*Statistical Science*
2020, Vol. 35, No. 3, 367–370
https://doi.org/10.1214/19-STS741
Main article: https://doi.org/10.1214/19-STS699


# Commentary on Yu et al.: Opportunities and Challenges for Matching Methods in Large Databases

**Elizabeth A. Stuart and Benjamin Ackerman**

In their paper titled "Matching Methods for Observational Studies Derived from Large Administrative Databases," authors Ruoqi Yu, Jeffrey Silber and Paul Rosenbaum [13] discuss matching methods in the age of "big data." Matching methods, such as Mahalanobis distance matching, exact matching and propensity score matching, are well established design strategies for reducing bias due to observed characteristics in nonexperimental studies. Fundamentally, matching methods aim to create matched groups of units (often individuals) who are similar to one another on a (sometimes large) set of covariates. Exact matching aims to do that matching on each variable simultaneously (e.g., find for each treated individual a control individual with the same values of age, gender, baseline health status, etc.). Other matching methods summarize the covariates into a low-dimensional summary (e.g., the probability of receiving treatment, known as the propensity score [9]) and match units on that summary. The simplest form of matching is 1:1 matching, where each treated subject is matched to 1 comparison subject similar on the variables used in the matching. Many variations on 1:1 matching exist, including variable ratio matching, full matching and methods that allow controls to be used as a match more than once ("matching with replacement"). Recent advances in matching include "fine balance," which aims to exactly match the marginal distributions of the covariates rather than requiring each pair match exactly [8]. See Stuart (2010) [11] for the background and details of some of those approaches. (Note that while Yu et al. [13] use the term "control" to refer to the reference group, we prefer the term "comparison" group in the nonexperimental context, to distinguish it from the control group in a randomized controlled trial. For current purposes, readers can consider the "control" and "comparison" terms equivalent.)

Weighting, in particular inverse probability of treatment weighting [1], is another common strategy for handling observed covariates in nonexperimental studies. Some researchers view it as preferable to matching because it does not "throw away data" as matching can appear to do (e.g., a study with 1000 treated and 2000 comparison subjects may end up using only 1000 of the comparison subjects in a 1:1 matched design). However, there are a number of benefits to considering matching. First, it is often an attractive approach for nonexperimental study design because of its strong design aspects: it is straightforward, for example, to show the similarity of the matched groups and for even nontechnical readers to intuitively see that the individuals that remain in the analysis "look similar" to one another, at least on the observed covariates [4]. In contrast, with weighting it can be challenging for readers to interpret what it means when, for example, a subject receives a "weight" of 1.3; what is 1.3 people? Finally, it is important to note that the apparent change in sample size does not always lead to reduced power; with weighting the key metric would be the effective sample size, not the number of subjects that remain in the analysis. The total sample size in this case can thus sometimes be a not fully informative metric. For these reasons, matching should be considered as a key tool and often quite appropriate strategy in the nonexperimental study design toolbox and we are happy to see this paper, which helps provide a method for implementing matching in large-scale data.

While the increasing availability of large administrative data (such as electronic health records (EHR), medical claims data, large-scale educational datasets or social media data) allows for unprecedented inference and research, its high dimensional nature poses computational challenges that make certain existing matching methods either infeasible to implement or unfavorable given the context and time needed to complete. In fact, such data often lead to a conundrum: large numbers of covariates observed, such that being able to match on many or all of them can help satisfy the key assumption of unconfounded treatment assignment [9], but computationally it may be challenging to actually do that matching.

These large and comprehensive datasets thus provide opportunity for significant bias reduction due to observed covariates, but there is still the issue of how to actually deal with all of those covariates. The approach proposed by Yu et al. (2020) [13] provides a potential solution to

*Elizabeth A. Stuart is Professor of Mental Health, Biostatistics, and Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, 21205, USA (e-mail: estuart@jhu.edu). Benjamin Ackerman is a Doctoral Candidate in the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, 21205, USA.*

this problem, by allowing researchers to match large numbers of units on large numbers of covariates. With smaller, more "traditional" nonexperimental data that have limited pretreatment covariates available, researchers rely on conducting sensitivity analyses to address concerns of unobserved confounding. While sensitivity analyses are useful and important, the ability to account for all possible potential confounders is key for satisfying unconfoundedness. The method proposed by Yu et al. therefore helps make the most of big data for drawing causal inferences. (We also applaud the authors for still doing an analysis of sensitivity to unobserved confounding, on top of the close matching on a large set of observed covariates.)

To illustrate the complications of implementing matching with large administrative data, Yu et al. (2020) [13] present a toy example to demonstrate the methodological challenges with current practice. The authors include a helpful figure (Figure 1 in [13]) depicting each method with an accompanying bipartite graph, a visual that represents the treated individuals on one side as dots (nodes), connected to all of their candidate control node matches on the other side by line segments (edges). They first present the simple case where the treated nodes are connected to all possible control node matches such that all possible edges are present, also known as a complete and dense bipartite graph. This becomes impractical for optimal matching in a single optimization when there are, say, 30,000 treated people to match to 60,000 controls in an EHR sample. Next, they illustrate a common current practice of binning the data and performing an exact match within each bin, reducing the computing needs. However, such binning could either (1) arbitrarily define bins by covariate levels regardless of their scientific basis or (2) lead to bins too small to match in, especially when binning by a variable with many categories with varying size. A specific example that can cause challenges are when binning by ICD-9 or ICD-10 diagnosis codes in EHR data. ICD codes are alphanumeric strings that define a standardization method for recording all of a patient's diagnoses, and are primarily used for filing billing claims with health insurance companies. ICD codes provide a wealth of consistently recorded information on a patient's health, and matching on ICD codes can help ensure that comparisons are being made across individuals with similar states of health. However, binning by those codes may cause challenges, as certain codes are so common that their resulting bins are too large, while rarer codes may need to have their bins merged because they are too small.

To avoid such arbitrary binning, the authors then turn to the use of a caliper as an alternative way to remove edges between treated and control units that are not close enough matches. Imposing a caliper on the propensity score limits the possible control matches for a given treated node to only those within a certain interval around the treated node's propensity score. However, the choice of the wrong (or too tight) caliper can lead to infeasible matches, and thus choosing a caliper can be quite tricky in practice.

To solve these challenges, the authors propose a matching method that determines an optimal caliper that reduces the number of edges in the bipartite graph but still leads to a feasible match. However, when there is extensive overlap between the distributions of propensity scores for treated and control groups, the number of viable control matches per treated subject can still be quite large, even after determining the optimal caliper. Therefore, the proposed method also identifies the smallest number of nearest neighbors needed for each treated subject, $\nu$, such that pair matching is still feasible. This significantly reduces the number of bipartite graph edges and, therefore, the number of computations necessary to obtain a matched sample. While it is not required to use $\nu$, the authors note the benefit in researchers knowing that a match is still feasible if the graph is restricted to at most $\nu$ nearest neighbors per treated subject. The authors apply an iterative variation of Glover's algorithm to first determine the optimal caliper, and then to calculate $\nu$ (the details of which are available in their paper).

A particularly nice aspect is that this approach can also be combined with exact matching on a nominal covariate (the authors use exact matching on sex as an example); this can be particularly useful when there is a covariate believed to be particularly predictive of the outcome(s) of interest. As one example of that type of setting, in a study of suicide prevention centers in Denmark, it was determined that it was particularly crucial to obtain exact matches on two key covariates that are strongly predictive of repeat suicide attempts: previous deliberate self-harm attempt and whether the individual had any psychiatric disorder [3]. This ability to prioritize certain covariates in the matching is a key benefit of matching in general, and of this proposed new approach as well. An interesting direction for future work will be to help readers think through what variables to force an exact match or fine balance on, or whether the methods could be adapted to automatically prioritize balance on certain variables, such as those most strongly related to the outcome(s) of interest (e.g., as discussed in Stuart et al. (2013) [12]).

We want to particularly applaud the authors for already making available software to implement the methods in R. In our experience this has been a key limitation in terms of implementing many of the newer and more sophisticated matching methods. Having the `bigmatch` package available, along with guidance and documentation for researchers to properly use the package's functions, will greatly facilitate the use of these important methods.

On that note, we see many potential uses for the proposed method in the public health and education fields we

work in. For example, the suicide prevention study by Erlangsen et al. (2015) [3] was able to get good covariate balance on 31 covariates in a sample of 5678 treated and 17,034 comparison individuals. However, an even larger number of covariates were available in the extensive Danish registry data used in that study; using the proposed strategies it is possible we could have obtained even better covariate balance on an even larger set of covariates. Another potential application could be examining the effectiveness of medication treatments for Type II Diabetes on mortality. While a previous study used propensity score matching to assess the causal effect of taking metformin on lowering mortality rates among individuals with Type II Diabetes, adjusting for around 20 covariates in a sample of 20,000 subjects [10], one could imagine applying the methods proposed in this paper to an entire population of individuals with diabetes in a health care system, using a much richer set of measured covariates. For instance, some health care systems collect information on family histories of illnesses in their EHR through provider notes or as questionnaire items during patient visits. Social determinants of health are also becoming increasingly included in EHR as health systems see the potential to use EHR to further improve their patient population's health. These additional pieces of information could contribute greatly to Type II Diabetes research.

There are also some clear extensions of the method that could be important. Matching methods are sometimes used in longitudinal settings known as comparative interrupted time series or difference in difference designs, in which researchers wish to estimate the causal effect of some event (e.g., the adoption of a new policy, or the initiation of a new intervention) by comparing outcomes before and after the event occurrence. Matching is complex in that case because of the multiple time points, etc., but matching on a large set of confounders may be just as important here as in the simple single time point case. Extending these methods to that case may be useful; Daw and Hatfield (2018) [2] provide useful guidance on careful variable selection when matching in these settings to avoid introducing additional bias from regression to the mean. It would also be quite interesting to consider extending these matching methods when using longitudinal EHR data where the treatment of interest is time-varying.

We also, though, want to highlight other challenges in using EHR or other complex data for estimating causal effects. The methods proposed are useful once the research question has been articulated clearly, the covariates identified, outcomes defined, etc. In our experience the steps leading to that are almost as difficult as determining what to do with the data once it is "ready to go." Statisticians would be well served to get engaged in those earlier stages that take data from, for example, "raw" EHR data, to clean analyzable data. This begins in the early stages of data acquisition with the nontrivial tasks of extracting data from free-form text provider summaries and laboratory results in pdf outputs into "tidy" format for databases, and subsequently inspecting the data quality. Data recorded at different health centers and by different health care professionals may vary in quality, consistency, use of units and formatting and phrasing for categorical variables, which could lead to issues of measurement error. Additionally, it may be common to find that data are missing, potentially not at random, and often in ways related to variation in practices across health care settings and even individual providers as they enter data during patient visits.

Other data challenges stem from the continuous process by which data are collected and processed for patients; for instance, a patient may visit a clinic on January 1, at which time a provider orders laboratory tests, but the test results are not recorded in the patient's EHR until January 20. Making decisions on how to align patient vitals, diagnoses, laboratory tests, etc. with visit dates or other time points may have downstream effects with the eventual analyses conducted. Additionally, there may be biases due to frequency of visits (i.e., informed presence bias) [6]. Lastly, EHR data are not always collected with the intention of conducting research. Practices such as upcoding, where providers include a multitude of ICD codes for insurance claims purposes only, may lead to inaccurate depictions of a patient's current health state [5, 7]. The scenarios under which such processes may lead to bias in a non-experimental study should be carefully considered.

In conclusion, we applaud the authors on a very nice paper and contribution. We hope that it serves both to provide a useful and rigorous new approach for implementing matching methods in nonexperimental studies in large-scale datasets, as well as an example for how to conduct a rigorous nonexperimental study, including assessment of sensitivity to an unobserved confounder. We can only hope that future researchers pay as much attention to designing a study to reduce confounding due to observed factors, and to analyses of sensitivity to an unobserved confounder, as do Yu et al. (2020) [13]. We look forward to trying out these methods in our own projects.

## REFERENCES

[1] Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat. Med.* **34** 3661–3679. MR3422140 https://doi.org/10.1002/sim.6607

[2] Daw, J. R. and Hatfield, L. A. (2018). Matching and regression to the mean in difference-in-differences analysis. *Health Serv. Res.* **53** 4138–4156. https://doi.org/10.1111/1475-6773.12993

[3] Erlangsen, A., Lind, B. D., Stuart, E. A., Qin, P., Stenager, E., Larsen, K. J., Wang, A. G., Hvid, M., Nielsen, A. C. et al. (2015). Short-term and long-term effects of psychosocial therapy for people after deliberate self-harm: A register-based, nationwide multicentre study using propensity score matching. *The Lancet Psychiatry* **2** 49–58.

[4] GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statist. Sci.* **22** 153–164. MR2408951 https://doi.org/10.1214/088342306000000691

[5] GERUSO, M. and LAYTON, T. (2015). Upcoding: Evidence from Medicare on squishy risk adjustment. Technical Report, National Bureau of Economic Research.

[6] GOLDSTEIN, B. A., BHAVSAR, N. A., PHELAN, M. and PENCINA, M. J. (2016). Controlling for informed presence bias due to the number of health encounters in an electronic health record. *Am. J. Epidemiol.* **184** 847–855.

[7] ROSE, S. (2016). A machine learning framework for plan payment risk adjustment. *Health Serv. Res.* **51** 2358–2374. https://doi.org/10.1111/1475-6773.12464

[8] 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. *J. Amer. Statist. Assoc.* **102** 75–83. MR2345534 https://doi.org/10.1198/016214506000001059

[9] 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. MR0742974 https://doi.org/10.1093/biomet/70.1.41

[10] ROUSSEL, R., TRAVERT, F., PASQUET, B., WILSON, P. W., SMITH, S. C., GOTO, S., RAVAUD, P., MARRE, M., PORATH, A. et al. (2010). Metformin use and mortality among patients with diabetes and atherothrombosis. *Arch. Intern. Med.* **170** 1892–1899.

[11] STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

[12] STUART, E. A., LEE, B. K. and LEACY, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J. Clin. Epidemiol.* **66** S84–S90.

[13] YU, R., SILBER, J. H. and ROSENBAUM, P. R. (2020). Matching methods for observational studies derived from large administrative databases. *Statist. Sci.* **35** 338-355.