PROJECT MUSE®

Matching with attention to effect modification in a data challenge

Luke Keele, Samuel D. Pimentel

Observational Studies, Volume 5, Issue 2, 2019, pp. 83-92 (Article)

Published by University of Pennsylvania Press

DOI: <https://doi.org/10.1353/obs.2019.0004>

➔ *For additional information about this article*

<https://muse.jhu.edu/article/793359/summary>

Matching with attention to effect modification in a data challenge

Luke Keele

*University of Pennsylvania
3400 Spruce St, Philadelphia, PA 19130*

luke.keele@uphs.upenn.edu

Samuel D. Pimentel

*University of California, Berkeley
367 Evans Hall, Berkeley, CA 94720*

spi@berkeley.edu

Abstract

For this data challenge, we implement a data analytic approach based on matching. First, we use matching to control for observed confounders. Next, we implement a randomization test that combines results from subgroups of matched pairs formed using potential effect modifiers. This analysis, which can be conducted in either a planned or an exploratory manner, is designed both to detect effect modification in specific subgroups and to exploit any variation in effect size to make the study more robust to bias from a hidden confounder. We find that accounting for possible effect modification does make the study results less sensitive to hidden bias. However, we also find that an exploratory analysis that does not fully exploit outcome information fails to fully discover possible effect modifiers.

Keywords: matching, effect modification, sensitivity analysis

1. Methodology and Motivation

1.1 Matching in observational studies

Matching is a flexible and intuitive tool for designing effective observational studies. In its simplest form, matching involve pairing each subject from a group receiving a treatment or exposure of interest to a control subject that appears similar on observed pre-treatment covariates. When observed covariates include all relevant confounders and sufficiently similar pairs are available, the resulting paired design may be analyzed as though it were a randomized study. Analyzing observational data using matched comparisons has many attractive features. For instance, problems with covariate overlap are quickly identified (Stuart 2010), the separation of design from analysis ensures valid inference even if multiple matched designs are initially considered (Rubin 2007), and the preservation of the original units of analysis permits helpful qualitative input from scientific collaborators (Rosenbaum 2002).

For our purposes, two such features are particularly important. One is the natural way matched designs capture information about treatment effect variation. Specifically, each distinct matched pair difference is an estimate of the treatment effect at a particular point in multidimensional covariate space. A single matched design may permit both an average treatment effect estimate over the entire sample and a more granular study of treatment effect heterogeneity or, in the parlance of the matching literature, effect modification by

pre-treatment covariates. The second key advantage of matching methods is the set of tools available for conducting sensitivity analysis after matching in order to assess the impact of possible violations of the strong ignorability, or “no unmeasured confounders,” assumption. Since some form of this uncheckable assumption underlies all analysis of observational data, effective methods for causal inference should be able to demonstrate some robustness to a violation of this assumption. Sensitivity analysis can provide just such a certificate, or identify situations in which results are very fragile with respect to failures of ignorability.

1.2 Identifying effect modification in matched designs

Effect modification occurs when the magnitude of the treatment effect varies with the value of one or more pretreatment covariates, so that certain subgroups of matched pairs experience a different effect than others. A recent literature has studied methods for detecting and leveraging effect modification in matched observational studies (Hsu et al. 2013, 2015; Lee et al. 2018). These works identify two basic strategies for studying effect modification. In one strategy, a small set of subgroups, perhaps identified by one or two nominal variables, are identified a priori based on scientific considerations. In addition to analyzing the data for the entire match, investigators conduct inference separately within each of these subgroups. This strategy is very similar to standard subgroup analysis in social and medical applications.

The second approach is to identify subgroups likely to experience effect modification in an adaptive way, from the data itself. Hsu et al. (2013) describe one way to do this by fitting a CART model to matched pair differences using a large set of pre-treatment covariates, and measuring treatment effects within the subgroups identified by the leaves of the resulting tree model. Importantly, these trees are fit not on the actual matched pair outcome differences but on their absolute values, or on the ranks of the absolute values. Using the signed outcomes to select subgroups for inference would violate the design principle of outcome blinding that is fundamental in matching approaches and would invalidate tests for treatment effects in the resulting groups. While the absolute values are an imperfect proxy for the actual matched pair differences, they preserve valid inference since randomization inference varies only the signs of these differences.

In principle these methods may be applied to any matched design, whether or not it was constructed with the study of effect modification in mind. However, a subtle aspect of the approaches just described is that they rely on pairs exactly matched for the potential effect modifier of interest. For example, if gender is a possible effect modifier, an effect may be estimated within a group of pairs in which both subjects are women and another effect may be estimated in group of pairs in which both subjects are men. But if the study also contains pairs that include both a woman and a man, these last pairs are not directly useful for studying effect modification. One way to understand this principle is to recognize that effect modification is a result of an interaction effect between a treatment variable and a pre-treatment covariate. Pairs matched exactly for the covariate remove the main effect of the pre-treatment covariate and allow study of the interaction, while outcome differences in pairs with non-identical covariates may be partly explained by the main effect of the covariate and partly explained by the interaction.

In practice, this requirement means that investigators with an interest in effect modification must take care to ensure high proportions of matched pairs share identical values of the potential effect modifiers when they construct the match. This requirement stands in contrast to recent recommendations and techniques for matching that focus on achieving marginal balance, a less stringent goal than exact matching (Zubizarreta 2012; Pimentel et al. 2015). Note that this requirement applies both when subgroups of interest are identified a priori and when the exploratory technique is used; in the latter case the CART model is fit only to the subset of pairs matched exactly on all the covariates used in the fit.

1.3 Impact of effect modification on inference and sensitivity to bias

While understanding effect modification for a given dataset is usually scientifically interesting in itself, another important benefit is reduced sensitivity to bias in testing for the overall presence or absence of a treatment effect in the entire sample. In particular, matched data are often analyzed by using randomization inference to test Fisher-style sharp null hypotheses of no treatment effect. Such hypotheses specify that no individual anywhere in the study receives treatment, in contrast to the more common Neyman hypothesis framework which specifies that an average effect is equal to zero but allows for nonzero individual effects. A rejection of the sharp null hypothesis of zero effect for a specific subgroup of the sample is also a rejection for the overall sharp null; note that this is not the case in a Neyman-style testing framework. If effect modification is structured so that certain small subgroups of the sample experience very large effects, a test that combines results from individual tests in several subgroups may be more powerful than a single test incorporating all the data. For example, Hsu et al. (2015) uses a truncated product of p-values generated from tests in several subgroups.

More important yet, analyses that pay attention to effect modification may also reduce sensitivity to unmeasured bias. To make this claim more clear, we briefly review Rosenbaum’s method of sensitivity analysis, described in greater detail in Rosenbaum (2005) and Rosenbaum (2002). Randomization inference in a matched design typically proceeds under the assumption that individuals in the same pair have the same true propensity to receive treatment; mathematically, their odds of treatment are assumed to be identical, allowing construction of a known and tractable randomization distribution. In sensitivity analysis, this assumption is relaxed so that odds of treatment for paired individuals may differ by up to a multiplicative factor Γ . For any specific choice of Γ , a worst-case analysis may be conducted to obtain the largest possible p-value or the widest possible confidence interval that might have been achieved under these settings. Analysts often repeat the analysis from many Γ -values and report a threshold value at which the hypothesis test ceases to reject or at which the confidence interval barely covers 0. For example, say we observe that the estimated p-value is 0.01. This p-value assumes $\Gamma = 1$, that is the unobserved confounder does not change the odds of treatment within matched pairs. To perform a sensitivity analysis, we increase Γ until the worst-case p-value reaches or exceeds 0.05. If this occurs for a relatively small value of Γ , we can conclude that a confounder with a small effect on the treatment odds would change our conclusions. However, if the value of Γ is large when the bounds include zero, then we have greater confidence that a weak confounder would not change the conclusions of the statistical analysis.

When effect modification is present, hypothesis tests for different subgroups of a sample may vary not only in their degree of statistical significance but in their sensitivity to bias. Results for multiple sensitivity analyses may be combined using truncated p-values just as standard test results are combined, resulting in an overall sensitivity analysis, and Hsu et al. (2013) demonstrate that in this setting the level of sensitivity to bias tends to be driven by the subgroup with the strongest or most robust effect. As such, combining the results of several subgroup analyses may produce more reliable inferences less sensitive to unmeasured confounding than the results of a single test for the entire sample.

1.4 Study design for the workshop data.

We demonstrate both matching and a search for effect modification in a synthetic dataset constructed from the National Study of Learning Mindsets for a workshop at the 2018 Atlantic Causal Inference Conference, described earlier in this journal issue in Carvalho et al. (2019). We approached the workshop data by forming a matched pair design and conducting randomization inferences and sensitivity analyses designed to detect and leverage possible effect modification. We first used methods based on a priori identification of two important effect modifiers, and then conducting an exploratory analysis designed to identify additional potential modifiers. Given the framing of the workshop data as an observational study, a primary concern guiding the design was the possibility of unmeasured confounding between the treatment and control groups.

When constructing the matched design itself, we sought to match exactly as frequently as possible on variables judged to be likely effect modifiers. While in general it is a difficult task to match on many variables exactly, in this case the large majority of covariates, including the two pre-identified possible effect modifiers X1 and X2, were measured at the school level. Since in addition each school contained proportionally similar numbers of treated and control students, we chose to match students within schools, thus ensuring that all pairs in the study shared identical values for all school-level covariates. (However, as discussed below, we also conducted an across-schools match for comparative purposes.) Within schools we matched students optimally on a robust Mahalanobis distance incorporating all student covariates. We also estimated a propensity score using all available covariates and imposed a matching caliper of 0.15 standard deviations of the estimated propensity scores, requiring that all pairs formed contain individuals whose estimated propensity scores differ by no more than that amount. This ensures that in the absence of unmeasured confounding, matched individuals have near-identical odds of treatment, an important assumption underlying later randomization inferences. The caliper restriction prevented some treated units in the study from being matched as described in the next section, but we retained a large proportion of the treated students.

To account for effect modification by X1 and X2, the a priori possible effect modifiers, we made two separate partitions of the matched pairs, one by quintiles of X1 and one by quintiles of X2. For each partition, we conduct a separate randomization test of the sharp null hypothesis using a member of the family of Huber-Maritz M-tests chosen to be insensitive to unmeasured bias (Rosenbaum 2007). We then combine the 5 p-values from these tests using the truncated product of p-values (Zaykin et al. 2002). We also conduct a sensitivity analysis for this procedure, reporting the threshold level of unobserved

confounding Γ at which the combined p-value ceases to be significant. For comparative purposes, we also conduct a Huber-Maritz M-test and sensitivity analysis for the overall sample. We also conducted exploratory analyses to detect effect modification using CART models fit to absolute matched pair differences, as discussed below.

2. Workshop Results

We first present a few details on the results of the two matches. First, there was very little observed bias in the data. To assess the quality of the match, we used a measure of standardized distance, which for a given variable is computed by taking the mean difference between matched patients and dividing by the pooled standard deviation before matching (Silber et al. 2001; Rosenbaum and Rubin 1985; Cochran and Rubin 1973). In matching, we generally seek to make all standardized differences less than 0.10 or less than one-tenth of a standard deviation, which is often considered an acceptable discrepancy (Silber et al. 2001; Rosenbaum and Rubin 1985; Cochran and Rubin 1973; Rosenbaum 2010).

Table 1: Balance Statistics for the Unmatched Data and Two Matched Comparisons

	Std. Diff Unmatched	Std. Diff. Within School Match	Std. Diff. Across School Match
S3	0.13	0.00	-0.01
X1	-0.10	0.00	0.01
X2	0.06	0.00	-0.00
X3	-0.00	0.00	0.00
X4	-0.02	0.00	-0.00
X5	0.07	0.00	-0.01
C1 Cat. 1	-0.04	0.00	0.01
C1 Cat. 2	-0.00	-0.02	-0.02
C1 Cat. 3	-0.02	0.03	0.02
C1 Cat. 4	0.04	-0.04	-0.07
C1 Cat. 5	0.01	0.03	0.02
C1 Cat. 6	0.03	-0.01	0.00
C1 Cat. 7	0.02	0.00	0.03
C1 Cat. 8	0.02	0.02	0.04
C1 Cat. 9	-0.03	0.02	0.02
C1 Cat. 10	0.01	0.02	0.03
C1 Cat. 11	-0.01	0.01	0.05
C1 Cat. 12	-0.01	0.02	0.02
C1 Cat. 13	-0.05	0.01	0.01
C1 Cat. 14	-0.02	-0.01	0.01
C1 Cat. 15	0.01	0.03	0.04
C2	0.05	0.01	0.00
C3	-0.10	-0.01	-0.03
XC Cat 1	0.01	0.00	0.00
XC Cat 2	-0.03	0.00	0.00
XC Cat 3	0.03	0.00	0.01
XC Cat 4	-0.05	0.00	-0.00
XC Cat 5	0.03	0.00	-0.00

Table 1 contains balance statistics for both the unmatched data and the two matches we implemented. First, it is clear that there is little overt bias in the unmatched data. For example, the largest standardized difference across treated and control groups *before*

matching was .13. The matching does further improve balance. After matching, with one exception, all the standardized differences were 0.05 or less. For the within school match, the standardized differences are zero for all discretized school-level variables since by design each pair is exactly matched on these variables. Table 2 further illustrates the primary difference between the two matches. For student level covariates, neither match exactly matches on student level covariates, but the within school match exactly matches on all school level covariates, which is a major advantage for studying effect modification by these covariates.

Table 2: Proportion of pairs matched exactly on each variable

	Across School Match	Within School Match
S3	0.79	0.78
C1	0.84	0.82
C2	0.90	0.87
C3	0.92	0.91
XC	0.98	1.00
X1	0.90	1.00
X2	0.90	1.00
X3	0.90	1.00
X4	0.90	1.00
X5	0.90	1.00

Table 3 contains the outcome estimates. We estimated treatment effects and confidence intervals after matching by inverting Huber-Maritz m-tests and computing Hodges-Lehmann estimates as described in Rosenbaum (2002); for the unadjusted case we used simple regression on the treatment indicator. The results from both matches are nearly identical with treatment effect estimates of 0.26 and 0.27 for the across and within school matches respectively. Moreover, consistent with the fact that there is very little overt bias, the unadjusted estimate does not differ much from the adjusted estimates. The sample sizes do vary some as a result of the matching. In the unmatched data, there were 3,384 treated units and 7,007 controls. In the across school match, we matched 3,375 of the treated units one to one. In the within school match, we matched 3,079 of the treated units one to one. Some loss of sample size is often a by-product of exact matching, especially when combined with a caliper restriction.

Next, we explored the possibility of effect modification by X1. For this analysis, we only focus on results from the within school match. For this analysis, we simply stratified the matched pairs by the quintiles of X1e, and estimated overall treatment effect within each quantile, again by inverting randomization tests and computing Hodges-Lehmann estimates. Table 4 contains the results from this stratified analysis. The point estimates indicate quite clearly that for students with lower X1 scores the intervention appears to be more effective. However, the confidence intervals overlap to an extent that these estimates that can't be differentiated statistically.

Next, we repeated this analysis for the X2 covariate. Again, we stratified the matched pairs by the quintiles of the X2 variable, and estimated overall treatment effect within each

Table 3: Results from Outcome Analysis

	Outcome
Unadjusted	0.30 [0.28 , 0.33]
Across School Match 1	0.26 [0.24 , 0.29]
Within School Match 2	0.27 [0.25 , 0.30]

Table 4: Treatment Effect Estimates Stratified by X1 Quantile

	Point Estimate	95% CI
Quantile 1	0.33	[0.27, 0.38]
Quantile 2	0.27	[0.21, 0.34]
Quantile 3	0.31	[0.26, 0.36]
Quantile 4	0.22	[0.16, 0.28]
Quantile 5	0.21	[0.14, 0.27]

quantile. Table 5 contains the results from this stratified analysis. Here, the pattern is less clear, however, treatment effect estimates in the lowest quantile are much lower. Again, the confidence intervals overlap to an extent that these estimates can't be differentiated statistically.

Table 5: Treatment Effect Estimates Stratified by X2 Quantile

	Point Estimate	95% CI
Quantile 1	0.17	[0.11, 0.24]
Quantile 2	0.30	[0.24, 0.36]
Quantile 3	0.29	[0.24, 0.34]
Quantile 4	0.33	[0.28, 0.39]
Quantile 5	0.24	[0.19, 0.30]

Next, we explored whether effect modification by these two covariates reduced sensitivity to bias from unobserved confounders. Table 6 contains the results from three different sensitivity analyses. In the first sensitivity analysis, we calculated the upper-bound on the one-sided p-value for unconditional treatment effect. Here, we find that $\Gamma = 2.51$. This implies that a hidden confounder would need to induce the odds of treatment within matched pairs to differ by a factor of 2.5 before the observed results could be fully explained under the null hypothesis. Next using the methods we outlined above, we conducted two additional sensitivity analyses that account for effect modification by the mindset and test score variables. In both cases, we find the results are less sensitive to hidden bias as the value for Γ is now 2.87 for the X2 effect modifier and 2.83 for the X3 effect modifier.

Table 6: Sensitivity Analysis for Unconditional and Conditional Treatment Effect Estimates

	Unconditional	X1 Effect Modifier	X2 Effect Modifier
Γ	2.51	2.87	2.83

To conduct a more exploratory analysis of effect modification, we calculated within-pair differences in outcome for each matched pair in our study, which we denote as τ_i . Following Hsu et al. (2013), we then regressed the rank of $|\tau_i|$ on possible effect modifiers using CART in order to identify subgroups with differential effects (without jeopardizing later inference by using the signs of the τ_i s). We fit a CART model both to all covariates, school- and student-level, using only those pairs matched exactly on all variables, and to all school-level covariates using all pairs (since all were matched exactly on school-level covariates). However, in neither case did the CART model make any splits after pruning using cross-validation, meaning that no subgroups for effect modification were identified.

3. Post-workshop analysis

We conducted one additional analysis after the workshop. Specifically, we were curious as to why the CART regression did not detect any effect modification. We suspected it was due to the fact that we used the unsigned differences in treated and control outcomes in the CART model. As we noted above, using the unsigned ranks is necessary so the statistical tests on which the sensitivity analysis are based have the correct level. To explore what role this might have played in our analysis, we re-fit the CART regression to all matched pairs using the signed within-pair matched differences as outcomes and the school-level covariates as regressors. We trimmed the fitted tree to ensure reduce overfitting. The results are in Figure 1. Now the tree splits the data on both the X1 covariate as well as one of the indicators for XC variable. Of course, a hypothesis test or sensitivity analysis based on this search procedure violates the design principle by using the same outcome information both to identify subgroups and to conduct inference.

4. Discussion

In our analysis, we demonstrated how to explore effect modification through a sensitivity analysis. Under this approach, investigators can both test whether the magnitude of the treatment effect varies with a baseline covariate and explore whether effect modification makes the results less sensitive to bias from hidden confounders. This approach is quite flexible, since it can be applied to both planned and exploratory analyses of effect modification. That is, one can use flexible tree-based methods to search for subgroups where treatment effects may be larger. Our analysis did reveal some interesting tradeoffs that are required when the analysis is exploratory. To yield valid tests, the exploratory model must be blind to some outcome information. However, we showed that ignoring the full outcome distribution caused us to overlook some evidence of effect modification.

Figure 1: The regression tree formed fitting matched pair differences in outcomes on base-line covariates

We argue that this result has two implications. First, it may be valuable to clarify conditions under which effect modification cannot be detected without signed data, as occurred here. This line of research could provide guidance about exactly when and how researchers pay a cost by blinding themselves to signs of pair differences. Second, we suggest a two-step analysis of effect modification when groups are to be selected from data. First, the CART model procedure can be done using partial outcome information to preserve the validity of formal confirmatory test. Second, a purely exploratory analysis can be conducted using the full outcome information. This allows recovery of strong signals in the data that may be hidden by the partially-blinded analysis, even if inference cannot easily be obtained.

Finally, a promising new direction for studying effect modification in matched sets, called the submax method, is presented in recent work by Lee et al. (2017). Simulations show that the submax method may have greater power to detect moderate effect modification than the approaches based on CART we describe here. While it is beyond the scope of our current investigation to adapt the submax framework to the problem in the data challenge, we recommend its further study and generalization.

References

- Carlos Carvalho, Avi Feller, Jared Miller, Spencer Woody, and David Yeager. Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge. *Observational Studies*, 5(21-35), 2019.
- William G. Cochran and Donald B. Rubin. Controlling bias in observational studies. *Sankhya-Indian Journal of Statistics, Series A*, 35:417–446, December 1973.
- Jesse Y Hsu, Dylan S Small, and Paul R Rosenbaum. Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association*, 108

- (501):135–148, 2013.
- Jesse Y Hsu, José R Zubizarreta, Dylan S Small, and Paul R Rosenbaum. Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika*, 102(4):767–782, 2015.
- Kwonsang Lee, Dylan S Small, and Paul R Rosenbaum. A powerful approach to the study of moderate effect modification in observational studies. *Biometrics*, 2017.
- Kwonsang Lee, Dylan S Small, Jesse Y Hsu, Jeffrey H Silber, and Paul R Rosenbaum. Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(2):535–546, 2018.
- Samuel D Pimentel, Rachel R Kelz, Jeffrey H Silber, and Paul R Rosenbaum. 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(510):515–527, 2015.
- Paul R. Rosenbaum. *Observational Studies*. Springer, New York, NY, 2nd edition, 2002.
- Paul R. Rosenbaum. Sensitivity analysis in observational studies. In Brian S. Everitt and David C. Howell, editors, *Encyclopedia of Statistics in Behavioral Science*, volume 4, pages 1809–1814. John Wiley and Sons, Chichester, UK, 2005.
- Paul R Rosenbaum. Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics*, 63(2):456–464, 2007.
- Paul R. Rosenbaum. *Design of Observational Studies*. Springer-Verlag, New York, 2010.
- Paul R. Rosenbaum and Donald B. Rubin. Constructing a control group using multivariate matched sampling methods. *The American Statistician*, 39(1):33–38, February 1985.
- Donald B Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 26(1):20–36, 2007.
- Jeffrey H Silber, Paul R Rosenbaum, Martha E Trudeau, Orit Even-Shoshan, Wei Chen, Xuemei Zhang, and Rachel E Mosher. Multivariate matching and bias reduction in the surgical outcomes study. *Medical Care*, 39(10):1048–1064, 2001.
- Elizbeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010.
- Dmitri V Zaykin, Lev A Zhivotovsky, Peter H Westfall, and Bruce S Weir. Truncated product method for combining p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 22(2):170–185, 2002.
- José R Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371, 2012.