MULTIPLE SAMPLING WHEN TREATMENT UNITS ARE MATCHED TO NUMEROUS CONTROLS USING PROPENSITY SCORE

Paper presented at the American Evaluation Association Annual Conference

November 7 – 10, 2007, Baltimore

Shu Liang

Paul Bellatty

Evaluation and Research

Oregon Department of Corrections

Abstract

Quantifying the impact of a social program requires sound experimental designs and sound statistical analyses. Generating accurate and unbiased treatment estimates requires comparability of treatment and control units. Ideally, individual units should be randomly assigned to the treatment or control groups so that pre-treatment differences are negated. Pragmatic and ethical considerations often prohibit random assignment. When random assignment is not available, propensity score matching is a useful tool for constructing a comparable control group. Often the number of potential control units far exceeds the number of treatment units. A single treatment unit may be matched with many control units. A common practice is to randomly select one of the potential control matches, create a control group with the same number of treatment units, and perform statistical analyses on this matched control-treatment sample. However, sometimes conclusion may vary with a different sample. This paper proposes an alternative: Use multiple sampling to generate treatment effect when a single treatment unit is matched to numerous controls.

Multiple Random Sampling When Treatment Units are Matched to Numerous Controls Using Propensity Scores

Introduction

Evaluation research uses methods that appropriately quantify the impact of a social program. Conclusions derived from a well designed study using reliable methods provide justification for continuation or termination of the program (Smith, 2000; See Bryson et al., 2002).

Generating accurate and unbiased treatment estimates requires comparability of treatment and control units. Ideally, individual units should be randomly assigned to the treatment or control groups. The randomization of units to control and treatment groups provides assurances that no systematic differences in observed or unobserved covariates (i.e., bias) exist between groups (D'agostino, 1998). Consequently, it allows comparability of the treatment and control groups and provides a reliable means of quantifying the effectiveness of programs.

Pragmatic and ethical considerations often prohibit the use of random assignment. "Even if it is feasible, the randomization may be comprised by noncompliance and other missing data problems" (Hirano et al., 2002, p2). Under these circumstances, evaluation researchers may find it practical and plausible to utilize an observational study. An observational study is "an empirical investigation of treatments, policies, or exposure and the effects they cause when assignment of treatments is not controlled" (Rosenbaum, 1995; cited in Quigley et al., 2003).

When researchers have no control over the assignment of the units to treatment or control group, differences in observed covariates may exist between the treatment and control groups (D'agostino, 1998). There could be other problems such as self-selection or systematic judgment by the researcher when selecting units to the treatment (Dehejia & Wabah, 1998). With all these drawbacks, a direct comparison of the outcomes of the treatment and control groups may be biased and misleading (D'agostino, 1998; Dehejia & Wabah, 1998; Hirano & Imbens, 2002).

Evaluation researchers must adjust for pre-treatment differences to ensure that the treatment and control groups are comparable. There are different ways of achieving comparability between treatment control groups. For example, evaluation researchers may incorporate data on observed covariates when estimating treatment effects through stratification or covariate adjustment. Evaluation researchers may also incorporate data on covariates into the study design through matched sampling (D'agostino, 1998). Matched sampling is a frequently used method to construct a probabilistically equivalent group (Heckman, 1989; Lalonde, 1986; Manski &Garfinkel, 1992; Rsonebaum, 1989; 1995; Rosenbaum & Rubin, 1983; see Quigley, 2003).

A specific matched sampling methodology that is becoming increasingly popular is propensity scoring. Simply stated, the propensity score for an individual unit refers to the probability that an individual unit would receive the treatment of interest based on the observed covariates of the individual. Evaluation researchers match individuals from the treatment group with a similar demographic twin who did not receive the treatment using propensity scoring. The fundamental idea is: "If we use the probability that a subject would have been treated (that is, the propensity score) to adjust our estimate of

the treatment effect, we can create a ‘quasi-randomized’ experiment. That is, if we find two subjects, one is in the treated group and one is in the control, with the same propensity score, then we could imagine that these two subjects were ‘randomly’ assigned to each group in the sense of being equally likely to be treated or control” (D’agostino, 1998, p 2267). Using propensity scores to construct a probabilistically equivalent control group has advantages over traditional matching methods. With traditional matching (selecting controls similar on all important characteristics), it is often difficult to find enough matches, even when there are only a small number of relevant covariates. This is “because the number of matching cells increases exponentially with number of covariates and cells could quickly become empty of treatment individuals, or control cases, or both” (O’Conniffe et al, 2000, p 288). Propensity score matching, on the other hand, summarizes all covariate information simultaneously into a single value irrespective of the number of covariates; as a result, propensity scoring can accommodate variability across a large number of observed covariates and still establish probabilistically equivalent groups (D’agostino, 1998; Dehejia & Wahba, 1998; Rosenbaum, 1995; Rosenbaum & Rubin, 1983, 1985).

While coming up with the idea of using propensity score to construct equivalent control groups has required highly innovative thinking, the actual creation of the propensity score is relatively simple. One method is to use a logistic regression, using all identified relevant factors as the independent variables and the group membership (being in the treatment or control) as the dependent variable. Based on the logistic model, a propensity score is created for each individual and this score represents the

likelihood of being in the treatment. For each treatment unit, a control matched unit is identified and selected for the control group.

Generally, there are far more units who did not receive the treatment (i.e., control group) than units who received treatment (i.e., the treatment group). Often, the methodology finds numerous similar twins in the control group for a given treatment individual. Researchers can refine the matching process so that fewer matches are produced. However, there can be negative effects associated with refining the matching process. As the search for a matching control unit is refined, control units cannot always be identified for each treatment unit. Treatment units without twins are eliminated from the analysis. A highly refined matching process may select very well-matched pairs but can substantially reduce the number of treatment individuals, rendering it difficult to generalize to all individuals receiving the treatment.

A less refined matched paired group of treatment and control units creates a different problem. Evaluation researchers may be confronted with a large number of potential matches in the control group for each participant in the treatment group. The large number of potential matches in the control group for each treatment unit implies that many units in the control group have the same propensity score. If equal sizes are preferred for both groups, how do researchers select the single matched control unit for each treatment unit?

Propensity score matching assumes that all pertinent covariates are included in creation of the propensity score. However, unrecognized or emergent factors often influence the outcome variable. Furthermore, even if we know some factors are relevant, they may not be available in the database. Two individuals who may appear

similar with limited data may actually differ substantially for some important variables. Having a large pool of control units to identify matched units has some benefits. Randomly selecting a matched control unit for every treatment unit can temper the influence of covariates not included in the database. Despite some benefits, the same methodology can be problematic. Results from a single matched comparison will never be identical to the next single matched comparison using the same prospective group of control units. Thus a large pool of prospective control units allows for more chances of matching to the treatment units; but at the same time, single sample matching rarely provides the same results for every matched sample, if the multiple matched control units have different outcomes. Theoretically, every random selection can provide different results with each randomly selected control sample. This issue prompted the development of the current methodology.

Proposed method

Randomly selecting control units for each treatment unit has benefits; however, different results for each matched sample are problematic. This paper suggests an additional methodology to eliminate the potential differences among analyses using single matched groups. The proposed methodology uses many randomly selected one-to-one matched samples and averages treatment effects across a large number of samples. Running 1,000 simulations with randomly selected control groups should provide more accurate estimates of treatment effectiveness. We believe treatment estimates based on a large number of potential random samples should be more closely approximate the true effect. Specifically, the methodology includes these steps:

(1) identify variables to match treatment and control groups; (2) create a propensity score for each individual using the logistic regression analysis (with variables identified in the previous step (1) as the independent variables and group membership as the dependent variable); (3) examine the size of the matched sample with different precision levels; (4) identify the best level of precision by considering the number of treatment units remaining in the sample and the quality of the matching; (5) create a "control" group by matching to those given treatment at the chosen level of precision and perform the pre-determined statistical analysis on the sample; and (6) perform multiple random sampling when numerous control individuals are matched with a single treatment individual.

An empirical example for illustration

To illustrate the methodology, two examples are presented. The first example prompted the proposed methodology of multiple random sampling. We were asked to evaluate a program at the Oregon Department of Corrections (DOC). The research question was: "Does participation in the program reduce the likelihood in recidivism?" The data included 1,747 inmates who were released from DOC between January 2004 and May 2005. Approximately, 300 participated in the program and the remaining 1,447 offenders did not participate. Since inmates were not randomly assigned to treatment or control conditions, we used propensity score matching methodology to obtain a comparable control group.

Step 1: Identify variables pertinent to the study. Fourteen factors were statistically related to recidivism including: (1) inmates' age at admission to the prison;

(2) being African-American; (3) being Hispanic; (4) time since release from prison; (5) percent of earned-time accrued; (6) number of prior incarcerations; (7) previous revocation; (8) prior theft conviction; (9) number of custody cycles; (10) severity of the crime associated with current incarceration; (11) sentence length; (12) drug and alcohol treatment needs at intake; (13) conviction of person-to-person crime; and (14) conviction of property or statutory crime.

Step 2. A logistic regression was performed to obtain a propensity score for each individual. The treatment condition is a dichotomous variable: being a participant was coded 1 and being a non-participant was coded 0. This variable served as the dependent variable of the logistic model and the above list of factors served as independent variables.

Step 3. We examined various levels of matching precision, ranging from the most restrictive matching (.00001), to moderate matching (.0001), to the least restrictive matching (.001). With the most restrictive matching (.00001), only 15 treatment individuals were match with controls. At .0001, sixty-seven treatment individuals were matched with sixty-seven controls. The matched number increased to 210 for each group when matching was done at .001. Matching at .001 appears to be more reasonable than the more restrictive matching. Table 1 gives a summary of the matched control-treatment numbers of various level of matching.

Table 1. Level of matching and number of observations

Level of matching	Treatment	Control
.001	210	210
.0001	67	67
.00001	15	15

Step 4. Does the matching at .001 provide similar treatment and control groups? To answer this question, treatment and control groups were assessed using the covariates before and after matching. Comparison among covariates with different measurement units is not appropriate. To ensure covariates are comparable, the mean difference between the groups was standardized (for the formula, see D'agostino, 1998). If control and treatment groups differ substantially for a single covariate, the standardized value would approximate 100 or -100. If the control and treatment groups were nearly identical for a given covariate, the standardized value would approximate 0. Before matching, the standardized differences for the covariates ranged from 2.8 to 55.8. The control and treatment groups differed on all covariates except three: being African-American, number of prior incarcerations, and prior theft conviction. The largest standardized values were for drug and alcohol needs (55.8), and percentage of earned-time (50.4). See Table 2 for a summary of the covariates before matching.

Table 2. Group comparisons before matching

Variable	Treatment (STD)		Control (STD)		Standardized Difference	P-value
AGE AT ADMISSION	33.78	(9.15)	32.48	(9.75)	13.72	0.03
BLACK	0.06	(0.23)	0.08	(0.27)	-10.06	0.09
TIME SINCE RELEASE	381.44	(127.14)	408.13	(149.46)	-19.24	0.00
EARNED TIME	1.70	(0.53)	1.39	(0.68)	50.39	0.00
PRIOR INCARCERATIONS	1.54	(2.17)	1.48	(2.28)	2.84	0.66
REVOCATION	0.76	(0.43)	0.71	(0.46)	13.09	0.04
THEFT CONVISTION	0.34	(0.47)	0.30	(0.46)	8.19	0.19
CUSTDODY CYCLES	1.88	(0.97)	1.66	(0.92)	22.82	0.00
SEVERITY OF CRIME	366.17	(88.11)	384.69	(92.15)	-20.53	0.00
SENTENCE LENGTH	2.82	(0.84)	3.05	(1.54)	-18.66	0.00
DRUG/ALCOHOL NEEDS	2.87	(0.43)	2.53	(0.77)	55.75	0.00
PERSON-TO-PERSON CRIME	0.20	(0.40)	0.31	(0.46)	-25.33	0.00
PROPPERTY CRIME	0.30	(0.46)	0.37	(0.48)	-13.21	0.04

After matching, no statistically significant differences existed between groups See Table 3 for a summary of the comparison.

Table 3. Group comparison after matching at .001

Variable	Treatment (STD)	Control (STD)	Standardized Difference	P-value
AGE AT ADMISSION	33.44 (9.10)	32.85 (9.07)	6.45	0.51
AFRICAN-AMERICAN	0.06 (0.24)	0.05 (0.21)	6.27	0.52
TIME SINCE RELEASE	388.19 (133.19)	381.06 (137.77)	5.26	0.59
EARNED TIME	1.60 (0.58)	1.63 (0.58)	-4.10	0.67
PRIOR INCARCERATIONS	1.56 (2.29)	1.82 (2.57)	-10.57	0.28
REVOCATION	0.75 (0.44)	0.76 (0.43)	-2.20	0.82
THEFT CONVICTION	0.33 (0.47)	0.39 (0.49)	-10.91	0.26
CUSTODY CYCLES	1.79 (0.95)	1.81 (1.05)	-1.91	0.85
SEVERITY OC FRIME	376.04 (91.75)	370.29 (89.78)	6.34	0.52
SENTENCE LENGTH	2.90 (0.86)	2.91 (0.91)	-1.08	0.91
DRUD/ALCOHOL NEEDS	2.82 (0.50)	2.77 (0.51)	11.34	0.25
PERSON-TO-PERSON CRIME	0.21 (0.41)	0.24 (0.43)	-6.79	0.49
PROPERTY CRIME	0.35 (0.48)	0.38 (0.49)	-4.94	0.61

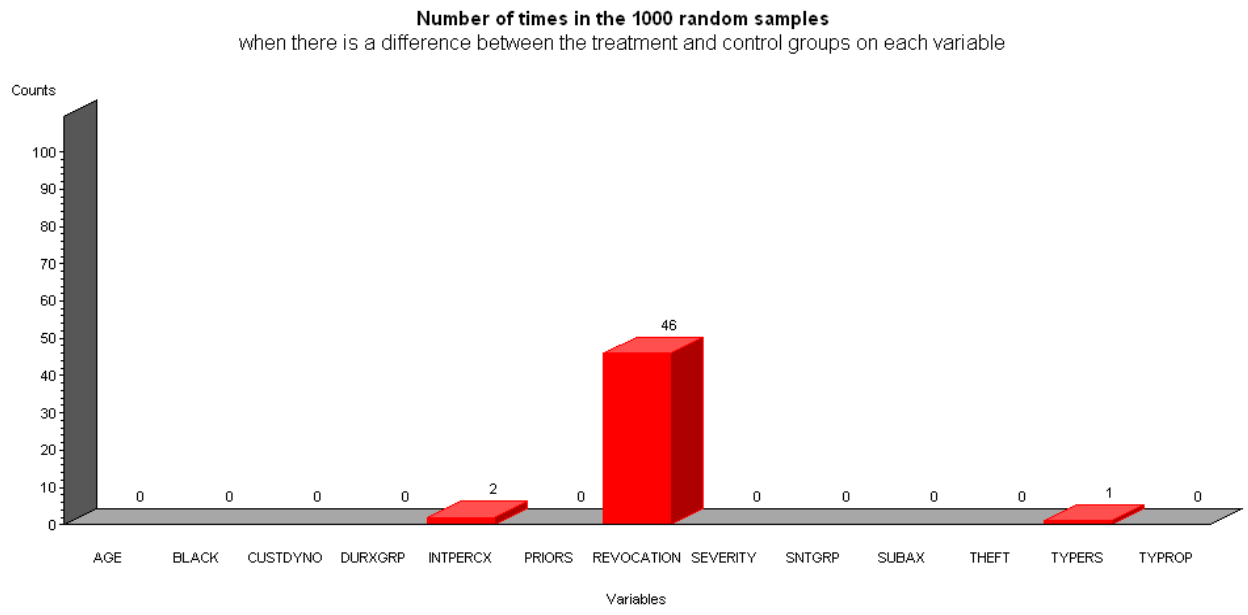
An additional logistic regression analysis was performed to assure comparability was achieved between control and treatment groups (Quigley et al., 2003). If the treatment units are well matched with the control units, the logistic regression should not be able to predict group membership. The results revealed no association between the set of covariates and group membership ($x^2(13) = 7.14$, $p=.90$); in addition, none of the individual covariates predicted membership. These analyses suggest a well-matched sample has been established.

Step 5. After adequate matching had been attained, statistical analysis (survival analysis) was performed on the single matched sample. The result indicated individuals in the treatment group were 35% less likely to recidivate than the individuals in the control group ($x^2(1) = 4.50$, $p=.034$). The conclusion with the single matched sample was the program was effective in reducing recidivism.

Step 6. A single random sample is often used to determine the effectiveness of many social programs. A better methodology might include multiple random sampling. If the single random sample includes control matches that are representative of all

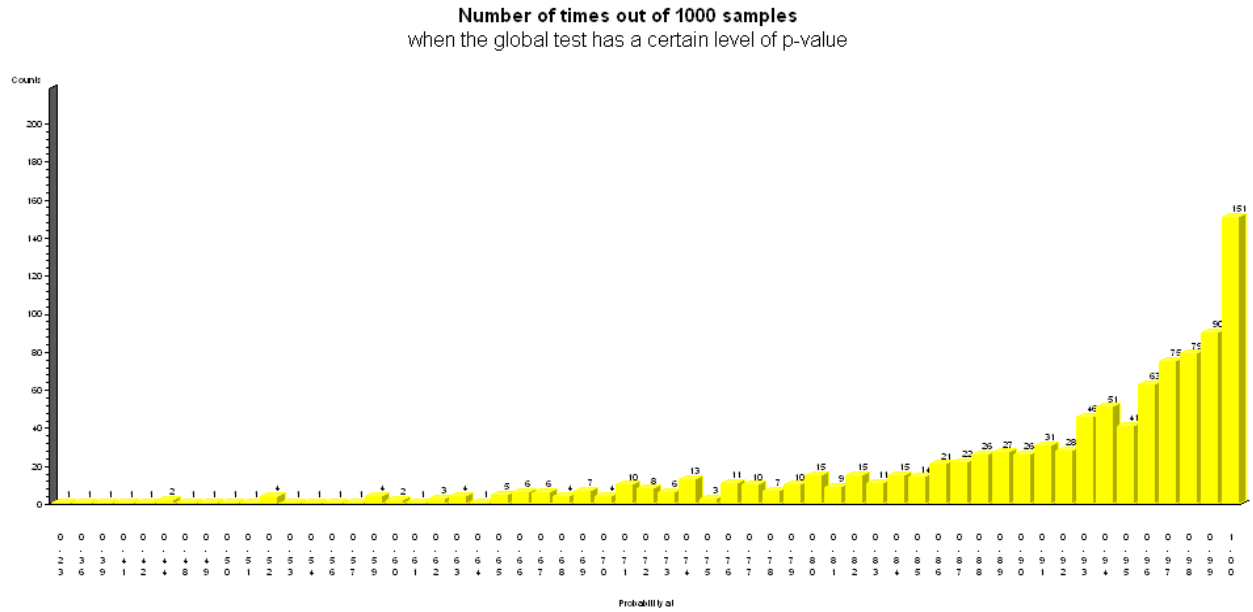
potential matches, the single and multiple random samples would provide similar conclusions. For this analysis, multiple random sampling was also performed. With multiple random sampling, tests on covariates and comparability of groups were performed. Figure 1 presents the distribution of each covariate when 1000 random samples were used as matched controls. Using the .05 significance level, there were no statistically significant differences with all covariates. Only forty-six samples showed a difference in revocation (by chance, we would expect 50).

Figure 1



For each sample, a logistic regression was performed to determine if groups could be predicted using the covariates. In none of the 1000 samples, treatment conditions could not be reliably predicted from the covariates. Figure 2 presents the distribution of the significance of the overall tests.

Figure 2



These analyses suggest all covariates were similar between the treatment and control groups. A Cox regression was performed on each of the 1000 samples and only 2.7% of the samples (27 out of 1000) included statistically significant treatment effects; most samples indicated no treatment effect. Using the methodology of multiple random sampling, the treatment as deemed ineffective. While the single-sample analysis revealed a significant treatment effect with a hazard ratio of .65, the mean treatment effect using the 1000 samples was .81; most treatment effects using the 1000 samples were not statistically significant. There was large variation in the estimates across the samples. The conclusion was the treatment was not effective and the hazard ratio estimate was likely to be about .81; but it could range from .58 to 1.20. Figure 3 presents the distribution of the significance of the tests for treatment effects using the 1000 samples and Figure 4 presents the distribution of the hazard ratio.

Figure 3

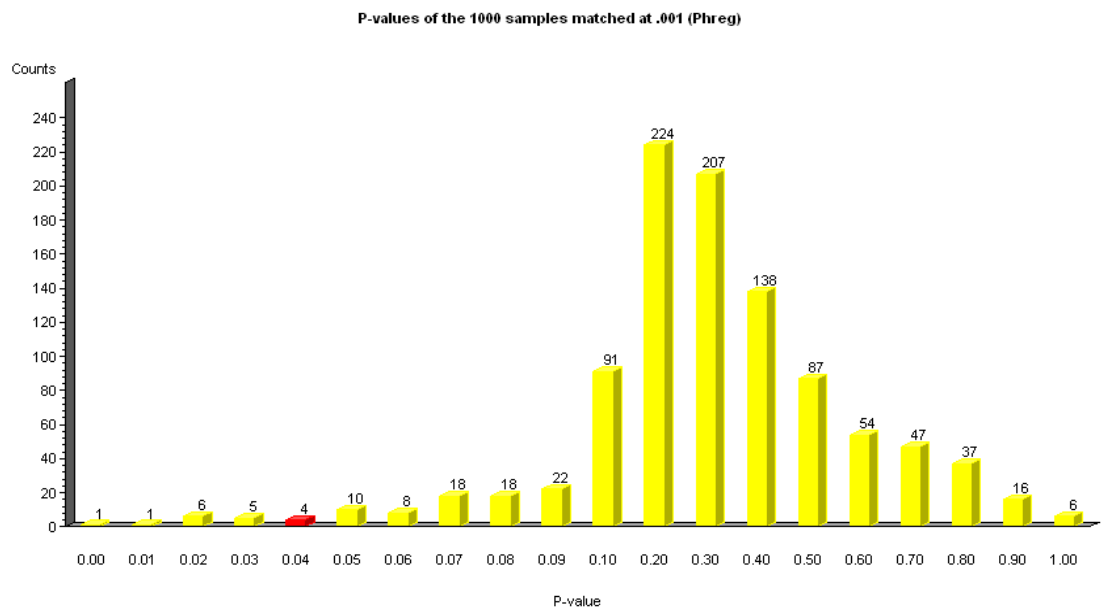
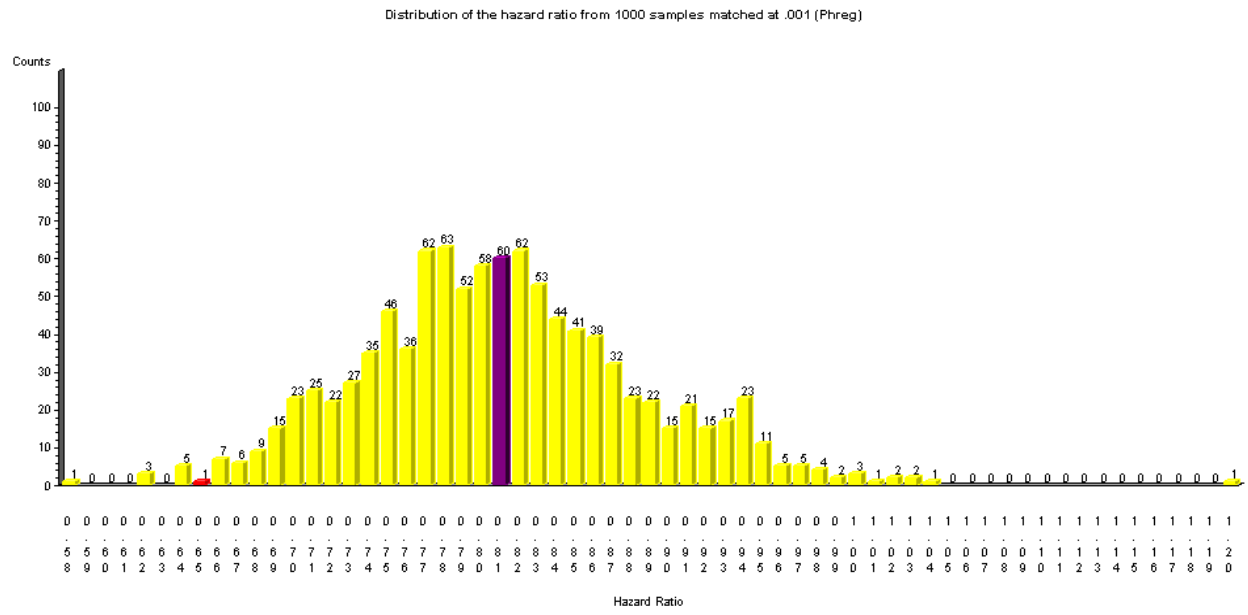


Figure 4



Validation of the method with a randomized experiment

When a single matched control-treatment comparison suggests a program is effective and a subsequent analysis of the same dataset suggests the program is ineffective, methodological as well as interpretive questions arise. For example, from which matched sample should the researcher draw the conclusion? The proposed methodology appears to be effective in dealing with the problem. The question remains: How do the results from a multiple sampling methodology compare with the results from a randomized experiment?

Randomized experiment

Oregon DOC researchers have access to a randomized study of parenting programs provided to offenders. In the study, inmates who are parents were randomly assigned to the treatment (participating in a parenting program) or control group. Data were available on 167 participants and 138 non-participants. Background information was available on fourteen covariates: (1) inmates' age at admission to the prison; (1) being an African American; (3) being a Hispanic; (4) time since release from the prison; (5) percent of earned-time accrued in prison; (6) number of prior incarcerations; (7) previous revocation; (8) prior theft conviction; (9) number of custody cycles; (10) severity of the crime prompting the incarceration; (11) sentence length; (12) drug and alcohol treatment need; (13) conviction of person-to-person crime; and (14) conviction of a property or statutory crime. The treatment group (parenting) and the control groups were similar on all covariates except time since release from the prison and amount of earned-time. The treatment group had shorter time since release for prison and less earned-time. Table 4 provides a descriptive summary of the comparison.

Table 4. Comparison of the randomized treatment and control group

VARIABLE	PARENTING	STD	CONTROL	STD	DIFF	STND DIFF	P-value
BLACK	0.17	0.37	0.10	0.30	0.07	19.43	0.09
HISPANIC	0.07	0.26	0.06	0.23	0.01	5.62	0.63
PRIOR INCARCERATIONS	0.66	1.25	0.80	1.41	-0.14	-10.49	0.36
REVOCATION	0.66	0.47	0.75	0.44	-0.08	-17.94	0.12
SEVERITY	365.86	103.24	379.48	96.24	-13.62	-13.65	0.24
MALE	0.44	0.50	0.46	0.50	-0.03	-5.34	0.64
SINCERELEASE	642.04	329.90	800.83	380.49	-158.8	-44.59	0.00
THEFT	0.38	0.49	0.42	0.50	-0.04	-7.54	0.51
AGE AT ADMISSTION	29.58	6.86	30.18	7.01	-0.60	-8.65	0.45
CUSTODY CYCLES	1.53	0.81	1.57	0.90	-0.03	-3.78	0.74
EARNED TIME	1.40	0.49	1.53	0.50	-0.13	-26.99	0.02
MENTAL PROBLEM	0.53	0.50	0.52	0.50	0.01	2.24	0.85
PERSON-TO-PERSON CRIME	0.53	0.50	0.51	0.50	0.03	5.13	0.66
SENTENCE LENGTH	23.69	21.08	19.86	16.76	3.83	20.09	0.08
SUBSTANCE TREATMENT NEED	1.95	1.33	1.83	1.38	0.11	8.31	0.47

A Cox regression analysis was conducted to examine the relationship between participation in the parenting program and likelihood to recidivate after release from the prison. All covariates were considered in the statistical mode. Only age at admission, number of prior incarcerations, and conviction of a person-to-person crime were significantly associated to recidivism and were included in the final model. After adjusting for these covariates, the treatment effect did not attain statistical significance. The estimated hazard ratio was 1.056 (p=.83), which means that the participants were as likely to recidivate as the non-participants. See Table 5 for results.

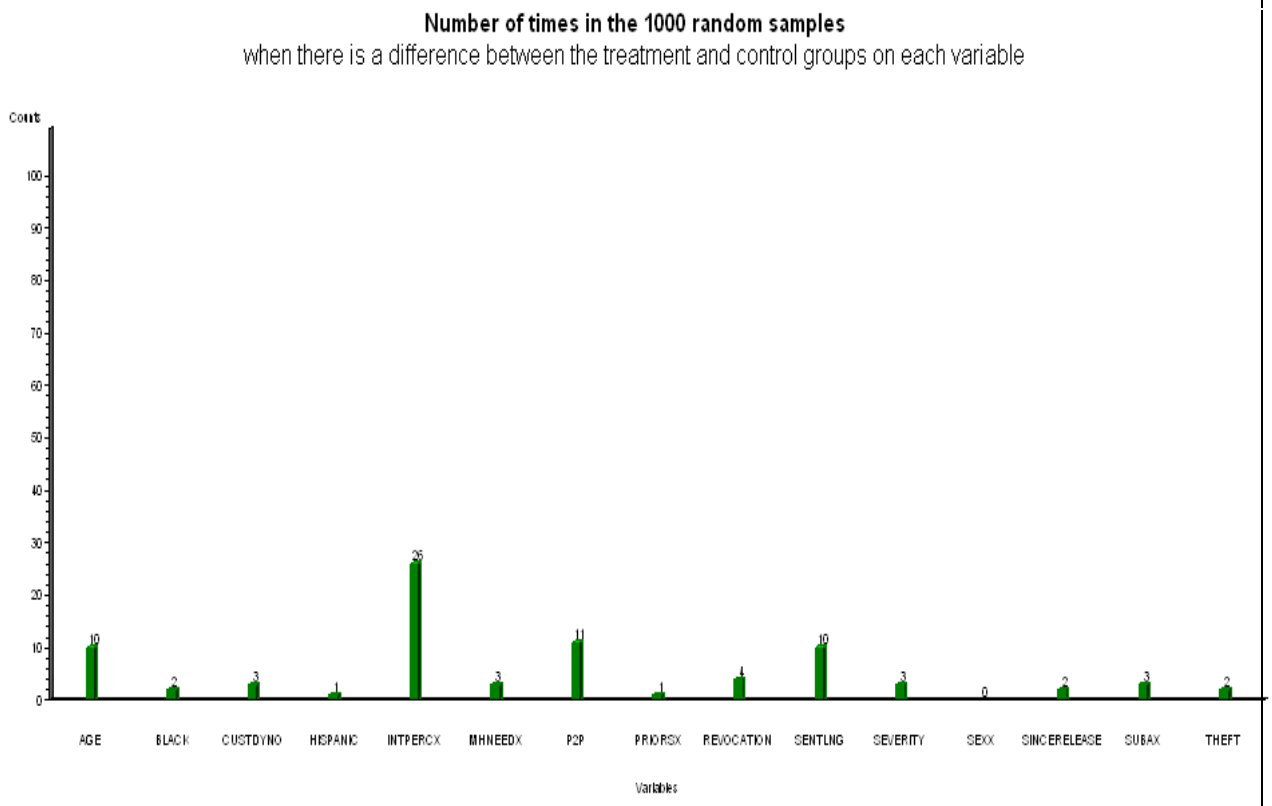
Table 5. Results of the Cox regression analysis

Analysis of Maximum Likelihood Estimates								
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Parenting	1	0.05491	0.25176	0.0476	0.8273	1.056	0.645	1.730
Prior incarceration	1	0.17791	0.07906	5.0642	0.0244	1.195	1.023	1.395
Age at admission	1	-0.06281	0.02277	7.6096	0.0058	0.939	0.898	0.982
Person-to-person crime	1	0.53082	0.26950	3.8796	0.0489	1.700	1.003	2.884

Multiple random sampling of matched controls

A pool of 10,568 inmates being released from Oregon DOC prisons was considered as potential controls. Using the fourteen covariates, a logistic regression was conducted to create a propensity score for each individual. One hundred and forty-seven treatment members were matched to at least one individual not receiving parenting classes. In many instances, one treatment member could be matched with many controls. Multiple samples (n=1000) were randomly selected using the .001 level of precision. Before conducting the planned statistical analysis, a comparison was made between groups for all covariates and all samples. The groups were determined comparable across all samples. For all covariates except earned-time, the percent of samples with statistically significant covariates seldom exceeded 1%. For earned-time, the percent was 2.6% (see Figure 5).

Figure 5



In addition to comparing the means of the covariates, a logistic regression was conducted to determine if group membership could be predicted. The result indicated that all covariates did not predict whether an individual would be a participant or a non-participant in any sample. The smallest p-value across the samples was .13 (see Figure 6). An examination of the discriminability index (coefficient gamma) indicated that these covariates did not distinguish the groups well. The gamma coefficient gamma was generally small with about 80% of the samples having a value less than .2 (Figure 7).

Figure 6

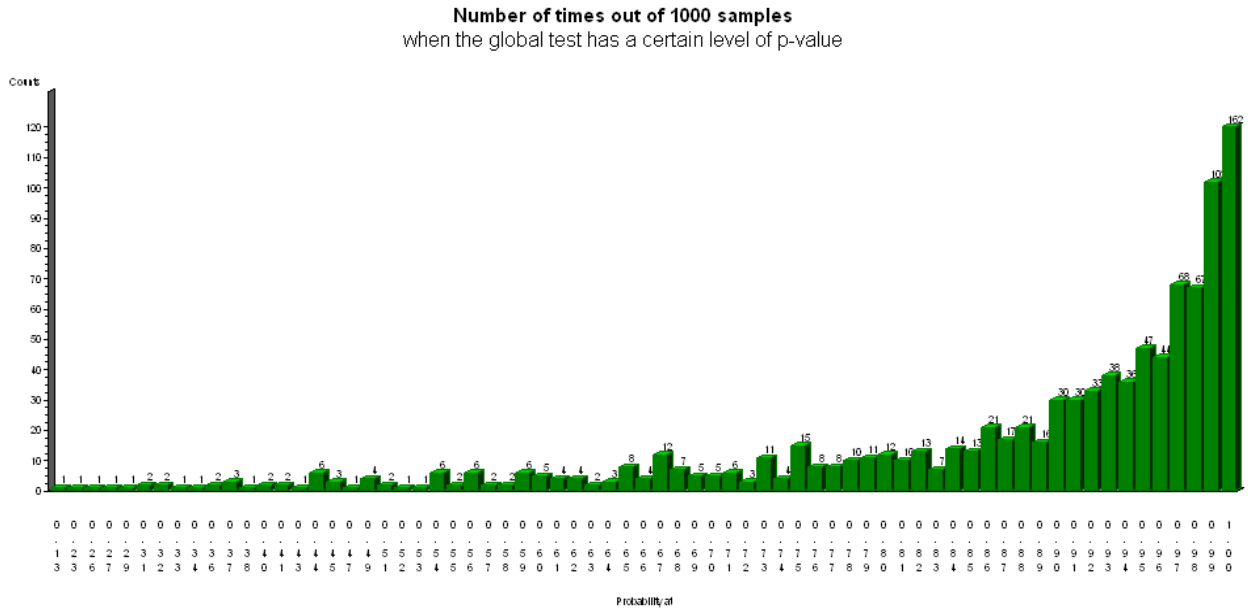
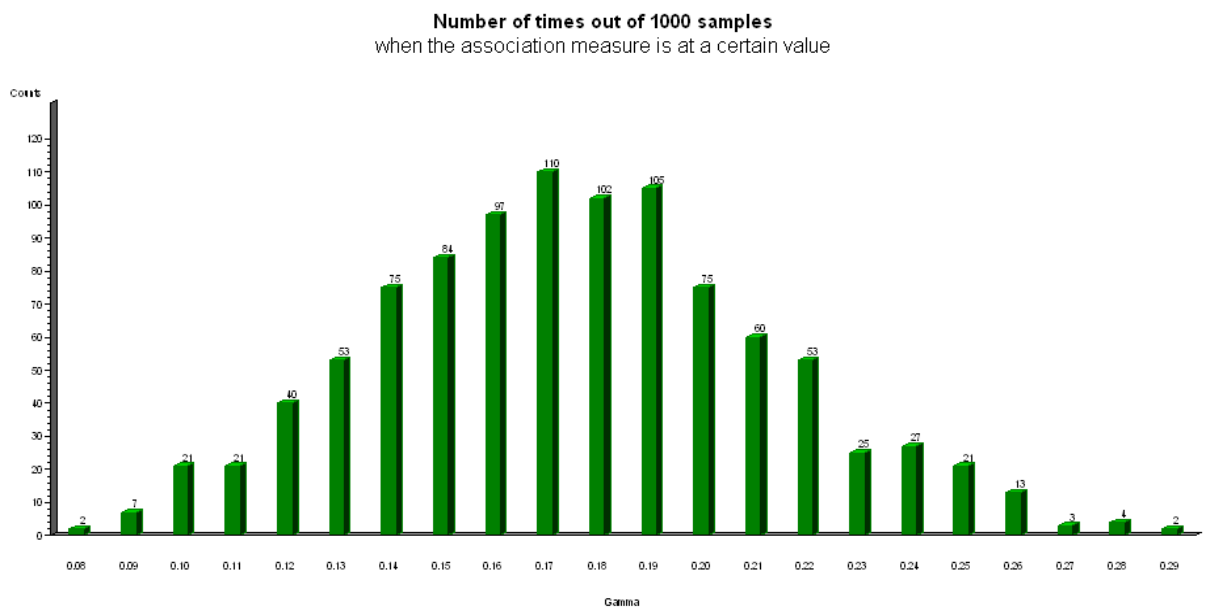
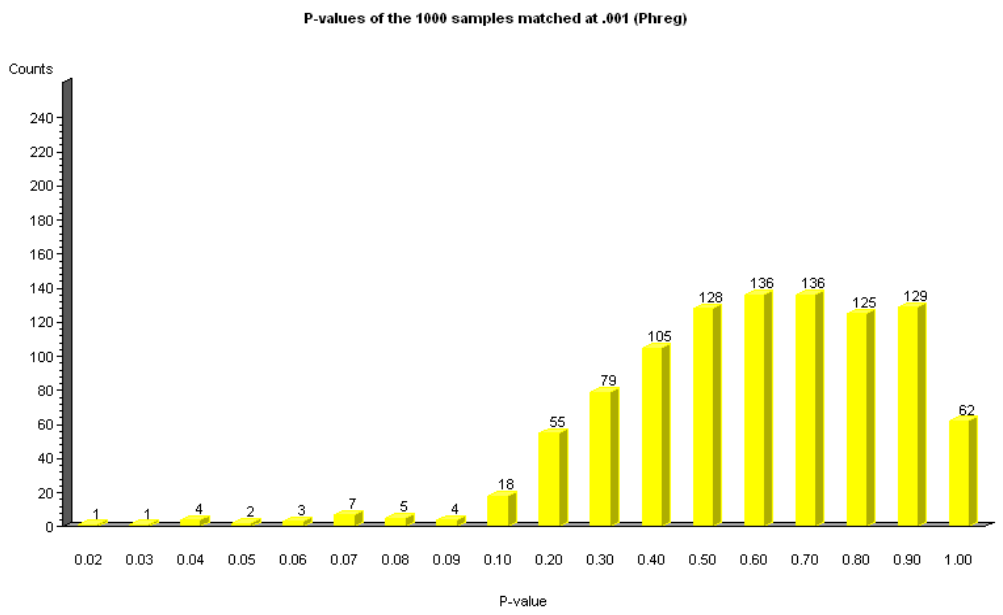


Figure 7



The groups were well balanced with all the covariates. A Cox regression was performed on each sample and there was little evidence of a treatment effect. Of the 1,000 samples, only eight had a statistically significant effect at .05. If the criterion was changed to .10, the number of samples with statistically significant treatment effects was 45 (Figure 8).

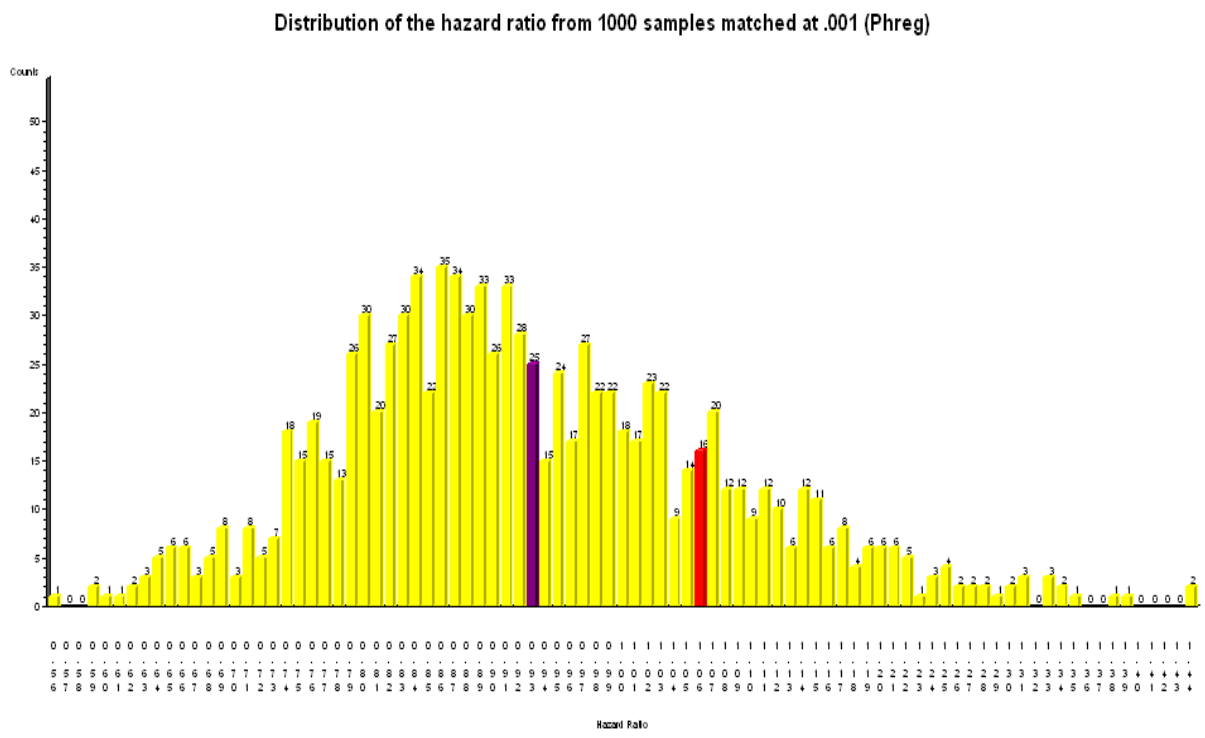
Figure 8



The mean treatment effect (hazard ratio) across the samples was .93. Although not identical to the 1.056 from the randomized sample (.93 is as close to 1.00 as 1.06 is), the conclusion is the same; there was no treatment effect. Figure 9 presents the distribution of the estimated effects involving the 1,000 samples. While the estimated 95% confidence interval for the treatment effect (hazard ratio) of the randomized sample was .65 to 1.73, the range of the estimate of the 1000 samples ran from .56 to 1.44. The

95% confidence interval of the matched sample whose hazard ratio equaled the mean of the 1000 samples was .56 to 1.60. Although the ranges derived from multiple random sampling is slightly smaller, the fundamental conclusion remains the same: There is no evidence of a treatment effect (see Figure 9).

Figure 9

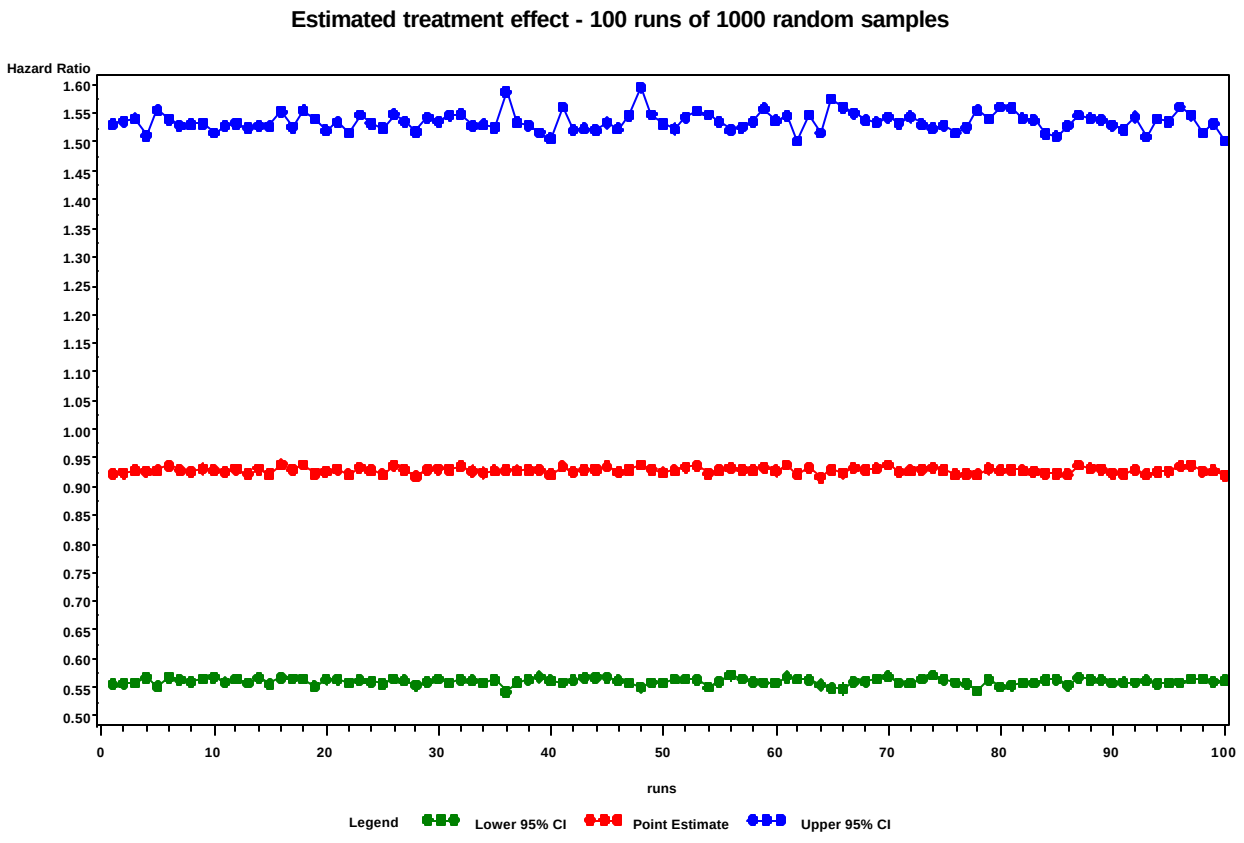


Consistency of the results with more repetition of the procedure

Based on the results of multiple sampling, the treatment effect was negligible with both methodologies. Will this conclusion remain the same with a different run of 1,000 samples? Stated differently, will the mean effect estimate have a similar range? One hundred repetitions of the multiple random sampling were conducted. Figure 10 provides a summary of the 100 repetitions of multiple sampling (n=1000 for each

repetition). The horizontal axis represents the order of repetitions. The middle dotted line is the mean of each 1000 samples. The lower and upper dotted lines are respectively for the lower and upper 95% confidence interval of each repetition. The graph clearly shows consistency of the solution across the repetitions. This implies multiple samplings provide similar results for all 100 sets of analyses performed. One multiple sampling analysis appears sufficient with data used.

Figure 10



Conclusion

Random assignment is generally the preferred experimental design if social and ethical issues can be addressed. When random assignment is not pragmatic, other designs or statistical procedures are appropriate. Some quasi-experimental designs adjust for differences between control and treatment groups and enable estimation of the treatment effects. Propensity scoring can match treatment units with control units and provide good estimates of program effectiveness. When multiple control units possess the same propensity scores as a single treatment unit, randomly selecting control units is often the logical solution. Random selection of control units rarely provides the same treatment estimates when multiple analyses are performed. To better estimate the actual treatment effect, many random samples of the control units provide improved estimates over the single matched control-treatment sample. This methodology provides accurate and consistent estimates of the treatment effects.

References

Bryson, A., Dorsett, R., & Purdon, S. (2002). *The use of propensity score matching in the evaluation of active labor market policies.* United Kingdom: Policy Studies and National Center for Social Research.

O'Conniffe, D., Gash, V., & O'Connell, P.J. (2000). *Evaluating state programs: "Natural experiments" and propensity scores*. The Economic and Social Review, 31, (4), 283-308.

D'agostino, R.B., Jr. (1998). *Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group.* Statistics in Medicine, 17, 2265-2281.

Dehejia, R.H. & Wahba, S. (1998). *Propensity score matching methods for non-experimental causal stidies* (National Bureau of Economic Research Working Paper No.6829). Cambrige, MA: National Bureau of Economic Research.

Heckman, J.J. (1989). *Causal inference and nonrandom samples*. Journal of Educational Statistics, 14, 159-168.

Lalonde, R. (1986). *Evaluating the econometric evaluations if training programs*. American Economic Review, 76, 604-620.

Manski, D.F. & Garfinkel, I. (1992). *Introduction.* In C. Manski & I. Garfinkel (Eds.), *Evaluating welfare and training programs (pp. 1-22).* Cambridge, MA: Harvard University Press.

Hirano, K. & Imbens, G.W. (2002). *Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization*. Downloaded from the internet.

Rosenbaum, P.R. (1995). *Observational studies*. NewYork: Springer-Verlag.

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

Rosenbaum, P.R. & Rubin, D.B. (1983). *The central role of the propensity score in observational studies for causal effects*. Biometrika, 70, 41-55.

Quigley, D.D., Munoz, J., & Jacknowitz, A. (2003). *Using a matched sampling methodology to evaluate program effects: An illustration from the university of California outreach programs.* Downloaded from the internet.

Smith, J. (2000). *A critical survey of empirical methods for evaluating active labor market policies,* Schweiz. Zeitschrift fur Volkswirtschaft und Statistik, 136, (3) 1-22.