CEO Matching and Weighting Estimator Checklist

Study Title:

Report type:

Contractor:

Criteria and Sub-criteria	Clear and Concise? Y/N
Study Characteristics	
Does the report clearly state the research question or questions of interest?	
Does the report identify itself as a propensity score matching-based evaluation?	
What type of intervention is being tested? (examples: curriculum, product, program, practice, or policy, etc.)	
What is the comparison group (or baseline reference group) against which outcomes will be measured?	
Is the comparison group appropriately described? Is use of the comparison group justified?	
What is/are the hypothesis/hypotheses that is/are to be tested?	
What are the characteristics of the study participants, such as their age, grade, race, ethnicity, gender and socioeconomic status?	
What is the location of the study, including indicators of the characteristics of the setting such as region, urbanity and school sizes?	
How was the location chosen?	
Intervention	
Does the report describe the intervention (program, practice, or policy) in sufficient detail for readers to know what is being tested?	
Does the report describe the actual implementation of the intervention studied, including adaptations of content, level and variation in duration and intensity, and technical assistance to program implementers/managers?	
Does the report describe similarities or differences between the intervention studied and other interventions commonly used for similar purposes, including qualities such as duration and intensity, content and delivery, and required and available technical assistance to program implementers/managers?	
Does the report describe the fidelity of implementation of the intervention?	
Does the report present information on the cost of the intervention (if applicable)?	
Comparison Group Conditions	
Does the report describe the comparison condition (counterfactual)? If it includes an intervention, does it describe the comparison intervention and provide details on the actual implementation experience?	
Study Setting	
Does the report include a description of the time period and location of the study, including characteristics of the setting such as region, urbanity, or the size of the project?	
Participants	
Does the report describe the characteristics of the study participants, such as their age, race-ethnicity, gender, and socioeconomic status?	
If the study participants include members of special populations (such as persons with disabilities and dislocated workers), does it describe the process and criteria used to identify those participants, along with their proportion in the study sample?	
Study Design and Analysis	

Criteria and Sub-criteria	Clear and Concise? Y/N
Does the report describe the type of matching process that will be used and the sample size?	
Is the choice of the matching estimator reasonable and justified?	
Is there any attempt to justify the identification strategy?	
Is the justification reasonable?	
Does the report present a histogram of the common support among treatment and control observations, in order to ensure matching is being correctly implemented?	
Does the report discuss the process by which common support is established, including a discussion of any trimming procedures?	
Does the report present tests of balance after matching, along with adjusted and unadjusted means of the comparison group covariates and means for the participants?	
Does the report discuss the model of treatment assignment (probit, logit, etc.)?	
Does the report present all observed variables that are used and discuss potentially omitted variables that may affect the procedure?	
Does the report clearly describe the outcome measures used (whether or not the outcome measure is standardized, how data is collected, etc.)?	
Does the study report the extent to which data is missing?	
Does the report account for missing data (case deletion, nonresponse weights, imputation) for both outcomes control variables?	
Does the report handle missing data in a reasonable way?	
Does the report discuss the bandwidth or "tuning" parameter and how it is selected?	
Are the results sensitive to the choice of tuning parameter?	
Does the report discuss how standard errors for the impacts were estimated?	
Does the report discuss precision and consistency of standard error estimates?	
Is there any reason why treatment assignment may be dynamic? Does the report discuss this?	
Sample Attrition/Nonresponse	
For each key outcome measure, does it include a diagram (e.g. Consort diagram) or table that shows a clear pathway to the final analytic study sample for that outcome, including:	
• Numbers of sites or individuals randomly assigned to intervention and control groups?	
• Numbers for whom outcome data was collected?	
• Numbers of individuals or site that attrited from sample, and reasons for attrition (moved away, absent, refused, site closed)?	
Does the discussion of attrition and response rates include the extent to which the rates attrition and nonresponse differ for the treatment and comparison groups?	
If there is differential attrition after random assignment, does the report mention that as a potential threat to internal validity?	
Tests for Pre-Intervention Treatment and Comparison Group Equivalence	
Does the report provide documentation of sample equivalence (1) at baseline for all randomized sample units (i.e. the initial sample), and (2) for the treatment-comparison analysis (final) sample?	
Does the documentation include sample sizes, means, and standard deviations for key background characteristics and for baseline (pre-intervention) measures of the key outcomes (or closely associated variables)?	
Analytic Approach	

Criteria and Sub-criteria	Clear and Concise? Y/N
Does the report adequately describe the approach to using impact estimates, including models to estimate effects (e.g. regression, ANCOVA, or HLM model) and their appropriateness for the data structure? Do they appropriately account for stratification and clustering?	
If the treatment and comparison groups were not equivalent at baseline, are the characteristics that differed between the treatment and comparison groups included as covariates in the multivariate analysis?	
Does the report provide any rationale for examining subgroups studied, and if so, any approach to estimating effects for sample subgroups?	
Does the report clearly describe any sensitivity analyses conducted?	
If appropriate, does the report account for multiple comparisons by adjusting the critical statistical value to account for the analysis of multiple outcomes within the same domain or use of the same intervention or comparison groups in multiple analyses of the same outcomes?	
Does the report provide any rationale for examining subgroups studied, and approach to estimating effects for sample subgroups?	
Are the results for all outcome measure reported (not just those with significant or positive effects)?	
Is the reporting of results of outcome measures complete (reporting of sample sizes, means, SDs, confidence intervals, significance test results)?	
Are the strengths and limitations of the analyses presented clearly?	
Results	
Are results from the model appropriately presented and discussed?	
Are the results presented in an objective manner? Have they been "cherry-picked"?	
Does the report include the sample sizes, means and standard deviations for key background characteristics and for baseline measures of the key outcomes for the analytic sample? Are the results presented separately for the treatment/control groups?	
Was multiple hypothesis testing conducted? How many outcomes were there?	
If multiple hypothesis testing is conducted, does the report adjust the statistical critical value?	
Is there attrition in the study? Is it large or small?	
Does attrition differ by treatment status?	
How is attrition treated in the study? Is the decision reasonable?	
Does the report provide standard errors in addition to stars/bolding to indicate levels of statistical significance?	
Does the report indicate the duration of time over which outcomes are measured?	
Is the period over which outcomes are estimated sufficiently long enough to effectively capture the effects of the problem?	
Conclusions	
Are the conclusions consistent with the research questions asked?	
Are the conclusions based on objective reporting of information?	
Does the report reach appropriate conclusions or are results overstated and/or not supported by appropriate evidence?	
Does the report make note of any limitations?	
Does the report address sources of potential bias or imprecision?	
Are the conclusions drawn reasonable and/or useful to the implementing agency?	
General Comments	
Is the report concise and clear? Can it be understood by the intended audience?	

Criteria and Sub-criteria	Clear and Concise? Y/N
Did the report identify clearly what is conjecture, speculation or opinion—and the sources of such views?	

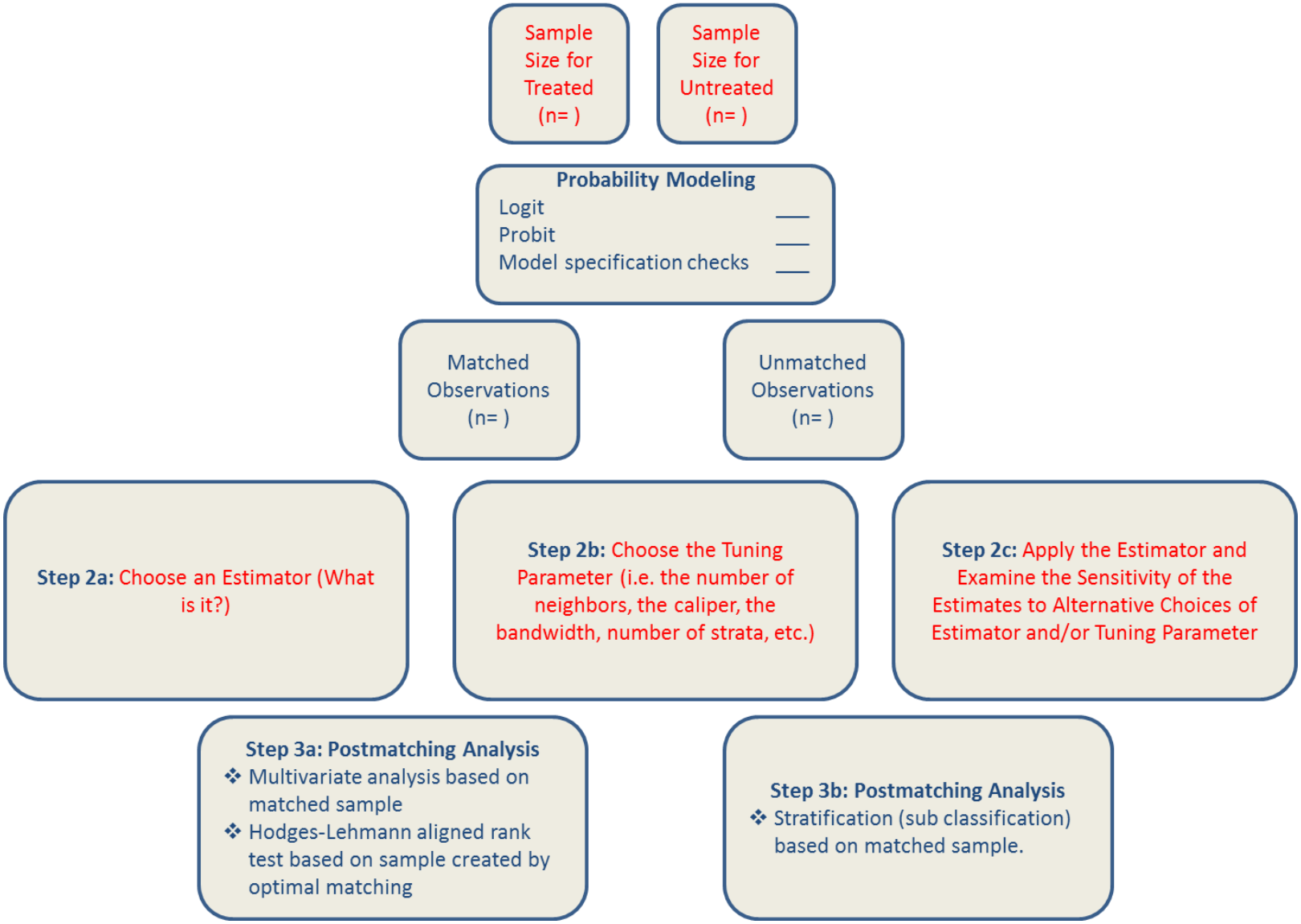
Estimator Ranking (1=strongest, 3= weakest)	Matching Estimator	Overview	Strength	Weakness	Distance Metrics Used	Key Citations
	Mahalanobis Distance Matching (MDM)	MDM is employed by randomly ordering subjects, and then calculating the distance between the first treated subject and all controls.	MDM is a useful estimator to detect outliers, especially in development of linear regression.	Because MDM is not based on a one-dimensional score, it may be difficult to find close matches when many covariates are included in the model. When number of covariates increases, the average Mahalanobis distance between observations increases as well.	Mahalanobis Distance	
	Kernel Matching	This method constructs a match for each program participant using a weighted average over multiple persons in the comparison group.	Because more information is used, lower variance is achieved.	There is a possibility that the observations used are bad matches. Hence, the proper imposition of the common support condition is of major importance. The choice of the kernel function also matters for whether or not all of the comparison units receive a non-zero weight.		Heckman, Ichimura and Todd (1997, 1998).
	Local Linear Matching	Local linear matching is a generalized version of kernel matching, proposed by Heckman, Ichimura and Todd (1997).	Local linear estimation has a faster rate of convergence near boundary points and greater robustness to different data design densities. Therefore local linear regression performs better than kernel estimation in cases where the nonparticipant observations on the propensity score P fall on one side of the participant observations.	Similar weakness to kernel matching		Todd (2008); Heckman, Ichimura and Todd (1997, 1998).

Estimator Ranking (1=strongest, 3= weakest)	Matching Estimator	Overview	Strength	Weakness	Distance Metrics Used	Key Citations
	Inverse Propensity Weighting (IPW)	Inverse propensity weighting uses the inverse of the propensity score to weight each observation in the treated group, and one minus the inverse of the propensity score to weight the controls.	Weighting is useful because it includes all the data (provided weights are non-zero) and does not depend on random sampling, thus providing replicability. Imbens et al. (2003) show that this weighting can produce unbiased estimates of the true treatment effect.	The method does not work well in practice, since observations with a very low probability of being treated have asymptotically large inverses as weights, causing the effect size to be dominated by this value and a high variance in the results (Posner and Ash). Also, unbiasedness requires the weights to be calculated using the true propensity score, which may be hard to estimate in practice.		Imbens et al. (2003)
	Double Robust Estimation	Double Robust Estimation combines outcome regression with weighting by the propensity score such that the estimator of treatment effect is robust to misspecification of one (but not both) of these models (Funk et al., 2011).	Double Robust Estimation allows for one of the two model specifications to be incorrect while still producing unbiased estimates.	If both of the models are incorrectly specified, then there will still be bias in the estimates, and it will be augmented more than if a single model had been used.		Funk et al. (2011)
	Coarsened Exact Matching (CEM)	CEM is a Monotonoic Imbalance Bounding (MIB) matching method --- which means that the balance between the treated and control group is chosen by the user ex ante. Therefore, adjusting the imbalance on one variable has no effect on the maximum imbalance of any other. CEM also strictly bounds through ex ante user choice both the degree of model dependence and the average treatment effect estimation error.	CEM eliminates the need for a separate procedure to restrict data to common empirical support, is robust to measurement error, works well for multicategory treatments, determining blocks in experimental designs, and evaluating extreme counterfactuals. When used properly with informative data, CEM can reduce model dependence and bias and improve efficiency.	Choosing the coarsening parameter appropriately is the primary issue to consider when running CEM. If the parameter is set too large, then information that might have been useful to produce better matches may be missed. If the parameter is set too small, then too many observations may be discarded without a chance for compensation during the analysis stage. Standard issues with matching estimators also apply, such as not matching on an important covariate (unless it is closely related to a variable that is matched on).		Iacus et al. (2011)
	Genetic Matching (GenMatch)	Genetic Matching is a method of multivariate matching that uses an evolutionary search algorithm to determine the weight each covariate is given. It is a generalization of MDM.	Diamond and Sekhon (2012) have shown that GenMatch improves covariate balance and may reduce bias.	The strengths of the estimator require that the "selection on observables" assumption hold, an assumption that is not easily testable in practice.		Diamond and Sekhon (2012)

Estimator Ranking (1=strongest, 3= weakest)	Matching Estimator	Overview	Strength	Weakness	Distance Metrics Used	Key Citations
	Nearest Neighbor Matching (NNM)	This method selects an individual from the comparison group as a matching partner for a treated individual that is closest in terms of propensity score.	NNM allows both with and without replacement in carrying out the estimate. Matching with replacement causes the average quality of matching to increase and the bias will decrease, but fewer cases will be used, reducing precision.	When performing NNM without replacement, estimates depend on the order in which observations get matches. Therefore, it is vital to ensure that ordering is random. Nearest neighbor technique faces the risk of imprecise matches if the closest neighbor is numerically distant.		
	Nearest Neighbor Matching with a Caliper	This method is a variant of nearest neighbor matching. Nearest neighbor matching faces the risk of bad matches, if the closest neighbor is far away. This can be avoided by imposing a tolerance level on the maximum propensity score distance (caliper).	Because NNM faces the risk of bad matches (if the closest neighbor is far away), imposing a caliper imposes a tolerance level on the maximum propensity score distance. Bad matches are avoided and hence the matching quality rises.	Often must make decision between two realities: While trying to maximize exact matches, cases may be excluded due to incomplete matching. While trying to maximize cases, more inexact matching typically results.		
	Radius Matching	Radius matching is a variation of caliper matching that attempts to use not only the nearest neighbor within each caliper but all of the units within the caliper. Radius matching is recommended when the control group is large and there is more than one nearest neighbor.	Radius matching uses only as many comparison units as are available within the caliper and therefore allows for usage of more (fewer) units when good matches are (are not) available.	A drawback of radius matching is that it is difficult to know a priori what choice for the tolerance level is reasonable.		
	Optimal Full Matching (OM)	Optimal Matching is the process of developing matched sets in such a way that the total sample distance of propensity scores is minimized. Optimal Full Matching is a method to generalize Optimal Matching to use all of the available comparison observations.	Optimal matching identifies matched sets in such a way that the process aims to optimize the total distance, and decisions made later take into consideration decisions made earlier.	The first problem with OM is how the OM algorithm uses "insertions" and "deletions." The second problem with OM is the lack of clear benchmarks that can be used to test the results.		
	Fine Balance	Fine balance refers to exact balancing of a nominal variable, often a variable with many discrete categories, but it does not require individually matched treated and control subjects for this variable. Fine balance creates a patterned distance matrix which is passed to a subroutine that optimally pairs the rows and columns of the matrix. See Rosenbaum et al. (2007).	Fine balance does not require individually matching on the propensity score but uses this score to balance participants instead on some meaningful variable.	The principal disadvantage of fine balancing is that it is a constraint on an optimization problem, namely the minimization of the total distance within matched sets, so one can obtain a better or lower minimum total distance by removing the constraint.		Rosenbaum et al. (2007)

Estimator Ranking (1=strongest, 3= weakest)	Matching Estimator	Overview	Strength	Weakness	Distance Metrics Used	Key Citations
	Stratification or Interval Matching	In this method, the range of variation of the propensity score is divided into intervals such that within each interval, treated and control units have, on average, the same propensity score.	Cochrane and Chambers (1965) have shown that five subclasses are often enough to remove 95% of the bias associated with one single covariate.	There is no ideal number of strata to use. One way to justify the choice of the number of strata is to check the balance of the propensity score within each stratum. If propensity score within each stratum is not balanced, the strata are too large and need to be split. In addition, the standard "weakness" is the choice of bandwidth, which leads to a bias-variance tradeoff.		Cochrane and Chambers (1965)

Flow Diagram of Propensity Score Matching Study



Key References Consulted:

Busso, M., DiNardo, J. and McCrary, J. (2011), "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators", Unpublished manuscript.

Caliendo, M. and Kopeinig, S. (2008), "Some Practical Guidance for the Implementation of Propensity Score Matching", *Journal of Economic Surveys*, 22(1), 31-72.

Cochrane, W. and Rubin, D.B. (1973), "Controlling Bias in Observational Studies", *Sankyha: The Indian Journal of Statistics, Series A (1961-2002)*, 35, 417-446.

Cochrane, W. and Chambers, S. (1965),"The Planning of Observational Studies of Human Populations", *Journal of the Royal Statistical Society*, Series A, 128, 234-266.

Diamond, A. and Sekhon, J. (2012), "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies", forthcoming in *Review of Economics and Statistics*

Frölich, M. (2004), "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators", *Review of Economics and Statistics* 86(1): 77-90.

Funk, M. J., Westreich, D., Wiesen, C., Sturmer, T., Brookhart, M. A. and Davidian, M. (2011), "Doubly Robust Estimation of Causal Effects", *American Journal of Epidemiology*, 1-7.

Heckman, J., Ichimura, H. and Todd, P.E. (1997), "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies*, 64, 605-654.

Heckman, J., Ichimura, H. and Todd, P.E. (1998), "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies*, 65, 261-294.

Hirano, K., Imbens, G. and Ridder G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *Econometrica* 71(4): 1161-1189

Hollister, M. (2009), "Is Optimal Matching Suboptimal?", *Sociological Methods Research*, 38, 235-264.

Huber, M., Lechner, M. and Wunsch, C. (2011), "How to Control for Many Covariates? Reliable Estimators Based on the Propensity Score", IZA Discussion Paper No. 5268.

Iacus, S., King, G. and Porro, G. (2011) "Causal Inference without Balance Checking: Coarsened Exact Matching", *Political Analysis*, 1-24.

Posner, M. and Ash, A. "Comparing Weighting Methods in Propensity Score Analysis", Unpublished Manuscript.

Rosenbaum, P.R. and Rubin, D.B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 1, 41-55.

Rosenbaum, P.R., Ross, R. and Silber, J. (2007), “Minimum Distance Matched Sampling With Fine Balance in an Observational Study of Treatment for Ovarian Cancer”, *Journal of the American Statistical Association*, 102, No. 477, Applications and Case Studies, 75-83.

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

Todd, P. E. (2008), "Matching Estimators", The New Palgrave Dictionary of Economics. Second Edition. Eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan