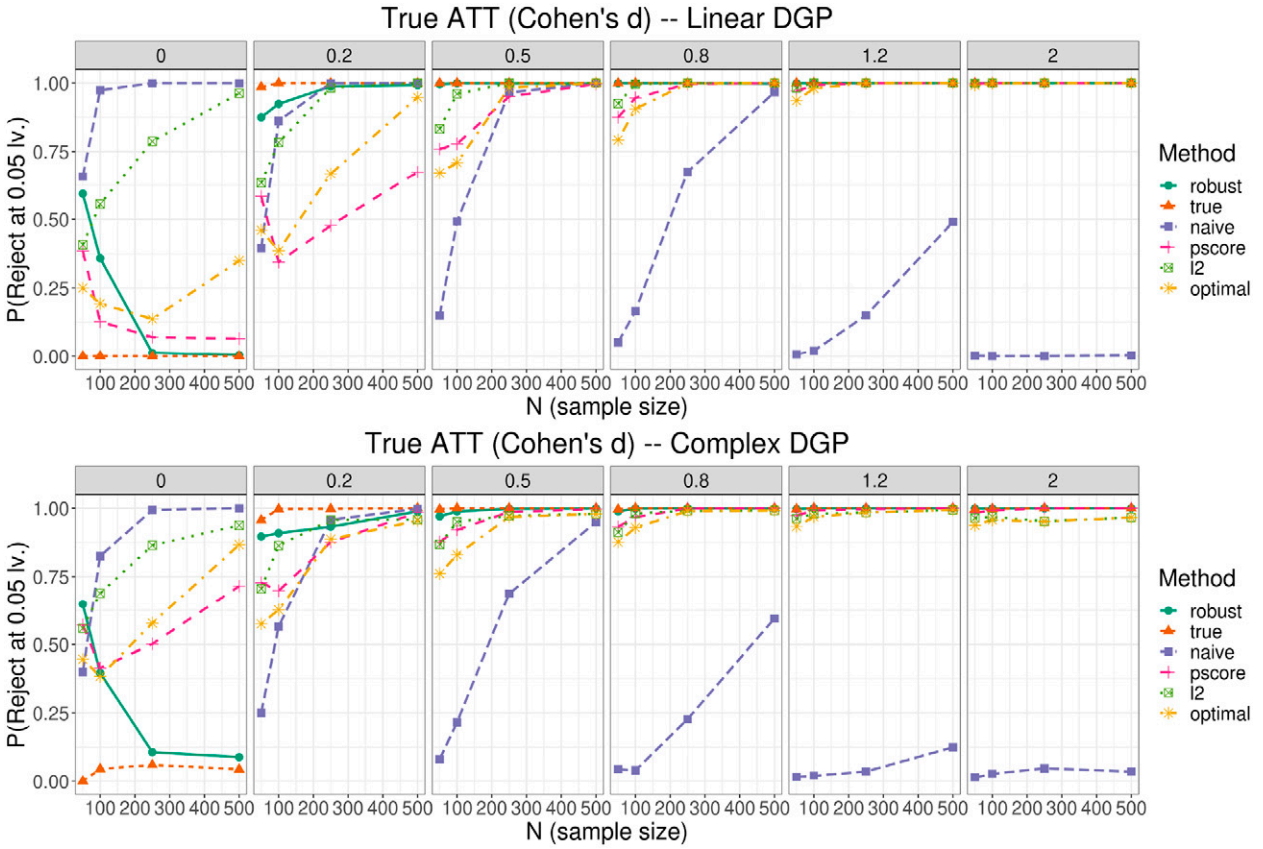
Name	Description	Citation
Robust	Robust test	This paper
True	True potential outcomes	Benchmark
Naive	No matching at all	Benchmark
Pscore	Propensity-score matching	Rosenbaum (1984)
L2	Matching on the L_2 distance of the covariates	e.g., Abadie and Imbens (2006)
Optimal	Optimal matching via Mixed Integer Programming	Zubizarreta (2012)

well as τ, are calibrated at each simulation so that the population ATT has a Cohen's d-statistic (defined as: $\frac{\mathbb{E}[Y(1)-Y(0)|T=1]}{Var[Y(1)-Y(0)|T=1]}$) of a prespecified value. In most settings, we will simulate data for $d = 0, 0.2, 0.5, 0.8, 1.2, 2$, which are commonly understood to correspond to null, tiny, weak, moderate, strong, and very strong treatment effects. For each setting, we simulate 1,000 data sets and compute the proportion in which each method rejects the null hypothesis of zero treatment effect at the 5% level as our estimate of the rejection rate of each method.

We compare against several commonly employed matching methods, as well as benchmark approaches, such as a completely naïve test without any matching and an idealized hypothesis test conducted on both potential outcomes, which is never possible in practice. We summarize the methods included in our simulation studies in Table 2.

To ensure the closest possible comparison, we constrain all methods to match all treated units to one control unit without replacement—that is $M = N^t$. Given that one of the main points of our paper is that there is not clear guidance on how to select hyperparameters for any given matching method, we try to use default or author-recommended values whenever possible for all the other methods tested. When this is

Figure 2. (Color online) Comparison of Performance of Robust McNemar's Test Against Several Other Matching Methods



Notes. First row: linear outcome DGP; second row: complex outcome DGP. Ideally, a test would not reject the null when the ATT is zero (leftmost panel), and the probability of rejection would increase as the ATT becomes stronger (four rightmost panels). The ideal method is labelled "true" and represented by the dashed line in the figure.

not possible, we set hyperparameters to mimic those of the robust test as much as possible.

In all simulations, the set of good matches considered by our tests will be defined by all those matches that satisfy a mean balance constraint, as well as a caliper on the propensity-score distance of the units matched. These constraints are implemented in our MIPs as follows:

- *Mean Balance*: $|\frac{1}{M}\sum_{i=1}^{N^t}\sum_{j=1}^{N^c} a_{ij}(x_{ip}^t - x_{jp}^c)| \le \epsilon \times (\sigma(x_p^t)/2 + \sigma(x_p^c)/2)$, where $\sigma_p(x_p^t)$, $\sigma_p(x_p^c)$ are the standard deviation of covariate x_p in the treated and control set, respectively (P constraints total).
- *Caliper*: $a_{ij}|e_i^t - e_j^c| \le \epsilon$ for $i = 1,\dots N^t$, $j = 1,\dots,N^c$, where e_i^t is the propensity score for treated unit i, and e_j^c is the analogue for control unit j ($N^t \times N^c$ constraints total).

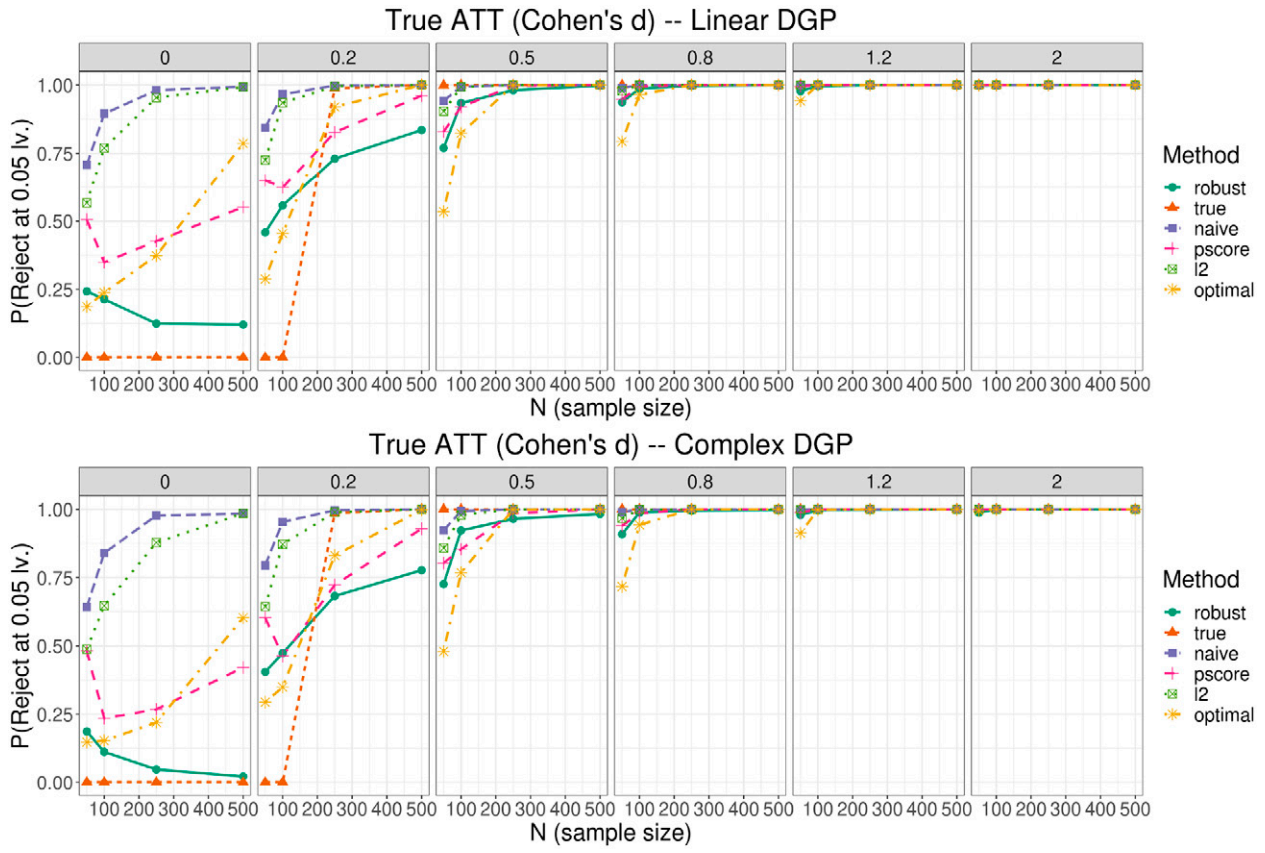
The value of ϵ is chosen to be the smallest feasible one on a grid between 0.01 and 1.

6.1. Results

We first present results from comparison of our robust McNemar's test against the other methods. For all methods, p-values were computed after matching using the conventional formula for exact p-values for McNemar's test (see Tamhane and Dunlop 2000).

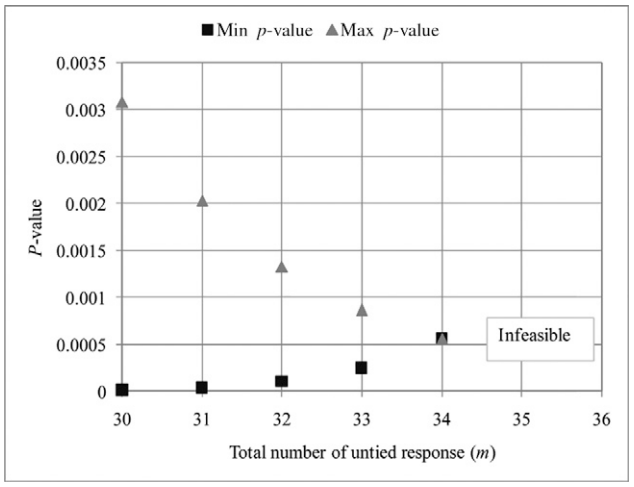
Results are reported in Figure 2. We see that the robust test performs well as sample size grows: It displays low probability of rejection when the treatment effect is zero, and probability of rejection grows as the strength of the treatment also grows. When there is no ATT in the population, the robust test does display an error rate, but this is only when the sample size is low (≤100), and the error rate decreases as the sample size grows, eventually converging with the one for the idealized McNemar's test performed with the true potential outcomes. This is not true for methods we compare against, which seem to display an *increase* in error rate (incorrect rejection) as the sample size increases. When an ATT is present in the population, the robust test displays statistical power comparable to or better than that of other matching methods, indicating that our bounding approach is not too extreme, given that matches are chosen well. Figure 3 reports results for the robust z-test in the simulated data settings introduced previously.

Figure 3. (Color online) Comparison of Performance of Robust z-Test Against Several Other Matching Methods



Notes. First row: linear outcome DGP; second row: complex outcome DGP. The ideal method is labelled "true" and represented by the dark dashed line in the figure.

Figure 4. (Case Study 1) Variation of McNemar's Test p-Values for Different m



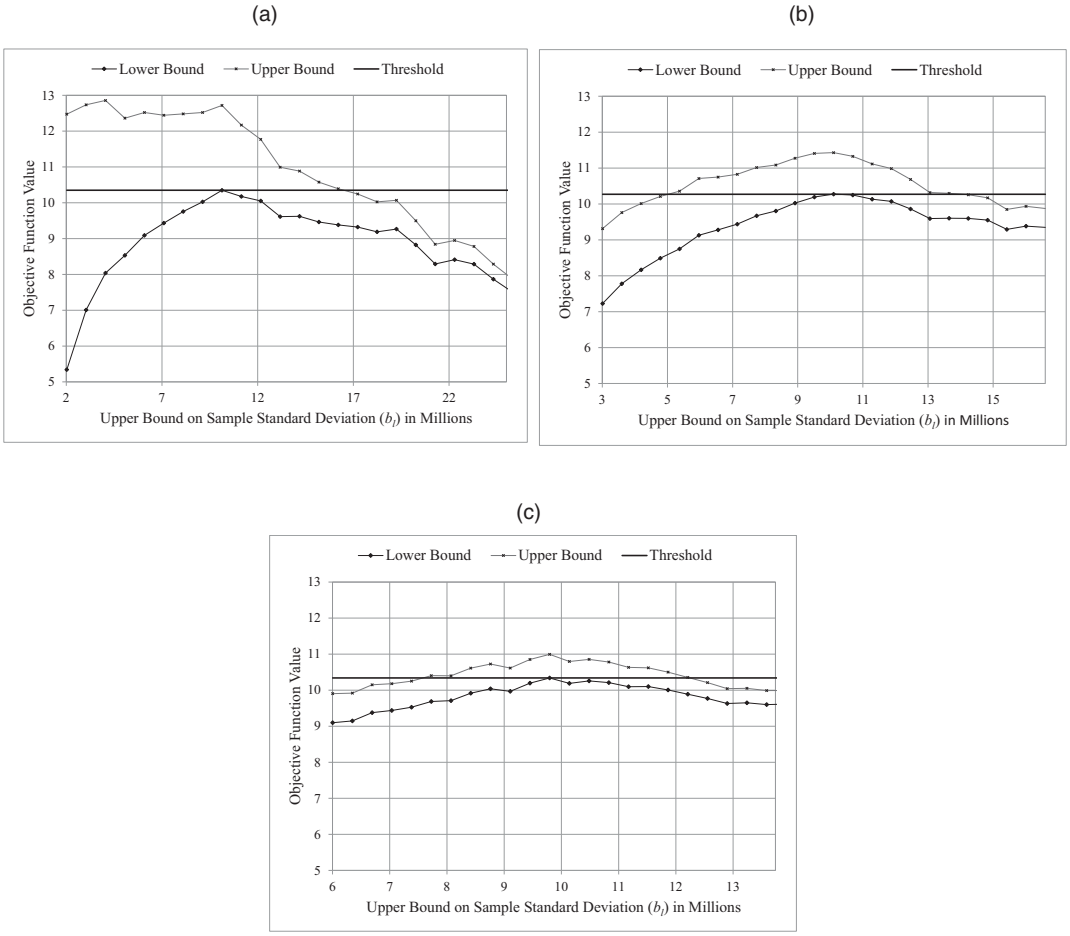
Note. As long as there are enough matched groups, then for *any* reasonable choice of match assignment, the null hypothesis is rejected.

This version of our robust test performs even more strongly than the robust McNemar's test, compared with other prominent matching methods. In this case, the robust z-test displays a good rejection rate, even at low sample values ($n = 50$) when the true ATT is zero, and is well powered as the ATT increases in strength, as well as when sample size grows.

7. Case Studies

In this section, we apply our proposed methods to two real-world data sets. We show that our test statistics can produce robust results in both data sets, which suggests that our methods can have wide practical applicability. Some of our results show that null hypotheses can be rejected robustly, whereas in some other cases, they show that there is not enough evidence to reject the null hypothesis once the additional uncertainty from the choice of matching procedure is quantified via our robust tests.

Figure 5. Upper and Lower Bounds for Maximum z-Test Objective Function Value over a Range of b_l (Bike Sharing Data, $n = 45$), Illustrating the Optimum Search Range at Various Steps in Algorithm 6 in the Online Appendix



Notes. The final value with three iterations is shown in (c). The final value for the maximization problem is between 10.34 and 10.99. (a) Initial mesh. (b) Refined mesh. (c) Finest mesh.

7.1. Case Study 1: The Effect of Smoking on Osteoporosis in Women

In this case study, we used Global Longitudinal Study of Osteoporosis in Women (GLOW) data used in the study of Hosmer et al. (2013). Each data point represents a patient, and the outcome is whether the person developed a bone fracture. The treatment is smoking. We match on several pretreatment covariates: age, weight, height, and BMI. As there are several more control than treated units, we test $\mathbb{H}_0$ for the ATT with the general formulation of McNemar's test introduced before, by matching almost all of the treated units.

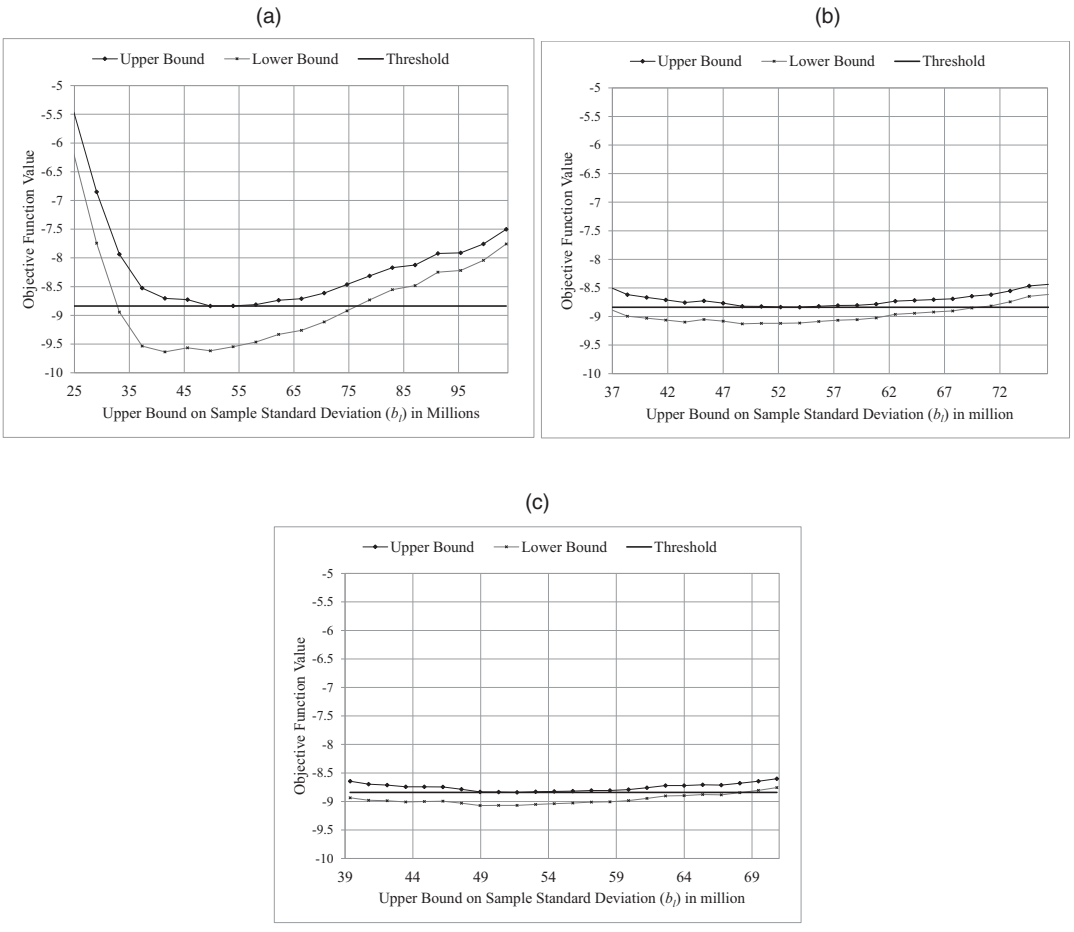
Figure 4 shows results for testing $\mathbb{H}_0$ with the general program in Formulation 1, without binning constraints. We include a balance constraint in the formulation, by requiring that matched units i, j respect $dist_{ij} \leq 0$, where $dist_{ij}$ is zero if the sum of the differences of all the covariates is six or less, and one otherwise. We choose the value six because this is the smallest caliper on the absolute distance between units that still permits all treated units to be matched. This figure lends evidence to the fact that $\mathbb{H}_0$ can be rejected robustly when the match assignment induces better balance in the data. We can conclude this because both min and max p-values are below 0.01 at each value of M, signifying that results would be statistically significant and positive using any good-quality match assignment.

This case study shows the ability of our tests to detect effects when they are robust and present, even with a small number of observations.

7.2. Case Study 2: The Effect of Mist on Bike-Sharing Usage

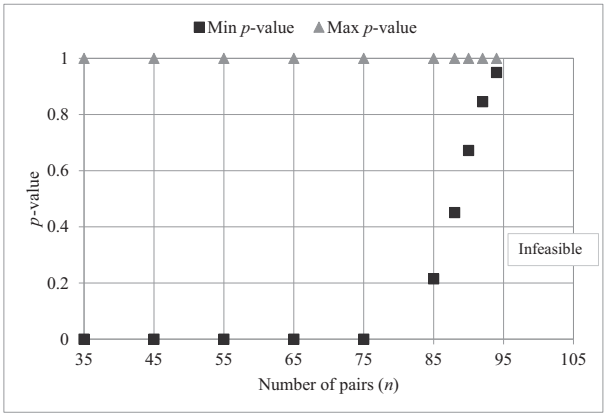
In our second case study, we used two years (2011–2012) of bike-sharing data from the Capital Bike Sharing system (see Fanaee-T and Gama 2014) from Washington, DC. We study the effect of misty weather on the number of bikes rented. Control covariates include *Season*, *Year*, *Workday*, *Temperature*, *Humidity*, and *Wind Speed*. Additional information on this case study is available in Online Appendix K.

Figure 6. Upper and Lower Bounds for Minimum z-Test Objective Function Value over a Range of b_l (Bike Sharing Data, $n = 90$), Illustrating the Optimum Search Range at Various Steps in Algorithm 6 in the Online Appendix



Notes. The final value with three iterations is found in Figure 5(c). The final value for the minimization problem is between −8.84 and −9.06. (a) Initial mesh. (b) Refined mesh. (c) Finest mesh.

Figure 7. Variation of z-Test Optimum p-Values for Different N (Case Study 2)



Figures 5 and 6 show the upper and lower bounds for the maximum objective function value for different b_l with $n = 30$ for the maximization problem in Figure 5 and the minimization problem in Figure 6. These figures illustrate the meshes at different scales within the algorithm. We computed p-values for Z^+ and Z^- under several counts of matched pairs: $M = 30, 50, 70, 90, 110$. For $M = 30$ through $M = 90$, the p-value for Z^+ was zero, and the p-value for Z^- was one, whereas both p-values become one when the number of matches is 110. This is shown in Figure 7. The problem becomes infeasible for a larger number of matches. This illustrates that there is a lot of uncertainty associated with the choice of match assignment—that is, a reasonable experimenter choosing 90 matched pairs can find a p-value of ~0 and declare a statistically significant difference, whereas another experimenter can find a p-value of ~1 and declare the opposite. In this case, it is truly unclear whether mist has an effect on the total number of rental bikes. Figure 7 shows robust p-values for different amounts of matched units.

8. Conclusion

Believing hypothesis-test results conducted from matched-pairs studies on observational data can be perilous. These studies typically ignore the uncertainty associated with the choice of matching method and, in particular, how the experimenter or an algorithm chooses the matched groups: We have given both simulated and real-data evidence of this problem. We want to know that for *any* reasonable choice of experimenter who chooses the assignments, the result of the test would be the same. In this work, we have addressed the issue above by introducing robust test statistics that consider extrema over all possible good matches. This is justified because p-values obtained for the extrema must, by definition, include p-values obtained under any other possible good match. We have provided practical implementations of this principle for both discrete and continuous data, as well as theoretical and empirical analyses of the performance of our methods.

References

Abadie A, Imbens GW (2006) Large sample properties of matching estimators for average treatment effects. *Econometrica* 74(1):235–267.

Abadie A, Imbens GW (2011) Bias-corrected matching estimators for average treatment effects. *J. Bus. Econom. Statist.* 29(1):1–11.

Coker B, Rudin C, King G (2021) A theory of statistical inference for ensuring the robustness of scientific results. *Management Sci.* 67(10):6174–6197.

Diamond A, Sekhon JS (2013) Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Rev. Econom. Stat.* 95(3):932–945.

Fanaee-T H, Gama J (2014) Event labeling combining ensemble detectors and background knowledge. *Progress Artificial Intelligence* 2(2-3):113–127.

Holland PW (1986) Statistics and causal inference. *J. Amer. Statist. Assoc.* 81(396):945–960.

Hosmer D, Lemeshow S, Sturdivant R (2013) *Applied Logistic Regression*, 3rd ed. (John Wiley and Sons Inc., Hoboken, NJ).

Iacus SM, King G, Porro G (2011) Multivariate matching methods that are monotonic imbalance bounding. *J. Amer. Statist. Assoc.* 106(493):345–361.

Iacus SM, King G, Porro G (2012) Causal inference without balance checking: Coarsened exact matching. *Polit. Anal.* 20(1):1–24.

Imai K, King G, Stuart EA (2008) Misunderstandings between experimentalists and observationalists about causal inference. *J. Roy. Statist. Soc. Ser. A* 171(2):481–502.

Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, Cambridge, UK).

Morgan SL, Winship C (2007) *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* (Cambridge University Press, Cambridge, UK).

Morucci M, Rudin C (2020) Matching bounds: How choice of matching method affects treatment effect estimates and what to do about it. Preprint, submitted September 6, https://arxiv.org/abs/2009.02776.

Rosenbaum PR (1984) Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Assoc.* 79(387):565–574.

Rosenbaum P (1989) Optimal matching for observational studies. *J. Amer. Statist. Assoc.* 84:1024–1032.

Rosenbaum P (2010) Evidence factors in observational studies. *Biometrika* 97(2):333–345.

Rosenbaum PR (2012) Optimal matching of an optimally chosen subset in observational studies. *J. Comput. Graphical Statist.* 21:57–71.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Ed. Psych.* 5(66):688–701.

Rubin D (2007) The design vs. the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statist. Med.* 26(1):20–36.

Tamhane A, Dunlop D (2000) *Statistics and Data Analysis* (Prentice Hall, Englewood Cliffs, NJ).

Zubizarreta J (2012) Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* 107(500):1360–1371.