This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association*, 26 February 2020, available online: http://www.tandfonline.com/10.1080/01621459.2020.1720693.*

Optimal tradeoffs in matched designs comparing US-trained and internationally-trained surgeons

Samuel D. Pimentel and Rachel R. Kelz*

December 23, 2019

Abstract

Does receiving a medical education outside the United States impact a surgeon's performance? We study this question by matching operations performed by internationally-trained surgeons to those performed by US-trained surgeons in reanalysis of a large health outcomes study. An effective matched design must achieve several goals, including balancing covariate distributions marginally, ensuring units within individual pairs have similar values on key covariates, and using a sufficiently large sample from the raw data. Yet in our study, optimizing some of these goals forces less desirable results on others. We address such tradeoffs from a multi-objective optimization perspective by creating matched designs that are Pareto optimal with respect to two goals. We provide general tools for generating representative subsets of Pareto optimal solution sets and articulate how they can be used to improve decision-making in observational study design. In the motivating surgical outcomes study, formulating a multi-objective version of the problem helps us balance an important variable without sacrificing two other design goals, average closeness of matched pairs on a multivariate distance and size of the final matched sample. **Keywords:** matching; Pareto optimality; fine balance; network flow optimization; multi-objective optimization.

*Samuel D. Pimentel is Assistant Professor of Statistics, University of California, Berkeley, Berkeley, CA 94720 (spi@berkeley.edu) and Rachel R. Kelz is Professor of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 (rapaporr@uphs.upenn.edu). This work was partially supported by the Department of Defense under FA9550-11-C-0028, Army Research Office, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a. This work is derived in part from a limited data set supplied by the Florida Agency for Health Care Administration, but the Agency specifically disclaims responsibility for any analysis, interpretations, or conclusions that may be created as a result of the limited dataset. Although the AMA Masterfile data is the source of the raw physician data, the tables and tabulations were prepared by the authors and do not reflect the work of the AMA. The authors declare no conflicts. The authors also thank Kristina Simmons, Chris Wirtalla, and Luke Keele for substantial assistance with data management, and Paul Rosenbaum, Dylan Small, reading group participants at the University of Pennsylvania and UC Berkeley, and attendees of the 2016 Wharton-INSEAD Doctoral Student Consortium and the 2017 Joint Statistical Meetings for helpful comments.

1 Introduction: best choices when design goals conflict

1.1 Comparing US-trained and internationally-trained surgeons

Roughly 15% of the surgeons practicing in the United States received medical training outside the US. Does the variable content and quality of international medical training, compared to standardized US medical education, cause internationally-trained surgeons to perform worse in the operating room? A randomized trial in which operations are randomly assigned to be performed by either an international medical graduate (IMG) or a US medical graduate (USMG) could answer this, but would not likely be performed. Zaheer et al. (2017) instead compared the performance of IMGs to USMGs in an observational study, using a large database of electronic health records. The observational study, while feasible, faces challenges that the randomized trial avoids. IMGs tend to practice in different settings than USMGs, with different mixes of patients. The randomized trial ensures IMGs and USMGs see similar patient populations on average, but in the observational study the groups may differ on important covariates before treatment. Matching can often eliminate such differences on observed covariates by selecting new comparison groups from the raw data (Rosenbaum and Rubin 1985; Stuart 2010). In this study, operations conducted by IMGs, our treated units, can be matched to similar operations conducted by USMGs, which serve as controls. Matched samples are constructed using only covariates (including attributes of patients, surgeons, and procedures known prior to the surgery itself) and not outcomes, to ensure valid causal inference (Rubin 2007).

1.2 Design goals in matching surgical operations

A matching design in this study must accomplish several different goals to permit causal inference about the effect of surgical training. One important goal is achieving balance on pre-treatment covariates. Balance is achieved if the distribution of an observed covariate in the group of selected USMG operations after matching is made similar to the distribution of the same covariate in the group of IMG operations after matching, ignoring which operations are matched to each other. Balanced matches more closely resemble the randomized study, in which IMG and USMG operations

have identical covariate distributions on both observed and unobserved covariates.

Another goal is close or exact matching in pairs, which requires that the operation performed by the IMG and the operation performed by the USMG in each individual pair have similar or identical values on key covariates. Matching closely on important prognostic variables — those that are highly predictive of adverse outcomes — helps reduce within-pair heterogeneity in outcomes, which can make inference less sensitive to biases from unobserved confounding variables (Rosenbaum 2005). Exact matching is also an important step in detecting effect modification (Hsu et al. 2015), and closely-matched pairs can sometimes be used for in-depth qualitative analyses that help improve the larger study (Rosenbaum and Silber 2001). Matching exactly on a variable in every case produces balance on this variable as a byproduct; however, exact matching on many covariates at once is not generally possible for more than a small fraction of the data. A multivariate distance, such as the Mahalanobis distance (Rubin 1980), may be used to summarize closeness on many variables between paired units.

A third important goal is maximizing the number of observations included in the final matched design. Comparing matched designs that adjust equally well for observed covariates in settings with a uniform treatment effect, tests based on the larger matched study will have greater power than those based on the smaller study. In addition, retaining the large majority of the IMG operations is important in our study for purposes of external validity. To the degree that the IMG operations in the initial data are an approximately representative sample of IMG operations in the United States as a whole, retaining the large majority of these IMG patients help ensure that the results we obtain generalize to the national level.

1.3 Surgical experience and conflict among design goals

Sometimes improvements to one design goal may come at a cost to another. Consider one important covariate in our study, surgical experience. It is natural to expect that surgeons who have been practicing for many years may conduct surgeries differently from those relatively new to the field and experience different rates of success. Hence to isolate effects of differences in surgeon training background, we wish for a matched design with balanced distributions of surgical experience be-

tween the sample of IMG operations and the sample of USMG operations. This variable is highly imbalanced in the raw data — IMGs tend to have more experience than USMGs — so it must be emphasized when constructing matched pairs.

Unfortunately, placing greater emphasis on balance for surgical experience generally involves accepting diminished success in other design goals. For instance, consider eliminating the matched pair with the largest imbalance in surgical experience from our matched study. The study's overall imbalance on experience improves, but this gain is achieved by reducing the size of the matched sample. For further discussion of such tradeoffs between match quality and sample size, see Resa and Zubizarreta (2016). As another example, the pairings of study subjects that best improve balance on surgical experience may not be the same as the pairings that best emphasize closeness on other important covariates. In all cases operations are matched exactly on hospital ID, so IMG operations can only be paired to USMG operations in the same hospital. Since the pool of operations available for matching in each hospital is small, a choice may be necessary between low-distance pairings and pairings that attain balance on experience.

A constructed example illustrating this situation is given in Figure 1. In hospital 1, an operation by an experienced IMG surgeon has 2 potential matches. One available USMG operation has a low pair distance Δ_1, but is performed by an inexperienced surgeon. The other pairing maintains balance on surgeon experience, but accepts a pair distance higher by ϵ_1. The situation is repeated in Hospital 2, with distances Δ_2 and $\Delta_2 + \epsilon_2$. In each hospital, a conflict arises between optimizing pair distances and optimizing balance on surgical experience.

The two matches shown in Figure 1 represent the choices of common matching algorithms. Multivariate distance matching, targeting the distances alone without reference to marginal balance, will choose the pairing with smaller distance in both hospitals (Rosenbaum and Rubin 1985), while matching with fine balance, which instead prioritizes balance strictly over pairwise distances, will select the experienced USMG operation in both hospitals (Rosenbaum et al. 2007). However, if ϵ_1 is very small and ϵ_2 is very large, the cost of prioritizing balance over pairwise closeness in Hospital 1 is negligible, while in Hospital 2 it is is very high. Here a match that chooses the second pairing in Hospital 1, and the first pairing in Hospital 2 is clearly superior to either of the matches in the

figure; it avoids paying the high cost ϵ_2, but is able to partly balance the surgical procedure variable at negligible cost. The most desirable design turns out to be not an extreme solution produced by strictly prioritizing one goal as in standard methods, but an intermediate solution that properly weighs the relative cost of the two goals. The example deals with a few individual patients. In the IMG-USMG example, many similar situations aggregate to cause a conflict at a large scale.

1.4 Dealing with tradeoffs in the IMG-USMG study

To apply standard matching methods in the IMG-USMG study, we would begin by defining an objective function and selecting a match that achieves the optimal value of that objective function. However, multiple design goals, each associated with its own distinct objective function, are present. If we somehow knew in advance which weighted combination of the different objective functions will produce the most globally attractive match, we could define that combination as our objective function. In the absence of such knowledge, it is instead necessary to compute and examine multiple matches that vary the relative importance of, or the tradeoff between, the different design goals, selecting a best option from this set. Note that selecting among many matches computed from the data does not introduce bias as long as outcome analysis is postponed until after the selection of a final match (Rubin 2007). For examples of selection among matched designs in the literature, see Hansen (2004) and Zubizarreta et al. (2013).

In a multi-objective setting, the goal of optimal matching is no longer to produce a single match that optimizes a single objective function, but to produce an "admissible" class of matches that vary the tradeoff between multiple goals across the full range of relative weightings between the different goals. A natural tool for defining this class of solutions is Pareto optimality. A solution to a multi-objective optimization problem is Pareto optimal if there exists no other solution with equally good performance on all objective functions and strictly better performance on at least one. Computing and examining the Pareto optimal set gives a full description of the tradeoff between the goals and allows an investigator to make an informed judgment about how to resolve it. For example, by looking at all matches Pareto optimal for balance on experience and exact matching on important covariates, we might recognize an opportunity to improve the rate of exact matching

by a large amount by accepting a very minor decrease in balance.

Given two quantitative design goals, such as those in our IMG-USMG study, we characterize the set of Pareto optimal solutions as solutions to a parametric family of single-objective combinatorial optimization problems. In general these single-objective problems belong to an NP-hard class and are not computationally tractable; however, we show that many of their solutions can be recovered via a different parametric family of problems for which solutions can be obtained in polynomial time. Using structural features of the Pareto optimal solution set, we describe how to create a global summary of the kinds of matches available in the IMG-USMG study based on a small subset of Pareto-optimal points, and we use this framework to select a match that strikes a harmonious balance between the different design goals in the IMG-USMG study.

The match produced by this method allows a fully-design-based comparison of IMG and USMG patient outcomes. In the original study, Zaheer et al. (2017) used matching to address imbalances in characteristics of patients and type of surgery performed, but did not match directly on surgical experience. This potentially important variable remained imbalanced after matching, and concerns about its impact were addressed by fitting generalized linear models to the data after matching. By using our new methods for tradeoff analysis in the design stage, we eliminate the need for a secondary model-based analysis. We replicate a portion of the original study and provide stronger confirmation of the original results, demonstrating equivalent performance, as measured by patient mortality, between IMGs and USMGs.

1.5 Outline

Section 2 introduces a formal optimization representation for matching problems, defines Pareto optimality, and describes two single-objective optimization problems that play a key role in the solution to the tradeoffs problem for the IMG-USMG study. Section 3 describes the design stage of the IMG-USMG study, including a strategy for efficiently exploring the family of Pareto optimal solutions and the use of this strategy to strike appropriate tradeoffs between imbalances in surgical experience, close pairings on other covariates, and sample size. Section 4 uses the resulting match to evaluate the relative performance of IMGs and USMGs, addressing the possibility of unmeasured

confounding via sensitivity analysis. Section 5 concludes and suggests other contexts in which this approach is useful.

2 Formal framework and tools for exploring tradeoffs

2.1 Matching as a network flow optimization problem

We view matching has as a special case of the minimum-cost network flow optimization problem, in which a graph consisting of nodes and directed edges is given. Each node may supply or demand units of "flow," and flow may be sent across directed edges, obeying capacity constraints associated with the edges and paying edge-specific costs per unit flow. The task is to route all flow from supply nodes to demand nodes, paying the minimum total cost over all possible routings; see for example Ford and Fulkerson (1962). A simple matching problem may be represented as a network flow optimization across a modified bipartite graph between the treated units and controls, and more complex matching problems with diverse design goals can also be encoded as network flow problems using expanded network representations. For example, network flow problems can incorporate calipers on propensity scores (Rosenbaum 1989), fine and near-fine balance constraints (Yang et al. 2012, online appendix), refined balance constraints (Pimentel et al. 2015), and full matching (Hansen and Klopfer 2006).

We will consider network flow problems in a generalized way, using $\mathcal{N}$ and $\mathcal{E}$ to represent the node and edge sets of the network, $\mathbf{x}$ to represent a vector of real-valued decision variables corresponding to the flows chosen across the different edges in $\mathcal{E}$, a linear function $f(\mathbf{x}) = \sum_{e \in \mathcal{E}} c_e x_e$ to represent the objective function implied by the costs in the network, and a set of decision vectors $\mathcal{F}$ to represent the collection of solutions allowed under the constraints that flow is conserved at each node and edge flow capacities are respected (Bertsekas 1998). We can then state the network flow problem as follows:

Problem 1. *Minimum-cost Network Flow Optimization*

$$\min \sum_{e \in \mathcal{E}} c_e x_e \qquad s.t. \qquad \mathbf{x} \in \mathcal{F}.$$

For a more detailed description of the minimum-cost network flow problem and its relation to matching, see Section A of the online supplement. Many polynomial-time algorithms for solving network flow problems exist, including the RELAX algorithm (Bertsekas et al. 1994) which is efficient in practice, implemented in FORTRAN, and available in R for academic use via the `optmatch` package (Hansen and Klopfer 2006).

A special feature of network flow problems is that when edge capacities and node supplies are integral, some optimal solution will always be integral and will be recovered by a certain class of solving algorithms, even without an explicit integer constraint on the decision variables (for details see Papadimitriou and Steiglitz 1982, sec. 13.2). Note, however, that when additional side constraints are imposed beyond constraint set $\mathcal{F}$ this guarantee no longer holds in general. For example, suppose that in the IMG-USMG study we wished to put an upper limit a on the number of IMG operations that could be matched to USMG operations involving a different surgical procedure. This constraint can be represented mathematically as a linear constraint requiring the sum of flow over certain edges (those pairing treated units to controls with a different surgical procedure) to be less than or equal to a. However, there is no known rule for constructing a network flow problem that incorporates this requirement into the standard constraint set $\mathcal{F}$. In order to produce relevant answers in matching problems in these settings, an explicit integer constraint on the decision variables must be added. Generally this makes the problem much more computationally difficult. In fact, the class of network flow problems with linear objective functions, general linear side constraints and 0-1 integer constraints on flow values is NP-complete, meaning that no polynomial-time solution algorithm or network-flow formulation is known to exist (Bertsekas 1998, §8.4).

2.2 Pareto optimal solutions for two quantitative goals

Suppose now that in a matching problem in the form of Problem 1, we have two linear objective functions of the optimization variables $\mathbf{x}$ instead of one:

$$f_1(\mathbf{x}) = \sum_{e \in \mathcal{E}} \psi_e x_e \qquad \text{and} \qquad f_2(\mathbf{x}) = \sum_{e \in \mathcal{E}} \gamma_e x_e \tag{1}$$

Denote the bivariate function $(f_1(\mathbf{x}), f_2(\mathbf{x}))$ by $\mathbf{f}(\mathbf{x})$. In the IMG-USMG example, we will use functions f_1 and f_2 at different points to represent measures of imbalance on surgical experience, close pairing on a multivariate covariate distance, and sample size. For examples of other design goals and their representations, see Section 5. We now define Pareto optimality in terms of $\mathbf{f}$; for similar definitions see Zeleny (1974) and Miettinen (1999).

Definition 1. *A solution* $\mathbf{x}$ *to Problem 1 is Pareto optimal for function* $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$ *if there exists no dominating* $\mathbf{x}'$*, i.e. a solution* $\mathbf{x}'$ *such that*

$$f_i(\mathbf{x}') \leq f_i(\mathbf{x}) \quad \textit{for all } i \textit{ and} \quad f_i(\mathbf{x}') < f_i(\mathbf{x}) \quad \textit{for some } i \in \{1, 2\}.$$

A Pareto optimal point $(f_1(\mathbf{x}), f_2(\mathbf{x})) \in \mathbb{R}^2$ is the image of a Pareto optimal solution $\mathbf{x}$ under $\mathbf{f}$. Let $\underline{\mathbf{f}}$ and $\overline{\mathbf{f}}$ be the Pareto optimal points associated with the smallest possible value of f_1 and smallest value of f_2, respectively, in the Pareto optimal set. These endpoints must exist since the original network flow problem has a finite feasible region, and by the definition of Pareto optimality they are also unique. Our goal is to explore the full set of Pareto optimal points and their associated solutions (that is, at least one solution per distinct point).

2.3 Two optimization problems for finding Pareto optimal points

To determine Pareto optimal points and their solutions, we rely on two optimization problems, each with a single objective function and a tuning parameter. The first problem minimizes f_1 under a linear constraint on f_2; we define it as follows for some nonnegative integer a.

Problem 2 (Directly Constrained Problem)**.**

$$\min f_1(\mathbf{x}) \qquad s.t. \qquad \mathbf{x} \in \mathcal{F}, \quad f_2(\mathbf{x}) \leq a, \quad \mathbf{x} \in \mathbb{Z}^{|\mathcal{E}|}$$

Call this problem $\mathcal{Q}(a)$ and represent its objective value at $\mathbf{x}$ by $Q(\mathbf{x}, a)$. The optimal objective value for Problem $\mathcal{Q}(a)$ is a decreasing step function over feasible values of a (since the set of all feasible solutions to the original network problem is finite). Optimality for this problem is closely linked to Pareto optimality for $\mathbf{f}$, as shown by the following result. All proofs are deferred to Section F of the online supplement.

Theorem 1. *Any Pareto optimal solution* $\mathbf{x}$ *for* $\mathbf{f}$ *must also be optimal for* $\mathcal{Q}\left(f_2(\mathbf{x})\right)$*, and all optimal solutions of* $\mathcal{Q}\left(f_2(\mathbf{x})\right)$ *are also Pareto optimal with identical values of* $\mathbf{f}$*.*

Theorem 1 tells us that if we solved $\mathcal{Q}(a)$ for all possible values of a, we would obtain the full set of Pareto optimal points and see the complete tradeoff relationship between our design goals. However, the directly constrained problem lacks a guaranteed polynomial-time solution algorithm and is not easily solvable in practice. As such we introduce the second problem, which incorporates f_2 into the objective function using a penalty parameter $\rho > 0$.

Problem 3 (Penalized Problem)**.**

$$\min f_1(\mathbf{x}) + \rho f_2(\mathbf{x}) \qquad s.t. \qquad \mathbf{x} \in \mathcal{F} \tag{2}$$

Represent Problem 3 by $\mathcal{P}(\rho)$. In addition, $P(\mathbf{x}, \rho)$ will represent the value of the objective function evaluated for particular values of $\mathbf{x}$ (assuming $\mathbf{x} \in \mathcal{F}$) and ρ. The optimal objective value for this problem is a nondecreasing piecewise linear function, since $P(\mathbf{x}, \rho)$ is linear for fixed $\mathbf{x}$ and the set of integer solutions in $\mathcal{F}$ is finite. The interpretation of parameter ρ is less straightforward than for parameter a in Problem $\mathcal{Q}$, which is a constraint on an interpretable objective function. However, Problem $\mathcal{P}(\rho)$ is a standard network flow problem — we have not introduced additional constraints, only changed the edge costs — so it is computationally feasible with an integral optimal solution. In addition, Theorem 2 shows that solving the penalized problem $\mathcal{P}(\rho)$ also produces an optimal solution to the directly constrained problem $\mathcal{Q}(a)$ for some a.

Theorem 2. *Suppose* $\mathbf{x}^*$ *is an integer-valued optimal solution for the penalized problem* $\mathcal{P}(\rho)$*. Then* $\mathbf{x}^*$ *is Pareto optimal as long as* $\rho > 0$*, and* $\mathbf{x}^*$ *is also optimal for the directly constrained problem* $\mathcal{Q}(a^*)$ *where*

$$a^* = f_2(\mathbf{x}^*)$$

Remark: Notice that this theorem only makes a claim in one direction; any solution to the penalized problem is also a solution to a directly constrained problem, but it may not be the case that all solutions to the directly constrained problem also solve a penalized problem. See Section C in the online supplement for further discussion. Also, the lack of guaranteed Pareto optimality

when $\rho = 0$ is not a major concern, since the most extreme Pareto points can always be recovered for some $\rho > 0$ (see Theorem C2 in the online supplement).

Theorem 2 shows that solving the penalized problem $\mathcal{P}(\rho)$ also produces an optimal solution to the directly constrained problem $\mathcal{Q}(a)$ for some a. By solving $\mathcal{P}(\rho)$ repeatedly for various values of ρ we may obtain a range of solutions to $\mathcal{Q}(a)$ for different a-values and observe the tradeoff between f_1 and f_2.

2.4 A small example

We now return to the illustrative example in Figure 1. In Table 1, we provide specific values for the parameters Δ_1, Δ_2, ϵ_1, and ϵ_2 defining the matching distances and solve Problems 2 and 3 repeatedly to identify a spectrum of possible matches that trade off our goals of matching closely within pairs (as measured by the within-pair distances) and of balancing surgical experience (IXP vs. XP) marginally. Here we define f_1 as the sum of pairwise covariate distances in the selected match, and we define f_2 as the total number of excess USMG operations across surgical experience categories (computed by taking the difference in the IMG and USMG counts for operations at each experience level and adding up the positive differences). Quantity f_2 is a rescaling of the total variation distance between the distribution of surgical experience in the two groups (Pimentel et al. 2015) and will be equal to 0 exactly when counts in both experience bins are equal across groups.

Representing this problem in network form produces a bipartite graph with directed edges from each treated unit to all controls in the same hospital, each of maximum capacity one and of cost given by the distance matrix. We define f_1 as the total cost incurred over these edges. Additional nodes and edges are added to the network for each category of surgical experience using a simple version of the network structure detailed in Pimentel et al. (2015). A specific edge carries the excess number of controls among matched operations by experienced surgeons, and another edge does the same for matched operations by inexperienced surgeons. f_2 is defined as the sum of flow across these two edges. To directly constrain f_2, we solve an integer program that limits the flow over the set of these two edges, and to penalize it instead, we assign a cost ρ to these edges.

Table 2 shows a tradeoff curve obtained by solving the penalized problem for different penalty

values ρ, and also solutions obtained by solving the directly constrained problem directly for several values of a (possible in this context since the problem is small). Here there exist exactly three distinct solutions, each of which can be recovered by either $\mathcal{Q}(a)$ or $\mathcal{P}(\rho)$ for the right parameter choice. Examining the tradeoff curve, we see that moving from Match 3 to Match 2 allows a huge decrease in overall covariate distance – much larger than the total covariate distance in Matches 1 or 2 – while accepting only a single additional mismatch on the binary covariate. In addition, Match 2 is much more desirable than Match 1, which accepts twice as large an imbalance on the marginal variable for an infinitesimal improvement in covariate distance. However, Match 2 cannot be produced by strictly prioritizing one of our goals over the other.

3 Matching US-trained and foreign-trained surgeons

3.1 Context and Design Goals

Zaheer et al. (2017) conducted a study comparing outcomes of general surgery patients treated by international medical graduates (IMGs) to those of patients treated by US medical graduates (USMGs). Their analysis used a dataset containing all general surgery operations listed in inpatient hospital discharge records in the states of New York and Florida during the period 2008-2011, a total of over 970,000 unique operations (New York State Department of Health 2008-2011; Florida Agency for Health Care Administration 2008-2011); the data was joined to an American Medical Association database of physician information to identify IMG and USMG surgeons (American Medical Association 2013). We re-analyze a subset of this data, operations performed in the Orlando metropolitan area (Orange, Lake, Osceola, Seminole, Volusia, Flagler, and Sumter counties), by matching IMG patients to USMG patients and comparing mortality rates in the matched sample. The selected dataset includes 20,024 operations received by IMG patients and 58,270 operations received by USMG patients. The operations differ systematically in several important pre-treatment attributes; for example, 59% of IMG operations were conducted on emergency room admissions, compared to only 51% of USMG operations. For a descriptive table containing information about the selected data, see Section B of the online supplement.

A first step in designing an effective matched design for this study is determining criteria under which matches will not be allowed at all. It is well established that variation across hospitals in practice and quality of care is closely related to the quality of surigcal outcomes, and since the proportion of operations performed by IMGs varies substantially across hospitals in our dataset, matches are only allowed between operations in the same hospital. In addition, matches are only allowed between operations whose estimated propensity scores, fitted using standard logistic regression on surgical procedure, coarsened Elixhauser comorbidity index (Elixhauser et al. 1998), patient age, emergency status, sex, race, and median income of patient's ZIP code, differed by less than 0.25 standard deviations. Homogeneity of propensity scores within pairs is especially important since we conduct inference by permuting treatment assignments within pairs, relying heavily on similarity of propensity for treatment within pairs (Rosenbaum 2002b, Ch. 3). We do not include hospital indicators in the propensity score fit; the best way to incorporate such a blocking variable when estimating a propensity score for caliper matching remains an interesting open question. Under these requirements, 94% of the IMG operations are able to be matched to USMG operations without replacement. Imposing these restrictions leads to both statistical and computational benefits, ensuring that the matching problem is sparse in the sense of Pimentel et al. (2015).

A key goal of the matched study was to achieve balance on surgical experience, measured in years of practice. In the raw data, IMG operations tend to be performed by surgeons with much more experience than USMG operations; for example, 25% of IMG operations were performed by surgeons with over 30 years of experience, while only 4% of USMG operations were. Unless the distribution of surgical experience can be made to look similar within the final matched samples, it will be difficult to determine whether any observed differences in mortality are actually due to differences in performance between USMGs and IMGs, or to differences in surgeons' experience levels. This goal is challenging, however, because of the extreme dissimilarity of the initial groups' experience distributions. To quantify imbalance in a way conducive to representation in a network flow problem, we create a categorical variable with a different category for each decile of surgical experience in the dataset as a whole and measure the total variation distance between the empirical

distributions of this covariate in the matched IMG and the matched USMG operations.

Several variables besides surgical experience are scientifically important: surgical procedure received, a coarsened measure of the Elixhauser comorbidity index (describing the patient's overall health at the time of surgery), an indicator of emergency room admission, and patient age and sex. Paired operations should be similar on these variables whenever possible, in order to reduce heterogeneity in matched pair differences and thus reduce sensitivity to unmeasured bias (Rosenbaum 2005). Accordingly, we compute a robust Mahalanobis distance (Rosenbaum 2010, §8.3) for each possible IMG-USMG pair, measuring the similarity of the operations with respect to these five variables, and seek to choose pairs with small values of this distance.

Finally, it will be desirable to retain as many IMG-USMG pairs as possible in the final matched design. Since some hospitals do not contain sufficient USMG operations to provide partners for all available IMG operations, at least a few IMG operations must be excluded. However, as discussed in Section 1.2, retaining a large proportion of IMG operations will help maintain a degree of external validity for our study, and may lead to greater power as well.

For details on how exact matching constraints, caliper constraints, and each of these objective functions may be represented in a network flow problem, see Section A of the online supplement.

3.2 Exploring the family of Pareto optimal matches

Given our three objective functions, how are we to explore the space of possible attractive matching designs? Consider first the tradeoff between any two of the three objective functions, labeled f_1 and f_2.The tools of Section 2.2 allow us to produce several individual Pareto optimal solution efficiently by choosing some penalty values ρ and optimizing $f_1(\mathbf{x}) + \rho f_2(\mathbf{x})$ for each. This collection of points is only a small subset of the full Pareto optimal set. However, because the set of Pareto optimal sets is highly structured, relatively small subsets of Pareto optimal points can actually place informative bounds on the locations of the points not yet computed. This leads to an effective approach for exploring the Pareto optimal set.

The intuition behind our solution is that additional Pareto optimal points of interest must lie "between" those already computed. To be more explicit, first consider any collection of distinct

Pareto optimal points and their associated matches (we distinguish here between distinct Pareto optimal solutions, which can share identical objective values, and distinct Pareto optimal points, which by definition do not). If we arrange the points so that their objective values for f_1 are strictly increasing, then their objective values for f_2 must necessarily be decreasing. To be concrete, if f_1 represents the average pair distance and f_2 represents the average imbalance, arranging Pareto optimal matches in order of increasing pair distance is the same as arranging them by decreasing imbalance.

The following result builds on this idea by describing how additional Pareto optimal points not yet computed must fit into such an ordering.

Theorem 3. *Let $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots, \mathbf{z}^{(m)}$ be a sequence of Pareto optimal points ordered to be increasing in f_1, with $\mathbf{z}^{(1)} = \underline{\mathbf{f}}$ and $\mathbf{z}^{(m)} = \overline{\mathbf{f}}$. Then any additional Pareto optimal point $\mathbf{z} = (z_1, z_2)$ satisfies*

$$z_1^{(i)} < z_1 < z_2^{(i+1)} \quad \text{and} \qquad z_2^{(i+1)} < z_2 < z_2^{(i)} \qquad \text{for some } i \in \{1, \ldots, m-1\}.$$

The proof follows directly from the definitions of Pareto optimality and of "endpoints" $\underline{\mathbf{f}}$ and $\overline{\mathbf{f}}$ given in Section 2.2. Theorem 3 suggests that if one could obtain a collection of Pareto optimal points, including $\underline{\mathbf{f}}$ and $\overline{\mathbf{f}}$, such that any one member of the collection was near another, one could conclude with confidence that any additional Pareto points would also be fairly similar to some already in the set.

This principle suggests that it should not be necessary to compute all Pareto optimal points in any problem of substantial size — identifying a sufficiently dispersed subset of the Pareto optimal set gives a detailed picture of the full spectrum of possible tradeoffs. We suggest a simple approach for exploring the Pareto optimal tradeoff curve: solve for a large value of ρ and a small value of ρ. If the two matches differ substantially on either imbalance or average pair distance, then solve for a grid of a few intermediate ρ-values, producing a collection of solutions from an increasing sequence of ρ-values $\rho_1, \ldots, \rho_m$. Then, for any pair of neighboring ρ-values (ρ_i, ρ_{i+1}) whose matches differ substantially, solve for a new grid of ρ-values between these ones. This process can be repeated several times to identify a sequence of distinct Pareto points. If each point is close to its neighbors on both objective functions, we can be confident that any additional Pareto optimal point will be similar to one already computed. Notice that since $\mathcal{P}(\rho)$ is a standard network flow problem for

any relevant value of ρ, computation simply involves repeated use of efficient standard network flow solvers. For additional remarks and technical results concerning this iterative approach to exploring the Pareto optimal set, see Section C of the online supplement.

Theorem 3 assumes that the initial collection of Pareto optimal points includes the endpoints $\overline{\mathbf{f}}$ and $\underline{\mathbf{f}}$. These solutions are guaranteed to be recovered by choosing sufficiently large and small values of ρ as shown in Section C of the online supplement. For any ad hoc value of $\rho > 0$, it is also easy to test if a solution is an endpoint, since at endpoint $\underline{\mathbf{f}}$ ($\overline{\mathbf{f}}$), the objective value f_1 (f_2) must agree with the objective value achieved by optimizing that function alone and ignoring the other objective completely. For example, in the tradeoff between balance and close pairing in the IMG-USMG study, we chose an initial value of ρ small enough that the average pair distance achieved was identical to that in a simple Mahalanobis distance match without balance constraints. In other cases it is easy to tell when an endpoint has been reached by the nature of the objective function; for example, imbalance on surgical experience cannot decrease below zero.

All the procedures discussed so far focus on varying the relative emphasis on two objective functions, holding other aspects of the optimization problem fixed. Why not consider all three of our objective functions — imbalance on surgical experience, average within-pair covariate distance, and sample size — at once? Although the theory of Pareto optimality extends naturally to problems with three or even K objective functions, efficient exploration of the Pareto optimal set in this setting is much more difficult. In particular, it is no longer generally possible to establish a simple ordering among Pareto optimal solutions that holds for all the objective functions at once. Accordingly, we focus on varying two objective functions at a time, holding the other fixed. For more discussion of the possibilities and challenges of problems with more than two objective functions, see Section D of the online supplement.

3.3 Balance vs. close pairing tradeoff

We now put this strategy into practice to examine specific tradeoffs in the IMG-USMG study. Our starting point is a standard off-the-shelf matching design, in which we seek to minimize the average Mahalanobis distance in a matched pair, while retaining as many IMG operations as possible,

without explicit attention to marginal balance. This match achieves very good rates of close or exact matching on important prognostic variables: in 92% of pairs the IMG and USMG performed the same procedure, and in 96% of pairs the IMG and USMG patients had identical coarsened Elixhauser indices. However, surgical experience remains highly imbalanced. We first consider whether it is possible to address the large imbalance on surgical experience by relaxing emphasis on close pairing while maintaining as many IMGs in the match as possible; retaining almost all of the IMG operations helps ensure distributional similarity between the operations in our matched design and the overall distribution of IMG operations in our data, which will ensure a degree of external validity.

Let f_1 be the sum of the Mahalanobis distances for the IMG-USMG operation pairs used in the network flow solutions. Let f_2 be the total variation distance between the matched IMG and USMG distributions across deciles of surgical experience. The directly-constrained problem $\mathcal{Q}(a)$ corresponds to setting a maximal level of total variation imbalance a and finding the minimal total of within-pair edge costs under this constraint; the penalized problem $\mathcal{P}(\rho)$ instead sets a cost of ρ on each unit imbalance. We solve the penalized problem for a shrinking grid of ρ-values as described in Section 3.2.

To ensure IMG operations are retained where possible, retaining maximal sample size, we assign a high cost to portions of the network that are used to remove treated units from the match (for details see Section A of the online supplement). This cost remains fixed as ρ is varied, these costs can be considered to be part of f_1, so that f_1 is actually a linear combination of the total pair distances and the number of excluded units. However, since the number of excluded operations remained fixed across the matches computed, the specific costs do not actually play a role in the tradeoff, and we ignore their costs when reporting values for f_1.

Table 3 and Figures 2(a) and 2(b) summarize the matches, each containing 18,888 matched pairs. In the table, each column gives the match produced under a different ρ-value. Reading from left to right shows how various design metrics change with increasing ρ, or (equivalently) as greater emphasis is placed on balance relative to close pairing. Match E gives a match with strict priority on balance over close pairing, the solution that would have been produced by matching under a

near fine balance constraint as in Yang et al. (2012); Match A approximates the solution that would be produced by a minimal-distance optimal match, without any accounting for balance. The intermediate matches, however, do not correspond to matches produced by any standard method. Figures 2(a) and 2(b) include points representing additional matches not included in the table for a more complete picture of the tradeoff.

Theorem 2 allows us to interpret each solution without reference to ρ. For example, Match C achieves a total variation imbalance of 3820 operations; therefore this match is optimal for the directly constrained problem $\mathcal{Q}(3280)$, which allows at most an imbalance of 3280 on experience deciles. This imbalance is equivalent to a scaled total variation distance of 0.20; Pareto optimality is preserved whether we consider the scaled or the raw imbalance as the objective function of interest. As ρ is increased, values of f_2 in the optimal solutions decrease while values of f_1 increase; from the perspective of the directly-constrained problem, this shows how the optimal choice of f_1 increases as f_2 is constrained below lower and lower thresholds a.

There are likely hundreds more distinct Pareto-optimal solutions for these two objective functions besides the points shown in Table 3 or Figures 2(a) and 2(b). However, Theorem 3 tells us that points not yet computed must lie in a constrained region between two adjacent points we have already computed. Since the points we have computed are dispersed broadly across the range of the tradeoff between matches A and E with relatively small gaps between them, not much additional information about the relative costs of balance and close pairing in this data will be gained by computing additional points. For example, the point for match B and the point just to its right in Figure 2(a) are sufficiently close that we know that any additional point will be relatively similar to one we have already computed.

In accordance with the increasing emphasis on balance, the magnitude of standardized differences on both experience overall and individual deciles of experience decrease as the value of ρ is increased from left to right in Table 3. However, the standardized difference on experience never drops below 0.22, so balance remains quite poor even in Match E. In addition, the increasing emphasis on balance at the expense of close pairing appears to have very little effect on the rate at which individuals are matchzd exactly on the variables composing the Mahalanobis distance. In

the bottom part of the table, there is almost no change in the proportion of exact matching on any variable of interest.

While there is a mathematical tradeoff between balance on experience and close matching on other variables, the minimal impact of that tradeoff on close pairing means that we can safely select Match E and get improved balance on experience "for free," without appreciable negative impact on close pairing. However, the balance on experience remains poor even when it is strictly prioritized over close pairing. Further modifications must be made to the match before it is acceptable for analysis.

3.4 Balance vs. sample size tradeoff

Balance cannot be improved sufficiently in this match merely by focusing less on close pairing; however, substantial gains may still be possible if we are willing to reduce the size of the matched sample. Accordingly, we now turn to a different tradeoff using the same network flow problem. We define a new objective function f_1 as a linear combination of the goals in the previous tradeoff, given by the sum of the Mahalanobis distances for the pair edges used in the network flow solutions and a constant M multiplied by the the total variation imbalance in surgical experience. Guided by Match E in Table 3, we choose M to be 449.86 to specify a strict priority of balance over close pairing, although we could choose any larger value of M; the purpose of weighting imbalances by M is simply to enforce a strict priority, specifying that among individuals we keep in the match, their balance on surgical experience is overridingly more important than their proximity on a multivariate distance. This is similar to the use of a large fixed penalty for excluding units in the the balance-versus-pairing tradeoff just considered.

Next, we define f_2 to be the number of IMG operations excluded from the match. Minimizing f_2 now corresponds to maximizing sample size, and the directly constrained problem $\mathcal{Q}(a)$ now gives the match with minimal deviations from balance of size at least a. The penalized problem $\mathcal{P}(\rho)$ corresponds to setting the cost (on the associated edge in the network flow problem) of excluding an IMG from the match uniformly to ρ.

We again solve $\mathcal{P}(\rho)$ for a grid of ρ-values, giving results in Table 4 and Figures 2(c) and 2(d).

Now the penalty ρ describes the importance of retaining each individual IMG operation rather than avoiding imbalances, and it decreases from left to right in Table 4 to show increasing attention to balance (relative to sample size). Note that Match F is almost identical to Match E, since both matches use a strict prioritization of the three design goals (sample size, balance, and finally close pairing), so that Table 4 essentially picks up where Table 3 left off. In addition, many of the values of ρ illustrating the tradeoff are similar in value to the fixed imbalance penalty M, suggesting that much of the tradeoff relationship can be understood by considering cases when the exclusion penalty is either just larger or just smaller than the imbalance penalty.

As in the tradeoff explored in Section 3.3, we see the values of the two objective functions change in opposite directions as attention to balance increases, but this time the magnitude of the changes are consequential for both balance and sample size. Standardized differences on experience are reduced from unacceptably high values to zero across the spectrum of matches, while the percentage of IMG operations excluded ranges from 6% to 65%. In addition, as plots 2(c) and 2(d) highlight, there is a desirable intermediate region of the spectrum where imbalance on experience has been reduced almost to zero but the large majority of IMG operations are retained in the match. Matches I-L all fall into this category, each reducing the standardized difference on experience almost to zero while retaining close to 80% of the IMG operations. These matches are similar enough that there is not a strong reason to prefer one of them over the others; however, Match K (achieved by setting $\rho = 448.74$) achieves the minimal value for the standardized difference on surgical experience among this group of matches and we use it for our outcome analysis in Section 4.2.

Note that Match F and Match N correspond to matches that could have been produced by standard technologies. Match F ($\rho = 500$) is equivalent to matching under a near fine balance constraint, and Match N (which was not produced using the tradeoff algorithm) is a strict exact match on experience decile. Without the tradeoff algorithm, users might choose in practice between one of these two matches, each of which performs poorly with respect to one of the two important design goals. The tradeoff algorithm allows exploration of the desirable intermediate region and discovery of Match K.

3.5 Computation time

All of the matches described in Section 4.3 and 4.4 were run on a desktop computer in sequence over a period of several days. While several of the individual matches took multiple hours to complete, prohibiting a real-time interactive study of the Pareto optimal point set, computational demands were reasonable in practice. For a more in-depth analysis of the runtime of the individual matches in this problem and a discussion of potential improvements in efficiency, see Section E of the online supplement.

4 Outcome analysis: comparing surgeon performance

4.1 Review: sensitivity of matched comparisons to unmeasured bias

Suppose our matching design succeeds in pairing operations whose probabilities of being performed by an IMG are identical, and consider all permutations of IMG-assignment such that exactly one operation in each matched pair is assigned to an IMG. Because no operation in a pair is more likely than the other to be treated by an IMG, all such permutations are equally likely. Under a sharp null hypothesis that each patient's outcome would remain identical had its IMG/USMG assignment been reversed, we can calculate the value of a given test statistic for every possible permutation, which provides a distribution for randomization inference (Fisher 1935). When comparison groups in an observational study differ on an unobserved covariate, however, probabilities of treatment assignment may not be equal within matched pairs, and randomization inference as described above will be biased. Sensitivity analyses probe studies to find the degree of unmeasured bias needed to overturn the results (Rosenbaum 2010). Formally, in a matched analysis with I pairs, let π_{ij} be the probability of treatment by an IMG for individual $j \in \{1, 2\}$ in pair i, and suppose

$$\Gamma^{-1} \leq \frac{\pi_{i1}/(1-\pi_{i1})}{\pi_{i2}/(1-\pi_{i2})} \leq \Gamma$$

for some value Γ. If $\Gamma = 1$ the original randomization test is valid, but if $\Gamma > 1$ it may not be. For any fixed $\Gamma > 1$, any of a collection of possible permutation distributions may be the true

randomization distribution, but procedures exist to find the worst-case p-value over this collection. In a sensitivity analysis, worst-case p-values are computed for larger and larger values of Γ until the study's results cease to be significant. This threshold Γ is referred to as the sensitivity of a study to unmeasured bias. In testing scenarios where a null hypothesis of zero effect is initially rejected, a sensitivity analysis describes the degree of confounding necessary to explain the observed results when no effect is present. In equivalence testing, on the other hand, the null hypothesis states that an effect exceeds a certain threshold size; if such a hypothesis is initially rejected (indicating only small effects are consistent with the data), a sensitivity analysis describes the degree of confounding necessary to mask a true effect at least as large as the threshold. In practice sensitivity analysis for an equivalence test is performed by conducting separate sensitivity analyses for two one-sided tests and combining the results, as described in Rosenbaum and Silber (2009b). For more discussion of sensitivity analysis, see Gilbert et al. (2003), Egleston et al. (2009), Hsu and Small (2013), and Fogarty and Small (2016).

4.2 Results for IMG-USMG match

We focus on in-hospital mortality as an outcome measure. In the matched sample produced by Match K in 4, the patient mortality rate for IMG operations is 1.5% while the mortality rate for USMG operations is 1.7%. We first test the sharp null hypothesis of no effect of treatment by an IMG using McNemar's test, which gives a p-value of 0.14. Since no difference significant at the 0.05 level is detected, we next perform a test of equivalence using the two-one-sided test procedure, attempting to reject null hypotheses that the effect is larger in magnitude in either direction than a given effect threshold. As a measure of effect size we use the attributable effect (Rosenbaum 2002a), and as an effect threshold we use one fourth of the age-75 mortality rate in 2008 as in Pimentel et al. (2015), which corresponds to a difference of approximately 162 total deaths between the matched IMG and USMG samples. Note that the familywise error rate is controlled even though we perform both one-sided tests without multiple testing correction (Goeman et al. 2010). The equivalence test p-value is below $1 \cdot 10^{-10}$, so we reject the hypothesis that the true effect size exceeds 162 deaths.

Might bias due to unmeasured confounding variables be masking a true differential effect on

mortality? We address this concern via a sensitivity analysis as described in Section 4.1 for the test of equivalence. The sensitivity analysis shows that the equivalence result remains significant at the 0.05 level for a Γ-value of 1.7, but ceases to hold when $\Gamma = 1.8$. An unmeasured bias of $\Gamma = 1.7$ could be produced a binary confounding variable perfectly correlated with patient mortality that also raised odds of treatment by an IMG by a factor of 1.7; such a bias could also be produced by a binary confounding variable that, when present, multiplies the odds of in-hospital mortality by eight while doubling the odds of treatment by an IMG (for more details on this calculation see Rosenbaum and Silber 2009a). Thus our overall finding is that IMGs and USMGs perform equivalently with respect to in-hospital patient mortality, and that the result is robust to a moderate degree of unmeasured confounding.

These results are broadly consistent with the findings of the original study (Zaheer et al. 2017). They go beyond them in at least two important ways. First, the original study accounted for the role of surgical experience only post hoc via adjustment by a parametric linear model, while our approach explicitly balances this potentially important confounder as part of the match itself. Second, we conduct a test of equivalence and an associated sensitivity analysis. Rather than declaring a failure to detect any positive effect that is robust to unmeasured confounding, we are able to positively establish the absence of an effect of a certain standard size under minor to moderate unobserved confounding.

5 Discussion

Understanding the impact of surgical training on performance is an important step in determining the key medical and institutional factors driving quality of care for surgical patients. The findings presented here help establish equivalence on key performance metrics between surgeons trained outside the United States and those trained domestically. These results are consistent with previous findings in studies that compared patient outcomes for internationally- and domestically-trained physicians (Ozuah et al. 2001; Norcini et al. 2010; Zaheer et al. 2017). Our work stands out among this literature in its handling of observed confounding variables entirely through a carefully-chosen matching design, in obviating the need to fit and interpret parametric regression models, and in

its sophisticated handling of unobserved bias via a sensitivity analysis for an equivalence test. We hope that our contributions will help dispel stigma that internationally-trained surgeons have been shown to face (Moore and Rhodenbaugh 2002).

On the methodological side, the connection between the directly-constrained problem and the penalized problem illustrated by Theorem 2 provides valuable insight into the common practice of using penalties in matched designs. Penalties are most often added to covariate distances to discourage individual matches that violate some design goal, as in Silber et al. (2001), Haviland et al. (2007), Rosenbaum et al. (2007), and Rosenbaum (2010, §8.4), but they may also be used to encourage compliance with more general design goals, as in Zubizarreta (2012) and Pimentel et al. (2015). Most of these works choose penalties to be very large to strictly enforce the goal in question with priority over other design elements. The results in Sections 2.2-2.3 generalizes the notion of enforcing a constraint via penalization both by allowing for a wide range of possible design goals and by giving clear interpretations for designs in which the penalties are not so large as to enforce strict priorities. A notable exception to the penalization literature's focus on strict priorities is Rosenbaum (2012); the Pareto optimality result proved there for a specific tradeoff setting is a special case of Theorem 2.

Section 3 describes the exploration of two specific tradeoffs in the IMG-USMG study using the penalized problem. What other tradeoffs in matching might merit study in this manner? In studies of hospital processes, one might wish to match exactly on a group of process-related variables such as emergency room status and surgical procedure but might be content with marginal balance of comorbidities. Tradeoffs between a pairwise distance based on the process-related variables and balance on each of several individual comorbidities, much like the tradeoff in Section 3.3, could be considered. More generally, one might study the tradeoff between minimizing average matched-pair covariate distances and matching exactly on a nominal variable. Exact matching rate could be quantified using the flow over network edges that connect inexact matches on the nominal variable. The penalized problem would set the cost of these edges to the penalty ρ, and intermediate solutions would give the minimal average-distance matches among designs that match exactly at at least the achieved rate. In a similar vein, one could use the algorithm to evaluate the implications of using

one pairwise similarity metric, such as the difference in propensity scores, versus another, such as the Mahalanobis distance.

The tradeoff framework may also be employed in instrumental variable studies to understand tradeoffs between sample size and instrument strength (Keele and Morgan 2016). Consider a study with a continuous instrument in which subjects are first divided into a high-dose group and a low-dose group (perhaps discarding some subjects with mid-range doses), and the goal is to form high-dose-low-dose pairs that maximize the difference in dose, subject to other constraints; for an example of such a study see Yang et al. (2014). Here matching distances for potential high-dose-low-dose pairs could be defined by inverses of the actual dose difference, so that similar doses would be associated with large distances and vice versa. The sum of these distances (describing instrument strength) forms one dimension of the tradeoff, while the number of individuals excluded forms the other. Theorem 2 tells us that the individual matches produced by solving the penalized problem will have optimal separation on the instrument (as defined by the distance) among matches of equal or greater cardinality. This approach to managing the IV strength/sample size tradeoff is similar in some respects to repeated near-far matching with varying numbers of sinks (Baiocchi et al. 2010, 2012); however, it works with predefined low-dose and high-dose groups rather than forming nonbipartite matches, and unlike near-far matching it can support balance constraints.

Yet another potential tradeoff concerns the maximal sized of matched sets in variable-ratio matching (Ming and Rosenbaum 2000). Suppose one wishes to form matched sets of potentially varying sizes so that most or all controls belong to some set. Two standard options are fixed-ratio matching and unconstrained variable-ratio matching. The former approach may force poor matches and introduce bias, while the latter may use data inefficiently (Hansen 2004). The tradeoff algorithm can help explore options between these extremes using the network of Rosenbaum (1989) for unconstrained variable-ratio matching, which is based around a familiar bipartite graph of treated nodes and control nodes, but includes an additional source-like node called the overflow node that sends extra flow through treated units matching to many controls. The edges connecting the overflow node to each treated unit govern the maximum allowed control:treated matching ratio, or matching ratio cap, via their capacities. Split each of these edges into an edge of small capacity

with no cost and an edge of high capacity, and consider the tradeoff between the sum of flow across these latter edges against the sum of matched pair distances. Solving $\mathcal{P}(\rho)$ with extreme values of ρ should recover fixed-ratio and unconstrained variable-ratio matches. Intermediate matches can be interpreted, via Theorem 2, as minimum-bias matches with no more than a fixed number of controls matched in excess of a matching ratio cap. In large problems, this may allow more fine-grained control of the tradeoff than can be achieved by varying the ratio cap.

References

American Medical Association (2013), "AMA Physician Masterfile," http://www.ama-assn.org/ama/pub/about-ama/physician-data-resources/physician-masterfile.page, accessed March 2016.

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., Small, D. S., Yang, L., Polsky, D., and Groeneveld, P. W. (2012), "Near/far matching: a study design approach to instrumental variables," *Health Services and Outcomes Research Methodology*, 12, 237–253.

Bertsekas, D. P. (1998), *Network optimization: continuous and discrete models*, Belmont, MA: Athena Scientific.

Bertsekas, D. P., Tseng, P., et al. (1994), *RELAX-IV: A faster version of the RELAX code for solving minimum cost flow problems*, Massachusetts Institute of Technology, Laboratory for Information and Decision Systems Cambridge, MA.

Egleston, B. L., Scharfstein, D. O., and MacKenzie, E. (2009), "On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death," *Biometrics*, 65, 497–504.

Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. (1998), "Comorbidity measures for use with administrative data," *Medical care*, 8–27.

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

Florida Agency for Health Care Administration (2008-2011), "Hospital Inpatient Discharge Data," http://www.floridahealthfinder.gov/researchers/researchers.aspx, accessed March 2016.

Fogarty, C. B. and Small, D. S. (2016), "Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming," *Journal of the American Statistical Association*, 111, 1820–1830.

Ford, Jr, L. R. and Fulkerson, D. R. (1962), *Flows in networks*, Princeton, NJ: Princeton University Press.

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003), "Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials," *Biometrics*, 59, 531–541.

Goeman, J. J., Solari, A., and Stijnen, T. (2010), "Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority," *Statistics in medicine*, 29, 2117–2125.

Hansen, B. B. (2004), "Full matching in an observational study of coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618.

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.

Haviland, A., Nagin, D. S., and Rosenbaum, P. R. (2007), "Combining propensity score matching and group-based trajectory analysis in an observational study." *Psychological methods*, 12, 247.

Hsu, J. Y. and Small, D. S. (2013), "Calibrating sensitivity analyses to observed covariates in observational studies," *Biometrics*, 69, 803–811.

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.

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.

Miettinen, K. (1999), *Nonlinear multiobjective optimization*, Kluwer Academic Publishers.

Ming, K. and Rosenbaum, P. R. (2000), "Substantial gains in bias reduction from matching with a variable number of controls," *Biometrics*, 56, 118–124.

Moore, R. A. and Rhodenbaugh, E. J. (2002), "The unkindest cut of all: are international medical school graduates subjected to discrimination by general surgery residency programs?" *Current surgery*, 59, 228–236.

New York State Department of Health (2008-2011), "Statewide Planning and Research Cooperative System (SPARCS) Non-Identifying Data," http://www.health.ny.gov/statistics/sparcs/#datainfo, accessed March 2016.

Norcini, J. J., Boulet, J. R., Dauphinee, W. D., Opalek, A., Krantz, I. D., and Anderson, S. T. (2010), "Evaluating the quality of care provided by graduates of international medical schools," *Health Affairs*, 29, 1461–1468.

Ozuah, P. O., Curtis, J., and Dinkevich, E. (2001), "Physical examination skills of US and international medical graduates," *Jama*, 286, 1021–1021.

Papadimitriou, C. H. and Steiglitz, K. (1982), *Combinatorial optimization: algorithms and complexity*, Courier Corporation.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), "Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons," *Journal of the American Statistical Association*, 110, 515–527.

Resa, M. d. l. A. and Zubizarreta, J. R. (2016), "Evaluation of Subset Matching Methods and Forms of Covariate Balance," *Statistics in Medicine*, 35, 4961–4979.

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

— (2002a), "Attributing effects to treatment in matched observational studies," *Journal of the American Statistical Association*, 97, 183–192.

— (2002b), *Observational Studies*, New York, NY: Springer.

— (2005), "Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies," *The American Statistician*, 59, 147–152.

— (2010), *Design of Observational Studies*, New York, NY: Springer.

— (2012), "Optimal Matching of an Optimally Chosen Subset in Observational Studies," *Journal of Computational and Graphical Statistics*, 21, 57–71.

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.

Rosenbaum, P. R. and Rubin, D. B. (1985), "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *The American Statistician*, 39, 33–38.

Rosenbaum, P. R. and Silber, J. H. (2001), "Matching and thick description in an observational study of mortality after surgery," *Biostatistics*, 2, 217–232.

— (2009a), "Amplification of sensitivity analysis in matched observational studies," *Journal of the American Statistical Association*, 104, 1398–1405.

— (2009b), "Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units," *Journal of the American Statistical Association*, 104, 501–511.

Rubin, D. B. (1980), "Bias Reduction Using Mahalanobis-Metric Matching," *Biometrics*, 36, 293–298.

— (2007), "The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials," *Statistics in medicine*, 26, 20–36.

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Even-Shoshan, O., Chen, W., Zhang, X., and Mosher, R. E. (2001), "Multivariate matching and bias reduction in the surgical outcomes study," *Medical care*, 1048–1064.

Stuart, E. A. (2010), "Matching methods for causal inference: A review and a look forward," *Statistical Science*, 25, 1–21.

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.

Yang, F., Zubizarreta, J. R., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2014), "Dissonant conclusions when testing the validity of an instrumental variable," *The American Statistician*, 68, 253–263.

Zaheer, S., Pimentel, S. D., Simmons, K. D., Kuo, L. E., Datta, J., Williams, N., Fraker, D. L., and Kelz, R. R. (2017), "Comparing International and United States Undergraduate Medical Education and Surgical Outcomes Using a Refined Balance Matching Methodology," *Annals of Surgery*, 265, 916–922.

Zeleny, M. (1974), *Linear multiobjective programming*, vol. 95, Springer Science & Business Media.

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.

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

Table 1: Matching matrix for the example in Figure 1 with specific values given for the distance parameters Δ_i and ϵ_i. Each row corresponds to an IMG operation and each column corresponds to a USMG operation, and each operationis given a name containing a hospital index (a number 1 through 2) and a letter indexing operations within the same hospital. The font of the row/column labels indicates the experience level of the surgeon performing the operation; labels in regular font refer to operations by experienced surgeons (XP) while those of inexperienced surgeons (IXP) have labels in italics.

	1Y	1Z	*2Y*	2Z
1A	$\Delta_1 = 6$	$\Delta_1 + \epsilon_1 = 7$	-	-
2A	-	-	$\Delta_2 = 3$	$\Delta_2 + \epsilon_2 = 83$

Table 2: Solutions to the matching problem described in Table 1 for several different tradeoffs between the two design goals. The first column of the table shows the matched pairs for Hospitals 1 and 2 in each match. The two central columns of the table show measures of success on the two design goals, close pairing and balance on a binary indicator of surgical experience (measured by the number of total excess USMG operations in each category). The plot shows the tradeoff between these objectives graphically. The final two columns of the table indicate whether and how the match in each row may be produced by solving $\mathcal{Q}(a)$ and $\mathcal{P}(\rho)$.

	Matches	f_1	f_2	Optimal for $\mathcal{Q}(a)$ with $a \in$	Optimal for $\mathcal{P}(\rho)$ with $\rho \in$
1	(*1A*, *1Y*), (*2A*, *2Y*)	9	2	$[2, \infty)$	$[0, 1]$
2	(*1A*, *1Y*), (*2A*, 2Z)	10	1	$[1, 2)$	$[1, 80]$
3	(*1A*, 1Z), (*2A*, 2Z)	90	0	$[0, 1)$	$[80, \infty)$

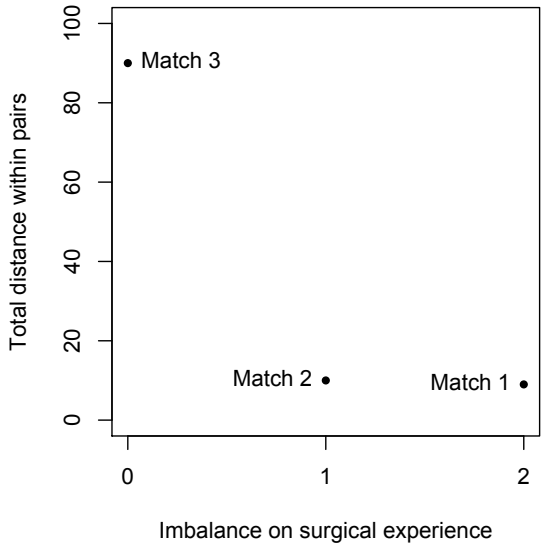
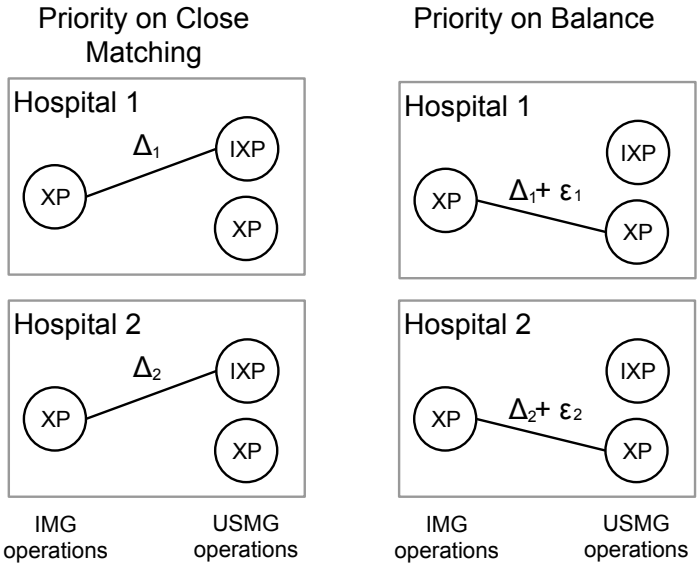


Figure 1: Illustration of a hypothetical tradeoff between close pairing on a covariate distance and covariate balance in the IMG-USMG match. XP stands for treatment by an experienced surgeon and IXP for treatment by an inexperienced surgeon. Matches are restricted to occur within hospitals. Optimal balance is achieved when the IMG operations in Hospitals 1 and 2 are matched to the USMG operations with XP; however, optimal closeness in pairs is achieved instead when the IMG operations are matched to the USMG operations with IXP.



	Attention to balance (vs. close pairing)				
	Very low ⟵		⟶ Very high		
Imbalance penalty ρ	0.01	1	5	8	449.86
Match label	A	B	C	D	E
Avg. pairwise Mahalanobis distance	1.02	1.03	1.11	1.18	1.23
f_1 (Sum of pairwise distances)	19241.24	19412.02	20960.35	22205.27	23241.56
f_2 (Total variation imbalance on exp. deciles)	4926	3784	3280	3075	2982
Scaled TV imbalance	0.26	0.20	0.17	0.16	0.16
	Standardized differences				
Experience	0.47	0.32	0.26	0.23	0.22
Experience <5 yrs	-0.11	-0.06	-0.03	-0.02	-0.02
Experience 5-7 yrs	-0.03	0.00	0.00	0.00	0.00
Experience 8-10 yrs	-0.05	-0.00	-0.00	0.00	-0.00
Experience 11-12 yrs	-0.08	-0.01	-0.00	0.00	0.00
Experience 13-14 yrs	-0.17	-0.16	-0.15	-0.14	-0.13
Experience 15-17 yrs	-0.32	-0.34	-0.30	-0.29	-0.30
Experience 18-20 yrs	-0.17	-0.16	-0.16	-0.15	-0.15
Experience 21-24 yrs	-0.02	0.00	0.00	0.00	0.00
Experience 25-30 yrs	0.07	0.00	0.00	0.00	0.00
Experience 31+ yrs	0.58	0.47	0.39	0.36	0.35
Experience missing	0.28	0.26	0.25	0.25	0.25
Elixhauser index	0.01	0.01	0.01	0.01	0.01
Emergency admission	-0.01	-0.01	-0.02	-0.02	-0.02
Sex	0.02	0.02	0.02	0.02	0.03
Age	0.01	0.02	0.02	0.02	0.02
Largest over 32 procedure types	-0.06	-0.06	-0.06	-0.06	-0.06
	Proportion matched exactly				
Procedure type	0.92	0.92	0.91	0.91	0.90
Coarsened Elixhauser index	0.96	0.96	0.95	0.95	0.95
Emergency admission	0.98	0.98	0.98	0.98	0.98
Sex	0.96	0.96	0.95	0.95	0.95
Age quintile	0.75	0.74	0.73	0.72	0.72
Experience decile	0.08	0.08	0.08	0.08	0.08

Table 3: Summary of matches. The five central columns show different matches that vary the tradeoff between achieving balance on deciles of surgical experience and minimizing the sum of matched-pair differences. Match A ($\rho = 0.01$) approximates a pure distance-based match that ignores balance on experience deciles completely, while Match E ($\rho = 449.86$) prioritizes balance on experience over forming close pairs. The second and third rows of the table show the values of the objective functions f_1 and f_2, corresponding to the sum of pairwise distances and the imbalance on experience deciles, respectively; the first and fourth rows give f_1 (ignoring the cost of excluded units) and f_2, respectively normalized by the number of matched samples (18,888). The next 17 rows of the table describe marginal balance on various covariates in the data, especially surgical experience and its deciles, using standardized differences. The final six rows give the proportion of pairs matched (out of 18,888 total pairs) exactly on individual discrete variables. Note that the coarsened Elixhauser index differs from the Elixhauser index only in that it combines all individuals with 3 or more comorbidities into one large category.

	Attention to balance (vs. sample size)								
			Very low	⟵	⟶	Very high			Exact
Exclusion penalty ρ	500	454.36	449.86	449.3	449.02	448.74	350	0.01	match
Match label	F	G	H	I	J	K	L	M	N
Prop. IMG operations dropped	0.06	0.09	0.15	0.17	0.18	0.18	0.21	0.65	0.62
f_2 (IMG operations dropped)	1136	1824	3061	3496	3564	3605	4118	13000	12450
f_1 (linear combination)	1.36E6	1.05E6	4.93E5	2.97E5	2.66E5	2.48E5	1.89E4	13.5	2.18E4
TV imbalance, exp. deciles	2982	2294	1057	622	554	513	0	0	0
Scaled TV imbalance	0.16	0.13	0.06	0.04	0.03	0.03	0.00	0.00	0.00
	Standardized differences								
Experience	0.23	0.16	0.05	0.01	0.00	0.00	-0.01	-0.00	0.02
Experience <5 yrs	-0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Experience 5-7 yrs	0.00	0.00	0.00	0.00	0.00	-0.00	0.00	0.00	0.00
Experience 8-10 yrs	-0.00	-0.00	-0.00	0.00	0.00	0.00	0.00	0.00	0.00
Experience 11-12 yrs	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	-0.00	0.00
Experience 13-14 yrs	-0.12	-0.09	-0.03	-0.00	-0.00	0.00	-0.00	0.00	0.00
Experience 15-17 yrs	-0.28	-0.26	-0.16	-0.12	-0.11	-0.11	0.00	0.00	0.00
Experience 18-20 yrs	-0.15	-0.11	-0.04	-0.01	-0.01	-0.00	-0.00	0.00	0.00
Experience 21-24 yrs	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Experience 25-30 yrs	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Experience 31+ yrs	0.35	0.26	0.09	0.03	0.02	0.02	0.00	0.00	0.00
Experience missing	0.25	0.23	0.18	0.16	0.16	0.15	0.00	0.00	0.00
Elixhauser index	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Emergency admission	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	0.00	-0.00
Sex	0.03	0.02	0.02	0.02	0.02	0.02	0.03	0.00	-0.00
Age	0.02	0.02	0.02	0.02	0.02	0.02	0.02	-0.00	0.04
Largest of 32 proc. types	-0.06	0.03	0.04	0.04	0.04	0.04	0.04	0.00	-0.05
	Proportion matched exactly								
Procedure type	0.90	0.93	0.92	0.92	0.92	0.92	0.91	1.00	0.78
Coarsened Elixhauser index	0.95	0.96	0.96	0.95	0.95	0.95	0.95	1.00	0.87
Emergency admission	0.98	0.98	0.98	0.98	0.98	0.98	0.97	1.00	0.93
Sex	0.95	0.95	0.96	0.96	0.96	0.96	0.95	1.00	0.89
Age quintile	0.72	0.73	0.74	0.73	0.73	0.73	0.72	0.97	0.54
Experience decile	0.08	0.08	0.09	0.10	0.10	0.10	0.10	0.10	1.00

Table 4: Summary of matches. The eight central columns show different matches that vary the tradeoff between achieving balance on deciles of surgical experience and maximizing the size of the matched sample. Match F ($\rho = 500$) strictly prioritizes retaining all IMG operations possible, while Match M ($\rho = 0.01$) freely sacrifices sample size to achieve improved balance on experience. Match N, computed using a different method (strict exact matching on deciles of experience), is provided for comparison. The second and third rows of the table show the values of the objective functions f_2 and f_1, corresponding to the number of IMG operations excluded and a linear combination of the sum of matched distances in pairs and the total variation imbalance on experience deciles scaled by 449.86, respectively. The first row gives f_2 as a proportion of the total number of available IMG operations (20,024) while the fourth and fifth rows give the raw and scaled (by the number of matched samples) total variation imbalance. The remainder of the table is analogous to Table 3 and is described in its caption.

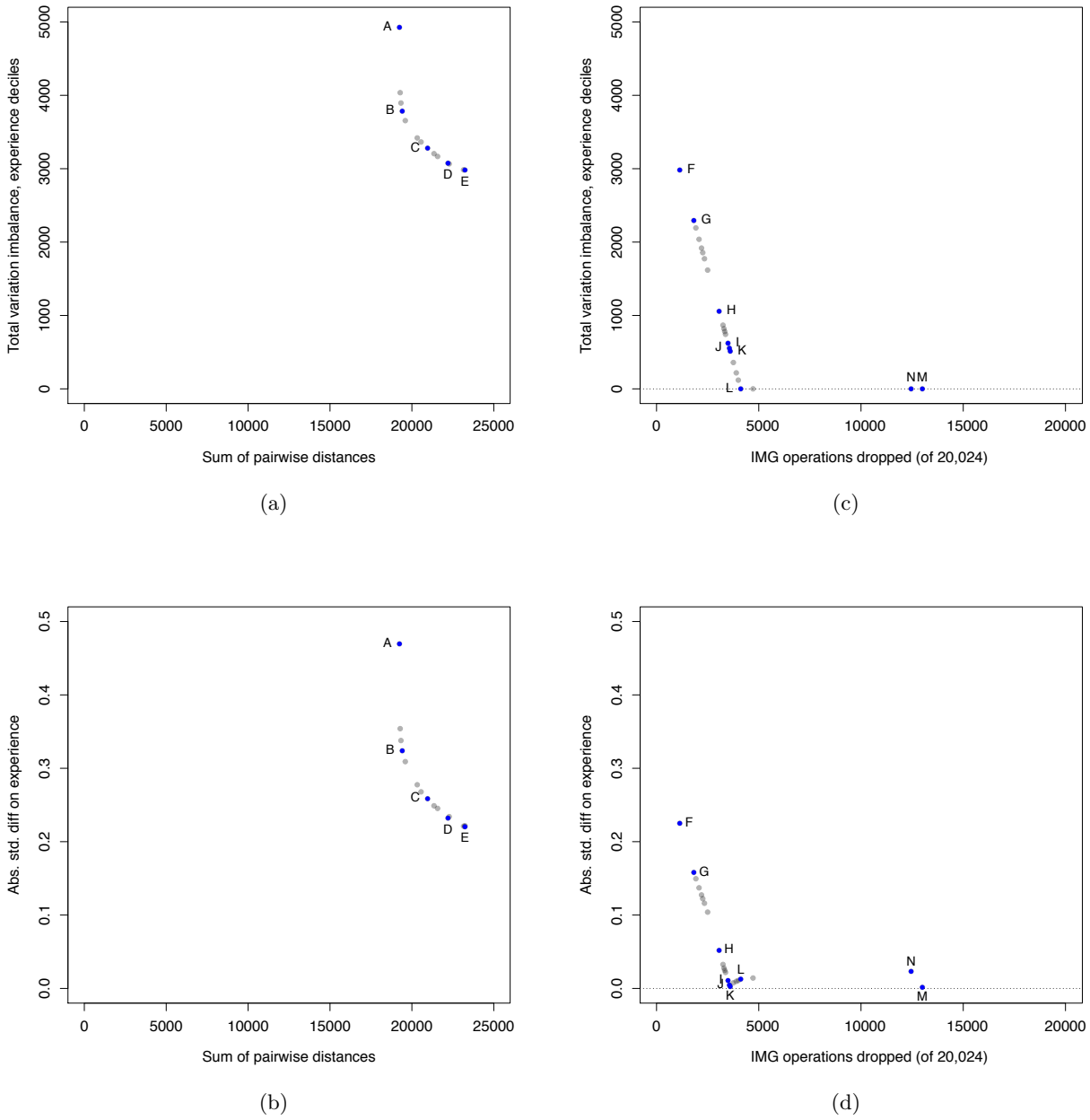


Figure 2: Plots showing tradeoffs in the matched comparison between balance on experience decile and either the sum of matched pairwise distances (subplots (a) and (b)) or the number of IMG operations excluded (subplots (c) and (d)). Each dot represents a different match, and the dark-colored dots are described in Tables 3 and 4. Plots (a) and (c) use total variation imbalance on deciles of surgical experience (a value ranging from 0 to the number of matched samples) on the y-axis; this is one of the objective functions being optimized directly. In contrast plots (b) and (d) use the absolute standardized difference, a different balance measure, for their y-axes.

Online supplement to "Optimal tradeoffs in matched designs comparing US-trained and internationally-trained surgeons"

Samuel D. Pimentel and Rachel R. Kelz*

December 23, 2019

Abstract

This supplement contains several components. Section A gives a detailed specification of the mapping between the matching problems in the main manuscript and their representation as network flow algorithms. Section B gives a richer description of the variables in the surgical outcomes dataset. Section C provides supporting results and additional context for the iterative approach to exploring the Pareto optimal solution set in the main mansucript. Section D generalizes results in the main paper specific to two-objective tradeoffs to settings with more than two objective functions, and discusses challenges in applying these results in practice. Section E gives an in-depth analysis of runtime for the matches conducted in the main manuscript. Finally, Section F contains proofs for technical results throughout the main manuscript and the supplement.

A Network flow problems

In this section we describe in detail the connection between matching and minimum-cost network flow algorithms, and give a detailed description of the networks used to conduct the tradeoff analysis in the main manuscript.

*Samuel D. Pimentel is Assistant Professor of Statistics, University of California, Berkeley, Berkeley, CA 94720 (spi@berkeley.edu) and Rachel R. Kelz is Associate Professor of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104 (rapaporr@uphs.upenn.edu).

A.1 Detailed characterization of matching as minimum-cost network flow

To characterize network algorithms formally, we adopt the setting of Bertsekas (1998). Consider a set of nodes $\mathcal{N}$ and a set of directed edges $\mathcal{E}$. The edge connecting node $n_1 \in \mathcal{N}$ to distinct node $n_2 \in \mathcal{N}$ is given by $e = (n_1, n_2)$ (only one such edge exists for each i, j combination). Each edge $e \in \mathcal{E}$ has a lower capacity of zero and a nonnegative integer upper capacity of d_e, and a nonnegative cost c_e. Finally, each node i has an integer supply $b_i \in \mathbb{Z}$. The decision variables $\mathbf{x} \in \mathbb{R}^{|\mathcal{E}|}$, which describe flow across each edge in $\mathcal{E}$, are constrained so that the total flow into a node must be equal to the total flow out of that node (minus any supply), and so that no edge carries negative flow or more flow than its upper capacity d_e. Formally, we may represent this constraint set $\mathcal{F}$ as follows:

$$\mathcal{F} = \left\{ \mathbf{x} : \sum_{k:(i,k)\in\mathcal{E}} x_{(i,k)} - \sum_{j:(j,i)\in\mathcal{E}} x_{(j,i)} = b_i \;\; \forall i \in \mathcal{N}, \quad 0 \leq x_e \leq d_e \; \forall e \in \mathcal{E} \right\}$$

To represent matching as a network flow problem, define a network with one node for each of T treated units $\tau_1, \tau_2, \ldots, \tau_T$ with a supply of 1, and one node for each of C control units $\kappa_1, \kappa_2, \ldots, \kappa_C$ with a supply of 0. Connect each τ_i to each κ_j (forming a bipartite graph) by an edge with cost c_{ij}, where $c_{ij} \geq 0$ for all $i = 1, \ldots, T$, $j = 1, \ldots, C$ is a distance describing the desirability of the pairing (τ_i, κ_j). Add an additional "sink" node ω with a supply of $-T$ (i.e. a demand for T units of flow) and connect each κ_j to ω by an edge with cost zero. All edges in the network have an upper capacity of one.

To see how matches are obtained from a solution to the network flow problem, suppose first that all x_e values were restricted to be integers. Now solutions to this network flow problem must send one unit of flow from each treated unit τ_i to one κ_j, and no κ_j can be matched to multiple τ_i since at most one unit of flow can pass on from each κ_j to the sink ω. Since the only edges with nonzero costs are the treatment-control edges with associated matching distances c_{ij}, the optimal flow solution will minimize the total cost of treatment-control pairings. The optimal set of treated-control pairs will be identified by the T edges of form $e = (\tau_i, \kappa_j)$ for which the associated x_e variable is set equal to 1. In fact the problem does not include a constraint forcing the x_e variables to take on integer values; however, as discussed in Section 2.1 of the main manuscript, some optimal

solution will always be integral and will be recovered by a certain class of solving algorithms.

The network just described implements a simple matching problem for pairing every treated unit to a control unit without replacement, allowing any treated unit to match to any control. By modifying the network, other constraints or objective functions can be built into the problem. For instance, suppose we wish to impose constraints which permit matches only between treated and control units sharing identical values of a nominal variable (*exact matching*) or only between treated and control units with values of a continuous variable that differ by no more than a fixed amount (*caliper matching*). These constraints can be represented by removing from the network any edges (τ_i, κ_j) associated with pairings that would violate the constraint. Sections A.2 and A.3 discuss more involved modifications to the network used to constrain or optimize marginal imbalances and sample size, respectively. Networks with these modifications are used to conduct the tradeoff analyses in Sections 3.3 and 3.4 of the main manuscript.

A.2 Optimizing marginal balance in a network flow algorithm

Yang et al. (2012, online appendix) and Pimentel et al. (2015) extend the basic network flow algorithm described in Section A.1 to allow constraints on marginal balance of one or more nominal covariates. Here we review the algorithm of Pimentel et al. (2015), which we build upon to implement the tradeoff analyses in the main manuscript.

Consider a series of nested refined balance covariates $\nu_1, \ldots, \nu_K$ each mapping each treated or control unit to a category in the set $\{\lambda_{k1}, \ldots, \lambda_{kL_k}\}$. We would like our algorithm to pay attention to marginal balance on each of these variables, measured by total variation distance. As before, the network contains treated nodes $\mathcal{T} = \{\tau_1, \ldots, \tau_T\}$ with supply one each, control nodes $\mathcal{C} = \{\kappa_1, \ldots, \kappa_C\}$ with supply zero, and a sink node ω with demand T. We include edges (τ_i, κ_j) with associated distances c_{ij} and upper capacity one; all possible treated-control edges may be included as in the network of Section A.1, or some strictly smaller set, perhaps defined by some combination of exact matching or caliper constraints. In addition, the network includes three nodes $(\lambda_{k\ell}, \lambda'_{k\ell}, \lambda''_{k\ell})$ for each category $\ell = 1, \ldots, L_k$ of each balance level $k = 1, \ldots, K$, all with supply 0, and edges connecting these three nodes into a "triangle" for each category: $(\lambda_{k\ell}, \lambda'_{k\ell})$ with infinite

capacity and cost Υ^{K-k+1} (where Υ is a network-wide penalty parameter), $(\lambda'_{k\ell}, \lambda''_{k\ell})$ with infinite capacity and cost zero, and $(\lambda_{k\ell}, \lambda''_{k\ell})$ with capacity $d_{k\ell} = |\{\tau_i \in \mathcal{T} : \nu_k(\tau_t) = \lambda_{k\ell}\}|$ and cost zero. The fine balance "triangles" with $k > 1$ and $\ell \in 1, \ldots, L_k$ are connected to each other via edges $(\lambda''_{k\ell}, \lambda_{(k-1)\ell'})$ (with infinite capacity and zero cost) where $\lambda_{(k-1)\ell'}$ is the potentially coarser category in level $k-1$ in which the category $\lambda_{k\ell}$ nests. Finally, each control node κ_c is connected to the level-K balance node $\lambda_{K\ell}$ such that $\nu_K(\kappa_j) = \lambda_{K\ell}$ by edges of zero cost and capacity one, and each level-1 balance node $\lambda''_{1\ell}$ is connected to the sink ω by edges of zero cost and infinite capacity.

As demonstrated in Pimentel et al. (2015), whenever $\Upsilon > 0$, an optimal flow through the network just described measures the total variation distance between the treated and control samples selected for the kth nominal variables via the sum of flows across the $(\lambda_{k\ell}, \lambda'_{k\ell})$ edges, which we denote as the *imbalance edges*, for all ℓ. Intuitively, as much flow as possible is routed over the costless $(\lambda_{k\ell}, \lambda''_{k\ell})$ edges, so that all that remains in the penalized edges is the discrepancy of units for category $k\ell$, discrepancies which sum to the total variation distance. Pimentel et al. (2015) recommend setting Υ to a very large value in order to constrain the algorithm to produce optimal balance in a hierarchical manner across the nominal variables. In Sections 3.3 of the main manuscript, we use a similar network structure with $K = 1$ nominal variables, but instead of using the imbalance edges to enforce a constraint we use the flow across these edges as our second objective function f_2.

A.3 Excluding treated units in a network flow algorithm

The network used for tradeoffs in Sections 3.3 and 3.4 of the main manuscript differs somewhat from the original network proposed in Pimentel et al. (2015), which requires all treated units in the problem to be matched. We modify this network to allow exclusion of treated units. It is important to note that although the network algorithm in Pimentel et al. (2015) supports 1:k matching for integers $k > 1$, the modifications described here only apply to one-to-one matching problems (where $k = 1$).

For ease of exposition, assume first that no balance constraint is present, so that λ-nodes are absent from the network completely and control nodes κ_c are all connected directly to the sink ω by edges of capacity one. For each treated unit t, we add an edge of capacity one between treated

node τ_t and the sink ω. Now for any treated unit t, a feasible flow exists in which the flow produced at τ_t travels directly to the sink via this new *bypass* edge, and none of the edges (τ_t, κ_c) for any c carry any flow. If no cost is assigned to the bypass edges, the empty match that excludes all treated individuals is optimal, so in general some penalty is placed on the bypass edges.

Finally, we consider the case where balance constraints are also present. If we add bypass edges connecting treated nodes directly to the sink as in the previous paragraph, the balance constraints will no longer work properly when treated individuals are excluded, because the capacities of the $(\lambda_{k\ell}, \lambda''_{k\ell})$ edges have been set with reference to the prevalence of category k_ℓ in the full treated sample. When treated individuals are excluded, the empirical distribution of covariate ν_k in the matched treated group also changes and this must be accounted for when calculating imbalances between matched groups. This problem is easily addressed by changing the bypass edges to connect treated nodes τ_t to the category of covariate ν_K associated with τ_t. Routing flow from excluded treated units through the nodes corresponding to those units' covariate values effectively subtracts out the excluded individuals from the capacity constraints and ensures that balance is being calculated and penalized correctly for the reduced sample.

In Section 3.3 of the main manuscript, a very high fixed penalty is placed on the bypass edges to ensure treated units are excluded from the match only when it is otherwise impossible to match them. In Section 3.4, the total flow across the bypass edges, equal to the number of treated units excluded, is used as one of the objective functions.

B Summary table for surgery outcomes dataset

Table B1 contains average values and standard differences for variables deemed scientifically important in the surgical outcomes dataset used in Section 3 of the main manuscript. As described in greater detail in the main manuscript, the dataset is a subset of the larger pool of records used in Zaheer et al. (2017) and contains only operations performed in the Orlando metropolitan area. To see the impact of various matched designs on the comparability of the IMG and the USMG samples, it is productive to compare this table to Tables 3 and 4 in the main manuscript.

	USMG operation average (n = 58,270)	IMG operation average (n = 20,024)	Standardized difference
Experience	14.4	19.6	0.536
Experience <5 yrs	0.112	0.092	-0.066
Experience 5-7 yrs	0.101	0.084	-0.061
Experience 8-10 yrs	0.135	0.082	-0.173
Experience 11-12 yrs	0.097	0.069	-0.102
Experience 13-14 yrs	0.094	0.043	-0.203
Experience 15-17 yrs	0.123	0.052	-0.254
Experience 18-20 yrs	0.099	0.048	-0.194
Experience 21-24 yrs	0.118	0.123	0.017
Experience 25-30 yrs	0.069	0.122	0.182
Experience 31+ yrs	0.039	0.235	0.593
Experience missing	0.012	0.050	0.218
Elixhauser index	2.00	1.90	-0.054
Emergency admission	0.506	0.588	0.166
Male sex	0.413	0.427	0.027
Age	56.2	54.9	-0.071

Table B1: Summary information for the surgical outcomes dataset used in Section 4 of the main text, consisting of all 2008-2011 Orlando-area operations in the data of Zaheer et al. (2017). The first two columns give average variable values for operations by USMGs and IMGs respectively, or proportions when the variable in question is binary. The third column gives the difference between the averages standardized by the pooled estimate of the standard deviation of the variable in question.

C Searching iteratively for Pareto optimal solutions

Section 3.2 of the main manuscript describes an iterative procedure for computing a broad range of Pareto optimal solutions for objective functions f_1 and f_2., involving repeatedly solving the penalized problem $\mathcal{P}(\rho)$ (Problem 3 in the main manuscript). In this section we provide additional comments and technical results relating to this procedure.

First, suppose one solves $P(\rho)$ for a certain value ρ. Intuitively, solving again with a larger value ρ should place more priority on function f_2, which counts the flow through the penalized edges, and less on f_1. Conversely, decreasing ρ should emphasize f_1 at the expense of f_2. The iterative procedure implicitly relies on this fact. The following result confirms it formally.

Theorem C1. *Let $\mathbf{x}^{(1)}$ be a solution to $\mathcal{P}(\rho_1)$, and let $\mathbf{x}^{(2)}$ be a solution to $\mathcal{P}(\rho_2)$ where $\rho_1 < \rho_2$. Then either*

$$\mathbf{f}(\mathbf{x}^{(1)}) = \mathbf{f}(\mathbf{x}^{(2)}), \quad \textit{or} \quad f_1(\mathbf{x}^{(2)}) > f_1(\mathbf{x}^{(1)}) \textit{ and } f_2(\mathbf{x}^{(2)}) < f_2(\mathbf{x}^{(1)})$$

Second, the bounding relationship described by Theorem 3 of the main manuscript depends on knowledge of the extreme Pareto points $\underline{\mathbf{f}}$ and $\overline{\mathbf{f}}$. The following result shows that points $\underline{\mathbf{f}}$ and $\overline{\mathbf{f}}$ can be found by solving the penalized problem with sufficiently large and small penalties.

Theorem C2. *There exist numbers $\underline{\rho}, \overline{\rho} \in (0, \infty)$ with $\underline{\rho} < \overline{\rho}$ such that*

$$\mathbf{f}(\mathbf{x}) = \underline{\mathbf{f}} \textit{ if } \mathbf{x} \textit{ solves } \mathcal{P}(\rho) \textit{ for any } \rho \leq \underline{\rho} \qquad \textit{and} \qquad \mathbf{f}(\mathbf{x}) = \overline{\mathbf{f}} \textit{ if } \mathbf{x} \textit{ solves } \mathcal{P}(\rho) \textit{ for any } \rho \geq \overline{\rho}$$

Finally, Theorem 2 of the main manuscript does not rule out the possibility that certain solutions to $\mathcal{Q}(a)$ may not be solutions to $\mathcal{P}(\rho)$ for any choice of ρ. These solutions will not be recovered by the iterative procedure, which relies entirely on solving problem $\mathcal{P}$. However, the inability to recover all solutions to the directly constrained problem from the penalized problem is generally a minor concern. Practically speaking, in larger matching problems the bounds of Theorem 3 of the main manuscript often rule out the existence of Pareto optimal solutions materially different than those obtained by solving penalized problems. If desired, stronger bounds on the location of additional Pareto optimal points can be proved using linear relaxations of the directly constrained

problem.

D Generalization of key results to K-dimensional tradeoffs

In this section we discuss generalization of tradeoff analysis to the case where more than two objective functions are present.

D.1 Pareto optimal solutions for K quantitative goals

Suppose that in a matching problem we have K design goals instead of two, and that we can represent each of these goals by an objective function f_k:

$$f_k(\mathbf{x}) = \sum_{e \in \mathcal{E}} \psi_e^{(k)} x_e \qquad \text{for each } k = 1, \ldots, K \tag{5}$$

As before, denote the K-variate function $(f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_K(\mathbf{x}))$ by $\mathbf{f}(\mathbf{x})$. Our previous definition of Pareto optimality requires only minor modification to apply to the K-goal case.

Definition D1. *A solution* $\mathbf{x}$ *to a network flow problem (as given in Problem 1 of the main paper) is Pareto optimal for function* $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_K(\mathbf{x}))$ *if there exists no dominating* $\mathbf{x}'$*, i.e. a solution* $\mathbf{x}'$ *such that*

$$f_k(\mathbf{x}') \leq f_k(\mathbf{x}) \quad \textit{for all } k \in \{1, \ldots, K\} \textit{ and} \quad f_k(\mathbf{x}') < f_k(\mathbf{x}) \quad \textit{for some } k \in \{1, \ldots, K\}.$$

Next we consider the multidimensional extensions of the directly constrained and penalized problems. Consider the following optimization problem, for some vector $\mathbf{a} \in \mathbb{Z}^{K-1}$ with nonnegative elements.

Problem D1 (Directly Constrained Problem, K goals)**.**

$$\min f_1(\mathbf{x}) \qquad s.t. \qquad \mathbf{x} \in \mathcal{F}, \quad \{f_k(\mathbf{x}) \leq a_{k-1}\}_{k=2}^{K}, \quad \mathbf{x} \in \mathbb{Z}^{|\mathcal{E}|} \tag{6}$$

Extending our previous notation, call this problem $\mathcal{Q}(\mathbf{a})$ and represent its objective value at $\mathbf{x}$ by $Q(\mathbf{x}, \mathbf{a})$ and its optimal solution for a given integer vector $\mathbf{a}$ as $Q^*(a)$. Note that this problem is a special case of the "ϵ-constraint" method for producing Pareto optimal solutions described by Miettinen (2008) for more general settings.

Theorem D3. *Any Pareto optimal solution* $\mathbf{x}$ *for* $\mathbf{f}$ *must also be optimal for* $\mathcal{Q}(f_2(\mathbf{x}), \ldots, f_K(\mathbf{x}))$*, and all optimal solutions of* $\mathcal{Q}(f_2(\mathbf{x}), \ldots, f_K(\mathbf{x}))$ *are also Pareto optimal with identical values of* $\mathbf{f}$*.*

As in the $K = 2$ case, Theorem D3 tells us that if we solved $\mathcal{Q}(\mathbf{a})$ for all possible values of $\mathbf{a}$, we would obtain the full set of Pareto optimal points and see the complete tradeoff relationship between our design goals. Next we consider the generalized version of the penalized problem, for some vector $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{K-1})$ with $\boldsymbol{\rho} \geq 0$ (i.e. all vector elements ρ_i are nonnegative).

Problem D2 (Penalized Problem, K goals)**.**

$$\min f_1(\mathbf{x}) + \sum_{k=2}^{K} \rho_{k-1} f_k(\mathbf{x}) \qquad s.t. \qquad \mathbf{x} \in \mathcal{F} \tag{7}$$

Extending previous notation, represent Problem D2 by $\mathcal{P}(\boldsymbol{\rho})$. In addition, $P(\mathbf{x}, \boldsymbol{\rho})$ will represent the value of the objective function evaluated for particular values of $\mathbf{x}$ (assuming $\mathbf{x} \in \mathcal{F}$) and $\boldsymbol{\rho}$, and $P^*(\boldsymbol{\rho})$ will represent the optimal objective value for a given fixed $\boldsymbol{\rho}$. Problem $\mathcal{P}(\boldsymbol{\rho})$ remains a standard network flow problem. This solution approach is a special case of the "weighting method" described by Miettinen (2008) for more general multiobjective optimization settings.

Theorem D4. *Suppose* $\mathbf{x}^*$ *is an integer-valued optimal solution for the penalized problem* $\mathcal{P}(\boldsymbol{\rho})$*. Then* $\mathbf{x}^*$ *is Pareto optimal whenever* $\boldsymbol{\rho} > 0$*, and* $\mathbf{x}^*$ *is also optimal for the directly constrained problem* $\mathcal{Q}(\mathbf{a}^*)$ *where*

$$\mathbf{a}^* = (f_2(\mathbf{x}^*), \ldots, f_K(\mathbf{x}^*))$$

D.2 Results helpful in exploring Pareto set

Unfortunately other aspects of the set of Pareto optimal points do not generalize so conveniently to the multidimensional case. For instance, points in Pareto-optimal sets for $K > 2$ objectives no

longer have a strong ordering in multiple goals at once; ordering points according to the objective value achieved on f_1 will no longer necessarily order them by objectives achieved on any other single goal f_k.

Theorem 3 in the main manuscript and Theorems C1 - C2 also fail to generalize in an attractive way when $K > 2$. As in the $K = 2$ case, each additional Pareto-optimal point imposes bounds on the locations of additional Pareto points, but the bounds are far less restrictive for large K; in particular, a point in K-space can only be ruled out as a Pareto point if it strictly dominates or is strictly dominated by some Pareto point, i.e. if it lies in 2 of the 2^K quadrants obtained when viewing a particular Pareto point as the origin. Even when we take the union over many such forbidden sets, the set of possible Pareto points may remain quite large. Extreme high and low values for each individual objective k still exist, and it is possible to prove that these values will be achieved by setting the associated penalty ρ_{k-1} sufficiently large or small relative to an arbitrary fixed value for the other ρ_js. However, these values may no longer be achieved at unique Pareto points. In addition, if we increase (decrease) a single penalty ρ_{k-1} we can show that the value of objective f_k in the new optimal solution is either identical or strictly larger (smaller) than in the previous solution, and that the linear combination of the other goals, with weights given by the other ρ_{j-1}s, will also remain identical or change in the opposite direction; however, we have no guarantee for the direction in which any other objective f_j will change.

Different approaches to exploring Pareto optimal solutions for $K > 2$ goals in general settings are discussed in the multiobjective optimization literature, and may be adaptable to the case of matching in observational studies. In particular, this literature considers strategies for finding all extreme points of the Pareto optimal set and for using sampling techniques to compute an informative range of intermediate solutions (Ringuest 1992, ch. 9 and 11). Further work is needed to consider these methods in the specific context of matching and to evaluate their computational cost and practical value.

E Runtime analysis for matches in surgical data

Figure E1 describes the amount of time required to run each of the 14 matches based on surgical data that are described in Tables 3 and 4 of the main manuscript. Each match is computed by solving an appropriate version of the penalized problem (Problem 3 in the main manuscript). The results clearly indicate that for problems of this size and complexity, tradeoff analysis cannot be done interactively in a single sitting, and generating a subsample of the full Pareto-optimal curve by sequentially solving the penalized problem on a single machine requires substantially more computation time than would have been needed to compute any individual match. These requirements are not disqualifying factors for the method's relevance in practice, however. While many of the matches demand hours of computation time, they do not individually demand days or weeks, and since each match may be computed independently of others, they may be run simultaneously in parallel. In addition, the design stage of an observational study usually involves computing and comparing several different matches anyway; the machinery developed here makes it natural to group repeated matches into batches, which may ultimately increase workflow efficiency.

It is also clear from Figure E1 that not all matches are equal in their computational demands. In particular, most of the matches computed for the close pairing vs. balance tradeoff required 6-9 hours of computation time, while those computed for the balance vs. sample size tradeoff generally required 1.0-1.5 hours. The contrast between runtimes for match E and match F is particularly striking, since the matches produced by the two specifications are almost identical. This difference is not explained by the number of observations in the raw dataset, which is constant across all these matches. Instead, we hypothesize that the difference is driven by the maximum edge cost in the network. This number plays an important role in worst-case performance bounds for many minimum-cost network flow algorithms (Schrijver 2003, section 12.5), and in this problem it appears to have practical importance as well. As described in Section 4.3 of the main manuscript, the largest edge costs for the problems studying the close pairing vs. balance tradeoff are those for the exclusion edges and are equal to 2000. In contrast, as described in Section 4.4, the largest edge costs for the problems studying the balance vs. sample size tradeoff are the maximum of 449.86 and ρ, a number that is just below 450 for most of the matches F-N. In further support of

the connection between maximum edge cost and runtime, we note that when we ran the penalized problem for the balance vs. sample size tradeoff is run with a ρ-value and maximum edge cost an order of magnitude larger ($\rho = 4498.6$) it requires about 15 hours of computation time.

A detailed study of the role of maximum edge costs in explaining computational burden is beyond the scope of the current work. However, the potential for such a connection has practical implications for end users of the tradeoffs algorithm. For instance, in solving the penalized problem, it may be more efficient to start with small values of ρ and work one's way up into a region of interest rather than starting with large ρ-values.

While the size and general structure of the network flow problem representing the matching task remained fairly constant across the versions considered in this work, these aspects of the problem certainly influence computational costs. The work of Pimentel et al. (2015) suggests that major computational gains may arise from increasing the sparsity of the network problem, or in other words reducing the number of edges in the network relative to the number of nodes. For example, in the current design an IMG operation may be matched to any USMG operation in the same hospital whose propensity score obeys the caliper restriction; sparsity could be increased and computation accelerated by additionally requiring calipers on other continuous variables or permitting matches only between certain categories of operations.

F Proofs

Proof of Theorem 1. If the first claim fails, there is some solution $\mathbf{x}'$ such that

$$f_2(\mathbf{x}') \leq f_2(\mathbf{x}) \qquad \text{and} \qquad f_1(\mathbf{x}') < f_1(\mathbf{x})$$

which contradicts the Pareto optimality of $\mathbf{x}$. For the second claim, note that all optimal solutions $\mathbf{x}'' \neq \mathbf{x}$ for $\mathcal{Q}(f_2(\mathbf{x}))$ satisfy $f_1(\mathbf{x}'') = f_1(\mathbf{x})$, or either $\mathbf{x}$ or $\mathbf{x}''$ would not be optimal. Also $f_2(\mathbf{x}'') \leq f_2(\mathbf{x})$, by feasibility for $\mathcal{Q}(f_2(\mathbf{x}))$, and $f_2(\mathbf{x}'') \geq f_2(\mathbf{x})$, by Pareto optimality of $\mathbf{x}$. Thus $\mathbf{f}(\mathbf{x}'') = \mathbf{f}(\mathbf{x})$ and $\mathbf{x}''$ is also Pareto optimal. □

Proof of Theorem 2. First prove Pareto optimality. Suppose that $\rho > 0$ and that $\mathbf{x}'$ dominates $\mathbf{x}^*$ with respect to objective functions f_1 and f_2. Then $P(\mathbf{x}', \rho') < P(\mathbf{x}^*, \rho')$ for any $\rho' > 0$ (since

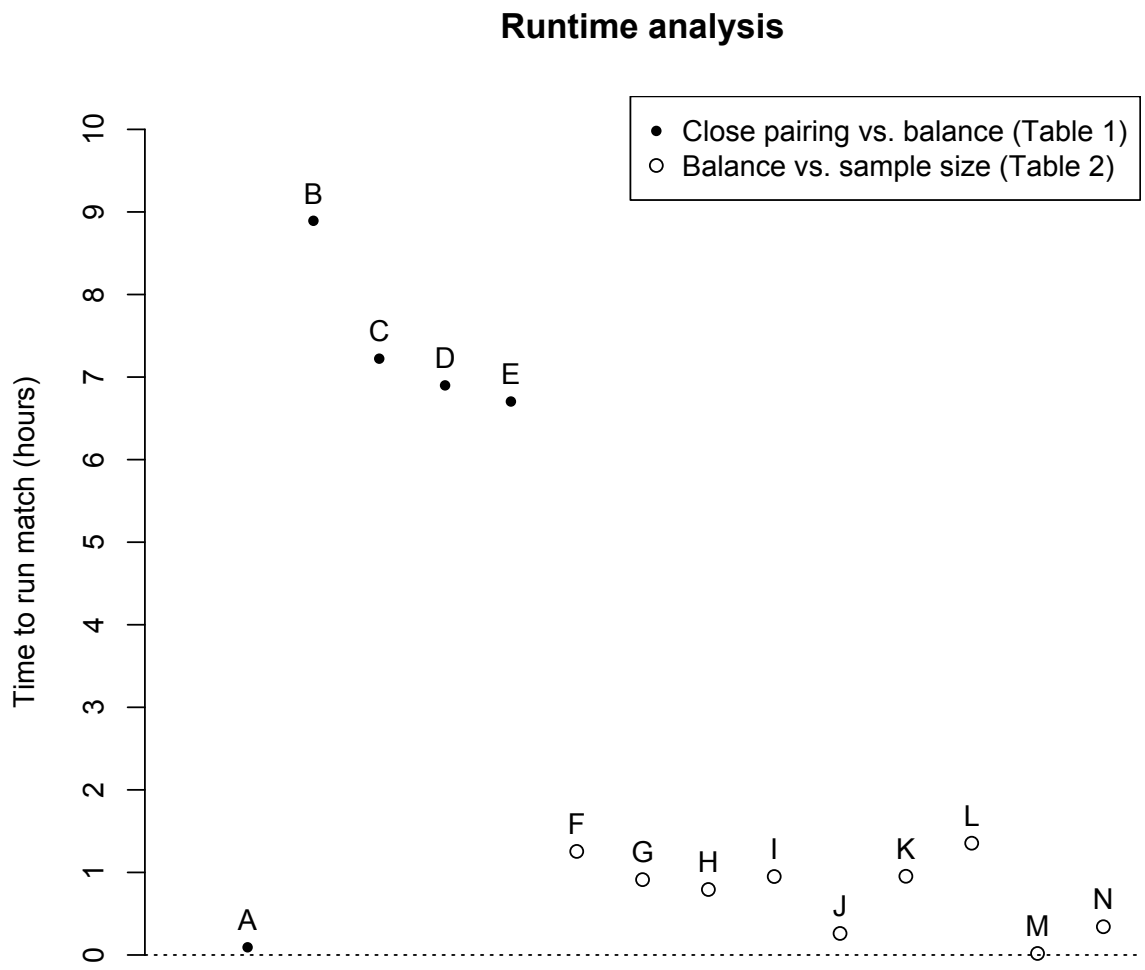


Figure E1: Runtimes (in hours) for matches displayed in Tables 3 and 4 of the main manuscript, measured on an iMac Pro with a 3.2 GHz Intel Xeon Wu processor and 32 GB of RAM. Each match was produced by solving a version of the penalized problem (Problem 3 in the main manuscript): matches A-E were computed as part of the close pairing vs. balance tradeoff analysis described in Section 4.3, while matches F-N were computed as part of the balance vs. sample size tradeoff described in Section 4.4.

the objective function $P(\mathbf{x}, \rho)$ is a linear combination of the two objectives with nonzero weights), a contradiction since $\mathbf{x}^*$ is optimal for $P(\rho)$. Therefore no dominating solution exists and $\mathbf{x}^*$ is Pareto optimal.

For the second claim, since $\mathbf{x}^*$ is feasible for $\mathcal{P}(\rho)$ with $f_2(\mathbf{x}^*) = a^*$, it is also feasible for $\mathcal{Q}(a^*)$. Choose any $\mathbf{x}' \neq \mathbf{x}^*$ such that $\mathbf{x}'$ is feasible for $\mathcal{Q}(a^*)$. Solution $\mathbf{x}'$ is feasible for $\mathcal{P}(\rho)$ (since it necessarily satisfies $\mathbf{x}' \in \mathcal{F}$). By optimality of $\mathbf{x}^*$ for $\mathcal{P}(\rho)$, $P(\mathbf{x}^*, \rho) \leq P(\mathbf{x}', \rho)$, i.e.

$$f_1(\mathbf{x}^*) + \rho f_2(\mathbf{x}^*) \leq f_1(\mathbf{x}') + \rho f_2(\mathbf{x}')$$

By feasibility of $\mathbf{x}'$, $f_2(\mathbf{x}') \leq a^* = f_2(\mathbf{x}^*)$, $f_1(\mathbf{x}^*) \leq f_1(\mathbf{x}')$. Since f_1 is the objective function for $\mathcal{Q}(a^*)$ and $\mathbf{x}'$ is freely chosen from its feasible region, $\mathbf{x}^*$ is optimal for $\mathcal{Q}(a^*)$. □

Proof of Theorem C1. Since $\mathbf{x}^{(1)}$ is optimal for $\mathcal{P}(\rho_1)$ and $\mathbf{x}^{(2)}$ is optimal for $\mathcal{P}(\rho_2)$, we know $P(\mathbf{x}^{(2)}, \rho_2) \leq P(\mathbf{x}^{(1)}, \rho_2)$ and $P(\mathbf{x}^{(1)}, \rho_1) \leq P(\mathbf{x}^{(2)}, \rho_1)$. These facts in turn imply

$$P(\mathbf{x}^{(2)}, \rho_2) - P(\mathbf{x}^{(2)}, \rho_1) \leq P(\mathbf{x}^{(1)}, \rho_2) - P(\mathbf{x}^{(1)}, \rho_1)$$

By canceling $f_1(\mathbf{x})$ terms and dividing both sides by $(\rho_2 - \rho_1)$, this implies $f_2(\mathbf{x}^{(2)}) \leq f_2(\mathbf{x}^{(1)})$. In addition, by expanding both sides of the inequality $P(\mathbf{x}^{(1)}, \rho_1) \leq P(\mathbf{x}^{(2)}, \rho_1)$ and collecting ρ_1 terms we can show the following:

$$\rho_1 \left[f_2\left(\mathbf{x}^{(1)}\right) - f_2\left(\mathbf{x}^{(2)}\right) \right] \leq f_1\left(\mathbf{x}^{(2)}\right) - f_1\left(\mathbf{x}^{(1)}\right)$$

The left-hand side is nonnegative since $f_2(\mathbf{x}^{(2)}) \leq f_2(\mathbf{x}^{(1)})$, which means $f_1(\mathbf{x}^{(1)}) \leq f_2(\mathbf{x}^{(1)})$ also. If the former inequality is strict, the latter must be also, while if $f_2(\mathbf{x}^{(2)}) = f_2(\mathbf{x}^{(1)})$, then $f_1(\mathbf{x}^{(1)}) = f_2(\mathbf{x}^{(1)})$ also, or $P(\mathbf{x}^{(1)}, \rho_2) < P(\mathbf{x}^{(2)}, \rho_2)$ which contradicts an earlier fact. □

Proof of Theorem C2. Throughout we use the notation of Section A.1. For some ordering $e_1, \ldots, e_{|\mathcal{E}|}$ of the edges in $\mathcal{E}$, define $\mathcal{V} \subset \mathbb{Z}^{|\mathcal{E}|}$ by $\mathcal{V} = \{-d_{e_1}, -d_{e_1}+1, \ldots, d_{e_1}-1, d_{e_1}\} \times \{-d_{e_2}, \ldots, d_{e_2}\} \times \ldots, \times \{-d_{e_{|\mathcal{E}|}}, \ldots, d_{e_{|\mathcal{E}|}}\}$ and represent the cost coefficients in f_1 and f_2 by vectors $\boldsymbol{\psi} = (\psi_{e_1}, \ldots, \psi_{e_{|\mathcal{E}|}})$ and $\boldsymbol{\gamma} = (\gamma_{e_1}, \ldots, \gamma_{e_{|\mathcal{E}|}})$. Let $\underline{\rho} = \left(\min_{\mathbf{v} \in \mathcal{V}, \mathbf{v}^T\boldsymbol{\psi} > 0} \mathbf{v}^T\boldsymbol{\psi} - \epsilon\right) / \left(\sum_{e \in \mathcal{E}} \gamma_e d_e\right)$ for some $\epsilon > 0$ such that $\underline{\rho}$ is positive (since $\mathcal{V}$ is a finite set some such ϵ always exists). Let $\underline{\mathbf{x}}$ be the solution to $\mathcal{P}(\underline{\rho})$. If a distinct network flow solution $\mathbf{x} \neq \underline{\mathbf{x}}$ satisfies $f_1(\mathbf{x}) < f_1(\underline{\mathbf{x}})$, then $P(\underline{\rho}, \mathbf{x}) < P(\underline{\rho}, \underline{\mathbf{x}})$ since $f_1(\underline{\mathbf{x}}) - f_1(\mathbf{x}) \geq \min_{\mathbf{v} \in \mathcal{V}, \mathbf{v}^T\boldsymbol{\psi} > 0} \mathbf{v}^T\boldsymbol{\psi} > \underline{\rho} \sum_{e \in \mathcal{E}} \gamma_e d_e \geq \underline{\rho}(f_2(\mathbf{x}) - f_2\underline{\mathbf{x}}))$, but this is a con-

tradiction, so $\mathbf{f}(\underline{\mathbf{x}}) = \underline{\mathbf{f}}$ (with Pareto optimality following from Theorem 2). Similarly, let $\overline{\rho} = \left(\sum_{e\in\mathcal{E}} \psi_e d_e\right) / \left(\min_{\mathbf{v}\in\mathcal{V},\mathbf{v}^T\gamma>0} \mathbf{v}^T\gamma\right) + \epsilon$ and let $\mathbf{x}$ be the solution to $\mathcal{P}(\overline{\rho})$. If $\mathbf{x} \neq \overline{\mathbf{x}}$ satisfies $f_2(\mathbf{x}) < f_2(\overline{\mathbf{x}})$, then $P(\overline{\rho}, \mathbf{x}) < P(\overline{\rho}, \overline{\mathbf{x}})$ since $\overline{\rho}(f_2(\overline{\mathbf{x}}) - f_2(\mathbf{x})) > \sum_{e\in\mathcal{E}} \psi_e d_e \geq f_1(\mathbf{x}) - f_1(\overline{\mathbf{x}})$. This is a contradiction so by Theorem 2, $\mathbf{f}(\overline{\mathbf{x}}) = \overline{\mathbf{f}}$. □

Proof of Theorem D3. If the first claim fails, there is some solution $\mathbf{x}'$ such that

$$f_1(\mathbf{x}') < f_1(\mathbf{x}) \qquad \text{and} \qquad f_k(\mathbf{x}') \leq f_k(\mathbf{x}) \quad \text{for all } k = 2, \ldots, K$$

which contradicts the Pareto optimality of $\mathbf{x}$. For the second claim, note that all optimal solutions $\mathbf{x}'' \neq \mathbf{x}$ for $\mathcal{Q}(f_2(\mathbf{x}), \ldots, f_K(\mathbf{x}))$ satisfy $f_1(\mathbf{x}'') = f_1(\mathbf{x})$, or either $\mathbf{x}$ or $\mathbf{x}''$ would not be optimal. Also $f_k(\mathbf{x}'') \leq f_k(\mathbf{x})$ for all $k \in \{2, \ldots, K\}$, by feasibility for $\mathcal{Q}(f_2(\mathbf{x}), \ldots, f_K(\mathbf{x}))$, and $f_k(\mathbf{x}'') \geq f_l(\mathbf{x})$ for all $k \in \{2, \ldots, K\}$, by Pareto optimality of $\mathbf{x}$. Thus $\mathbf{f}(\mathbf{x}'') = \mathbf{f}(\mathbf{x})$ and $\mathbf{x}''$ is also Pareto optimal. □

Proof of Theorem D4. First prove Pareto optimality. Suppose that $\boldsymbol{\rho} > 0$ and that $\mathbf{x}'$ dominates $\mathbf{x}^*$ with respect to the objective functions in (5). Then $P(\mathbf{x}', \boldsymbol{\rho}') < P(\mathbf{x}^*, \boldsymbol{\rho}')$ for any $\boldsymbol{\rho}' > 0$ (since the objective function $P(\mathbf{x}, \boldsymbol{\rho}')$ is a linear combination of the K objectives with nonzero weights), a contradiction since $\mathbf{x}^*$ is optimal for $P(\boldsymbol{\rho})$. Therefore no dominating solution exists and $\mathbf{x}^*$ is Pareto optimal.

For the second claim, since $\mathbf{x}^*$ is feasible for $\mathcal{P}(\boldsymbol{\rho})$ with $(f_2(\mathbf{x}^*), \ldots, f_K(\mathbf{x}^*)) = \mathbf{a}^*$, it is also feasible for $\mathcal{Q}(\mathbf{a}^*)$. Choose any $\mathbf{x}' \neq \mathbf{x}^*$ such that $\mathbf{x}'$ is feasible for $\mathcal{Q}(\mathbf{a}^*)$. Solution $\mathbf{x}'$ is feasible for $\mathcal{P}(\boldsymbol{\rho})$ (since it necessarily satisfies $\mathbf{x}' \in \mathcal{F}$). By optimality of $\mathbf{x}^*$ for $\mathcal{P}(\boldsymbol{\rho})$, $P(\mathbf{x}^*, \boldsymbol{\rho}) \leq P(\mathbf{x}', \boldsymbol{\rho})$, i.e.

$$f_1(\mathbf{x}^*) + \sum_{k=2}^{K} \rho_{k-1} f_k(\mathbf{x}^*) \quad \leq \quad f_1(\mathbf{x}') + \sum_{k=2}^{K} \rho_{k=1} f_k(\mathbf{x}') \tag{8}$$

By feasibility of $\mathbf{x}'$, $f_k(\mathbf{x}') \leq a_k^* = f_k(\mathbf{x}^*)$. This in turn implies

$$\sum_{k=2}^{K} \rho_{k-1} f_k(\mathbf{x}') \quad \leq \quad \sum_{k=2}^{K} \rho_{k-1} f_k(\mathbf{x}^*)$$

In order for this statement and statement (8) to hold simultaneously, it must also be true that $f_1(\mathbf{x}^*) \leq f_1(\mathbf{x}')$. Since f_1 is the objective function for $\mathcal{Q}(\mathbf{a}^*)$ and $\mathbf{x}'$ is freely chosen from its feasible region, $\mathbf{x}^*$ is optimal for $\mathcal{Q}(\mathbf{a}^*)$. □

References

Bertsekas, D. P. (1998), *Network optimization: continuous and discrete models*, Belmont, MA: Athena Scientific.

Miettinen, K. (2008), "Introduction to Multiobjective Optimization: Noninteractive Approaches," in *Multiobjective Optimization: Interactive and Evolutionary Approaches*, eds. Branke, J., Deb, K., Miettinen, K., and Słowiński, R., Berlin Heidelberg: Springer-Verlag, chap. 1, pp. 1–26.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), "Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons," *Journal of the American Statistical Association*, 110, 515–527.

Ringuest, J. L. (1992), *Multiobjective optimization: behavioral and computational considerations*, Norwell, MA: Kluwer Academic Publishers.

Schrijver, A. (2003), *Combinatorial optimization : polyhedra and efficiency.*, Algorithms and combinatorics: 24, Berlin ; New York : Springer, c2003.

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.

Zaheer, S., Pimentel, S. D., Simmons, K. D., Kuo, L. E., Datta, J., Williams, N., Fraker, D. L., and Kelz, R. R. (2017), "Comparing International and United States Undergraduate Medical Education and Surgical Outcomes Using a Refined Balance Matching Methodology," *Annals of Surgery*, 265, 916–922.