Submitted to manuscript (Please, provide the manuscript number!)

Hypothesis Tests That Are Robust to Choice of Matching Method

Marco Morucci
Dept. of Political Science, Duke University, 140 Science Dr, Durham, NC 27708, USA marco.morucci@duke.edu

Md. Noor-E-Alam
Dept. of Mechanical and Industrial Engineering, Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA md.alam@neu.edu

Cynthia Rudin
Dept. of Computer Science, Duke University, 308 Research Drive, Durham, NC 27708, USA cynthia@cs.duke.edu

A vast number of causal inference studies test hypotheses on treatment effects after treatment cases are matched with similar control cases. The quality of matched data is usually evaluated according to some metric, such as balance; however the same level of match quality can be achieved by different matches on the same data. Crucially, matches that achieve the same level of quality might lead to different results for hypothesis tests conducted on the matched data. Experimenters often specifically choose not to consider the uncertainty stemming from how the matches were constructed; this allows for easier computation and clearer testing, but it does not consider possible biases in the way the assignments were constructed. What we would really like to be able to report is that no matter which assignment we choose, as long as the match is sufficiently good, then the hypothesis test result still holds. In this paper, we provide methodology based on discrete optimization to create robust tests that explicitly account for this variation. For binary data, we give both fast algorithms to compute our tests and formulas for the null distributions of our test statistics under different conceptions of matching. For continuous data, we formulate a robust test statistic, and offer a linearization that permits faster computation. We apply our methods to real-world datasets and show that they can produce useful results in practical applied settings.

Key words: causal inference, observational studies, hypothesis test, matched pairs design, discrete optimization, integer programming.

1. Introduction

As massive and varied amounts of observational data are accumulating in healthcare, internet marketing, and governance, these data are increasingly used for understanding important cause-and-effect relationships. We might want to know whether a policy causes people to use fewer public services, or we might want to know whether a particular drug causes a side effect, or whether a view of an internet advertisement results in an increased chance of purchase. Controlled trials with randomized treatment and control populations are often small and expensive, and not often
possible due to the ethics of offering the treatment, or perhaps not possible due to the fact that the
treatment happened only in the past. Observational Inference addresses this issue by supposing
that treatment is administered based exclusively on a set of attributes related to both outcomes
and treatments, such that when the values of these attributes are held fixed, then treatment is
assigned uniformly at random.

Classically, assignments of treatment and control units to matches are constructed using a fixed
design (Rosenbaum 2010), without regard to the outcome (Rubin 2007, 2008). Practically, this
could be a major flaw in the current paradigm. 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 matches, 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 pairs, would get results that
disagree. The use of matching for covariate adjustments is often advocated in place of parametric
models (Ho et al. 2007) because estimates with the former do not depend on arbitrary assumptions
about the parametric form of the outcome distribution. Unfortunately, variation in estimates due
to arbitrary choice of matching method injects the same type of uncertainty into the results that
matching was supposed to remove in the first place. The problem of model choice becomes the
problem of match choice.

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. 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.

Finding a set of matched pairs that obey certain conditions is purely a data mining problem.
Similar subfields of data mining, where the goal is to locate optimal subsets of data, include asso-
ciation rule mining and event detection. For these types of problems, modern discrete optimization
techniques have rarely been used, though there is one recent precedent in the literature for matched
pairs, namely the line of work by Zubizarreta (2012). Zubizarreta et al. (2013, 2014). Optimization
techniques have major advantages over other types of approaches: (i) they are extremely flexible
and allow the experimenter to match on very complex conditions, such as quantiles of attributes,
which network optimization methods for matching cannot handle; (ii) they can be computationally efficient, depending on the strength of the integer programming formulation – strong formulations have relaxations close to the set of feasible integer solutions, and (iii) mixed-integer linear programs have guarantees on the optimality of the solution – in fact, they produce upper and lower bounds on the value of the optimal solution.

After formalization of our framework in Section 2, we offer a robust formulation for McNemar’s statistic for binary outcomes in In Section 3 we introduce a general Integer Linear Program for general matching problems, and a non-ILP algorithm time algorithm for matching problems with a particular type of constraint. We then give methods to compute p-values for this test under several null hypotheses as well as conceptions of matching. We provide exact formulas for randomization and conditional distributions of the statistic.

In Section 4 we outline an ILP to compute the robust version of the canonical z-test in a general case. We offer a linearized formulation for the program that makes it solvable with any popular ILP solver. Finally, we present two case studies that apply our robust statistics to real-world datasets in Section 5.

2. Matching for Robust Tests

Throughout this paper we adopt the potential outcomes framework (see Holland 1986, Rubin 1974). For each unit \( i = 1, \ldots, N \), we have potential outcomes \( Y_i(t) \), 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 \( Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i) \). To conduct our hypothesis test we have observations: \((x^t_1, y^t_1), \ldots, (x^t_{N^t}, y^t_{N^t}) \) and \((x^c_1, y^c_1), \ldots, (x^c_{N^c}, y^c_{N^c}) \), with \( X \) being a set of covariates that take value in some discrete set \( C \). We make the classical assumptions of Conditional Ignorability of treatment assignment and Stable Unit Treatment Value (SUTVA):

Assumption 1. (Conditional Ignorability) For any two units, \( i, j \), we assume that: \( \Pr(T_i = 1 | X_i = x) = \Pr(T_j = 1 | X_j = x) = e(x) \), where \( e \) is some function mapping from covariate space, \( C \) to \((0, 1)\).

Assumption 2. (SUTVA) For all units, \( i: Y_i(t_1, \ldots, t_N) = Y_i(t_i) \).

A matching operator determines which control is assigned to which treatment. In this paper we focus on one-to-one matching, therefore we define the matching operator as follows.

Definition 1. (Matching Operator) A matching operator \( a : \{1, \ldots, N^t\} \to \{1, \ldots, N^c, \emptyset\} \) obeys the following: if \( i \neq k \) and \( a(i) \neq \emptyset \) then \( a(i) \neq a(k) \). That is, no two treatment units \( i \) and \( k \) are assigned to the same control unit.
We define the size of the matching, i.e: the number of matched pairs, to be $M = \sum_{i=1}^{N_t} I(a(i) \neq \emptyset)$, with $I(E)$ representing the indicator function for the event $E$ throughout the paper. The set of all matching operators is $A = \text{set of all possible assignments } \{a\}$. Throughout the rest of the paper we slightly abuse notation by also using $a$ to represent a $N^t \times N^c$ matrix of matches such that $a_{ij} = 1$ if $i$ and $j$ are matched together.

Throughout the paper we will be interested in testing the following null hypotheses, starting from hypotheses on the population of experimental units, with outcomes, treatment indicators and covariates respectively denoted by $Y, T, X$:

\[ H_{ATE}^0 := E_X[E_{Y|X}[Y(1) - Y(0)|X]] = 0, \] (1)

\[ H_{ATT}^0 := E_X[E_{Y|X}[Y(1) - Y(0)|X,T = 1]] = 0. \] (2)

The first two are null hypotheses of 0 average treatment effects, one on the whole sample, and the other only on the treated units. All these tests presuppose potential outcomes $Y(t)$ to be random variables drawn from some distribution, we also consider randomization inference for finite samples, in this case potential outcomes are assumed to be nonrandom and fixed at the same value under any treatment assignment:

\[ H_{\text{sharp}}^0 := Y_i(1) = Y_i(0) \text{ for all } i. \] (3)

This is Fisher’s classic sharp null hypothesis of no treatment effect, and while it is a more stringent null hypothesis than the hypothesis of no average treatment effect, it is widely used in experimental and non-experimental research today. This hypotheses, unlike the prior ones, presupposes potential outcomes to be fixed and nonrandom.

For now, consider a generic test statistic $\psi_y(a)$, where $y = (y_1^t, \ldots, y_{N_t}^t, y_1^c, \ldots, y_{N_c}^c)$. This particular notation for $\psi$ is used to emphasize the dependence of the test statistic on the matches, $a$, by denoting that $\psi$ is always evaluated on the same dataset $y$, but with possibly varying matches.

### 2.1. Matching Analyses and Their Sensitivity to Arbitrary Choices

Existing studies that use matching for hypothesis testing all roughly follow the following template:

1. Choose a test statistic
2. Define criteria that acceptable matches should satisfy (e.g. covariate balance)
3. Apply one or several matching algorithms to the data, generating several sets of matches that satisfy the previously defined criteria
4. Arbitrarily choose one among all the matches that satisfy the requirements (e.g. best balance)
5. Compute the test statistic and its p-value on the matched data
This procedure is simple and computationally attractive, however it does not explicitly incorporate uncertainty arising from step 4: it is the case that values of the test statistic computed under different but equally good matches could differ. Since these matches all satisfy the quality requirements imposed by the analyst, they are indistinguishable from one another and so are the test statistic values resulting from them.

Each of the algorithms that could be used for matching could also be choosing among the different, equally good solutions that it finds arbitrarily or basing on arbitrary choices of the analyst that are commonly thought not to influence results. All of the algorithms in Table 1 suffer from this issue: sensitivity to seemingly arbitrary choices by the experimenter that go unaccounted for in the regular test statistic and p-value calculations. Consider the popular greedy Nearest-Neighbor matching procedure: the order in which observations are matched matters for the resulting assignment as observations that are matched earlier will have a greater pool of potential matches to choose from. Optimal assignment algorithms are subject to arbitrary implementation choices and could have to choose among multiple optima and near-optima, and most often will do so arbitrarily.

A natural question to ask at this point is whether the issue raised above affects existing results of matching studies or is actually mitigated in practice. Some evidence is provided by Morucci and Rudin (2018), who offer empirical evidence of the extent of this problem by replicating several social science studies that use matching and performing the same hypothesis tests with several popular matching algorithms, showing that hypothesis test results are influenced by choice of matching method.

2.2. Proposed Approach

We propose to use the range of test statistics produced by good matches as a test statistic itself. First we define a set of good assignments as $\mathcal{A}_{\text{good}} \subseteq \mathcal{A}$ where all $a \in \mathcal{A}_{\text{good}}$ obey quality constraints

| Algorithm                  | Advantage | Sensitivity to arbitrary choices |
|----------------------------|-----------|----------------------------------|
| Greedy NN-Matching         | Fast      | Yes (e.g: affected by arbitrary transformations of the data such as permutations) |
| Optimal Matching (Rosenbaum 1989) | Optimal   | Yes but Moderate (e.g: arbitrary choice among many possible nearly optimal solutions) |
| Optimal With Selection (Rosenbaum 2012) | Optimal for both groups | Yes (same problem as Optimal matching with larger decision space) |

Table 1 Some matching algorithms and their sensitivity to arbitrary choices. The phrase “larger decision space” is intended to mean that the number of decision variables is larger in the case of Optimal matching with selection, as it now involves treatment units as well.
besides those in $A$. $A_{\text{good}}$ intuitively represents the set of assignments arising from all reasonable matching methods, we refer to a match in $A_{\text{good}}$ as a good assignment. Then, the procedure is as follows:

**Robust Procedure:** Over all assignments in $A_{\text{good}}$, choose the two that respectively minimize and maximize the value of the test statistic of interest, and compute the maximal and minimal value of the test statistic:

$$
\psi^+ := \max_{a \in A_{\text{good}}} \psi_y(a), \quad \psi^- := \min_{a \in A_{\text{good}}} \psi_y(a).
$$

(4)

The range between $\psi^+$ and $\psi^-$ includes, by definition, all the possible values of $\psi$ that could be obtained on data $y$ by changing the matching assignment. In this sense, the pair $(\psi^+, \psi^-)$ is robust to analyst choice of matching method, and to arbitrary choices made by the matching algorithm itself. If both test statistics successfully reject the null hypothesis of interest, then it can be concluded that the null can be rejected robustly and independently of the chosen matching method.

Denoting the test statistic as a random variable $\Psi$, and the actual computed realization in the data with $\psi$, several kinds of robust p-values can be computed by the joint and marginal distributions of $(\Psi^+, \Psi^-)$ under any of the hypotheses of interest.

This robust approach is justified as follows: First, we do not want to consider how likely a certain assignment $a$ 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. Second, we do not want to consider statistics of the set of $A_{\text{good}}$, such as the average value of $\psi$ over the set of $A_{\text{good}}$. This is simply a special case of the previous point, where we assume that all good assignments are equally likely to be chosen, which is clearly not the case. Third, there is no sound statistical theory underpinning the idea that results averaged over a set of good matches should be more accurate than results from one single match. Fourth, the idea that our tests might lead to results that are too extreme is fundamentally inaccurate, as there is no reason why these results should be considered outlying or extreme even if most of the matches in $A_{\text{good}}$ lead to similar results that are different than those at $(\psi^+, \psi^-)$. 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 $\psi$
is circular logic: if a set of assignments is included in $A_{good}$, then it must be thought to be good
enough according to the analyst’s criteria. Excluding this assignment because it produces values of
$(\psi^+, \psi^-)$ that are too extreme, is tantamount to excluding observations from the dataset because
they are not producing the desired result.

2.3. Computing Null Distributions for Robust Test Statistics

Here we discuss two different conceptions of the role of matching in observational causal inference
that give rise of two different types of distributions for our test statistics.

**Matching as preprocessing (MaP)** This is the approach of [Ho et al. (2007)] and [Rosenbaum
(2010)], who assume that matching is a deterministic, preprocessing step. Our goal is to approxi-
mate a draw from a target distribution, $f(Y(1), Y(0), X)$, but we cannot estimate this distribution
as we never observe $Y(1)$ and $Y(0)$ jointly. Matching, together with assumptions on the form of
$f$, allows us to solve this problem by treating a sample of matched pairs as an approximate draw
from $f$. In this approach the matched data is then analyzed as if it were drew directly from $f$. Test
statistics obtained with any match in $A_{good}$ are treated as if computed on a sample from $f$: at
no point in this process does the unmatched data and the matching algorithm explicitly enter
the uncertainty computation. Because of the nature of this approach, the null distribution of the
test statistic $\psi$ is computed post-matching, and is always the same under any of the matches in
$A_{good}$. In this sense, the range of p-values produced in this way is a Hacking Interval (Coker et al.
2018), a non-probabilistic interval of summary statistics (such as p-values) that could be derived
from any manipulation of the data. Given that matching is a simple preprocessing step in this
framework, choice of matches in $A_{good}$ is always deterministic. Once we have chosen the set of
matches that we think best approximates a draw from $f$, then any test statistic $\Psi$ computed on it
has distribution independent of the matches.

**Matching as Estimator (MaE):** In this case matching is taken to fully be part of the data
generating process that leads to the observed value of the test statistic. This approach is taken by
[Abadie and Imbens (2006)]. In this case we still assume the matched data to represent a random
draw from our target distribution of interest, $f$, however we take into full account the initial
distribution that the unmatched data were drawn from. While in the case of MaP we assumed
that matched data comes from a certain statistical experiment, in this case we assume that data is
randomly drawn before matching. This implies that the distribution of the statistic is affected by
the matching procedure, and as such, standard error and p-value calculations are affected by the
matching procedure. For this problem we focus on the testing $H_{0 \text{sharp}}$ and $H_{0 \text{ATE}}$. For binary data, we will consider null distributions for $(\Psi^+, \Psi^-)$ under both hypotheses that take into consideration additional uncertainty stemming from matches being made post-treatment, as well as strategies for computing one and two-tailed p-values.

2.4. Constraining the Quality of the Matches

Determining the quality requirements that matches should satisfy to be usable is ultimately up to the experimenter, however there are several general types of constraints that matches in $A_{\text{good}}$ should most often obey. Let $dist(x, x')$ be a metric on the space of $x$, some of these constraints are:

- (Calipers) When $a(i) \neq \emptyset$ then $dist(x_t^{a(i)}, x_c^{a(i)}) \leq \epsilon$.
- (Covariate balance, mean of chosen treatment units similar to mean of control group) For all covariates $p$ we have:

$$\left| \frac{1}{M} \sum_{\{i: a(i) \neq \emptyset\}} x_{tp}^i - \frac{1}{M} \sum_{\{j: a(j) \neq \emptyset\}} x_{cp}^j \right| \leq \epsilon_p. \quad (5)$$

- (Maximizing the fitness of the matches) In general, one can optimize any measure of user-defined fitness for the assignment (including those defined above), and then constrain $A_{\text{good}}$ to include all other feasible assignments at or near that fitness level, by including the following constraints:

$$\text{Fitness}(a, \{x_t^i\}_{i=1}^{N_t^i}, \{x_c^i\}_{i=1}^{N_c^i}) \geq \text{Maxfit} - \epsilon,$$

where Maxfit is precomputed as: $\text{Maxfit} = \max_{a \in A} \text{Fitness}(a, \{x_t^i\}_{i=1}^{N_t^i}, \{x_c^i\}_{i=1}^{N_c^i})$. If one desires the range of results for all maximally fit pairs and no other pairs, $\epsilon$ can be set to 0.

In our examples we use calipers mostly for ease of notation, but replacing constraints with those above (or other constraints) is trivially simple using MIP software.

A special type of constraints are Exclusively Binning Constraints: these are constraints that can be expressed as a division of the data into strata. These strata are sets of observations, such that one unit belongs exclusively to one set, and if exclusively units within the same stratum are matched together, then resulting matches satisfy all the constraints. More formally, we define exclusively binning constraints as follows:

**Definition 2.** (Exclusively Binning Constraints) The constraints on $A_{\text{good}}$ are exclusively binning if there exists a partition $S$ of $\{1, \ldots, N\}$ such that:

$$\forall a \in A_{\text{good}} : (a(i) = j \implies i, j \in S),$$

$$\forall a \in A : (a(i) = j, \implies i, j \in S) \implies a \in A_{\text{good}},$$

for some subset $S \in S$. 
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]. In practice, several types of constraints on the quality of matches, chiefly balance, as defined in Equation (5), can be implemented as a coarsening choice on the covariates [Iacus et al. 2011], and as such as exclusively binning constraints.

In what follows, we provide special cases of the Robust Procedure for two specific hypothesis tests, McNemar’s test for binary outcomes and the z-test for real-valued outcomes. We outline strategies for computing these statistics as well as their distributions under the hypotheses of interest. We begin with McNemar’s test.

3. Robust McNemar’s Test

We begin with the problem of hypothesis testing for binary outcomes, that is \( Y(t) \in \{0, 1\} \). Let \( B(a) = \sum_{i=1}^{N^t} \sum_{j=1}^{N^c} a_{ij} y_i^t (1 - y_j^c) \) be the count of matched pairs in which the treated unit has outcome 1 and the control unit has outcome 0 under assignment \( a \), and let \( C(a) = \sum_{i=1}^{N^t} \sum_{j=1}^{N^c} a_{ij} y_j^c (1 - y_i^t) \) be the number of matched pairs in which the treated unit has outcome 0 and the control unit has outcome 1. We choose to use the following test statistic for both hypotheses:

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

We correct the denominator by adding one to the count of pairs, as it may happen that a matched set exists such that \( B(a) + C(a) = 0 \) in cases in which matches can be made in ways that allow that. This is the only difference between our formulation of the test statistic and its classical formulation (see Tamhane and Dunlop 2000). The robust test statistic can be defined as the pair:

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

In what follows, we outline two strategies to compute \((\chi^+, \chi^-)\) subject to the constraints that define \(\mathcal{A}_{good}\), one when the constraints are very general, and one when the constraints are exclusively binning. The first relies on common mixed integer programming techniques and is flexible enough to be implemented with any common MIP solver. The second exploits the form of the exclusively binning constraints to solve the problem with a fast polynomial time algorithm. Under MaP both algorithms can be used to test all of the hypotheses above, while under MaE, we provide exact expressions for the distribution of \((\chi^+, \chi^-)\) under \(\mathbb{H}_0^{\text{sharp}}\) in Section 3.3 and \(\mathbb{H}_0^{\text{ATE}}\) with a fixed number of treated units in Section 3.4.
3.1. Optimizing McNemar’s statistic with general constraints

In this section we give an Integer Linear Program (ILP) formulation that optimizes $\chi$ with a predefined number of matches $M$. This formulation can be adapted for testing either $H_{ATE}^0$ or $H_{ATT}^0$ under MaP by changing the number of matches made, and can be used to test $H_{sharp}^0$ with any number of matches.

One can show that the number of pairs where the same outcome is realized for treatment and control is irrelevant, so we allow it to be chosen arbitrarily, with no constraints or variables defining it. The total number of pairs $M$ is also not relevant for this test. Therefore, we choose only the total number of untied responses ($B$ and $C$-pairs) $m$. The problem can be solved either for a specific $m$ or globally for all $m$ by looping over all possible values of $m$ until the problem becomes infeasible. In most practical data analysis scenarios the largest feasible $m$ is chosen.

Formulation 1: General ILP Formulation for McNemar’s Test

Maximize/Minimize

$$\chi(a) = \left[ \frac{B(a) - C(a) - 1}{\sqrt{m + 1}} \right]$$

subject to:

$$\sum_{i=1}^{N_t} \sum_{j=1}^{N_c} a_{ij} y_i^t (1 - y_j^c) = B(a)$$  \hspace{1cm} \text{(Total number of first type of discordant pairs)}  \hspace{1cm} (7)$$

$$\sum_{i=1}^{N_t} \sum_{j=1}^{N_c} a_{ij} y_j^c (1 - y_i^t) = C(a)$$  \hspace{1cm} \text{(Total number of second type of discordant pairs)}  \hspace{1cm} (8)$$

$$B(a) + C(a) = m$$  \hspace{1cm} \text{(Total number of discordant pairs)}  \hspace{1cm} (9)$$

$$\sum_{i=1}^{N_t} a_{ij} \leq 1 \hspace{1cm} \forall j \hspace{1cm} \text{(Match each control unit at most once)}  \hspace{1cm} (10)$$

$$\sum_{j=1}^{N_c} a_{ij} \leq 1 \hspace{1cm} \forall i \hspace{1cm} \text{(Match each treatment unit at most once)}  \hspace{1cm} (11)$$

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

(Additional user-defined match quality constraints.)  \hspace{1cm} (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, hypothesis $H_{ATT}^0$ can be tested by adding constraint $\sum_{i=1}^{N_t} \sum_{j=1}^{N_c} a_{ij} = N_t$, only in the case where $N_t \leq N_c$, that is matching all of the treatment units, while $H_{ATE}^0, H_{sharp}^0$ can be tested with any reasonable number of matches.
3.2. Robust McNemar’s Test With Exclusively Binning Constraints.

If the constraints on $A_{\text{good}}$ are exclusively binning, then $\chi$ can be optimized quickly and easily in both directions if the partition $S$ is constructed first. Before showing this, we introduce an IP formulation for McNemar’s test with exclusively binning constraints:

Formulation 2: IP formulation for McNemar’s test with exclusively binning constraints

Maximize/Minimize $\chi(a) = \left[ \frac{B(a) - C(a) - 1}{\sqrt{B(a) + C(a)} + 1} \right]$

subject to:

1. $\sum_{i=1}^{N_t} \sum_{j=1}^{N_c} a_{ij} y_t^i (1 - y_c^j) = B(a)$ (Total number of first type of discordant pairs) (14)
2. $\sum_{i=1}^{N_t} \sum_{j=1}^{N_c} a_{ij} y_c^j (1 - y_t^i) = C(a)$ (Total number of second type of discordant pairs) (15)
3. $\sum_{i=1}^{N_t} a_{ij} \leq 1 \forall j$ (Match each control unit at most once) (16)
4. $\sum_{i=1}^{N_c} a_{ij} \leq 1 \forall i$ (Match each treatment unit at most once) (17)
5. $\sum_{i\in S_l, j\in S_l} a_{ij} = \min(N_t^l, N_c^l)$ $\forall l$ (Make as many matches as possible) (18)
6. $a_{ij} \in \{0, 1\} \forall i, j$ (Defines binary variable $a_{ij}$) (19)

(Additional user-defined exclusively binning constraints.) (20)

Note that analysts can remove this part of the formulation and introduce a fixed number of matches as constraint in a way similar to Constraint (9) if desired.

This problem can be solved in linear time and without one of the canonical IP solution methods by using the fact that the strata defined by the exclusively binning constraints can each be optimized separately, once the direction of the resulting statistic is known. In stratum $S_l$ there will be $U_l$ treated units with outcome $Y_i = 1$, $\eta_l$ control units with outcome $Y_j = 1$, $V_l = N_t^l - U_l$ treatment units with outcome 0, and $\nu_l = N_c^l - \eta_l$ control units with outcome 1. This is summarized in Figure 1. To ensure that as many units as possible are matched, within each stratum we make exactly $M_l = \min(N_t^l, N_c^l)$ matches. We would like to make the matches within each stratum such that $\chi$ is either maximized or minimized. Algorithm 1 maximizes $\chi$, and an exactly symmetric algorithm for minimizing $\chi$ is given in Appendix A. As it is clear from their definitions, these algorithms do not require any of the conventional MIP solving techniques and as such are much
Algorithm 1: Maximize $\chi$ with Exclusively Binning Constraints

Data: Positive integer vectors $(U_1, \ldots, U_L)$, $(V_1, \ldots, V_L)$, $(\eta_1, \ldots, \eta_L)$, $(\nu_1, \ldots, \nu_L)$

Result: Maximal $\chi$ statistic value over all possible matches in $\mathcal{A}_{\text{good}}$.

1. for $l = 1, \ldots, L$ do
2. \hspace{1em} $M_l = \min(N_l^t, N_l^c)$
3. \hspace{1em} $U_l^+ := U_l - \max(U_l - N_l^c, 0)$
4. \hspace{1em} $V_l^+ := M_l - U_l^+$
5. \hspace{1em} $\eta_l^+ := \max(\eta_l - \max(N_l^c - N_l^t, 0), 0)$
6. \hspace{1em} $\nu_l^+ := M_l - \eta_l^+$
7. end
8. $TE = \sum_{l=1}^{L} U_l^+ - \eta_l^+$
9. if $TE \geq 0$ then
10. \hspace{1em} for $l = 1, \ldots, L$ do
11. \hspace{2em} $A_l^+ := \min(U_l^+, \eta_l^+)$
12. \hspace{2em} $D_l^+ := \min(\nu_l^+, V_l^+)$
13. \hspace{2em} $B_l^+ := \min(U_l^+ - A_l^+, \nu_l^+ - D_l^+)$
14. \hspace{2em} $C_l^+ := \min(\eta_l^+ - A_l^+, V_l^+ - D_l^+)$
15. \hspace{1em} end
16. end
17. else
18. \hspace{1em} for $l = 1, \ldots, L$ do
19. \hspace{2em} $B_l^+ := \min(U_l^+, \nu_l^+)$
20. \hspace{2em} $C_l^+ := \min(\eta_l^+, V_l^+)$
21. \hspace{2em} $A_l^+ := \min(U_l^+ - B_l^+, \eta_l^+ - C_l^+)$
22. \hspace{2em} $D_l^+ := \min(\nu_l^+ - B_l^+, V_l^+ - C_l^+)$
23. \hspace{1em} end
24. end
25. return \(\frac{\sum_{l=1}^{L} B_l^+ - C_l^+ - 1}{\sqrt{\sum_{l=1}^{L} B_l^+ + C_l^+ + 1}}\)

faster: the running time of Algorithm 1 is clearly linear in $N$, the number of units. This constitutes a substantial speed up over solving the problem with a regular MIP solver. The following theorem states the correctness of the algorithm for solving Formulation 2:
Theorem 1. (Correctness of Algorithm 1) Algorithm 1 globally solves Formulation 2 for the Maximal value of $\chi$.

Proof. See Appendix A.

In the following sections, we offer exact formulas and computational algorithms for the distribution of the test statistic values obtained with the results in this section, and when matches are considered as part of the statistic.

3.3. Randomization Distribution of $(\chi^+, \chi^-)$ Under Sharp Null Hypothesis and Exclusively Binning Constraints

Now that we can compute $(\chi^+, \chi^-)$ on a given dataset we must derive a formula for its distribution under $\mathbb{H}_0^{\text{sharp}}$, so that we can test the hypothesis in a MaE framework, that is when the matching procedure contributes to the uncertainty in the final value of the statistic of interest. Recall that $\mathbb{H}_0^{\text{sharp}}$ is Fisher’s sharp null hypothesis of no treatment effect, that is $Y_i(1) = Y_i(0) \forall i$. In this section we provide an exact expression for the joint null distribution of the test statistics under $\mathbb{H}_0^{\text{sharp}}$ for the case in which constraints are exclusively binning and $S$ can be constructed prior to matching.

As is standard in randomization inference, we hold potential outcomes for all of the units in our sample to be fixed and nonrandom: the only randomness in our calculation will stem from the treatment assignment distribution. Note that all of the probability statements that follow are made conditional on $X$ and, for simplicity, we omit the relevant conditional notation from them.

Our approach to constructing this randomization distribution is as follows: we assume that any set of units that could be matched together under a good assignment has the same probability of receiving the treatment, and that, after treatment is assigned, matches are made to compute $(\chi^+, \chi^-)$ with the program in Formulation 2.

Since all good matches are made within the same stratum, we can assume that units within the same stratum receive the treatment with equal probability. This implies that Assumption 1 becomes:

Assumption 3. (Stratified Conditional Ignorability) For all $l = 1, \ldots, L$: $\Pr(T_i = 1|i \in S_l) = e_l$, with $0 < e_l < 1$.

Under Assumption 3 and under $\mathbb{H}_0^{\text{sharp}}$, the data generating process for the count variables in each stratum is as follows:

1. $U_l|\mathbb{H}_0^{\text{sharp}} \overset{iid}{\sim} \text{Bin}(e_l, N^1_l)$ (number of treated units with outcome 1) \hspace{1cm} (21)
2. $\eta_l = N^1_l - U_l$ (number of control units with outcome 1) \hspace{1cm} (22)
3. $V_l|\mathbb{H}_0^{\text{sharp}} \overset{iid}{\sim} \text{Bin}(e_l, N^0_l)$ (number of treated units with outcome 0) \hspace{1cm} (23)
4. $\nu_l = N^0_l - V_l$ (number of control units with outcome 0) \hspace{1cm} (24)
\[ N^t_l = U_l + V_l \] (total number of treated units) \hspace{1cm} (25)
\[ N^c_l = \nu_l + \eta_l \] (total number of control units), \hspace{1cm} (26)

where \( N^t_l \) denotes the number of units in stratum \( l \) that have \( Y_i(1) = Y_i(0) = 1 \), \( N^0_l = N_l - N^t_l \) denotes the number of units in the same stratum with potential outcomes \( Y(1) = Y(0) = 0 \), and \( e_l = \Pr(T_i = 1|i \in S_l) \). Note that we now do not know how many units are treated and untreated in each stratum with certainty; this is equivalent to assuming that Bernoulli trials are performed within each stratum. In order to compute the distribution of \( (\chi^+, \chi^-) \) and an estimate of \( \hat{e}_l \) is needed for each stratum. Under assumption 3 this probability can be estimated consistently and in an unbiased way with \( \hat{e}_l = \frac{N^t_l}{N^t_l} \), or other equivalent estimators.

For convenience, we also introduce truncated versions of the variables above, which explicitly take into account the fact that exactly \( M_l \) matches are made within each stratum, letting \( G_l^- = \max(N^t_l - N^c_l, 0) \) and \( G_l^+ = \max(N^c_l - N^t_l, 0) \):

\[ M_l = \min(N^t_l, N^c_l) \] \hspace{1cm} (27)
\[ U_l^+ := U_l - \max(U_l - N^c_l, 0) \] \hspace{1cm} (28)
\[ U_l^- := \max(U_l - G_l^-, 0), 0) \] \hspace{1cm} (29)
\[ \eta_l^+ := \max(\eta_l - G_l^+, 0) \] \hspace{1cm} (30)
\[ \eta_l^- := \eta_l - \max(\eta_l - N^t_l, 0). \] \hspace{1cm} (31)

We now introduce a randomization distribution of the truncated variables in any stratum. Note that, by definition, these distributions are independent from one another, which is something that we will use in deriving an expression for the distribution of \( (\chi^+, \chi^-) \). As a reminder, \( M_l = \min(N^t_l, N^c_l) \) denotes the number of matches made within each stratum. The theorems below can appear notationally complex, however they lend themselves well to implementation for computing the distributions of interest.

**Lemma 1.** (*Randomization Distribution of Truncated Variables*) For \( l = 1, \ldots, L \) let \( (N^t_l, N^0_l, e_l) \) be fixed and known and let \( \mathbb{D} = (U_l, V_l, \eta_l, \nu_l, N^t_l, N^c_l) \) be drawn from the data generating process of equations (25) - (26). Let \( a, b, c, d, m \) be elements of \( \{0, \ldots, \min(N^0_l, N^t_l)\} \) and let \( m \in \{0, \ldots, N_l\} \).

The variables \( (U_l^+ \!, \! U_l^- \!, \! \eta_l^+ \!, \! \eta_l^- \!, \! M_l) \) have the following joint distribution:

\[
\Pr(U_l^+ = a, U_l^- = b, \eta_l^+ = c, \eta_l^- = d, M_l = m | \mathbb{H}_0^{\text{sharp}}) = \\
\sum_{j=0}^{N^t_l} \sum_{k=0}^{N^0_l} \Pr(U_l = j | \mathbb{H}_0^{\text{sharp}}) \Pr(V_l = k | \mathbb{H}_0^{\text{sharp}}) \Pr(U_l = j | U_l = k, V_l = k) \Pr(U_l^- = b | U_l = j, V_l = k) \times \Pr(\eta_l^+ = c | U_l = j, V_l = k) \Pr(\eta_l^- = d | U_l = j, V_l = k) \Pr(M_l = m | U_l = j, V_l = k),
\] \hspace{1cm} (32)
where:

\[ \mathbb{I}(U^+_i = a|U_i = j, V_i = k) = \begin{cases} 
1 & \text{if } j = N_i - k - a, \; k \leq N_i - 2a \\
1 & \text{if } j = a, \; k \leq N_i - 2a \\
0 & \text{otherwise.}
\end{cases} \] (33)

\[ \mathbb{I}(U^-_i = b|U_i = j, V_i = k) = \begin{cases} 
1 & \text{if } j = N_i - 2k - b, \; k \leq N_i - 2b \\
1 & \text{if } j = b, \; k \leq \frac{N_i - 2b}{2} \\
1 & \text{if } b = 0, \; j \geq \frac{N_i}{2} - k, \; k \geq \frac{N_i - j}{2} \\
1 & \text{if } b = 0, \; j = 0, \; k \leq \frac{N_i}{2} \\
0 & \text{otherwise.}
\end{cases} \] (34)

\[ \mathbb{I}(\eta^+_i = c|U_i = j, V_i = k) = \begin{cases} 
1 & \text{if } j = c + N_i^0 - 2k, \; k \geq \frac{N_i^0 - N_i^1 + 2c}{2} \\
1 & \text{if } j = N_i^1 - c, \; k \geq \frac{N_i^0 - N_i^1 + 2c}{2} \\
1 & \text{if } c = 0, \; j \leq N_i^0 - 2k, \; k \geq \frac{N_i - 2j}{2} \\
1 & \text{if } c = 0, \; j = N_i^1, \; k \geq \frac{N_i^0 - N_i^1}{2} \\
0 & \text{otherwise.}
\end{cases} \] (35)

\[ \mathbb{I}(\eta^-_i = d|U_i = j, V_i = k) = \begin{cases} 
1 & \text{if } j = d - k, \; k \geq 2d - N_i^1 \\
1 & \text{if } j = N_i^1 - d, \; k \geq 2d - N_i^1 \\
0 & \text{otherwise.}
\end{cases} \] (36)

\[ \mathbb{I}(M_i = m|U_i = j, V_i = k) = \begin{cases} 
1 & \text{if } j = m - k, \; m \leq N_i^1 \\
1 & \text{if } j = N_i - k - m, \; m \leq \frac{N_i}{2} \\
0 & \text{otherwise.}
\end{cases} \] (37)

and \( \Pr(U_i = j|\mathbb{H}^\text{sharp}_0) = \text{Bin}(j, N_i^1, c_i) \) and \( \Pr(V_i = k|\mathbb{H}^\text{sharp}_0) = \text{Bin}(k, N_i^0, c_i) \).

**Proof** See Appendix B.2

Note that, while the domain of the distribution above cannot be expressed by a simpler formula, it is simple and computationally fast to enumerate it for finite \( N_i \). Finally, the joint null distribution of \( (\chi^+, \chi^-) \) is given in the following theorem:

**Theorem 2.** (Randomization Distribution of \( (\chi^+, \chi^-) \)). For \( l = 1, \ldots, L \) let \( (N_i^1, N_i^0, c_i) \) be fixed and known and let data, \( D = (U_i, V_i, \eta_i, \nu_i, N_i^1, N_i^0) \), be drawn from the data generating process of Equations (21) - (26). Let \( \chi^+ \) be the maximum of Formulation 2 on \( D \), and let \( \chi^- \) be the minimum of the problem on the same variables. Let \( N^m = \sum_{i=1}^L \min(N_i^1, N_i^0) \) and define: \( X_{N^m} = \left\{ \frac{b-c-1}{\sqrt{b+c+1}} : b, c \in \{0, \ldots, N^m\} \right\} \), Additionally define:

\[ A(y) = \left\{ a = (a_1, \ldots, a_L) : \sum_{i=1}^L a_i = y, \; a_i \in \{0, \ldots, y\} \right\}, \] (38)

\[ B(y, s) = \left\{ b = (b_1, \ldots, b_L) : \sum_{i=1}^L b_i = \left( \frac{y-1}{s} \right)^2 - 1, \; b_i \in \left\{ 0, \ldots, \left( \frac{y-1}{s} \right)^2 - 1 \right\} \right\}, \] (39)

\[ C(x) = \left\{ c = (c_1, \ldots, c_L) : \sum_{i=1}^L c_i = x, \; c_i \in \{0, \ldots, x\} \right\}, \] (40)
Let $H$ be the Cartesian product of the sets above, such that each element of $H(x, y, r, s)$ be the Cartesian product of the sets above, such that each element of $H(x, y, r, s)$ is a 4-tuple of vectors: $(a, b, c, d)$. Let $N^1 = \sum_{i=1}^L N_i^1$; the pmf of $(\chi^+, \chi^-)$, for two values $(s, r) \in \mathcal{X}_N^m \times \mathcal{X}_N^m$ is:

\[
\Pr(\chi^- = s, \chi^+ = r \mid \mathbb{I}_{m\text{har}}^r) = \sum_{x = -N^m}^{N^m} \sum_{y = -N^m}^{N^m} \sum_{(a, b, c, d) \in H(x, y, r, s)} \prod_{l=1}^L \begin{cases} 
    h_1(a_l, b_l, c_l, d_l) & \text{if } x < 1, y < 1 \\
    h_2(a_l, b_l, c_l, d_l) & \text{if } x \geq 1, y < 1 \\
    h_3(a_l, b_l, c_l, d_l) & \text{if } x \geq 1, y \geq 1 \\
    0 & \text{otherwise}
\end{cases}
\]

where:

\[
h_1(a_l, b_l, c_l, d_l) = \mathbb{I}(|a_l| = b_l) \sum_{m=0}^{N_l^1/2} \sum_{j=0}^{m} \{ \\
    \Pr\left(U^-_i = a_i + j, U^+_i = \frac{2m - d_i + c_i}{2}, \eta^-_i = j, \eta^+_i = \frac{2m - d_i - c_i}{2}, M_i = m \right) \\
    + \Pr\left(U^-_i = a_i + j, U^+_i = \frac{d_i + c_i}{2}, \eta^-_i = j, \eta^+_i = \frac{d_i - c_i}{2}, M_i = m \right) \mathbb{I}(m \neq d_i) \}
\]

\[
h_2(a_l, b_l, c_l, d_l) = \mathbb{I}(|a_l| = b_l) \mathbb{I}(|c_l| = d_l) \\
    + \sum_{m=0}^{N^1/2} \sum_{j=0}^{m} \sum_{k=0}^{m} \{ \Pr(\chi^+ = a_l + k, \eta^-_i = j, \eta^+_i = k, M_i = m) \}
\]

\[
h_3(a_l, b_l, c_l, d_l) = \mathbb{I}(|c_l| = d_l) \sum_{m=0}^{N_l^1/2} \sum_{j=0}^{m} \{ \\
    \Pr\left(U^-_i = \frac{2m - b_i + a_i}{2}, U^+_i = c_l + j, \eta^-_i = \frac{2m - b_i - a_i}{2}, M_i = m \right) \\
    + \Pr\left(U^-_i = \frac{b_i + a_i}{2}, U^+_i = c_l + j, \eta^-_i = \frac{b_i - a_i}{2}, \eta^+_i = j, M_i = m \right) \mathbb{I}(m \neq b_i) \}
\]

Proof. See Appendix B.3.

Note that the exact form of the probabilities in $h_1, h_2, h_3$ is given by Lemma 1. The distribution is essentially a product of many binomial distributions truncated to be defined only on portions of their domains that include possible ranges allowed by the constraints on $A_{\text{good}}$.

There are several possibilities for computing a lookup table for the distribution in Theorem 2 for a given dataset: a naïve procedure based on brute force enumeration of the quantities in Equation (42) has computational complexity $O(2^{N^m} + (N^m)^2)$. The main bottleneck is having to compute the sets $A, B, C, D$ and their Cartesian product, $H$. Computing these sets alone is a version of the subset sum problem and as such NP-hard and solvable, at best, in pseudo-polynomial time using
dynamic programming algorithms. Even with a better method for computing these sets, computing their Cartesian product takes \( O(2^{N^m} + (N^m)^2) \) time, and each value within this set needs to then be iterated over \( N^m \) times at worst.

Fortunately, the complexity of computing this distribution can be significantly reduced by noting that the sum over the set is simply the convolution of all the independent distributions of the four parts of the range in each stratum. Recall that the distributions of matched pairs in each stratum are independent of each other because of SUTVA, and because optimization of \( \chi \) is performed almost independently in each stratum (for more details on this point see Appendix A). Since the calculation is a convolution of independent discrete probabilities, it can be sped up with multidimensional Fast Fourier Transforms. The basic row-column mDFT algorithm reduces the complexity of having to generate \( H(x, y, r, z) \) to \( O((N^m)^4 \log(N^m)^4) \). This, coupled with simple enumeration to create lookup tables for distributions of \( g_1, g_2, g_3 \) in each stratum, as well as the final range distribution yields an algorithm that creates a probability table for the null distribution of \( (\chi^+, \chi^-) \) in \( O((N^m)^4 \log(N^m)^4) \). This implies that the worst-case time of computing the distribution in Theorem 2 is polynomial.

We use the theorems in this section to compute p-values under MaE for \( \chi^+ \) and \( \chi^- \). Procedures to derive these p-values from the distributions given in this and in the following sections are detailed in Section 3.5. Before turning to these procedures, we introduce a formula for the distribution of \( (\chi^+, \chi^-) \) under \( H_0^{ATE} \) in the following section.

### 3.4. Distribution of \( (\chi^+, \chi^-) \) Under \( H_0^{ATE} \) and Exclusively Binning Constraints

In this section, we seek a formulation for \( \Pr(\chi^- = s, \chi^+ = r | H_0^{ATE}) \), the distribution of robust McNemar’s statistics under exclusively binning constraints, and a fixed number of treated units, \( N^*_t \) in each stratum. This distribution is needed to test \( H_0^{ATE} \), which differs from the hypothesis studied before because it allows potential outcomes to be treated as random quantities, and is weaker in that it requires treated and control outcomes to be equal only on average. As before, all the probabilistic statements that follow are conditional on \( X \).

In order to compute exact distributions for \( (\chi^+, \chi^-) \), we must require some kind of consistency in the behavior of \( Y \) for units in the same stratum. This is explicitly stated in the following assumption:

**Assumption 4. (Distribution of \( Y \) (Discrete))** Let \( S_i \in S \), for all units \( i \in S_i \) we assume: \( \Pr(Y_i = 1 | T_i = 1, i \in S_i) = p_t^i \) and \( \Pr(Y_i = 1 | T_i = 0, i \in S_i) = p_c^i \).

This is a common assumption in matching and it is satisfied fully only when units in the same stratum have the same exact values of \( X \). It is simply equivalent to stating that potential outcomes
for all units in stratum $i$ come from the same distribution. This formulation can, in principle, accept two different probabilities for treatment and control groups, and does not require conditional ignorability of treatment assignment to be used. This distribution can be used to compute a p-value under $H_0^{ATE}$ by requiring: $p_l^t = p_l^c = p_l$ for all $l$. In order for this null distribution to correctly estimate the conditional null hypothesis Assumption 3 must hold. Note that this assumption implies $H_0$ for all units in stratum $i$.

Obtain $H_0$ value under ignorability of treatment assignment to be used. This distribution can be used to compute a p-value for two different probabilities for treatment and control groups, and does not require conditional ignorability of treatment assignment to be used. This distribution can be used to compute a p-value under $H_0^{ATE}$ by requiring: $p_l^t = p_l^c = p_l$ for all $l$. In order for this null distribution to correctly estimate the conditional null hypothesis Assumption 4 must hold. Note that this assumption implies $H_0$ for all units in stratum $i$.

An estimate of $p_l^t$ and $p_l^c$ can be produced easily under Assumption 4 only. It can be done in an unbiased and consistent way with $\hat{p}_l^t = \frac{\sum_{i \in S_l} w_l^t}{N_l}$, or any other estimator of choice. When wanting to test the null hypothesis that $p_l^t = p_l^c$ one can use $\hat{p}_l^t = \hat{p}_l^c = \frac{\sum_{i \in S_l} w_l^t + \sum_{j \in S_l} w_j^c}{N_l}$.

Recall that exactly $M_i = \min(N_l^t, N_l^c)$ matches are made, with $M = \sum_{l=1}^L M_i$. Now denote with $N_l^t$ the number of treated units in stratum $S_l$ and with $N_l^c$ the number of control units in that same stratum. For each stratum, $l = 1, \ldots, L$, under Assumption 4, the data are generated as follows:

$$U_l | A_4 \xleftarrow{iid} \text{Bin}(p_l^t, N_l^t)$$

$$V_l = N_l^t - U_l$$

$$\eta_l | A_4 \xleftarrow{iid} \text{Bin}(p_l^c, N_l^c)$$

$$\nu_l = N_l^c - \eta_l.$$  

The joint distribution of the truncated variables introduced in Equations (28)-(31) under this DGP is given in the following lemma:

**Lemma 2.** (Distribution of Truncated Variables under Assumption 4.) For all $l = 1, \ldots, L$, let $(N_l^t, N_l^c, p_l^t, p_l^c)$ be fixed and known, and let $G_l^+ = \max(N_l^c - N_l^t, 0)$, $G_l^- = \max(N_l^t - N_l^c, 0)$. Let data $\mathbb{D} = \{(U_l, V_l, \eta_l, \nu_l)\}_{l=1}^L$ be drawn according to Equations (46)-(49). The truncated variables have the following joint distributions:

$$\Pr(U_l^- = a, U_l^+ = b) = \begin{cases} 
\Pr(U_l = b) & \text{if } b < \min(N_l^c, G_l^+) + 1, \ a = 0 \\
\Pr(U_l = b) & \text{if } b = a + G_l^-, \ 0 < a < N_l^c - G_l^- \\
\Pr(N_l^c \leq U_l \leq G_l^+) & \text{if } b = N_l^c, \ a = 0 \\
\Pr(U_l = a + G_l^-) & \text{if } b = N_l^c, \ a \geq \max(N_l^c - G_l^-, 1) \\
0 & \text{otherwise.}
\end{cases}$$  

(50)
Pr(\eta^-_i = a, \eta^+_i = b) = \begin{cases} 
\Pr(\eta_i = a) & \text{if } a < \min(N^+_i, G^+_i + 1), \ b = 0 \\
\Pr(\eta_i = a) & \text{if } a = b + G^+_i, \ 0 < b < N^+_i - G^+_i \\
\Pr(N^+_i \leq \eta_i \leq G^+_i) & \text{if } a = N^+_i, \ b = 0 \\
\Pr(\eta_i = b + G^+_i) & \text{if } a = N^+_i, \ b \geq \max(N^+_i - G^+_i, 1) \\
0 & \text{otherwise.} 
\end{cases}
(51)

where \Pr(U_i = x|A_i) = \text{Bin}(x, p^+_i, N^+_i), \ \Pr(x \leq U_i \leq y|A_i) = \sum_{z=x}^{y} \text{Bin}(z, p^+_i, N^+_i), \ \Pr(\eta_i = x|A_i) = \text{Bin}(x, p^+_i, N^+_i), \ \text{and } \Pr(x \leq \eta_i \leq y|A_i) = \sum_{z=x}^{y} \text{Bin}(z, p^+_i, N^+_i).

Proof See Appendix C.1

Using Lemma 3 these distributions are easy to enumerate and to construct a lookup table for. They are also clearly symmetrical and their form implies that \Pr(U^-_i = a, U^+_i = b, \eta^-_i = c, \eta^+_i = d) = \Pr(U^-_i = a, U^+_i = b) \Pr(\eta^-_i = c, \eta^+_i = d). This fact is going to allow us to derive an expression for the distribution of \((\chi^+, \chi^-)\), given in the following theorem.

**Theorem 3. (Distribution of \((\chi^+, \chi^-)\) under Assumption 2)** For all \(l = 1, \ldots, L\), let \((N^+_i, N^-_i, p^+_i, p^-_i)\) be fixed and known, and let \(M_l = \min(N^+_i, N^-_i)\), \(M = \sum_{l=1}^{L} M_l\). Let data \(D = \{(U_i, V_i, \eta_i, \nu_i)\}_{i=1}^{L}\) be drawn according to Equations (46)–(49). Let \(\chi^+\) be the maximum of Formulation 2 on \(D\) and let \(\chi^-\) be the minimum of Formulation 2 also on \(D\). Let \(X_M := \left\{ \frac{k-\nu-1}{\sqrt{M+1}} : b, c \in \{0, \ldots, M\} \right\}\). Additionally, let \(A(y), B(y, s), C(x), D(x, r), H(x, y, r, s)\) be defined as in Theorem 2. The pmf of \((\chi^+, \chi^-)\), under Assumption 4 for two values \(s, r \in X_M\) is given by:

\[
\Pr(\chi^- = s, \chi^+ = r|X, A_4) = \sum_{x=-M}^{M} \sum_{y=-M}^{M} \sum_{(a, b, c, d) \in H(x, y, r, z)} \prod_{l=1}^{L} \left\{ g_1(a_l, b_l, c_l, d_l) \quad \text{if } x < 1, y < 1 \\
g_2(a_l, b_l, c_l, d_l) \quad \text{if } x \geq 1, y < 1 \\
g_3(a_l, b_l, c_l, d_l) \quad \text{if } x \geq 1, y \geq 1 \\
0 \quad \text{otherwise,} 
\right. 
\]
(52)

where:

\[
g_1(a_l, b_l, c_l, d_l) = \mathbb{I}(|a_l| = b_l) \sum_{j=0}^{M_l} \left[ \Pr\left(U^-_i = a_l + j, U^+_i = \frac{2M_l - d_l + c_l}{2}\right) \Pr\left(\eta^-_i = j, \eta^+_i = \frac{2M_l - d_l - c_l}{2}\right) \\
+ \Pr\left(U^-_i = a_l + j, U^+_i = \frac{d_l + c_l}{2}\right) \Pr\left(\eta^-_i = j, \eta^+_i = \frac{d_l - c_l}{2}\right) \mathbb{I}(M_l \neq d_l) \right] 
\]
(53)

\[
g_2(a_l, b_l, c_l, d_l) = \mathbb{I}(|a_l| = b_l) \mathbb{I}(|c_l| = d_l) \sum_{j=0}^{M_l} \sum_{k=0}^{M_l} \Pr(U^-_i = a_l + j, U^+_i = c_l + k) \Pr(\eta^-_i = j, \eta^+_i = k) 
\]
(54)

\[
g_3(a_l, b_l, c_l, d_l) = \mathbb{I}(|c_l| = d_l) \sum_{k=0}^{M_l} \left[ \Pr\left(U^-_i = \frac{2M_l - b_l + a_l}{2}, U^+_i = c_l + k\right) \Pr\left(\eta^-_i = \frac{2M_l - b_l - a_l}{2}, \eta^+_i = k\right) \\
+ \Pr\left(U^-_i = \frac{b_l + a_l}{2}, U^+_i = c_l + k\right) \Pr\left(\eta^-_i = \frac{b_l - a_l}{2}, \eta^+_i = k\right) \mathbb{I}(M_l \neq b_l) \right] 
\]
(55)

The probabilities \(\Pr(U^-_i = a, U^+_i = b)\) and \(\Pr(\eta^-_i = a, U^+_i = b)\) are given in Lemma 2 and depend on \(N^+_i, N^-_i, p^+_i, p^-_i\).
The pmf above can be computed in polynomial time for finite \( N \) with the same DFT scheme introduced in Section 3.3.

Once we have formulas for the distribution of \((\chi^+, \chi^-)\) under the null hypotheses of interest, we can then use them to compute p-values for the statistics introduced before. In the next section, we outline strategies to do so under both conceptions of matching.

3.5. Computing p-values for McNemar’s test

Here we discuss methods for computing p-values for all the hypotheses of interest under both conceptions of matching that we discuss in the paper.

\( H_0^{ATE} \) and \( H_0^{ATT} \) under MaP can be tested by approximating the p-value of \((\chi^+, \chi^-)\) using the common normal asymptotic approximation used for the regular McNemar’s statistic if the number of discordant \( B \) and \( C \) pairs is large enough, that is, letting \( \Phi \) denote the standard normal CDF:

\[
\text{p-val}_{MaP} = \min(2(1 - \Phi(\chi^+)), 2(1 - \Phi(\chi^-))),
\]

where the values of \((\chi^+, \chi^-)\) and the computation used to obtain them will differ depending on whether \( H_0^{ATE} \) or \( H_0^{ATT} \) is being tested.

\( H_0^{ATE} \) and \( H_0^{sharp} \) under MaE: The joint randomization distribution of \((\chi^+, \chi^-)\) under \( H_0^{sharp} \) given in Theorem 2. From this distribution, the marginal distribution of \( \chi^+ \) or \( \chi^- \) can be computed by summing over the other statistic:

\[
\Pr(\chi^+ = r | H_0^{sharp}) = \sum_{s \in \mathcal{X}_{Nm}} \Pr(\chi^- = s, \chi^+ = r | H_0^{sharp}),
\]

\[
\Pr(\chi^- = s | H_0^{sharp}) = \sum_{r \in \mathcal{X}_{Nm}} \Pr(\chi^- = s, \chi^+ = r | H_0^{sharp}).
\]

We recommend computing these distributions by first generating a lookup table for \( \Pr(\chi^-, \chi^+ | H_0^{sharp}) \) from Theorem 2 using the mDFT method described at the end of Section B.3 and then summing over values of the marginalized statistic explicitly using the table. Once these distributions are computed MaE p-values can be obtained for an observed value of \((\chi^+, \chi^-), (s, r)\) with:

\[
\text{p-val}_{MaE} = \min(2\Pr(\chi^+ \leq r | H_0^{sharp}), 2\Pr(\chi^- \geq s | H_0^{sharp})).
\] (56)

Only one direction of each statistic can be considered because, by definition, the other must be either smaller or larger. The procedure for obtaining a p-value under \( H_0^{ATE} \) is the same, except...
the distribution of \((\chi^+, \chi^-)\) to be employed is the one given in Theorem 3, for the case in which \(p^t_l = p^c_l\) for each \(l\).

\[ H^\text{sharp}_0 \text{ under MaP} \]: This test can be performed by obtaining \(a^+\) with the algorithms detailed in Section 3 and then simulating different treatment assignments with units matched together having the same probability of being assigned to treatment, computing the test statistic \(\chi\) after each iteration and storing the resulting value as a sample. After a sufficient number of iterations, the randomization distribution of \(\chi^+\) should be approximated well by the samples drawn. This entire process must then be repeated to obtain the distribution of \(\chi^-\) (For more details on post-matching randomization inference see Rosenbaum (2010)). P-values can then be computed with the formula in Eq. (56) starting from this distribution.

In this section we have introduced a version of McNemar’s test for binary outcomes that is robust to choice of matching method. We have given algorithms to compute our statistics as well as formulas and methods to compute null distributions and p-values for them under the two different conceptions of matching that we outlined before. In the next section, we introduce a robust version of the z-test for continuous outcomes.

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 0, when outcomes are real-valued, that is: \(Y(t) \in \mathbb{R}\). For this test we focus on MaP, that is the case in which matching is treated as a simple data preprocessing device. After computing the robust z-statistic, a p-value for it can be obtained with the canonical Normal asymptotic approximation. In what follows, \(M\) is the total number of pairs, and \(\hat{\sigma}\) is the sample standard deviation of the differences, \(y^t_i - y^c_{a(i)}\). The z-score is:

\[
z_y(a) = \frac{d_a \sqrt{M}}{\hat{\sigma}_a},
\]

where:

\[
d_a = \frac{1}{M} \sum_{i=1}^{N^t} \sum_{j=1}^{N^c} y^t_i - y^c_{a(i)}, \quad \text{and} \quad \hat{\sigma}_a = \sqrt{\frac{1}{M} \sum_{i=1}^{N^t} \sum_{j=1}^{N^c} (y^t_i - y^c_j)^2 \sigma_{ij} - \bar{d}_a^2}.
\]

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

\[
z^+ := \max_{a \in A_{\text{good}}} \frac{d_a \sqrt{M}}{\hat{\sigma}_a}, \quad z^- := \min_{a \in A_{\text{good}}} \frac{d_a \sqrt{M}}{\hat{\sigma}_a}, \tag{57}
\]
We provide general ILP formulations for computing the robust statistic under any kind of constraint on $A_{good}$ 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 (that we will discuss) can handle testing of $H_0^{ATE}$ and $H_0^{ATT}$ for MaP and $H_0^{sharp}$ for both MaP. For the purposes of testing under MaE, we propose the use of simulation methods for sampling $(z^+, z^-)$ under $H_0^{sharp}$.

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

Robust $z$-tests under general match quality constraints can be formulated as ILPs that can be solved with any of the common solvers. The $z$ statistic is clearly not linear in the decision variables (the matches). If one were to optimize it directly, a solution could be approximated using a MINLP (mixed-integer nonlinear programming solver) but guarantees on the optimality of the solution could not necessarily be made. 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 to 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 as follows:

$$\text{Maximize/Minimize}_a \sum_{i=1}^{N_t} \sum_{j=1}^{N_c} (y_i^t - y_j^c) a_{ij},$$

which is now linear. The quantity in $[58]$ is the estimated treatment effect. At the same time, we will limit the sum of squared differences term 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:

$$\sum_{i=1}^{N_t} \sum_{j=1}^{N_c} (y_i^t - y_j^c)^2 a_{ij} \leq b_l.$$

Putting this together, the new formulation is an ILP.

Formulation 3: ILP formulation for $z$-test

$$\text{Maximize/Minimize}_a \sum_{i=1}^{N_t} \sum_{j=1}^{N_c} (y_i^t - y_j^c) a_{ij} \quad (\text{Treatment effect})$$
subject to:

\[
\sum_{i=1}^{N_t} \sum_{j=1}^{N_c} (y_{it} - y_{cj})^2 a_{ij} \leq b_l \quad \text{(Upper bound on sample variance)} \tag{61}
\]

\[
\sum_{i=1}^{N_t} \sum_{j=1}^{N_c} a_{ij} = M \quad \text{(Choose } M \text{ pairs)} \tag{62}
\]

\[
\sum_{i=1}^{N_t} a_{ij} \leq 1 \quad \forall j \quad \text{(Match each control unit at most once)} \tag{63}
\]

\[
\sum_{j=1}^{N_c} a_{ij} \leq 1 \quad \forall i \quad \text{(Match each treatment unit at most once)} \tag{64}
\]

\[
a_{ij} \in \{0, 1\} \quad \forall i, j \quad \text{(Defines binary variable } a_{ij}) \tag{65}
\]

\[
\text{(Additional user-defined covariate balance constraints.)} \tag{66}
\]

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 \(H_0^{ATT}\) under full matching can be formulated by setting \(M = N_t\) in Formulation 3, which can also be modified to form the problem of choosing both treatment and control populations simultaneously, to handle the setting of Rosenbaum (2012). Here, the mean is taken over the same portion of the population as Rosenbaum (2012), which is the region of overlap between the control and treatment populations, removing extreme regions. This setting can be handled by looping over values of \(M\) until the program becomes infeasible.

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-score and p-value. If we denote the solution of the maximization problem as \(a_i\), where \(a_i\) is still also indexed by \(ij\), we will then be able to bound the value of \(z\). Shortly we will use Theorem 4 to prove bounds on the z-score as follows:

\[
\max_i \frac{\bar{d}_{a_i} \sqrt{M}}{\sqrt{\frac{1}{M} b_l - (d_{a_i})^2}} \leq \max_a z(a) \leq \max_i \frac{\bar{d}_{a_i} \sqrt{M}}{\sqrt{\frac{1}{M} b_{l-1} - (d_{a_i})^2}}, \tag{67}
\]

Using these bounds, can now formulate an algorithm to choose progressively finer meshes for \(b_l\) to maintain the guarantee on the quality of the solution, repeatedly solving the ILP Formulation.

Algorithm 2 works by first solving the ILP Formulation by relaxing (removing) the first constraint (upper bound on sample standard deviation), and using the resulting matches to compute the first upper bound on the standard deviation, \(b^n\). The algorithm then creates a coarse mesh \(b_1, \ldots, b_L\) where \(b_1 < b_l < b_L\). We want the interval \([b_1, b_L]\) to be wide enough to contain the true value of
Algorithm 2: Maximize $z$ with general constraints.

**Data:** Set of real vectors $\{(y_{t1}, \ldots, y_{tN_t^l})\}_{l=1}^L$ and $\{(y_{c1}, \ldots, y_{cN_c^l})\}_{l=1}^L$

Additional data parameters for optimization, such as covariates $\mathbf{D}$.

Set of constraints on Formulation 3: $\mathbf{W}$,

ILP Solver for Formulation 3, $F(\mathbf{W}, (\mathbf{D}), \mathbf{y})$

A Procedure to generate an increasingly finer, ordered sequence of reals $b_1, \ldots, b_L$ from lower and upper bounds, $G(Lb, Ub)$.

**Result:** $N_t^l \times N_c^l$ binary matrix of matches, $\mathbf{a}$.

1. Solve Formulation 3 by removing the upper bound constraint: $a_0 := F(\mathbf{W}, (\mathbf{D}), \mathbf{y})$;
2. $b_l := \sum_{i=1}^{N_t} \sum_{j=1}^{N_c} (y_{ti} - y_{cj})^2 a_{0ij}$;
3. $b = (b_1, \ldots, b_L) := G(b_1, B_L)$
4. while $\max_l \left\{ \frac{d_{a_l \sqrt{M}}}{\sqrt{b_{l-1} - (d_{a_l})^2}} - \max_l \left\{ \frac{d_{a_l \sqrt{M}}}{\sqrt{b_{l} - (d_{a_l})^2}} \right\} \right\} \leq \epsilon$ do
5. \hspace{1em} for $l = 1, \ldots, \mathbf{L}$ do Solve Formulation 3: $a_l := F(\mathbf{W}, b_l, \mathbf{y})$;
6. \hspace{1em} $b^u := \max_{l \in \{1, \ldots, L\}} \frac{d_{a_l \sqrt{M}}}{\sqrt{b_{l} - (d_{a_l})^2}}$
7. \hspace{1em} for $l = 2, 4, 6, \ldots, L$ do
8. \hspace{2em} if $\frac{d_{a_l \sqrt{M}}}{\sqrt{b_{l-1} - (d_{a_l})^2}} > b^u$ then $(b_{l1}, \ldots, b_{LL}) := G(b_{l-1}, b_l)$;
9. \hspace{1em} end
10. \hspace{1em} end
11. $b := (b_{11}, \ldots, b_{LL})$
12. return any one of the $a_l$.

$f_2(\mathbf{a}^*) := \sum_{i=1}^{N_t} \sum_{j=1}^{N_c} (y_{ti}^* - y_{cj}^*)^2 a^*_{ij}$, where $\mathbf{a}^* \in \text{arg max } z(\mathbf{a})$, 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 $b_l$ could be chosen evenly spaced, though they do not need to be. Note that the choice of $(b_1, \ldots, b_L)$ at each iteration does not affect the optimality of the solution, only the speed at which it is obtained. The algorithm then computes the solution to the ILP Formulation as well as upper and lower bounds for the solution using (67).

Correctness of the algorithm follows directly from optimality of the solutions at the bounds introduced before. This is guaranteed by the following:

**Theorem 4.** (Optimal Solution of ILP with Upper Bound on Variance) Consider the optimization problem

$$x^* \in \text{arg max } _x F(f_1(x), f_2(x)),$$
where \( f_1 \) and \( f_2 \) are real-valued functions of \( x \in X \), \( F \) is monotonically increasing in \( f_1 \) and monotonically decreasing in \( f_2 \). Assume we are given \([b_1, b_2, ..., b_l, ..., b_L]\) that span a wide enough range so that \( x^* \) obeys:

\[
b_{l^*-1} \leq f_2(x^*) \leq b_{l^*} \text{ for some } l^* \in \{1, ..., L\}.
\]

Define \( x_l \) as follows:

\[
x_l \in \text{arg max}_{x : f_2(x) \leq b_l} f_1(x),
\]

where the equality follows because \( F \) monotonically increases in \( f_1 \). Then

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

**Proof**  See Appendix E.

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 Formulations 1 and 2 is exactly of the form of Theorem 4, where

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

\[
f_2(a) = \sum_{i=1}^{N^t} \sum_{j=1}^{N^c} (y_i - y_j)^2a_{ij},
\]

and

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

The algorithm is correct because for each \( l \) we determine whether the interval \([b_{l-1}, b_l]\) can be excluded because it provably does not contain a \( f_2(a) \) value corresponding to the maximum value of \( z(a) \). In particular, we know from the bounds in (67) and from the theorem that if the upper bound on the objective for a particular \( b_{l'} \) is lower than all lower bounds for the optimal solution \( a^* \) 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

\[
\frac{d_{a_l} \sqrt{M}}{\sqrt{\frac{1}{M}b_{l-1} - (d_{a_l})^2}} < \max_l \frac{d_{a_l} \sqrt{M}}{\sqrt{\frac{1}{M}b_{l} - (d_{a_l})^2}}.
\]

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 using the previous procedure, thus creating finer and finer meshes, and the process repeats on these finer meshes until the desired tolerance (\( \epsilon \)) is achieved. The extra constraints on \( a \) in the
ILP Formulation are also compatible with Theorem 4. Thus, the bounds in (67) are direct results of Theorem 4 applied to the z statistic as an objective function.

P-values under MaP can be computed in the same way as for McNemar’s statistic outlined in Section 3.5. In the case of \( H_0^{ATE} \) and \( H_0^{ATT} \) under MaP, the standard normal asymptotic approximation holds for \((Z^+, Z^-)\) as well, and ranges of p-values can be produced in the way described before.

Now that we are equipped with two robust tests for matched data, we look at some practical problems: in the next section we apply our methods to data from existing observational studies.

5. Case Studies

In this section we apply our proposed methods to two real world datasets. We show that our test statistics can produce robust results in both datasets, which suggests that our methods can have wide practical applicability. Some of our results show that null hypotheses can be rejected robustly, while in some other cases they show that there isn’t enough evidence to reject the null so once the additional uncertainty introduced in the testing procedure by choice of matching method is accounted for by our robust tests.

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

In this case study we used GLOW (Global Longitudinal study of Osteoporosis in Women) data used in the study of Hosmer et al. (2013). Each data point represents a patient and the outcome is whether or not the person developed a bone fracture. The treatment is smoking. We match on several pre-treatment covariates: age, weight, height, and BMI. We first test \( H_{sharp}^0 \) and \( H_{ATE}^0 \) under MaE with the exclusively binning formulation of McNemar’s test, we then test \( H_{ATE}^0 \) under MaP with the general formulation.

To test \( H_{sharp}^0 \) and \( H_{ATE}^0 \) under MaE, we create strata by coarsening the matching covariates to be at most \( \epsilon \% \) of their full-sample standard deviation away from each other. This is equivalent to imposing a balance constraint of \( \epsilon \% \) of each covariate’s standard deviation (\( \sigma \)) in the program in Formulation 2, with the definition of balance given in Equation 5. Results are presented for choices of \( \epsilon \) ranging from 0 to 1 in intervals of 0.1 in Figure 2.

The top left panel in figure shows MaE p-values obtained with the methodology described in Section 3.5 for computed value of \((\chi^+, \chi^-)\), shown in the top right panel. Both hypotheses are rejected at the 0.05 threshold for \( \epsilon = 0.5, 0.8, 1.1 \). Matching is infeasible for values of \( \epsilon \) lower than 0.5. In general the figure shows that our tests are successful in rejecting the null hypotheses when
Figure 2 MaE results for GLOW data at different levels of balance. The top-left panel shows p-values, the top-right panel shows values of $(\chi^+, \chi^-)$, the bottom-left panel shows counts of matched units and the bottom-right panel shows counts of strata. All these quantities are plotted at multiple levels of required balance in the matches, represented by $\epsilon$. Values of $\epsilon$ range from 0 to 1 in intervals of 0.1. The striped shaded area on the left of each panel represents values of $\epsilon$ for which very few matches ($< 10$) were made. The dotted shaded area on the right of each panel represents a region where matches were highly imbalanced $\epsilon > 1.2 \sigma_p$.

there is better balance among the matched groups (smaller $\epsilon$), while the tests fail to reject at larger values of $\epsilon$: this is because the set of potential matches that achieve balance at most $\epsilon \sigma_p$ for each covariate is larger at larger values of $\epsilon$. Our test statistics correctly detect this possibility and account for the greater uncertainty in the matching procedure at these values. This is also confirmed by the values of $(\chi^+, \chi^-)$ in the top right panel: as the imbalance grows $\chi^+$ diverges farther away from $\chi^-$. In general, these results show that there is evidence to reject $H_{\text{ATE}}^0$ and $H_{\text{sharp}}^0$ when a sufficient number of good quality matches is made.

Figure 3 shows results for testing $H_{\text{ATE}}^0$ under MaP with the general program in Formulation 1, without binning constraints. We include a non-binning constraint in the formulation, by requiring that matched units $i, j$ respect $dist_{ij} \leq 0$, where $dist_{ij}$ is 0 if the sum of the differences of all the covariates is 6 or less and 1 otherwise. This figure lends more evidence to the fact that $H_{\text{ATE}}^0$ can be rejected robustly when the matches made induce 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 with any matching algorithm and set of hyperparameters used with it.

Figure 2 and 3 highlight different features of the matching methods that they employ, the first showing results for different levels of match quality and the second for a small range of matched units, and therefore are not directly comparable. We note, however, that there is correspondence between results for numbers of total matches between 30 and 40, and for matches of good quality (non-shaded area in Fig. 2). This suggests that p-values are generally low for both tests as long as a sufficient number of good-quality matches are made, lending evidence to the fact that the no-effect hypotheses can be rejected.

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

In our second case study, we used 2 years (2011-2012) of bike sharing data from the Capital Bike Sharing (CBS) system (see Fanaee-T and Gama 2014) from Washington DC comprised of 3,807,587 records over 731 days, from which we chose 247 treatment days and 463 control days according to the weather as follows: The control group consists of days with Weather 1: Clear, Few clouds, Partly cloudy, and Partly cloudy. The treatment group consists of days with Weather 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist. The covariates for matching are as follows: Season (Spring; Summer; Fall; Winter), Year (2011; 2012), Workday (No; Yes), Temperature (maximum 41 degree Celsius), Humidity (maximum 100 percent), Wind speed (maximum 67). The outcome is the total number of rental bikes. We computed distance between days as follows: \( \text{dist}_{ij} = 1 \) if covariates season, year and workday were the same, and the differences in temperature, humidity and wind speed are less or equal to 2, 5 and 5, respectively.
for treated unit $i$ and control unit $j$, 0 otherwise. Since the outcome variable is continuous, we focus on testing $H_0^{ATE}$ and $H_0^{ATT}$ under MaP by producing ranges of p-values with the method introduced in Section 4.

Figures 4 and 5 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 4 and minimization problem in Figure 5. 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 $M = 50$, the p-value for $Z^+$ was 0 and the p-value for $Z^-$ was 1, while both p-values become 1 when the number of matches is 110. The problem becomes infeasible for a larger number of matches. This illustrates that there is a lot of uncertainty associated with the choice of experimenter – one experimenter choosing 90 matched pairs can find a $P$-value of $\sim 0$ and declare a statistically significant difference while another experimenter can find a $P$-value of $\sim 1$ and declare the opposite. In this case it is truly unclear whether or not mist has an effect on the total number of rental bikes.

6. Conclusion and Discussion

Believing hypothesis test results conducted from matched pairs studies on observational data can be dangerous. 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 pairs. In one of our case studies, we showed that it is possible to construct matched pairs so that the treatment seems to be effective with high significance, and yet another set of matched pairs exists where the estimated treatment effect is completely insignificant. 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 theoretically justified because p-values obtained for the extrema must, by definition, include p-values obtained under any other possible good match.

For binary data, we have provided ways to compute robust test statistics and p-values under two different conceptions of matching, either as a simple preprocessing device (MaP), or as full part of the data generating process (MaE). In the first case we have given a general ILP that can obtain extreme match under arbitrary constraints. In the second case, we have defined a special category of constraints for which we provide linear-time algorithms to obtain extreme matches, as well as null distributions that take the matching process into account.

---

1 Optimization models for this case study have been implemented in AMPL [Fourer et al. 2002], and solved with the solver CPLEX [ILOG 2007]. The bike sharing and GLOW datasets are publicly available for the purpose of reproducibility. The reported solution time with a X64-based PC with Intel(R) Core(TM) i7-4790 CPU running at 3.60 GHz with 16 GB memory and mip gap of 0.001 to solve a single instance is less than 1 second for all the tests.
Figure 4  Upper and lower bounds for maximum $z$-test objective function value over a range of $b_i$ (Bike Sharing data, $N=30$), illustrating the optimum search range at various steps Algorithm 2. The final optimum is found in Panel 4c. The final value for the maximization problem is between 31.41 and 31.48.

For continuous data, we have provided a general algorithm to compute maximal and minimal $Z$-test values for arbitrary match quality constraints, as well as theoretical guarantees of the global optimality of the solution produced by our algorithm.

Our case studies highlight the practical utility of our algorithms, showing that they are capable to detect additional uncertainty introduced by the choice of matching method in experimental results. It is not hard to see why this uncertainty has to be taken into account when drawing scientific conclusions, and our proposed method represents a step in that direction.

**Acknowledgments**

The authors express their gratitude to the Natural Sciences and Engineering Research Council of Canada (NSERC) for their financial support of this research.
Figure 5 Upper and lower bounds for minimum z-test objective function value over a range of $b_l$ (Bike Sharing data, $N=30$), illustrating the optimum search range at various steps Algorithm 7 (see appendix). The final optimum is found in Panel 4c. The final value for the maximization problem is between -73.01 and -73.47.
References

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

Chen, D.-S., R.G. Batson, Y. Dang. 2011. *Applied Integer Programming: Modeling and Solution*. Wiley.

Coker, B., C. Rudin, G. King. 2018. A Theory of Statistical Inference for Ensuring the Robustness of Scientific Results. *ArXiv e-prints*.

Courtney, M.E., G.R. Cusick. 2010. Crime during the transition to adulthood: How youth fare as they leave out-of-home care in illinois, iowa, and wisconsin, 2002-2007. ICPSR27062-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Dehejia, R.H., S. Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**(448) 1053–1062.

Dehejia, R.H., S. Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* **84**(1) 151–161.

Fanæe-T, Hadi, Joao Gama. 2014. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence* **2**(2-3) 113–127.

Fourer, R., D.M. Gay, B.W. Kernighan. 2002. *Ampl: A modeling language for mathematical programming*. Duxbury Press, Cole Publishing Co.

Ho, Daniel E., Kosuke Imai, Gary King, Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15**(3 (Summer)) 199–236.

Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* **81**(396) 945–960.

Hosmer, D.W., S. Lemeshow, R.X. Sturdivant. 2013. *Applied Logistic Regression: Third Edition*.. John Wiley and Sons Inc.

Iacus, Stefano M, Gary King, Giuseppe Porro. 2011. Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association* **106**(493) 345–361.

Iacus, Stefano M, Gary King, Giuseppe Porro. 2012. Causal inference without balance checking: Coarsened exact matching. *Political analysis* **20**(1) 1–24.

ILOG. 2007. *Cplex 11.0 user’s manual*. ILOG, Inc.

Imai, Kosuke, Gary King, Elizabeth A Stuart. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)* **171**(2) 481–502.

Lalonde, R. 1986. Evaluating the econometric evaluations of training programs. *American Economic Review* **76**(4) 604–620.
Morgan, Stephen L., Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.

Morucci, Marco, Cynthia Rudin. 2018. Matching bounds: How choice of matching method affects treatment effect estimates and what to do about it. *Working Paper*.

Noor-E-Alam, M., A. Mah, J. Doucette. 2012. Integer linear programming models for grid-based light post location problem. *European Journal of Operational Research* **222**(1) 17–30.

Rosenbaum, Paul R. 2012. Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics* **21** 57–71.

Rosenbaum, P.R. 1989. Optimal matching for observational studies. *Journal of the American Statistical Association* **84** 1024–1032.

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

Rubin, D.B. 2007. The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine* **26**(1) 20–36.

Rubin, D.B. 2008. For objective causal inference, design trumps analysis. *Annals of Applied Statistics* **2**(3) 808–840.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **5**(66) 688–701.

Tamhane, A.C., D.D. Dunlop. 2000. *Statistics and Data Analysis*. Prentice Hall, New Jersey.

Winston, W.L., M. Venkataramanan. 2003. *Introduction to Mathematical Programming, (4th ed.).* Thomson (Chapter 9).

Wolsey, L.A. 1998. *Integer Programming*. Wiley-Interscience, Series in Discrete Mathematics and Optimization, Toronto.

Zubizarreta, J.R. 2012. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association* **107**(500) 1360–1371.

Zubizarreta, J.R., R.D. Paredes, P.R. Rosenbaum. 2014. Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics* **8**(1) 204–231.

Zubizarreta, J.R., D.S. Small, N.K. Goyal, S. Lorch, P.R. Rosenbaum. 2013. Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Annals of Applied Statistics* **7**(1) 25–50.
Appendix

A. Algorithms that maximize and minimize χ under Exclusively Binning Constraints

In this section we re-introduce Algorithm 1 together with a symmetric procedure for minimizing McNemar’s statistic in Formulation 2. We give explicit proofs of the correctness of the algorithms for solving Formulation 2 in Subsection A.1.

As a reminder, we are in the following situation: we have $N_t$ treated units and $N_c$ control units measured on $P$ features, $X$ that take value in a finite set, $X$. All these units also have an outcome $Y \in \{0,1\}$ and treatment $T \in \{0,1\}$. Since the constraints on the optimization problem are exclusively binning, then based on the values of $X$ we group the units into $L$ strata such that each stratum $S_1, \ldots, S_L$ contains $N_t^l$ and $N_c^l$ units. In each stratum there will be $U_l$ treated units with outcome $Y = 1$, $V_l$ control units with outcome 0, and $\eta_l = N_c^l - \eta_l$ control units with outcome 1. We then would like to create pairs of units within each stratum, such that each pair contains exactly one treated unit ($T_i = 1$) and one control unit ($T_j = 0$). Once we have created the pairs we compute:

- $A_l$: the number of matched pairs in stratum $l$ such that the treated unit has outcome 1 and the control unit has outcome 0. We refer to these pairs as $A$-pairs.
- $B_l$: the number of matched pairs in stratum $l$ such that the treated unit has outcome 1 and the control unit 0. We refer to these pairs as $B$-pairs.
- $C_l$, the number of matched pairs in stratum $l$ such that the treated unit has outcome 0 and the control unit 1. We refer to these pairs as $C$-pairs.
- $D_l$, the number of matched pairs in stratum $l$ such that both the treated and control units have outcome 0, referred to as $D$-pairs.

The following tables summarize the data and the statistics we are interested in:

| $T$ | $Y$ | $Y_c$ |
|-----|-----|-----|
| 1   | 1   | 1   |
| 0   | 0   | 0   |

| $U_l$ | $V_l$ | $\eta_l$ | $\nu_l$ |
|-------|-------|----------|---------|

| $l$ |
|-----|

We then take the sums of $B_l$ and $C_l$ across all strata to obtain $B = \sum_{l=1}^{L} B_l$ and $C = \sum_{l=1}^{L} C_l$. Finally, we use these two quantities to compute:

$$\chi = \frac{TE - 1}{\sqrt{SD + 1}} = \frac{B - C - 1}{\sqrt{B + C + 1}}.$$  

We would like to make the matches within each stratum such that $\chi$ is either maximized or minimized, assuming that we must match as many units as possible. Throughout the rest of this document we use a $+$ superscript to denote values of $A$, $B$, $C$, $D$ and $\chi$ output by maximizing $\chi$ and a $-$ superscript to denote the corresponding values obtained by minimizing it. We limit our analysis to the case in which the maximum number of feasible matches needs to be achieved, that is, exactly $M = \sum_{l=1}^{L} \min(N_t^l, N_c^l)$ must to be made:
Algorithm 3: ComputeMaximizedSD

Data: Positive integers: $U, V, \eta, \nu$

Result: Positive integers: $A, B, C, D$

1. Make $B := \min(U, \nu)$ B-pairs.
2. Make $C := \min(\eta, V)$ C-pairs.
3. Make $A := \min(U - B, \eta - C)$ A-pairs.
4. Make $D := \min(\nu - B, V - C)$ D-pairs.
5. Return $A, B, C, D$

Algorithm 4: ComputeMinimizedSD

Data: Positive integers: $U, V, \eta, \nu$

Result: Positive integers: $A, B, C, D$

1. Make $A := \min(U, \eta)$ A-pairs.
2. Make $D := \min(\nu, V)$ D-pairs.
3. Make $B := \min(U - A, \nu - D)$ B-pairs.
4. Make $C := \min(\eta - A, V - D)$ C-pairs.
5. Return $A, B, C, D$.

A.1. Correctness of the Optimization Algorithms

What follows is a proof of the algorithms’ correctness. The proof is structured into 4 different claims and a theorem equivalent to Theorem 4 following directly from these claims. Before introducing these claims it is useful to summarize the various ways in which units can be matched and unmatched to lead to different pair types: the cells in Table 2 show what pairs can be created by unmaking two other pairs and matching their units across. For example, if we unmake an $A$-pair and a $D$-pair, we are left with a treated and a control unit with $Y = 1$ and a treated and a control unit with $Y = 0$: if we match them across we obtain a $B$ and a $C$-pair. In the rest of this document, we refer to this operation of unmaking two pairs and matching across their units as exchanging pair one with pair two. Note also that, if the number of treated units to be matched equals the number of control units, and all units are matched in some way, the only way we can change those matches is by performing one or more of the operations detailed in the table. This is, of course, only possible if the required pairs are present among the existing matches for example, we cannot unmake a $A$-pair if there are none made already. For claims 1-3, assume that any two units can be matched together. Since we don’t explicitly consider the different strata in these claims, we omit the $l$ subscript from the notation.

Claim 1. Suppose we want to maximize $\chi$ subject to the constraint that we must make as many matches as possible, that is we must make $M = \min(N^t, N^c)$ matches, and suppose that $N^c > N^t$ so that $M = N^t$. Then it
Algorithm 5: Maximize $\chi$ with exclusively binning constraints.

**Data:** Positive integer vectors $(U_1, \ldots, U_L)$, $(V_1, \ldots, V_L)$, $(\eta_1, \ldots, \eta_L)$, $(\nu_1, \ldots, \nu_L)$

**Result:** Maximal $\chi$ statistic value

1. for $l = 1, \ldots, L$ do
2. $M_l := \min(N^c_l, N^t_l)$
3. $U^+_l := U_l - \max(U_l - N^c_l, 0)$
4. $V^+_l := M_l - U^+_l$
5. $\eta^+_l := \max(\eta_l - \max(N^c_l - N^t_l, 0), 0)$
6. $\nu^+_l := M_l - \eta^+_l$
7. end
8. $TE^+ = \sum_{l=1}^L U^+_l - \eta^+_l$
9. if $TE^+ \geq 1$ then
10. for $l = 1, \ldots, L$ do
11. $(A_l^+, B_l^+, C_l^+, D_l^+) := \text{ComputeMinimizedSD}(U^+_l, V^+_l, \eta^+_l, \nu^+_l)$
12. end
13. end
14. else
15. for $l = 1, \ldots, L$ do
16. $(A_l^+, B_l^+, C_l^+, D_l^+) := \text{ComputeMaximizedSD}(U^+_l, V^+_l, \eta^+_l, \nu^+_l)$
17. end
18. end
19. return $\chi^+ = \frac{\sum_{l=1}^L B^+_l - C^+_l - 1}{\sqrt{\sum_{l=1}^L B^+_l + C^+_l + 1}}$

is always optimal to leave unmatched $\min(N^c - N^t, \eta)$ control units with outcome 1 and $\max(N^c - N^t - \eta, 0)$ control units with outcome 0. Suppose instead that $N^t \geq N^c$, then it is always optimal to leave unmatched $\min(N^t - N^c, V_l)$ treated units with outcome 0 and $\max(U_l - N^t, 0)$ treated units with outcome 1.

**Proof.** We will show that, if $N^c > N^t$ and we must make $N^t$ matches, then it is optimal to first leave unmatched as many control units with outcome 1 as possible, that is $\min(N^c - N^t, \eta)$, and, if after these units have been excluded, there still are more control than treated units, to leave unmatched the remaining control units with outcome 0. Suppose initially that $N^c = N^t + 1$, fix $\chi = \frac{B-C}{\sqrt{B+C+2}}$ and assume that there are exactly $N^t - 1$ already matched treatment units and at least two leftover control units. Since we must make exactly $N^t$ matches we can only match the leftover treatment unit with one of the two controls.

Assume that the two control units are $u_1$ with outcome $Y_{u_1} = 1$ and $u_0$ with $Y_{u_0} = 0$. There are two possible scenarios: first, the currently unmatched treatment unit has outcome 1: if we match it with $u_0$ we get a $B$-pair with corresponding value of $\chi$:

$$\chi_{u_0} = \frac{B-C}{\sqrt{B+C+2}}$$
Algorithm 6: Minimize $\chi$ with exclusively binning constraints.

**Data:** Positive integer vectors $(U_1, \ldots, U_L)$, $(V_1, \ldots, V_L)$, $(\eta_1, \ldots, \eta_L)$, $(\nu_1, \ldots, \nu_L)$

**Result:** Maximal $\chi$ statistic value

1. for $l = 1, \ldots, L$
   2. $M_l := \min(N^t_l, N^c_l)$
   3. $U^-_l := \max(U_l - \max(N^t_l - N^c_l, 0), 0)$
   4. $V^-_l := M_l - U^-_l$
   5. $\eta^-_l := \eta_l - \max(\eta_l - N^t_l, 0)$
   6. $\nu^-_l := M_l - V^-_l$
3. end

8. $TE^- = \sum_{l=1}^L U^-_l - \eta^-_l$
9. if $TE^- \geq 1$ then
   10. for $l = 1, \ldots, L$
       11. $(A^-_l, B^-_l, C^-_l, D^-_l) := \text{ComputeMaximizedSD}(U^-_l, V^-_l, \eta^-_l, \nu^-_l)$
   12. end
13. end
14. else
   15. for $l = 1, \ldots, L$
       16. $(A^-_l, B^-_l, C^-_l, D^-_l) := \text{ComputeMinimizedSD}(U^-_l, V^-_l, \eta^-_l, \nu^-_l)$
   17. end
18. end
19. return $\chi^- = \frac{\sum_{l=1}^L B^-_l - C^-_l - 1}{\sqrt{\sum_{l=1}^L B^-_l + C^-_l + 1}}$

| A | B | C | D |
|---|---|---|---|
| A | A,A | A,B | A,C | B,C |
| B | A,B | B,B | A,D | B,D |
| C | A,C | A,D | C,C | C,D |
| D | B,C | B,D | C,D | D,D |

Table 2  What pairs can be created by exchanging matches. Cells are the resulting pairs when a pair in the left margin is exchanged with a pair in the top row.

if we match the treatment unit with $u_1$ we get a $A$-pair, and:

$$\chi_{u_1} = \frac{B - C - 1}{\sqrt{B + C + 1}}$$

With some algebra we can see that, $\chi_{u_0} \geq \chi_{u_1}$, which implies that we always gain more from matching a treated unit with outcome 1 to a control unit with outcome 0. In this case, leaving $u_1$ unmatched is the optimal choice.

Now suppose that the leftover treatment unit has outcome 0. Then, if we match it with $u_0$ we form a $D$-pair and get:

$$\chi_{u_0} = \frac{B - C - 1}{\sqrt{B + C + 1}}$$
no change from the initial $\chi$. If we match this treatment unit with $u_1$ we have formed a $C$-pair instead get:

$$\chi_{u_1} = \frac{B - C - 2}{\sqrt{B + C + 2}}.$$ 

again some algebra reveals that $\chi_{u_0} \geq \chi_{u_1}$: the value of $\chi$ is maximized by choosing to match the treatment unit with $u_0$ in this case as well. This shows that, when there is a choice of multiple control units to match with a treatment unit the control with outcome 1 should always be left out if we wish to maximize $\chi$.

We now show that, if $N^t = N^c + 1$ then it is always optimal to leave unmatched a treated unit with outcome 0 instead of one with outcome 1 if $\chi$ is to be maximized. Let there be one unmatched control unit and two candidate treatment units for it to be matched with, $u_1$ such that $Y^t_{u_1} = 1$ and $u_0$ such that $Y^t_{u_0} = 0$. Let the value of $\chi$ without those units be $\chi = \frac{B + C - 1}{\sqrt{B + C + 1}}$. If the unmatched control has outcome 1 and we match it to $u_0$ we end up with a $C$-pair and an updated value of $\chi$ equal to:

$$\chi_{u_0} = \frac{B - C - 2}{\sqrt{B + C + 2}}.$$ 

If we instead match the control unit with outcome 1 to $u_0$ we produce an $A$-pair, and the following value of $\chi$:

$$\chi_{u_1} = \chi.$$ 

Then, with some algebra we can see that $\chi_{u_0} \leq \chi_{u_1}$, implying that, in order to maximize $\chi$, the optimal strategy is to leave $u_0$ unmatched. If the control unit to be matched has outcome 0 and we match it with $u_0$ we get a $D$-pair and

$$\chi_{u_0} = \chi.$$ 

If we instead match the control unit with outcome 0 to $u_1$ we get a $B$-pair and:

$$\chi_{u_1} = \frac{B - C}{\sqrt{B + C + 2}}.$$ 

Again, with some algebra we see that $\chi_{u_1} \geq \chi_{u_0}$, implying that leaving $u_0$ unmatched is optimal in this case as well. This shows that leaving treated units with outcome 0 unmatched if $N^t = N^c + 1$ is the optimal strategy to maximize $\chi$.

For the case in which $N^t = N^c + k$ we can proceed by induction on the number of matched control units: assume inductively that the optimal choice at $k - 1$ is to leave unmatched a treated unit with outcome 0. At match $k$ we will have a value $\chi_k$ and one unmatched control unit and two candidate treatments. By the above, the value of $\chi_k$ that we would get by leaving the treated unit with outcome 0 unmatched is always larger than the one we would get by leaving the unit with outcome 1 unmatched. Because of this it is optimal to leave the treated unit with outcome 0 unmatched also for the $k^{th}$ match. This proves that, regardless of the difference between the number of treated and control units, it is optimal to leave as many treated units with outcome 0 as possible unmatched. If $N^t \geq N^c$, the largest possible number of treated units with outcome 0 that can be left unmatched is clearly $\min(N^t - N^c, V)$, either we exhaust the difference between $N^t$ and $N^c$ by not matching treatment units with outcome 0, or we exhaust all $V$ treated
units with outcome 0 and still have leftover treated units that don’t have a match in the control group. In this second case, we will have to make up this difference by leaving unmatched treated units with outcome 1 in excess: the precise amount of which is \( N^t - N^c - V = N^t - N^c - N^t + U = U - N^c \). A symmetrical argument shows that, if \( N^c = N^t + k \) it is always optimal to leave as many control units with outcome 1 as possible unmatched, before leaving units with outcome 0 unmatched, and that the largest possible amount of control units with outcome 1 that can be left unmatched in this case is \( \min(N^c - N^t, \eta) \), and the amount of control units with outcome 0 is \( N^c - N^t - \eta \), in case all \( \eta \) control units with outcome 1 are left unmatched without being able to exhaust the difference between unmatched units.

Finally, note that the proof for the case in which we wish to minimize \( \chi \) is exactly symmetrical to this one. □

**Claim 2.** If \( N^t = N^c = M \), and exactly \( M \) matches must be made, then \( B^+ - C^+ = B^- - C^- = U - \eta \) independently of how units are matched.

**Proof.** Let \( W \) be the matching in which all units are paired in a way such that \( B = \min(U, \nu) \) and \( C = \min(\eta, V) \). Since \( M = U + V = \eta + \nu \), then \( U \geq \nu \iff \eta \geq V \), so it must be that either: \( B = U, C = \eta \) or \( B = V, C = \nu \). By definition of \( V \) and \( \nu \) we have that in both cases: \( B - C = U - \eta \). Now we will prove that this equality must hold for any other match in which all units are matched. Let \( B \) be the number of \( B \)-pairs created with that matching and \( C \) the number of \( C \)-pairs. Consider any other match \( W' \neq W \) also satisfying the fact that all \( M \) treatment and control units are matched and let \( B' \) and \( C' \) be the counts of \( B \) and \( C \)-pairs generated by \( W' \). Since exactly the same number of units are matched in \( W \) and \( W' \) it must be that there exists some sequence of exchange operations that, if applied to \( W \) generates \( W' \). We now proceed by induction on \( k \), the number of operations applied to \( W \) to get to \( W' \), starting with \( k = 1 \). Note first, by Table 2 that the only operations that can alter the number of \( B \) and \( C \)-pairs are: exchanging an \( A \)-pair with a \( D \)-pair, obtaining a \( B \)-pair and a \( C \)-pair, and exchanging a \( B \)-pair with a \( C \)-pair, obtaining an \( A \)-pair and a \( D \)-pair. In the first case, we unmake an \( A \)-pair and a \( D \)-pair in \( W \) and use those units to make a \( B \) and a \( C \)-pair in \( W' \) so the respective counts are now \( B' = B + 1 \) and \( C' = C + 1 \), which implies that \( B' - C' = B - C \). In the other case we have \( B' = B - 1 \), \( C' = C - 1 \) and \( B' - C' = B - C \). Finally, suppose inductively that \( B^{(k)} - C^{(k)} = B - C \) after the \( k \)th exchange operation. Again, by Table 2 the only operations that can alter the counts of \( B \) and \( C \) are exactly the two discussed in the base case ; and by the same reasoning, we conclude that \( B^{(k+1)} - C^{(k+1)} = B - C \). □

**Claim 3.** If \( N^t = N^c = M \) and exactly \( M \) matches must be made, then algorithms 3 and 4 respectively make the matches that maximize and minimize \( B + C \).

**Proof.** Consider algorithm 4 first. We will only show the statement for this algorithm as the proof of the correctness of the other algorithm is exactly symmetrical to this. Let \( W^* \) be the match output by this
algorithm and let \( A^*, B^*, C^*, D^* \) be the respective pair counts under \( W^* \). By lines 1-4 of algorithm 4 we know that:

\[
\begin{align*}
A^* &= \min(U, \eta) \\
D^* &= \min(V, \nu) \\
B^* &= \min(U - A^*, \nu - D^*) \\
C^* &= \min(\eta - A^*, V - D^*)
\end{align*}
\]

By definition of \( U, V, \eta, \nu, A^*, D^* \) and \( B^*, C^* \) are the largest possible number of \( A \) and \( D \)-pairs that can be made with the given units. It can be seen from the definitions above that one of \( C^* \) or \( B^* \) will always be 0: in case \( A^* = U \) then \( U - A^* = 0 \) and \( B^* = \min(U - A^*, \nu - D^*) = 0 \). In case \( A^* = \eta \) we have \( \eta - A^* = 0 \) and \( C^* = \min(\eta - A^*, V - D^*) = 0 \) by consequence. Since exactly \( M \) units must be matched, and \( N^t = N^c = M \) by assumption, then the only operations allowed to change the matches from \( W^* \) are those in table 2 as there are no leftover units unmatched. Note that the only operation in the table that would allow for a decrease in \( B \) and \( C \) is exchanging \( B \) with \( C \), but this operation can never be performed on \( W^* \) as either it contains no \( C \)-pairs or no \( B \)-pairs. Then it must be that \( B^* \) and \( C^* \) are the smallest number of \( B \) and \( C \)-pairs that can be made with the existing units and thus that they minimize \( B + C \). □

A.1.1. Proof of Theorem 1

Proof. Consider first the problem of maximizing \( \chi \). The observations are divided into \( l \) strata such that \( N_i^t \) treated units and \( N_i^c \) control units are in each stratum: we can only match units that are in the same stratum. If we denote the count of \( B \)-pairs and \( C \)-pairs in stratum \( l \) with \( B_l \) and \( C_l \) then the objective function for the problem is

\[
\chi = \frac{TE}{\sqrt{SD}} = \frac{\sum_{i=1}^{L}(B_i - C_i) - 1}{\sqrt{\sum_{i=1}^{L}B_i + C_i + 1}}.
\]

As \( TE \) is a separable function of \( B_l \) and \( C_l \), optimizing the former equals maximizing each of the latter individually, the same is true for \( SD \). Note first, that, by the form of the objective function, we must make as many matches as possible. This number of matches is exactly \( M = \sum_{i=1}^{L} M_i = \sum_{i=1}^{L} \min(N_i^t, N_i^c) \) because, by constraints (10) and (11), each treatment and control unit can only be matched once. By Constraint (18) we know that exactly \( M_i = \min(N_i^t, N_i^c) \) matched pairs must be constructed in each stratum, therefore, units in excess of \( M_i \) must be discarded in each stratum. By Claim 1 if there are more control units to be matched than treated units in one stratum it is always optimal to leave unmatched the amounts of units described in Claim 1 in each stratum, independently of how matches are made in other strata by separability of \( TE \). The algorithm does this explicitly when defining \( U_l^+, V_l^+, \eta_l^+, \nu_l^+ \) at lines 2-5. These are updated counts of units to be matched, such that, for all \( l \), \( U_l^+ + V_l^+ = \eta_l^+ + \nu_l^+ = M_l \) by definition of these quantities, therefore, the count of treated and control units to be matched is equal in all strata and equal to \( M_l \). Note that definitions of the counts given in lines 2-5 of Algorithm 4 are precisely the initial number of each outcome-treatment pair minus the optimal amount of units to be left unmatched given in Claim 1. By Claim 2 we know that if the number of treated units to be matched is equal to the number of control units to be matched, as it is the case after line 5 of the algorithm, then \( TE = \sum_{i=1}^{L} U_i^+ - \eta_i^+ \) independently of how matches are made. Because
of this, $TE$ can be considered fixed at this point, and maximizing $\chi$ equates with maximizing $SD$ if $TE < 0$ and minimizing it if $TE \geq 0$: this is checked explicitly by the algorithm at line 8. Lastly, if the algorithm calls $\text{ComputeMinimizedSD}$ if $TE \geq 0$ and $\text{ComputeMaximizedSD}$ if $TE < 0$: these two procedures are shown to correctly maximize and minimize $B_l + C_l$ in each stratum by Claim 3. Note that $SD = \sum_{l=1}^{L} B_l + C_l$ is also separable in the strata, and therefore it can be optimized globally by separately optimizing $B_l + C_l$ in each stratum. Since $\sqrt{SD}$ is monotonic in $B_l + C_l$, we know that this quantity is also maximized or minimized in this way. This shows that Algorithm 5 globally maximizes $\chi$ over the set of allowed matches. Finally, note that constraints (7), and (8) are all not violated by definition of these quantities, and constraints (10) and (11) are not violated because no two units are matched more or than once. Finally, the additional exclusively binning constraints are obeyed by definition, as matches are made exclusively within each stratum. The proof of the correctness of Algorithm 6 is exactly symmetrical to this one: this symmetry is made apparent by the fact that running Algorithm 6 on the data is equivalent to flipping the treatment indicator and then running Algorithm 5 on the resulting data. □

B. Randomization Distribution of $(\chi^+, \chi^-)$ Under Exclusively Binning Constraints

In this section we use the algorithm above to derive a randomization distribution for $(\chi^+, \chi^-)$ under the null hypothesis of no treatment effect. This allows us to test $\mathbb{H}_0^{\text{sharp}}$ under two types of uncertainty: in the data itself, and in the choices made by the analyst in a MaE model. As a reminder, under $\mathbb{H}_0^{\text{sharp}}$, potential outcomes are assumed to be fixed and invariant between treatment regimes. Under exclusively binning constraints, the units are divided into $S_1, \ldots, S_l$ strata, and matches are allowed only in the same stratum.

Given $L$ different levels in which the covariates are grouped, and that there are $N_t^l$ treatment units and $N_c^l$ control units in each stratum $l$. From the two assumptions before it follows that the data are generated as follows:

$$U_l \mid \mathbb{H}_0^{\text{sharp}} \overset{iid}{\sim} \text{Bin}(e_l, N_t^l),$$

$$\eta_l = N_t^l - U_l,$$

$$V_l \mid \mathbb{H}_0^{\text{sharp}} \overset{iid}{\sim} \text{Bin}(e_l, N_c^l),$$

$$\nu_l = N_c^l - V_l,$$

$$N_t^l = U_l + V_l,$$

$$N_c^l = \nu_l + \eta_l,$$

As stated in the paper, all these quantities have interpretations in our matching framework. Specifically, within one stratum, I.E., at one level of $X$, called $x_l$: $U_l$ is the number of treated units with outcome 1, $V_l$ is the number of units with $T = 1$ and $Y = 0$, $\eta_l$ is the number of units with $T = 0$ and $Y = 1$ and $\nu_l$ is the number of units with $T = 0$ and $Y = 0$. The null hypothesis of no treatment effect is encoded in the fact that the distributions of $U_l$ and $\eta_l$ differ only in the number of trials and not in the probability of a success. Note now that these pair counts are random variables: the algorithms make pairs to purposefully obtain optimal values of $\chi$, which in turn depends on the random data. Recall that $B$ and represents the number of pairs
such that the treated unit has outcome 1 and the control unit has outcome 0 and \( C \) the total number of pairs in which the control unit has outcome 1 and the treated unit 0.

Recall that the test statistic is defined as follows:

\[
\chi^+ = \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} = \frac{B^+ - C^+ - 1}{\sqrt{B^+ + C^+ + 1}}, \quad \chi^- = \frac{TE^- - 1}{\sqrt{SD^- + 1}} = \frac{B^- - C^- - 1}{\sqrt{B^- + C^- + 1}},
\]

where \( B^+ \) is the count of matched pairs produced by algorithm 5 such that the treated unit has outcome 1 and the control unit in the pair has outcome 0, \( C^+ \) is the count of pairs where the opposite is true, and \( B^- \) and \( C^- \) are the analogues produced by the minimization algorithm. For convenience, we also introduce "truncated" versions of the variables above, letting \( G_i^- = \max(N_i - N_i^+, 0) \) and \( G_i^+ = \max(N_i^+ - N_i, 0) \):

\[
U_i^+ = U_i - \max(U_i - N_i^+, 0) \\
V_i^+ = M_i - U_i^+ \\
U_i^- = \max(U_i - G_i^-, 0), 0) \\
V_i^- = M_i - U_i^- \\
\eta_i^+ = \max(\eta_i - G_i^+, 0) \\
\nu_i^+ = M_i - \eta_i^+ \\
\eta_i^- = \eta_i - \max(\eta_i - N_i^-, 0) \\
\nu_i^- = M_i - \eta_i^- .
\]

These definitions correspond to those introduced at lines 2-5 of Algorithms 5 and 6. Figure 6 summarizes the variables in our framework as well as how the maximization and minimization algorithms operate.

---

**B.1. Simplified Representations for McNemar’s Statistic**

The following statement gives us simplified forms for the values of \( TE \) and \( SD \) output by the maximization and minimization algorithms respectively, we state it as a claim as it requires a simple yet nontrivial proof:
Claim 4. Let $TE^+$, $SD^+$ and $TE^−$ and $SD^−$ represent the values of the numerator and denominator of $\chi$ as defined in Eq. (68), and output by Algorithms 5 and 6 respectively. Let $M_i = \min(N_i^+, N_i^-)$ denote the number of matches that are made in each stratum. Then they can be written as follows:

$$TE^+ = \sum_{i=1}^{L} TE_i^+ = \sum_{i=1}^{L} U_i^+ - \eta_i^+$$

$$TE^- = \sum_{i=1}^{L} TE_i^- = \sum_{i=1}^{L} U_i^- - \eta_i^-$$

$$SD^+ = \begin{cases} S^+ = \sum_{i=1}^{L} [U_i^+ - \eta_i^+] & \text{if } TE^+ \geq 1 \\ R^+ = \sum_{i=1}^{L} M_i - [U_i^+ + \eta_i^+ - M_i] & \text{if } TE^+ < 1 \end{cases}$$

$$SD^- = \begin{cases} S^- = \sum_{i=1}^{L} [U_i^- - \eta_i^-] & \text{if } TE^- < 1 \\ R^- = \sum_{i=1}^{L} M_i - [U_i^- + \eta_i^- - M_i] & \text{if } TE^- \geq 1 \end{cases}$$

Proof. Note first that the definition of $TE^+$ and $TE^−$ follows directly from lines 3-5 and 8 of Algorithms 5 and 6 respectively. As for the definition of $SD$ we will prove only the claim for $SD^+$ as the proof for $SD^−$ is exactly symmetrical. Proving the claim implies showing that, denoting with $B^+$ and $C^+$ the count of $B$ and $C$ pairs that maximize $\chi$:

$$B^+ + C^+ = SD^+.$$ 

Now define:

$$SD_i^+ = \begin{cases} S_i^+ = |U_i^+ - \eta_i^+| & \text{if } TE^+ \geq 1 \\ R_i^+ = M_i - |U_i^+ + \eta_i^+ - M_i| & \text{if } TE^+ < 1 \end{cases}$$

It is clear from this definition that $SD^+ = \sum_{i=1}^{L} SD_i^+$, and since $SD_i^+$ has the same definition in all strata, it suffices to prove that $B_i^+ + C_i^+$ can be written as in Eq. (73) for one stratum to prove the equality in Eq. (71). The rest of the proof is concerned with establishing this result.

Consider the computation for $SD^+$ occurring in Alg. 5 the algorithm checks explicitly if $TE \geq 1$ at line 8 and, if true it calls Alg. 4 on inputs $(U_i^+, V_i^+, \eta_i^+, \nu_i^+)$ to generate the matched pair counts $A_i^+, B_i^+, C_i^+, D_i^+$, if false it generates the same quantities by calling Alg. 3 on the same inputs.

Consider now the case in which $TE^+ \geq 1$, and Algorithm 4 is called with $(U_i^+, V_i^+, \eta_i^+, \nu_i^+)$ as inputs. We know that the algorithm returns the following counts for $B$ and $C$ pairs:

$$B_i^3 = \min(U_i^+ - A, V_i^+ - D)$$

(By line 3 of Alg. 4)

$$= \min(U_i^+ - \min(U_i^+, \eta_i^+), \nu_i^+ - \min(\nu_i^+, V_i^+)),$$

(By lines 1 and 2 of Alg. 4) (74)

and:

$$C_i^3 = \min(\eta_i^+ - A, V_i^+ - D)$$

(By line 4 of Alg. 4)

$$= \min(\eta_i^+ - \min(U_i^+, \eta_i^+), V_i^+ - \min(\nu_i^+, V_i^+)),$$

(By lines 1 and 2 of Alg. 4) (75)
Case 1: \( U_i^+ \geq \eta_i^+ \).

Note first that, in this case:

\[
U_i^+ \geq \eta_i^+ \quad \text{(76)}
\]

\[
\implies M_i - \eta_i^+ \geq M_i - U_i^+ \quad \text{(Add } M_i \text{ to both sides) (77)}
\]

\[
\implies \nu_i^+ \geq V_i^+. \quad \text{(By definition of } \nu_i^+, V_i^+) \quad \text{(78)}
\]

Because of this we can see that:

\[
B_i \Theta + C_i \Theta = \min(U_i^+ - \min(U_i^+, \eta_i^+), \nu_i^+ - \min(\nu_i^+, V_i^+))
\]

\[
+ \min(\eta_i^+ - \min(U_i^+, \eta_i^+), V_i^+ - \min(\nu_i^+, V_i^+)) \quad \text{(By (74) and (75))}
\]

\[
= \min(U_i^+ - \eta_i^+, \nu_i^+ - V_i^+) + \min(\eta_i^+ - \eta_i^+, V_i^+ - V_i^+) \quad \text{(By assumption of this case and result in (78))}
\]

\[
= \min(U_i^+ - \eta_i^+, M_i - U_i^+ - (M_i - \eta_i^+)) \quad \text{(By definition of } V_i^+ \text{ and } \nu_i^+) \quad \text{(78)}
\]

\[
= U_i^+ - \eta_i^+
\]

\[
= |U_i^+ - \eta_i^+| \quad \text{(By assumption of this case)}
\]

\[
= S_i^+. \quad \text{(Definition of } S_i^+) \quad \text{(78)}
\]

Case 2: \( U_i^+ < \eta_i^+ \).

Note that in this case Eq. (78) implies that: \( \nu_i^+ < V_i^+ \). Because of this we have:

\[
B_i \Theta + C_i \Theta = \min(U_i^+ - \min(U_i^+, \eta_i^+), \nu_i^+ - \min(\nu_i^+, V_i^+))
\]

\[
+ \min(\eta_i^+ - \min(U_i^+, \eta_i^+), V_i^+ - \min(\nu_i^+, V_i^+)) \quad \text{(By (74) and (75))}
\]

\[
= \min(U_i^+ - U_i^+, \nu_i^+ - \nu_i^+) + \min(\eta_i^+ - U_i^+, V_i^+ - \nu_i^+) \quad \text{(By assumption of this case and result in (78))}
\]

\[
= \min(\eta_i^+ - U_i^+, M_i - U_i^+ - (M_i - \eta_i^+)) \quad \text{(By definition of } V_i^+ \text{ and } \nu_i^+) \quad \text{(78)}
\]

\[
= \eta_i^+ - U_i^+
\]

\[
= |U_i^+ - \eta_i^+| \quad \text{(By assumption of this case)}
\]

\[
= S_i^+. \quad \text{(Definition of } S_i^+) \quad \text{(78)}
\]

This shows \( SD_i^+ = S_i^+ \) in the case in which \( TE \geq 1 \), the first of Eq. (71).

The proof is similar for the case in which \( TE < 1 \). Now we must show that, in this case \( SD_i^+ = R^+ \). In this case Algorithm 3 is called with \( U_i^+, V_i^+, \eta_i^+, \nu_i^+ \) as inputs at line 16 of Algorithm 5. On those inputs, Algorithm 3 will return the following counts of \( B \) and \( C \) pairs:

\[
B_i \Theta = \min(U_i^+, \nu_i^+) \quad \text{(By line 1 of Alg. 3) (79)}
\]

and:

\[
C_i \Theta = \min(\eta_i^+, V_i^+). \quad \text{(By line 1 of Alg. 3) (80)}
\]
We proceed as before with two separate cases. **Case 1**: $U_i^+ \geq \nu_i^+$.

First note that:

$$\nu_i^+ = M_i - \eta_i^+ \quad \text{(By definition of } \nu_i^+)$$

$$\leq U_i^+ \quad \text{(By assumption of this case)}$$

$$\implies \eta_i^+ \geq M_i - U_i^+$$

$$\implies \eta_i^+ \geq V_i^+. \quad \text{(By definition of } V_i^+ \text{)} \quad (81)$$

Second, the assumption of this case also implies:

$$\nu_i^+ = M_i - \eta_i^+ \quad \text{(By definition of } \nu_i^+)$$

$$\leq U_i^+ \quad \text{(By assumption of this case)}$$

$$\implies U_i^+ + \eta_i^+ \geq M_i$$

$$\implies U_i^+ + \eta_i^+ - M_i \geq 0$$

$$\implies U_i^+ + \eta_i^+ - M_i = |U_i^+ + \eta_i^+ - M_i|. \quad \text{(Def. of absolute value)} \quad (82)$$

Putting these results together we obtain:

$$B^{(3)} + C^{(3)} = \min(U_i^+, \nu_i^+) + \min(\eta_i^+, V_i^+) \quad \text{(By } (79) \text{ and } (80)) \quad (83)$$

$$= \nu_i^+ + V_i^+ \quad \text{(By assumption of this case and } (81)) \quad (84)$$

$$= M_i - U_i^+ + M_i - \eta_i^+ \quad \text{(By definition of } V_i^+ \text{ and } \nu_i^+) \quad (85)$$

$$= M_i - (U_i^+ + \eta_i^+ - M_i) \quad (86)$$

$$= M_i - |U_i^+ + \eta_i^+ - M_i| \quad \text{(By } (82)) \quad (87)$$

$$= R_i^+. \quad \text{(Definition of } R_i^+) \quad (88)$$

**Case 2**: $U_i^+ < \nu_i^+$.

This assumption of this case together with $\nu_i^+ < V_i^+$. Note also that:

$$\nu_i^+ = M_i - \eta_i^+ \quad \text{(By By definition of } \nu_i^+)$$

$$> U_i^+ \quad \text{(By assumption of this case)}$$

$$\implies U_i^+ + \eta_i^+ - M_i < 0$$

$$\implies M_i - U_i^+ - \eta_i^+ > 0$$

$$\implies M_i - U_i^+ - \eta_i^+ = |M_i - U_i^+ - \eta_i^+|. \quad \text{(Def. of absolute value)} \quad (89)$$

With the two results above we obtain:

$$B^{(4)} + C^{(4)} = \min(U_i^+, \nu_i^+) + \min(\eta_i^+, V_i^+) \quad \text{(By } (79) \text{ and } (80)) \quad (90)$$

$$= U_i^+ + \eta_i^+ \quad \text{(By assumption of this case and } (81)) \quad (91)$$

$$= M_i - M_i + U_i^+ + \eta_i^+ \quad \text{(Add and subtract } M_i) \quad (92)$$
This proves that $SD_I^+ = R_I^+$ in the case in which $TE^+ < 1$, the second case in Equation (71). □

B.2. Proof of Lemma 1

Proof Throughout the proof we maintain Assumption 3 and all of the probability statements to follow are conditional on it holding. The proof is simple and follows by inspecting the definitions of the truncated variables. Note first that the form for $Pr(U_i^+ = a, U_i^- = b, \eta_i^+ = c, \eta_i^- = d, M_i = m|H_0^{sharp})$ given in Eq. (92) follows from the law of total probability, independence of $U_i$ and $V_i$ and conditional independence of $U_i^+, V_i^+, \eta_i^+, \eta_i^-, M_i$ given $U_i$ and $V_i$. It remains to show that the forms for the indicator functions in Equations (32)–(37) are those in the theorem. This can be done by inspecting the definitions of the truncated variables and by expanding them out into conditions on $U_i$ and $V_i$. In what follows we refer to the following definitions of the quantities employed, listed here with reference to where they are introduced in the paper:

\[
U_i \sim Bin(N_i^1, c_i) \quad \text{(Given in Eq. (21))}
\]

\[
\eta_i = N_i^1 - U_i \quad \text{(Given in Eq. (22))}
\]

\[
V_i \sim Bin(N_i^0, c_i) \quad \text{(Given in Eq. (23))}
\]

\[
\nu_i = N_i^0 - V_i \quad \text{(Given in Eq. (24))}
\]

\[
N_i^c = \eta_i + \nu_i \quad \text{(Given in Eq. (25))}
\]

\[
U_i^+ = U_i - \max(U_i - N_i^c, 0) \quad \text{(Given in Eq. (26))}
\]

\[
U_i^- = \max(U_i - G_i^-, 0) \quad \text{(Given in Eq. (27))}
\]

\[
\eta_i^+ = \max(\eta_i - G_i^+, 0) \quad \text{(Given in Eq. (28))}
\]

\[
\eta_i^- = \eta_i - \max(\eta_i - N_i^c, 0) \quad \text{(Given in Eq. (29))}
\]

\[
M_i = \min(N_i^1, N_i^c) \quad \text{(Given in Eq. (30))}
\]

\[
G_i^+ = \max(N_i^1 - N_i^c, 0) \quad \text{(Given in Eq. (31))}
\]

\[
G_i^- = \max(N_i^c - N_i^c, 0).
\]

We now derive conditions on $U_i$ and $V_i$ that lead to the realization of the event $U_i^+ = a$, we do so by expanding the definition of $U_i^+$ and considering all the cases it entails separately. Starting with the definition of $U_i^+$ we have:

\[
U_i^+ = U_i - \max(U_i - N_i^c, 0) \quad \text{(By definition of $U_i$)}
\]

\[
= U_i - \max(U_i - \eta_i - \nu_i, 0) \quad \text{(By definition of $N_i^c$)}
\]

\[
= U_i - \max(U_i - N_i^1 + U_i - N_i^0 + V_i, 0) \quad \text{(By definition of $\eta_i, \nu_i$)}
\]

\[
= U_i - \max(2U_i + V_i - N_i, 0). \quad \text{(Because $N_i^1 + N_i^0 = N_i$)}
\]
There are now two cases for the event $U^+_i = a$, depending on how the max function in the definition of $U^+_i$ is resolved:

**Case 1:**

\[
\max(2U_i + V_i - N_i, 0) = 2U_i + V_i - N_i
\]  \hspace{1cm} (98)

Because of this our event of interest can be written as:

\[
a = U^+_i = U_i - 2U_i - V_i + N_i \quad \text{(By (97))}
\]

\[
\implies U_i = N_i - V_i - a, \quad (99)
\]

which is the first condition in the first case of Equation (33). Note that the condition of this case given in Eq. (98) implies that $2U_i + V_i - N_i \geq 0$, and we can use the result in (99) to expand this condition as follows:

\[
0 \leq 2U_i + V_i - N_i \quad \text{(By (98))}
\]

\[
= 2(N_i - V_i - a) + V_i - N_i \quad \text{(By (99))}
\]

\[
= N_i - V_i - 2a
\]

\[
\implies V_i \leq N_i - 2a,
\]

which is the second condition in the first case of Equation (33).

**Case 2:**

\[
\max(2U_i + V_i - N_i, 0) = 0
\]  \hspace{1cm} (100)

Using this case and the result in (97), we can write the event of interest as

\[
a = U^+_i = U_i, \quad (101)
\]

the first condition in the second case of Equation (33). Second, the fact that the max function equals 0 in Case 2, implies that $2U_i + V_i - N_i \leq 0$ is another condition for this case. Using the result in (101) this can be simplified as:

\[
0 \geq 2U_i + V_i - N_i \quad \text{(By (100))}
\]

\[
= 2a + V_i - N_i \quad \text{(By (101))}
\]

\[
\implies V_i \leq N_i - 2a,
\]

which is the second condition in the second case of Equation (33).

We proceed in the same manner for the event $U^-_i = b$, in Equation (34) we start by expanding the definition of $U^-_i$:

\[
U^-_i = \max(U_i - \max(N_i^l - N_i^0, 0), 0)
\]  \hspace{1cm} (By definition of $U^-_i$)

\[
= \max(U_i - \max(U_i + V_i - \eta_i - \nu_i, 0), 0) \quad \text{(By definition of $N_i^l$, $N_i^0$)}
\]

\[
= \max(U_i - \max(U_i + V_i - N_i^l + U_i - N_i^0 + V_i, 0), 0) \quad \text{(By definition of $\eta_i$, $\nu_i$)}
\]

\[
= \max(U_i - \max(2(U_i + V_i) - N_i, 0), 0). \quad \text{(Because $N_i = N_i^l + N_i^0$)} \quad (102)
\]
Because of the two max functions we have four different possibilities for the value of \( U_l^- \), all dependent on how the max functions resolve. As we did before, we can derive conditions on \( U_l \) and \( V_l \) by studying these four cases separately. We start each of the four cases by listing the ways the inner max and the outer max resolve.

**Case 1:**

\[
\begin{align*}
\max(2(U_l + V_l) - N_l, 0) &= 2(U_l + V_l) - N_l \quad (103) \\
\max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) &= U_l - \max(2(U_l + V_l) - N_l, 0). \quad (104)
\end{align*}
\]

First, we can use both conditions to simplify the event \( U_l^- = b \) as follows:

\[
b = U_l^- = U_l - 2(U_l + V_l) + N_l \quad (105)
\]

\[
\implies U_l = N_l - 2V_l - b, \quad (106)
\]

which is the first condition in the first case of Equation (34). Note now that the condition in (104) implies that \( U_l - \max(2(U_l + V_l) - N_l, 0) \geq 0 \), combining this with the above we obtain:

\[
0 \leq U_l - \max(2(U_l + V_l) - N_l, 0)
= b \quad \text{(By (105))}
\geq 0, \quad \text{(By definition)}
\]

so this condition is always satisfied because of how we restrict the domain of \( b \). The condition in (103) implies that \( 2(U_l + V_l) - N_l \geq 0 \), again we use the result in (106) to expand this as follows:

\[
0 \leq 2(U_l + V_l) - N_l
= 2(N_l - 2V_l - b + V_l) - N_l \quad \text{(By (106))}
= N_l - V_l - 2b
\implies V_l < N_l - 2b,
\]

the second condition in the first case of Eq. (34).

**Case 2:**

\[
\begin{align*}
\max(2(U_l + V_l) - N_l, 0) &= 0 \quad (107) \\
\max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) &= U_l - \max(2(U_l + V_l) - N_l, 0). \quad (108)
\end{align*}
\]

The first condition in the second case of Eq. (34) follows from using these conditions with the event \( U_l^- = b \):

\[
b = U_l^- = \max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) \quad \text{(By (102))}
= U_l - \max(2(U_l + V_l) - N_l, 0) \quad \text{(By (108))}
= U_l. \quad \text{(By (107))}
\]


The condition in (108) implies that $U_l - \max(2(U_l + V_l) - N_l) \geq 0$, using the above we see that this condition is always satisfied in this case:

$$0 \leq U_l - \max(2(U_l + V_l) - N_l, 0)$$

$$= U_l \quad \text{(By (107))}$$

$$= b \quad \text{(By (109))}$$

$$\geq 0. \quad \text{(By definition)}$$

Because of this, this condition is omitted from the formulation in Eq. (34). Finally, condition (107) implies:

$$0 \geq 2(U_l + V_l) - N_l$$

$$= 2(b + V_l) - N_l \quad \text{(By (109))}$$

$$\Rightarrow V_l \leq \frac{N_l - 2b}{2},$$

the second condition of the case.

**Case 3:**

$$\max(2(U_l + V_l) - N_l, 0) = 2(U_l + V_l) - N_l \quad (110)$$

$$\max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) = 0. \quad (111)$$

First, we have the event $U_l^- = b$ taking form:

$$b = U_l^- = \max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) \quad \text{(By (102))}$$

$$= 0, \quad \text{(By (111))}$$

which leads us to the first condition in case 3 of Equation (34). For the second condition, start with (110), then we have:

$$2(U_l + V_l) - N_l \geq 0 \quad \text{(By (110))}$$

$$\Rightarrow U_l \geq \frac{N_l}{2} - V_l, \quad (112)$$

which is the form of the second condition of Case 4 in (34). Finally, from Condition (111) we have:

$$0 \geq U_l - \max(2(U_l + V_l) - N_l, 0)(\text{By (111)})$$

$$= U_l - 2(U_l + V_l) + N_l \quad \text{(By (110))}$$

$$\Rightarrow V_l \geq \frac{N_l - U_l}{2}. \quad (113)$$

Clearly, (113) is the last condition in the fourth case of equation (34).

**Case 4:**

$$\max(2(U_l + V_l) - N_l, 0) = 0 \quad (114)$$

$$\max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) = 0. \quad (115)$$
First, the event $U_i^- = b$ takes the following form in this case:

$$
b = U_i^- = \max(U_i - \max(2(U_i + V_i) - N_i, 0), 0) \quad \text{(By definition of } U_i^-)$$
$$= 0, \quad \text{(By (115))}
$$
which is the first condition in the fourth case of Eq. (34). Second, we have:

$$0 \geq U_i - \max(2(U_i + V_i) - N_i, 0) \quad \text{(By (115))}$$
$$= U_i \quad \text{(By (114))}
$$
$$\implies U_i \leq 0, \quad (116)$$

but since $U_i$ is a binomial random variable its value can never be less than 0. Because of this the second condition in the case becomes:

$$U_i = 0. \quad (117)$$

Finally we have:

$$0 \geq 2(U_i + V_i) - N_i \quad \text{(By (114))}$$
$$= 2V_i - N_i \quad \text{(By (117))}
$$
$$\implies V_i \leq \frac{N_i}{2}, \quad (118)$$

the third condition in the fourth case of (34).

As we did before we can now expand the definition of $\eta_i^+$ to derive conditions on $U_i$ and $V_i$ that lead to the realization of the event $\eta_i^+ = c$. Starting with the definition of $\eta_i^+$ we have:

$$\eta_i^+ = \max(\nu_i - G_i^+, 0) \quad \text{(By definition of } \eta_i^+)
$$
$$= \max(\nu_i - \max(N_i^c - N_i^l, 0), 0) \quad \text{(By definition of } G_i^+) \quad (119)$$
$$= \max(\nu_i - \max(\nu_i + U_i - V_i, 0), 0) \quad \text{(By definition of } N_i^l, N_i^c)$$
$$= \max(N_i^l - U_i - \max(N_i^l - U_i + N_i^0 - V_i - U_i - V_i, 0), 0) \quad \text{(By definition of } \nu_i, \nu_l)$$
$$= \max(N_i^l - U_i - \max(N_i - 2(U_i + V_i), 0), 0). \quad (120)$$

Since there are two nested max functions in the definition of $\eta_i^+$, there will be 4 cases that correspond to how the max functions are resolved; each one of these cases is going to represent a different value of $\eta_i^+$. Below we study each case separately and show that they lead to the four cases in the indicator function of Eq. (35).

**Case 1:**

$$\max(N_i^l - U_i - \max(N_i - 2(U_i + V_i), 0), 0) = N_i^l - U_i - \max(N_i - 2(U_i + V_i), 0) \quad (121)$$
$$\max(N_i - 2(U_i + V_i), 0) = N_i - 2(U_i + V_i). \quad (122)$$
First, use both condition to derive a form for the event $\eta_i^+ = c$:

$$c = \eta_i^+ = \max(N^{1}_i - U_i - \max(N_i - 2(U_i + V_i), 0), 0) \quad \text{(By (120))}$$

$$= N^{1}_i - U_i - N_i + 2(U_i + V_i) \quad \text{(By (121) and (122))}$$

$$= U_i + 2V_i - N_i^0$$

$$\implies U_i = c + N_i^0 - 2V_i, \quad (123)$$

which is the first condition in the first case of Equation (35). Second, we use the result above to rewrite the first condition:

$$0 \leq N_i - 2(U_i + V_i) \quad \text{(By (122))}$$

$$= N_i - 2(c + N_i^0 - 2V_i + V_i) \quad \text{(By (123))}$$

$$= N_i^1 - N_i^0 - 2c + 2V_i$$

$$\implies V_i \geq \frac{N_i^0 - N_i^1 + 2c}{2},$$

which is the second condition in the first case in Equation (35).

**Case 2:**

$$\max(N_i^1 - U_i - \max(N_i - 2(U_i + V_i), 0), 0) = N_i^1 - U_i - \max(N_i - 2(U_i + V_i), 0) \quad (124)$$

$$\max(N_i - 2(U_i + V_i), 0) = 0 \quad (125)$$

First, use the second condition to find a form for the event $\eta_i^+ = c$

$$c = \eta_i^+ = \max(N_i^1 - U_i - \max(N_i - 2(U_i + V_i), 0), 0) \quad \text{(By (120))}$$

$$= N_i^1 - U_i \quad \text{(By (124) and (125))}$$

$$\implies U_i = N_i^1 - c \quad (126)$$

Now use the second condition together with the above to derive:

$$0 \geq N_i - 2(U_i + V_i) \quad \text{(By (125))}$$

$$= N_i - 2(N_i^1 - c + V_i) \quad \text{(By (126))}$$

$$= N_i^0 - N_i^1 + 2c - 2V_i$$

$$\implies V_i \geq \frac{N_i^0 - N_i^1 + 2c}{2}.$$ 

This, and Equation (126) are the conditions in the second case of Equation (35).

**Case 3:**

$$\max(N_i^1 - U_i - \max(N_i - 2(U_i + V_i), 0), 0) = 0 \quad (127)$$

$$\max(N_i - 2(U_i + V_i), 0) = N_i - 2(U_i + V_i). \quad (128)$$
The first condition of Case 3 in Eq. (35) is given by:
\[ c = \eta_1^+ = \max(N_i^1 - U_i - \max(N_i - 2U_i + V_i), 0), 0 \]  
(By (120))
\[ = 0. \]  
(By (127))

Now use (128) to rewrite the (127):
\[ 0 \geq N_i^1 - U_i - \max(N_i - 2U_i + V_i), 0 \]  
(By (127))
\[ = N_i^1 - U_i - N_i + 2(U_i + V_i) \]  
(By (128))
\[ = U_i - N_i^0 + 2V_i \]
\[ \implies U_i \leq N_i^0 - 2V_i, \]  
(129)
this is the second condition of the third case in Eq. (35). The final condition in the case is given by:
\[ 0 \leq N_i - 2(U_i + V_i) \]  
(By (128))
\[ \implies V_i \geq \frac{N_i - 2U_i}{2}. \]

**Case 4:**

\[ \max(N_i^1 - U_i - \max(N_i - 2U_i + V_i), 0), 0) = 0 \]  
(130)
\[ \max(N_i - 2(U_i + V_i), 0) = 0. \]  
(131)

The first condition in the fourth case of Eq. (35) is obtained as follows:
\[ c = \eta_1^+ = \max(N_i^1 - U_i - \max(N_i - 2(U_i + V_i), 0), 0) \]  
(By (120))
\[ = 0. \]  
(By (130))

The second condition in the same case can be obtained by starting from (130):
\[ 0 \geq N_i^1 - U_i - \max(N_i - 2U_i + V_i), 0) \]  
(By (130))
\[ = N_i^1 - U_i \]  
(By (131))
\[ \implies U_i \geq N_i^1. \]

Since \( U_i \) is a binomial random variable with number of trials \( N_i^1 \), it can never be greater then \( N_i^1 \), so the condition above becomes:
\[ U_i = N_i^1, \]  
(132)
which is the second condition in the fourth case of Eq. (35). Now rearrange the terms in the second condition and use the result above to obtain the final condition in the case:
\[ 0 \geq N_i - 2(U_i + V_i) \]  
(By (132))
\[ = N_i - 2N_i^1 + 2V_i \]  
(By (132))
\[ = N_i^0 - N_i^1 + 2V_i \]  
(By (132))
\[ \because N_i^1 + N_i^0 = N_i \]
\[ \implies V_i \geq \frac{N_i^0 - N_i^1}{2}. \]
This concludes the derivation of Eq. (35).

As before, we derive the conditions on $U_i$ and $V_i$ that lead to the realization of the event $\eta_i^- = d$ by expanding its definition:

\[
\eta_i^- = \eta_i - \max(\eta_i - N_i^1, 0) \quad \text{(By definition of $\eta_i^-$)}
\]

\[
= N_i^1 - U_i - \max(N_i^1 - U_i - U_i - V_i, 0) \quad \text{(By definition of $\eta_i$ and $N_i^1$)}
\]

\[
= N_i^1 - U_i - \max(N_i^1 - 2U_i - V_i, 0). \quad \text{(133)}
\]

Using this expansion, we see that there are two possible definitions for the event $\eta_i^- = d$ in terms of $U_i$ and $V_i$, both depending on how the max function is evaluated.

**Case 1:**

\[
\max(N_i^1 - 2U_i - V_i, 0) = N_i^1 - 2U_i - V_i. \quad \text{(134)}
\]

First, we use the above to rewrite the event $\eta_i^- = d$ for this case:

\[
d = \eta_i^- = N_i^1 - U_i - \max(N_i^1 - 2U_i - V_i, 0) \quad \text{(By 133)}
\]

\[
= N_i^1 - U_i - N_i^1 + 2U_i + V_i \quad \text{(By 134)}
\]

\[
= U_i + V_i
\]

\[
\Rightarrow U_i = d - V_i. \quad \text{(135)}
\]

Second, we use the result just derived to rewrite the condition for the max in (134) in this case:

\[
0 \leq N_i^1 - 2U_i - V_i \quad \text{(By 134)}
\]

\[
= N_i^1 - 2(d - V_i) - V_i \quad \text{(By 135)}
\]

\[
\Rightarrow V_i \geq 2d - N_i^1.
\]

This and (135) are, respectively, the first and second conditions on the first case in Equation (36).

**Case 2:**

\[
\max(N_i^1 - 2U_i - V_i, 0) = 0. \quad \text{(136)}
\]

First, we use the above to rewrite the event $d = \eta_i^-$ for this case:

\[
d = \eta_i^- = N_i^1 - U_i - \max(N_i^1 - 2U_i - V_i, 0) \quad \text{(By 133)}
\]

\[
= N_i^1 - U_i \quad \text{(By 136)}
\]

\[
\Rightarrow U_i = N_i^1 - d. \quad \text{(137)}
\]
Second, we rewrite the condition on the max with the above result:

\[ 0 \geq N_i^1 - 2U_i - V_i \quad \text{(By (136))} \]
\[ = N_i^1 - 2(N_i^1 - d) - V_i \quad \text{(By (137))} \]
\[ \implies V_i \geq 2d - N_i^1. \]

These are the condition in the second case of Equation (36).

Finally we have the event \( M_i = m \): conditions on \( U_i \) and \( V_i \) that lead to its realization can again be established by expanding its definition, this will lead to the form of \( M_i \) in Equation (37). Starting with the definition of \( M_i \):

\[ M_i = \min(N_i^1, N_i^0) \quad \text{(By definition of \( M_i \))} \]
\[ = \min(U_i + V_i, \eta_i + \nu_i) \quad \text{(By definition of \( N_i^1, N_i^0 \))} \]
\[ = \min(U_i + V_i, N_i^1 - U_i + N_i^0 - V_i) \quad \text{(By definition of \( \eta_i, \nu_i \))} \]
\[ = \min(U_i + V_i, N_i - (U_i + V_i)). \quad \text{(Because \( N_i^1 + N_i^0 = N_i \))} \quad (138) \]

Because of this we see that the definition of \( M_i \) is composed of two cases that depend on how the min function is resolved.

**Case 1:**

\[ \min(U_i + V_i, N_i - (U_i + V_i)) = U_i + V_i. \quad (139) \]

First, use the expanded definition of \( M_i \) with the case above to rewrite the event \( M_i = m \):

\[ m = M_i = \min(U_i + V_i, N_i - (U_i + V_i)) \quad \text{(By (138))} \]
\[ = U_i + V_i \quad \text{(By (139))} \quad (140) \]
\[ \implies U_i = m - V_i. \]

Second, we use the result just introduced to rewrite the condition on the min for this case:

\[ U_i + V_i \leq N_i - (U_i + V_i) \quad \text{(By (139))} \]
\[ \implies 0 \geq U_i + V_i - N_i + (U_i + V_i) \]
\[ = 2U_i + 2V_i - N_i \]
\[ = 2m - N_i \quad \text{(By (140))} \]
\[ \implies m \leq N_i / 2. \]

leading us to both conditions in the first case of Equation (37).

**Case 2:**

\[ \min(U_i + V_i, N_i - (U_i + V_i)) = N_i - (U_i + V_i). \quad (141) \]
Again, we use the case above to rewrite the definition of the event $M_i = m$:

$$m = M_i = \min(U_i + V_i, N_i - (U_i + V_i)) \quad \text{(By (138))}$$

$$= N_i - (U_i + V_i) \quad \text{(By (141))}$$

$$\implies U_i = N_i - V_i - m. \quad \text{(142)}$$

Second, we use the above to rewrite the condition for this case:

$$U_i + V_i \geq N_i - (U_i + V_i) \quad \text{(By (141))}$$

$$\implies N_i - V_i - m + V_i \geq m \quad \text{(By (142) and (143))}$$

$$\implies m \leq N_i/2.$$ 

These are the two conditions in the definition of $M_i = m$ in equation (37).

Finally, the forms in equation (32) are simply obtained by conditioning on the event $U_i = j, V_i = k$ for any of the definitions above. This concludes the proof of the lemma. □

**B.3. Proof of Theorem 2**

All the probability statements throughout the proof are made conditionally on $X$ and $H^{\text{sharp}}_0$, for this reason we omit the conditional notation from these statements. Recall also that all the quantities representing counts of units in each stratum are nonnegative integers.

Note first that, for any stratum $l$:

$$0 \leq U^+_l = U_i - \max(U_i - N^+_i, 0) \quad \text{(By definition of $U^+_i$)}$$

$$\leq N^+_i, \quad \text{(By definition of $U_i$)} \quad \text{(144)}$$

and

$$0 \leq \eta^+_l = \max(\eta_i - G^+_i, 0) \quad \text{(By definition of $\eta^+_i$)} \quad \text{(145)}$$

$$= \max(N^+_i - U_i - G^+_i, 0) \quad \text{(By definition of $\eta_i$)} \quad \text{(146)}$$

$$\leq N^+_i. \quad \text{(147)}$$

This implies that:

$$TE^+ = \sum_{l=1}^{L} U^+_l - \eta^+_l \quad \text{(By definition of $TE^+$)}$$

$$\leq \sum_{l=1}^{L} N^+_l - 0, \quad \text{(By (144) and (35))}$$

$$= N^1$$

and:

$$TE^+ = \sum_{l=1}^{L} U^+_l - \eta^+_l \quad \text{(By definition of $TE^+$)}$$

$$\geq \sum_{l=1}^{L} 0 - N^1_i \quad \text{(By (33) and (147))}$$

$$= -N^1,$$
where recall that $N^1 = \sum_{l=1}^{L} N^1_l$, as defined in the statement of the theorem. By the same argument we can see that $-N^1 \leq TE^- \leq N^1$. This fact is useful to bound the domain of $TE^+$ and $TE^-$: while it is likely that not all integers between $-N^1$ and $N^1$ have positive probability for $TE^+$ and $TE^-$ under the distribution of $(\chi^+, \chi^-)$, we know that no integers outside the range above will have positive mass under that distribution.

To bound the domain of $\chi = \frac{b-c-1}{b+c+1}$ note that, by definition, $B_l$ represents the number of matched pairs in stratum $l$ that have outcome 1 for the treated unit and 0 for the control unit, independently of how matches are made. Because of this we know that $B_l$ can never be greater than either the number of treated units with outcome 1 or control units with outcome 0 in stratum $l$, and, therefore: $B_l \leq \min(N^1_l, N^0_l)$, independently of how many units are assigned to treatment. For the same reason we conclude that $N^m = \sum_{l=1}^{L} \min(N^1_l, N^0_l)$, using these facts we can see that:

$$\chi \in X_{N^m} = \left\{ \frac{b-c-1}{b+c+1} : b, c \in \{0, \ldots, N^m\} \right\}.$$

Note finally that, since $\chi \in X_{N^m}$ regardless of how matches are made, we know that $\chi^+ \in X_{N^m}$ and $\chi^- \in X_{N^m}$ because $\chi^+$ and $\chi^-$ are special cases of $\chi$ in which matches are made with Algorithms 5 and 6 respectively. Because of this we conclude that $(\chi^+, \chi^-) \in X_{N^m} \times X_{N^m}$. This set is fast to enumerate computationally and summations over its elements can be performed efficiently. While the distribution of $(\chi^+, \chi^-)$ likely does not place positive probability over all values in $X_{N^m} \times X_{N^m}$, we shall consider values in this set and derive conditions under which they have positive probability as well as what their probability is under this distribution.

Begin now by writing the pmf of the range $(\chi^+, \chi^-)$. As stated in Theorem 2 assume that $N^1_l, N^0_l$ are fixed and known for all strata $l = 1, \ldots, L$. For any two values $(r, s) \in X_{N^m} \times X_{N^m}$ we have:

$$\Pr(\chi^- = s, \chi^+ = r) = \Pr \left( \frac{TE^- - 1}{\sqrt{SD^- + 1}} = s, \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} = r \right)$$

$$= \Pr \left( \frac{TE^- - 1}{\sqrt{SD^- + 1}} = s, \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} = r \right)$$

(By def. of $(\chi^+, \chi^-)$)

$$= \Pr \left( \frac{TE^- - 1}{\sqrt{SD^- + 1}} = s, \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} = r \right)$$

(By Claim 4)

$$= \Pr \left( \frac{TE^- - 1}{\sqrt{SD^- + 1}} = s, \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} = r \right)$$

$$= \Pr \left( \frac{TE^- - 1}{\sqrt{SD^- + 1}} = s, \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} = r \right)$$

$$= \Pr \left( \frac{TE^- - 1}{\sqrt{SD^- + 1}} = s, \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} = r \right)$$

$$= f_1(r, z) + f_2(r, z) + f_3(r, z) + f_4(r, z) \quad \text{(149)}$$

The equality in (149) follows by the representation of $SD^+$ and $SD^-$ in (71) and (72) of Claim 4 respectively. Now we work with each of the four parts separately.

Starting with $f_3$, we now show that it must always be that $f_3 = 0$. Consider the event set $\{TE^- \geq 1, TE^+ < 1\}$. By definition of $\chi^+$, we know that, in this case:

$$\chi^+ = \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} \geq 0$$

(By definition of $\chi^-$)

Because of the event $TE^- \geq 1$
and:

\[ \chi^+ = \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} \]

(By definition of \( \chi^+ \))

\[ < 0. \]

(Because of the event \( TE^+ < 1 \))

Therefore, the event \( \{ TE^- \geq 1, TE^+ < 1 \} \) implies the event \( \{ \chi^+ < \chi^- \} \), but this is a contradiction, since \( \chi^+ \) and \( \chi^- \) are, respectively, the maximum and minimum of the same optimization problem (Formulation 2). Because of this it will never be the case that \( \chi^+ < \chi^- \). It must be, then, that the probability of this event under the distribution of interest is 0, and that, therefore, the probability of the event \( \{ TE^- \geq 1, TE^+ < 1 \} \) is also 0. Since \( f_3 \) is the probability of the event set \( \{ TE^- \geq 1, \frac{TE^- - 1}{\sqrt{R^- + 1}} = s, TE^+ < 1, \frac{TE^+ - 1}{\sqrt{R^+ + 1}} = r \} \), which is included in the 0-probability set above, then it must be that \( f_3 = 0 \).

This leaves us with \( f_1, f_2, f_4 \) to write out. For convenience we repeat the definitions of the sets introduced in the statement of the theorem:

\begin{align*}
\mathcal{A}(y) &= \left\{ \mathbf{a} = (a_1, \ldots, a_l) : \sum_{i=1}^l a_i = y, \ a_i \in \{0, \ldots, y\} \right\}, \quad (150) \\
\mathcal{B}(y, s) &= \left\{ \mathbf{b} = (b_1, \ldots, b_l) : \sum_{i=1}^l b_i = \left( \frac{y - 1}{s} \right)^2 - 1, \ b_i \in \left\{0, \ldots, \left( \frac{y - 1}{s} \right)^2 - 1 \right\} \right\}, \quad (151) \\
\mathcal{C}(x) &= \left\{ \mathbf{c} = (c_1, \ldots, c_l) : \sum_{i=1}^l c_i = x, \ c_i \in \{0, \ldots, x\} \right\}, \quad (152) \\
\mathcal{D}(x, r) &= \left\{ \mathbf{d} = (d_1, \ldots, d_l) : \sum_{i=1}^l d_i = \left( \frac{x - 1}{r} \right)^2 - 1, \ d_i \in \left\{0, \ldots, \left( \frac{x - 1}{r} \right)^2 - 1 \right\} \right\}. \quad (153)
\end{align*}

In addition, let \( \mathcal{H}(x, y, r, s) = \mathcal{A}(y) \times \mathcal{B}(y, s) \times \mathcal{C}(x) \times \mathcal{D}(x, r) \) be the Cartesian product of the sets above, such that each element of \( \mathcal{H}(x, y, r, s) \) is a 4-tuple of vectors each of length \( L \): \( (\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) \). We are now ready to derive \( f_1(r, s) \):

\begin{align*}
f_1(r, s) &= \Pr \left( TE^- < 1, \frac{TE^- - 1}{\sqrt{S^- + 1}} = s, TE^+ < 1, \frac{TE^+ - 1}{\sqrt{R^+ + 1}} = r \right) \\
&= \sum_{y=-N^1}^0 \sum_{x=-N^1}^0 \Pr \left( \sum_{i=1}^L T E^-_i = y, \frac{T E^-_i - 1}{\sqrt{S^- + 1}} = s, T E^+_i = x, \frac{T E^+_i - 1}{\sqrt{R^+ + 1}} = r \right) \\
&= \sum_{y=-N^1}^0 \sum_{x=-N^1}^0 \Pr \left( \sum_{i=1}^L T E^-_i = y, S^-_i = \left( \frac{y - 1}{s} \right)^2 - 1, T E^+_i = x, R^+_i = \left( \frac{x - 1}{r} \right)^2 - 1 \right) \\
&= \sum_{y=-N^1}^0 \sum_{x=-N^1}^0 \Pr \left( \sum_{i=1}^L T E^-_i = y, \sum_{i=1}^L S^-_i = \left( \frac{y - 1}{s} \right)^2 - 1, \sum_{i=1}^L T E^+_i = x, \sum_{i=1}^L R^+_i = \left( \frac{x - 1}{r} \right)^2 - 1 \right) \quad (157)
\end{align*}

\begin{align*}
&= \sum_{y=-N^1}^0 \sum_{x=-N^1}^0 \sum_{(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) \in \mathcal{H}(x, y, r, s)} \Pr \left( T E^-_1 = a_1, S^-_1 = b_1, T E^+_1 = c_1, R^+_1 = d_1, \ldots, \\
&\quad T E^-_L = a_L, S^-_L = b_L, T E^+_L = c_L, R^+_L = d_L \right) \\
&= \sum_{y=-N^1}^0 \sum_{x=-N^1}^0 \sum_{(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) \in \mathcal{H}(x, y, r, s)} \prod_{i=1}^L \Pr \left( T E^-_i = a_i, S^-_i = b_i, T E^+_i = c_i, R^+_i = d_i \right) \quad (159)
\end{align*}
Line (154) follows from the definition of \( f_1 \) given in (149). Equation (156) follows from rearranging the terms in the previous line. Equation (158) follows from the definitions in Claim 4. Equation (159) follows from the definition of \( \mathcal{H}(x,y,r,s) \), and equation (160) follows from independence of \( \text{TE}_i^+, \text{TE}_i^-, \mathcal{S}_i^+, \mathcal{S}_i^- \), \( R_i^+, R_i^- \) in stratum \( l \) from the same quantities in any other stratum. This independence is evident from the definitions of these quantities given in Claim 4. The same exact derivation steps leads us to a similar definition for \( f_2 \):

\[
\begin{align*}
    f_2(r,s) &= \Pr \left( \frac{\text{TE}^-}{\sqrt{S}} < 1, \frac{\text{TE}^- - 1}{\sqrt{S} + 1} = s, \frac{\text{TE}^+ - 1}{\sqrt{S} + 1} = r \right) \\
    &= \sum_{y=-N^1}^{N^1} \sum_{x=1}^{N^1} \Pr \left( \text{TE}^- = y, \frac{\text{TE}^- - 1}{\sqrt{S} + 1} = s, \frac{\text{TE}^+ - 1}{\sqrt{S} + 1} = r \right) \\
    &= \sum_{y=-N^1}^{N^1} \sum_{x=1}^{N^1} \Pr \left( \text{TE}^- = y, S^- = \left( \frac{y-1}{s} \right)^2 - 1, \text{TE}^+ = x, S^+ = \left( \frac{x-1}{r} \right)^2 - 1 \right) \\
    &= \sum_{y=-N^1}^{N^1} \sum_{x=1}^{N^1} \Pr(TE_i^- = a_1, S_i^- = b_1, TE_i^+ = c_1, S_i^+ = d_1, \ldots) \\
    &= \sum_{y=-N^1}^{N^1} \sum_{x=1}^{N^1} \sum_{(a,b,c,d) \in \mathcal{H}(x,y,r,s)} \prod_{l=1}^{L} \Pr \left( TE_i^- = a_l, S_i^- = b_l, TE_i^+ = c_l, S_i^+ = d_l \right).
\end{align*}
\]

In the case of \( f_4 \) we can follow the same steps to derive:

\[
\begin{align*}
    f_4(r,s) &= (\text{TE}^- \geq 1, \frac{\text{TE}^- - 1}{\sqrt{R} + 1} = s, \text{TE}^+ \geq 1, \frac{\text{TE}^+ - 1}{\sqrt{S} + 1} = r) \\
    &= \sum_{y=1}^{N^1} \sum_{x=1}^{N^1} \Pr \left( \text{TE}^- = y, \frac{\text{TE}^- - 1}{\sqrt{R} + 1} = s, \text{TE}^+ = x, \frac{\text{TE}^+ - 1}{\sqrt{S} + 1} = r \right) \\
    &= \sum_{y=1}^{N^1} \sum_{x=1}^{N^1} \Pr \left( \text{TE}^- = y, R^- = \left( \frac{y-1}{s} \right)^2 - 1, \text{TE}^+ = x, S^+ = \left( \frac{x-1}{r} \right)^2 - 1 \right) \\
    &= \sum_{y=1}^{N^1} \sum_{x=1}^{N^1} \Pr(TE_i^- = a_1, R_i^- = b_1, TE_i^+ = c_1, S_i^+ = d_1, \ldots) \\
    &= \sum_{y=1}^{N^1} \sum_{x=1}^{N^1} \sum_{(a,b,c,d) \in \mathcal{H}(x,y,r,s)} \prod_{l=1}^{L} \Pr \left( TE_i^- = a_l, R_i^- = b_l, TE_i^+ = c_l, S_i^+ = d_l \right). 
\end{align*}
\]

Note that \( f_1, f_2 \) and \( f_4 \) are nonzero on disjoint parts of the sets \( \{-N^1, \ldots, N^1\} \), as well as the fact that the definition of \( \mathcal{A}(y), \mathcal{B}(y,z), \mathcal{C}(x), \mathcal{D}(x,r) \) and \( \mathcal{H}(x,y,r,s) \) doesn’t change for any of the functions. For these reasons we can write the pmf of the range as in Eq (42). It remains to show that the three functions \( h_1, h_2, h_3 \) have the form stated in the theorem. This can be accomplished by expanding the inner probabilities of \( \text{TE}_i^+, \text{TE}_i^-, \mathcal{S}_i^+, \mathcal{S}_i^- \), \( \mathcal{D}_i^+, \mathcal{D}_i^- \) in each stratum. Before doing this, it will be useful to note that, for any stratum \( l \):

\[
0 \leq M_l = \min(N_i^l, N_i^-) \quad \text{(By definition of } M_l) \]
Because of this we consider values of $M_i$ in the integer range $0, \ldots, N_i/2$. Beginning with $\Pr(T_{E_i} = a_i, S_i^- = b_i, T_{E_i}^+ = c_i, R_i^+ = d_i)$:

$$\Pr(T_{E_i} = a_i, S_i^- = b_i, T_{E_i}^+ = c_i, R_i^+ = d_i)$$

(161)

$$= \sum_{m=0}^{N_i/2} \Pr(U_i^- - \eta_i = a_i, |U_i^- - \eta_i| = b_i, U_i^+ - \eta_i = c_i, M_i - |U_i^+ + \eta_i^- - M_i| = d_i, M_i = m)$$

(162)

$$= \sum_{m=0}^{N_i/2} \Pr(U_i^- - \eta_i = b_i, U_i^+ - \eta_i = c_i, m - |U_i^+ + \eta_i^- - m| = d_i, M_i = m|U_i^- - \eta_i = a_i) \Pr(U_i^- - \eta_i = a_i)$$

(163)

$$= \sum_{m=0}^{N_i/2} \mathbb{I}(|a_i| = b_i) \Pr(U_i^+ - \eta_i = c_i, m - |U_i^+ + \eta_i^- - m| = d_i, M_i = m|U_i^- - \eta_i = a_i) \Pr(U_i^- - \eta_i = a_i)$$

(164)

$$= \sum_{m=0}^{N_i/2} \mathbb{I}(|a_i| = b_i) \Pr(U_i^- - \eta_i = a_i, U_i^+ - \eta_i = c_i, |U_i^+ + \eta_i^- - m| = m - d_i, M_i = m)$$

(165)

$$= h_1(a_i, b_i, c_i, d_i).$$

(170)

Equation (162) follows from the representations of $TE$ and $SD$ given in Claim 1. Equation (163) from the definition of conditional probability, Equation (165) from the fact that $a$ and $b$ are constants and therefore they are independent from the other quantities in the equation. Equation (166) follows from the definition of conditional probability. Equations (167) follows from the fact that the event $\{M_i - |U_i^+ + \eta_i^- - M_i| = d_i\}$ is equal to the event $\{M_i - U_i^+ - \eta_i^- + M_i = d_i\} \cup \{M_i + U_i^+ + \eta_i^- - M_i = d_i, M_i \neq d_i\}$, and therefore the probability of its occurrence is equal to the sum of the probability of these two events. Equation (168) follow by rearranging the terms in the previous line, and Equation (169) from summing over values of $\eta_i^-$. The final line of the derivation is from the definition of $h_1$ given in the statement of the theorem. The following derivations for $h_2$ and $h_3$ follow exactly the same steps, starting with $h_2$:

$$\Pr(T_{E_i} = a_i, S_i^- = b_i, T_{E_i}^+ = c_i, S_i^+ = d_i)$$
Finally, $h_3$ can be derived from:

$$
\Pr(TE_i^- = a_i, R_i^- = b_i, TE_i^+ = c_i, S_i^+ = d_i) \\
= \sum_{m=0}^{N/2} \Pr(U_i^- - \eta_i^- = a_i, M_i = m | U_i^- + \eta_i^- - M_i| = b_i, U_i^+ - \eta_i^+ = c_i, |U_i^+ - \eta_i^+| = d_i, M_i = m) \\
= \sum_{m=0}^{N/2} \Pr(U_i^- - \eta_i^- = a_i, M_i = m | U_i^- + \eta_i^- - M_i| = b_i, |U_i^+ + \eta_i^+| = d_i, M_i = m | U_i^+ - \eta_i^+ = c_i) \Pr(U_i^+ - \eta_i^+ = c_i) \\
= \sum_{m=0}^{N/2} \Pr(U_i^- - \eta_i^- = a_i, m - |U_i^- + \eta_i^- - m| = b_i, |U_i^+ + \eta_i^+| = d_i, M_i = m | U_i^+ - \eta_i^+ = c_i) \Pr(U_i^+ - \eta_i^+ = c_i) \\
= \sum_{m=0}^{N/2} \Pr(U_i^- - \eta_i^- = a_i, m - |U_i^- + \eta_i^- - m| = b_i, |U_i^+ + \eta_i^+| = d_i, M_i = m | U_i^+ - \eta_i^+ = c_i) \Pr(U_i^+ - \eta_i^+ = c_i) \\
= \sum_{m=0}^{N/2} \Pr(U_i^- - \eta_i^- = c_i, M_i = m) \\
= h_3(a_i, b_i, c_i, d_i).
$$

This concludes the derivation of all the quantities in Theorem [3].
C. Proof of Theorem 3

C.1. Proof of Lemma 2

Proof. We will prove the result for $\Pr(U_i^- = a, U_i^+ = b)$ only, as the proof for $\Pr(\eta_i^- = a, \eta_i^+ = b)$ is symmetrical. Recall first that we are now in a case in which the number of treated and control units in each stratum, $N_i^t$ and $N_i^c$ respectively, are fixed. We use this together with Assumption 3 to write the marginal distributions of $U_i^+$ and $U_i^-$:

$$\Pr(U_i^+ = b) = \Pr(U_i - \max(U_i - N_i^c, 0) = b) = \begin{cases} \Pr(U_i = b) & \text{if } b < N_i^c \\ \Pr(U_i \geq N_i^c) & \text{if } b = N_i^c \end{cases}$$

$$\Pr(U_i^- = a) = \Pr(\max(U_i - G_i^-, 0) = a) = \begin{cases} \Pr(U_i \leq G_i^-) & \text{if } a = 0 \\ \Pr(U_i = a + G_i^-) & \text{if } a > 0 \end{cases}$$

In the above, $G_i^- = \max(N_i^t - N_i^c, 0)$, a fixed quantity. The first equality in both formulas follows from the definition of $U_i^+$ and $U_i^-$ given in Equations 28 and 29 respectively. The two cases of each definition follow directly from these definitions. Given the marginal distributions above, we can easily find distributions for $U_i^+$ and $U_i^-$ conditional on $U_i = x$:

$$\mathbb{I}(U_i^+ = b | U_i = x) = \begin{cases} 1 & \text{if } b < N_i^c, x = b \\ 1 & \text{if } b = N_i^c, x \geq N_i^c \\ 0 & \text{otherwise}. \end{cases}$$ (171)

$$\mathbb{I}(U_i^- = a | U_i = x) = \begin{cases} 1 & \text{if } a = 0, x \leq G_i^- \\ 1 & \text{if } a > 0, x = a + G_i^- \\ 0 & \text{otherwise}. \end{cases}$$ (172)

This is because, if the number of treated and control units in each stratum is fixed, randomness in $U_i^+$ and $U_i^-$ comes only from $U_i$, as is evident from the definitions of $U_i^+$ and $U_i^-$ given in Eqns. 33 and 34. Therefore once that is fixed the two quantities become constant and, as such, independent of each other. Given this fact, it is evident that $\Pr(U_i^- = a, U_i^+ = b | U_i = x) = \mathbb{I}(U_i^- = a, U_i^+ = b | U_i = x) = \mathbb{I}(U_i^- = a | U_i = x)\mathbb{I}(U_i^+ = b | U_i = x)$. This product of indicator functions can also be written as four conditions, each one representing the intersection of the event sets on which both distributions place nonzero probability:

$$\mathbb{I}(U_i^- = a | U_i = x)\mathbb{I}(U_i^+ = b | U_i = x) = \begin{cases} 1 & \text{if } b < N_i^c, x = b \text{ and } a = 0, x \leq G_i^- \\ 1 & \text{if } b < N_i^c, x = b \text{ and } a > 0, x = a + G_i^- \\ 1 & \text{if } b = N_i^c, x \geq N_i^c \text{ and } a = 0, x \leq G_i^- \\ 1 & \text{if } b = N_i^c, x \geq N_i^c \text{ and } a > 0, x = a + G_i^- \\ 0 & \text{otherwise}. \end{cases}$$

This is evident from Equations 171 and 172. The event sets in the equation can be simplified further as some of their elements are redundant, we show this in the following.

**Case 1:** $b < N_i^c, x = b$ and $a = 0, x \leq G_i^-$. First, using the second and fourth condition we can write $b \leq G_i^-$. Then, since $b$ has to be less than both $G_i^-$ and $N_i^c$ we can rewrite the condition as: $b < \min(N_i^c, G_i^- + 1)$, where we replace the equality on $G_i^-$ with $b$ being strictly less than $G_i^- + 1$. This is possible because $G_i^+$ is, by definition, a nonnegative integer. The
final event here becomes: \( x = b, b < \min(N_i^c, G_i^- + 1), a = 0 \).

**Case 2:** \( b < N_i^c, x = b \) and \( a > 0, x = a + G_i^- \).

First, using the first, second and fourth condition we have: \( N_i^c > b = x = a + G_i^- \), which we rewrite as the event: \( x = b, b = a + G_i^-, a < N_i^c - G_i^- \). With this the final event becomes: \( x = b, b = a + G_i^-, 0 < a < N_i^c - G_i^- \).

**Case 3:** \( b = N_i^c, x \geq N_i^c \) and \( a = 0, x \leq G_i^- \).

Here we can combine the second and fourth conditions to obtain: \( N_i^c \leq x \leq G_i^- \). This leads to the event: \( b = N_i^c, N_i^c \leq x \leq G_i^-, a = 0 \).

**Case 4:** \( b = N_i^c, x \geq N_i^c \) and \( a > 0, x = a + G_i^- \).

First from the second and fourth events we have: \( x = a + G_i^- \geq N_i^c \), which can be used to obtain the event: \( a \geq N_i^c - G_i^- \). We also rewrite the third condition as \( a \geq 1 \), leading to the following final representation for the event of this case: \( b = N_i^c, a \geq \max(N_i^c - G_i^-, 1), x = a + G_i^- \).

Finally, we can put all these events together to obtain a simplified formulation for the joint conditional distribution of \( U_i^+, U_i^- \):

\[
\Pr(U_i^- = a, U_i^+ = b | U_i = x) = \begin{cases} 
1 & \text{if } x = b, b < \min(N_i^c, G_i^- + 1), a = 0 \\
1 & \text{if } x = b, b = a + G_i^-, 0 < a < N_i^c - G_i^- \\
1 & \text{if } b = N_i^c, N_i^c \leq x \leq G_i^-, a = 0 \\
1 & \text{if } b = N_i^c, a \geq \max(N_i^c - G_i^-, 1), x = a + G_i^- \\
0 & \text{otherwise.}
\end{cases}
\]

The marginal distribution of \( U_i^+, U_i^- \) can now be found by using the law of total probability to sum over \( U_i^+ \):

\[
\Pr(U_i^- = a, U_i^+ = b) = \sum_{x=0}^{N_i^c} \Pr(U_i = x) \Pr(U_i^+ = a, U_i^- = b | U_i = x)
\]

\[
= \sum_{x=0}^{N_i^c} \begin{cases} 
\Pr(U_i = x) & \text{if } x = b, b < \min(N_i^c, G_i^- + 1), a = 0 \\
\Pr(U_i = x) & \text{if } x = b, b = a + G_i^-, 0 < a < N_i^c - G_i^- \\
\Pr(U_i = x) & \text{if } b = N_i^c, N_i^c \leq x \leq G_i^-, a = 0 \\
\Pr(U_i = x) & \text{if } b = N_i^c, a \geq \max(N_i^c - G_i^-, 1), x = a + G_i^- \\
0 & \text{otherwise}
\end{cases}
\]

\[
= \begin{cases} 
\Pr(U_i = b) & \text{if } b < \min(N_i^c, G_i^- + 1), a = 0 \\
\Pr(U_i = b) & \text{if } b = a + G_i^-, 0 < a < N_i^c - G_i^- \\
\Pr(U_i = b) & \text{if } b = N_i^c, a = 0 \\
\Pr(U_i = a + G_i^-) & \text{if } b = N_i^c, a \geq \max(N_i^c - G_i^-, 1) \\
0 & \text{otherwise}
\end{cases}
\]
To obtain the final line the condition on $x$ in the event is substituted in the probability. This concludes the derivation of the pmf of $U_l^+, U_l^-$ and the proof of this lemma. □

C.2. Proof of Theorem 3

Proof The proof follows closely the template of the proof of Theorem 2. Note first that in this case $N_l^i$ and $N_l^c$ are fixed and known for all strata $l$. We first bound the values of $U_l^+$ and $\eta_l^+$:

$$0 \leq U_l^+ = M_l - V_l^+ \quad \text{(By definition of } V_l^+)$$

$$\leq M_l, \quad \text{(Both } M_l \text{ and } V_l^+ \text{ are nonnegative integers)}$$

and:

$$0 \leq \eta_l^+ = M_l - \nu_l^+ \quad \text{(By definition of } \nu_l^+)$$

$$\leq M_l, \quad \text{(Both } M_l \text{ and } \nu_l^+ \text{ are nonnegative integers)}$$

From this it follows that: $-M_l \leq TE_l^+ = U_l^+ - \eta_l^+ \leq M_l$. Recall that $M = \sum_{l=1}^M M_l$ and that $TE^+ = \sum_{l=1}^L TE_l^+$; it must be that $-M \leq TE^+ \leq M$. An exactly symmetric argument shows that $-M \leq TE^- \leq M$.

Because of the above the values that we will consider for $TE^+$ and $TE^-$ in this proof are the integers ranging between $-M$ and $M$. Finally, as detailed in the proof of Theorem 2, it must be that $B$ and $C$ are always less than $M$, the total number of matches made, as they represent counts of matched pairs. This allows us to bound the set of values that have positive probability under the distribution of $(\chi^+, \chi^-)$:

$$\mathcal{X}_M = \left\{ \frac{b - c - 1}{\sqrt{b} + c + 1}; b, c \in \{0, \ldots, M\} \right\}.$$

Given the set above, we will consider values of $(\chi^+, \chi^-)$ in the set $\mathcal{X}_M$, even though not all of the set’s elements will have positive probability under the distribution of interest. As done in Theorem 2 we derive a set of conditions on values within $\mathcal{X}_M$ that determine whether or not those values have positive probability and, if yes, what is their probability under the distribution of $(\chi^+, \chi^-)$ under Assumption 4.

The steps leading to the form of (52) in the main statement of the theorem are exactly the same as those in the proof of Theorem 2 with the sole difference that the possible values of $TE^+$ and $TE^-$ now are between $-M$ and $M$. Because of this, the summations over $x$ and $y$ both in the definition of the joint probability of interest and in the functions, $f_1, f_2, f_3, f_4$, introduced in the proof of Theorem 2 range from $-M$ to $M$ in this case. In particular, the summation over $x$ and $y$ in $f_1$ ranges from $-M$ to 0 in this case, the summation over $x$ in $f_2$ ranges from 1 to $M$, the summation over $y$ in $f_2$ now ranges from $-M$ to 0, and the summations over $x$ and $y$ in $f_4$ range from 1 to $M$ in this case. The definitions of $f_1, f_2, f_4$ are also exactly the same as those in the proof of Theorem 2 including the fact that $f_3 = 0$. The argument made for this in the proof of Theorem 2 applies exactly in the same way for this case.
The only other difference is in the inner probabilities for the joint distribution of the statistics in each stratum. These inner probabilities result in equations (53)-(55), and can be derived as follows:

$$
\begin{align*}
\Pr(TE_i^-=a_i, S_i^- = b_i, TE_i^+ = c_i, R_i^+ = d_i) \\
= \Pr(U_i^- - \eta_i^- = a_i, |U_i^- - \eta_i^-| = b_i, U_i^+ - \eta_i^+ = c_i, |U_i^+ + \eta_i^+ - M_i| = d_i) \\
= \Pr(U_i^- - \eta_i^- = a_i, |a_i| = b_i, U_i^+ - \eta_i^+ = c_i, |U_i^+ + \eta_i^+ - M_i| = M_i - d_i) \\
= \mathbb{I}(|a_i| = b_i) \Pr(U_i^- - \eta_i^- = a_i, U_i^+ - \eta_i^+ = c_i, |U_i^+ + \eta_i^+ - M_i| = M_i - d_i) \\
= \mathbb{I}(|a_i| = b_i) \left[ \Pr(U_i^- - \eta_i^- = a_i, U_i^+ - \eta_i^+ = c_i, |U_i^+ + \eta_i^+ - M_i| = M_i - d_i) + \Pr(U_i^- - \eta_i^- = a_i, U_i^+ - \eta_i^+ = c_i, |U_i^+ + \eta_i^+ - M_i| = M_i - d_i) \right] \\
= \mathbb{I}(|a_i| = b_i) \left[ \Pr(U_i^- = a_i + \eta_i^-, U_i^+ = c_i + \eta_i^+, \eta_i^- = \frac{2M_i - d_i - c_i}{2}) \\
+ \Pr(U_i^- = a_i + \eta_i^-, U_i^+ = c_i + \eta_i^+, \eta_i^+ = \frac{d_i - c_i}{2}) \mathbb{I}(M_i \neq d_i) \right] \\
= \mathbb{I}(|a_i| = b_i) \sum_{j=0}^{M_i} \Pr(U_i^- = a_i + j, U_i^+ = \frac{d_i + c_i}{2}) \Pr(\eta_i^- = j, \eta_i^+ = \frac{2M_i - d_i - c_i}{2}) \\
+ \Pr(U_i^- = a_i + j, U_i^+ = \frac{2M_i - d_i - c_i}{2}) \Pr(\eta_i^- = j, \eta_i^+ = \frac{d_i - c_i}{2}) \\
= g_1(a_i, b_i, c_i, d_i). \\
\end{align*}
$$

In the above, Equation (173) follows from the definitions of $TE_i^-, S_i^-, TE_i^+, R_i^+$ given in Claim 4. Equation (174) follows from rearranging the terms in the last equality in the parentheses, and from plugging in the equality for $U_i^- - \eta_i^-$. (175) follows from the fact that $a_i$ and $b_i$ are both constants and so the event $|a_i| = b_i$ is independent from all the others inside the parentheses. Equation (176) follows from the fact that the event $\{|U_i^+ + \eta_i^+ - M_i| = M_i - d_i\} = \{U_i^+ + \eta_i^+ - M_i = M_i - d_i\} \cup \{U_i^+ + \eta_i^+ - M_i = d_i - M_i, M_i \neq d_i\}$. This follows from the definition of the absolute value function applied to positive integers. Note that the two sets are disjoint, so the probability of their events can be added and that $M_i$ and $d_i$ are fixed quantities, so they are independent of all the other events in the statement. Equation (177) follows from rearranging the terms in the equalities for $\eta_i^-$. Equation (178) follows from summing over all values of $\eta_i^-$ and from plugging in equalities for $\eta_i^+$ into the equation for $U_i^+$. Finally, (179) follows from the definition of $g_1(a_i, b_i, c_i, d_i)$ given in the statement of the theorem. The derivations for $g_2$ and $g_3$ follow exactly the same steps. For $g_2$ we have:

$$
\begin{align*}
\Pr(TE_i^- = a_i, S_i^- = b_i, TE_i^+ = c_i, S_i^+ = d_i) \\
= \Pr(U_i^- - \eta_i^- = a_i, |U_i^- - \eta_i^-| = b_i, U_i^+ - \eta_i^+ = c_i, |U_i^+ + \eta_i^+ - M_i| = d_i) \\
= \Pr(U_i^- - \eta_i^- = a_i, |a_i| = b_i, U_i^+ - \eta_i^+ = c_i, |a_i| = d_i) \\
= \mathbb{I}(|a_i| = b_i) \mathbb{I}(|a_i| = d_i) \Pr(U_i^- - \eta_i^- = a_i, U_i^+ - \eta_i^+ = c_i) \\
= \mathbb{I}(|a_i| = b_i) \mathbb{I}(|a_i| = d_i) \Pr(U_i^- = a_i + \eta_i^-, U_i^+ = c_i + \eta_i^+) \\
= \mathbb{I}(|a_i| = b_i) \mathbb{I}(|a_i| = d_i) \sum_{j=0}^{M_i} \sum_{k=0}^{M_i} \Pr(U_i^- = a_i + k, U_i^+ = c_i + j) \Pr(\eta_i^- = k, \eta_i^+ = j) \\
= g_2(a_i, b_i, c_i, d_i).
\end{align*}
$$
Finally, $g_3$ can be derived with the same steps as $g_1$:

$$\Pr(TE^- = a_t, R^-_i = b_t, TE^+_i = c_i, S^+_i = d_i)$$

$$= \Pr(U^-_i - \eta^- = a_t, M_t - |U^-_i + \eta^- - M_t| = b_t, U^+_i + \eta^+ = c_i, |U^+_i - \eta^+| = d_i)$$

$$= \Pr(U^-_i - \eta^- = a_t, |U^-_i + \eta^- - M_t| = M_t - b_t, U^+_i + \eta^+ = c_i, |c_i| = d_i)$$

$$= \Pr(|\eta^-| = d_i) \Pr(U^-_i - \eta^- = a_t, |U^-_i + \eta^- - M_t| = M_t - b_t, U^+_i - \eta^+ = c_i)$$

$$= \Pr(U^-_i - \eta^- = a_t, U^+_i + \eta^+ - M_t = M_t - b_t, U^+_i + \eta^+ = c_i + \eta^+)$$

$$+ \Pr(U^-_i = a_t, U^+_i + \eta^+ - M_t = b_t - M_t, U^+_i + \eta^+ = c_i + \eta^+) \mathbb{I}(M_t \neq b_i)$$

$$= \Pr(U^-_i - \eta^- = a_t, U^+_i + \eta^+ = c_i + \frac{2M_t - a_t - b_t}{2}, U^+_i - \eta^+ = c_i + \frac{b_t - a_t}{2})$$

$$\mathbb{I}(M_t \neq b_i)$$

$$= \Pr(U^-_i = a_t, U^+_i + \eta^+ = c_i + k) \Pr(\eta^- = \frac{b_t - a_t}{2}, \eta^+ = k)$$

$$+ \Pr(U^-_i = a_t, U^+_i + \eta^+ = c_i + k) \Pr(\eta^- = \frac{2M_t - b_t - a_t}{2}, \eta^+ = k) \mathbb{I}(M_t \neq b_i)$$

$$= g_3(a_t, b_t, c_i, d_i).$$

This concludes the proof of Theorem $3$.  

D. Algorithms for $Z$ minimization.

E. Proof of Theorem $4$

Proof Because the $b_t$'s are defined on a pre-specified grid, we know the maximum value of $F$ may not occur at one of the grid points. Since by definition of $x_t$ we know that $f_2(x_t) \leq b_t$, and since $F$ is decreasing in its second argument, we have $F(f_1(x_t), b_t) \leq F(f_1(x_t), f_2(x_t))$ for each $t$, and taking a max over all $t$:

$$\max_{1 \leq t \leq L} F(f_1(x_t), b_t) \leq \max_{1 \leq t \leq L} F(f_1(x_t), f_2(x_t)) \leq \max_x F(f_1(x), f_2(x)).$$

This is the left inequality of the bound. The rest of the proof deals with the right inequalities. First, it is true that:

$$f_1(x^*) = \max_{x: f_2(x) = f_2(x^*)} f_1(x).$$

(180)

If this were not true then either $f_1(x^*) < \max_{x: f_2(x) = f_2(x^*)} f_1(x)$ or $f_1(x^*) > \max_{x: f_2(x) = f_2(x^*)} f_1(x)$. If the first inequality were true then

$$F(f_1(x^*), f_2(x^*)) < F(\max_{x: f_2(x) = f_2(x^*)} f_1(x), f_2(x^*)),$$

which contradicts the definition of $x^*$. The second option also cannot be true as we know there exists a solution $x^*$ so that the maximum is attained with $f_1$ and $f_2$ values $f_1(x^*)$ and $f_2(x^*)$. So we can say that (180) holds.

From (180) and using $l^*$ defined in the statement of the theorem, we can derive:

$$f_1(x^*) = \max_{x: f_2(x) = f_2(x^*)} f_1(x) \leq \max_{x: f_2(x) \leq f_2(x^*)} f_1(x) \leq \max_{x: f_2(x) \leq l^*} f_1(x) = f_1(x^*),$$

(181)
Algorithm 7: Minimize \( z \) with general constraints.

**Data:** Set of real vectors \( \{(y_{i1}, \ldots, y_{iN_i})\}_{i=1}^L \) and \( \{(y_{i1}^c, \ldots, y_{iN_i}^c)\}_{i=1}^L \)

Additional data parameters for optimization, such as covariates \( D \),

Set of constraints on Formulation 3: \( W \),

ILP Solver for Formulation 3, \( F(W, (D), y) \)

A Procedure to generate a increasingly finer, ordered sequence of reals \( b_1, \ldots, b_L \) from lower and upper bounds, \( G(Lb, Ub) \).

**Result:** \( N^t \times N^c \) binary matrix of matches, \( a \).

1. Minimize Formulation 3 by removing the upper bound constraint: \( a_0 := F(W, (D), y) \);
2. \( b_l := \sum_{i=1}^{N^t} \sum_{j=1}^{N^c} (y_{ij}^c - y_{ij})^2 a_{0ij} \);
3. \( b = (b_1, \ldots, b_L) := G(b_1, B_L) \);
4. while \( \min_{l \in \{1, \ldots, L\}} \frac{d_{ai} \sqrt{M}}{\sqrt{b_l - (d_{ai})^2}} - \min_{l \in \{1, \ldots, L\}} \frac{d_{ai} \sqrt{M}}{\sqrt{b_{l-1} - (d_{ai})^2}} \leq \epsilon \) do
   5. for \( l = 1, \ldots, L \) do Minimize Formulation 3: \( a_l := F(W, (D, b_l), y) \);
   6. \( b^a := \min_{l \in \{1, \ldots, L\}} \frac{d_{ai} \sqrt{M}}{\sqrt{b_l - (d_{ai})^2}} \)
   7. for \( l = 2, 4, 6, \ldots, L \) do
      8. if \( \frac{d_{ai} \sqrt{M}}{\sqrt{b_{l-1} - (d_{ai})^2}} < b^a \) then \( (b_{l1}, \ldots, b_{LL}) := G(b_{l-1}, b_l) \);
   9. end
10. \( b := (b_{l1}, \ldots, b_{LL}) \)
11. end
12. return any one of the \( a_l \).

where we used that the set \( \{ x : f_2(x) = f_2(x^*) \} \) is smaller than the set \( \{ x : f_2(x) \leq f_2(x^*) \} \) which is smaller than \( \{ x : f_2(x) \leq b_t \_ \} \), since \( f_2(x^*) \leq b_t \_ \) by definition of \( l^* \). Thus, \( f_1(x^*) \leq f_1(x_{t-1}) \). Now,

\[
F(f_1(x^*), f_2(x^*)) \leq F(f_1(x^*), b_{t-1}) \\
\leq F(f_1(x_{t-1}), b_{t-1}) \\
\leq \max F(f_1(x_l), b_{l-1}).
\]

Here the first inequality above follows from the definition of \( l^* \), \( b_{t-1} \leq f_2(x^*) \), and the fact that \( F \) decreases in the second argument. The second inequality comes from [181] and the fact that \( F \) is increasing in its first argument. The third inequality follows from taking a maximum over all \( l \) rather than using \( l^* \). The proof is complete. □

F. Case Study Details

F.1. Additional Figures for Case Study 1

F.2. Additional Case Study: Training Program Evaluation

In this case study, we used training program evaluation data described in [Dehejia and Wahba (1999)], and [Dehejia and Wahba (2002)], which were drawn from [Lalonde (1986)]. This data set contains 15,992 control
Figure 7 Upper and lower bounds for maximum $z$-test objective function value over a range of $b_l$ (Case Study 1: $n=30$), illustrating the range for the finer mesh in the next step of the algorithm, which is found in Figure 8.

Figure 8 Upper and lower bounds for maximum $z$-test objective function value over a range of $b_l$ (Case Study 1: $n=30$), at the finest mesh. The final value for the maximization problem is between 31.41 and 31.48.

units and 297 treatment units. The covariates for matching are as follows: age, education, Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if married, 0 otherwise), ndegree (1 if no degree, 0 otherwise), and earnings in 1975. The outcome is the earnings in 1978. The treatment variable is whether an individual receives job training. We computed distance between units as follows: $\text{dist}_{ij} = 1$ if covariates Black, Hispanic, married and ndegree were the same, and the differences in age, education and

Hypothesis Tests That Are Robust to Choice of Matching Method
Morucci, Noor-E-Alam, and Rudin: *Hypothesis Tests That Are Robust to Choice of Matching Method*

Article submitted to: manuscript no. (Please, provide the manuscript number!)

Figure 9  Upper and lower bounds for minimum $z$-test objective function value over a range of $b_l$ (Case Study 1: $n=30$), illustrating the range for the finer mesh in the next step of the algorithm, which is found in Figure 10.

Figure 10  Upper and lower bounds for minimum $z$-test objective function value over a range of $b_l$ (Case Study 1: $n=30$), at the finest mesh. The final value for the minimization problem is between -73.01 and -73.47.

earnings in 1975 were less or equal to 5, 4 and 4000, respectively, for treated unit $i$ and control unit $j$, 0 otherwise.

In the Figure 12 the $P$-value upper bounds are 1 and the $P$-value lower bounds are 0, illustrating that there is a lot of uncertainty associated with the choice of experimenter – one experimenter choosing 287 matched
pairs can find a P-value of $\sim 0$ and declare a statistically significant difference while another experimenter can find a P-value of $\sim 1$ and declare the opposite. In this case it is truly unclear whether or not training has an effect on the earnings. To sanity check whether a reasonably sized effect would have been detected had one been present, we injected synthetic random noise (with normal distribution of mean $\simeq$ $10,000 and standard deviation $\simeq$ $100$) on the treatment outcome, and the z-test (see Figure ??) robustly detects the treatment effect before the solutions become infeasible.
F.3. Additional Case Study: Crime and Transition into Adulthood

In this case study we have used the data from a U.S. Department of Justice study regarding crime during the transition to adulthood, for youths raised in out-of-home care (see Courtney and Cusick 2010). Each observation represents a youth, and the outcome is whether he or she committed a violent crime over the 3 waves of the study.

The (binary) covariates are as follows: hispanic, white, black, other race, alcohol or drug dependency, mental health diagnosis, history of neglect, entry over age of 12, entry age under 12, in school or employed, prior violent crime, any prior delinquency. The “treatment” variable is whether or not the individual is female; in particular we want to determine whether being female (controlling for race, criminal history, school/employment and relationship with parents) influences the probability of committing a violent crime. Here dist$_{ij}$ = 1 whenever all covariates of treated unit $i$ are the same as those of the control unit $j$, 0 otherwise.

Figure 11 is constructed in an analogous way to Figure 13 (using McNemar’s test rather than the $z$-test) showing the total number of discordant pairs along the x axis. Here, any matched pairs assignment would show a significant difference for the risks of violence between males and females. This difference becomes more pronounced as the number of pairs increases. Thus, the outcome is robust to the choice of experimenter.

G. Integer Linear Programming Basics

ILP techniques have become practical for many large-scale problems over the past decade, due to a combination of increased processor speeds and better ILP solvers. Any type of logical condition can be encoded as linear constraints in an ILP formulation with binary or integer variables. Consider two binary variables $x \in \{0, 1\}$ and $y \in \{0, 1\}$. The logical condition “if $y = 0$ then $x = 0$” can be simply encoded as

$$x \leq y.$$
Note that this condition imposes no condition on $x$ when $y = 1$. Translating if-then constraints into linear constraints can sometimes be more complicated; suppose, we would like to encode the logical condition that if a function $f(w)$ is greater than 0, then another function $g(w)$ is greater or equal to 0. We can use the following two linear equations to do this, where $\theta$ is a binary variable and $M$ is a positive number that is larger than the maximum values of both $f$ and $g$:

\[-g(w) \leq M\theta\]
\[f(w) \leq M(1 - \theta).\]

In order for $f(w)$ to be positive, then $\theta$ must be 0, in which case, $g(w)$ is then restricted to be positive. If $f(w)$ is negative, $\theta$ must be 0, in which case no restriction is placed on the sign of $g(w)$. (See for instance the textbook of Winston and Venkataramanan (2003), for more examples of if-then constraints).

ILP can capture other types of logical statements as well. Suppose we would like to incorporate a restriction such that the integer variable $S_i$ takes a value of $K$ only if $i = t$, and 0 otherwise. The following four if-then constraints can be used to express this statement, where $\lambda_1$ and $\lambda_2$ are binary variables:

\[\lambda_1 = 1 \text{ if } i + 1 > t\]
\[\lambda_2 = 1 \text{ if } t + 1 > i\]
\[S_i = k \text{ if } \lambda_1 + \lambda_2 > 1\]
\[S_i = 0 \text{ if } \lambda_1 + \lambda_2 < 2.\]

Each of these if-then constraints (4)-(7) can be converted to a set of equivalent linear equations, similar to what we described above. (See also Noor-E-Alam et al. (2012) and Winston and Venkataramanan (2003)).

There is no known polynomial-time algorithm for solving ILP problems as they are generally NP-hard, but they can be solved in practice by a number of well-known techniques (Wolsey (1998)). The LP relaxation of an ILP provides bounds on the optimal solution, where the LP relaxation of an ILP is where the integer constraints are relaxed and the variables are allowed to take non-integer (real) values. For instance, if we are solving a maximization problem, the solution of the LP relaxation can serve as an upper bound, since it solves a problem with a larger feasible region, and thus attains a value at least as high as that of the more restricted integer program. ILP solvers use branch-and-bound or cutting plane algorithms combined with other heuristics, and are useful for cases where the optimal integer optimal solution is not attained by the LP relaxation. The branch-and-bound algorithms often use LP relaxation and semi-relaxed problems as subroutines to obtain upper bounds and lower bounds, in order to determine how to traverse the branch-and-bound search tree (Chen et al. 2011, Wolsey 1998). The most popular ILP solvers such as CPLEX, Gurobi and MINTO each have different versions of branch-and-bound techniques with cutting plane algorithms and problem-specific heuristics.