Statistics and Public Policy

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uspp20

Optimal Matching for Observational Studies That Integrate Quantitative and Qualitative Research

Ruoqi Yu, Dylan S. Small, David Harding, José Aveldanes & Paul R. Rosenbaum

To cite this article: Ruoqi Yu, Dylan S. Small, David Harding, José Aveldanes & Paul R. Rosenbaum (2021) Optimal Matching for Observational Studies That Integrate Quantitative and Qualitative Research, Statistics and Public Policy, 8:1, 42-52, DOI: 10.1080/2330443X.2021.1919260

To link to this article: https://doi.org/10.1080/2330443X.2021.1919260

Published online: 17 Jun 2021.

STATISTICS AND PUBLIC POLICY
2021, VOL. 8, NO. 1, 42–52
https://doi.org/10.1080/2330443X.2021.1919260

Optimal Matching for Observational Studies That Integrate Quantitative and Qualitative Research

Ruoqi Yu[a], Dylan S. Small[a], David Harding[b], José Aveldanes[b], and Paul R. Rosenbaum[a]

[a]Wharton School, University of Pennsylvania, Philadelphia, PA; [b]Department of Sociology, University of California at Berkeley, Berkeley, CA

ABSTRACT

A quantitative study of treatment effects may form many matched pairs of a treated subject and an untreated control who look similar in terms of covariates measured prior to treatment. When treatments are not randomly assigned, one inevitable concern is that individuals who look similar in measured covariates may be dissimilar in unmeasured covariates. Another concern is that quantitative measures may be misinterpreted by investigators in the absence of context that is not recorded in quantitative data. When text information is automatically coded to form quantitative measures, examination of the narrative context can reveal the limitations of initial coding efforts. An existing proposal entails a narrative description of a subset of matched pairs, hoping in a subset of pairs to observe quite a bit more of what was not quantitatively measured or automatically encoded. A subset of pairs cannot rule out subtle biases that materially affect analyses of many pairs, but perhaps a subset of pairs can inform discussion of such biases, perhaps leading to a reinterpretation of quantitative data, or perhaps raising new considerations and perspectives. The large literature on qualitative research contends that open-ended, narrative descriptions of a subset of people can be informative. Here, we discuss and apply a form of optimal matching that supports such an integrated, quantitative-plus-qualitative study. The optimal match provides many closely matched pairs plus a subset of exceptionally close pairs suitable for narrative interpretation. We illustrate the matching technique using data from a recent study of police responses to domestic violence in Philadelphia, where the police report includes both quantitative and narrative information.

ARTICLE HISTORY
Received April 2020
Accepted April 2021

KEYWORDS
Causal inference; Narrative description; Optimal matching; Threshold algorithms

1. Strengthening Causal Inference by Integrating Methodologies

1.1. Effects Caused by Treatments

The effect of a treatment on an individual is a comparison of two potential responses, the individual's response if assigned to treatment and the individual's response to control; see Rubin (1974). This causal effect is not observed: we see the response of an individual under treatment or control, not both, so we do not see the effect of the treatment. Inference about the effects caused by treatments is comparatively straightforward in randomized trials that use random numbers to divide a finite population into treated and control groups; see Fisher (1935). Inference about causal effects is much more difficult when treatments are not randomly assigned, because in the absence of randomization treated and control groups may differ prior to treatment in terms of covariates that were not observed.

Mervyn Susser wrote that if evidence of cause and effect suffers from certain limitations that are not a consequence of a limited sample size, then we should not try to dispel these limitations by increasing the sample size or exactly replicating the study; rather, we should seek evidence that suffers from different limitations. He wrote

> The epidemiologist … seeks … consistency of results in a variety of repeated tests. … Consistency is present if the result is not dislodged in the face of diversity in times, places, circumstances, and people, as well as of research design (Susser 1987, p. 88) … The strength of the argument rests on the fact that diverse approaches produce similar results (Susser 1973, pp. 148).

For related discussion, see Rosenbaum (2001), Rosenbaum (2021), Lawlor, Tilling, and Davey Smith (2016), and Munafò and Davey-Smith (2018).

A mixed-method approach combines qualitative and quantitative research, each having certain strengths. The strength of qualitative methods such as in-depth interviewing lies in assessing what traditional surveys cannot: how respondents understand events in their lives, how they use language, the complexity of respondents' beliefs and attitudes, and extended accounts of their behaviors (Riessman 2012). Furthermore, qualitative interview studies provide coherence, richness, and context for understanding the complexity of respondents' experiences that quantitative surveys cannot—by design—provide (Weiss 1994). On the other hand, quantitative methods enable hypothesis

CONTACT Dylan S. Small dsmall@wharton.upenn.edu Department of Statistics, The Wharton School, University of Pennsylvania, Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340.

testing, representativeness and generalizability, and the testing of causal mechanisms (Frankel 1983; Aneshensel 2013).

The advantages of using both qualitative and quantitative methods stem from the complementary strengths of the two types of methods (Creswell and Plano Clark 2007; Axinn and Pearce 2006). For example, qualitative data on how research participants understand their experiences can be used to help interpret quantitative results (Roth and Mehta 2002), similar results from both methods can provide greater confidence in a given study's conclusions (Stenger et al. 2014), or qualitative data can inform the measurement of quantitative data, such as when in-depth interview data are used to inform the development of close-ended survey questions (Crede and Berrego 2013). Despite the advantages of mixed-methods research, Creswell and Plano Clark (2007) cautions that careful consideration of research design for effective integration of qualitative and quantitative methods is essential. Although we believe that the integration of qualitative and quantitative research is useful, our goal here is not to argue for that claim, but rather to illustrate a new matching technique that facilitates such an integration.

1.2. Application: Confiscating Guns During Incidents of Domestic Violence

The application in Section 3 concerns a study of the effects of confiscating guns present during incidents of domestic violence in Philadelphia; see Small, Sorenson, and Berk (2019). In this application, the police officer fills in a structured form for each incident and provides a text note containing further information. The structured form provides quantitative information, but the text is also interesting and useful. The texts are vivid in a way the forms are not. As described later, Table 4 contains ten such texts. However, there are hundreds of such texts in the application, and in many medical applications there may be tens of thousands of texts.

How can many texts inform a quantitative analysis? Reading the texts, we realize they reflect the natural concerns of police officers, not always the concerns of researchers. The texts carefully distinguish what the police saw as opposed to what they were told. The texts contain details about the gun that might serve to identify it. Threats that mention the gun are carefully recorded and distinguished from threats that do not mention the gun. Perhaps quantitative data can be extracted from texts, but an algorithm that viewed the texts as a bag-of-words might misunderstand the context in which those words appear. In a city not without racial tensions, the word "black" appears in many texts, but it almost invariably distinguishes a black pistol from a silver pistol, presumably for later identification. The word "shot" often appears, although in different texts it is people or car tires or front doors that are shot. How best can we make use of texts without reading them all?

1.3. Matching Many People With Narrative Description of a Subset of Pairs

Among methods of adjustment for covariates, matching leaves people intact as people. Intact people can be met, interviewed, recorded, quoted, or described in narrative terms. Text describing intact people may be read as text. Do people who look comparable in measured quantitative covariates still seem broadly comparable in narrative descriptions? Do narrative descriptions contain relevant aspects not contained in the quantitative data, perhaps aspects that can be encoded to form enhanced quantitative data? Do quantitative measures correctly represent situations and events in light of narrative information that may also be available? Did efforts to encode text data in quantitative form fail to capture important aspects of the text?

Rosenbaum and Silber (2001) argued that it is natural to strengthen a matched quantitative comparison of many pairs by adding qualitative description of a subset of matched pairs. Their full study matched 830 patients who had died following surgery to patients who survived, examining these case-control pairs. The matching used quantitative data from medical chart abstraction in an effort to compare case and control patients similar prior to hospital admission. In a pilot study, they compared, for 38 pairs: (i) quantitative data obtained by chart abstraction, and (ii) a detailed reading of the medical charts for these pairs. This comparison led to improved use of the quantitative data—revised definitions of cancer and congestive heart failure—and a new match incorporating these revisions. The side-by-side reading of charts for matched pairs was critical: unremarkable, accurately abstracted charts were revealed to be poorly matched only by seeing that the paired charts described people who differed in notable respects. Prior to revision, the cancer definitions were correct, but comparisons of distinct individuals regarded as similar by those definitions revealed that the original definitions were inadequate to capture important distinctions in a complex disease. For instance, consider a binary variable, "history of melanoma." We might see that as a correct description of one patient who had a skin malignancy surgically removed without subsequent evidence of disease. We might see that also as a correct description of another patient who is dying of metastatic, stage IV melanoma. However, if we had paired these patients together, we would immediately see that "history of melanoma" fails to capture the enormous difference between these two patients, so a revision in the variable is required. We would see the inadequacy of "history of melanoma" as initially defined only if we compared two people who were the same in terms of this poorly defined variable, while also seeing that they were very different in the medical chart.

For a related discussion of purposeful, comparative selection of cases for qualitative comparisons, see Seawright and Gerring (2008) and Tarrow (2010). It is well-understood that case selection in small-N studies cannot rely on the statistical properties of random sampling that provide generalizability and estimation of sampling error, nor can it rely on comparisons to fully control for confounding influences, as one would find in randomized controlled experiments or matching study designs. Instead, small-N qualitative studies tend to rely on "purposive" selection of cases for comparisons that can help researchers make important, even if ungeneralizable, discoveries (Lieberson 1991). Such strategies focus on comparisons across key participant categories for the research question at hand, or sample for range (identify subcategories of the group under study and ensure that a given number of people in that

category are interviewed) to ensure that theoretically important cases are included in the data (George and Bennett 2005; Small 2009). Because much qualitative work seeks to inductively identify and examine processes or mechanisms rather than providing a generalizable description, case selection may also prioritize cases where various hypothesized mechanisms are likely to be present and amenable to study (Lamont and White 2005).

1.4. What Can and What Cannot Be Learned From Close Pairs?

What can and what cannot be learned from qualitative, narrative or thick description of a moderate number of closely matched pairs? Perhaps, there are between five and fifty pairs, depending upon the context. Why, given the nature of modern matching techniques, is it important to compare pairs that are exceptionally close in terms of the quantitative data?

Subtle, though perhaps important, differences cannot be detected with small sample sizes. A reduction in 30-day mortality following surgery from 5% to 4% would be important, but to have 80% power to detect such a difference between two independent binomial proportions requires a sample of 6745 patients in each of two groups. Therefore, it is important to keep the entire matched sample.

Close examination of a small number of pairs can reveal that something that should never happen often happens. It is easy to misinterpret quantitative data, but our goal is to avoid misinterpretation. Misinterpretations of quantitative data are often discovered by attempts to give a coherent narrative account of a few intact people and comparing that narrative with the representation of these people in quantitative data. What appears to be a missing answer to a question may be nothing of the sort, because particular answers to several screening questions prevent the asking of another question. Positive biopsy results are preserved in elaborate and expensive detail one data file, but negative results are not preserved, so the existence of a negative result must be deduced by combining that file with another file that records bills for biopsies. There are mistakes you can make when interpreting a column of numbers that you could not possibly make when describing an intact person.

A subset of closely matched pairs can reveal what unrelated individuals cannot. Rosenbaum and Silber (2001) found that an initial pairing of surgical deaths and surgical survivors had made perfectly correct statements about deaths and about survivors as individuals, but that these correct statements were too coarse to be useful. The quantitative statement was confirmed as a true statement about an individual when compared with the medical chart. Stare at patients one at a time and no problem is evident. When a death and a survivor were compared, however, their true similarity in quantitative data revealed the quantitative data to be making statements that were true but too coarse. Typically, the patient who died had more severe health problems prior to surgery than did the ostensively similar patient who survived. This prompted informed efforts to make subtler distinctions among patients in the quantitative data.

2. Optimal Matching for Quantitative Plus Qualitative Comparisons

2.1. Match Many People, and Produce a Subset of Very Close Pairs for Qualitative Comparison

How should we match if we want a quantitative comparison of many, say T, pairs, and a narrative description of a subset of $F \geq 1$ pairs? For instance, T might be hundreds or thousands or tens of thousands of pairs, while F might be thirty pairs. Because considerable time and effort will be expended upon the F pairs, we want this subset of pairs to be exceptionally close in terms of pretreatment covariates. We want to judge the adequacy of the observed covariates using pairs that are exceptionally close in terms of the observed covariates.

In a matched sample, covariates are balanced if the distribution of age, say, is almost the same among the T treated individuals and their T matched controls, and the same is true of the distribution of income, and so on. In contrast, a single pair is close if the two people in the pair have nearly the same age, the same income, and so on. Modern matching methods often use propensity scores, fine balance constraints, and similar devices to balance many covariates in treated and control groups, but they do this with pairs that need not be close person by person. For instance, if there are 30 binary covariates, it is often possible to balance them, but these define 2^{30} or about a billion categories of people and it is often not possible to find pairs that are uniformly close. However, it may be possible to find a subset of such pairs.

Why focus on a subset of close pairs, rather than a representative subset of pairs? Balancing covariates, with propensity scores or fine balance constraints, removes imbalances in measured covariates in the study as a whole, but this control for measured covariates is not evident in every pair. Covariate balance is easy to check quantitatively in the study as a whole, so qualitative research is not needed for that purpose. In contrast, it is not easy to check quantitatively whether the available observed covariates constitute an adequate and accurate representation of the situation under study. When using qualitative comparisons in such checks, it is useful to eliminate differences in the observed covariates, so problems that quantitative research can reliably address—covariate imbalance—do not obscure information that only qualitative comparisons provide.

A theorem about propensity scores says that if it suffices to match for certain observed covariates, say a hundred observed covariates, then it suffices to match for a single covariate, namely the propensity score that conditions on these covariates (Rosenbaum and Rubin 1983, Theor. 2 and 3). This is true in a specific sense: two matched individuals with the same propensity score may be very different in terms of the hundred covariates, but the differences are not systematic, so they tend to balance out over many matched pairs. From this specific quantitative perspective, two methods—namely matching closely for the propensity score or matching closely for each of the hundred covariates—are different, but they are not extremely different, and this is fortunate because what the second method requires cannot be done. When thinking about the comparability of two matched samples, this theorem is highly relevant: it says that the systematic bias in matched samples due to these hundred measured

covariates can be removed by pairing for the propensity score. In matched samples, control for a hundred covariates is secured by controlling for one covariate, leaving the rest to balance out. It is, however, not helpful to qualitatively compare two individuals in a pair who differ greatly on a hundred covariates, with the mere reassurance that these enormous differences will balance out in other pairs that are not subjected to qualitative evaluation. For qualitative research, pairs close on covariates are helpful, but if there are many covariates, close pairs are exceedingly rare.

The proposed algorithm integrates quantitative matching and examination of qualitative information, often information in narrative form. The match balances many covariates and is close on a few key covariates, but it contains as an integral component a subset of pairs that are exceptionally close in quantitative terms, that is, in terms of most measured covariates. The quantitative analysis claims pairs in this subset are comparable. Does the qualitative comparison of this subset of pairs concur? Does narrative description of exceptionally close pairs support or raise concerns about the comparability of the individuals being compared? Does it raise concerns that some covariates are misinterpreted and require revised definitions? When text data exists for everyone, does reading the narrative as a narrative—rather than feeding the text into a machine learning algorithm—suggest ways to encode certain features of the text so that it may become part of the quantitative analysis?

Why insist that qualitative research focus on a subset of the pairs in the quantitative match? Integration of qualitative and quantitative research means each method provides a check on the other, and to do that they must take different approaches to looking at the same worldly situation, checking the same assumptions about that situation. The method we propose produces both a match of many pairs balancing covariates for quantitative analysis, plus a subset of extremely close pairs for qualitative examination. The extremely close pairs were eligible as pairs in the quantitative matched sample—those close pairs were built using the same covariates, with no additional information that the qualitative researcher might have had. If the qualitative researcher picked pairs who seemed especially cooperative or especially interesting or especially well-suited for comparison or especially good at illustrating some point that the qualitative researcher wishes to make, then that may or may not yield valuable qualitative research, but it limits its value as a check or critique of the quantitative half of the investigation.

2.2. Matching is an Aspect of Design, Not of Analysis

Matching is a feature of the design of an observational study, completed prior to examination or use of outcomes; then, when outcomes are examined, a planned primary analysis is conducted Imbed and Rubin (2007). In the example in §3, both the quantitative and qualitative aspects of design are based on a police form collected at baseline, with no access to later outcome data.

Separating design from analysis means that an investigator cannot shop among many analyses for a preferred conclusion, thereby invalidating the properties of statistical procedures. However, because matching is completed without access to outcome information, no bias is introduced into the subsequent quantitative analysis if the design steps outlined in Section 2.1 are repeatedly improved in response to repeated qualitative comparisons of pretreatment characteristics of a subset of pairs.

We consider matching without replacement in this paper. Matching without replacement has the feature of being a transparent design that is analogous to a blocked experiment. Another approach to matching is matching with replacement (Abadie and Imbens 2006; Stuart 2010; Imbens and Rubin, 2015, chap. 18.9). Rosenbaum (2017a, sec. 1.2) discusses comparisons of matching without replacement vs. with replacement.

2.3. Notation

Our goal is to find a match for T pairs that balances covariates for which there are a smaller number F of exceptionally close pairs for interviews. The method uses a threshold technique from Rosenbaum (2017a), which develops an idea of Garfinkel (1971). The method is available in an R package `thickmatch`.

There are T treated individuals, $\tau_1, \ldots, \tau_T$, and $C \geq T$ potential controls, $\gamma_1, \ldots, \gamma_C$. Each treated individual will be paired with a different control, yielding T matched sets consisting of $2T$ distinct individuals. There is a distance $\delta_{tc} \geq 0$ between treated individual τ_t and potential control γ_c, based on their observed covariates, so $\delta_{tc} = 0$ if these two people are identical on all of the observed covariates. The exact form of the distance does not matter in the discussion that follows. Commonly, this distance is some form of Mahalanobis distance focused on important covariates with a caliper on the propensity score computed from all covariates; it may incorporate other considerations (e.g., calipers on individual covariates). An optimal or minimum-distance match $\mu(\cdot)$ assigns each τ_t to a different control, specifically to control γ_c with $c = \mu(\tau_t)$, forming T pairs, $\{\tau_t, \gamma_{\mu(\tau_t)}\}$, in such a way that the sum of the T within-pair distances, $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$, is minimized. Various algorithms (e.g., Bertsekas 1981) implemented in various `R` packages solve this problem. There are $C!/(C-T)!$ possible pairings $\mu(\cdot)$—an enormous number—but algorithms exist that can find the best $\mu(\cdot)$ in $O(C^3)$ arithmetic steps; so, it is entirely practical. The optimal match we constructed in the example took a fraction of one second.

Commonly, there are many covariates, including continuous covariates, with the consequence that the $T \times C$ distances δ_{tc} are all distinct. For instance, if one of the covariates had a Normal distribution and the Mahalanobis distance was used, then the probability that at least two δ_{tc} take the same numerical value is zero. The procedure is slightly easier to describe if the δ_{tc} are untied, so we assume this, discussing the minor consequences of a few ties in §2.6.

Let $\mathcal{I} \subseteq \{\tau_1, \ldots, \tau_T\}$ be a subset of the treated individuals, perhaps most commonly all of them, $\mathcal{I} = \{\tau_1, \ldots, \tau_T\}$. For narrative description, we want F treated individuals in $\mathcal{I}$ to be exceptionally well matched, with small distances to their matched controls, and subject to doing that, we want to minimize the total of all T within-pair distances, $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$. This means that the total distance $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$ will typically be a little larger than the minimum in the previous paragraph, but these F pairs will be very close. This can give us a large matched sample

for quantitative analysis and a subset of F pairs for qualitative interpretation.

Any match $\mu(\cdot)$, good or bad, pairs the individuals in $\mathcal{I}$ to distinct controls, and we will interview or otherwise narratively describe the $2F$ individuals in these F closest pairs. So each treated individual $\tau_t \in \mathcal{I}$ is assigned a different control γ_c with $c = \mu(\tau_t)$, and this matched pair, (τ_t, γ_c), has covariate distance $\delta_{t,\mu(\tau_t)} = \delta_{t,c}$. We may sort these within-pair distances into increasing order, and we can look at the Fth largest within-pair distance in this order. So, for any match $\mu(\cdot)$, good or bad, let Δ_μ be the Fth-order statistic of the within-pair distances, $\delta_{t,\mu(\tau_t)}$, for the treated individuals in $\tau_t \in \mathcal{I}$, if we use match $\mu(\cdot)$. If $\mathcal{I} = \{\tau_1, \ldots, \tau_T\}$ and $F = 30$, then Δ_μ is the largest of the $F = 30$ smallest distances in the match, the largest distance for the $F = 30$ pairs that will be described. We want a match $\mu(\cdot)$ that minimizes this quantile Δ_μ, and among all matches that do that, we want a match that minimizes the total distance over all pairs, $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$. This is a constrained version of the original optimization problem: minimize $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$, but subject to the constraint that Δ_μ is as small as possible.

2.4. The Algorithm

The method first determines the smallest possible Δ_μ by a threshold algorithm similar to that of Garfinkel (1971). Then, knowing the best Δ_μ, the algorithm solves the optimal matching problem with a revised, penalized distance.

We start with a guess κ of Δ_μ, perhaps quite a poor guess. We create new distances δ'_{tc}, where $\delta'_{tc} = 1$ if $\delta_{tc} > \kappa$ and $\delta'_{tc} = 0$ if $\delta_{tc} \leq \kappa$. We find an optimal matching to minimize $\sum_{t=1}^{T} \delta'_{t,\mu(\tau_t)}$. If this minimum is $\sum_{t=1}^{T} \delta'_{t,\mu(\tau_t)} \leq T - F$, then we found at least F pairs at distance of at most κ, so we revise κ to be a smaller number and try again. If this minimum is $\sum_{t=1}^{T} \delta'_{t,\mu(\tau_t)} > T - F$, then it is impossible to find F distinct pairs with a distance of at most κ, so we revise κ to be a larger number and try again. A binary search quickly finds a close upper bound for Δ_μ.

Let β be a large number, called a penalty. Define new distances $\delta''_{tc} = \delta_{tc}$ if $\delta_{tc} \leq \Delta_\mu$ and $\delta''_{tc} = \delta_{tc} + \beta$ if $\delta_{tc} > \Delta_\mu$. We then obtain our desired match $\mu(\cdot)$ by finding a minimum distance match for the new distances δ''_{tc}. For large enough β, the algorithm finds a match that minimizes the original distance $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$ subject to the constraint that we have F pairs with distances of at most Δ_μ , where Δ_μ is as small as possible; see Rosenbaum (2017a, prop. 1). The penalty β should be large relative to $\max_{t,c} \delta_{tc}$. Although theorems often assume β is enormous, it often aids computational stability to take β to be large but not enormous, perhaps $\beta = 50 \times \max_{t,c} \delta_{tc}$. This whole algorithm is implemented in an R package `thickmatch` on CRAN. It is user-friendly and computationally efficient.

2.5. Other Matched Designs

Our threshold algorithm for quantile-constrained optimization has been described for matched pairs, but there are many other matched designs. In the simplest departure from matched pairs, each treated individual is matched to two controls. Instead, treated individuals might be matched to a variable number of controls, that number being dependent on the number of controls available as the covariates change; see, for instance, Pimentel, Yoon, and Keele (2015).

In full matching, each matched set contains either one treated individual and one or more controls, or one control and one or more treated individuals. It is possible to show that a full matching is the optimal form of matching, the form that arises when minimizing the sum of the within-set covariate distances (Rosenbaum 1991). See Hansen (2004) for an application and Hansen and Klopfer (2006) for an efficient algorithm. A full match can match all controls or some controls. In the analysis step, a full match must be weighted to estimate treatment effects, reflecting the unequal sizes of the matched sets.

In each of these matched designs, the algorithm of Section 2.4 for matched pairs may be adapted as follows. We describe the adaptation for full matching, but the same approach works generally. The algorithm in Section 2.4 is used to build a matched pair design which minimizes the Fth quantile, Δ_μ, in matched pairs, exactly as before. Therefore, the maximum distance for F distinct pairs has been minimized. New distances are defined as $\delta''_{tc} = \delta_{tc}$ for the F selected pairs and $\delta''_{tc} = \delta_{tc} + \beta$ in all other cases, where β is a large penalty. An optimal full match with these revised distances will (i) force inclusion of the selected F distinct pairs as a constraint, (ii) minimize the total of the within-set treatment-control distances, δ_{tc}, over all full matchings that satisfy the constraint (i).

The `thickmatch` package in R implements this extension of the method, and the example in §3 uses full matching. See Section 4 for additional features of the `thickmatch` package useful in other applications. For a different use of threshold matching algorithms in observational studies, see Yu et al. (2020).

2.6. Tied Distances

We have assumed that there are no ties which is the case if at least one of the covariates is from a continuous distribution and the distance, δ_{tc}, is the Mahalanobis distance. If there are many discrete covariates and no continuous covariates, then ties will be extremely uncommon, but they will have positive probability. This section briefly indicates the small consequences of a few ties among the δ_{tc}.

With or without ties among the δ_{tc}, the algorithm finds F pairs that are as close as possible; that is, F pairs whose maximum distance, Δ_μ, has been minimized. With any such fixed set of F pairs, the algorithm minimizes the total of treatment-control distances within matched sets. If the distances are never tied, then there are, by definition, exactly F distances below the Fth quantile in the matched pair data, Δ_μ, and the set of F closest pairs is unique. If there are ties at the Fth quantile, it is possible that there are several pairs with distance exactly equal to Δ_μ, with the consequence that the description "the closest F pairs" actually describes several distinct sets of F pairs.

Suppose that the investigator wanted to interview $F = 30$ pairs, but some distances δ_{tc} were equal. The algorithm might produce 28 pairs with $\delta_{tc} < \Delta_\mu$ and 3 tied pairs with $\delta_{tc} = \Delta_\mu$. There would then be three essentially equivalent ways to select $F = 30$ close pairs, taking the 28 pairs with $\delta_{tc} < \Delta_\mu$ plus 2 of

the three pairs with $\delta_{tc} = \Delta_\mu$. The algorithm, as we described it in Section 2.5, uses one arbitrary choice of these three equivalent pairings. Obviously, this issue of ties is a very minor issue if there are only a few tied distances with $\delta_{tc} = \Delta_\mu$.

3. Application: Confiscating Guns Present During Domestic Violence

3.1. Background

A variety of laws restrict the possession of guns by individuals convicted of domestic violence or under a restraining order. Pennsylvania went further than this, taking action on the basis of the perspective of an arresting officer, prior to any court decision. The statute requires the arresting officer to "seize all weapons used by the defendant in the commission of the alleged offence," if the officer has observed "recent physical injury to the victim or other corroborative evidence" [18 PA. Cons. Stat. x 2711(b)]. How the statute operates in practice and its effects are matters for investigation.

Small, Sorenson, and Berk (2019) examined data from 2013 provided by the Philadelphia Police Department concerning domestic violence incidents in the city. They focused on the 220 domestic violence incidents in 2013 that involved a firearm, noting that a firearm was removed in 52 of these 220 incidents. That is, the treatment, firearm removal, was applied in 52 of 220 domestic violence incidents that involved a firearm, the remaining $220 - 52 = 168$ incidents being controls.

Baseline data for matching and design came from a form that a police officer must complete when responding to a domestic violence incident, whether or not an arrest is made. The form contains some items with check-boxes, some boxes to be filled in, and a hand-written narrative, about a paragraph in length, describing the incident. Would it be helpful to read some closely matched narratives for treatment-control pairs? Would it change our inference for the outcome or make it stronger? This might be done several times, in an iterative attempt to improve both the coding and the match. We will illustrate this in Sections 3.3-3.6. First, we discuss in Section 3.2 why it may be valuable to actually read a few closely matched narratives rather than purely apply machine learning methods to classify the narratives.

3.2. Narrative is More Than Words

We hope to extract useful information from the text paragraphs. The paragraphs contain quite a bit of information, are sometimes vivid, are shaped by the issues that concern police during domestic violence, and are written in a hurry with spelling errors, abbreviations, indecipherable words, and so on.

Would it help to use a machine learning algorithm to lend structure to the text paragraphs? To illustrate some of the difficulties algorithms face when conceptualizing domestic violence based on a narrative paragraph, we applied several algorithms. It is conceivable that the algorithms would perform in a different way with a larger dataset, but the data used here are all of the data for one year in one of the largest cities in the United States. Our basic suggestion is a human reading of a few matched narratives can be helpful in guiding automated procedures.

First, we applied correlated topic models using the `topicmodels` package in R; see Blei and Lafferty (2007). As suggested by Blei and Lafferty (2007) and Srivastava and Sahami (2009), we preprocessed the text by removing 174 "stop words" on a list in the `tm` package in `R`, such as "about" and "the." Also, we removed words with very low frequencies. With five topics, Table 1 shows the top ten words in each topic produced by `topicmodels`.

The topics are not ideal. The words "compl," "comp," and "complainat" distinguish the five topics, but all refer to complainant, the injured party who may have called the police. Topics 1, 2, and 4 include "offender." More or less by definition, in an incident of domestic violence, there is almost invariably a complainant and an offender, and so these words do not make useful distinctions. Topics 3 and 5 include "police," but by definition the police are present in every one of these incidents. Topics 2 and 4 include "handgun," topic 3 includes "gun," but by definition there is a gun involved in every one of these incidents, and handguns predominate. The symbol "xxx" denotes identifying material that was deleted before the data were made available to us. The term "black" appears in topics 1 and 5, but reading narratives reveals that it very likely refers to the color of a gun—often black versus silver—reflecting the concern the police exhibit in carefully describing guns to aid in their identification later. The words "stated" and "states" appear, reflecting the concern the police exhibit in carefully distinguishing what they saw and what they were told, and who told them what. The words "male" and "female" appear, but most of these incidents involve heterosexual couples. It is difficult to know what to make of words like "left" and "pulled" : Is it "pulled the trigger and left the house" or "pulled her hair with his left hand."

Separately, we applied, first, support vector machines, then random forests, in an effort to predict the treatment, namely the confiscation of a gun. Many of the words dubbed important

Table 1. Topic modeling of text paragraphs on the police reports of domestic violence incidents.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	Compl	Comp	Offender	Gun	Compl
2	States	Offender	Compl	States	xxx
3	xxx	Head	Police	Compl	Comp
4	Stated	Pulled	Gun	Male	Stated
5	Male	Face	States	Handgun	Male
6	House	Handgun	Stated	Pulled	Black
7	Black	Left	House	xxx	Verbal
8	Pulled	Complainat	xxx	Pointed	States
9	Female	Verbal	Location	Offender	Police
10	Offender	Boyfriend	Pointed	Stated	Boyfriend

NOTE: The table shows ten important words for each of five topics.

Table 2. Covariate balance, before and after matching.

	Weighted Mean			Standardized Difference	
	Treated	Matched control	All controls	After matching	Before matching
Sample size	52	148	168	52 vs. 148	52 vs. 168
Offender Covariate					
Male	0.87	0.86	0.88	0.03	-0.05
Age (in years)	36.83	36.43	32.06	0.03	**0.41**
History of substance abuse	0.13	0.10	0.18	0.10	-0.12
Under court supervision	0.02	0.02	0.10	0.00	**–0.35**
History of domestic violence	0.27	0.27	0.42	0.00	**–0.32**
History of domestic violence reported to police	0.15	0.12	0.26	0.07	**–0.27**
Fled scene	0.37	0.37	0.80	0.00	**–0.99**
Arrested at scene	0.75	0.75	0.22	0.00	**1.24**
Victim Covariate					
Female	0.83	0.85	0.89	-0.07	-0.19
Age (in years)	33.52	32.92	30.19	0.06	**0.32**

NOTES: Because matched sets vary in size, means are weighted to reflect frequencies in the treated group. The standardized difference is the difference in means divided by the standard deviation prior to matching. Standardized differences above 0.2 are in **bold**.

were similar to those in Table 1, a key exception being "recovered" : when the word "recovered" appeared, the police were likely to have confiscated the gun. Of course, this may refer to "recovered the gun" as an alternative way of saying "confiscated the gun."

Again, the algorithms might perform differently in a larger dataset or with more guidance. The methods we propose entail reading a carefully chosen subset of narratives, and this reading might help in guiding `topicmodels` or similar software. For example, such a method might be useful in guiding the matching with text data methods discussed in Mozer et al. (2020) and Roberts, Stewart, and Nielsen (2020). Reading narratives and building topics are complementary, not competing activities. In particular, as seen below, reading narratives suggests a couple of themes built from interconnected words, such as {son, daughter, baby, infant, child, father, mother, parent}, any of which suggest that a child was in some way connected to the domestic violence incident. In thinking about whether or not to confiscate a gun, the presence of a child might be relevant.

3.3. A First Match

Table 2 shows covariate imbalance in an optimal full match, using the method in Section 2.5, where full matching is described. The match was constructed using the `thickmatch` package in R, and it took less than one second, more precisely 0.30 seconds. The `thickmatch` package was written for this article and it is publicly available at CRAN. The match includes all 52 treated incidents, in which the firearm was confiscated, and 148 control incidents. The match included 24 treatment-control pairs, 2 sets with two treated individuals each matched to one control, 3 sets with one treated matched to two controls, 3 sets with one treated matched to three controls, and 9 sets of widely varying sizes. Because the matched sets vary in size, the control data are weighted to reflect the frequencies in the treated group.

Table 2 shows the mean of each covariate in the treated group, in the entire unweighted control group, and in the matched and weighted control group. It also shows the difference in means, before and after matching and weighting, in units of the standard deviation before matching. Before matching, there are many substantial differences between the treated and control groups. In particular, the gun is less likely to be confiscated if the offender fled the scene, and more likely to be confiscated if the offender was arrested at the scene. After matching and weighting, the treated and control groups appear similar in terms of these observed covariates.

3.4. Five Exceptionally Close Pairs

As discussed in Section 2.5, the optimal full match in Section 3.3 was built to include five exceptionally close pairs. Table 3 compares the five pairs in terms of the covariates in Table 2. Certainly, these five pairs are as promised: they are exceptionally close on the observed covariates. Close pairs represent common situations, but a method for requiring variety in pairs is described in Section 5.

To illustrate the method, Table 4 contains the narrative paragraphs for the five pairs in Table 3. Despite being similar in terms of observed covariates, the narratives describe quite different situations. In the treated incident in the first pair, the offender shot the complainant in the chest, while in the matched control incident, it was her front door that was shot. The two incidents in the fifth pair happen to be similar: both record verbal threats to shoot or kill someone. Some of the narratives mention children, a baby in SP2ID 36606, a daughter in 44968.

One finds in Table 4 that the police carefully record certain aspects. The gun is often carefully described, presumably for later identification, but this is not likely to be a useful covariate for matching. The narrative often indicates whether there were threats to kill or shoot someone. The narrative often mentions whether children are involved. These are plausible considerations when deciding to confiscate a firearm, aspects not reflected in covariates.

3.5. Close or Representative Pairs?

Table 3 considered pairs that were very similar in terms of measured covariates. Modern matching methods—for example,

Table 3. Covariates for 5 optimally close pairs.

	Pair 1		Pair 2		Pair 3		Pair 4		Pair 5	
ID	3425	13992	24963	17473	35793	36606	40857	25918	44968	25791
Treatment	1	0	1	0	1	0	1	0	1	0
Offender Covariate										
Male	1	1	1	1	1	1	1	1	1	1
Age	17	17	33	32	29	30	55	57	39	38
H. substance abuse	0	0	0	0	0	0	0	0	0	0
Court supervision	0	0	0	0	0	0	0	0	0	0
H. domestic violence	0	0	0	0	0	0	0	0	0	0
H. domestic violence reported to police	0	0	0	0	0	0	0	0	0	0
Fled scene	1	1	1	1	0	0	0	0	0	0
Arrested at scene	0	0	0	0	1	1	1	1	1	1
Victim Covariate										
Female	1	1	1	1	1	1	1	1	1	1
Age	16	18	29	28	31	32	46	48	26	26

NOTE: Treatment is 1 for treated, 0 for control. "History of" is shortened to "H.'

Table 4. Narrative paragraphs from the police report for 5 exceptionally close pairs.

SP2ID	Treatment	Description of Incident
3425	1	Comp states that offender was over her house. they started argues male pushed her then pulled out a blk revolver. and shot her in the chest. offender threw the gun on the flair. the fled the loc. comp transported to chop by xxx. dr. xxx ait. stable cond. 22 cal recover bus with white tape on hand grip serial # scratches off live rounds
13992	0	Comp states while sleeping on the couch in the living room she heard a large pop sounding like a gunshot she also observed damage and a large whole to front door. witness states she observed a white vehicle leaving the scene from the second floor window. comp states that she has been recieving threat from ex boyfriend through email and text and that he drives a white car. xxx on loc. put out flash over radio further invest male taken into custody complait taken for interview
24963	1	Compl. states she came home to mention home address and found another woman in the house with her husband the offender & compl. began argueing at which point the offender left the house to get into compl veh (pa tag xxx) the compl. followed offender (husband) outside to veh when offender produce a black/silver handgun and fired two shots in compl direction. compl was not hit by gun fire. offender fled location with unk female
17473	0	Complt said offender came to above location. pointed a gun at her and said he would shoot her. male had a black hand gun and pointed it at the complainat. male goa compl said she dated the male for three months and attempted to end the relationship.
35793	1	Abv. compl states she had a dispute w/offender having dispute over relationship. when he entered prop she tired to make him leave he grabbed her arm & her shirt she turned around and saw he pulled a black handgun out of a blue duffle bag and stated when she scream for cousin (b/w witness) "your cousin will get it to." male fled loc when witness saw him smash out both mirrors of compl's veh & walk up & down xxx. concerned citizen (did not want involvement) showed
36606	0	Comp states that the offender was intoxicated and came to her house (apt) and pointed a black hand gun at her and took their 3 month baby out of her apt and drove away. defendant was stopped on the xxx block of xxx st the baby was in the vehicle and returned to the mother
40857	1	Ric for person w/ gun b/m blue shirt ten pants female compl. police arrived as location compl stated male offender started a verbal argument because she wouldn't have sex with him. while arguing the male went up sticks retrieved his stectar pistol & pointed it at the compl male stated he did nothing wrong because the gun was fake. gun was modified seal. gun placed on property receipt xxx male arrested
25918	0	Compl. states that she is the payee for offender's disability check. compl. had her car serviced before dropping offender's money off. offender met compl. in apartment's parking lot. offender punched compl. and started fighting compl. witness below attempted to break fight up. offender grabbed compl. and dragged her back to his apartment. witness called police. compl. grabbed at outside of frame of apartment door knocking screendoor out. compl. states that she saw a black handgun on offender's coffee table. compl said that she was going to tell police about the gun. offender struck compl in back of head with gun. compl. was able to break free and ran and awaited police arrival
44968	1	Complainat states to police that she and husband had a verbal altercation when husband became irate pushing complainat. complainat states she then shoved him away when husband pulled out a black handgun waving it in her and daughter's direction stating "stay out of my face" "i'm going to shoot you both" complainat states she ran to bathroom and called 911 male taken into custody
25791	0	Compl states while walking home from the grocery store the offender compl's ex boyfriend started to follow the compl and brandish a black colored handgun stating "wheres your boyfriend i'm going to kill him" compl states she proceeded to run and the off. chased her. firing (1) one time in her direction. compl states off. then fled on foot in the direction of his home at the other end of the block xxx. police arrived on

methods using propensity scores or fine balance—do not focus on the production of pairs that are very similar, but rather on treated and control groups that are similar in aggregate as groups, as they were in Table 2. Modern methods want the mean age, the quartiles of age, and so on, to be similar in matched groups, but they do not insist that matched individuals be similar in age. With just ten 3-category covariates, there are $3^{10} = 59,049$ types of people, so close pairs are invariably very rare. The pairs in Table 3 are unusual in being so close—they are close by construction, and it would be impossible to construct many such close pairs.

For comparison with the close pairs in Table 3, we also looked at five randomly picked pairs in our matched sample. In the treated half of the first such pair, the police had recovered a firearm from a male offender who was 82 years old, with a female victim who was 64. The matched control incident in that pair had failed to recover a firearm from a male offender who was 47 and a female victim who was 48. In the treated half of the pair, the 82-year-old offender had remained at home, and was arrested at the scene, while in the control half the 47 year old offender had left before the police arrived and so was not arrested. The 82 year old did not talk of killing the victim, but the 47 year old did. This pair is one of many that balance the covariates—there is nothing intrinsically wrong with this pair—but the two halves of the pair are so dissimilar that little is learned by comparing them in qualitative terms.

3.6. Iteration and Revision of the Match

In light of Section 3.4 and Table 4, we created two new binary covariates, using the computer to search the text for keywords. The keywords were permitted to differ by one letter from the intended keyword. The first covariate was 1 if either the word "kill" appeared or the phrase "shoot you" appeared; otherwise, this covariate was zero. The second covariate was 1 if any of the following words appeared: child, baby, daughter, son, dad, mom, father, mother, daddy, mommy, pregnant; otherwise, the second covariate was zero.

These two covariates were included as covariates in the distance. The resulting match, with two additional covariates, exhibited covariate balance similar to Table 2: all covariates, including the two new covariates, had standardized differences after matching of at most 0.1.

For this new match, a new table of narratives for five close pairs was constructed. Two of the five pairs were the same as in Table 4, and three were different, reflecting the impact of the two new covariates. Each narrative is, of course, unique; however, there were no obvious themes among these unique narratives.

Although we cannot demonstrate this formally, our sense is that the two covariates built by reading paired narratives make sensible distinctions among domestic violence incidents. These distinctions seemed to reflect distinctions that the police were also making when writing their paragraphs, for instance, whether the offender had threatened to kill or shoot someone.

Using a different match, Small, Sorenson, and Berk (2019) did several analyses. They did several analyses for a binary indicator of at least one subsequent call to the police for domestic violence, and for a count of the number of such calls, and used several methods of analysis. We redid these several analyses with our two new matched samples, namely the initial match and the match that added two new covariates derived from the text. The results are shown in Table 5. Adjustment for the two new text covariates slightly increased the point estimate of the size of the effect, but the confidence intervals and *P*-values were similar. In summary, close reading of the narratives for a few closely matched pairs raised a concern that we might have missed important sources of confounding. We generated new covariates to capture the potential sources of confounding identified from reading the narratives and we matched on these covariates in addition to the covariates included in our initial match. With this new match, we obtained similar inferences about the effect of gun removal as with our original match, enhancing our confidence in the inferences from our original match.

The results do not provide evidence that gun removal at the scene of an intimate partner violence (IPV) incident reduces reports to police of subsequent IPV. If anything, the results suggest gun removal increases the likelihood of subsequent IPV reports to the police. Small, Sorenson, and Berk (2019) suggested three possible explanations for this finding: (i) IPV victims might feel safer and less fearful of retribution without the gun in the home and be more comfortable to call for assistance; (ii) IPV victims might perceive the police as having responded in a helpful way in the index incident (i.e., they removed the gun) and, thus, have more confidence in police and be more willing to call for assistance; or (iii) an abuser who previously wielded a gun as a silent threat, when no longer having access to a gun after one is removed at the scene, may resort to more frequent and aggressive threats which leads to more reported incidents. Other explanations are possible and future research is warranted, see Small, Sorenson, and Berk (2019) for further discussion.

4. Aspects of the `thickmatch` Package Relevant to Other Datasets

The `thickmatch` package in R integrates close matching of a subset of pairs with additional standard aspects of modern matching. We mention these additional aspects only briefly as they are reviewed elsewhere (Rosenbaum 2020a, 2020b). The `thickmatch` package integrates these aspects, because one optimization produces the entire match, including T balanced sets and F extremely close pairs.

Constraints may be imposed on a matching algorithm to ensure that the match has certain desired property such as balancing certain categorical variables. Often, much of the work in achieving a satisfactory match is done by the constraints,

Table 5. Outcome analysis for the match in Small, Sorenson, and Berk (2019), our initial match and our new match adjusting for the two new covariates derived from the text as well.

	Binary outcome: conditional logistic regression			
	Odds multiplier	95% CI	asymp. *p*	perm. *p*
Small, Sorenson and Berk (2019)	3.30	(0.96,11.34)	0.06	0.10
Our initial match	3.74	(0.88,15.80)	0.07	0.12
The match after adjusting for key words	4.32	(1.08,17.36)	0.04	0.06
	Binary outcome: Mantel-Haenszel test			
	Odds multiplier	95% CI	asymp. *p*	perm. *p*
Small, Sorenson and Berk (2019)	3.30	(0.80,14.03)	0.11	0.09
Our initial match	3.73	(0.75,22.53)	0.09	0.14
The match after adjusting for key words	4.32	(0.91,24.46)	0.07	0.04
	Count outcome: weighted robust poisson regression			
	Count multiplier	95% CI	asymp. *p*	perm. *p*
Small, Sorenson and Berk (2019)	6.41	(1.33,30.99)	0.02	0.01
Our initial match	5.58	(1.69,18.39)	0.00	0.03
The match after adjusting for key words	5.69	(1.63,19.87)	0.01	0.03

NOTE: asymp. p: asymptotic *P*-value; perm. p: permutation *P*-value.

not by closely pairing individuals in terms of a distance, δ_{tc}. For qualitative comparison of a subset of pairs, small distances are helpful for pairs in this subset. However, for quantitative comparison of treated and control groups, it is the comparability of the groups, not close pairs, that matters most, and constraints often target comparable groups.

Constraints may be imposed either as hard or soft constraints. If a hard constraint cannot be achieved, then the problem is infeasible, and no match is produced. If a soft constraint is imposed instead, then a match is always produced, but if necessary it may slightly violate the constraints. For instance, the `thickmatch` package implements both exact and near-exact matching for nominal covariates. Exact matching for gender would fail if the number of control women was one woman short of what is needed for exact matching, while near-exact matching would produce a match with one pair in which a man was matched to a woman. Soft constraints are typically implemented by adjusting the objective function, say $\sum_{t=1}^{T} \delta_{t,\mu(\tau_t)}$, so that violations of a constraint are severely penalized by dramatically increasing the objective function's value. Also, the `thickmatch` package provides the option of a caliper, implemented as a soft constraint, on a score such as the propensity score, thereby minimizing the number of times the caliper is violated.

Fine balance for a nominal covariate occurs if the marginal distribution of the nominal covariate is the same in the treated and control groups, ignoring whether treated individuals are paired to controls for this nominal covariate. Fine balance may be implemented as a hard constraint, but may then generate infeasibility. Near-fine balance comes as close to fine balance as the data will allow, and it is implemented as a soft-constraint. The `thickmatch` package implements near-fine balance for pair matching and for matching with a fixed number of controls. To implement fine balance or near-fine balance with a variable number of controls, the method of Pimentel, Yoon, and Keele (2015) should be used instead.

The `thickmatch` package was conceived to implement modern matching techniques while generating F pairs that are exceptionally close, so that qualitative comparison of this subset of pairs might be particularly informative. However, nothing in the implementation requires F to be small, and the package can be used to achieve other purposes not described here. For instance, a large F may be used to impose either a minimax or upper quantile constraint on pair distances, as proposed in Rosenbaum (2017a).

There are several soft constraints in the `thickmatch` package, each implemented as a penalty on the objective function. The package gives the user control over the relative size of these penalties, so the user sets the priorities when some constraints cannot be satisfied. For instance, the user can say that the caliper on the propensity score is more important than fine balance.

5. Discussion

The method in Section 2 simultaneously produces a conventional matched sample and an exceptionally close subsample of F pairs for narrative description or other qualitative studies. The goal is to match in a way that facilitates integration of quantitative and qualitative research. In the example, Table 3 shows that F exceptionally close pairs were indeed found. Two extensions of the matching method are described below.

5.1. Ensuring Variety in the Narratives

We might wish to ensure that the F close pairs that will be described are not too similar to one another. This is easily accomplished: apply the method in Section 2.4 several times in several subpopulations defined by one or two covariates. In the example, most couples were heterosexual, with a female complainant and a male offender. The method just described could be used to ensure that a few narratives in Table 4 described other situations by (i) matching within the subpopulation of cases involving a female complainant and a male offender and (ii) matching within the subpopulation of cases involving other situations, specifying for each match that a few exceptionally close pairs be provided.

5.2. Alternative Matching Algorithms

The algorithm in Section 2.4 used a threshold method to optimize a quantile Δ_μ of the within-pair distances, then imposed that quantile as a constraint, minimizing the total distance subject to that constraint using a minimum cost flow algorithm. The minimum cost flow algorithm could be replaced by a different algorithm. For instance, the threshold method may be applied to nonbipartite matching, as discussed by Lu et al. (2011), or to matching using mixed integer programming, as discussed by Zubizarreta (2012). In this way, a qualitative component may be added to a variety of other matching techniques.

Acknowledgment

We are grateful to Susan Sorenson and the Philadelphia Police Department for access to the police response data.

References

Abadie, A., and Imbens, G. W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235-267. [45]

Aneshensel, C. (2013), *Theory-based Data Analysis for the Social Sciences* (2nd ed.), Thousand Oaks, CA: Sage. [43]

Axinn, W. G., and Pearce, L. D. (2006), *Mixed Method Data Collection Strategies*, Cambridge: Cambridge University Press. [43]

Bertsekas, D. P. (1981), "A New Algorithm for the Assignment Problem," *Mathematical Programming*, 21, 152–171. [45]

Blei, D. M., and Lafferty, J. D. (2007), "A Correlated Topic Model of Science," *Annals of Applied Statistics*, 1, 17–35. [47]

Crede, E., and M. Borrego. (2012), "Learning in Graduate Engineering Research Groups of Various Sizes," *Journal of Engineering Education*, 101, 565-589. []

Creswell, J. W., and Plano Clark, V. L. (2007), *Designing and Conducting Mixed Methods Research*, Thousand Oaks, CA: Sage. [43]

Fisher, R. A. (1935), *Design of Experiments*, Edinburgh: Oliver and Boyd. [42]

Frankel, M. (1983), "Sampling Theory," in *Handbook of Survey Research*, eds. P.H. Rossi, J. Wright, and A.B. Anderson, 21-52, New York: Academic Press. [43]

Garfinkel, R. S. (1971), "An Improved Algorithm for the Bottleneck Assignment Problem," *Operations Research*, 19, 1747-1751. [45,46]

George, A. L. and Bennett, A. (2005), *Case Studies and Theory Development in the Social Sciences*, Cambridge, MA: MIT Press. [44]

Hansen, B. B. (2004), "Full Matching in an Observational Study of Coaching for the SAT," *Journal of the American Statistical Association*, 467, 609–618. [46]

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. [46]

Imbens, G. W., and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York, NY: Cambridge University Press. [45]

Lamont, M., and White, P. (2005), *Workshop on Interdisciplinary Standards for Systematic Qualitative Research*, Report Prepared for the National Science Foundation, Washington, DC. [44]

Lawlor, D. A., Tilling, K., and Davey Smith, G. (2016), "Triangulation in Aetiological Epidemiology," *International Journal of Epidemiology*, 45, 1866–1886. [42]

Lieberson, S. (1991), "Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases," *Social Forces*, 70, 307–320. [43]

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011), "Optimal Nonbipartite Matching and Its Statistical Applications," *The American Statistician*, 65, 21-30. [51]

Mozer, R., Miratrix, L., Kaufman, A. R., and Anastasopoulos, L. J. (2020). "Matching With Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality," *Political Analysis*, 28, 445–468. [48]

Munafò, M. R. and Davey-Smith, G. (2018), "Repeating Experiments is Not Enough," *Nature*, 553, 399-401. [42]

Pimentel, S. D., Yoon, F. and Keele, L. (2015), "Variable-ratio Matching With Fine Balance in a Study of the Peer Health Exchange," *Statistics in Medicine* 34, 4070–4082. [46,51]

Riessman, C. K. (2012). "Analysis of Personal Narratives," in *The Sage Handbook of Interview Research: The Complexity of the Craft*, eds. J. F. Gubrium, J. A. Holstein, A. B. Marvasti, and K. D. McKinney, 367–380. Thousand Oaks, CA: Sage. [42]

Roberts, M. E., Stewart, B. M., and Nielsen, R. A. (2020), "Adjusting for Confounding With Text Matching," *American Journal of Political Science*, 64, 887-903. [48]

Rosenbaum, P. R. (1991), "A Characterization of Optimal Designs for Observational Studies," *Journal of the Royal Statistical Society*, Series B, 53, 597-610. [46]

——— (2001), "Replicating Effects and Biases," *The American Statistician*, 55, 223-227. [42]

——— (2017a), "Imposing Minimax and Quantile Constraints on Optimal Matching in Observational Studies," *Journal of Computational and Graphical Statistics*, 26, 66-78. [45,46,51]

——— (2020a), "Modern Algorithms for Matching in Observational Studies," *Annual Review of Statistics and Its Application*, 7, 143–176. [50]

——— (2020b), *Design of Observational Studies* (2nd ed.), New York: Springer. [50]

——— (2021), *Replication and Evidence Factors in Observational Studies*, New York: Chapman and Hall/CRC. [42]

Rosenbaum, P. R. and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [44]

Rosenbaum, P. R., and Silber, J.H. (2001), "Matching and Thick Description in an Observational Study of Mortality After Surgery," *Biostatistics*, 2, 217–232. [43,44]

Roth, W. D., and Mehta, J. D. (2002), "The Rashomon Effect: Combining Positivist and Interpretivist Approaches in the Analysis of Contested Events," *Sociological Methods and Research*, 31, 131–173. [43]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [42]

——— (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. [45]

Seawright, J., and Gerring, J. (2008), "Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options," *Political Research Quarterly*, 61, 294-308. [43]

Small, D. S., Sorenson, S. B., and Berk, R. A. (2019), "After the Gun: Examining Police Visits and Intimate Partner Violence Following Incidents Involving a Firearm," *Journal of Behavioral Medicine*, 42, 591–602. [43,47,50]

Small, M. L. (2009), "'How Many Cases Do I Need?' On Science and the Logic of Case Selection in Field-Based Research." *Ethnography*, 10, 5–38. [44]

Srivastava, A. N. and Sahami, M. (eds.) (2009), *Text Mining: Classification, Clustering, and Applications*, New York: Chapman and Hall/CRC. [47]

Stenger, K. M., Ritter-Gooder, P. K., Perry, C., and Albrecht, J. A. (2014), "A Mixed Methods Study of Food Safety Knowledge, Practices and Beliefs in Hispanic Families With Young Children," *Appetite*, 83, 194–201. [43]

Stuart, E. A. (2010), "Matching Methods for Causal Inference: A Review and a Look Forward," *Statistical Science*, 25, 1–21. [45]

Susser, M. (1973), *Causal Thinking in the Health Sciences: Concepts and Strategies in Epidemiology*, New York: Oxford University Press. [42]

——— (1987), "Falsification, Verification and Causal Inference in Epidemiology: Reconsideration in the Light of Sir Karl Popper's Philosophy," in *Epidemiology, Health and Society: Selected Papers*, ed. M. Susser, 82–93, New York: Oxford University Press. [42]

Tarrow, S. (2010), "The Strategy of Paired Comparison: Toward a Theory of Practice," *Comparative Political Studies*, 43, 230–259. [43]

Weiss, R. S. (1994), *Learning from Strangers: The Art and Method of Qualitative Interview Studies*, New York: The Free Press. [42]

Yu, R., Silber, J. H., and Rosenbaum, P. E. (2020), "Matching Methods for Observational Studies Derived From Large Administrative Databases," *Statistical Science*, 35, 338–355. [46]

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. [51]