Statistical Science
2020, Vol. 35, No. 3, 371–374
https://doi.org/10.1214/20-STS790
Main article: https://doi.org/10.1214/19-STS699

Rejoinder: Matching Methods for Observational Studies Derived from Large Administrative Databases

Ruoqi Yu, Jeffrey H. Silber and Paul R. Rosenbaum

1. OUTLINE

We thank the discussants for their insightful and generous comments. We organized our reply around a few themes, rather than responding to issues one by one. In Section 2, we recap the major elements of the paper in light of the discussion. Then Section 3 reviews the several goals of matching. Finally, Section 4 discusses open questions.

2. RECAP

First, let us restate the main themes of the paper.

- *Network optimization*. In our paper, each matched sample is obtained by optimizing a criterion subject to constraints. Specifically, each match is obtained as a minimum cost flow in a network, a rich but special family of integer programs that can be solved in polynomial time; see Bertsekas (1998) and Korte and Vygen (2012). There are other approaches to matching that leave the world of polynomial-time optimization algorithms, and these have both advantages and disadvantages (Zubizarreta, 2012; Karmakar, Small and Rosenbaum, 2019), but they are not discussed in our paper.
- *The constraints do most of the covariate balancing*. It is not possible to closely pair individuals for many covariates. It is possible to form treated and control groups with similar distributions of many covariates; that is, it is possible to balance many low-dimensional summaries of high-dimensional covariates. Balancing of covariates is largely achieved by the constraints, not by minimizing the within pair covariate distance. The balancing constraints include: (i) calipers on the rank of the scalar propensity score, (ii) near-fine balance constraints for a nominal covariate, perhaps with thousands of levels, (iii) possibly other balance constraints (Zubizarreta, 2012; Yu and Rosenbaum, 2019). If the constraints do most of the work, then finding the constraints that achieve your objectives is a central aspect of matching. In contrast, covariate distances used for pairing should focus on a few key covariates highly predictive of the outcome (Rosenbaum, 2005; Zubizarreta, Paredes and Rosenbaum, 2014).
- *Optimization is not recommendation*. As our example illustrates, the standard practice is to build several optimal matched samples, then pick the best one. There is no contradiction here: an optimal match is the solution to an optimization problem, not a recommended match. The practical goal is a match that is good in several senses, not best in one overriding sense, so each optimal solution is merely an approximation to the practical goal. Optimization is an aid to judgement, not a substitute for judgement. It is possible to produce the set of Pareto optima for a multi-objective optimization problem as a potentially useful guide (Pimentel and Kelz, 2020; Rosenbaum, 2012), but ultimately the investigators must pick one match, so the basic structure is unchanged: practical judgement is used to pick the most satisfactory of several optimally matched samples. Several optimal solutions provide points on a map by which judgement can steer among multiple objectives. Matching is part of the design of the study, completed prior to the examination of outcomes (Rubin, 2007).
- *Guarantee feasibility; guarantee speed*. There is no point in trying to solve an optimization problem subject to constraints if no solution satisfies the constraints. A fast implementation of Glover's (1967) algorithm permits certain types of constraints to be checked for feasibility at negligible cost. These include combinations of: (i) exact match constraints for a nominal covariate, perhaps with many levels, (ii) a caliper on the rank of the propensity score, (iii) a near-neighbor count on the propensity score. A threshold algorithm—a binary search—rapidly finds the tightest feasible constraint with negligible error, thereby guiding optimal

Ruoqi Yu is PhD student, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: ruoqiyu@wharton.upenn.edu). Jeffrey H. Silber is Professor, Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: silber@email.chop.edu; URL: https://cor.research.chop.edu/). Paul R. Rosenbaum is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, USA (e-mail: rosenbaum@wharton.upenn.edu).

matching, avoiding infeasibility. The `bigmatch` package finds the tightest feasible caliper and then pauses to ask you: Now that you know the tightest feasible caliper, what caliper would you like to use? A smaller caliper is infeasible, but a larger caliper might better balance several objectives. The investigator controls and knows the degree of sparsity of the network before lengthy computations begin. Stated differently, the investigator knows and controls ν, so that Proposition 6 provides the relevant guide to the situation at hand.

- *Do not compute all $T \times C$ distances*. Recall that there are T treated units τ, C potential controls γ, a covariate distance $\delta(\tau,\gamma)$ between each treated unit τ and each potential control γ, where $T \leq C$ so that $T \times C = O(C^2)$; commonly, $T \propto C$. Threshold algorithms, of the type proposed by Garfinkel (1971), have been used before in matching to identify or impose a tightest feasible constraint on a minimum distance match (Rosenbaum, 2017); however, previous implementations required the computation of all $T \times C$ distances, $\delta(\tau,\gamma)$, so these algorithms were computationally infeasible in large matching problems. Although $T \times C = O(C^2)$ appears small compared to the theoretical computational time bound $O(C^2 \log C)$ for optimizing a sparse network, the bound is a theoretical worst case rarely encountered in practice, while computing all $T \times C$ distances, $\delta(\tau,\gamma)$, actually takes $T \times C$ steps. Although this is not a theorem, our practical experience is that the computer spends most of its time calculating distances, $\delta(\tau,\gamma)$, and creating the network, and it spends much less time on optimization. If it is a mistake to pair a treated unit and a control with very different values of the propensity score, then why compute a Mahalanobis distance, $\delta(\tau,\gamma)$, between two such units? The `bigmatch` package avoids the computation of most distances, $\delta(\tau,\gamma)$, because the bipartite graph has been made sparser before any distances are computed. In the example, $T = 38{,}841$, $C = 159{,}527$, so there are $T \times C = 6.2 \times 10^9$ distances, but most of these were never computed.
- *The big gains come from near-fine balance*. In Table 2 of the article, the 463 Principal Procedure categories, the 973 Principal Diagnosis categories and their $473 \times 973 = 450{,}499$ interaction categories are better balanced in our final matched sample than they would have been in the most balanced of 10,000 randomized experiments built from the same data. This result is produced in large part by the near-fine balance constraint. The big gains from solving one sparse optimization problem, rather than many smaller dense optimization problems, come from the possibility of using near-fine balance on a large scale. A small but important innovation in the `bigmatch` package is the implementation of near-fine balance as a soft constraint, ensuring that the feasibility guaranteed by Glover's algorithm is not lost when the additional near-fine balance constraint is imposed.

3. MATCHING HAS SEVERAL OBJECTIVES

The term "convincing evidence" has two meanings, a normative meaning and a descriptive meaning, that can be in conflict. At times, evidence that should convince an audience does not do so. At other times, evidence that does convince an audience should not do so. In a successful scientific study, the normative and descriptive meanings do not diverge. Evidence is convincing precisely to the extent that it ought to be convincing, and the reasons it ought to convince coincide with the reasons it does convince. Achieving this is the first, the most important objective of matching.

The audience for a matched observational study is rarely an audience of statisticians. Surgeons are the audience for a study of surgery. Oncologists are the audience for a study of oncology. Health policy makers are the audience for a study of health policy. Each such audience has technical expertise that statisticians lack. A matched study is intended to be open to view by relevant experts, open to knowledgeable critical assessment. A matched study straightforwardly shows who was compared to whom, in what senses they were comparable and in what other senses they might differ. A matched study straightforwardly compares outcomes in treated and control groups whose degree of comparability is open to view. If a study has limitations, then these limitations are open to view by experts in the field. The audience does not need to understand a matching algorithm to critically appraise the resulting matched sample. Too often, statistical analysis based on elaborate models and methodology is little more than a report by the study's authors of an essentially private experience that they had with data. This does not happen with a matched study if properly designed and conducted.

A randomized clinical trial has a primary statistical analysis stated in the study's protocol. Matching sets up the basic comparison prior to examination of outcomes, thereby framing a primary analysis. A primary analysis does not preclude secondary and exploratory analyses; rather, a primary analysis distinguishes such analyses. Tukey (1980), p. 24, wrote: "I see no real alternative, in most truly confirmatory studies, to having a single main question—in which a question is specified by all of design, collection, monitoring and analysis." Secondary analyses are distinguished from exploratory analyses in the sense that: (i) secondary analyses are planned before examining outcomes, (ii) in some appropriate sense, secondary analyses control the frequency of errors when one or more analyses continue beyond the primary analysis.

The central problem in an observational study is that adjustments for measured covariates may fail to render comparable treated and control groups, because these groups may differ with respect to covariates that were not measured. Potential bias from unmeasured covariates is the central problem no matter how adjustments are made for measured covariates. Double or triple adjustment for measured covariates leaves the central problem untouched. However, it is known that if matching balances many covariates but pairs closely for a few key covariates predictive of the outcome, then insensitivity to unmeasured covariates is increased, that is, the design sensitivity is increased (Rosenbaum, 2005; Zubizarreta, Paredes and Rosenbaum, 2014). Typically, propensity scores and fine balance try to balance many covariates, while distances and exact matching try to ensure close pairing for a few key covariates.

At times, the magnitude of a treatment effect varies with the level of a measured covariate; that is, there is effect modification. When this occurs, conclusions may be insensitive to larger unmeasured biases at some levels of this covariate, because larger effects are typically insensitive to larger unmeasured biases. A principled, planned secondary analysis may demonstrate this in a sample matched for this covariate (Hsu et al., 2015; Lee, Small and Rosenbaum, 2018).

In their discussion, Stuart and Ackerman correctly emphasize the importance of the early planning of large observational studies. When the sample size is large, theory and experience suggest that it is often helpful to create a 10% planning sample that is used and discarded, together with a focused and planned 90% analysis sample (Heller, Rosenbaum and Small, 2009; Zhang et al., 2011). Often, the loss of a small part of a large sample is inconsequential when the central problems are biases that do not diminish with increasing sample sizes. Zhang et al. (2011) used a small sample of about 13,000 of their matched pairs for planning, leaving an analysis sample of about 120,000 pairs. The 13,000 pairs were useful in planning and were not missed in analysis.

4. DIRECTIONS

4.1 Full Matching and Related Techniques

As Fredrickson, Errickson and Hansen observe in their discussion, matching with a fixed number of controls, and even pair matching, are not always possible. The issue turns on Frank Yoon's entire number, defined to be $\text{ent}(\mathbf{x}) = \{1 - \Pr(Z = 1|\mathbf{x})\}/\Pr(Z = 1|\mathbf{x})$ where $\Pr(Z = 1|\mathbf{x})$ is the propensity score: in large samples, we expect to see $\text{ent}(\mathbf{x})$ controls available per treated individual at value $\mathbf{x}$ of the observed covariates. In concept in large samples, pair matching is feasible if $\text{ent}(\mathbf{x}) > 1$ for all $\mathbf{x}$, matching 2-to-1 is feasible if $\text{ent}(\mathbf{x}) > 2$, etc. In contrast, in concept in large samples, full matching is feasible whenever $\mathbf{x}$ has the same support in treated and control groups. Ways of combining variable-ratio matching with fine balance have only begun to be studied, but show promise; see Pimentel, Yoon and Keele (2015). What is a good way to combine full matching with fine balance in large problems without computing most distances, $\delta(\tau, \gamma)$?

An alternative approach when $\text{ent}(\mathbf{x}) < 1$ for some $\mathbf{x}$ is to stay with pair matching but redefine the study population using $\mathbf{x}$. A promising method is discussed by Fogarty et al. (2016).

4.2 Beyond Calipers and Counts of Nearest Neighbors

Calipers, exact matching and counts of nearest neighbors have been in the statistical literature for decades. However, a doubly convex bipartite graph can be produced in other ways than by sorting on exact match categories and a propensity score. Are there better ways? Traditional calipers are symmetric, but the biases that they are intended to remove are asymmetric. Suppose that the treated group tends to be older than the controls, prior to matching. A symmetric caliper might allow a control to be at most five years older or five years younger than a treated individual. Given the initial direction of the bias, within this symmetric caliper, controls will tend to be younger, on average. An asymmetric caliper might allow a control to be up to eight years older or up to two years younger than a treated individual. The length of the caliper is ten years in both cases, symmetric or asymmetric, but the asymmetric caliper is tolerant of mismatches that work against the bias. Yu and Rosenbaum (2019), Table 3, show that asymmetric calipers can remove much more bias if the degree of asymmetry is selected carefully. Glover's algorithm could be used to rapidly select both the length of the caliper for the rank of the propensity score and its degree of asymmetry.

More generally, directional penalties adjust distances to favor pairs that work against the direction of biases (Yu and Rosenbaum, 2019). Suppose that treated individuals, τs, are commonly older than controls, γs, prior to matching. The simplest directional penalty slightly increases the distance $\delta(\tau, \gamma)$ whenever τ is older than γ. Directional penalties do not affect the computational complexity of network optimization algorithms, nor do they require computation of $T \times C$ unpenalized distances, $\delta(\tau, \gamma)$, so directional penalties can be used in large matching problems. Directional penalties are closely related to Lagrangians used in integer programming (Yu and Rosenbaum, 2019, Section 2.5).

REFERENCES

BERTSEKAS, D. P. (1998). *Network Optimization*. Athena Scientific, Belmont, MA.

FOGARTY, C. B., MIKKELSEN, M. E., GAIESKI, D. F. and SMALL, D. S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *J. Amer. Statist. Assoc.* **111** 447–458. MR3538678 https://doi.org/10.1080/01621459.2015.1112802

GARFINKEL, R. S. (1971). An improved algorithm for the bottleneck assignment problem. *Oper. Res.* **19** 1747–1751.

GLOVER, F. (1967). Maximum matching in a convex bipartite graph. *Nav. Res. Logist. Q.* **14** 313–316.

HELLER, R., ROSENBAUM, P. R. and SMALL, D. S. (2009). Split samples and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **104** 1090–1101. MR2750238 https://doi.org/10.1198/jasa.2009.tm08338

HSU, J. Y., ZUBIZARRETA, J. R., SMALL, D. S. and ROSENBAUM, P. R. (2015). Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika* **102** 767–782. MR3431552 https://doi.org/10.1093/biomet/asv034

KARMAKAR, B., SMALL, D. S. and ROSENBAUM, P. R. (2019). Using approximation algorithms to build evidence factors and related designs for observational studies. *J. Comput. Graph. Statist.* **28** 698–709. MR4007751 https://doi.org/10.1080/10618600.2019.1584900

KORTE, B. and VYGEN, J. (2012). *Combinatorial Optimization: Theory and Algorithms*, 5th ed. *Algorithms and Combinatorics* **21**. Springer, Heidelberg. MR2850465 https://doi.org/10.1007/978-3-642-24488-9

LEE, K., SMALL, D. S. and ROSENBAUM, P. R. (2018). A powerful approach to the study of moderate effect modification in observational studies. *Biometrics* **74** 1161–1170. MR3908134

PIMENTEL, S. D. and KELZ, R. R. (2020). Optimal tradeoffs in matched designs comparing US-trained and internationally trained surgeons. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2020.1720693

PIMENTEL, S. D., YOON, F. and KEELE, L. (2015). Variable-ratio matching with fine balance in a study of the Peer Health Exchange. *Stat. Med.* **34** 4070–4082. MR3431322 https://doi.org/10.1002/sim.6593

ROSENBAUM, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist.* **59** 147–152. MR2133562 https://doi.org/10.1198/000313005X42831

ROSENBAUM, P. R. (2012). Optimal matching of an optimally chosen subset in observational studies. *J. Comput. Graph. Statist.* **21** 57–71. MR2913356 https://doi.org/10.1198/jcgs.2011.09219

ROSENBAUM, P. R. (2017). Imposing minimax and quantile constraints on optimal matching in observational studies. *J. Comput. Graph. Statist.* **26** 66–78. MR3610408 https://doi.org/10.1080/10618600.2016.1152971

RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat. Med.* **26** 20–36. MR2312697 https://doi.org/10.1002/sim.2739

TUKEY, J. W. (1980). We need both exploratory and confirmatory. *Amer. Statist.* **34** 23–25.

YU, R. and ROSENBAUM, P. R. (2019). Directional penalties for optimal matching in observational studies. *Biometrics* **75** 1380–1390. MR4041838 https://doi.org/10.1111/biom.13098

ZHANG, K., SMALL, D. S., LORCH, S., SRINIVAS, S. and ROSENBAUM, P. R. (2011). Using split samples and evidence factors in an observational study of neonatal outcomes. *J. Amer. Statist. Assoc.* **106** 511–524. MR2847966 https://doi.org/10.1198/jasa.2011.ap10604

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107** 1360–1371. MR3036400 https://doi.org/10.1080/01621459.2012.703874

ZUBIZARRETA, J. R., PAREDES, R. D. and ROSENBAUM, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann. Appl. Stat.* **8** 204–231. MR3191988 https://doi.org/10.1214/13-AOAS713