Using Approximation Algorithms to Build Evidence Factors and Related Designs for Observational Studies

Bikram Karmakar, Dylan S. Small and Paul R. Rosenbaum[1]

University of Pennsylvania, Philadelphia

Abstract. Observational or nonrandomized studies of treatment effects are often constructed with the aid of polynomial-time algorithms that optimally form matched treatment-control pairs or matched sets. Because each observational comparison may potentially be affected by bias, investigators often reinforce a single comparison with an additional comparison that is unlikely to be affected by the same biases, for instance using multiple control groups or evidence factors or control+instrument designs. Use of two comparisons affected by different biases may detect bias if the two comparisons disagree, or may show that two comparisons with different weakness concur in their conclusions. Even this simplest addition — a second comparison — creates design problems without polynomial-time solutions. Faced with a problem that no polynomial-time algorithm can solve, a so-called "approximation algorithm" is a type of compromise: it provides a solution in polynomial time that is provably not much worse than the unattainable optimal solution. Building upon existing techniques for related problems in operations research, we develop an approximation algorithm for minimum distance matching with near-fine balance for three comparison groups. This algorithm is a practical approach to most observational designs that add a second comparison. The method is applied to an observational study of the effects of side airbags on injury severity in the US Fatality Analysis Reporting System. For many car makes and models, side airbags were initially unavailable, then later available as optional equipment for an additional fee, then still later provided as standard equipment. Within sets matched for make and model of car, for safety belt use, for direction of impact,

[1]Bikram Karmakar is a doctoral student and Dylan S. Small and Paul R. Rosenbaum are professors in the Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340 US. 11 October 2018. dsmall@wharton.upenn.edu.

and other covariates, we compare crashes in these three periods, where each comparison has different limitations. The method is implemented in the `R` package `approxmatch`, whose example reproduces some of the calculations.

Keywords: Approximation algorithm; causal inference; evidence factors; matching; multiple control groups; observational study.

1 The need for approximation algorithms when constructing an observational design

1.1 Optimal constructions with polynomial-time algorithms

An algorithm is said to solve a problem in polynomial time if there is a polynomial, say n^3, such that the algorithm can solve any instance of the problem of size n in at most κn^3 arithmetic operations, where κ is a constant, and in this case the algorithm runs in $O\left(n^3\right)$-time. Generally, the constant multiplier, κ, depends upon the programming language, computer and other details, while the exponent, here 3, does not, so the focus of attention is on the exponent. Saying that a problem cannot be solved in polynomial time means saying that there are always problem instances such that it takes more than κn^a arithmetic operations to solve them, no matter how the constants κ and a are picked. If a problem cannot be solved in polynomial time, then large problems may be virtually impossible to solve. See Papadimitriou and Steiglitz (1982) or Korte and Vygen (2012) for general discussion of combinatorial optimization algorithms and their performance. In this section, we first mention common uses of polynomial-time algorithms to construct observational studies, then point out that even some very simple and important designs cannot be built in this way.

Many methods of constructing a design for an observational study solve a combinatorial optimization problem using a polynomial-time algorithm. For instance, many treatment-

control matching problems involving n people are reexpressed as optimal assignment problems or minimum cost flow problems in a network and are solved by algorithms that run in $O\left(n^3\right)$-time. In the simplest case, matched pairs are found to minimize the total over pairs of the covariate distance between the treated and control individuals within each pair; see, for instance, Rosenbaum (1989). The covariate distance might combine a propensity score with some form of Mahalanobis distance, or other techniques. The algorithm need not construct pairs: it may construct matched sets with two controls matched to each treated individual, or it may be a full matching in which each matched set contains either one treated subject and one or more controls, or one control and one or more treated individuals; see, for instance, Hansen and Klopfer (2006). It is often useful to add fine-balance or near-fine balance constraints to minimum distance matching: these minimize the total distance within pairs subject to the constraint that a nominal variable, perhaps with many categories such as ICD-10 surgical procedure, is as balanced as possible; see, for instance, Rosenbaum, Ross and Silber (2007), Yang et al. (2012) and Pimentel, Yoon and Keele (2015) for methods and Silber et al. (2016) for an application. These techniques are implemented in `R` in Hansen's `optmatch` package (Hansen and Klopfer 2006; Hansen, 2007) and Pimentel's `rcbalance` package (Pimentel, 2016, 2017). Zubizarreta (2012) enlarges the scope of matching techniques in his `designmatch` package using mixed integer programming methods that often perform well despite lacking an explicit time bound (Zubizarreta and Kilcioglu 2016); Keele, Titiunik and Zubizarreta (2015) provide an application to enhancing regression discontinuity designs through matching.

A second polynomial-time algorithm used to design observational studies involves minimum distance nonbipartite matching; see Lu et al. (2011) and the references given there. Where bipartite (or two-part) matching pairs individuals from two groups, treated or control, nonbipartite (or, awkwardly, not-two-part) matching begins with a single population.

Nonbipartite matching splits a single population into nonoverlapping matched pairs in such a way that the total distance within pairs is minimized. Derigs' (1988) algorithm is available in the `R` package `nbpMatching` (Beck et al. 2016). One use of nonbipartite matching is in strengthening a weak instrument; see Baiocchi et al. (2010) and Keele and Morgan (2016) for discussion and Lorch et al. (2012) for an application.

For general discussion of matching in observational studies, see Rosenbaum (2010, Part II; 2017, Chapter 11) and Stuart (2010).

1.2 Closely related problems are much more difficult

If instead of matching two groups, treated and control, to minimize the distance within pairs, as in §1.1, we wished to match three groups of equal size to form minimum distance matched triples — the 3-dimensional assignment problem of Pierskalla (1968) — then the problem is believed to be very difficult and is classified among problems believed to have no solution by a polynomial-time algorithm; see Crama and Spieksma (1992, Theorem 1). Crama and Spieksma (1992, §3) proposed several approximation algorithms for matching everyone in three groups of initially equal size. These algorithms are not immediately applicable to statistical problems, because: (i) comparison groups are rarely of equal size prior to matching, (ii) we may want multiple controls from some groups, and (iii) one typically imposes additional constraints, such as fine-balance or near-fine balance for certain nominal variables; see §3.2. In §3, we employ ideas from Crama and Spieksma (1992) to produce a polynomial time approximation algorithm that incorporates features (i)-(iii). Before discussing the algorithm, we motivate its use in an application in §2.

2 Motivating Application: Effects of Side Airbags in Crashes

2.1 Side airbags: at first unavailable, then optional, then standard

According to the Insurance Institute for Highway Safety (http://www.iihs.org/iihs/ratings/safety-features), side airbags protecting both the driver's head and torso were unavailable in 1997, and were standard equipment on 97.9% of cars in the 2017 model year. Over 20 years, more and more makes and models of cars gradually added side airbags, often offering them initially as optional equipment for an additional fee, later providing them as standard equipment. For instance, the Volvo C70 did not have side airbags in 1998, had them as optional equipment from 1999 to 2002, and then had them as standard equipment in 2003. The Nissan Altima did not have side airbags in 2006, had them as optional equipment from 2007-2009, and had them as standard equipment in 2010. It is not attractive to judge the safety effects of side airbags by comparing a Volvo C70 to a Nissan Altima, because side airbags are only one difference between these vehicles and their drivers. Perhaps Volvos attract drivers concerned with safety, with the possible consequence that Volvos are driven differently from, say, Dodge Chargers.

More attractive is to compare crashes of the same make and model of car in eras that differ in availability of side airbags. For instance, one might compare a crash of a 1998 Volvo C70 to a crash of a 2003 Volvo C70, the latter having a side airbag. Presumably, the decision to purchase a Volvo C70 in 1998 rather than in 2003 mostly reflects an individual's need to acquire or replace a car in those years, not a greater concern with automotive safety by the purchaser of the 2003 model with side airbags. Nonetheless, the world changed in many ways between 1998 and 2003, not just in the addition of side airbags to Volvo C70s, so the wide separation in time raises other concerns. The comparison of the era before side airbags and the era with side airbags as standard equipment is, in certain respects, an

attractive natural experiment, but it is not a perfect one.

In contrast, the middle era of optional side airbags has both attractions and concerns. Models built in the same or adjacent years may be more similar than models separated by five years. An attraction of the optional era is that one could compare a 2002 Volvo C70 to a 2003 Volvo C70, or a 1998 Volvo C70 to a 1999 Volvo C70. Indeed, one could compare two 2000 Volvo C70's, one with a side airbag, the other without. The concern is that, during the optional period, 1999-2002 for Volvo C70s, some drivers paid extra for side airbags and others declined to do so, perhaps indicating their different levels of concern with traffic safety. Perhaps a driver concerned with automotive safety expresses that concern in more than one way, say buying a side airbag, driving soberly and slowly, and abstaining from tailgating, so that comparisons within the optional era confound side airbags with other safety behavior. Presumably, the decision to purchase a Volvo C70 in 2000, rather than 1998 or 2003, again mostly reflects an individual's need to acquire a car in 2000, but the secondary decision, paying extra for side airbags in 2000, could be strongly related to other unmeasured safe driving behaviors. So the optional era nudges people towards side airbags, but it does not determine whether they acquire a side airbag or not. One conventional strategy would compare the optional era to one of the other two eras, viewing the optional era as an instrument or instrumental variable for the purchase of a side airbag, thereby side-stepping the decisions of individual drivers about paying extra for a side airbag; see Baiocchi et al. (2014) for a review of instruments. Viewing the optional era in this way, an assumed exclusion restriction would attribute changes in injuries over eras to changes in the changed frequency of side airbags, leading to an estimate of the complier average effect of side airbags; see Angrist et al. (1996).

We will form matched sets consisting of crashes involving the same make and model of car, one before side airbags were available for this make and model, one after side airbags

became standard equipment, and between 1 or 3 crashes in the optional period. As noted above for the Volvo C70 and Nissan Altima, the years involved vary from one make and model to another. We tried to find 3 crashes in the optional period because most buyers did not buy side airbags during the optional period. However, if there were too few crashes to form 1-3-1 matched sets, perhaps because the optional period was brief, we formed 1-1-1 matched triples instead. Ultimately, there were 978 matched sets of the form 1-3-1, and 1398 sets of the form 1-1-1. The data came from the US Fatality Analysis Reporting System, described in §2.2.

2.2 The Fatality Analysis Reporting System

Provided by the US National Highway Traffic Safety Administration, the US Fatality Analysis Reporting System (FARS) records information about motor vehicle accidents with at least one fatality. The FARS data contain extensive information about the vehicles involved, some information about the occupants of the vehicles including a measure of severity of injury, 0 for uninjured to 4 for death, and some information about the circumstances of the crash. No doubt, crashes in FARS are unrepresentative of all crashes, because every crash in FARS was severe enough to cause at least one death.

Some care is needed when using FARS data to examine the effects of safety equipment. A crash is recorded only if there is at least one fatality. In FARS, a crash involving a lone driver hitting a tree is always lethal for the driver, not because driving alone is dangerous, nor because trees are invariably deadly, but because a crash involving just one person is recorded in FARS only if that person died. More subtly, if safety equipment prevents all deaths in a crash, then it also prevents the accident from being recorded in FARS, whereas removing the safety equipment might have caused a death, so the same crash would be recorded in FARS. For discussion of issues that arise when a treatment can cause data to

go uncollected, see Rosenbaum (2005).

We looked at crashes involving at least two vehicles between 1995 and 2015. We included cars, minivans, SUVs and pickup trucks, but excluded motor cycles and large trucks. In such crashes, we picked at random one vehicle with at least one death, discarding that vehicle. The remaining vehicle or vehicles may or may not have had a death. We do know, however, that data on the remaining vehicles that we studied would have been collected by FARS whether or not side airbags or their absence caused or prevented a death in the vehicle, because the discarded vehicle would, in either case, have prompted FARS to collect data about this accident. This selection process makes the vehicles we examined unrepresentative of vehicles in FARS, but it avoids a tautological source of bias. We may hope that the selection process makes vehicles unrepresentative in a parallel manner in the three eras that we examine, the era prior to side airbags, the optional era, and the era when they were standard equipment. For brevity in tables and figures, these three eras are called "none", "optional" and "all", and by definition the years involved vary from one make and model to another. We considered only makes and models that had cars in all three eras, so that, for instance, we would exclude a new make of car that was first sold in 2012 with standard side airbags.

Many car models had no cars in one or more eras, "none", "optional" and "all". For instance, a new car model might have had side airbags from the beginning. A discontinued car model might never have had side airbags. A car model might go from "none" to "all" with no "optional" period. We required a car model to have at least 40 cars in each of the three periods, "none", "optional" and "all", and we used these 40 cars to define the eras. There were 31,505 cars that qualified. This number was slightly reduced by 2,282 cars due to missing data on key variables. For each car model, we matched the smallest group, "none", "optional" or "all", to the two larger groups. Where possible, we matched

1-3-1, none-optional-all, because most cars in the optional period did not have side airbags, so selecting three cars increased the chance that one had side airbags. If 1-3-1 was not possible because 3 optional era cars were not available, we matched 1-1-1. This yielded 2376 matched sets. In the end, we had 978 matched 1-3-1 sets and 1398 matched 1-1-1 sets, where $2376 = 1398 + 978$, with a total of 9084 cars in these 2376 matched sets.

The matching used the new algorithm in §3. The `R` package `approxmatch` implements the procedure in §3; Karmaker (2017). That package includes a data frame called `Dodgeram` with 6953 crashes involving Dodge Ram trucks. The examples in the documentation for the `kwaymatching` function in the `approxmatch` package in `R` reproduce the 3-way matching of the Dodge Ram trucks.

2.3 Matched crashes

Table 1 describes covariate balance after matching. In Table 1, each matched set counts the same, so the optional period for a 1-1-1 triple contributes one driver age for the optional period, but a 1-3-1 matched set contributes one average of three driver ages. Table 1 shows characteristics of the driver, such as age, and of the crash, such the direction of impact and whether a fire or explosion occurred. By definition, we did not match for the model year, nor did we match for the crash year. The years are out of balance by definition: for each make and model, the none-era precedes the optional era which precedes the all-era. In the optional era, about 18% of owners had purchased cars with side airbags. We also matched for some additional variables not shown in Table 1, such as characteristics of the right front passenger if there was one, and the stated highway speed limit which may or may not have been heeded.

Table 1 shows the balance achieved by matching, but it does not contrast the situations before and after matching. Figures 1 and 2 show this contrast, with the left bar showing

the situation before matching, the right bar showing the situation after matching. We hope to see that the right bars, after matching, are of similar height, and indeed they are. Plots use `ggplot2` (Wickham 2009). Notably in Figure 1, belt use and age are similar after matching, but they were different before matching. The mean driver's age was about four years younger in the "none" period than in the "all" period, perhaps because the baby-boomers are aging, and there was also about a 12% increase in use of safety belts over this period. Also, there were more female drivers in the later "all" period. Of course, the match controlled for age, gender and restraint use, but there could be other factors that were not measured. Notably in Figure 2, the direction of impact shifted slightly as the periods passed, with an increase in rear and right impacts, and a decrease in front impacts and rollovers. Recall that FARS records only lethal crashes, so this is a change in the pattern of lethal crashes, not necessarily a change in the pattern of all impacts. Again, the matching removed these differences.

The match used a distance that satisfied the triangle inequality, and it finely balanced several covariates and was exact for the presence of a right front passenger. Specifically, the match near-finely balanced indicators of rollover and fire occurrence during the crash. Following the idea in §4, the distance was the weighted sum of distances involving the absolute difference in logits of the propensity score, a rank based Mahalanobis distances for occupant characteristics, and two other rank based Mahalanobis distances for direction of impact and safety belt use; see the examples in the `approxmatch` package in `R`. Weighting several distances permits the combined distance to satisfy the triangle inequality while giving the analyst control of the relative importance of variables in the match. In principle, finding an optimal 1-1-1 match or 1-3-1 match is for all intents impossible in large problems. The match was produced using the approximation algorithm we describe and develop in §3. The approximation algorithm runs in polynomial time.

3 Problem, algorithm, and guarantee

3.1 Matching structure and distance

There are three disjoint sets of units, $\mathcal{I} = \{1, \ldots, I\}$, $\mathcal{J} = \{1, \ldots, J\}$ and $\mathcal{K} = \{1, \ldots, K\}$ with $I \leq J \leq K$. In §2, the sets $\mathcal{I}$, $\mathcal{J}$, and $\mathcal{K}$ were eligible car crashes in the three eras, "none", "optional" and "all".

There is a distance, $\delta_{ij} \geq 0$, between pairs of units, $i \in \mathcal{I}$ and $j \in \mathcal{J}$, a distance $\delta'_{ik} \geq 0$ between pairs of units in $i \in \mathcal{I}$ and $k \in \mathcal{K}$, and a distance $\delta''_{jk} \geq 0$ between pairs of units in $j \in \mathcal{J}$ and $k \in \mathcal{K}$. Write $\boldsymbol{\delta}$ for the vector of $IJ + IK + JK$ distances, δ_{ij}, δ'_{ik}, δ''_{jk}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. Although it is suggestive to call the numbers in $\boldsymbol{\delta}$ distances, they are required to have some but not all of the properties of distances in a metric space. Precisely, we require without further mention that the entries in $\boldsymbol{\delta}$ be nonnegative, possibly infinite, and satisfy a part of the triangle inequality, namely $\delta''_{jk} \leq \delta_{ij} + \delta'_{ik}$, and we call such a $\boldsymbol{\delta}$ a "matching distance array", or briefly a "distance". Notice that the triangle inequality bounds distances δ''_{jk} between units in $j \in \mathcal{J}$ and $k \in \mathcal{K}$, but need not bound δ_{ij} for $i \in \mathcal{I}$ and $j \in \mathcal{J}$, nor δ'_{ik} for $i \in \mathcal{I}$ and $k \in \mathcal{K}$. We need $\delta''_{jk} \leq \delta_{ij} + \delta'_{ik}$, but not other implications of the triangle inequality, because our algorithm makes δ_{ij} and δ'_{ik} small, then concludes that δ''_{jk} could not be very big by virtue of this inequality. Distances are not defined for two units in $\mathcal{I}$, nor for two units in $\mathcal{J}$, nor for two units in $\mathcal{K}$; rather, distances are defined between units in different sets. We understand $\delta''_{jk} \leq \delta_{ij} + \delta'_{ik}$ to hold trivially if either $\delta_{ij} = \infty$ or $\delta'_{ik} = \infty$. For instance, if individual i has covariates $\mathbf{x}_i$, individual j has covariates $\mathbf{x}_j$, and $\boldsymbol{\Sigma}$ is the covariance matrix of the covariates, then the Mahalanobis distance defined as $\delta_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$ satisfies the triangle inequality. Although the Mahalanobis distance yields a formula, we do not assume that a single formula produced the δ_{ij}, δ'_{ik}, δ''_{jk}, leaving open the possibility

that δ''_{jk} is defined differently from δ_{ij} or δ'_{ik}, requiring only that $\delta''_{jk} \leq \delta_{ij} + \delta'_{ik}$. Section 4 discusses various useful distances that satisfy the required conditions.

Let $\rho \geq 1$ be an integer such that $\rho \leq J/I$ and let $\kappa \geq 1$ be an integer such that $\kappa \leq K/I$. A common and basic case is $(\rho, \kappa) = (1, 1)$. We want to construct a closely matched and balanced blocked study such that each block contains 1 unit from $\mathcal{I}$, ρ units from $\mathcal{J}$ and κ units from $\mathcal{K}$, and no units appear more than once. If $(\rho, \kappa) = (1, 1)$, then we want matched triples with one unit each from $\mathcal{I}$, $\mathcal{J}$ and $\mathcal{K}$.

More precisely, a $(1 + \rho + \kappa)$-tuple $B = (i,\ j_1,\ j_2, \ldots,\ j_\rho,\ k_1,\ k_2, \ldots,\ k_\kappa)$ is a (ρ, κ)-block if $i \in \mathcal{I},\ j_1 \in \mathcal{J}, \ldots,\ j_\rho \in \mathcal{J},\ k_1 \in \mathcal{K},\ \ldots,\ k_\kappa \in \mathcal{K}$ where $j_1, \ldots,\ j_\rho$ are distinct and $k_1, \ldots,\ k_\kappa$ are distinct. A (ρ, κ)-design $\mathcal{B}$ is a collection of blocks B such that every unit $i \in \mathcal{I}$ appears in exactly one block, and each $j \in \mathcal{J}$ and each $k \in \mathcal{K}$ appears in at most one block $B \in \mathcal{B}$.

A block $B = (i,\ j_1,\ j_2, \ldots,\ j_\rho,\ k_1,\ k_2, \ldots,\ k_\kappa)$ has total between group distance

$$\delta_B = \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} + \sum_{m=1}^{\kappa} \delta'_{i,k_m} + \sum_{\ell=1}^{\rho} \sum_{m=1}^{\kappa} \delta''_{j_\ell,k_m}. \tag{1}$$

If δ_B is very small, then i is typically close to $j_1,\ j_2, \ldots,\ j_\rho$ and close to $k_1,\ k_2, \ldots,\ k_\kappa$, each j in the block is close to each k in the block. The design $\mathcal{B}$ has total distance $\delta_{\mathcal{B}} = \sum_{B \in \mathcal{B}} \delta_B$. We would prefer a design in which $\delta_{\mathcal{B}}$ is small.

The distance in (1) worries about distances between $\mathcal{I}$, $\mathcal{J}$, and $\mathcal{K}$, not distances within them. To see why this is reasonable, consider a simple case. Suppose that δ_{ij} is the absolute difference in age and $\rho = 2$. If i has age 25, j_1 has age 24 and j_2 has age 26, then $\sum_{\ell=1}^{\rho} \delta_{i,j_\ell}$ in (1) is $|25 - 24| + |25 - 26| = 2$. In contrast, if i is 24, j_1 is 25 and j_2 is 26, then $\sum_{\ell=1}^{\rho} \delta_{i,j_\ell}$ is $|24 - 25| + |24 - 26| = 3$. So we prefer the first triple of ages to the second, and this makes sense because in the second distribution of ages i is younger than both j_1 and j_2, so they are inferior as a control group. Had the first term in (1) included

a distance between the two units from $\mathcal{J}$, with three terms instead of $\delta_{i,j_1} + \delta_{i,j_2}$ in the first sum in (1), then $|25-24| + |25-26| + |24-26| = 4$ for both distributions of age, so the distance would not represent our preference for the first distribution of ages.

3.2 Fine balance and near-fine balance of nominal categories

Fine balance entails equating, in two or more groups, the marginal distributions of a nominal covariate, often a covariate with many levels, without worrying about whether individuals are paired for this covariate. A very large completely randomized experiment balances the marginal distributions of covariates without pairing individuals, and fine balance aims at an analogous form of balance for an observed nominal covariate. Fine balance can ensure that many nominal categories are balanced, while permitting the pairing to focus on other covariates strongly associated with the outcome.

There are $C \geq 1$ mutually exclusive and exhaustive categories, $\mathcal{C}_1, \ldots, \mathcal{C}_C$, such that every unit, $i \in \mathcal{I}$ belongs to exactly one category, say $i \in \mathcal{C}_c$ for one specific c, and the same is true for every unit, $j \in \mathcal{J}$ and every unit $k \in \mathcal{K}$. For example, $\mathcal{C}_1$ might be the set of females and $\mathcal{C}_2$ might be the set of males. More commonly, $\mathcal{C}_1, \ldots, \mathcal{C}_C$ might represent dozens or hundreds of categories, say principal surgical procedures or car models. The trivial case $C = 1$ places everyone in the same category, and it will permit a single theorem to cover the situations with categories ($C > 1$) and without categories ($C = 1$). Write f_{cg} for the number of units in category c from group $g = 1$, 2, 3, where group 1 is $\mathcal{I}$, group 2 is $\mathcal{J}$ and group 3 is $\mathcal{K}$.

If $\mathcal{B}$ is a (ρ, κ)-design, then it exhibits a certain degree of imbalance with respect to the categories $\mathcal{C}_1, \ldots, \mathcal{C}_C$. Write $\mathbf{f}_\mathcal{B}$ for a $C \times 3$ matrix of counts defined as follows. For (ρ, κ)-design $\mathcal{B}$, the count, $f_{\mathcal{B}cg}$ in row c and column g is the number of units in $\mathcal{B}$ in category c from group g. So, by definition, the column totals are $\sum_{c=1}^{C} f_{\mathcal{B}c1} = I$, $\sum_{c=1}^{C} f_{\mathcal{B}c2} = \rho I$ and

$\sum_{c=1}^{C} f_{\mathcal{B}c2} = \kappa I$. Because $\mathcal{B}$ takes everyone in group $\mathcal{I}$, we have $f_{c1} = f_{\mathcal{B}c1}$ for each c, and because $\mathcal{B}$ takes some people from groups $\mathcal{J}$ and $\mathcal{K}$, we have $f_{c2} \geq f_{\mathcal{B}c2}$ and $f_{c3} \geq f_{\mathcal{B}c3}$.

The (ρ, κ)-design $\mathcal{B}$ is finely balanced if $f_{\mathcal{B}c1} = f_{\mathcal{B}c2}/\rho = f_{\mathcal{B}c3}/\kappa$ for $c = 1, \ldots, C$; that is, every matched group has the same proportion of individuals in each category. Fine balance is discussed in Rosenbaum (1989, §3.2; Rosenbaum 2010, §10). Fine balance is not always feasible. Write $f_{c\min} = \min(f_{c1},\, f_{c2}/\rho,\, f_{c3}/\kappa)$. If $f_{c\min} = f_{c1}$ for $c = 1, \ldots, C$, then fine balance is feasible, and if it is feasible, then we wish to require it. The (ρ, κ)-design $\mathcal{B}$ is near-fine or exhibits near-fine balance if $f_{\mathcal{B}c1} \geq f_{c\min}$, $f_{\mathcal{B}c2}/\rho \geq f_{c\min}$ and $f_{\mathcal{B}c3}/\kappa \geq f_{c\min}$ for $c = 1, \ldots, C$. Near-fine balance is always feasible, and when fine balance is feasible, near-fine balance implies fine-balance. In a sense, near-fine balance exhibits minimal deviation from fine balance. Requiring near-fine balance when $C = 1$ imposes no constraint, and in this case the focus is entirely on minimizing distance, $\delta_{\mathcal{B}}$, with no concern for balance over categories. This definition of near-fine balance is similar to the definition in Yang et al. (2012) but has been adjusted to permit three groups instead of a treated and a control group.

The traditional three-dimensional assignment problem is to minimize $\delta_{\mathcal{B}}$ with $C = 1$, $\rho = \kappa = 1$, and $I = J = K$, and even in this simplest case, finding an optimal solution is not practical; see Crama and Spieksma (1992). Consider the general problem of finding a near-fine (ρ, κ)-design $\mathcal{B}$ with a small distance, $\delta_{\mathcal{B}}$. We propose an approximation algorithm for general C, ρ, κ, I, J, K: it finds a near-fine (ρ, κ)-design with a total distance, $\delta_{\mathcal{B}}$, that is at most $1 + \max(\rho, \kappa)$ times the minimum distance for all near-fine (ρ, κ)-designs. The approximation algorithm runs in $O\left(K^3\right)$ time in the worst case, and for matched triples, $\rho = \kappa = 1$, it produces a value of $\delta_{\mathcal{B}}$ that is at most twice the true minimum as $1 + \max(\rho, \kappa) = 2$.

3.3 An approximation algorithm

Define a partial block $P = (i,\ j_1,\ j_2, \ldots,\ j_\rho)$ to be $\rho + 1$ distinct units with $i \in \mathcal{I}$ and $j_1,\ j_2, \ldots,\ j_\rho \in \mathcal{J}$. A set $\mathcal{P}$ of partial blocks will be called acceptable if it is the initial part of some near-fine (ρ, κ)-design $\mathcal{B}$; however, this needs to be said with a bit more care. A set $\mathcal{P}$ of partial blocks P is compatible with a near-fine (ρ, κ)-design $\mathcal{B}$ if each block $B = (i,\ j_1,\ j_2, \ldots,\ j_\rho,\ k_1,\ k_2, \ldots,\ k_\kappa) \in \mathcal{B}$ has an initial segment $P = (i,\ j_1,\ j_2, \ldots,\ j_\rho)$ that is a partial block in $\mathcal{P}$. A set $\mathcal{P}$ of partial blocks is defined to be acceptable if its partial blocks are compatible with at least one near fine (ρ, κ)-design $\mathcal{B}$. If $\mathcal{P}$ is acceptable, then it necessarily follows that its partial blocks are nonoverlapping and $f_{\mathcal{B}c1} \geq f_{c\,\min}$, $f_{\mathcal{B}c2}/\rho \geq f_{c\,\min}$. An unexciting but nonetheless important subtlety here is that $f_{\mathcal{B}c1}$ and $f_{\mathcal{B}c2}$ are determined by $\mathcal{P}$ without reference to group $\mathcal{K}$, but $f_{c\,\min}$ is defined in a way that involves group $\mathcal{K}$.

The following two-step algorithm first assigns j's in $\mathcal{J}$ to each $i \in \mathcal{I}$ to produce an acceptable set $\mathcal{P}$ of partial blocks $P = (i,\ j_1,\ j_2, \ldots,\ j_\rho)$; then, it assigns k's in $\mathcal{K}$ to each partial block, P.

Step 1: Match $\mathcal{I}$ and $\mathcal{J}$ to form an acceptable set $\mathcal{P}$ of partial blocks of minimal distance

$$\sum_{(i,\ j_1,\ j_2, \ldots,\ j_\rho) \in \mathcal{P}} \sum_{\ell=1}^{\rho} \delta_{i,j_\ell}. \tag{2}$$

Step 2: Extend each partial block $P = (i,\ j_1,\ j_2, \ldots,\ j_\rho) \in \mathcal{P}$ into a block

$$B = (i,\ j_1,\ j_2, \ldots,\ j_\rho,\ k_1,\ k_2, \ldots,\ k_\kappa)$$

so that the resulting collection of blocks $\mathcal{B}$ is a near-fine (ρ, κ)-design $\mathcal{B}$ that minimizes

$$\sum_{(i,\, j_1,\, j_2,\ldots,\, j_\rho)\in\mathcal{P}} \sum_{m=1}^{\kappa} \left(\delta'_{i,k_m} + \sum_{\ell=1}^{\rho} \delta''_{j_\ell,k_m} \right).$$

Step 1 matches individuals in $\mathcal{I}$ to individuals in $\mathcal{J}$ forming partial blocks, while Step 2 takes those partial blocks and matches each partial block to individuals in $\mathcal{K}$. In other words, the entire procedure consists of two matches, one of individuals to individuals, the other of partial blocks to individuals. The entire procedure is suboptimal because Step 1 makes decisions with no allowance for their consequences in Step 2, but we will show that the procedure's mistakes are limited in size. It is the triangle inequality that limits the size of the errors.

Consider the requirement of near-fine balance, namely the requirement that $f_{\mathcal{B}c1} \geq f_{c\min}$, $f_{\mathcal{B}c2}/\rho \geq f_{c\min}$, and $f_{\mathcal{B}c1}/\kappa \geq f_{c\min}$. This definition of near-fine balance refers to all three groups because $f_{c\min}$ is derived from all three groups. The definition of $f_{c\min}$ ensures that the requirement of near-fine balance is feasible: a (ρ, κ)-design $\mathcal{B}$ exhibiting near-fine balance always exists, albeit perhaps with an infinite total distance if some distances δ are infinite. Because Step 1 requires the partial blocks to be acceptable, they are compatible with near-fine balance; that is, the partial blocks are the initial parts of the blocks of some (ρ, κ) design $\mathcal{B}$ that exhibits near-fine balance. Step 2 requires that these partial blocks be extended to complete blocks so that the resulting (ρ, κ)-design $\mathcal{B}$ exhibits near-fine balance. So, by the definitions of Steps 1 and 2, the algorithm returns a (ρ, κ)-design $\mathcal{B}$ exhibiting near-fine balance, albeit perhaps with an infinite total distance.

Can Steps 1 and 2 be performed? Indeed they can, and in computational time that is $O\left(K^3\right)$. Step 1 can be done in $O\left(K^3\right)$ steps by solving a minimum cost network flow problem, matching elements of $\mathcal{I}$ to ρ elements of $\mathcal{J}$ with a requirement of near-

fine balance defined by the given values $f_{c\min}$; see Rosenbaum (1989, §3.2) or Yang et al. (2012, Appendix) with very minor changes to accommodate the value $f_{c\min}$ obtained from all three groups. (Specifically, in the language of these references, the capacity of the edge from the category c node to the sink is set to $\rho f_{c\min}$ in Step 1 and set to $\kappa f_{c\min}$ in Step 2, with one additional bypass node with capacity $\rho I - \rho \sum f_{c\min}$ in Step 1 or capacity $\kappa I - \kappa \sum f_{c\min}$ in Step 2 delivering the remaining flow to the sink.) The match in Step 2 can also be done in $O\left(K^3\right)$ steps by solving a minimum cost network flow problem, now matching each $P \in \mathcal{P}$ to κ elements of $\mathcal{K}$, so that the selected controls from $\mathcal{K}$ again exhibit near-fine balance defined by the given values of $f_{c\min}$.

In brief, the approximation algorithm entails solving two minimum cost flow problems, each of which runs in $O\left(K^3\right)$ steps, so the approximation algorithm itself is a polynomial time algorithm that runs in $O\left(K^3\right)$ steps.

3.4 A bound on the error of approximation

Proposition 1 says that the polynomial-time algorithm in §3.3 finds a near-fine design $\mathcal{B}$ whose total distance $\delta_{\mathcal{B}}$ is at most a fixed multiple of the distance $\delta_{\overline{\mathcal{B}}}$ for the unattainable optimal near-fine design, say $\overline{\mathcal{B}}$. The multiplier, $1 + \max(\rho, \kappa)$, in Proposition 1 equals 2 in the common case with $\rho = \kappa = 1$.

The proof of Proposition 1 extends certain ideas from Crama and Spieksma (1992) for the simpler 3-dimensional assignment problem to matching with (i) near-fine balance, (ii) unequal initial groups, $I \leq J \leq K$, and (iii) matching with multiple controls, $\rho \geq 1$ and $\kappa \geq 1$. One device they use with $I = J = K$ cannot be used here: Step 1 of our algorithm must start with the smallest group, $\mathcal{I}$.

Proposition 1 *Let $\overline{\mathcal{B}}$ be a near fine (ρ, κ)-design whose distance $\delta_{\overline{\mathcal{B}}}$ is minimal. Let $\mathcal{B}$ be produced by Steps 1 and 2 in §3.3. Then $\mathcal{B}$ is a near-fine (ρ, κ)-design with $\delta_{\mathcal{B}} \leq$*

$\{1 + \max(\rho, \kappa)\}\ \delta_{\overline{\mathcal{B}}}$.

Proof. Let $\mathcal{P}$ be the set of partial blocks $P = (i,\ j_1,\ j_2, \ldots,\ j_\rho)$ in $\mathcal{B}$ produced in Step 1. We now define a compromise $\widetilde{\mathcal{B}}$ between $\mathcal{B}$ and $\overline{\mathcal{B}}$, anchoring the compromise by units $i \in \mathcal{I}$. Define $\widetilde{\mathcal{B}}$ to be a (ρ, κ)-design with near-fine balance whose blocks $\widetilde{B} = (i,\ j_1,\ j_2, \ldots,\ j_\rho,\ k_1, \ldots,\ k_\kappa)$ are such that $P = (i,\ j_1,\ j_2, \ldots,\ j_\rho) \in \mathcal{P}$ is a partial block of $\mathcal{P}$, and for some $\ell_1, \ldots,\ \ell_\rho \in \mathcal{J}$, $\overline{B} = (i,\ \ell_1, \ldots,\ \ell_\rho,\ k_1, \ldots,\ k_\kappa) \in \overline{\mathcal{B}}$ is a block of the optimal design. By the triangle inequality, we always have

$$\delta''_{j_\ell, k_m} \leq \delta_{i, j_\ell} + \delta'_{i, k_m} \tag{3}$$

By definition,

$$\delta_{\mathcal{B}} = \sum_{(i,\ j_1,\ j_2, \ldots,\ j_\rho,\ k_1, \ldots,\ k_\kappa) \in \mathcal{B}} \left(\sum_{\ell=1}^{\rho} \delta_{i, j_\ell} + \sum_{m=1}^{\kappa} \delta'_{i, k_m} + \sum_{\ell=1}^{\rho} \sum_{m=1}^{\kappa} \delta''_{j_\ell, k_m} \right).$$

Now $\mathcal{B}$ and $\widetilde{\mathcal{B}}$ have the same partial blocks, $P = (i,\ j_1,\ j_2, \ldots,\ j_\rho)$, but for these fixed partial blocks, Step 2 completed the partial blocks in $\mathcal{B}$ as $(i,\ j_1,\ j_2, \ldots,\ j_\rho,\ k_1, \ldots,\ k_\kappa)$ where the k's were chosen to minimize $\sum_{(i,\ j_1,\ j_2, \ldots,\ j_\rho,\ k_1, \ldots,\ k_\kappa) \in \mathcal{B}} \left(\sum_{m=1}^{\kappa} \delta'_{i, k_m} + \sum_{\ell=1}^{\rho} \sum_{m=1}^{\kappa} \delta''_{j_\ell, k_m} \right)$; therefore,

$$\delta_{\mathcal{B}} \leq \delta_{\widetilde{\mathcal{B}}} = \sum_{(i,\ j_1,\ j_2, \ldots,\ j_\rho,\ k_1, \ldots,\ k_\kappa) \in \widetilde{\mathcal{B}}} \left(\sum_{\ell=1}^{\rho} \delta_{i, j_\ell} + \sum_{m=1}^{\kappa} \delta'_{i, k_m} + \sum_{\ell=1}^{\rho} \sum_{m=1}^{\kappa} \delta''_{j_\ell, k_m} \right). \tag{4}$$

Applying the triangle inequality (3) to $\sum_{\ell=1}^{\rho} \sum_{m=1}^{\kappa} \delta''_{j_\ell, k_m}$ in (4) yields

$$\delta_{\mathcal{B}} \leq \delta_{\widetilde{\mathcal{B}}} \leq \sum_{(i,\ j_1,\ j_2, \ldots,\ j_\rho,\ k_1, \ldots,\ k_\kappa) \in \widetilde{\mathcal{B}}} \left\{ (1 + \kappa) \sum_{\ell=1}^{\rho} \delta_{i, j_\ell} + (1 + \rho) \sum_{m=1}^{\kappa} \delta'_{i, k_m} \right\}$$

$$\leq \{1+\max(\rho,\kappa)\} \sum_{(i,\, j_1,\, j_2,\ldots,\, j_\rho,\, k_1,\ldots,\, k_\kappa)\in\widetilde{\mathcal{B}}} \left\{ \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} + \sum_{m=1}^{\kappa} \delta'_{i,k_m} \right\} . \tag{5}$$

Because $\mathcal{B}$ and $\widetilde{\mathcal{B}}$ have the same partial blocks, $P = (i,\ j_1,\ j_2, \ldots,\ j_\rho)$, and Step 1 picked the $(j_1,\ j_2, \ldots,\ j_\rho)$ in these blocks to minimize $\sum_{(i,\, j_1,\, j_2,\ldots,\, j_\rho,\, k_1,\ldots,\, k_\kappa)\in\mathcal{B}} \sum_{\ell=1}^{\rho} \delta_{i,j_\ell}$,

$$\sum_{(i,\, j_1,\, j_2,\ldots,\, j_\rho,\, k_1,\ldots,\, k_\kappa)\in\widetilde{\mathcal{B}}} \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} = \sum_{(i,\, j_1,\, j_2,\ldots,\, j_\rho,\, k_1,\ldots,\, k_\kappa)\in\mathcal{B}} \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} \leq \sum_{(i,\, j_1,\, j_2,\ldots,\, j_\rho,\, k_1,\ldots,\, k_\kappa)\in\overline{\mathcal{B}}} \sum_{\ell=1}^{\rho} \delta_{i,j_\ell}. \tag{6}$$

By the definition of $\widetilde{B}$,

$$\sum_{(i,\, j_1,\, j_2,\ldots,\, j_\rho,\, k_1,\ldots,\, k_\kappa)\in\widetilde{\mathcal{B}}} \sum_{m=1}^{\kappa} \delta'_{i,k_m} = \sum_{(i,\, j_1,\, j_2,\ldots,\, j_\rho,\, k_1,\ldots,\, k_\kappa)\in\overline{\mathcal{B}}} \sum_{m=1}^{\kappa} \delta'_{i,k_m}. \tag{7}$$

Combining (5), (6) and (7) yields

$$\begin{aligned}
\delta_{\mathcal{B}} &\leq \{1+\max(\rho,\kappa)\} \sum_{(i,\, j_1,\, j_2,\ldots,\, j_\rho,\, k_1,\ldots,\, k_\kappa)\in\overline{\mathcal{B}}} \left\{ \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} + \sum_{m=1}^{\kappa} \delta'_{i,k_m} \right\} \\
&\leq \{1+\max(\rho,\kappa)\} \sum_{(i,\, j_1,\, j_2,\ldots,\, j_\rho,\, k_1,\ldots,\, k_\kappa)\in\overline{\mathcal{B}}} \left\{ \sum_{\ell=1}^{\rho} \delta_{i,j_\ell} + \sum_{m=1}^{\kappa} \delta'_{i,k_m} + \sum_{\ell=1}^{\rho}\sum_{m=1}^{\kappa} \delta''_{j_\ell,k_m} \right\} \\
&= \{1+\max(\rho,\kappa)\}\ \delta_{\overline{\mathcal{B}}}.
\end{aligned}$$

■

3.5 The bound is tight

For $\rho = \kappa = 1$, or 1-1-1 matching, the bound in Proposition 1 is $\delta_{\mathcal{B}} \leq \{1+\max(\rho,\kappa)\}\ \delta_{\overline{\mathcal{B}}} = 2\,\delta_{\overline{\mathcal{B}}}$. This bound cannot be improved without additional assumptions. To see this, consider $\mathcal{I} = \{i\}$, $\mathcal{J} = \{j_1, j_2\}$, $\mathcal{K} = \{k_1, k_2\}$, where $x_i = 1$, $x_{j_1} = x_{k_1} = 0$, $x_{j_2} = 2-\epsilon$, and $x_{k_2} = 3-2\epsilon$, where $0.01 > \epsilon > 0$, and the distance between a and b is $|x_a - x_b|$, as

depicted in (8).

x	0	–	1	–	$(2-\epsilon)$	–	$(3-2\epsilon)$
unit	j_1, k_1		i		j_2		k_2

(8)

The optimal 1-1-1 match $\overline{\mathcal{B}}$ is (i, j_1, k_1) with $\delta_{\overline{\mathcal{B}}} = |x_i - x_{j_1}| + |x_i - x_{k_1}| + |x_{j_1} - x_{k_1}| = 1 + 1 + 0 = 2$. The approximation algorithm would first pair (i, j_2) with a distance of $|x_{j_2} - x_i| = |2 - \epsilon - 1| = 1 - \epsilon < 1 = |x_{j_1} - x_i|$. Then, the approximation algorithm would pair (i, j_2) with k_2 rather than with k_1 because $|x_i - x_{k_2}| + |x_{j_2} - x_{k_2}| = (2 - 2\epsilon) + (1 - \epsilon) = 3 - 3\epsilon < |x_i - x_{k_1}| + |x_{j_2} - x_{k_1}| = 1 + 2 - \epsilon = 3 - \epsilon$. So the approximation algorithm would yield $\mathcal{B}$ consisting (i, j_2, k_2) with distance $\delta_{\mathcal{B}} = |x_{j_2} - x_i| + |x_{k_2} - x_i| + |x_{j_2} - x_{k_2}| = |1 - \epsilon| + |2 - 2\epsilon| + |1 - \epsilon| = 4 - 4\epsilon$. Because $\epsilon > 0$ can be made arbitrarily small, the best bound is $\delta_{\mathcal{B}} = 2\,\delta_{\overline{\mathcal{B}}}$. Problems with $\delta_{\mathcal{B}} = 2\,\delta_{\overline{\mathcal{B}}}$ of any size can be constructed by replicating this example with x's that are spaced apart by, say, 10 units for each replicate, say at 10, 11, $12 - \epsilon$ and $13 - 2\epsilon$ for the second replicate.

4 Covariate distances that satisfy the needed triangle inequality

Statistical matching has often used covariate distances, but is typically indifferent about whether those distances satisfy the triangle inequality. Recall that a matching distance array $\boldsymbol{\delta}$ only requires nonnegative, possibly infinite numbers such that $\delta''_{jk} \leq \delta_{ij} + \delta'_{ik}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, with some convenient consequences when δ''_{jk} is defined differently from δ_{ij} and δ'_{ik}. In this section, we briefly discuss some options in defining $\boldsymbol{\delta}$.

If $\boldsymbol{\delta}$ and $\widetilde{\boldsymbol{\delta}}$ are two matching distance arrays, then so is $w\boldsymbol{\delta} + \widetilde{w}\widetilde{\boldsymbol{\delta}}$ for any nonnegative real numbers $w \geq 0$, $\widetilde{w} \geq 0$. For instance, a positively weighted combination of two or more Mahalanobis distances involving different, perhaps overlapping, sets of covariates, yields a

new distance array. This permits some covariates to receive greater emphasis, others to receive less. If a first match using distances $\boldsymbol{\delta}$ leaves an unsatisfactory imbalance for a few covariates, then use of a second distance $w\boldsymbol{\delta} + \widetilde{w}\widetilde{\boldsymbol{\delta}}$ may fix the problem if $\widetilde{\boldsymbol{\delta}}$ emphasizes the problematic covariates.

With three groups, we may estimate a two-dimensional propensity score; see Imai and van Dyk (2004). A matching distance array is obtained as a positively weighted combination of a Mahalanobis distance for the two dimensional propensity score and a Mahalanobis distance for all or for a subset of covariates used in defining the scores.

A robust, rank-based variant of the Mahalanobis distance is often used to limit the influence of outliers and of rare binary covariates; see Rosenbaum (2010, Chapter 8). This rank-based distance is not a metric on the Euclidean space of covariates, but it produces nonnegative numbers that satisfy $\delta^{''}_{jk} \leq \delta_{ij} + \delta^{'}_{ik}$, so it yields a matching distance array.

In (1), the distance is a total, but when $\rho \geq 2$ or $\kappa \geq 2$, this total emphasizes the groups who will have more controls in the match. It is straightforward to alter given distances so that groups, rather than individuals within groups, receive equal emphasis. For any matching distance array $\boldsymbol{\delta}$, define $\widetilde{\delta}_{ij} = \delta_{ij}/\rho$, $\widetilde{\delta}^{'}_{ik} = \delta^{'}_{ik}/\kappa$ and $\widetilde{\delta}^{''}_{jk} = \delta^{''}_{jk}/(\rho\kappa)$. Because $\boldsymbol{\delta}$ is a distance array satisfying $\delta^{''}_{jk} \leq \delta_{ij} + \delta^{'}_{ik}$ and $\rho \geq 1$, $\kappa \geq 1$, it follows that $\widetilde{\boldsymbol{\delta}}$ is a matching distance array satisfying $\widetilde{\delta}^{''}_{jk} \leq \widetilde{\delta}_{ij} + \widetilde{\delta}^{'}_{ik}$. Computing (1) with $\widetilde{\boldsymbol{\delta}}$ in place of $\boldsymbol{\delta}$ yields

$$\widetilde{\delta}_B = \sum_{\ell=1}^{\rho} \widetilde{\delta}_{i,j_\ell} + \sum_{m=1}^{\kappa} \widetilde{\delta}^{'}_{i,k_m} + \sum_{\ell=1}^{\rho}\sum_{m=1}^{\kappa} \widetilde{\delta}^{''}_{j_\ell,k_m} = \frac{1}{\rho}\sum_{\ell=1}^{\rho} \delta_{i,j_\ell} + \frac{1}{\kappa}\sum_{m=1}^{\kappa} \delta^{'}_{i,k_m} + \frac{1}{\rho\kappa}\sum_{\ell=1}^{\rho}\sum_{m=1}^{\kappa} \delta^{''}_{j_\ell,k_m},$$

so the revised distance is an average rather than a total. The distinction between $\boldsymbol{\delta}$ and $\widetilde{\boldsymbol{\delta}}$ will matter when ρ is large, because Step 2 of the approximation algorithm has one distance from $\mathcal{I}$ and ρ distances from $\mathcal{J}$ in each block distance to $k \in \mathcal{K}$. When ρ is large, $\widetilde{\boldsymbol{\delta}}$ or some compromise between $\boldsymbol{\delta}$ and $\widetilde{\boldsymbol{\delta}}$ may be more appropriate than $\boldsymbol{\delta}$ alone as a distance

array.

Often, we wish to match exactly for a covariate, say belted or unbelted driver in Table 1. Because of the near-fine balance constraints in §3.2, matching exactly for one covariate with $E \geq 2$ categories and balancing another covariate with $C \geq 2$ categories is not the same as splitting the matching problem into E separate matching problems, say matching belted drivers, and separately matching unbelted drivers, because splitting attempts the more difficult task of balancing the $E \times C$ joint categories. When either E or C or both are large, as is often true, near-fine balance of $E \times C$ joint categories often entails tolerating larger deviations from fine balance than balancing C categories. When feasible, how can exact matching for E categories of one nominal variable be combined with near-fine balance for C categories of another?

To implement exact matching with near-fine balance constraints, start with any matching distance array $\boldsymbol{\delta}$, so $\delta''_{jk} \leq \delta_{ij} + \delta'_{ik}$. Define $\widetilde{\delta}_{ij} = \infty$ if i and j differ in their exact match category, otherwise $\widetilde{\delta}_{ij} = \delta_{ij}$, similarly define $\widetilde{\delta}'_{ik} = \infty$ if i and k differ in exact match category, otherwise $\widetilde{\delta}'_{ik} = \delta'_{ik}$, but define $\widetilde{\delta}''_{jk} = \delta''_{jk}$. Trivially, $\widetilde{\delta}''_{jk} \leq \widetilde{\delta}_{ij} + \widetilde{\delta}'_{ik}$, so $\widetilde{\boldsymbol{\delta}}$ is also a matching distance array. Also trivially, if i and j are in the same exact category, and if i and k in the same exact category, then j and k are in the same exact category. It follows that a near-fine balance (ρ, κ)-design $\mathcal{B}$ is exactly matched for the E-category variable if and only if $\widetilde{\delta}_{\mathcal{B}} < \infty$. Let $\overline{\mathcal{B}}$ be a near fine (ρ, κ)-design whose distance $\widetilde{\delta}_{\overline{\mathcal{B}}}$ is minimal. If $\widetilde{\delta}_{\overline{\mathcal{B}}} = \infty$, then there is no near-fine (ρ, κ)-design $\mathcal{B}$ that is exactly matched for the E-category variable. If $\widetilde{\delta}_{\overline{\mathcal{B}}} < \infty$ for the optimal design, then Proposition 1 implies that the near-fine (ρ, κ)-design $\mathcal{B}$ produced by the approximation algorithm has finite total distance, $\widetilde{\delta}_{\mathcal{B}} < \infty$, so that design is also exactly matched for the E category variable. Additionally, if $\widetilde{\delta}_{\mathcal{B}} < \infty$, then the bound, $\widetilde{\delta}_{\overline{\mathcal{B}}} \leq \widetilde{\delta}_{\mathcal{B}} < \{1 + \max(\rho, \kappa)\}\, \widetilde{\delta}_{\overline{\mathcal{B}}}$ in Proposition 1 has avoided the infinite distances, and is referring to entries from the original matching

distance array, $\boldsymbol{\delta}$, for a match that is constrained to be both exact for the E category variable and near-finely balanced for the C category variable. In practice, distances that mismatch the E exact categories are increased not to ∞, but rather increased by a large finite penalty, so infeasibility of near-fine balance joint with exact matching is recognized when such a penalized distance appears in the match; see the discussion of "almost exact" matching in Rosenbaum (2010, §9.2).

5 Distinguishing effects of side airbags and unmeasured biases

In every nonrandomized or observational study of treatment effects, observed associations may reflect effects caused by the treatment under study, or biases in who was treated, or a combination of the two. A strength of the design in §2 is that it provides several comparisons relevant to the effects of side airbags. These several comparisons may concur, strengthening evidence that associations are effects caused by side airbags, or they may clash suggesting that some or all associations are not effects of side airbags. A design that always encourages causal conclusions, a design that cannot suggest caution and restraint, is not a good design. An easy way to publish false causal conclusions is to decline to look for evidence that might reveal bias if present.

A study contains two or more evidence factors if it provides two or more tests of the null hypothesis of no treatment effect that would be (essentially) independent were there no effect. A formal discussion of evidence factors involves technical issues that we do not present here; see Rosenbaum (2011; 2015, §3; 2017a; 2017b, §7). Recall that only 18% of studied vehicles in the optional period had side airbags.

Expressed more formally, our comparison contains several evidence factors, strict controls, and potential instruments. The "none" era is a control for the treated "all" era, because make and model determined the presence or absence of a side airbag in those eras,

but those two eras are separated by several years. This study has two evidence factors: (i) all-versus-other-eras and (ii) optional-versus-none. The availability of side airbags during the optional era can serve as an instrument for the purchase of a side airbag when the optional era is compared to the other eras. Additionally, there is another comparison within the optional era, comparing cars with or without side airbags; however, these are typically comparisons between, rather than within, matched sets. We do not present these analyses in detail, because two graphs tell the whole story.

Figure 3 shows the injury severity sustained by the driver, broken down by era: "none", "optional" or "all". There is a substantial and statistically significant reduction in injury severity between "none" and "all", but this is not plausibly an effect caused by side airbags, because the entire decline is already present in the "optional" era, when only 18% of owners had purchased vehicles with side airbags. Figure 4 is confined to the optional period, ignoring the matching, comparing vehicles with and without side airbags when owners had a choice about buying them. In Figure 4, the distribution of injuries looks similar with and without side airbags. In brief, the three-group design shows that some associations between side airbags and injuries are not plausibly effects caused by side airbags. With two rather than three groups, the same associations for two groups might mistakenly have been taken to be effects caused by side airbags.

6 Discussion

6.1 Summary

Observational studies often attempt to examine, possibly strengthen, a causal inference by making two comparisons instead of one treatment-versus-control comparison. The two comparisons may use two control groups, or two evidence factors, or combine a treated-control comparison with a instrument. Optimal construction of the relevant designs is

essentially impossible; see Crama and Spieksma (1992, Theorem 1). We have proposed a polynomial time approximation algorithm that produces a near-fine (ρ, κ)-design whose distance is not much greater than the unattainable optimal design.

6.2 Analyses with evidence factors or multiple control groups

The algorithm in Proposition 1 can be used to create two evidence factors, as in the example, or two control groups. The appropriate analysis depends upon the nature of the groups.

Two evidence factors provide two essentially independent tests of one null hypothesis of no treatment effect. These two independent tests and sensitivity analyses may be combined in a single analysis that combines independent P-values, such as the truncated product of Zaykin et al. (2002). The case $\rho = 1$, $\kappa \geq 1$, is discussed in detail in Rosenbaum (2011), and it is illustrated in the `sensitivitymv` package in `R`; see Rosenbaum (2015b). The general case $\rho \geq 1$, $\kappa \geq 1$, involves a stratified test (Rosenbaum 2018), rather than a matched test with multiple controls, and it may be implemented using the `senstrat` package in `R`; see also Rosenbaum (2017a) for elaboration.

Two control groups entail dependent tests. A simple strategy controls the family-wise error rate by testing several hypotheses in order, quitting when a null hypothesis is accepted. First, the treated group is compared to a combined control group, then to each control group separately, and finally comparisons attempt to show that the treated group differs more from both control groups than the control groups differ from each other; see Rosenbaum (2008) for specifics.

6.3 More than three groups

In most observational studies, finding two enlightening evidence factors or control groups is already a challenge, so we have focused on this situation, with a treated group and two comparison groups, or three groups in total. The situation with four or more groups is essentially parallel, and is implemented in the `approxmatch` package in `R`; see Crama and Spieksma (1992) for discussion of the simple case of groups of equal size without fine balance constraints. Step 2 of the algorithm is applied again, now matching blocks of the three group design to individuals from the fourth group, and so on. The approximation bound in Proposition 1 becomes worse because the triangle inequality is used again to bound some distances that were not optimized.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996), "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association*, 91, 444-455.

Baiocchi, M., Small, D. S., Lorch, S. and Rosenbaum, P. R. (2010), "Building a stronger instrument in an observational study of perinatal care for premature infants," *Journal of the American Statistical Association*, 105, 1285-1296.

Baiocchi, M., Cheng, J. and Small, D. S. (2014), "Instrumental variable methods for causal inference," *Statistics in Medicine*, 33, 2297-2340.

Beck, C., Lu, B. and Greevy, R. (2016), "nbpMatching: Functions for optimal non-bipartite matching," *R package version 1.5.1*, `https://CRAN.R-project.org/package=nbpMatching`.

Crama, Y. and Spieksma, F. C. R. (1992), "Approximation algorithms for three-dimensional assignment problems with triangle inequalities," *European Journal of Operational Research*, 60, 273-279.

Derigs, U. (1988), "Solving non-bipartite matching problems via shortest path techniques," *Annals of Operations Research*, 13, 225-261.

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

Hansen, B. B. (2007). "Flexible, optimal matching for observational studies," *R News*, 7, 18-24.

Hansen, B. B. and Klopfer, S. O. (2006), "Optimal full matching and related designs via network flows," *Journal of Computational and Graphical Statistics*, 15, 609–627. (`R` package `optmatch`)

Imai, K. and van Dyk, D. A. (2004), "Causal inference with general treatment regimes: generalizing the propensity score," *Journal of the American Statistical Association*, 99, 854-866.

Karmakar, B. (2017), "approxmatch: Approximately optimal fine balance matching with multiple groups," *R package version 1.0*, `https://CRAN.R-project.org/package=approxmatch`.

Keele, L. and Morgan, J. W. (2016). "How strong is strong enough? Strengthening instruments through matching and weak instrument tests," *The Annals of Applied Statistics*, 10, 1086-1106.

Keele, L., Titiunik, R. and Zubizarreta, J. R. (2015), "Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178, 223-239.

Korte, B. and Vygen, J. (2012), *Combinatorial Optimization* (5^{th} edition), New York: Springer.

Lorch, S. A., Baiocchi, M., Ahlberg, C. E. and Small, D. S. (2012), "The differential impact of delivery hospital on the outcomes of premature infants," *Pediatrics*, 130, 1-9.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011), "Optimal nonbipartite matching and its

statistical applications," *American Statistician*, 65, 21-30. (`R` package `nbpmatching`)

Papadimitriou, C. H. and Steiglitz, K. (1982), *Combinatorial Optimization: Algorithms and Complexity*, New York: Dover.

Pierskalla, W. P. (1968), "The multidimensional assignment problem," *Operations Research*, 16, 422-431.

Pimentel, S.D. (2016), "Large, sparse optimal matching with R package rcbalance," *R package version 1.8.5*, `https://CRAN.R-project.org/package=rcbalance`.

Pimentel, S.D. (2017), "rcbalance: Large, sparse optimal matching with refined covariate balance," *Observational Studies*, 2, 4-23.

Pimentel, S. D., Yoon, F. and Keele, L. (2015). "Variable-ratio matching with fine balance in a study of the Peer Health Exchange, " *Statistics in Medicine*, 34, 4070-4082.

R Core Team, (2018), "R: A language and environment for statistical computing," Vienna: R Foundation for Statistical Computing, `http://www.R-project.org`.

Rosenbaum, P.R. (1989), "Optimal matching in observational studies," *Journal of the American Statistical Association*, 84, 1024-1032.

Rosenbaum, P. R. (2005), "Attributable effects in case^2-studies," *Biometrics*, 61, 246-253.

Rosenbaum, P. R. (2008), "Testing hypotheses in order," *Biometrika*, 95, 248-252.

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

Rosenbaum, P. R. (2011), "Some approximate evidence factors in observational studies," *Journal of the American Statistical Association*, 106, 285-295.

Rosenbaum, P. R. (2015a), "How to see more in observational studies: Some new quasi-experimental devices," *Annual Review of Statistics and its Applications,* 2, 21-48.

Rosenbaum, P. R. (2015b), "Two R packages for sensitivity analysis in observational studies," *Observational Studies*, 1, 1-17.

Rosenbaum, P. R. (2017a), "The general structure of evidence factors in observational

studies," *Statistical Science,* 32, 514-530.

Rosenbaum, P. R. (2017b), *Observation and Experiment: An Introduction to Causal Inference*, Cambridge, MA: Harvard University Press.

Rosenbaum, P. R. (2018), "Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels," *Annals of Applied Statistics*, to appear. (`R` package `senstrat`)

Rosenbaum, P. R., Ross, R. N. and Silber, J. H. (2007). "Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer," *Journal of the American Statistical Association*, 102, 75-83.

Silber, J.H., Rosenbaum, P.R., McHugh, M.D., Ludwig, J.M., Smith, H.L., Niknam, B.A., Even-Shoshan, O., Fleisher, L.A., Kelz, R.R. and Aiken, L.H. (2016), "Comparison of the value of nursing work environments in hospitals across different levels of patient risk," *JAMA Surgery*, 151(6), 527-536.

Stuart, E. A. (2010), "Matching methods for causal inference," *Statistical Science*, 25: 1-21.

US National Highway Traffic Safety Administration, *US Fatality Analysis Reporting System*, https://www.nhtsa.gov/research- data/fatality- analysis- reporting- system- fars.

Vazirani, V. V. (2010), *Approximation Algorithms*, New York: Springer.

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, New York: Springer.

Williamson, D. P. and Shmoys, D. B. (2011), *The Design of Approximation Algorithms*, New York: Cambridge University Press.

Yang, D., Small, D. S., Silber, J. H., and Rosenbaum, P. R. (2012), "Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes," *Biometrics*, 68, 628-636. (Archived `R` package `finebalance`)

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., & Weir, B. S. (2002), "Truncated product method for combining P-values. *Genetic Epidemiology*, 22(2), 170-185. (`truncatedP`

function in `R` package `sensitivitymv`)

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

Zubizarreta, J. R. and Kilcioglu, C. (2016), "designmatch: Construction of optimally matched samples for randomized experiments and observational studies that are balanced and representative by design," *R package version 0.2.0*, `https://CRAN.R-project.org/package=des`

Table 1: Balance on covariates in 1398 matched 1-1-1 matched triples and 978 matched 1-3-1 sets. Cars were also matched for make and model. Group "none" refers to an era when side airbags were not available for this make/model, "optional" to an era when side airbags were optional for this make/model, and "all" to an era when all cars of this make/model had side airbags. For driver's age, crash year, and model year, values are means; otherwise, they are percentages. A mean is computed within a matched set, then averaged over sets.

	Driver			Direction of Impact				Roll-	Fire	Year	
Group	Age	Female	Belted	Left	Right	Front	Rear	over		Crash	Model
None	40	35	87	14	13	65	7	4	1	2006	1996
Optional	40	37	88	14	13	66	6	4	1	2010	2004
All	41	37	87	14	12	65	7	4	1	2012	2010

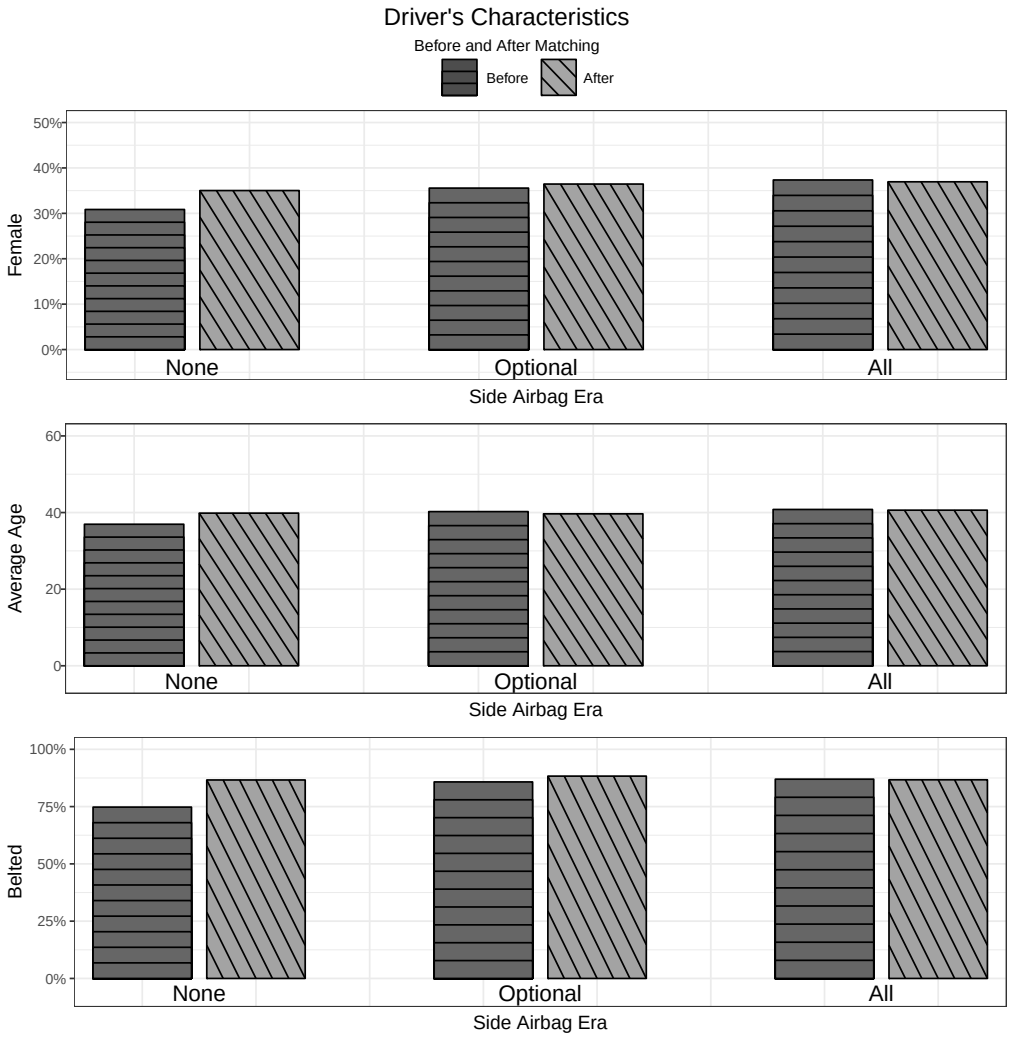

Figure 1: Driver's characteristics before matching and balance of the characteristics after matching across the three eras, None, Optional, and All. Bars of similar height after matching indicate that matching has balanced the covariate.

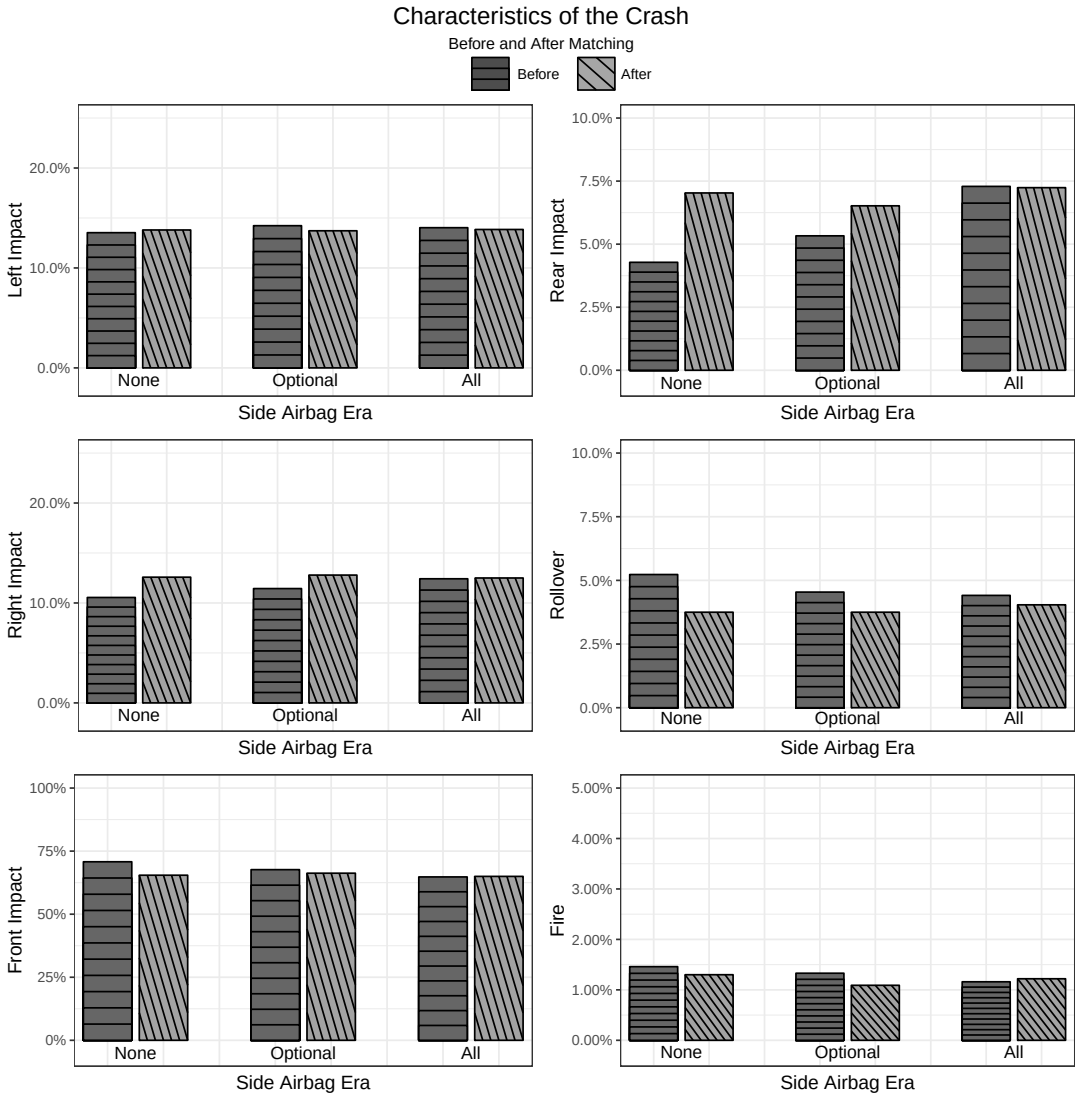

Figure 2: Characteristics of the crash before matching and balance of the characteristics after matching across the three eras, None, Optional, and All. Bars of similar height after matching indicate that matching has balanced the covariate.

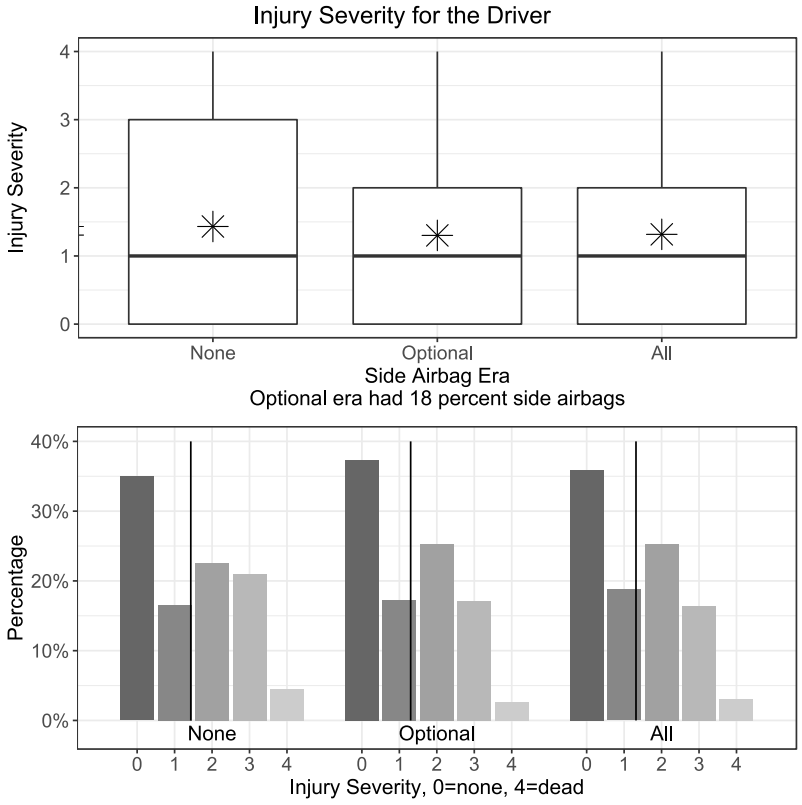

Figure 3: Injury severity in matched crashes grouped by the three eras, None, Optional, and All. The star in the boxplot and the vertical line in the barplot represent the mean.

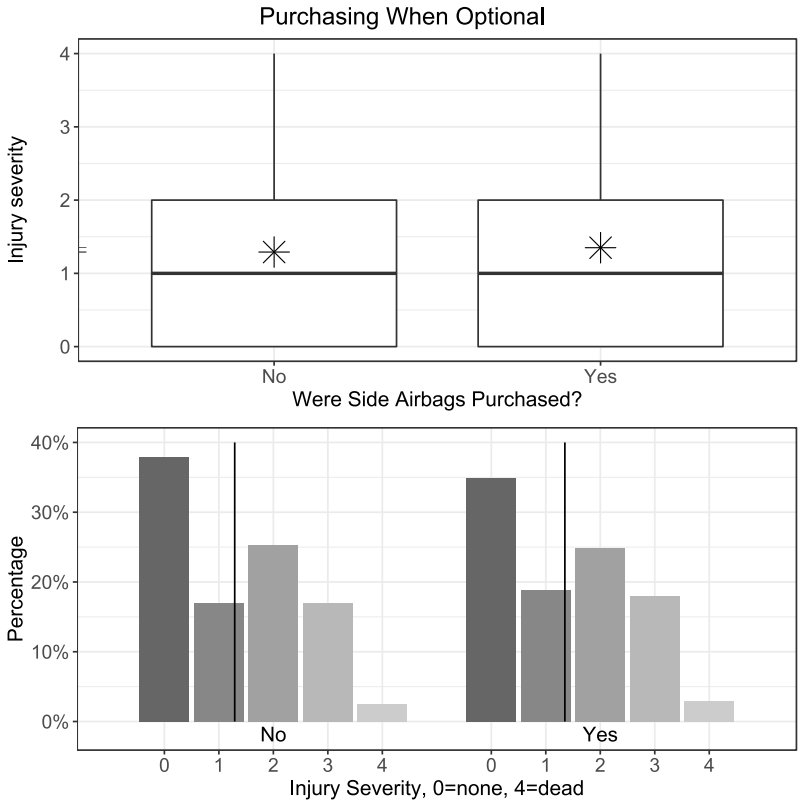

Figure 4: Unmatched crashes in the optional era grouped by whether side airbags were present. The star in the boxplot and the vertical line in the barplot represent the mean. This interesting but unmatched comparison is not an evidence factor.