Appendix A: 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 the uncertainty set $\mathcal{A}_{\text{good}}$ should most often obey. Let $dist(x, x')$ be a metric on the space of x, some of these constraints are:

- (Calipers) When $\mathbf{a}(i) \neq \emptyset$ then $\text{dist}(x_i^t, x_{\mathbf{a}(i)}^c) \leq \epsilon$.
- (Covariate balance, mean of chosen treatment units similar to mean of control group) $\forall$ covariates p we have:

$$\left| \frac{1}{M} \sum_{\{i:\mathbf{a}(i)\neq\emptyset\}} x_{ip}^t - \frac{1}{M} \sum_{\{j:\mathbf{a}(j)\neq\emptyset\}} x_{jp}^c \right| \leq \epsilon_p. \tag{31}$$

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

$$\text{Fitness}(\mathbf{a}, \{x_i^t\}_{i=1}^{N^t}, \{x_i^c\}_{i=1}^{N^c}) \geq \text{Maxfit} - \epsilon,$$

where Maxfit is precomputed as: $\text{Maxfit} = \max_{\mathbf{a}\in\mathcal{A}} \text{Fitness}(\mathbf{a}, \{x_i^t\}_{i=1}^{N^t}, \{x_i^c\}_{i=1}^{N^c})$. If one desires the range of results for all maximally fit pairs and no other pairs, ϵ can be set to 0.

We note that, to constitute a ILP formulation, the chosen type of fitness function must be expressed as some linear function of the matching indicators. While it is technically true that not all fitness functions can be encoded in this way, many popular constraints for matching can be. For example, Zubizarreta (2012) gives linear constraint formulations for a number of popular match quality constraints. However, despite the fact that many match assignments may exist that satisfy these constraints, these methodologies produce only one, and it is not chosen uniformly at random from any known distribution.

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.

Appendix B: Proof of Theorem 2

Proof We prove the result for statement (1), as the proof of statement (2) is exactly symmetrical. Because the b_l'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_l we know that $f_2(x_l) \leq b_l$, and since F is decreasing in its second argument, we have $F(f_1(x_l), b_l) \leq F(f_1(x_l), f_2(x_l))$ for each l, and taking a max over all l:

$$\max_{l\in 1\ldots L} F(f_1(x_l), b_l) \leq \max_{l\in 1\ldots L} F(f_1(x_l), f_2(x_l)) \leq \max_x F(f_1(x), f_2(x)).$$

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

$$f_1(x^*) = \max_{x: f_2(x)=f_2(x^*)} f_1(x). \tag{32}$$

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

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

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 (32) holds.

From (32) and using l^* defined in the statement of the theorem, we can derive:

$$f_1(x^*) = \max_{x:f_2(x)=f_2(x^*)} f_1(x) \leq \max_{x:f_2(x)\leq f_2(x^*)} f_1(x) \leq \max_{x:f_2(x)\leq b_{l^*}} f_1(x) = f_1(x_{l^*}), \tag{33}$$

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 $\{x : f_2(x) \leq b_{l^*}\}$, since $f_2(x^*) \leq b_{l^*}$ by definition of l^*. Thus, $f_1(x^*) \leq f_1(x_{l^*})$. Now,

$$\begin{aligned} F(f_1(x^*), f_2(x^*)) &\leq F(f_1(x^*), b_{l^*-1}) \\ &\leq F(f_1(x_{l^*}), b_{l^*-1}) \\ &\leq \max_l F(f_1(x_l), b_{l-1}). \end{aligned}$$

Here the first inequality above follows from the definition of l^*, $b_{l^*-1} \leq f_2(x^*)$, and the fact that F decreases in the second argument. The second inequality comes from (33) 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.

Appendix C: Algorithms that maximize and minimize χ under Exclusively Binning Constraints

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.

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

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, $\mathcal{X}$. 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_l^t and N_l^c units. In each stratum there will be U_l treated units with outcome $Y_i = 1$, η_l control units with outcome $Y_j = 1$, $V_l = N_l^t - U_l$ treatment units with outcome 0, and $\nu_l = N_l^c - \eta_l$ 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 $Y = 1$ and the control unit has $Y = 0$. 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:

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

	Y		
		1	0
T	1	U_l	V_l
	0	η_l	ν_l

	Y_c		
		1	0
Y_t	1	A_l	B_l
	0	C_l	D_l

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 χ 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 χ output by maximizing χ 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 $M = \sum_{l=1}^{L} \min(N_l^t, N_l^c)$ must to be made:

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

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

subject to:

$$\sum_{i=1}^{N^t} \sum_{j=1}^{N^c} a_{ij} y_i^t (1 - y_j^c) = B(\mathbf{a}) \qquad \text{(Total number of first type of discordant pairs)} \tag{34}$$

$$\sum_{i=1}^{N^t} \sum_{j=1}^{N^c} a_{ij} y_j^c (1 - y_i^t) = C(\mathbf{a}) \qquad \text{(Total number of second type of discordant pairs)} \tag{35}$$

$$\sum_{i=1}^{N^t} a_{ij} \leq 1 \qquad \forall j \quad \text{(Match each control unit at most once)} \tag{36}$$

$$\sum_{j=1}^{N^c} a_{ij} \leq 1 \qquad \forall i \quad \text{(Match each treatment unit at most once)} \tag{37}$$

$$\sum_{i \in S_l} \sum_{j \in S_l} a_{ij} = \min(N_l^t, N_l^c) \qquad \forall l \quad \text{(Make as many matches as possible)} \tag{38}$$

$$a_{ij} \in \{0, 1\} \qquad \forall i, j \quad \text{(Defines binary variable } a_{ij}\text{)} \tag{39}$$

(Additional user-defined **exclusively binning** constraints.) (40)

		Y	
		1	0
T	1	U_l	V_l
	0	η_l	ν_l

		Y_c	
		1	0
Y_t	1	A_l	B_l
	0	C_l	D_l

Figure 1 Summary of pair counts in stratum l.

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 $Y_i = 1$, η_l control units with outcome $Y_j = 1$, $V_l = N_l^t - U_l$ treatment units with outcome 0, and $\nu_l = N_l^c - \eta_l$ 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 $M_l = \min(N_l^t, N_l^c)$ matches. We would like to make the matches within each stratum such that χ is either maximized or minimized. Algorithm 1 maximizes χ,

Algorithm 1: Maximize χ 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 χ statistic value over all possible matches in $\mathcal{A}_{\text{good}}$.

for $l=1,\ldots,L$ **do**
  $M_l = \min(N_l^t, N_l^c)$
  $U_l^+ := U_l - \max(U_l - N_l^c, 0)$
  $V_l^+ := M_l - U_l^+$
  $\eta_l^+ := \max(\eta_l - \max(N_l^c - N_l^t, 0), 0)$
  $\nu_l^+ := M_l - \eta_l^+$
end
$TE = \sum_{l=1}^{L} U_l^+ - \eta_l^+$
if $TE \geq 0$ **then**
  for $l=1,\ldots,L$ **do**
    $A_l^+ := \min(U_l^+, \eta_l^+)$
    $D_l^+ := \min(\nu_l^+, V_l^+)$
    $B_l^+ := \min(U_l^+ - A, \nu_l^+ - D)$
    $C_l^+ := \min(\eta_l^+ - A, V_l^+ - D)$
  end
end
else
  for $l=1,\ldots,L$ **do**
    $B_l^+ := \min(U_l^+, \nu_l^+)$
    $C_l^+ := \min(\eta_l^+, V_l^+)$
    $A_l^+ := \min(U_l^+ - B, \eta_l^+ - C)$
    $D_l^+ := \min(\nu_l^+ - B, V_l^+ - C)$
  end
end
return $\frac{\sum_{l=1}^{L} B_l^+ - C_l^+ - 1}{\sqrt{\sum_{l=1}^{L} B_l^+ + C_l^+ + 1}}$

Algorithm 2: ComputeMaximizedSD

Data: Positive integers: U, V, η, ν

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

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

Algorithm 3: ComputeMinimizedSD

Data: Positive integers: U, V, η, ν

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

Make $A := \min(U, \eta)$ A-pairs.

Make $D := \min(\nu, V)$ D-pairs.

Make $B := \min(U - A, \nu - D)$ B-pairs.

Make $C := \min(\eta - A, V - D)$ C-pairs.

Return A, B, C, D.

Algorithm 4: Maximize χ 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 χ statistic value

for $l = 1, \ldots, L$ **do**
  $M_l := \min(N_l^t, N_l^c)$
  $U_l^+ := U_l - \max(U_l - N_l^c, 0)$
  $V_l^+ := M_l - U_l^+$
  $\eta_l^+ := \max(\eta_l - \max(N_l^c - N_l^t, 0), 0)$
  $\nu_l^+ := M_l - \eta_l^+$
end
$TE^+ = \sum_{l=1}^{L} U_l^+ - \eta_l^+$
if $TE^+ \geq 1$ **then**
  for $l = 1, \ldots, L$ **do**
    $(A_l^+, B_l^+, C_l^+, D_l^+) := ComputeMinimizedSD(U_l^+, V_l^+, \eta_l^+, \nu_l^+)$
  end
end
else
  for $l = 1, \ldots, L$ **do**
    $(A_l^+, B_l^+, C_l^+, D_l^+) := ComputeMaximizedSD(U_l^+, V_l^+, \eta_l^+, \nu_l^+)$
  end
end
return $\chi^+ = \frac{\sum_{l=1}^{L} B_l^+ - C_l^+ - 1}{\sqrt{\sum_{l=1}^{L} B_l^+ + C_l^+ + 1}}$

C.1. Correctness of the Optimization Algorithms

As it is clear from their definitions, these algorithms do not require any of the conventional MIP solving techniques and as such are much 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 6. *(Correctness of Algorithm 1) Algorithm 1 globally solves Formulation 2 for the Maximal value of* χ.

Algorithm 5: Minimize χ 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 χ statistic value

for $l=1,\ldots,L$ **do**
  $M_l := \min(N_l^t, N_l^c)$
  $U_l^- := \max(U_l - \max(N_l^t - N_l^c, 0), 0)$
  $V_l^- := M_l - U_l^-$
  $\eta_l^- := \eta_l - \max(\eta_l - N_l^t, 0)$
  $\nu_l^- := M_l - V_l^-$
end
$TE^- = \sum_{l=1}^{L} U_l^- - \eta_l^-$
if $TE^- \geq 1$ **then**
  for $l=1,\ldots,L$ **do**
    $(A_l^-, B_l^-, C_l^-, D_l^-) := ComputeMaximizedSD(U_l^-, V_l^-, \eta_l^-, \nu_l^-)$
  end
end
else
  for $l=1,\ldots,L$ **do**
    $(A_l^-, B_l^-, C_l^-, D_l^-) := ComputeMinimizedSD(U_l^-, V_l^-, \eta_l^-, \nu_l^-)$
  end
end
return $\chi^- = \frac{\sum_{l=1}^{L} B_l^- - C_l^- - 1}{\sqrt{\sum_{l=1}^{L} B_l^- + C_l^- + 1}}$

What follows is a proof of the algorithms' correctness. The proof is structured into 4 different claims and a theorem equivalent to Theorem 6 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 1 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

	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 1 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.

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 χ 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 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 $\max(U_l - N_l^c, 0)$ 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 $\chi = \frac{B-C-1}{\sqrt{B+C}}$ 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 χ:

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

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 χ. 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 χ 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 χ.

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 χ 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 $Y_{u_0} = 0$. Let the value of χ 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 χ 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 χ:

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

Then, with some algebra we can see that $\chi_{u_0} \leq \chi_{u_1}$, implying that, in order to maximize χ, 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 χ.

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 χ_k and one unmatched control unit and two candidate treatments. By the above, the value of χ_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 η 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 χ 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 ν 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 1 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 kth exchange operation. Again, by Table 1 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 2 and 3 respectively make the matches that maximize and minimize $B + C$.*

Proof. Consider Algorithm 3 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 3 we know that:

$$A^* = \min(U, \eta)$$
$$D^* = \min(V, \nu)$$
$$B^* = \min(U - A^*, \nu - D^*)$$
$$C^* = \min(\eta - A^*, V - D^*)$$

By definition of U, V, η, ν, A^* and D^* 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 1, 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 $B + C$.

C.1.1. Proof of Theorem 6

Proof. Consider first the problem of maximizing χ. The observations are divided into l strata such that N_l^t treated units and N_l^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

$$\chi = \frac{TE}{\sqrt{SD}} = \frac{\sum_{l=1}^{L}(B_l - C_l) - 1}{\sqrt{\sum_{l=1}^{L} B_l + C_l + 1}}.$$

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 $M = \sum_{i=1}^{L} M_l = \sum_{i=1}^{L} \min(N_l^t, N_l^c)$ because, by constraints (10) and (11), each treatment and control unit can only be matched once. By Constraint (38) we know that exactly $M_l = \min(N_l^t, N_l^c)$ matched pairs must be constructed in each stratum, therefore, units in excess of M_l 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 1 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 $TE = \sum_{l=1}^{L} U_l^+ - \eta_l^+$ independently of how matches are made. Because of this, TE can be considered fixed at this point, and maximizing χ 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 $ComputeMinimizedSD$ if $TE \geq 0$ and $ComputeMinimizedSD$ 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 4 globally maximizes χ 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 5 is exactly symmetrical to this one: this symmetry is made apparent by the fact that running Algorithm 5 on the data is equivalent to flipping the treatment indicator and then running Algorithm 4 on the resulting data.

Appendix D: Algorithm for Maximization of Z statistic

Algorithm 6 is useful for maximization of Z. Minimization of Z can be accomplished with a symmetric algorithm.

Algorithm 6: Maximize z with general constraints.

Data: Set of real vectors $\{(y^t_{l_1},\ldots,y^t_{l_{N^t_l}})\}_{l=1}^L$ and $\{(y^c_{l_1},\ldots,y^c_{l_{N^c_l}})\}_{l=1}^L$,
Initial lower bound on the variance, b_1, Additional data parameters for optimization, such as covariates, $\mathbf{D}$.

Result: $N^t \times N^c$ binary matrix of matches, $\mathbf{a}^*$.

Initialize: maximize Formulation 3 by removing the upper bound constraint in (19), denote solution by $\mathbf{a}^{(0)}$;
Compute initial upper bound on variance: $b_L^{(0)} := \sum_{i=1}^{N^t}\sum_{j=1}^{N^c}(y_i^t - y_j^c)^2 a_{ij}^{(0)}$;
Create initial grid of length L: $\mathbf{b} = (b_1^{(0)},\ldots,b_L^{(0)})$;
Start iteration counter, $iter = 1$
if $\bar{d}_{\mathbf{a}^{(0)}} \geq 0$ **then**
  while $\max_l \frac{\bar{d}_{\mathbf{a}_l^{(iter)}}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l-1}^{(iter)} - (\bar{d}_{\mathbf{a}_l^{(iter)}})^2}} - \max_l \frac{\bar{d}_{\mathbf{a}_l^{(iter)}}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l}^{(iter)} - (\bar{d}_{\mathbf{a}_l^{(iter)}})^2}} \geq \epsilon$ **do**
    Increment iteration counter: $iter = iter + 1$
    for $l = 1,\ldots,L$ **do** Maximize Formulation 3, using $b_l^{(iter-1)}$ as upper bound on variance. Denote solution with $\mathbf{a}_l^{(iter)}$;
    Compute lower bound on z-score $z_{LB} := \max_{l\in 1,\ldots,L} \frac{\bar{d}_{\mathbf{a}_l^{(iter)}}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l}^{(iter-1)} - (\bar{d}_{\mathbf{a}_l^{(iter)}})^2}}$
    for $l = 2,4,6,\ldots,L$ **do**
      Apply Theorem 2: **if** $\frac{\bar{d}_{\mathbf{a}_l^{(iter)}}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l-1}^{(iter-1)} - (\bar{d}_{\mathbf{a}_l^{(iter)}})^2}} > z_{LB}$ **then** refine the grid between $b_{l-1}^{(iter-1)}$ and $b_l^{(iter-1)}$, obtaining a new set of grid points;
    end
    Concatenate all refined grid points into a new grid, denoted by $\mathbf{b}^{(iter)} = (b_1^{(iter)},\ldots,b_L^{(iter)})$, and sorted in increasing order.
  end
  return $\mathbf{a}^* = \arg\max_{l=1,\ldots,L} \frac{\bar{d}_{\mathbf{a}_l^{(iter)}}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l}^{(iter)} - (\bar{d}_{\mathbf{a}_l^{(iter)}})^2}}$
end
if $\bar{d}_{\mathbf{a}^{(0)}} < 0$ **then**
  Flip the treatment indicator (1 becomes 0, 0 becomes 1) for convenience in handling only positive treatment effects.
  while $\min_l \frac{\bar{d}_{\mathbf{a}_l^{(iter)}}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l}^{(iter)} - (\bar{d}_{\mathbf{a}_l^{(iter)}})^2}} - \min_l \frac{\bar{d}_{\mathbf{a}_l^{(iter)}}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l-1}^{(iter)} - (\bar{d}_{\mathbf{a}_l^{(iter)}})^2}} \leq \epsilon$ **do**
    Increment iteration counter: $iter = iter + 1$
    for $l = 1,\ldots,L$ **do** Minimize Formulation 3, using $b_l^{(iter-1)}$ as upper bound on variance. Denote solution with $\mathbf{a}_l^{(iter)}$;
    Compute upper bound on z-score $z_{UB} := \min_{l\in 1,\ldots,L} \frac{\bar{d}_{\mathbf{a}_l^{(iter)}}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l}^{(iter-1)} - (\bar{d}_{\mathbf{a}_l^{(iter)}})^2}}$
    for $l = 2,4,6,\ldots,L$ **do**
      Apply Theorem 2: **if** $\frac{\bar{d}_{\mathbf{a}_l^{(iter)}}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l-1}^{(iter-1)} - (\bar{d}_{\mathbf{a}_l^{(iter)}})^2}} < z_{UB}$ **then** refine the grid between $b_{l-1}^{(iter-1)}$ and $b_l^{(iter-1)}$, obtaining a new set of grid points;
    end
    Concatenate all refined grid points into a new grid, denoted by $\mathbf{b}^{(iter)} = (b_1^{(iter)},\ldots,b_L^{(iter)})$, and sorted in increasing order.
  end
  return $\mathbf{a}^* = \arg\min_{l=1,\ldots,L} \frac{\bar{d}_{\mathbf{a}_l^{(iter)}}\sqrt{M}}{\sqrt{\frac{1}{M}b_{l}^{(iter)} - (\bar{d}_{\mathbf{a}_l^{(iter)}})^2}}$.
end

Appendix E: Randomization Distribution of (χ^+, χ^-) Under Exclusively Binning Constraints

In this section we use the algorithm above to derive a randomization distribution for (χ^+, χ^-) under the null hypothesis of no treatment effect. This allows us to test $\mathbb{H}_0^{sharp}$ 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 $\mathbb{H}_0^{sharp}$, potential outcomes are assumed to be fixed and invariant between treatment regimes. Under exclusively binning constraints, the units are divided into $S_1, \dots S_l$ 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_l^t treatment units and N_l^c control units in each stratum l. From the two assumptions before it follows that the data are generated as follows:

$$U_l | \mathbb{H}_0^{sharp} \overset{iid}{\sim} Bin(e_l, N_l^1), \tag{41}$$

$$\eta_l = N_l^1 - U_l, \tag{42}$$

$$V_l | \mathbb{H}_0^{sharp} \overset{iid}{\sim} Bin(e_l, N_l^0), \tag{43}$$

$$\nu_l = N_l^0 - V_l, \tag{44}$$

$$N_l^t = U_l + V_l, \tag{45}$$

$$N_l^c = \nu_l + \eta_l, \tag{46}$$

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 $Y=0$, η_l is the number of units with $T=0$ and $Y=1$ and ν_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 η_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 χ, 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:

$$\chi^+ = \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} = \frac{B^+ - C^+ - 1}{\sqrt{B^+ + C^+ + 1}}, \quad \chi^- = \frac{TE^- - 1}{\sqrt{SD^- + 1}} = \frac{B^- - C^- - 1}{\sqrt{B^- + C^- + 1}}, \tag{47}$$

where B^+ is the count of matched pairs produced by Algorithm 4 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 $G_l^- = \max(N_l^t - N_l^c, 0)$ and $G_l^+ = \max(N_l^c - N_l^t, 0)$:

$$U_l^+ = U_l - \max(U_l - N_l^c, 0) \tag{48}$$

$$V_l^+ = M_l - U_l^+ \tag{49}$$

$$U_l^- = \max(U_l - G_l^-, 0), 0) \tag{50}$$

$$V_l^- = M_l - U_l^- \tag{51}$$

$$\eta_l^+ = \max(\eta_l - G_l^+, 0) \tag{52}$$

$$\nu_l^+ = M_l - \eta_l^+ \tag{53}$$

$$\eta_l^- = \eta_l - \max(\eta_l - N_l^t, 0) \tag{54}$$

$$\nu_l^- = M_l - \eta_l^- . \tag{55}$$

These definitions correspond to those introduced at lines 2-5 of Algorithms 4 and 5. Figure 2 summarizes the variables in our framework as well as how the maximization and minimization algorithms operate.

(a) χ Minimization. (b) χ Maximization.

Figure 2 Matching procedures within one stratum. Matches are made above the dotted line and represented by the double arrow. Units below the line are discarded.

E.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 χ as defined in Eq (47), and output by algorithms 4 and 5 respectively. Let $M_l = \min(N_l^t, N_l^c)$ denote the number of matches that are made in each stratum. Then they can be written as follows:*

$$TE^+ = \sum_{l=1}^{L} TE_l^+ = \sum_{l=1}^{L} U_l^+ - \eta_l^+ \tag{56}$$

$$TE^- = \sum_{l=1}^{L} TE_l^- = \sum_{l=1}^{L} U_l^- - \eta_l^- \tag{57}$$

$$SD^+ = \begin{cases} S^+ = \sum_{l=1}^{L} |U_l^+ - \eta_l^+| & \text{if } TE^+ \geq 1 \\ R^+ = \sum_{l=1}^{L} M_l - |U_l^+ + \eta_l^+ - M_l| & \text{if } TE^+ < 1 \end{cases} \tag{58}$$

$$SD^- = \begin{cases} S^- = \sum_{l=1}^{L} |U_l^- - \eta_l^-| & \text{if } TE^- < 1 \\ R^- = \sum_{l=1}^{L} M_l - |U_l^- + \eta_l^- - M_l| & \text{if } TE^- \geq 1 \end{cases} . \tag{59}$$

Proof. Note first that the definition of TE^+ and TE^- follows directly from lines 3-5 and 8 of Algorithms 4 and 5 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 χ:

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

Now define:

$$SD_l^+ = \begin{cases} S_l^+ = |U_l^+ - \eta_l^+| & \text{If } TE^+ \geq 1 \\ R_l^+ = M_l - |U_l^+ + \eta_l^+ - M_l| & \text{If } TE^+ < 1 \end{cases}. \tag{60}$$

It is clear from this definition that $SD^+ = \sum_{l=1}^{L} SD_l^+$, and since SD_l^+ has the same definition in all strata, it suffices to prove that $B_l^+ + C_l^+$ can be written as in Eq. (60) for one stratum to prove the equality in Eq. (58). The rest of the proof is concerned with establishing this result.

Consider the computation for SD^+ occurring in Alg. 4: the algorithm checks explicitly if $TE \geq 1$ at line 8 and, if true it calls Alg. 3 on inputs $(U_l^+, V_l^+, \eta_l^+, \nu_l^+)$ to generate the matched pair counts $A_l^+, B_l^+, C_l^+, D_l^+$, if false it generates the same quantities by calling Alg. 2 on the same inputs.

Consider now the case in which $TE^+ \geq 1$, and Algorithm 3 is called with $(U_l^+, V_l^+, \eta_l^+, \nu_l^+)$ as inputs. We know that the algorithm returns the following counts for B and C pairs:

$$\begin{aligned} B_l^{(3)} &= \min(U_l^+ - A, V_l^+ - D) && \text{(By line 3 of Alg. 3)} \\ &= \min(U_l^+ - \min(U_l^+, \eta_l^+), \nu_l^+ - \min(\nu_l^+, V_l^+)), && \text{(By lines 1 and 2 of Alg. 3)} \end{aligned} \tag{61}$$

and

$$\begin{aligned} C_l^{(3)} &= \min(\eta_l^+ - A, V_l^+ - D) && \text{(By line 4 of Alg. 3)} \\ &= \min(\eta_l^+ - \min(U_l^+, \eta_l^+), V_l^+ - \min(\nu_l^+, V_l^+)). && \text{(By lines 1 and 2 of Alg. 3)} \end{aligned} \tag{62}$$

Case 1 for $TE^+ \geq 1$: $U_l^+ \geq \eta_l^+$.
Note first that, in this case:

$$\begin{aligned} & U_l^+ \geq \eta_l^+ && && (63) \\ \implies & M_l - \eta_l^+ \geq M_l - U_l^+ && \text{(Add } M_l \text{ to both sides)} && (64) \\ \implies & \nu_l^+ \geq V_l^+. && \text{(By definition of } \nu_l^+, V_l^+) && (65) \end{aligned}$$

Because of this we can see that:

$$\begin{aligned} & B_l^{(3)} + C_l^{(3)} \\ & = \min(U_l^+ - \min(U_l^+, \eta_l^+), \nu_l^+ - \min(\nu_l^+, V_l^+)) \\ & + \min(\eta_l^+ - \min(U_l^+, \eta_l^+), V_l^+ - \min(\nu_l^+, V_l^+)) && \text{(By (61) and (62))} \\ & = \min(U_l^+ - \eta_l^+, \nu_l^+ - V_l^+) + \min(\eta_l^+ - \eta_l^+, V_l^+ - V_l^+) && \text{(By assumption of this case and result in (65))} \\ & = \min(U_l^+ - \eta_l^+, M_l - U_l^+ - (M_l - \eta_l^+)) && \text{(By definition of } V_l^+ \text{ and } \nu_l^+) \\ & = U_l^+ - \eta_l^+ \\ & = |U_l^+ - \eta_l^+| && \text{(By assumption of this case)} \\ & = S_l^+. && \text{(Definition of } S_l^+) \end{aligned}$$

Case 2 for $TE^+ \geq 1$: $U_l^+ < \eta_l^+$.

Note that in this case Eq. (65), implies that: $\nu_l^+ < V_l^+$. Because of this we have:

$$\begin{aligned}
&B_l^{(3)} + C_l^{(3)} \\
&= \min(U_l^+ - \min(U_l^+, \eta_l^+), \nu_l^+ - \min(\nu_l^+, V_l^+)) \\
&+ \min(\eta_l^+ - \min(U_l^+ - \eta_l^+), V_l^+ - \min(\nu_l^+, V_l^+)) && \text{(By (61) and (62))} \\
&= \min(U_l^+ - U_l^+, \nu_l^+ - \nu_l^+) + \min(\eta_l^+ - U_l^+, V_l^+ - \nu_l^+) && \text{(By assumption of this case and result in (65))} \\
&= \min(\eta_l^+ - U_l^+, M_l - U_l^+ - (M_l - \eta_l^+)) && \text{(By definition of } V_l^+ \text{ and } \nu_l^+) \\
&= \eta_l^+ - U_l^+ \\
&= |U_l^+ - \eta_l^+| && \text{(By assumption of this case)} \\
&= S_l^+. && \text{(Definition of } S_l^+)
\end{aligned}$$

This shows $SD_l^+ = S_l^+$ in the case in which $TE \geq 1$, the first of Eq. (58).

The proof is similar for the case in which $TE < 1$. Now we must show that, in this case $SD_l^+ = R^+$. In this case Algorithm 2 is called with $U_l^+, V_l^+, \eta_l^+, \nu_l^+$ as inputs at line 16 of Algorithm 4. On those inputs, Algorithm 2 will return the following counts of B and C pairs:

$$B_l^{(2)} = \min(U_l^+, \nu_l^+) \qquad \text{(By line 1 of Alg. 2)} \tag{66}$$

and

$$C_l^{(2)} = \min(\eta_l^+, V_l^+). \qquad \text{(By line 1 of Alg. 2)} \tag{67}$$

We proceed as before with two separate cases.

Case 1 for $TE < 1$: $U_l^+ \geq \nu_l^+$.

First note that:

$$\begin{aligned}
\nu_l^+ &= M_l - \eta_l^+ && \text{(By definition of } \nu_l^+) \\
&\leq U_l^+ && \text{(By assumption of this case)} \\
\implies \eta_l^+ &\geq M_l - U_l^+ \\
\implies \eta_l^+ &\geq V_l^+. && \text{(By definition of } V_l^+ \text{)}
\end{aligned} \tag{68}$$

Second, the assumption of this case also implies:

$$\begin{aligned}
\nu_l^+ &= M_l - \eta_l^+ && \text{(By definition of } \nu_l^+) \\
&\leq U_l^+ && \text{(By assumption of this case)} \\
\implies U_l^+ + \eta_l^+ &\geq M_l \\
\implies U_l^+ + \eta_l^+ - M_l &\geq 0 \\
\implies U_l^+ + \eta_l^+ - M_l &= |U_l^+ + \eta_l^+ - M_l|. && \text{(Def. of absolute value)}
\end{aligned} \tag{69}$$

Putting these results together we obtain:

$$\begin{aligned}
B_l^{(2)} + C_l^{(2)} &= \min(U_l^+, \nu_l^+) + \min(\eta_l^+, V_l^+) && \text{(By (66) and (67))} && (70)\\
&= \nu_l^+ + V_l^+ && \text{(By assumption of this case and (68))} && (71)\\
&= M_l - U_l^+ + M_l - \eta_l^+ && \text{(By definition of } V_l^+ \text{ and } \nu_l^+) && (72)\\
&= M_l - (U_l^+ + \eta_l^+ - M_l) && && (73)\\
&= M_l - |U_l^+ + \eta_l^+ - M_l| && \text{(By (69))} && (74)\\
&= R_l^+. && \text{(Definition of } R_l^+) && (75)
\end{aligned}$$

Case 2 for $TE < 1$: $U_l^+ < \nu_l^+$.

This assumption of this case together with (68) implies $\eta_l^+ < V_l^+$. Note also that:

$$\begin{aligned}
\nu_l^+ &= M_l - \eta_l^+ && \text{(By By definition of } \nu_l^+)\\
&> U_l^+ && \text{(By assumption of this case)}\\
\implies & U_l^+ + \eta_l^+ - M_l < 0\\
\implies & M_l - U_l^+ - \eta_l^+ > 0\\
\implies & M_l - U_l^+ - \eta_l^+ = |M_l - U_l^+ - \eta_l^+|. && \text{(Def. of absolute value)} && (76)
\end{aligned}$$

With the two results above we obtain:

$$\begin{aligned}
B_l^{(2)} + C_l^{(2)} &= \min(U_l^+, \nu_l^+) + \min(\eta_l^+, V_l^+) && \text{(By (66) and (67))} && (77)\\
&= U_l^+ + \eta_l^+ && \text{(By assumption of this case and (68))} && (78)\\
&= M_l - M_l + U_l^+ + \eta_l^+ && \text{(Add and subtract } M_l) && (79)\\
&= M_l - (M_l - U_l^+ - \eta_l^+) && && (80)\\
&= M_l - |M_l - U_l^+ - \eta_l^+| && \text{(By (76))} && (81)\\
&= M_l - |U_l^+ + \eta_l^+ - M_l| && \text{(By def. of absolute value)} && (82)\\
&= R_l^+. && \text{(Definition of } R_l^+) && (83)
\end{aligned}$$

This proves that $SD_l^+ = R_l^+$ in the case in which $TE^+ < 1$, the second case in Equation (58).

E.2. Lemma 1

LEMMA 1. *(Randomization Distribution of Truncated Variables) For* $l = 1, \ldots, L$ *let* (N_l^1, N_l^0, e_l) *be fixed and known and let* $\mathbb{D} = (U_l, V_l, \eta_l, \nu_l, N_l^t, N_l^c)$ *be drawn from the data generating process of equations* (41) – (46). *Let* a, b, c, d, m *be elements of* $\{0, \ldots, \min(N_l^0, N_l^2)\}$ *and let* $m \in \{0, \ldots, N_l\}$. *The variables* (U_l^+, U_l^-, η_l^+, η_l^-, M_l) *have the following joint distribution:*

$$\begin{aligned}
&\Pr(U_l^+ = a, U_l^- = b, \eta_l^+ = c, \eta_l^- = d, M_l = m | \mathbb{H}_0^{sharp})\\
&= \sum_{j=0}^{N_l^1} \sum_{k=0}^{N_l^0} \Pr(U_l = j | \mathbb{H}_0^{sharp}) \Pr(V_l = k | \mathbb{H}_0^{sharp}) \mathbb{I}(U_l^+ = a | U_l = j, V_l = k) \mathbb{I}(U_l^- = b | U_l = j, V_l = k)\\
&\qquad \times \mathbb{I}(\eta_l^+ = c | U_l = j, V_l = k) \mathbb{I}(\eta_l^- = d | U_l = j, V_l = k) \mathbb{I}(M_l = m | U_l = j, V_l = k), \qquad (84)
\end{aligned}$$

where:

$$\mathbb{I}(U_l^+ = a | U_l = j, V_l = k) = \begin{cases} 1 & \text{if } j = N_l - k - a,\ k \leq N_l - 2a \\ 1 & \text{if } j = a,\ k \leq N_l - 2a \\ 0 & \text{otherwise.} \end{cases} \tag{85}$$

$$\mathbb{I}(U_l^- = b | U_l = j, V_l = k) = \begin{cases} 1 & \text{if } j = N_l - 2k - b,\ k \leq N_l - 2b \\ 1 & \text{if } j = b,\ k \leq \frac{N_l - 2b}{2} \\ 1 & \text{if } b = 0,\ j \geq \frac{N_l}{2} - k,\ k \geq \frac{N_l - j}{2} \\ 1 & \text{if } b = 0,\ j = 0,\ k \leq \frac{N_l}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{86}$$

$$\mathbb{I}(\eta_l^+ = c | U_l = j, V_l = k) = \begin{cases} 1 & \text{if } j = c + N_l^0 - 2k,\ k \geq \frac{N_l^0 - N_l^1 + 2c}{2} \\ 1 & \text{if } j = N_l^1 - c,\ k \geq \frac{N_l^0 - N_l^1 + 2c}{2} \\ 1 & \text{if } c = 0,\ j \leq N_l^0 - 2k,\ k \geq \frac{N_l - 2j}{2} \\ 1 & \text{if } c = 0,\ j = N_l^1,\ k \geq \frac{N_l^0 - N_l^1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{87}$$

$$\mathbb{I}(\eta_l^- = d | U_l = j, V_l = k) = \begin{cases} 1 & \text{if } j = d - k,\ k \geq 2d - N_l^1 \\ 1 & \text{if } j = N_l^1 - d, k \geq 2d - N_l^1 \\ 0 & \text{otherwise.} \end{cases} \tag{88}$$

$$\mathbb{I}(M_l = m | U_l = j, V_l = k) = \begin{cases} 1 & \text{if } j = m - k,\ m \leq \frac{N_l}{2} \\ 1 & \text{if } j = N_l - k - m,\ m \leq \frac{N_l}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{89}$$

and $\Pr(U_l = j | \mathbb{H}_0^{sharp}) = Bin(j, N_l^1, e_l)$ *and* $\Pr(V_l = k | \mathbb{H}_0^{sharp}) = Bin(k, N_l^0, e_l)$.

Note that even though 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_l.

Proof Throughout the proof we maintain Assumption 4 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 $\Pr(U_l^+ = a, U_l^- = b, \eta_l^+ = c, \eta_l^- = d, M_l = m | \mathbb{H}_0^{sharp})$ given in Eq. (84) follows from the law of total probability, independence of U_l and V_l and conditional independence of $U_l^+, V_l^+, \eta_l^+, \eta_l^-, M_l$ given U_l and V_l. It remains to show that the forms for the indicator functions in Equations (85)–(89) 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_l and V_l. 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:

$$\begin{aligned} U_l &\sim Bin(N_l^1, e_l) \\ \eta_l &= N_l^1 - U_l \\ V_l &\sim Bin(N_l^0, e_l) \\ \nu_l &= N_l^0 - V_l \\ N_l^t &= U_l + V_l \\ N_l^c &= \eta_l + \nu_l \\ U_l^+ &= U_l - \max(U_l - N_l^c, 0) \end{aligned}$$

$$
\begin{aligned}
U_l^- &= \max(U_l - G_l^-, 0) \\
\eta_l^+ &= \max(\eta_l - G_l^+, 0) \\
\eta_l^- &= \eta_l - \max(\eta_l - N_l^t, 0) \\
M_l &= \min(N_l^t, N_l^c) \\
G_l^+ &= \max(N_l^c - N_l^t, 0) \\
G_l^- &= \max(N_l^t - N_l^c, 0).
\end{aligned}
$$

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

$$
\begin{aligned}
U_l^+ &= U_l - \max(U_l - N_l^c, 0) && \text{(By definition of } U_l) \\
&= U_l - \max(U_l - \eta_l - \nu_l, 0) && \text{(By definition of } N_l^c) \\
&= U_l - \max(U_l - N_l^1 + U_l - N_l^0 + V_l, 0) && \text{(By definition of } \eta_l, \nu_l) \\
&= U_l - \max(2U_l + V_l - N_l, 0). && \text{(Because } N_l^1 + N_l^0 = N_l) \qquad (90)
\end{aligned}
$$

There are now two cases for the event $U_l^+ = a$, depending on how the max function in the definition of U_l^+ is resolved:

Case 1 for $U_l^+ = a$:

$$
\max(2U_l + V_l - N_l, 0) = 2U_l + V_l - N_l. \qquad (91)
$$

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

$$
\begin{aligned}
a = U_l^+ &= U_l - 2U_l - V_l + N_l \qquad \text{(By (90))} \\
\implies U_l &= N_l - V_l - a, \qquad (92)
\end{aligned}
$$

which is the first condition in the first case of Equation (85). Note that the condition of this case given in Eq. (91) implies that $2U_l + V_l - N_l \geq 0$, and we can use the result in (92) to expand this condition as follows:

$$
\begin{aligned}
0 &\leq 2U_l + V_l - N_l && \text{(By (91))} \\
&= 2(N_l - V_l - a) + V_l - N_l && \text{(By (92))} \\
&= N_l - V_l - 2a \\
\implies V_l &\leq N_l - 2a,
\end{aligned}
$$

which is the second condition in the first case of Equation (85).

Case 2 for $U_l^+ = a$:

$$
\max(2U_l + V_l - N_l, 0) = 0. \qquad (93)
$$

Using this case and the result in (90), we can write the event of interest as

$$a = U_l^+ = U_l, \tag{94}$$

the first condition in the second case of Equation (85). Second, the fact that the max function equals 0 in Case 2, implies that $2U_l + V_l - N_l \leq 0$ is another condition for this case. Using the result in (94) this can be simplified as:

$$\begin{aligned} 0 &\geq 2U_l + V_l - N_l && \text{(By (93))} \\ &= 2a + V_l - N_l && \text{(By (94))} \\ \implies V_l &\leq N_l - 2a, \end{aligned}$$

which is the second condition in the second case of Equation (85).

We proceed in the same manner for the event $U_l^- = b$. In Equation (86) we start by expanding the definition of U_l^-:

$$\begin{aligned} U_l^- &= \max(U_l - \max(N_l^t - N_l^c, 0), 0) && \text{(By definition of } U_l^-) \\ &= \max(U_l - \max(U_l + V_l - \eta_l - \nu_l, 0), 0) && \text{(By definition of } N_l^t, N_l^c) \\ &= \max(U_l - \max(U_l + V_l - N_l^1 + U_l - N_l^0 + V_l, 0), 0) && \text{(By definition of } \eta_l, \nu_l) \\ &= \max(U_l - \max(2(U_l + V_l) - N_l, 0), 0). && \text{(Because } N_l = N_l^1 + N_l^0) \end{aligned} \tag{95}$$

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 for $U_l^- = b$:

$$\max(2(U_l + V_l) - N_l, 0) = 2(U_l + V_l) - N_l \tag{96}$$

$$\max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) = U_l - \max(2(U_l + V_l) - N_l, 0). \tag{97}$$

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

$$b = U_l^- = U_l - 2(U_l + V_l) + N_l \qquad \text{(By (95), (96), (97))} \tag{98}$$

$$\implies U_l = N_l - 2V_l - b, \tag{99}$$

which is the first condition in the first case of Equation (86). Note now that the condition in (97) implies that $U_l - \max(2(U_l + V_l) - N_l, 0) \geq 0$, combining this with the above we obtain:

$$\begin{aligned} 0 &\leq U_l - \max(2(U_l + V_l) - N_l, 0) \\ &= b && \text{(By (98))} \\ &\geq 0, && \text{(By definition)} \end{aligned}$$

so this condition is always satisfied because of how we restrict the domain of b. The condition in (96) implies that $2(U_l + V_l) - N_l \geq 0$, again we use the result in (99) to expand this as follows:

$$\begin{aligned} 0 &\leq 2(U_l + V_l) - N_l \\ &= 2(N_l - 2V_l - b + V_l) - N_l \qquad \text{(By (99))} \\ &= N_l - V_l - 2b \\ \implies V_l &< N_l - 2b, \end{aligned}$$

the second condition in the first case of Eq. (86).

Case 2 for $U_l^- = b$:

$$\max(2(U_l + V_l) - N_l, 0) = 0 \tag{100}$$

$$\max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) = U_l - \max(2(U_l + V_l) - N_l, 0). \tag{101}$$

The first condition in the second case of Eq. (86) follows from using these conditions with the event $U_l^- = b$:

$$\begin{aligned} b = U_l^- &= \max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) \qquad \text{(By (95))} \\ &= U_l - \max(2(U_l + V_l) - N_l, 0) \qquad \text{(By (101))} \\ &= U_l. \qquad \text{(By (100))} \end{aligned} \tag{102}$$

The condition in (101) 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:

$$\begin{aligned} 0 &\leq U_l - \max(2(U_l + V_l) - N_l, 0) \\ &= U_l \qquad \text{(By (100))} \\ &= b \qquad \text{(By (102))} \\ &\geq 0. \qquad \text{(By definition)} \end{aligned}$$

Because of this, this condition is omitted from the formulation in Eq. (86). Finally, condition (100) implies:

$$\begin{aligned} 0 &\geq 2(U_l + V_l) - N_l \\ &= 2(b + V_l) - N_l \qquad \text{(By (102))} \\ \implies V_l &\leq \frac{N_l - 2b}{2}, \end{aligned}$$

the second condition of the case.

Case 3 for $U_l^- = b$:

$$\max(2(U_l + V_l) - N_l, 0) = 2(U_l + V_l) - N_l \tag{103}$$

$$\max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) = 0. \tag{104}$$

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

$$\begin{aligned} b = U_l^- &= \max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) && \text{(By (95))} \\ &= 0, && \text{(By (104))} \end{aligned}$$

which leads us to the first condition in case 3 of Equation (86). For the second condition, start with (103), then we have:

$$\begin{aligned} 2(U_l + V_l) - N_l &\geq 0 \qquad \text{(By (103))} \\ \implies U_l &\geq \frac{N_l}{2} - V_l, \end{aligned} \tag{105}$$

which is the form of the second condition of Case 4 in (86). Finally, from Condition (104) we have:

$$\begin{aligned} 0 &\geq U_l - \max(2(U_l + V_l) - N_l, 0) \text{(By (104))} \\ &= U_l - 2(U_l + V_l) + N_l \qquad \text{(By (103))} \\ \implies V_l &\geq \frac{N_l - U_l}{2}. \end{aligned} \tag{106}$$

Clearly, (106) is the last condition in the fourth case of Equation (86).

Case 4 for $U_l^- = b$:

$$\max(2(Ul + V_l) - N_l, 0) = 0 \tag{107}$$

$$\max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) = 0. \tag{108}$$

First, the event $U_l^- = b$ takes the following form in this case:

$$\begin{aligned} b = U_l^- &= \max(U_l - \max(2(U_l + V_l) - N_l, 0), 0) \qquad \text{(By definition of } U_l^-) \\ &= 0, \qquad \text{(By (108))} \end{aligned}$$

which is the first condition in the fourth case of Eq. (86). Second, we have:

$$\begin{aligned} 0 &\geq U_l - \max(2(U_l + V_l) - N_l, 0) \qquad \text{(By (108))} \\ &= U_l \qquad \text{(By (107))} \\ \implies U_l &\leq 0, \end{aligned} \tag{109}$$

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

$$U_l = 0. \tag{110}$$

Finally we have:

$$\begin{aligned} 0 &\geq 2(U_l + V_l) - N_l \qquad \text{(By (107))} \\ &= 2V_l - N_l \qquad \text{(By (110))} \\ \implies V_l &\leq \frac{N_l}{2}, \end{aligned} \tag{111}$$

the third condition in the fourth case of (86).

As we did before, we can now expand the definition of η_l^+ to derive conditions on U_l and V_l that lead to the realization of the event $\eta_l^+ = c$. Starting with the definition of η_l^+ we have:

$$\begin{aligned}
\eta_l^+ &= \max(\eta_l - G_l^+, 0) && \text{(By definition of } \eta_l^+\text{)} \\
&= \max(\eta_l - \max(N_l^c - N_l^t, 0), 0) && \text{(By definition of } G_l^+\text{)} \qquad (112) \\
&= \max(\eta_l - \max(\eta_l + \nu_l - U_l - V_l, 0), 0) && \text{(By definition of } N_l^t, N_l^c\text{)} \\
&= \max(N_l^1 - U_l - \max(N_l^1 - U_l + N_l^0 - V_l - U_l - V_l, 0), 0) && \text{(By definition of } \eta_l, \nu_l\text{)} \\
&= \max(N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0), 0). && (113)
\end{aligned}$$

Since there are two nested max functions in the definition of η_l^+, 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 η_l^+. Below we study each case separately and show that they lead to the four cases in the indicator function of Eq. (87).

Case 1 for $\eta_l^+ = c$:

$$\max(N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0), 0) = N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0) \qquad (114)$$

$$\max(N_l - 2(U_l + V_l), 0) = N_l - 2(U_l + V_l). \qquad (115)$$

First, use both condition to derive a form for the event $\eta_l^+ = c$:

$$\begin{aligned}
c = \eta_l^+ &= \max(N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0), 0) && \text{(By (113))} \\
&= N_l^1 - U_l - N_l + 2(U_l + V_l) && \text{(By (114) and (115))} \\
&= U_l + 2V_l - N_l^0 \\
\implies U_l &= c + N_l^0 - 2V_l, && (116)
\end{aligned}$$

which is the first condition in the first case of Equation (87). Second, we use the result above to rewrite the first condition:

$$\begin{aligned}
0 &\le N_l - 2(U_l + V_l) && \text{(By (115))} \\
&= N_l - 2(c + N_l^0 - 2V_l + V_l) && \text{(By (116))} \\
&= N_l^1 - N_l^0 - 2c + 2V_l \\
\implies V_l &\ge \frac{N_l^0 - N_l^1 + 2c}{2},
\end{aligned}$$

which is the second condition in the first case in Equation (87).

Case 2 for $\eta_l^+ = c$:

$$\max(N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0), 0) = N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0) \qquad (117)$$

$$\max(N_l - 2(U_l + V_l), 0) = 0. \qquad (118)$$

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

$$\begin{aligned} c = \eta_l^+ &= \max(N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0), 0) && \text{(By (113))} \\ &= N_l^1 - U_l && \text{(By (117) and (118))} \\ \implies U_l &= N_l^1 - c. && (119) \end{aligned}$$

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

$$\begin{aligned} 0 &\geq N_l - 2(U_l + V_l) && \text{(By (118))} \\ &= N_l - 2(N_l^1 - c + V_l) && \text{(By (119))} \\ &= N_l^0 - N_l^1 + 2c - 2V_l \\ \implies V_l &\geq \frac{N_l^0 - N_l^1 + 2c}{2}. \end{aligned}$$

This, and Equation (119) are the conditions in the second case of Equation (87).

Case 3 for $\eta_l^+ = c$:

$$\max(N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0), 0) = 0 \tag{120}$$

$$\max(N_l - 2(U_l + V_l), 0) = N_l - 2(U_l + V_l). \tag{121}$$

The first condition of Case 3 in Eq. (87) is given by:

$$\begin{aligned} c = \eta_l^+ &= \max(N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0), 0) && \text{(By (113))} \\ &= 0. && \text{(By (120))} \end{aligned}$$

Now use (121) to rewrite the (120):

$$\begin{aligned} 0 &\geq N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0) && \text{(By (120))} \\ &= N_l^1 - U_l - N_l + 2(U_l + V_l) && \text{(By (121))} \\ &= U_l - N_l^0 + 2V_l \\ \implies U_l &\leq N_l^0 - 2V_l, && (122) \end{aligned}$$

this is the second condition of the third case in Eq. (87). The final condition in the case is given by:

$$\begin{aligned} 0 &\leq N_l - 2(U_l + V_l) && \text{(By (121))} \\ \implies V_l &\geq \frac{N_l - 2U_l}{2}. \end{aligned}$$

Case 4 for $\eta_l^+ = c$:

$$\max(N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0), 0) = 0 \tag{123}$$

$$\max(N_l - 2(U_l + V_l), 0) = 0. \tag{124}$$

The first condition in the fourth case of Eq. (87) is obtained as follows:

$$\begin{aligned} c = \eta_l^+ &= \max(N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0), 0) && \text{(By (113))} \\ &= 0. && \text{(By (123))} \end{aligned}$$

The second condition in the same case can be obtained by starting from (123):

$$\begin{aligned} 0 &\geq N_l^1 - U_l - \max(N_l - 2(U_l + V_l), 0) && \text{(By (123))} \\ &= N_l^1 - U_l && \text{(By (124))} \\ \implies U_l &\geq N_l^1. \end{aligned}$$

Since U_l is a binomial random variable with number of trials N_l^1, it can never be greater then N_l^1, so the condition above becomes:

$$U_l = N_l^1, \tag{125}$$

which is the second condition in the fourth case of Eq. (87). Now rearrange the terms in the second condition and use the result above to obtain the final condition in the case:

$$\begin{aligned} 0 &\geq N_l - 2(U_l + V_l) \\ &= N_l - 2N_l^1 + 2V_l && \text{(By (125))} \\ &= N_l^0 - N_l^1 + 2V_l && \text{(Because } N_l^1 + N_l^0 = N_l\text{)} \\ \implies V_l &\geq \frac{N_l^0 - N_l^1}{2}. \end{aligned}$$

This concludes the derivation of Eq. (87).

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

$$\begin{aligned} \eta_l^- &= \eta_l - \max(\eta_l - N_l^t, 0) && \text{(By definition of } \eta_l^-\text{)} \\ &= N_l^1 - U_l - \max(N_l^1 - U_l - U_l - V_l, 0) && \text{(By definition of } \eta_l \text{ and } N_l^t\text{)} \\ &= N_l^1 - U_l - \max(N_l^1 - 2U_l - V_l, 0). && (126) \end{aligned}$$

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

Case 1 for $\eta_l^- = d$:

$$\max(N_l^1 - 2U_l - V_l, 0) = N_l^1 - 2U_l - V_l. \tag{127}$$

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

$$\begin{aligned} d = \eta_l^- &= N_l^1 - U_l - \max(N_l^1 - 2U_l - V_l, 0) && \text{(By (126))} \\ &= N_l^1 - U_l - N_l^1 + 2U_l + V_l && \text{(By (127))} \\ &= U_l + V_l \\ \implies U_l &= d - V_l. && (128) \end{aligned}$$

Second, we use the result just derived to rewrite the condition for the max in (127) in this case:

$$\begin{aligned} 0 &\leq N_l^1 - 2U_l - V_l && \text{(By (127))} \\ &= N_l^1 - 2(d - V_l) - V_l && \text{(By (128))} \\ \implies V_l &\geq 2d - N_l^1. \end{aligned}$$

This and (128) are, respectively, the first and second conditions on the first case in Equation (88).

Case 2 for $\eta_l^- = d$:

$$\max(N_l^1 - 2U_l - V_l, 0) = 0. \tag{129}$$

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

$$\begin{aligned} d = \eta_l^- &= N_l^1 - U_l - \max(N_l^1 - 2U_l - V_l, 0) && \text{(By (126))} \\ &= N_l^1 - U_l && \text{(By (129))} \\ \implies U_l &= N_l^1 - d. && (130) \end{aligned}$$

Second, we rewrite the condition on the max with the above result:

$$\begin{aligned} 0 &\geq N_l^1 - 2U_l - V_l && \text{(By (129))} \\ &= N_l^1 - 2(N_l^1 - d) - V_l && \text{(By (130))} \\ \implies V_l &\geq 2d - N_l^1. \end{aligned}$$

These are the condition in the second case of Equation (88).

Finally we have the event $M_l = m$: conditions on U_l and V_l that lead to its realization can again be established by expanding its definition, this will lead to the form of M_l in Equation (89). Starting with the definition of M_l:

$$\begin{aligned} M_l &= \min(N_l^t, N_l^c) && \text{(By definition of } M_l) \\ &= \min(U_l + V_l, \eta_l + \nu_l) && \text{(By definition of } N_l^t, N_l^c) \\ &= \min(U_l + V_l, N_l^1 - U_l + N_l^0 - V_l) && \text{(By definition of } \eta_l, \nu_l) \\ &= \min(U_l + V_l, N_l - (U_l + V_l)). && \text{(Because } N_l^1 + N_l^0 = N_l) \end{aligned} \tag{131}$$

Because of this we see that the definition of M_l is composed of two cases that depend on how the min function is resolved.

Case 1 for $M_l = m$:

$$\min(U_l + V_l, N_l - (U_l + V_l)) = U_l + V_l. \tag{132}$$

First, use the expanded definition of M_l with the case above to rewrite the event $M_l = m$:

$$\begin{aligned} m = M_l &= \min(U_l + V_l, N_l - (U_l + V_l)) && \text{(By (131))} \\ &= U_l + V_l && \text{(By (132))} \\ \implies U_l &= m - V_l. \end{aligned} \tag{133}$$

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

$$\begin{aligned} U_l + V_l &\leq N_l - (U_l + V_l) && \text{(By (132))} \\ \implies 0 &\geq U_l + V_l - N_l + (U_l + V_l) \\ &= 2U_l + 2V_l - N_l \\ &= 2m - N_l && \text{(By (133))} \\ \implies m &\leq N_l/2. \end{aligned}$$

leading us to both conditions in the first case of Equation (89).

Case 2 for $M_l = m$:

$$\min(U_l + V_l, N_l - (U_l + V_l)) = N_l - (U_l + V_l). \tag{134}$$

Again, we use the case above to rewrite the definition of the event $M_l = m$:

$$\begin{aligned} m = M_l &= \min(U_l + V_l, N_l - (U_l + V_l)) && \text{(By (131))} \\ &= N_l - (U_l + V_l) && \text{(By (134))} \end{aligned} \tag{135}$$

$$\implies U_l = N_l - V_l - m. \tag{136}$$

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

$$\begin{aligned} U_l + V_l &\geq N_l - (U_l + V_l) && \text{(By (134))} \\ \implies N_l - V_l - m + V_l &\geq m && \text{(By (135) and (136))} \\ \implies m &\leq N_l/2. \end{aligned}$$

These are the two conditions in the definition of $M_l = m$ in equation (89).

Finally, the forms in equation (84) are simply obtained by conditioning on the event $U_l = j, V_l = k$ for any of the definitions above. This concludes the proof of the lemma.

E.3. Theorem 7

The joint null distribution of (χ^+, χ^-) is given in the following theorem:

THEOREM 7. *(Randomization Distribution of (χ^+, χ^-)). For $l = 1, \ldots, L$ let (N_l^1, N_l^0, e_l) be fixed and known and let data, $\mathbb{D} = (U_l, V_l, \eta_l, \nu_l, N_l^t, N_l^c)$, be drawn from the data generating process of Equations* (41) – (46)*. Let χ^+ be the maximum of Formulation 2 on $\mathbb{D}$, and let χ^- be the minimum of the problem on the*

same variables. Let $N^m = \sum_{l=1}^{L} \min(N_l^1, N_l^0)$ *and define:* $\mathcal{X}_{N^m} = \left\{ \frac{b-c-1}{\sqrt{b+c+1}} : b, c \in \{0, \dots, N^m\} \right\}$*, Additionally define:*

$$\mathcal{A}(y) = \left\{ \mathbf{a} = (a_1, \dots, a_l) : \sum_{l=1}^{L} a_l = y,\ a_l \in \{0, \dots, y\} \right\}, \tag{137}$$

$$\mathcal{B}(y,s) = \left\{ \mathbf{b} = (b_1, \dots, b_l) : \sum_{l=1}^{L} b_l = \left(\frac{y-1}{s}\right)^2 - 1,\ b_l \in \left\{ 0, \dots, \left(\frac{y-1}{s}\right)^2 - 1 \right\} \right\}, \tag{138}$$

$$\mathcal{C}(x) = \left\{ \mathbf{c} = (c_1, \dots, c_l) : \sum_{l=1}^{L} c_l = x,\ c_l \in \{0, \dots, x\} \right\}, \tag{139}$$

$$\mathcal{D}(x,r) = \left\{ \mathbf{d} = (d_1, \dots, d_l) : \sum_{l=1}^{L} d_l = \left(\frac{x-1}{r}\right)^2 - 1,\ d_l \in \left\{ 0, \dots, \left(\frac{x-1}{r}\right)^2 - 1 \right\} \right\}. \tag{140}$$

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:* $(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$*. Let* $N^1 = \sum_{l=1}^{L} N_l^1$*; the pmf of* (χ^+, χ^-)*, for two values* $(s,r) \in \mathcal{X}_{N^m} \times \mathcal{X}_{N^m}$ *is:*

$$\begin{aligned} &\Pr(\chi^- = s, \chi^+ = r | \mathbb{H}_0^{sharp}) = \\ &\sum_{x=-N^m}^{N^m} \sum_{y=-N^m}^{N^m} \sum_{(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) \in \mathcal{H}(x,y,r,s)} \prod_{l=1}^{L} \begin{cases} h_1(a_l, b_l, c_l, d_l) & \text{if } x < 1, y < 1 \\ h_2(a_l, b_l, c_l, d_l) & \text{if } x \geq 1, y < 1 \\ h_3(a_l, b_l, c_l, d_l) & \text{if } x \geq 1, y \geq 1 \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{141}$$

where:

$$\begin{aligned} h_1(a_l, b_l, c_l, d_l) = \mathbb{I}(|a_l| = b_l) \sum_{m=0}^{N_l/2} \sum_{j=0}^{m} \Bigg\{ & \\ \Pr\left(U_l^- = a_l + j, U_l^+ = \frac{2m - d_l + c_l}{2}, \eta_l^- = j, \eta_l^+ = \frac{2m - d_l - c_l}{2}, M_l = m \right) & \\ + \Pr\left(U_l^- = a_l + j, U_l^+ = \frac{d_l + c_l}{2}, \eta_l^- = j, \eta_l^+ = \frac{d_l - c_l}{2}, M_l = m \right) \mathbb{I}(m \neq d_l) \Bigg\} & \end{aligned} \tag{142}$$

$$\begin{aligned} h_2(a_l, b_l, c_l, d_l) = \mathbb{I}(|a_l| = b_l) \mathbb{I}(|c_l| = d_l) & \\ + \sum_{m=0}^{N/2} \sum_{j=0}^{m} \sum_{k=0}^{m} \left\{ \Pr(U_l^- = a_l + k, U_l^+ = c_l + j, \eta_l^- = k, \eta_l^+ = j, M_l = m) \right\} & \end{aligned} \tag{143}$$

$$\begin{aligned} h_3(a_l, b_l, c_l, d_l) = \mathbb{I}(|c_l| = d_l) \sum_{m=0}^{N_l/2} \sum_{j=0}^{m} \Bigg\{ & \\ \Pr\left(U_l^- = \frac{2m - b_l + a_l}{2}, U_l^+ = c_l + j, \eta_l^- = \frac{2m - b_l - a_l}{2}, \eta_l^+ = j, M_l = m \right) & \\ + \Pr\left(U_l^- = \frac{b_l + a_l}{2}, U_l^+ = c_l + j, \eta_l^- = \frac{b_l - a_l}{2}, \eta_l^+ = j, M_l = m \right) \mathbb{I}(m \neq b_l) \Bigg\}. & \end{aligned} \tag{144}$$

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 $\mathcal{A}_{\text{good}}$.

Proof. All the probability statements throughout the proof are made conditionally on X and $\mathbb{H}_0^{sharp}$, 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:

$$\begin{aligned} 0 \leq U_l^+ &= U_l - \max(U_l - N_l^c, 0) && \text{(By definition of } U_l^+\text{)} \\ &\leq N_l^1, && \text{(By definition of } U_l\text{)} \end{aligned} \tag{145}$$

and

$$0 \leq \eta_l^+ = \max(\eta_l - G_l^+, 0) \qquad \text{(By definition of } \eta_l^+\text{)} \tag{146}$$

$$= \max(N_l^1 - U_l - G_l^+, 0) \qquad \text{(By definition of } \eta_l\text{)} \tag{147}$$

$$\leq N_l^1. \tag{148}$$

This implies that:

$$\begin{aligned} TE^+ &= \sum_{l=1}^{L} U_l^+ - \eta_l^+ && \text{(By definition of } TE^+\text{)} \\ &\leq \sum_{l=1}^{L} N_l^1 - 0, && \text{(By (145) and (87))} \\ &= N^1 \end{aligned}$$

and:

$$\begin{aligned} TE^+ &= \sum_{l=1}^{L} U_l^+ - \eta_l^+ && \text{(By definition of } TE^+\text{)} \\ &\geq \sum_{l=1}^{L} 0 - N_l^1. && \text{(By (85) and (148))} \\ &= -N^1, \end{aligned}$$

where recall that $N^1 = \sum_{l=1}^{L} N_l^1$, 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 (χ^+, χ^-), we know that no integers outside the range above will have positive mass under that distribution.

To bound the domain of $\chi = \frac{B-C-1}{\sqrt{B+C+1}}$ 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: $B_l \leq \min(N_l^1, N_l^0)$, independently of how many units are assigned to treatment. For the same reason we conclude that $C_l \leq \min(N_l^1, N_l^0)$ in every stratum, independently of how many units are treated or how the matches are made. Recall that $N^m = \sum_{l=1}^{L} \min(N_l^1, N_l^0)$, using these facts we can see that:

$$\chi \in \mathcal{X}_{N^m} = \left\{ \frac{b-c-1}{\sqrt{b+c+1}} : b, c \in \{0, \ldots, N^m\} \right\}. \tag{149}$$

Note finally that, since $\chi \in \mathcal{X}_{N^m}$ regardless of how matches are made, we know that $\chi^+ \in \mathcal{X}_{N^m}$ and $\chi^- \in \mathcal{X}_{N^m}$ because χ^+ and χ^- are special cases of χ in which matches are made with Algorithms 4 and

5 respectively. Because of this we conclude that $(\chi^+, \chi^-) \in \mathcal{X}_{N^m} \times \mathcal{X}_{N^m}$. This set is fast to enumerate computationally and summations over its elements can be performed efficiently. While the distribution of (χ^+, χ^-) likely does not place positive probability over all values in $\mathcal{X}_{N^m} \times \mathcal{X}_{N^m}$, 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 (χ^+, χ^-). As stated in Theorem 7, assume that N_l^1, N_l^0 are fixed and known for all strata $l = 1, \ldots, L$. For any two values $(r, s) \in \mathcal{X}_{N^m} \times \mathcal{X}_{N^m}$ we have:

$$\begin{aligned}
\Pr(\chi^- = s, \chi^+ = r) &= \Pr\left(\frac{TE^- - 1}{\sqrt{SD^- + 1}} = s, \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} = r\right) && \text{(By def. of } (\chi^+, \chi^-)) \\
&= \Pr\left(TE^- < 1, \frac{TE^- - 1}{\sqrt{S^- + 1}} = s, TE^+ < 1, \frac{TE^+ - 1}{\sqrt{R^+ + 1}} = r\right) && \text{(By Claim 4)} \\
&+ \Pr\left(TE^- < 1, \frac{TE^- - 1}{\sqrt{S^- + 1}} = s, TE^+ \geq 1, \frac{TE^+ - 1}{\sqrt{S^+ + 1}} = r\right) \\
&+ \Pr\left(TE^- \geq 1, \frac{TE^- - 1}{\sqrt{R^- + 1}} = s, TE^+ < 1, \frac{TE^+ - 1}{\sqrt{R^+ + 1}} = r\right) \\
&+ \Pr\left(TE^- \geq 1, \frac{TE^- - 1}{\sqrt{R^- + 1}} = s, TE^+ \geq 1, \frac{TE^+ - 1}{\sqrt{S^+ + 1}} = r\right) && (150) \\
&= f_1(r, z) + f_2(r, z) + f_3(r, z) + f_4(r, z).
\end{aligned}$$

The equality in (150) follows by the representation of SD^+ and SD^- in (58) and (59) 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 χ^+, we know that, in this case:

$$\begin{aligned}
\chi^- &= \frac{TE^- - 1}{\sqrt{SD^- + 1}} && \text{(By definition of } \chi^-) \\
&\geq 0 && \text{(Because of the event } TE^- \geq 1)
\end{aligned}$$

and:

$$\begin{aligned}
\chi^+ &= \frac{TE^+ - 1}{\sqrt{SD^+ + 1}} && \text{(By definition of } \chi^+) \\
&< 0. && \text{(Because of the event } TE^+ < 1)
\end{aligned}$$

Therefore, the event $\{TE^- \geq 1, TE^+ < 1\}$ implies the event $\{\chi^+ < \chi^-\}$, but this is a contradiction, since χ^+ and χ^- 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:

$$\mathcal{A}(y) = \left\{ \mathbf{a} = (a_1, \ldots, a_l) : \sum_{l=1}^{L} a_l = y,\ a_l \in \{0, \ldots, y\} \right\}, \tag{151}$$

$$\mathcal{B}(y,s) = \left\{ \mathbf{b} = (b_1, \ldots, b_l) : \sum_{l=1}^{L} b_l = \left(\frac{y-1}{s}\right)^2 - 1,\ b_l \in \left\{ 0, \ldots, \left(\frac{y-1}{s}\right)^2 - 1 \right\} \right\}, \tag{152}$$

$$\mathcal{C}(x) = \left\{ \mathbf{c} = (c_1, \ldots, c_l) : \sum_{l=1}^{L} c_l = x,\ c_l \in \{0, \ldots, x\} \right\}, \tag{153}$$

$$\mathcal{D}(x,r) = \left\{ \mathbf{d} = (d_1, \ldots, d_l) : \sum_{l=1}^{L} d_l = \left(\frac{x-1}{r}\right)^2 - 1,\ d_l \in \left\{ 0, \ldots, \left(\frac{x-1}{r}\right)^2 - 1 \right\} \right\}. \tag{154}$$

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 $f_1(r,s)$:

$$f_1(r,s) = \Pr\left(TE^- < 1, \frac{TE^- - 1}{\sqrt{S^- + 1}} = s, TE^+ < 1, \frac{TE^+ - 1}{\sqrt{R^+ + 1}} = r \right) \tag{155}$$

$$= \sum_{y=-N^1}^{0} \sum_{x=-N^1}^{0} \Pr\left(TE^- = y, \frac{TE^- - 1}{\sqrt{S^- + 1}} = s, TE^+ = x, \frac{TE^+ - 1}{\sqrt{R^+ + 1}} = r \right) \tag{156}$$

$$= \sum_{y=-N^1}^{0} \sum_{x=-N^1}^{0} \Pr\left(TE^- = y, S^- = \left(\frac{y-1}{s}\right)^2 - 1, TE^+ = x, R^+ = \left(\frac{x-1}{r}\right)^2 - 1 \right) \tag{157}$$

$$= \sum_{y=-N^1}^{0} \sum_{x=-N^1}^{0} \Pr\left(\sum_{l=1}^{L} TE_l^- = y, \sum_{l=1}^{L} S_l^- = \left(\frac{y-1}{s}\right)^2 - 1, \sum_{l=1}^{L} TE_l^+ = x, \sum_{l=1}^{L} R_l^+ = \left(\frac{x-1}{r}\right)^2 - 1 \right) \tag{158}$$

$$= \sum_{y=-N^1}^{0} \sum_{x=-N^1}^{0} \sum_{(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) \in \mathcal{H}(x,y,r,s)} \Pr(TE_1^- = a_1, S_1^- = b_1, TE_1^+ = c_1, R^+ = d_1, \ldots, \tag{159}$$

$$TE_L^- = a_L, S_L^- = b_L, TE_L^+ = c_L, R^+ = d_L) \tag{160}$$

$$= \sum_{y=-N^1}^{0} \sum_{x=-N^1}^{0} \sum_{(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) \in \mathcal{H}(x,y,r,s)} \prod_{l=1}^{L} \Pr\left(TE_l^- = a_l, S_l^- = b_l, TE_l^+ = c_l, R_l^+ = d_l \right). \tag{161}$$

Line (155) follows from the definition of f_1 given in (150). Equation (157) follows from rearranging the terms in the previous line. Equation (159) follows from the definitions in Claim 4. Equation (160) follows from the definition of $\mathcal{H}(x,y,r,s)$, and equation (161) follows from independence of $TE_l^+, TE_l^-, S_l^+, S_l^-, R_l^+, R_l^-$ 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:

$$f_2(r,s) = \Pr\left(TE^- < 1, \frac{TE^- - 1}{\sqrt{S^-}} = s, TE^+ \geq 1, \frac{TE^+ - 1}{\sqrt{S^+}} = r \right)$$

$$= \sum_{y=-N^1}^{0} \sum_{x=1}^{N^1} \Pr\left(TE^- = y, \frac{TE^- - 1}{\sqrt{S^- + 1}} = s, TE^+ = x, \frac{TE^+ - 1}{\sqrt{S^+ + 1}} = r \right)$$

$$= \sum_{y=-N^1}^{0} \sum_{x=1}^{N^1} \Pr\left(TE^- = y, S^- = \left(\frac{y-1}{s}\right)^2 - 1, TE^+ = x, S^+ = \left(\frac{x-1}{r}\right)^2 - 1 \right)$$

$$= \sum_{y=-N^1}^{0}\sum_{x=1}^{N^1} \Pr\left(\sum_{l=1}^{L} TE_l^- = y, \sum_{l=1}^{L} S_l^- = \left(\frac{y-1}{s}\right)^2 - 1, \sum_{l=1}^{L} TE_l^+ = x, \sum_{l=1}^{L} S_l^+ = \left(\frac{x-1}{r}\right)^2 - 1\right)$$

$$= \sum_{y=-N^1}^{0}\sum_{x=1}^{N^1} \sum_{(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})\in\mathcal{H}(x,y,r,s)} \Pr(TE_1^- = a_1, S_1^- = b_1, TE_1^+ = c_1, S^+ = d_1, \ldots$$
$$TE_L^- = a_L, S_L^- = b_L, TE_L^+ = c_L, S^+ = d_L)$$

$$= \sum_{y=-N^1}^{0}\sum_{x=1}^{N^1} \sum_{(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})\in\mathcal{H}(x,y,r,s)} \prod_{l=1}^{L} \Pr\left(TE_l^- = a_l, S_l^- = b_l, TE_l^+ = c_l, S_l^+ = d_l\right).$$

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

$$f_4(r,s) = (TE^- \geq 1, \frac{TE^- - 1}{\sqrt{R^- + 1}} = s, TE^+ \geq 1, \frac{TE^+ - 1}{\sqrt{S^+ + 1}} = r)$$

$$= \sum_{y=1}^{N^1}\sum_{x=1}^{N^1} \Pr\left(TE^- = y, \frac{TE^- - 1}{\sqrt{R^- + 1}} = s, TE^+ = x, \frac{TE^+ - 1}{\sqrt{S^+ + 1}} = r\right)$$

$$= \sum_{y=1}^{N^1}\sum_{x=1}^{N^1} \Pr\left(TE^- = y, R^- = \left(\frac{y-1}{s}\right)^2 - 1, TE^+ = x, S^+ = \left(\frac{x-1}{r}\right)^2 - 1\right)$$

$$= \sum_{y=1}^{N^1}\sum_{x=1}^{N^1} \Pr\left(\sum_{l=1}^{L} TE_l^- = y, \sum_{l=1}^{L} R_l^- = \left(\frac{y-1}{s}\right)^2 - 1, \sum_{l=1}^{L} TE_l^+ = x, \sum_{l=1}^{L} S_l^+ = \left(\frac{x-1}{r}\right)^2 - 1\right)$$

$$= \sum_{y=1}^{N^1}\sum_{x=1}^{N^1} \sum_{(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})\in\mathcal{H}(x,y,r,s)} \Pr(TE_1^- = a_1, R_1^- = b_1, TE_1^+ = c_1, S^+ = d_1, \ldots$$
$$TE_L^- = a_L, R_L^- = b_L, TE_L^+ = c_L, S^+ = d_L)$$

$$= \sum_{y=1}^{N^1}\sum_{x=1}^{N^1} \sum_{(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})\in\mathcal{H}(x,y,r,s)} \prod_{l=1}^{L} \Pr\left(TE_l^- = a_l, R_l^- = b_l, TE_l^+ = c_l, S_l^+ = d_l\right).$$

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 (141). 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 $TE_l^+, TE_l^-, SD_l^+, SD_l^-$ in each stratum. Before doing this, it will be useful to note that, for any stratum l:

$$\begin{aligned} 0 \leq M_l &= \min(N_l^t, N_l^c) && \text{(By definition of } M_l\text{)} \\ &= \min(N_l^t, N_l - N_l^t) && \text{(By definition of } N_l^c\text{)} \\ &\leq N_l/2. \end{aligned}$$

Because of this we consider values of M_l in the integer range $0, \ldots, N_l/2$. Beginning with $\Pr(TE_l^- = a_l, S_l^- = b_l, TE_l^+ = c_l, R_l^+ = d_l)$:

$$\Pr(TE_l^- = a_l, S_l^- = b_l, TE_l^+ = c_l, R_l^+ = d_l) \tag{162}$$

$$= \sum_{m=0}^{N_l/2} \Pr(U_l^- - \eta_l^- = a_l, |U_l^- - \eta_l^-| = b_l, U_l^+ - \eta_l^+ = c_l, M_l - |U_l^+ + \eta_l^+ - M_l| = d_l, M_l = m) \tag{163}$$

$$= \sum_{m=0}^{N_l/2} \Pr(|U_l^- - \eta_l^-| = b_l, U_l^+ - \eta_l^+ = c_l, m - |U_l^+ + \eta_l^+ - m| = d_l, M_l = m | U_l^- - \eta_l^- = a_l)$$

$$\times \Pr(U_l^- - \eta_l^- = a_l) \tag{164}$$

$$= \sum_{m=0}^{N_l/2} \Pr(|a_l| = b_l, U_l^+ - \eta_l^+ = c_l, m - |U_l^+ + \eta_l^+ - m| = d_l, M_l = m | U_l^- - \eta_l^- = a_l)$$
$$\times \Pr(U_l^- - \eta_l^- = a_l) \tag{165}$$

$$= \sum_{m=0}^{N_l/2} \mathbb{I}(|a_l| = b_l) \Pr(U_l^+ - \eta_l^+ = c_l, m - |U_l^+ + \eta_l^+ - m| = d_l, M_l = m | U_l^- - \eta_l^- = a_l)$$
$$\times \Pr(U_l^- - \eta_l^- = a_l) \tag{166}$$

$$= \mathbb{I}(|a_l| = b_l) \sum_{m=0}^{N_l/2} \Pr(U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l, |U_l^+ + \eta_l^+ - m| = m - d_l, M_l = m) \tag{167}$$

$$= \mathbb{I}(|a_l| = b_l) \sum_{m=0}^{N_l/2} \Pr(U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l, U_l^+ - \eta_l^+ + m = m - d_l, M_l = m)$$
$$+ \Pr(U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l, U_l^+ - \eta_l^+ + m = d_l - m, M_l = m) \mathbb{I}(m \neq d_l) \tag{168}$$

$$= \mathbb{I}(|a_l| = b_l) \sum_{m=0}^{N_l/2} \Pr\left(U_l^- = a_l + \eta_l^-, U_l^+ = c_l + \eta_l^+, \eta_l^+ = \frac{2m - d_l - c_l}{2}, M_l = m\right)$$
$$+ \Pr\left(U_l^- = a_l + \eta_l^-, U_l^+ = c_l + \eta_l^+, \eta_l^+ = \frac{d_l - c_l}{2}, M_l = m\right) \mathbb{I}(m \neq d_l) \tag{169}$$

$$= \mathbb{I}(|a_l| = b_l) \sum_{m=0}^{N_l/2} \sum_{j=0}^{m} \Pr\left(U_l^- = a_l + j, U_l^+ = \frac{2m - d_l + c_l}{2}, \eta_l^- = j, \eta_l^+ = \frac{2m - d_l - c_l}{2}, M_l = m\right)$$
$$+ \Pr\left(U_l^- = a_l + j, U_l^+ = \frac{d_l + c_l}{2}, \eta_l^- = j, \eta_l^+ = \frac{d_l - c_l}{2}, M_l = m\right) \mathbb{I}(m \neq d_l) \tag{170}$$

$$= h_1(a_l, b_l, c_l, d_l). \tag{171}$$

Equation (163) follows from the representations of TE and SD given in Claim 4, Equation (164) from the definition of conditional probability, Equation (166) from the fact that a and b are constants and therefore they are independent from the other quantities in the equation. Equation (167) follows from the definition of conditional probability. Equations (168) follows from the fact that the event $\{M_l - |U_l^+ + \eta_l^+ - M_l| = d_l\}$ is equal to the event $\{M_l - U_l^+ - \eta_l^+ + M_l = d_l\} \cup \{M_l + U_l^+ + \eta_l^+ - M_l = d_l, M_l \neq d_l\}$, and therefore the probability of its occurrence is equal to the sum of the probability of these two events. Equation (169) follow by rearranging the terms in the previous line, and Equation (170) from summing over values of η_l^-. 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:

$$\Pr(TE_l^- = a_l, S_l^- = b_l, TE_l^+ = c_l, S_l^+ = d_l)$$
$$= \sum_{m=0}^{N_l/2} \Pr(U_l^- - \eta_l^- = a_l, |U_l^- - \eta_l^-| = b_l, U_l^+ - \eta_l^+ = c_l, |U_l^+ - \eta_l^+| = d_l, M_l = m)$$
$$= \sum_{m=0}^{N_l/2} \Pr(|U_l^- - \eta_l^-| = b_l, |U_l^+ - \eta_l^+| = d_l, M_l = m | U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l)$$
$$\times \Pr(U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l)$$
$$= \sum_{m=0}^{N_l/2} \Pr(|a_l| = b_l, |c_l| = d_l, M_l = m | U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l) \Pr(U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l)$$

$$
\begin{aligned}
&= \sum_{m=0}^{N_l/2} \mathbb{I}(|a_l| = b_l)\mathbb{I}(|c_l| = d_l) \Pr(M_l = m | U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l) \Pr(U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l) \\
&= \mathbb{I}(|a_l| = b_l)\mathbb{I}(|c_l| = d_l) \sum_{m=0}^{N_l/2} \Pr(U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l, M_l = m) \\
&= \mathbb{I}(|a_l| = b_l|)\mathbb{I}(|c_l| = d_l) \sum_{m=0}^{N_l/2} \sum_{j=0}^{m} \sum_{k=0}^{m} \Pr(U_l^- = a_l + k, U_l^+ = c_l + j, \eta_l^- = k, \eta_l^+ = j, M_l = m) \\
&= h_2(a_l, b_l, c_l, d_l).
\end{aligned}
$$

Finally, h_3 can be derived from:

$$
\begin{aligned}
&\Pr(TE_l^- = a_l, R_l^- = b_l, TE_l^+ = c_l, S_l^+ = d_l) \\
&= \sum_{m=0}^{N_l/2} \Pr(U_l^- - \eta_l^- = a_l, M_l - |U_l^- + \eta_l^- - M_l| = b_l, U_l^+ - \eta_l^+ = c_l, |U_l^+ - \eta_l^+| = d_l, M_l = m) \\
&= \sum_{m=0}^{N_l/2} \Pr(U_l^- - \eta_l^- = a_l, m - |U_l^- + \eta_l^- - m| = b_l, |U_l^+ - \eta_l^+| = d_l, M_l = m | U_l^+ - \eta_l^+ = c_l) \\
&\quad \times \Pr(U_l^+ - \eta_l^+ = c_l) \\
&= \sum_{m=0}^{N_l/2} \Pr(U_l^- - \eta_l^- = a_l, m - |U_l^- + \eta_l^- - m| = b_l, |c_l| = d_l, M_l = m | U_l^+ - \eta_l^+ = c_l) \\
&\quad \times \Pr(U_l^+ - \eta_l^+ = c_l) \\
&= \sum_{m=0}^{N_l/2} \mathbb{I}(|c_l| = d_l) \Pr(U_l^- - \eta_l^- = a_l, m - |U_l^- + \eta_l^- - m| = b_l, M_l = m | U_l^+ - \eta_l^+ = c_l) \\
&\quad \times \Pr(U_l^+ - \eta_l^+ = c_l) \\
&= \mathbb{I}(|c_l| = d_l) \sum_{m=0}^{N_l/2} \Pr(U_l^- - \eta_l^- = a_l, m - |U_l^- + \eta_l^- - m| = b_l, U_l^+ - \eta_l^+ = c_l, M_l = m) \\
&= \mathbb{I}(|c_l| = d_l) \sum_{m=0}^{N_l/2} \Pr(U_l^- - \eta_l^- = a_l, m - U_l^- - \eta_l^- + m = b_l, U_l^+ - \eta_l^+ = c_l, M_l = m) \\
&\qquad + \Pr(U_l^- - \eta_l^- = a_l, m + U_l^- + \eta_l^- - m = b_l, U_l^+ - \eta_l^+ = c_l, M_l = m)\mathbb{I}(m \neq b_l) \\
&= \mathbb{I}(|c_l| = d_l) \sum_{m=0}^{N_l/2} \Pr(U_l^- = a_l + \eta_l^-, U_l^- = m - b_l + m - \eta_l^-, U_l^+ = c_l + \eta_l^+, M_l = m) \\
&\qquad + \Pr(U_l^- = a_l + \eta_l^-, U_l^- = b_l - \eta_l^-, U_l^+ = c_l + \eta_l^+, M_l = m)\mathbb{I}(m \neq b_l) \\
&= \mathbb{I}(|c_l| = d_l) \sum_{m=0}^{N_l/2} \sum_{j=1}^{m} \Pr\left(U_l^- = \frac{2m - b_l + a_l}{2}, U_l^+ = c_l + j, \eta_l^- = \frac{2m - b_l - a_l}{2}, \eta_l^+ = j, M_l = m\right) \\
&\qquad + \Pr\left(U_l^- = \frac{b_l + a_l}{2}, U_l^+ = c_l + j, \eta_l^- = \frac{b_l - a_l}{2}, \eta_l^+ = j, M_l = m\right) \mathbb{I}(m \neq b_l) \\
&= h_3(a_l, b_l, c_l, d_l).
\end{aligned}
$$

This concludes the derivation of all the quantities in Theorem 7.

Appendix F: Distribution of (χ^+, χ^-) Under Exclusively Binning Constraints

To compute the distribution we seek, we define the quantities $p_l^t = \Pr(Y = 1 | T = 1, X = x_i, i \in S_l)$ and $\Pr(Y = 1 | T = 0, X = x_j, j \in S_l)$. We can encode $\mathbb{H}_0$ in this distribution by requiring: $p_l^t = p_l^c = p_l$ for all l. In order for the distribution of interest to correctly be estimated, Assumption 4 must hold. Note that this

assumption implies directly that $\Pr(Y(t)=1|X=x_i, i\in S_l) = \Pr(Y=1|T=t, X=x_i, i\in S_l)$. With it, and by requiring $p_l^t = p_l^c$ we obtain $\mathbb{H}_0$: $\Pr(Y=1|T=1, X=x_i, i\in S_l) = \Pr(Y(1)=1|X=x_i, i\in S_l) = \mathbb{E}(Y(1)|X=x_i, i\in S_l) = \mathbb{E}(Y(0)|X=x_i, i\in S_l) = \Pr(Y(0)=1|X=x_i, i\in S_l) = \Pr(Y=1|T=0, X=x_i, i\in S_l)$. As stated before, while Assumption 4 is necessary to obtain the distribution of (χ^+, χ^-).

An estimate of p_l^t and p_l^c can be produced in an unbiased and consistent way with $\hat{p}_l^t = \frac{\sum_{i\in S_l} y_i^t}{N_l^t}$, or any other estimator of choice. When one wants to test the null hypothesis that $p_l^t = p_l^c$, one can use $\hat{p}_l^t = \hat{p}_l^c = \frac{\sum_{i\in S_l} y_i^t + \sum_{j\in S_l} y_j^c}{N_l}$.

Recall that exactly $M_l = \min(N_l^t, N_l^c)$ matches are made, with $M = \sum_{l=1}^L M_l$. 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, $l = 1, \ldots, L$, the data are generated as follows:

$$U_l \overset{iid}{\sim} Bin(p_l^t, N_l^t) \tag{172}$$

$$V_l = N_l^t - U_l \tag{173}$$

$$\eta_l \overset{iid}{\sim} Bin(p_l^c, N_l^c) \tag{174}$$

$$\nu_l = N_l^c - \eta_l. \tag{175}$$

The joint distribution of the truncated variables introduced in Equations (48)-(51) under this DGP is given in the following lemma:

LEMMA 2. *(Distribution of Truncated Variables) 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* (172)-(175)*. The truncated variables have the following joint distributions:*

$$\Pr(U_l^- = a, U_l^+ = b) = \begin{cases} \Pr(U_l = b) & \text{if } b < \min(N_l^c, G_l^- + 1),\ a = 0 \\ \Pr(U_l = b) & \text{if } b = a + G_l^-,\ 0 < a < N_l^c - G_l^- \\ \Pr(N_l^c \le U_l \le G_l^-) & \text{if } b = N_l^c,\ a = 0 \\ \Pr(U_l = a + G_l^-) & \text{if } b = N_l^c,\ a \ge \max(N_l^c - G_l^-, 1) \\ 0 & \text{otherwise.} \end{cases} \tag{176}$$

$$\Pr(\eta_l^- = a, \eta_l^+ = b) = \begin{cases} \Pr(\eta_l = a) & \text{if } a < \min(N_l^t, G_l^+ + 1),\ b = 0 \\ \Pr(\eta_l = a) & \text{if } a = b + G_l^+,\ 0 < b < N_l^t - G_l^+ \\ \Pr(N_l^c \le \eta_l \le G_l^+) & \text{if } a = N_l^t,\ b = 0 \\ \Pr(\eta_l = b + G_l^+) & \text{if } a = N_l^t,\ b \ge \max(N_l^t - G_l^+, 1) \\ 0 & \text{otherwise.} \end{cases} \tag{177}$$

where $\Pr(U_l = x) = Bin(x, p_l^t, N_l^t)$, $\Pr(x \le U_l \le y) = \sum_{z=x}^y Bin(z, p_l^t, N_l^t)$, $\Pr(\eta_l = x) = Bin(x, p_l^c, N_l^c)$, *and* $\Pr(x \le \eta_l \le y) = \sum_{z=x}^y Bin(z, p_l^c, N_l^c)$.

Proof We will prove the result for $\Pr(U_l^- = a, U_l^+ = b)$ only, as the proof for $\Pr(\eta_l^- = a, \eta_l^+ = b)$ is symmetrical. Recall first that we are now in a case in which the number of treated and control units in each stratum, N_l^t and N_l^c respectively, are fixed. We use this together with Assumption 4 to write the marginal distributions of U_l^+ and U_l^-:

$$\Pr(U_l^+ = b) = \Pr(U_l - \max(U_l - N_l^c, 0) = b) = \begin{cases} \Pr(U_l = b) & \text{if } b < N_l^c \\ \Pr(U_l \ge N_l^c) & \text{if } b = N_l^c \end{cases}$$

$$\Pr(U_l^- = a) = \Pr(\max(U_l - G_l^-, 0) = a) = \begin{cases} \Pr(U_l \le G_l^-) & \text{if } a = 0 \\ \Pr(U_l = a + G_l^-) & \text{if } a > 0. \end{cases}$$

In the above, $G_l^- = \max(N_l^t - N_l^c, 0)$, a fixed quantity. The first equality in both formulas follows from the definition of U_l^+ and U_l^- given in Equations (48) and (49) respectively. The two cases of each definition follow directly from these definitions. Given the marginal distributions above, we can easily find distributions for U_l^+ and U_l^- conditional on $U_l = x$:

$$\mathbb{I}(U_l^+ = b | U_l = x) = \begin{cases} 1 & \text{if } b < N_l^c, x = b \\ 1 & \text{if } b = N_l^c, x \geq N_l^c \\ 0 & \text{otherwise.} \end{cases} \tag{178}$$

$$\mathbb{I}(U_l^- = a | U_l = x) = \begin{cases} 1 & \text{if } a = 0, x \leq G_l^- \\ 1 & \text{if } a > 0, x = a + G_l^- \\ 0 & \text{otherwise.} \end{cases} \tag{179}$$

This is because, if the number of treated and control units in each stratum is fixed, randomness in U_l^+ and U_l^- comes only from U_l, as is evident from the definitions of U_l^+ and U_l^- given in Eqns. (85) and (86). Therefore once that is fixed the two quantities become constant and, as such, independent of each other. Given this fact, it is evident that $\Pr(U_l^- = a, U_l^+ = b | U_l = x) = \mathbb{I}(U_l^- = a, U_l^+ = b | U_l = x) = \mathbb{I}(U_l^- = a | U_l = x)\mathbb{I}(U_l^+ = b | U_l = x)$. 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:

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

This is evident from Equations (178) and (179). 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: $b < N_l^c, x = b$ and $a = 0, x \leq G_l^-$.

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

Case 2: $b < N_l^c, x = b$ and $a > 0, x = a + G_l^-$.

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

Case 3: $b = N_l^c, x \geq N_l^c$ and $a = 0, x \leq G_l^-$.

Here we can combine the second and fourth conditions to obtain: $N_l^c \leq x \leq G_l^-$. This leads to the event: $b = N_l^c$, $N_l^c \leq x \leq G_l^-$, $a = 0$.

Case 4: $b = N_l^c, x \geq N_l^c$ and $a > 0, x = a + G_l^-$.

First from the second and fourth events we have: $x = a + G_l^- \geq N_l^c$, which can be used to obtain the event:

$a \geq N_l^c - G_l^-$. We also rewrite the third condition as $a \geq 1$, leading to the following final representation for the event of this case: $b = N_l^c$, $a \geq \max(N_l^c - G_l^-, 1)$, $x = a + G_l^-$.

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

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

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

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

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.

Using Lemma 2, these distributions are easy to enumerate and to construct a lookup table for. They are also clearly symmetrical and their form implies that $\Pr(U_l^- = a, U_l^+ = b, \eta_l^- = c, \eta_l^+ = d) = \Pr(U_l^- = a, U_l^+ = b) \Pr(\eta_l^- = c, \eta_l^+ = d)$. This fact is going to allow us to derive an expression for the distribution of (χ^+, χ^-), given in the following theorem.

THEOREM 8. *(Distribution of (χ^+, χ^-)) 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 $M_l = \min(N_l^t, N_l^c)$, $M = \sum_{l=1}^L M_l$. Let data $\mathbb{D} = \{(U_l, V_l, \eta_l, \nu_l)\}_{l=1}^L$ be drawn according to Equations* (172)-(175)*. Let χ^+ be the maximum of Formulation 2 on $\mathbb{D}$ and let χ^- be the minimum of Formulation 2 also on*

$\mathbb{D}$. *Let* $\mathcal{X}_M := \left\{ \frac{b-c-1}{\sqrt{b+c+1}} : b,c \in \{0,\dots,M\} \right\}$. *Additionally, let* $\mathcal{A}(y), \mathcal{B}(y,s), \mathcal{C}(x), \mathcal{D}(x,r), \mathcal{H}(x,y,r,s)$ *be defined as in Theorem 7. The pmf of* (χ^+,χ^-)*, for two values* $s,r \in \mathcal{X}_M$ *is given by:*

$$\Pr(\chi^- = s, \chi^+ = r|X) = \sum_{x=-M}^{M} \sum_{y=-M}^{M} \sum_{(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) \in \mathcal{H}(x,y,r,z)} \prod_{l=1}^{L} \begin{cases} g_1(a_l,b_l,c_l,d_l) & \text{if } x<1, y<1 \\ g_2(a_l,b_l,c_l,d_l) & \text{if } x \geq 1, y<1 \\ g_3(a_l,b_l,c_l,d_l) & \text{if } x \geq 1, y \geq 1 \\ 0 & \text{otherwise,} \end{cases} \tag{180}$$

where:

$$\begin{aligned} &g_1(a_l,b_l,c_l,d_l) = \\ &\mathbb{I}(|a_l| = b_l) \sum_{j=0}^{M_l} \Bigg[\Pr\left(U_l^- = a_l + j, U_l^+ = \frac{2M_l - d_l + c_l}{2}\right) \Pr\left(\eta_l^- = j, \eta_l^+ = \frac{2M_l - d_l - c_l}{2}\right) \\ &\qquad + \Pr\left(U_l^- = a_l + j, U_l^+ = \frac{d_l + c_l}{2}\right) \Pr\left(\eta_l^- = j, \eta_l^+ = \frac{d_l - c_l}{2}\right) \mathbb{I}(M_l \neq d_l) \Bigg] \end{aligned} \tag{181}$$

$$\begin{aligned} &g_2(a_l,b_l,c_l,d_l) = \\ &\mathbb{I}(|a_l| = b_l)\mathbb{I}(|c_l| = d_l) \sum_{j=0}^{M_l} \sum_{k=0}^{M_l} \Pr(U_l^- = a_l + j, U_l^+ = c_l + k) \Pr(\eta_l^- = j, \eta_l^+ = k) \end{aligned} \tag{182}$$

$$\begin{aligned} &g_3(a_l,b_l,c_l,d_l) = \\ &\mathbb{I}(|c_l| = d_l) \sum_{k=0}^{M_l} \Bigg[\Pr\left(U_l^- = \frac{2M_l - b_l + a_l}{2}, U_l^+ = c_l + k\right) \Pr\left(\eta_l^- = \frac{2M_l - b_l - a_l}{2}, \eta_l^+ = k\right) \\ &\qquad + \Pr\left(U_l^- = \frac{b_l + a_l}{2}, U_l^+ = c_l + k\right) \Pr\left(\eta_l^- = \frac{b_l - a_l}{2}, \eta_l^+ = k\right) \mathbb{I}(M_l \neq b_l) \Bigg]. \end{aligned} \tag{183}$$

The probabilities $\Pr(U_l^- = a, U_l^+ = b)$ *and* $\Pr(\eta_l^- = a, U_l^+ = b)$ *are given in Lemma 2 and depend on* $N_l^t, N_l^c, p_l^t, p_l^c$.

Proof The proof follows closely the template of the proof of Theorem 7. Note first that in this case N_l^t and N_l^c are fixed and known for all strata l. We first bound the values of U_l^+ and η_l^+:

$$\begin{aligned} 0 \leq U_l^+ &= M_l - V_l^+ && \text{(By definition of } V_l^+\text{)} \\ &\leq M_l, && \text{(Both } M_l \text{ and } V_l^+ \text{ are nonnegative integers)} \end{aligned}$$

and:

$$\begin{aligned} 0 \leq \eta_l^+ &= M_l - \nu_l^+ && \text{(By definition of } \nu_l^+\text{)} \\ &\leq M_l. && \text{(Both } M_l \text{ and } \nu_l^+ \text{ are nonnegative integers)} \end{aligned}$$

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 7, 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 (χ^+,χ^-):

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

Given the set above, we will consider values of (χ^+, χ^-) 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 7 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 (χ^+, χ^-) under Assumption 4.

The steps leading to the form of (180) in the main statement of the theorem are exactly the same as those in the proof of Theorem 7, 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 7, 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 tom 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 7, including the fact that $f_3 = 0$. The argument made for this in the proof of Theorem 7 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 (181)-(183), and can be derived as follows:

$$
\begin{aligned}
&\Pr(TE_l^- = a_l, S_l^- = b_l, TE_l^+ = c_l, R_l^+ = d_l) \\
&= \Pr(U_l^- - \eta_l^- = a_l, |U_l^- - \eta_l^-| = b_l, U_l^+ - \eta_l^+ = c_l, M_l - |U_l^+ + \eta_l^+ - M_l| = d_l) && (184)\\
&= \Pr(U_l^- - \eta_l^- = a_l, |a_l| = b_l, U_l^+ - \eta_l^+ = c_l, |U_l^+ + \eta_l^+ - M_l| = M_l - d_l) && (185)\\
&= \mathbb{I}(|a_l| = b_l) \Pr(U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l, |U_l^+ + \eta_l^+ - M_l| = M_l - d_l) && (186)\\
&= \mathbb{I}(|a_l| = b_l) \Big[\Pr(U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l, U_l^+ + \eta_l^+ - M_l = M_l - d_l) \\
&\qquad + \Pr(U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l, U_l^+ + \eta_l^+ - M_l = d_l - M_l) \mathbb{I}(M_l \neq d_l) \Big] && (187)\\
&= \mathbb{I}(|a_l| = b_l) \Big[\Pr\left(U_l^- = a_l + \eta_l^-, U_l^+ = c_l + \eta_l^+, \eta_l^+ = \frac{2M_l - d_l - c_l}{2}\right) \\
&\qquad + \Pr\left(U_l^- = a_l + \eta_l^-, U_l^+ = c_l + \eta_l^+, \eta_l^+ = \frac{d_l - c_l}{2}\right) \mathbb{I}(M_l \neq d_l) \Big] && (188)\\
&= \mathbb{I}(|a_l| = b_l) \sum_{j=0}^{M_l} \Pr\left(U_l^- = a_l + j, U_l^+ = \frac{d_l + c_l}{2}\right) \Pr\left(\eta_l^- = j, \eta_l^+ = \frac{d_l - c_l}{2}\right) \mathbb{I}(M_l \neq d_l) \\
&\qquad + \Pr\left(U_l^- = a_l + j, U_l^+ = \frac{2M_l - d_l + c_l}{2}\right) \Pr\left(\eta_l^- = j, \eta_l^+ = \frac{2M_l - d_l - c_l}{2}\right) && (189)\\
&= g_1(a_l, b_l, c_l, d_l). && (190)
\end{aligned}
$$

In the above, Equation (184) follows from the definitions of $TE_l^-, S_l^-, TE_l^+, R_l^+$ given in Claim 4. Equation (185) follows from rearranging the terms in the last equality in the parentheses, and from plugging in the equality for $U_l^- - \eta_l^-$, (186) follows from the fact that a_l and b_l are both constants and so the event $|a_l| = b_l$ is independent from all the others inside the parentheses. Equation 187 follows from the fact that the event $\{|U_l^+ + \eta_l^+ - M_l| = M_l - d_l\} = \{U_l^+ + \eta_l^+ - M_l = M_l - d_l\} \cup \{U_l^+ + \eta_l^+ - M_l = d_l - M_l, M_l \neq d_l\}$, 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_l and d_l are fixed quantities, so they are independent of all the other events in the statement. Equation (188) follows from rearranging the terms in the equalities for η_l^+. Equation (189) follows from summing over all values of η_l^- and from plugging in equalities for η_l^+ into the equation for U_l^+. Finally, (190) follows from the definition of $g_1(a_l, b_l, c_l, d_l)$ 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:

$$
\begin{aligned}
&\Pr(TE^- = a_l, S_l^- = b_l, TE^+ = c_l, S_l^+ = d_l) \\
&= \Pr(U_l^- - \eta_l^- = a_l, |U_l^- - \eta_l^-| = b_l, U_l^+ - \eta_l^+ = c_l, |U_l^+ - \eta_l^+| = d_l) \\
&= \Pr(U_l^- - \eta_l^- = a_l, |a_l| = b_l, U_l^+ - \eta_l^+ = c_l, |c_l| = d_l) \\
&= \mathbb{I}(|a_l| = b_l)\mathbb{I}(|c_l| = d_l)\Pr(U_l^- - \eta_l^- = a_l, U_l^+ - \eta_l^+ = c_l) \\
&= \mathbb{I}(|a_l| = b_l)\mathbb{I}(|c_l| = d_l)\Pr(U_l^- = a_l + \eta_l^-, U_l^+ = c_l + \eta_l^+) \\
&= \mathbb{I}(|a_l| = b_l)\mathbb{I}(|c_l| = d_l)\sum_{j=0}^{M_l}\sum_{k=0}^{M_l}\Pr(U_l^- = a_l + k, U_l^+ = c_l + j)\Pr(\eta_l^- = k, \eta_l^+ = j) \\
&= g_2(a_l, b_l, c_l, d_l).
\end{aligned}
$$

Finally, g_3 can be derived with the same steps as g_1:

$$
\begin{aligned}
&\Pr(TE^- = a_l, R_l^- = b_l, TE_l^+ = c_l, S_l^+ = d_l) \\
&= \Pr(U_l^- - \eta_l^- = a_l, M_l - |U_l^- + \eta_l^- - M_l| = b_l, U_l^+ - \eta_l^+ = c_l, |U_l^+ - \eta_l^+| = d_l) \\
&= \Pr(U_l^- - \eta_l^- = a_l, |U_l^- + \eta_l^- - M_l| = M_l - b_l, U_l^+ - \eta_l^+ = c_l, |c_l| = d_l) \\
&= \mathbb{I}(|c_l| = d_l)\Pr(U_l^- - \eta_l^- = a_l, |U_l^- + \eta_l^- - M_l| = M_l - b_l, U_l^+ - \eta_l^+ = c_l) \\
&= \mathbb{I}(|c_l| = d_l)\Big[\Pr(U_l^- = a_l + \eta_l^-, U_l^- + \eta_l^- - M_l = M_l - b_l, U_l^+ = c_l + \eta_l^+) \\
&\qquad + \Pr(U_l^- = a_l + \eta_l^-, U_l^- + \eta_l^- - M_l = b_l - M_l, U_l^+ = c_l + \eta_l^+)\mathbb{I}(M_l \neq b_l)\Big] \\
&= \mathbb{I}(|c_l| = d_l)\Big[\Pr(U_l^- = a_l + \eta_l^-, \eta_l^- = \frac{2M_l - a_l - b_l}{2}, U_l^+ = c_l + \eta_l^+) \\
&\qquad + \Pr(U_l^- = a_l + \eta_l^-, \eta_l^- = \frac{b_l - a_l}{2}, U_l^+ = c_l + \eta_l^+)\mathbb{I}(M_l \neq b_l)\Big] \\
&= \mathbb{I}(|c_l| = d_l)\sum_{k=0}^{M_l}\Pr\left(U_l^- = \frac{b_l + a_l}{2}, U_l^+ = c_l + k\right)\Pr\left(\eta_l^- = \frac{b_l - a_l}{2}, \eta_l^+ = k\right) \\
&\qquad + \Pr\left(U_l^- = \frac{2M_l - b_l + a_l}{2}, U_l^+ = c_l + k\right)\Pr\left(\eta_l^- = \frac{2M_l - b_l - a_l}{2}, \eta_l^+ = k\right)\mathbb{I}(M_l \neq b_l) \\
&= g_3(a_l, b_l, c_l, d_l).
\end{aligned}
$$

This concludes the proof of Theorem 8.

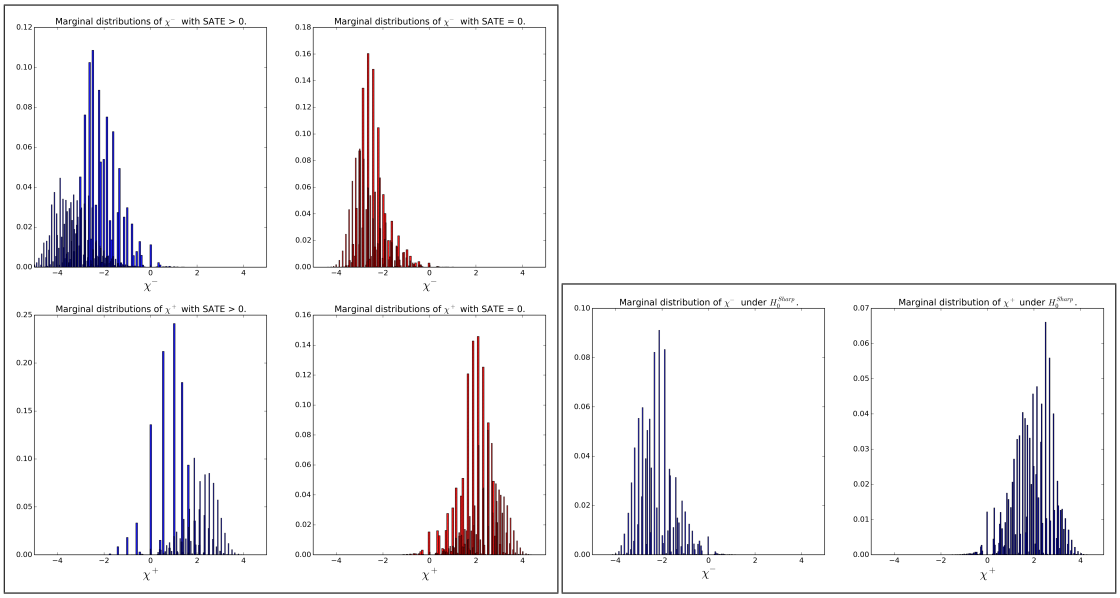
Appendix G: Proof of Theorem 5

Recall that, by Theorems 8 and 7, the distributions of matched pairs in each stratum are independent of each other because of SUTVA, and because optimization of χ is performed almost independently in each stratum. Since the calculation of the full joint distribution of McNemar's statistic 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 $\mathcal{H}(x, y, r, z)$ to $O(4mN^{4m} \log N)$ (see, e.g., Dudgeon and Mersereau 1983). Here N is the total number of units, m is the largest number of matches made in any of the strata. This last quantity is repeated four times, as there are four variables for which probabilities must be computed. These are the densities being convolved together in the summation. This, coupled with simple enumeration to create lookup tables for distributions of g_1, g_2, g_3 in each stratum (See Theorems 7 and 8 for the definition of these quantities), as well as the final range distribution yields an algorithm that creates a probability table for the null distribution of (χ^+, χ^-) in $O(4mN^{4m} \log N)$. This implies that the worst-case time of computing the distribution in Theorem 7 is polynomial.

Appendix H: Plots of exact distributions of (χ^+, χ^-) on a simulated dataset

Figure 3 shows marginal distributions of χ^+ and χ^- derived from the joint distributions in Theorems 7 and 8, respectively. We simulated datasets from the DGPs described in those sections and estimated the distributions with the formulas in Theorems 3 and 4, and the polynomial time algorithm suggested in Theorem 5. Panel (a) shows a location and scale difference in the distribution of (χ^+, χ^-) between when the null is true and when it is not. This demonstrates our tests' capacity to detect full-sample treatment effects.



(a) (b)

Figure 3 Marginal distributions of (χ^+, χ^-) on a simulated dataset with 150 units divided in 15 strata, each with more control than treated units. (a) General marginal distributions with random potential outcomes. The left column shows the distributions when a SATE is present in the whole sample, and the right column when the SATE is 0. (b) Null marginal distributions under $\mathbb{H}_0^{sharp}$

Appendix I: Additional Simulations

I.1. Performance of robust tests with increasing number of covariates

We present results of simulation studies under the same settings as those in Section 6, but we instead keep the number of units fixed at 200, and let the number of covariates grow.

Results for McNemar's test are reported in Figure 4, and results for the z-test are reported in Figure 5. We see that, largely, our tests tend to perform well even in the increasing P regime: both tests achieve the overall lowest error rate when the ATT is 0 (closest to the error rate of the idealized test that includes both true potential outcomes), and display good statistical power as the ATT increases in relative strength. The z-test is somewhat conservative when the true ATT is weak (Cohen's d=0.2), and when there are many covariates: this is a desirable pattern since the larger number of covariates makes the matching problem harder. Notably, the benchmark test performed on the true potential outcomes displays similar behavior to our robust z-test in Figure 5 when $d = 0.2$.

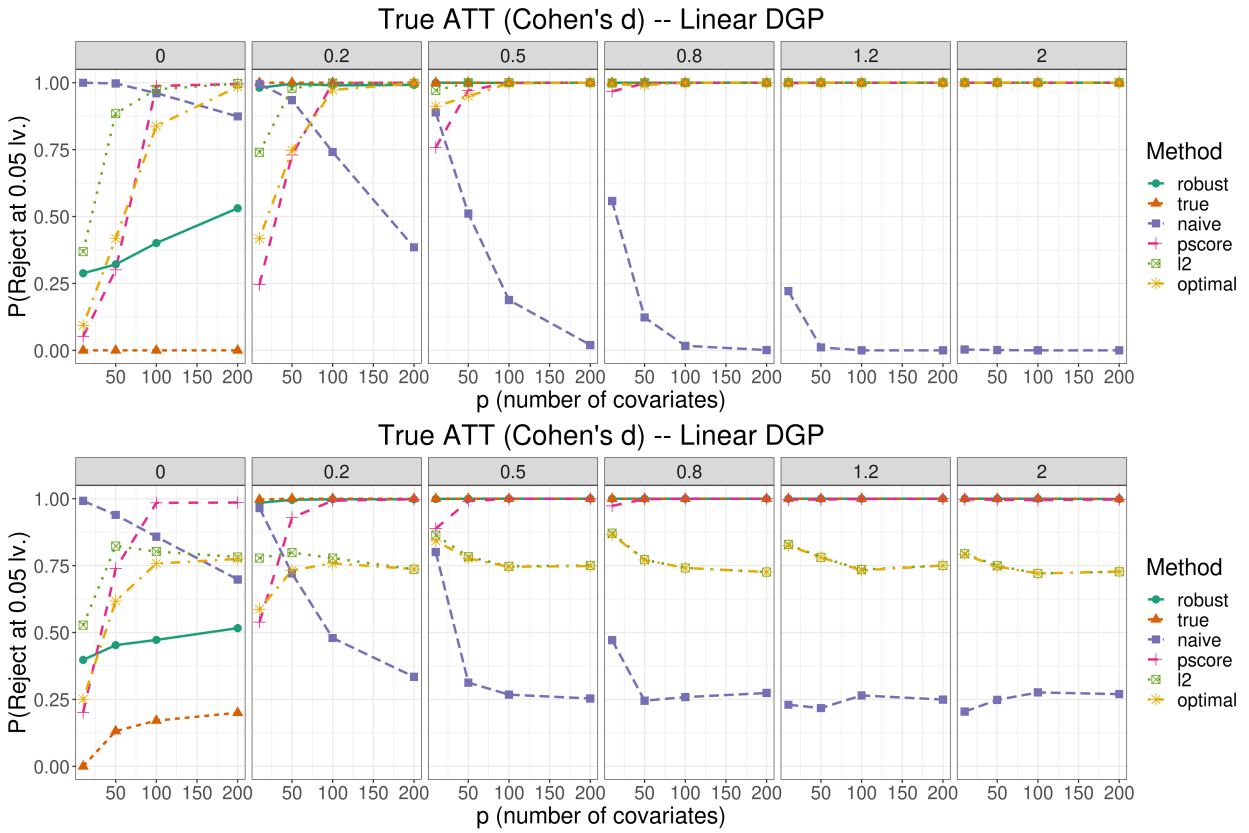


Figure 4 Performance of the robust McNemar's test on a sample size of $N = 200$ units with an increasing number of covariates. Top row: simple DGP, bottom row: complex DGP. The ideal method is labelled "true" and represented by the dark orange dashed line in the figure.

I.2. Performance of robust tests with Exclusively Binning Constraints

We present performance results of our robust tests when constraints are exclusively binning, i.e., when data is perfectly stratified. In this setting we partition the N units into strata such that each stratum contains

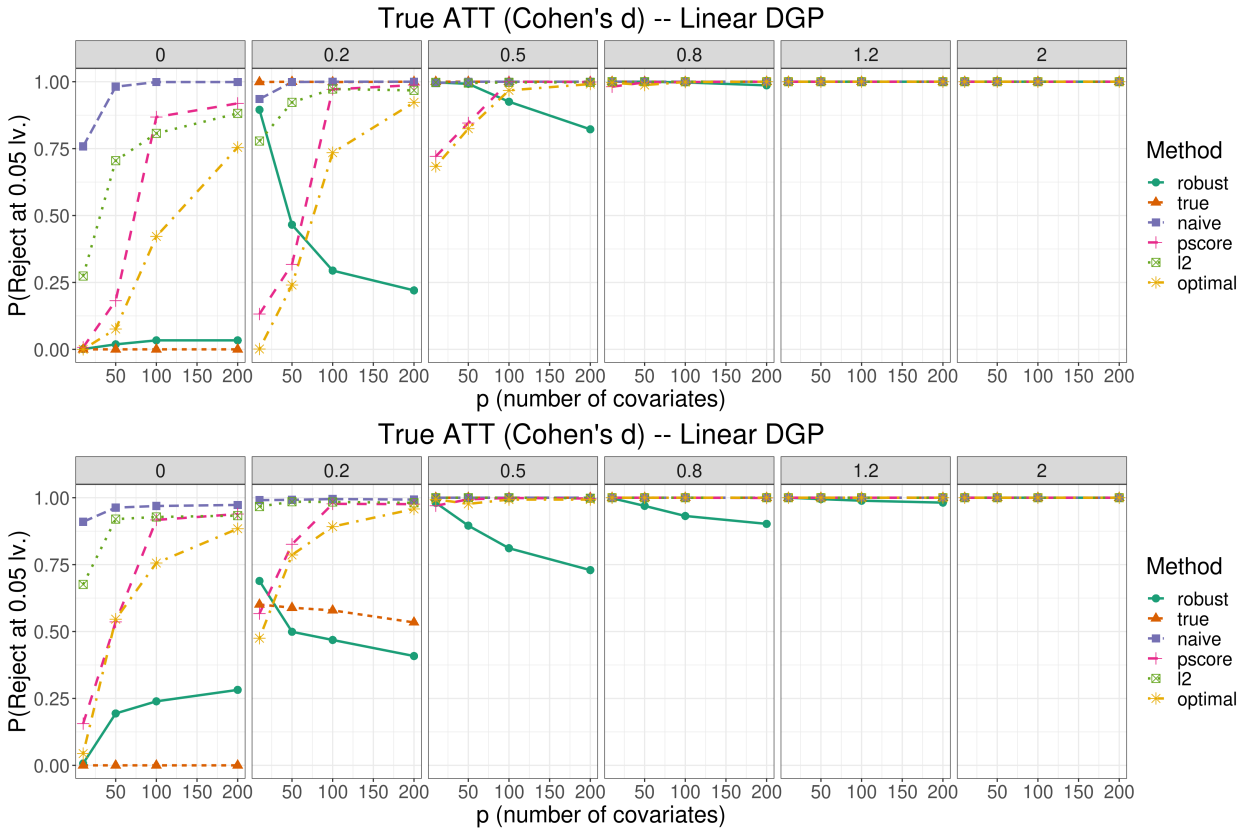


Figure 5 **Performance of the robust z-test on a sample size of $N = 200$ units with an increasing number of covariates. Top row: simple DGP, bottom row: complex DGP.**

exactly 25 units, for example, when $N = 100$ there will be 4 strata. We denote the number of strata for a given N with L_N, and use the variable $s_i = 1, \ldots, L_N$ to denote which stratum unit i is assigned to. For a set of fixed coefficients $\beta_1, \ldots, \beta_{L_N}$ that are calibrated as described prior, we generate:

$$e_i = \frac{1}{1 + \exp(-\beta_{s_i})}, \quad \epsilon_i \sim \text{Normal}(0, 1), \quad y_i^* = t_i \tau + \beta_{s_i} + \epsilon_i,$$

with treatments t_i assigned to exactly N^t units but each with probability e_i. For binary data, we set $y_i = \mathbb{I}[y_i^* > 0]$, and for continuous data $y_i = y_i^*$, as done in all the other simulations. All other settings are as in the simulations presented in Section 6 of the main paper.

Results for the robust McNemar's test are presented in Figure 6.

Results for the robust z-test are presented in Figure 7

We see that both test perform well under these types of constraints: error rate for the z-test is almost 0 when the ATT is 0, and the test achieves good statistical power as the ATT grows in strength. A similar patterns is visible for McNemar's test, though this test requires a slightly larger sample size ($N \geq 100$) in order to achieve a good false rejection rate.

I.3. Performance of robust z-test with different types of constraints

We study whether the type of match quality constraints used in our robust test affects both the error rate of the robust tests as well as their runtime. All simulation settings are kept same as before, except that our

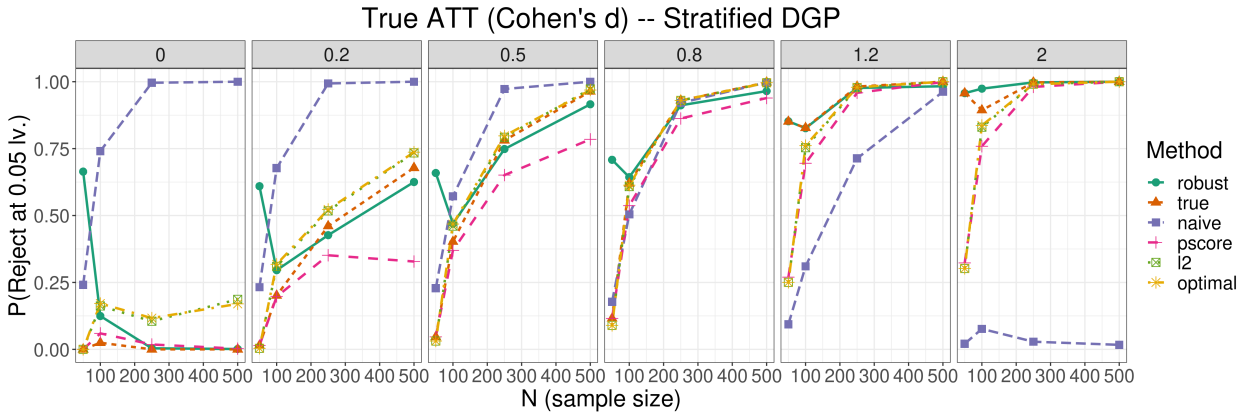


Figure 6 **Performance of robust McNemar's test under exclusively binning constraints.**

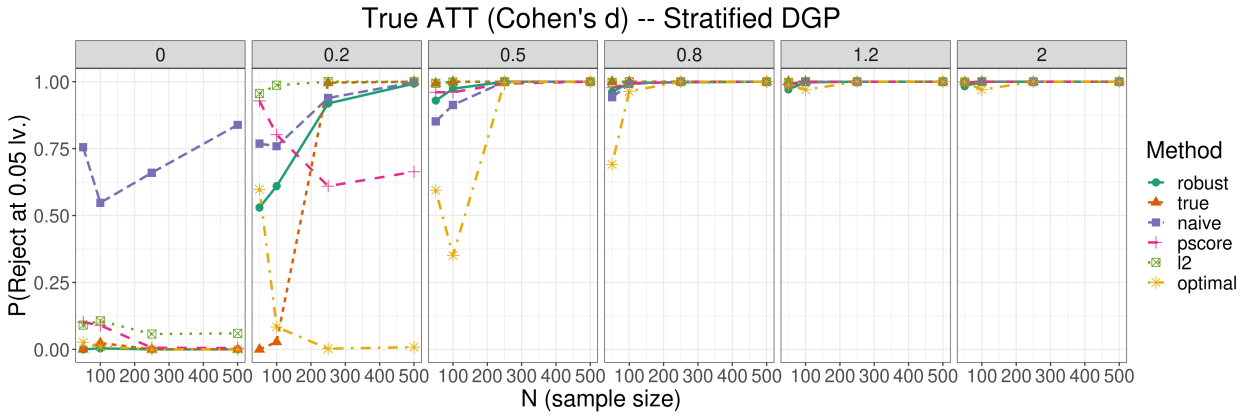


Figure 7 **Performance of robust Z test under exclusively binning constraints.**

robust test is ran 3 times per iteration, each time with a different type of constraint. The types of constraints we compare are as follows:

- **Moment**: We require balance between the first two moments of the empirical distributions of each covariate in the treated/control sets after matching. This is formulated as: $|\frac{1}{M}\sum_{i=1}^{N^t}\sum_{j=1}^{N^c} a_{ij}(x_{ip}^t - x_{jp}^c)| \leq \epsilon_p^{mean} \times (\sigma(x_p^t)/2 + \sigma(x_p^c)/2)$, for $p = 1, \ldots, P$, and $|\frac{1}{M}\sum_{i=1}^{N^t}\sum_{j=1}^{N^c} a_{ij}(x_{ip}^t - x_{jp}^c)^2| \leq \epsilon_p^{var} \times (\sigma^2(x_p^t)/2 + \sigma^2(x_p^c)/2)$, for $p = 1, \ldots, P$ ($2P$ constraints total).
- **Caliper**: We only allow units to be matched if their propensity scores are close enough. This is formulated as: $a_{ij}|e_i^t - e_j^c| \leq \epsilon^{cal}$ for $i = 1, \ldots N^t$, $j = 1, \ldots, N^c$ ($N^t \times N^c$ constraints total).
- **Quantile**: We require the empirical distributions of each covariate in the treated/control sets after matching to be similar in terms of their quantiles. For each covariate $p = 1, \ldots, P$, and a pre-specified sequence of quantiles $\gamma_{p1}, \ldots, \gamma_{pL}$, such that $\gamma_{p1} \leq \gamma_{p2} \leq \ldots \gamma_{pL}$, we require: $|\frac{1}{M}\sum_{i=1}^{N^t} a_{ij}\mathbb{I}[x_{ip}^t \leq \gamma_p] - \sum_{j=1}^{N^c}\sum_{i=1}^{N^t} a_{ij}\mathbb{I}[x_{jp}^c \leq \gamma_p]| \leq \epsilon_p^{qua}$ ($P \times L$ constraints total).

as previously done, tolerance values (ϵ_p^{mom}, ϵ^{cal}, ϵ_p^{qua}) are chosen as the smallest feasible values on an evenly spaced grid between 0.1 and 1, and $\sigma_p^2(x_p^t)$, $\sigma_p(x_p^t)$, $\sigma_p^2(x_p^c)$, $\sigma_p(x_p^c)$ are the variance and standard deviation of covariate x_p in the treated and control set respectively.

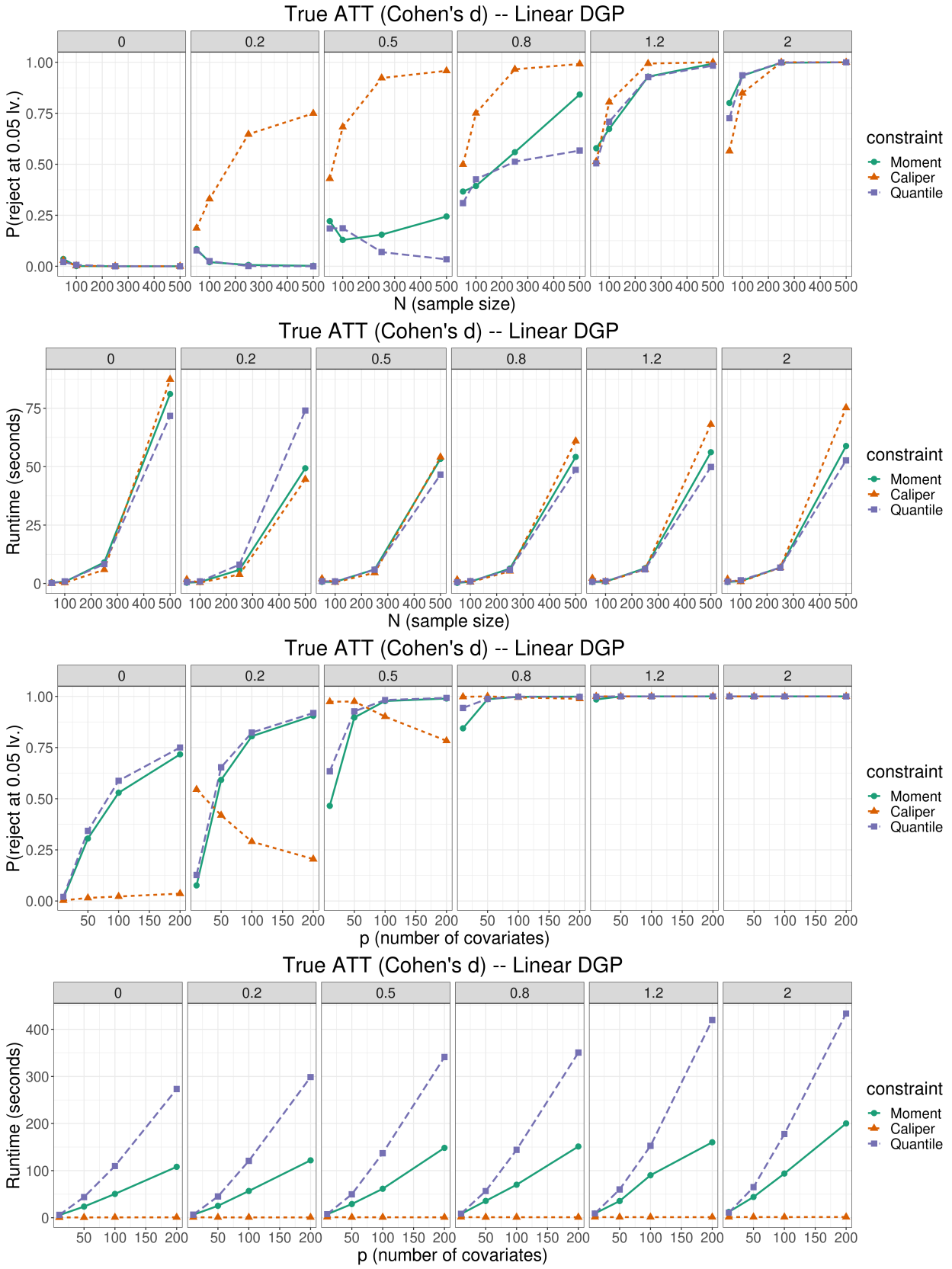


Figure 8 **Comparison of rejection rate and runtime of the robust z-test with different types of constraints.**

Results are reported in Figure 8. In both fixed P and fixed N regimes we see that the pscore caliper constraint seems to display the lowest rejection rate when the true ATT is 0, indicating that this type of constraint is particularly suitable for the settings we studied. Interestingly, the caliper constraint seems to also be the one able to achieve better statistical power as the true ATT increases in size in the fixed P regime, while it displays worse statistical power than the other two constraints in the fixed-N regime. In terms of runtime, there appears to not be virtually any difference between constraint types in the fixed-P regime, however the caliper constraints is much faster and unaffected by the number of covariates in the fixed-N regime, as the number of covariate grows, which is expected due to the propensity score being a 1-dimensional quantity for any amount of covariates. Also expectedly, the quantile constraint requires longer execution time as the number of covariates grows compared to the moment constraint.

The main takeaway from this excercise is that it is a good idea for analyst to include several constraints at once in their application of the robust tests, requiring several different metrics of balance to all be below a certain tolerance to admit a match assignment as good.

I.4. Performance of robust z-test with different MIP approximation methods

Since MIP solution algorithms have a runtime that is exponential in the number of decision variables (units to match, in our case), we have implemented and tested two different approximation methods to speed up computation for our robust tests. Both methods are based on first solving a linear relaxation of the original MIP, and then solving an integer program that has a solution closest to that of the relaxation. The two methods are described as follows:

Approximation 1 : This is the same approximation method used in Zubizarreta (2012). We first solve a linear relaxation of all the MIPs used for the robust z-test, (Formulation 2) and once an optimal solution is found via our proposed algorithm, we approximate a MIP solution by solving the following IP. Let $b_{11}, \ldots, b_{N^t N^c}$ be the continuous matching indicators output by solving the linear relaxation of Formulation 2. The approximation can be obtained by solving:

$$\begin{aligned}
&\max / \min \sum_{i=1}^{N^t} \sum_{j=1}^{N^c} a_{ij} b_{ij} \\
&\text{Subject to:} \\
&\quad \sum_{j=1}^{N^c} a_{ij} \leq 1, \quad i = 1, \ldots, N^t \\
&\quad \sum_{i=1}^{N^t} a_{ij} \leq 1, \quad j = 1, \ldots, N^c.
\end{aligned}$$

The approximation tries match as many units entirely as the linear relaxation matches fractionally, while not matching any unit more than once. This approximation provides a substantial speedup over the MIP, since it removes all the additional constraints, however the solution could be sub-optimal and some constraints may be violated in practice. Results from comparisons on simulated data in Figure 9 show that, in practice, this approximation is almost as reliable as solving the exact MIP, making it a viable option in practice.

Approximation 2 : Approximation 2 solves the same problem as Approximation 1, except that all the additional constraints (including user-defined balance constraints) in Formulation 2 are kept. This ensures that the solution found satisfies the constraints specified, but is more expensive in terms of computation time. As Figure 9 shows, this approximation performs even more closely to the MIP: gains in computation time are offset by almost no loss in performance.

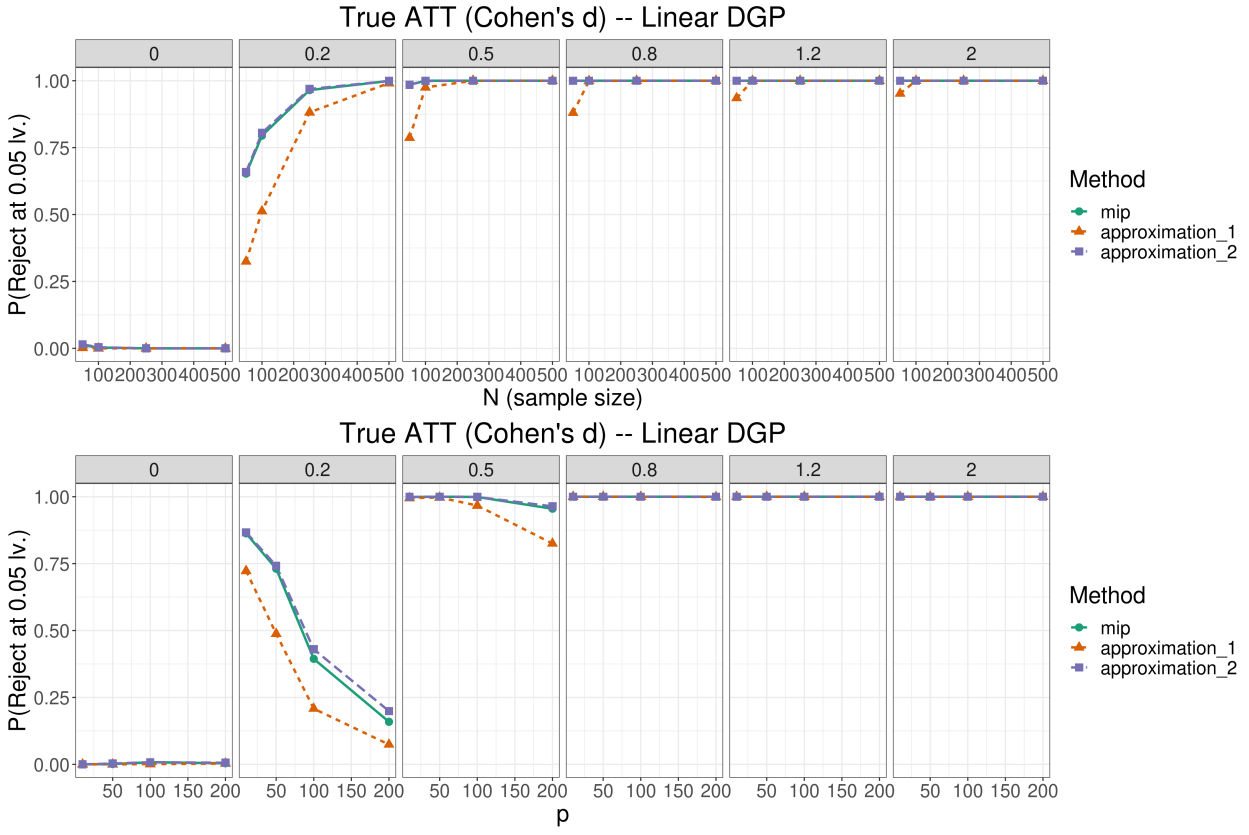


Figure 9 Performance of robust z-test with different MIP approximation methods.

I.5. Discussion of computational issues on large datasets

The problem we are solving is computationally hard. It can often be solved for about 1000 data points. If the number of data points is larger than this, we point out:

1. If the number of data points is large, the confidence bounds are likely to be narrow enough that an analysis like ours would not be necessary – the robust intervals would be small, given that any good match assignment would include a large amount of units. Our method would be overkill for such problems.

2. If the number of data points is too large, it is possible to break the domain into subsets of nearby points, with each subset having size ∼1000, and handle each subset separately.

3. Another option to deal with large datasets is to randomly split them into smaller datasets of size around 1000, run our robust test on each split, and finally average p-values to obtain a cumulative p-value.

4. We have also tested and implemented two different MIP approximation methods based on linear relaxations, which permit us to substantially lower the runtime of our algorithms. Both methods are outlined in Appendix I.4, and simulation results show that they output p-values similar to those output by the MIP.

We have conducted several simulations to assess the behavior of our proposed methods when there are many covariates. Sample results are available in Figure 5, and additional results for many covariates are available in Figures 8 and 9.

I.6. Comparison To Matching Bounds (Morucci and Rudin 2018)

We perform a comparison of our robust tests to Matching Bounds (Morucci and Rudin 2018), a method for a different facet of the issue of multiple match assignment disagreeing. There are several substantial differences between the Matching Bounds approach of Morucci and Rudin (2018) and the robust testing approach proposed in this work. Namely that Matching Bounds is a tool for finding ranges of **treatment effects**, our robust tests are a tool for finding ranges of **test statistics** (i.e., we consider not just treatment effects but the uncertainty in their estimates). An analyst interested in knowing what possible treatment effects could be obtained by changing matched samples should employ Matching Bounds, while an analyst interested in obtaining the range of test statistics (and p-values) that could be obtained by matching in different ways should use the robust tests. Specifically (Morucci and Rudin 2018) finds bounds on the raw value of the treatment effect of interest, our robust testing approach targets a test statistic for the treatment effect: this latter quantity depends both on TE value and on TE variance. This latter dependence implies that the optimum Z-value for a robust test will likely not be the same Z-value obtained by, for example, performing a Z-test via normal approximation on the set of units that maximizes/minimizes the TE as selected by solving the optimization problem in (Morucci and Rudin 2018). Specifically, a Z-test conducted on matches selected as in (Morucci and Rudin 2018) will find a sub-optimal solution to the Z-statistic optimization problem, potentially leading to increased chances of type-I error (incorrect rejection of the null.).

To further demonstrate the differences between the two methods and their purposes, we have performed a set of simulations whose settings are similar to those in Section 6 of our paper and results are reported in Figure 10. One can see the benefit of finding the range of test statistics, as compared to just the range of treatment effects: P-values output by MBs are not as conservative as those from the robust test when the ATT is 0, and the bounds on the ATT output by MBs are better than those output by the robust test when the ATT is 0.

Appendix J: Agreement of traditional matching methods in case studies 1 and 2

As a supplement to our case studies, we present p-values obtained after applying traditional matching algorithms to the datasets of Case Studies 1 and 2. We do so to gauge whether there is disagreement in rejection decisions between matching methods on these datasets.

We have constrained all the matching methods used to respect the caliper constraint we employ in constructing our robust test, e.g., units whose absolute distance in the covariates exceeds the thresholds defined in Section 7 are disallowed from being matched. This ensures that all three methods compared produce at the very least balance on the covariates up to that threshold. In this sense, all these methods produce similar balance.

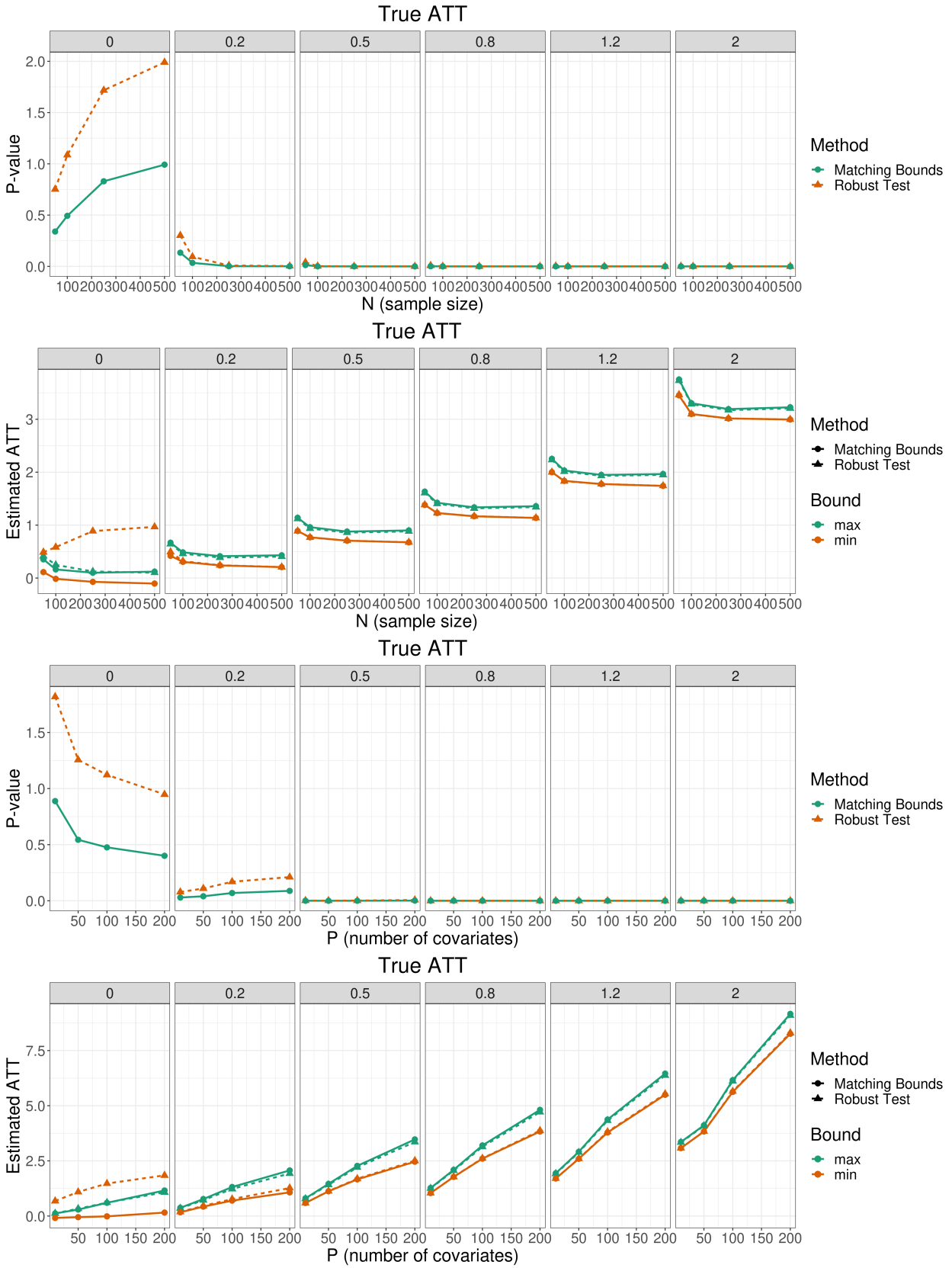


Figure 10 Comparison between Matching Bounds (Morucci and Rudin 2018) and the Robust Z-test. Top Row: The figure shows that if the true ATT is 0, our robust test have a high p-value (which is correct), and if the ATT is 0.2 or more, our robust test starts out conservatively and goes to 0 as N or P increases. Bottom Row: Our robust tests give uncertainty bounds on the ATT that are narrower than Matching Bounds because they optimize test statistics that depend on both the estimated mean and variance of the ATT.

Results of this comparison are presented in Table 2. We see that, generally methods tend to fail to reject on the GLOW dataset: this is because all methods fail to maximize the effective sample size of McNemar's test. They match units with the same outcome, and therefore all produce low-powered tests. Instead, we constrain our tests to output only matches that have a good effective sample size (almost all the treated units) – in this case we do indeed demonstrate that an effect is present.

For the Bike Sharing case study, we see that propensity score matching and L2 distance matching fail to reject the null hypothesis, while optimal constrained matching does in fact reject it. The disagreement here is likely due to the fact that several matched sets that meet the balance requirements are present, as we argue in our paper. Our application of the robust z-test to this dataset further corroborates this hypothesis by showing that, indeed, it cannot reject the null hypothesis over all possible good matches.

Table 2 P-values for McNemar and Z-test conducted on datasets for Case Studies 1 and 2 after matching with different traditional methods

	No Matching	Pscore	L2	Optimal
GLOW	0.00	1.00	1.00	0.42
Bike Share	0.00	0.20	0.32	0.00

Appendix K: Additional Information about Case Study 2

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: 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, 6 and 6, respectively for treated unit i and control unit j, 0 otherwise. Since the outcome variable is continuous, we focus on testing $\mathbb{H}_0^{SATT}$ by producing ranges of p-values for the Z-test. 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. 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.

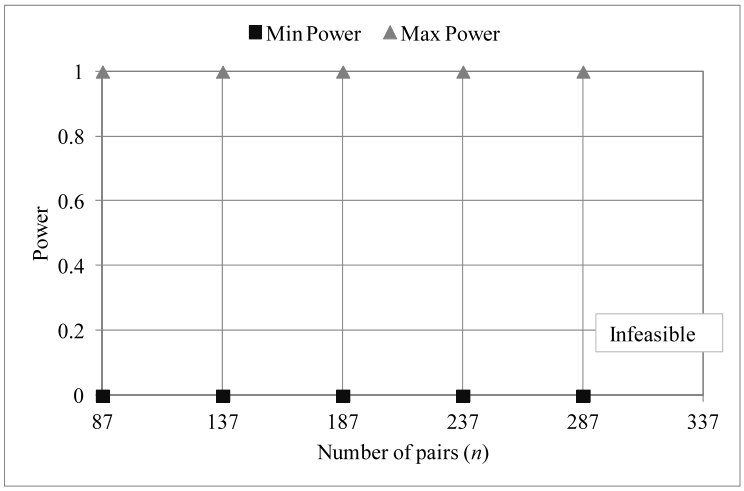


Figure 11 Variation of z-test optimum P-values for different n. (Case Study 3: Training Program)

Appendix L: Additonal Case Studies

L.1. Case Study 3: 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 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), nodegree (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: dist_{ij} =1 if covariates Black, Hispanic, married and nondegree were the same, and the differences in age, education and 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 11, 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 ~ 0 and declare a statistically significant difference while another experimenter can find a P-value of ~ 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 robustly detects the treatment effect before the solutions become infeasible.

L.2. Case Study 4: 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

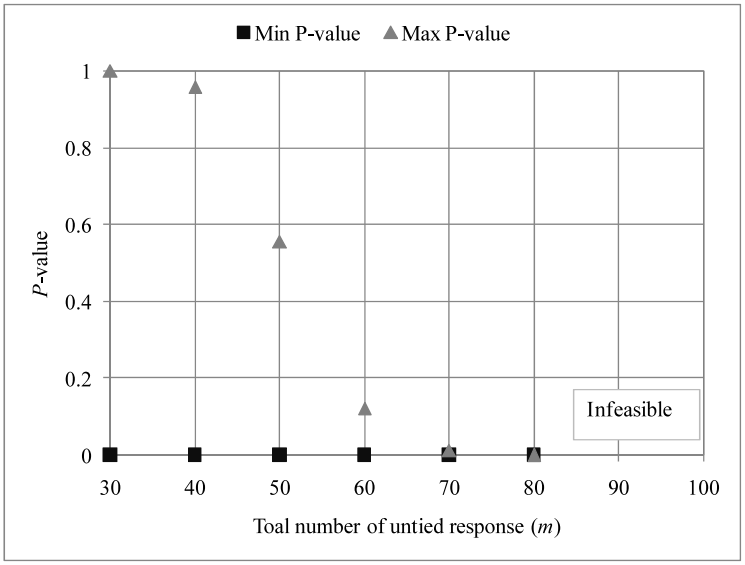


Figure 12 Variation of McNemar's test optimum P-value for different m. (Case Study 4: Crime and Transition)

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 7 is constructed in an analogous way to Figure 12 (using McNemar's test rather 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.

Appendix M: Robust Wilcoxon Signed Rank Test

In this section, we present an extension of our robust framework to the Wilcoxon Signed Rank Test. This extension is intended to show that our framework is general enough to be extendable to other test statistics, as well as provide one more robust test for practitioners.

The Wilcoxon Signed Rank Test is a more powerful test on the median than the sign test, but with a stronger assumption: it assumes that the population distribution is symmetric. We define the following parameters and decision variables to formulate the model:

- T_i is the the outcome of a treated observation i in the treatment group
- C_j is the the outcome of a control observation j in the control group
- s_{ij} is the ijth element of a matrix. It takes value 1 if the covariates of treated observation i and control observation j are similar enough to be a possible matched pair, otherwise 0

- δ_{ij} is a binary precomputed parameter, which is equal to 1 if T_i is greater than C_j, 0 otherwise
- g_{ijkl} is a binary precomputed parameter, which is equal to 1 if η_{ij} is greater than η_{kl}, where $\eta_{ij} = |T_i - C_j|$, 0 otherwise

There are three matrices of decision variables, namely:

- a_{ij} is a binary variable that is 1 if i and j are in the same pair, otherwise 0
- z_{ijkl} is a binary variable that is 1 if g_{ijkl}, a_{ij} and a_{kl} all are equal to 1, 0 otherwise. Intuitively z_{ijkl} is 1 only when ij and kl are both being used as pairs and when ij's absolute difference is larger than kl's absolute difference.
- h_{ij} is an integer variable whose value is the rank of pair ij. It is the count of kl pairs ranked beneath pair ij (plus one, so that the lowest rank is 1 rather than 0).

To compute the test statistic in the traditional way, one would compute absolute differences η_{ij} and rank order them. The test statistic is the sum of ranks of the positive differences.

The formulation is as follows :

$$\text{Maximize/Minimize}_{\mathbf{a},\mathbf{z},\mathbf{h}} \quad w_+(\mathbf{a},\mathbf{z},\mathbf{h}) = \sum_{i\in Q}\sum_{j\in R} h_{ij}\delta_{ij}$$

subject to:

$$g_{ijkl} + a_{ij} + a_{kl} - 2 \le z_{ijkl} \quad \forall i,j,k,l \qquad (z_{ijkl} \text{ is 1 only if } a_{ij},\ a_{kl} \text{ and } g_{ijkl} \text{ are 1}) \tag{191}$$

$$z_{ijkl} \le a_{ij} \quad \forall i,j,k,l \qquad (z_{ijkl} \text{ is 1 only if } a_{ij} \text{ is 1}) \tag{192}$$

$$z_{ijkl} \le a_{kl} \quad \forall i,j,k,l \qquad (z_{ijkl} \text{ is 1 only if } a_{kl} \text{ is 1}) \tag{193}$$

$$z_{ijkl} \le g_{ijkl} \quad \forall i,j,k,l \qquad (z_{ijkl} \text{ is 1 only if } g_{ijkl} \text{ is 1}) \tag{194}$$

$$a_{ij} \le s_{ij} \quad \forall i,j \qquad (\text{Choose only pairs that are allowed}) \tag{195}$$

$$\textstyle\sum_{i\in Q}\sum_{j\in R} a_{ij} = n \qquad (\text{Choose } n \text{ pairs}) \tag{196}$$

$$\textstyle\sum_{i\in Q} a_{ij} \le 1 \quad \forall j \qquad (\text{Choose at most one treatment }) \tag{197}$$

$$\textstyle\sum_{j\in R} a_{ij} \le 1 \quad \forall i \qquad (\text{Choose at most one control }) \tag{198}$$

$$a_{ij} \in \{0,1\} \quad \forall i,j \qquad (\text{Defines binary variable } a_{ij}) \tag{199}$$

$$h_{ij} = \textstyle\sum_{k\in Q}\sum_{l\in R} z_{ijkl} + a_{ij} \quad \forall i,j \qquad (\text{Defines } h_{ij}). \tag{200}$$

(optional covariate balance constraints).

Equations (191) to (194) are used to ensure that z_{ijkl}=1 when g_{ijkl}, a_{ij} and a_{kl} all are one. Equation (195) is used to maintain covariates constraint such that only when the precomputed parameter s_{ij} is 1 then a_{ij} is allowed to take a value of 1. Constraints (196)-(198) are the same constraints as in the sign test, to make sure we have n pairs with one treatment and control observation in each pair. Equation (199) defines binary variables a_{ij}. Equation (200) is used to calculate rank h_{ij} for each pair ij. If the pair ij is not being used then a_{ij} will be 0, which means z_{ijkl} will be 0 by Constraint (192). This means the only ij pairs that have positive heights are those that are being used as matched pairs. The objective will only add up heights of pairs ij for which the absolute differences are positive, which was precomputed as δ_{ij}.

The z-score is computed from the optimal value of w_+ through the following formula, and the pvalue is computed as usual.

$$z = \frac{w_+(\mathbf{a}, \mathbf{z}, \mathbf{h}) - n(n+1)/4 - 1/2}{\sqrt{n(n+1)(2n+1)/24}}.$$

Appendix N: 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 θ is a binary variable and M is a positive number that is larger than the maximum values of both f and g:

$$\begin{aligned} -g(w) &\leq M\theta \\ f(w) &\leq M(1-\theta). \end{aligned}$$

In order for $f(w)$ to be positive, then θ must be 0, in which case, $g(w)$ is then restricted to be positive. If $f(w)$ is negative, θ 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 λ_1 and λ_2 are binary variables:

$$\begin{aligned} &\lambda_1 = 1 \text{ if } i + 1 > t \\ &\lambda_2 = 1 \text{ if } t + 1 > i \\ &S_i = k \text{ if } \lambda_1 + \lambda_2 > 1 \\ &S_i = 0 \text{ if } \lambda_1 + \lambda_2 < 2. \end{aligned}$$

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.

References

Chen DS, Batson R, Dang Y (2011) *Applied Integer Programming: Modeling and Solution* (Wiley).

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

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

Dudgeon DE, Mersereau RM (1983) *Multidimensional digital signal processing*. Prentice-hall Signal Processing Series (Prentice-Hall).

Fanaee-T H, Gama J (2014) Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence* 2(2-3):113–127, ISSN 2192-6352.

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

ILOG (2007) Cplex 11.0 user's manual. ILOG, Inc.

Lalonde R (1986) Evaluating the econometric evaluations of training programs. *American Economic Review* 76(4):604–620.

Morucci M, Rudin C (2018) Matching bounds: How choice of matching method affects treatment effect estimates and what to do about it. *Working Paper* URL `https://arxiv.org/abs/2009.02776`.

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

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

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

Zubizarreta J (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.