

# Propensity Score Matching

## An Illustrative Analysis of Dose Response

E. MICHAEL FOSTER, PHD

**BACKGROUND.** Health services researchers are often interested in the effect of a treatment or a service in situations in which randomization is difficult or impossible. One useful alternative involves propensity score methods, a means for matching members of different groups based on a range of characteristics. Under certain assumptions, comparisons of the matched groups reveal the impact of the treatment of interest.

**OBJECTIVES.** This article reviews propensity score methods and illustrates their use in an analysis of dose response, the relationship between the volume of services received, and treatment outcomes. In mental health policy, this question is central to key issues such as parity.

**RESEARCH DESIGN.** Data for the illustrative analysis are taken from a well-known study of children's mental health services. This analysis estimates the impact of outpatient therapy based on comparisons of individuals receiving different treatment doses. Those comparisons are adjusted for preexisting observed differ-

ences among the groups using propensity score methods.

**SUBJECTS.** The study includes 301 participants aged 5 to 18 years treated at the study sites.

**MEASURES.** The analyses are based on family characteristics and the mental health status of children and adolescents reported in interviews with parents as well as administrative data on service use.

**RESULTS.** Analyses using propensity score matching suggest that added services improve treatment outcomes, especially child functioning. However, at least for the services and outcomes considered, the marginal benefits to high levels of treatment are limited.

**CONCLUSIONS.** These analyses illustrate the potential value of propensity score methods to health services researchers.

**Key words:** Econometrics; mental health; children's health; propensity score; dose; dose response; mental health services. (Med Care 2003;41:1183–1192)

For many questions of interest to health services researchers, randomization is difficult or impossible. For example, one might be interested in whether physician adherence to treatment guidelines improves patient satisfaction or in comparing individuals who receive psychotropic medications from family physicians with those under the care of psychiatrists. In those situations, randomizing individuals to different levels or types of care is

difficult. As a result, researchers must rely on naturalistic studies in which they infer the impact of different types or levels of treatment from differences among individuals receiving different care.

Inferring causality in studies of this type is difficult. Simple comparisons of individuals receiving different treatments are potentially misleading or biased in that they do not reveal the effect of

---

From the Department of Health Policy & Administration, Pennsylvania State University, Philadelphia, Pennsylvania.

Address correspondence and reprint requests to E.

---

Michael Foster, PhD, Associate Professor of Health Policy and Administration, The Pennsylvania State University, 116 Henderson Building North, University Park, PA 16802-6500. E-mail: emfoster@psu.edu

treatment per se. (More specifically, these comparisons are biased in small samples and inconsistent in statistical terms. The latter refers to the fact that the bias does not disappear as sample size increases.<sup>1</sup>) In many cases, these comparisons confound the effect of the treatment with that of the factors that lead individuals to select different treatments.

Health services researchers encounter this problem frequently. In mental health services, for example, researchers are often interested in questions of "dose response" or the link between the volume of services received and treatment outcomes. This question is central to a series of policy issues such as parity.<sup>2,3</sup> Insurance coverage for mental health services traditionally has been less generous than that for physical health. For example, the number of outpatient visits covered in a given year could be lower for mental health. Various forms of parity legislation such as the Mental Health Equitable Treatment Act have been proposed that would eliminate or reduce this disparity. The implications for patient well-being depend on dose response.

Randomizing individuals to different levels of treatment is difficult. One might assign individuals into different regimes in which treatment is more or less available or encouraged. This might be accomplished by subsidizing the use of treatment by or providing free transportation for one group. However, one cannot compel individuals to receive more treatment than they desire. Also, no one can generally prevent individuals from receiving additional services above and beyond those provided in a study. If, as seems likely, some individuals in the targeted group do not take advantage of the increased availability of treatment or pursue treatment outside the study, the comparison of the intervention and control groups still provides rather weak evidence of the effect of received treatment per se.

For that reason, the best evidence of how dosage affects outcomes is likely to come from naturalistic studies of individuals actually using different amounts of services. As discussed previously, however, interpreting the results of such studies is difficult. Fortunately, this area of research is an active one in which new and better methods are being developed. These methods include propensity score matching, which is increasingly common.<sup>4-8</sup> Under certain assumptions, this set of methods produces estimates of the impact of treatment by matching individuals receiving alternative levels or types of treatment.

Until recently, propensity score matching methods were limited to situations in which individuals either receive a treatment of interest or do not. Recently, however, these methods have been extended to multivariate treatments in which individuals could receive one of several treatments.<sup>9</sup> This extension makes the methods much more useful for examining dose-response questions in which a simple dichotomy is harder to define.

This article describes the propensity score methodology and applies it to data from the field of children's mental health services. The data are derived from a well-known study, the Fort Bragg Demonstration. That study was recently highlighted in the U.S. Surgeon General's report on mental health services.<sup>10</sup> We use these data to consider the question of whether increased outpatient visits improve mental health outcomes (symptomatology and functioning).

## Prior Research

Two areas of prior research serve as background for the subsequent analyses. The first involves propensity score methods. The second concerns the analysis of dose response in the area of children's mental health services.

## Propensity Score Methods

As noted previously, the central methodologic challenge in examining dose response involves distilling the effect of dose per se from that of the factors that lead individuals to use different amounts of services. This section discusses the means we used for doing so, propensity score matching. In this section, we outline the limitations of a regression approach, explain propensity score matching as an alternative, and discuss its strengths and weaknesses. Although we illustrate this discussion by referring to dose response, the method could be applied to a wide range of problems.

One means for adjusting between-group comparisons for preexisting differences involves ordinary regression. The various predictors of dose and outcomes could be included as regressors. Although common, such an approach has some limitations.<sup>6,7</sup> The first is that regression generally assumes a set of linear relationships between the covariates and the outcome of interest (in this

case, mental health status). Although this assumption can be relaxed (using interactions or polynomials), the models are fundamentally linear. A second, more subtle problem involves the so-called "support" or distribution of the covariates. Not only might the high- and low-dose groups differ in terms of the means of those variables, but the distribution of those variables could overlap relatively little. In that case, regression essentially projects the behavior of individuals in one group outside the observed range to form a comparison for the other at common values of the covariate. Such projections can be highly sensitive to functional form.

An alternative to regression involves statistical matching. These methods generally specify a function measuring the proximity of one case to another based on one or many characteristics.<sup>6,7</sup> Cases are then grouped to minimize the distance between matched cases. Unlike regression, the various forms of matching do not presume linearity. This approach also identifies problems with the support of the covariates. One could identify individuals in one group that are a poor match to anyone in the other group(s). One then might exclude those individuals from the analysis.

One challenge in matching is that the process is quite complex when many characteristics are involved. To address this problem, recent research has explored the use of the propensity score, the predicted probability that an individual receives the treatment of interest. Rosenbaum and Rubin<sup>7</sup> demonstrate that the propensity score captures all of the variance in the covariates relevant for adjusting between-group comparisons. As a result, one can simply match the 2 (or more) groups based on this single variable. As discussed subsequently, propensity score matching can be implemented in the form of probability weights for use in analyses of treatment and outcomes.

The principal benefit of propensity score matching, therefore, lies in convenience. In terms of statistical efficiency (or precision of estimation), these methods do not dominate conditioning on the covariates in all situations. In some instances, controlling or adjusting for all covariates produces more precise estimates<sup>11</sup> (eg, in which the probability of treatment is close to 0 or 1).

It is important to note that both ordinary regression and propensity score matching still rest on a critical assumption, "strong ignorability."<sup>8</sup> In behavioral terms, this assumption means that for common values of covariates, the choice of treat-

ment is not based on the benefits of alternative treatments. In other words, among individuals with the same characteristics used for matching, the model assumes that these individuals are sorted into different treatments as if randomly assigned. This assumption is rather strong; its plausibility depends on the particular treatment involved and on the range of covariates included in the analysis.

Building on the basic framework, the literature on propensity score matching has flourished in the last 20 years. It even has been applied to dose-related questions. In looking at the impact of an intervention, the Comprehensive Child Development Project, Hill et al. examined the impact of the intervention for individuals who participated beyond a minimal level. Although the intervention showed no overall long-term effects, propensity score matching revealed benefits for the highly engaged subgroup.<sup>12</sup>

Until recently, propensity score methods have been limited to 2-group situations such as a single treatment and a comparison group. This limitation has stunted the widespread adoption of the method. Imbens, however, extends the method to multigroup situations. His work builds on the fact that propensity score matching essentially involves a weighting scheme.<sup>9</sup> The weights are formed as the inverse of the predicted probability that an individual would make the choice he or she actually made. In instances in which 3 or more groups are of interest, one can use a polytomous choice model (such as multinomial or ordered logit) to estimate the probability that an individual selects one of the available treatments. The resulting predicted probabilities are then used to create weights that are used in subsequent analyses.<sup>9</sup>

### **Prior Research on Dose Response in Children's Mental Health Services**

Dose response has been an ongoing area of research within mental health services. Our review focuses on 4 recent studies.<sup>11,13-17</sup> Using data from the Fort Bragg study, Salzer et al.<sup>16</sup> examined the relationship between the number of outpatient visits and mental health outcomes, including symptomatology and functioning. Based on ordinary regression methods, their analyses included only baseline measures of the outcomes as covariates. Salzer and colleagues generally found no impact of dose; their findings were robust across a

range of outcomes and methods. (In a related paper, Andrade and colleagues found no relationship between outcomes and a series of dichotomous treatment categories.<sup>17</sup>)

Angold examined the impact of dose using data from a community study of children and youth in North Carolina.<sup>14</sup> That study gauged the impact of visits on symptoms and functioning. Data were available for 2 interviews before the period for which dose was measured. These data revealed that the conditions of individuals receiving higher doses were deteriorating before the period for which dose was defined. These results suggest that analyses that control only for level of mental health status could underestimate the impact of dose. Like earlier studies, the author found some variation in the effect of dose across outcomes: dose affected symptoms only. In addition, the effect of dose was nonlinear. In fact, clients who had only 1 or 2 sessions actually fared worse than the untreated. Real improvement required at least 8 sessions.

Foster provides a second analyses of outpatient use and client outcomes in the Fort Bragg data.<sup>13</sup> That article applied instrumental variables estimation to the problem of dose response. That method does not require ignorability and adjusts for unobserved or unmeasured characteristics that might be confounded with dose. The use of this method depends on finding an "instrumental variable," a characteristic that affects the outcome of interest (mental health status) *only through its effect on the regressor of interest* (dose).<sup>18</sup> Foster used study site and date of entry into the study as instrumental variables. Uncovering a relationship hidden in ordinary regression, instrumental variables estimation revealed that dose substantially improved functioning. Ordinary regression analyses seriously underestimated the impact of dose because of a sizable correlation between unobserved individual characteristics and dose. These results, however, depend on the validity of the instrumental variables themselves.

Bickman and colleagues also use instrumental variables estimation. They examined data from another prominent study in the field, the evaluation of a system of care in Stark County, Ohio.<sup>19</sup> Using instrumental variables estimation, the authors found no link between mental health outcomes and treatment dosage, which they measured as the total costs of services. These findings illustrate the potential limitations of instrumental variables estimation: the results are only as strong

as the instrumental variables are valid. In this case, the underlying assumptions were implausible; the authors assumed that a long list of characteristics, including a child's diagnosis or characteristics of his or her parents such as problems with substance abuse affected mental health outcomes only through dose. If this assumption is incorrect, the resulting estimates have poor statistical properties and are misleading.

## Data

This article examines data from the Fort Bragg Demonstration. In this section, we briefly describe the demonstration and discuss the sample analyzed and measures used.

### The Fort Bragg Demonstration

Initiated in 1990, the Fort Bragg Demonstration provided a test of the "Continuum of Care" philosophy of providing children with mental health services. (For an overview of the Fort Bragg Demonstration and Evaluation [FBEP], see Bickman et al.<sup>19</sup>) The FBEP evaluated the impact of the Demonstration on service use, costs, and mental health outcomes. It compared children and adolescents treated between 1990 and 1993 at Fort Bragg with similar individuals at 2 comparison sites.

### Sample

The focus of the FBEP was on the experiences of a sample of 984 children and adolescents aged 5 to 18 years who received services at the comparison and demonstration sites.<sup>19</sup> The analyses were limited to individuals receiving only outpatient services and for whom follow-up data were available. In particular, a total of 431 observations used only outpatient services in the 12 months after study entry. Of these observations, 130 were dropped from the analysis because 12-month follow-up data were unavailable. (To improve the precision of the parameter estimates involved (and of the resulting propensity scores), the full sample was used in the multinomial logit analysis presented subsequently. Also dropped from that analysis were 14 cases with missing data for the covariates.

The information used to identify individuals who received only outpatient services and to determine the number of outpatient visits received

within a year of entry into the study was derived from 2 sources. The primary source of utilization data for the comparison site was claims from the military's insurer (the payor at the comparison site). At the demonstration, a central clinic either arranged or provided all services and maintained a management information system tracking service use. (Based on a substudy involving a review of medical records for a sample of individuals at each site, the 2 sources appear fairly comparable in terms of their accuracy and completeness.<sup>19,20</sup>)

Because both sources were available from 1988 forward, these data also could be used to identify individuals who received mental health services before the start of the study.

## Measures

Interviews were conducted with children and their parents at entry into the study and at 6-month intervals thereafter. That information included client and family demographics such as age, gender, race (dummy coded for black vs. white), and parent education. Key mental health outcomes include symptomatology (measured by the Child Behavior Checklist [CBCL]<sup>21</sup>) and functioning (measured by the Child and Adolescent Functioning Assessment Scale [CAFAS]).<sup>22,23</sup> These measures provide outcomes for these analyses. As discussed subsequently, baseline scores on these measures also were used in calculating the propensity score as were prior service use and the demographic variables described previously.

## Results

As noted, individuals receiving different amounts of services might differ systematically, and one strategy is to adjust for various individual characteristics that potentially influence both dose and outcomes. The data provide a rich source of such information, and Table 1 describes the sample in terms of characteristics used in the analyses reported here.

For the purposes of the analyses here, we characterized dose with 7 categories to capture any nonlinearities between service use and outcomes (see Table 2). The boundaries were somewhat arbitrary, but provided a fairly even distribution of the sample across the resulting categories and captured nearly all the variance in dose (88%).

In the first stage of our analyses, we estimated a multinomial logit regression predicting dosage using the characteristics in Table 1 as explanatory variables. Table 3 presents the results of those analyses. The resulting coefficient estimates were used to calculate propensity scores for each dosage category. As discussed previously, the inverse of that probability was used to create a weight. The propensity score-adjusted estimates presented here incorporated these weights.

Table 4 describes levels of improvement over time using the 7 groups described previously. Improvement was measured as the change over time in the outcome measures. (Because the measures were scored so that higher scores indicated worse mental health, the actual scores generally decreased over time.) We assessed statistical significance for the overall relationship using ordinary regression that included 6 dummy variables (with the first, lowest-dose category serving as the omitted group). For the CBCL, we found no relationship between dose and symptomatology ( $P = 0.28$ ). The relationship between dose and functioning (CAFAS) was significant ( $P = 0.02$ ).

As discussed previously, the relationship between dose and functioning (or the lack of one between symptomatology and dose) could be explained by systematic differences among individuals receiving different doses. Table 4, therefore, also presents the findings based on propensity score matching (ie, using the probability weights described previously). Like in the raw comparisons, dosage was unrelated to symptomatology (CBCL) in the propensity score analysis. The relationship between dose and the CAFAS, however, remained significant in the propensity score analysis ( $P = 0.02$ ). The striking difference between the adjusted and unadjusted figures involved the second group, individuals receiving between 3 and 6 visits. When one accounts for the other covariates, this group's situation actually *deteriorated* during the 12-month follow-up period. One can see the relative position of this group most clearly in Figure 1, which plots the unadjusted and propensity score-adjusted figures.

This finding was consistent across a variety of alternative model specifications. The dip remained, for example, when we used an ordered logit to generate the predicted probabilities and the corresponding weights. In additional analyses, we increased the range of characteristics used to generate the propensity scores including, for example, clinical diagnosis and other demographics

TABLE 1. Sample Characteristics (n = 301)

Characteristic	Mean	Standard Deviation
Demonstration	66%	0.47
Female	41%	0.49
Age	10.15	3.51
Race (nonwhite = 1; white = 0)	23%	0.42
Parental Education		
College graduate	26%	0.44
High school graduate only*	15%	0.36
Previous service use	16%	0.36
Baseline functioning (CAFAS)	36.71	21.51
Baseline symptomatology (CBCL)	63.50	10.18

\*Omitted or baseline category: caretaker has some postsecondary education.

Information on the number of outpatient visits taken from CHAMPUS claims and management information system data from the Fort Bragg Demonstration. Other data collected as part of interviews conducted as part of the Fort Bragg Evaluation.

CAFAS = Child and Adolescent Functioning Assessment Scale; CBCL = Child Behavior Checklist.

such as family income. The resulting analyses were similar to those presented here. (See the third series in Fig. 1.) We also examined the sensitivity of our findings to the boundaries used to define dose categories. In supplemental analyses, we divided the second category into 2 groups (3 or 4 visits and 5 or 6 visits). The dip at 3 to 6 visits was clearly driven by the 3- or 4-visit group.

## Conclusions

In many circumstances, the best source of information on the effect of a treatment or service

TABLE 2. Dose Categories (n = 301)

	Percentage
1 or 2 Visits	12
3-6	13
7-11	16
12-18	16
19-24	12
25-35	14
37+	17
	100

involves comparisons of individuals who choose alternative treatments. Propensity score methods represent a promising means for improving such comparisons by providing a flexible and convenient way to adjust for preexisting between-group differences. These methods, however, adjust only for between-group differences in measured characteristics and are no panacea. In any particular application, the range of available measures largely determines the strength of the method. The "strong ignorability" assumption on which propensity score methods rest remains a relatively strong one, but analysts can make it more palatable by incorporating a wider array of individual characteristics in their analyses.

This discussion leaves open the question of how many covariates to actually include in developing propensity scores. One standard for judging the adequacy of the covariates is provided by prior research on the treatment of interest itself. In that light, the model used to determine propensity scores should incorporate the predisposing, enabling, and need factors prior research has identified as shaping service use. To the extent systematic determinants are omitted, and those characteristics affect or are correlated with the outcomes of interest, the analysis is subject to omitted variable bias. For this reason, the analyst should err on the side of caution in determining which covariates to include in adjusting between-group comparisons.

Propensity score methods recently have been extended to situations involving more than 2 groups. In this form, propensity score matching involves a form of weighting, and this article illustrates that approach using data from a prominent study in the field of children's mental health services. Our findings suggest that increases in the number of outpatient visits, after a point, improve functioning. Our findings also reveal some diminishing returns to additional visits at higher levels of use. This level is well below that at which visit maximums are typically set by insurance companies. Although subject to some imprecision in this analysis (as a result of sample size), the returns beyond 12 to 18 visits are uncertain at best, at least for these data and the outcomes considered here. Whether this finding would hold in larger, more representative datasets is an important question for future research. If the finding does replicate, it suggests the emphasis on parity could be a bit misplaced.

TABLE 3. Multinomial Logit Analysis of Dose (n = 417)

Covariate		Dose Category (no. of visits)						
		1-2	3-6	7-11	12-18	19-24	25-35	37+
Site (1 = demonstration; 0 = comparison)	Marginal effect	-0.216	-0.214	0.051	0.068	0.092	0.093	0.127
	SE	0.041	0.044	0.037	0.038	0.031	0.034	0.032
	P value	0.00	0.00	0.17	0.08	0.00	0.01	0.00
Gender (1 = male; 0 = female)	Marginal effect	0.003	-0.058	0.097	0.043	-0.034	-0.001	-0.051
	SE	0.026	0.033	0.044	0.044	0.035	0.039	0.036
	P value	0.91	0.08	0.03	0.33	0.34	0.99	0.15
Age	Marginal effect	0.000	-0.006	0.008	0.003	0.001	-0.001	-0.005
	SE	0.004	0.005	0.006	0.006	0.005	0.005	0.005
	P value	0.94	0.22	0.17	0.65	0.82	0.93	0.35
Race (nonwhite = 1; 0 = white)	Marginal effect	0.086	-0.024	-0.009	0.043	0.001	-0.067	-0.029
	SE	0.038	0.036	0.043	0.048	0.038	0.039	0.039
	P value	0.02	0.50	0.83	0.37	0.98	0.09	0.46
Parental education (Omitted or baseline category: caretaker has some postsecondary education)								
College graduate	Marginal effect	-0.001	0.038	-0.038	-0.030	0.088	-0.027	-0.029
	SE	0.032	0.049	0.053	0.056	0.059	0.051	0.051
	P value	0.96	0.45	0.47	0.59	0.14	0.60	0.57
High school graduate only	Marginal effect	-0.062	-0.060	-0.008	0.019	-0.045	0.034	0.122
	SE	0.024	0.036	0.045	0.049	0.038	0.047	0.052
	P value	0.01	0.09	0.85	0.69	0.24	0.48	0.02
Previous service use	Marginal effect	0.050	-0.078	-0.052	0.047	0.024	0.019	-0.011
	SE	0.043	0.034	0.050	0.062	0.051	0.053	0.048
	P value	0.24	0.02	0.30	0.45	0.63	0.72	0.83
Baseline functioning (CAFAS)	Marginal effect	0.001	0.000	-0.002	0.001	-0.002	0.001	0.000
	SE	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	P value	0.08	0.80	0.04	0.26	0.02	0.11	0.77
Baseline symptomatology (CBCL)	Marginal effect	-0.002	-0.002	-0.002	-0.004	0.001	0.002	0.007
	SE	0.001	0.002	0.002	0.002	0.002	0.002	0.002
	P value	0.08	0.23	0.46	0.11	0.52	0.47	0.00

LR chi<sup>2</sup> (54) = 168.91  
 Prob > chi<sup>2</sup> = 0.0000

Presented are the marginal effects derived from the multinomial logit regression coefficients. Those P values less than 0.05 are emboldened.

SE = Standard error; CAFAS = Child and Adolescent Functioning Assessment Scale; CBCL = Child Behavior Checklist.

The benefits of higher dose seem limited to functioning rather than symptomatology. This divergence could be either substantive or methodologic in nature. Such variation is common but poorly understood. (See, for example, the literature on dose response for adults.<sup>24-28</sup>) This variability is clearly an area for future research.

Like Angold, we find that individuals who participate in treatment for a few visits and then drop out fare poorly. The poor performance of this group could be explained by the fact that these children and/or their parents find the first few therapy sessions unsettling but might not persist long enough to experience the benefits of treatment. (In addition, those sessions could sensitize

parents to the children's problems.) Just as likely is that this finding reflects forms of selection not captured by the propensity score methodology. Expanding the list of covariates does diminish the drop-off for this group somewhat.

From a service-delivery perspective, this group deserves special attention regardless of the explanation. Clinicians could lack the resources to contact these patients for follow up, but managed care companies might use a form of low-intensity case management to encourage these clients to continue their treatment. Given the poor outcomes for these individuals, this practice could be cost-effective. Extending the episode of outpatient therapy could avoid costly inpatient care later.

TABLE 4. Impact of Dose on Key Outcomes\*

Dosage Level	Unadjusted		Adjusted <sup>†</sup>	
	CAFAS	CBCL	CAFAS	CBCL
1-2	12.42	6.73	7.59	6.28
3-6	10.57	8.98	-4.13	8.84
7-11	13.00	10.83	19.86	11.77
12-18	24.32	9.29	24.14	9.52
19-24	12.50	8.06	13.71	6.36
25-35	26.34	9.95	22.74	8.91
37+	19.11	9.76	17.94	9.17
R-square	0.06	0.02	0.10	0.04
Observations	270	301	260	290
Significance level <sup>‡</sup>	0.02	0.28	0.02	0.19

\*So that higher scores indicate greater improvement, we reverse coded the CAFAS and CBCL in calculating the measure of change.

<sup>†</sup>These estimates are propensity score-adjusted using the weighting scheme in Imbens.<sup>9</sup> The weights are based on the propensity scores (or predicted probabilities) calculated using the results of the multinomial logit model presented in Table 3.

<sup>‡</sup>Significance level refers to the null hypothesis that there is no variation across the dose categories.  
CAFAS = Child and Adolescent Functioning Assessment Scale; CBCL = Child Behavior Checklist.

Because these analyses are based on the same data as 2 other articles, some effort to reconcile the different findings is instructive. These results contradict those in Salzer et al.<sup>16</sup>; one explanation is that those authors control for only baseline differences in the key outcome measures. As Angold

demonstrates, substantial heterogeneity exists even among those individuals with similar scores at baseline. The apparent absence of a link between dose and outcome, therefore, could reflect the fact that individuals receiving higher doses were more seriously ill, masking the benefits of

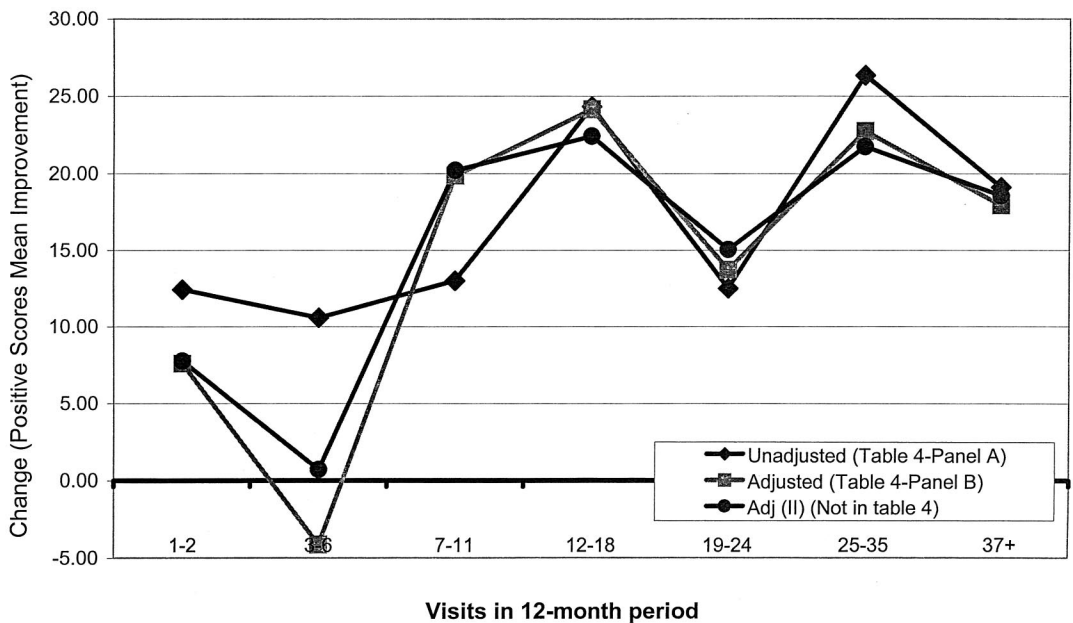


FIG. 1. Change over time—functioning.

added treatment. Other explanations are possible such as a more rigid functional form. By using 7 categories, the analyses presented here capture substantial nonlinearities that would be masked in ordinary regression (even one in which dose was log-transformed).

On the other hand, the results presented here are similar to those in the study by Foster.<sup>13</sup> As discussed, the instrumental variables estimation used in that paper relaxes the strong ignorability assumption but replaces it with other assumptions (involving the choice of instrumental variables). For either propensity score or instrumental variables methods, the assumptions involved are largely untestable.

How does one choose between the 2 methods? Obviously, the presence of a plausible instrumental variable is one criterion. If the instrument is contrived or a poor predictor of treatment status, then the properties of instrumental variables estimation are poor.<sup>1,8</sup> Another criterion involves the range of available covariates on which to match groups receiving different treatments. With a greater range of variables, the strong ignorability assumption becomes more plausible, and the propensity score methods become more attractive.

Another criterion involves the analyst's knowledge of the phenomenon involved. The instrumental variables approach does address a broader range of potential problems, including unobserved between-group differences.<sup>8</sup> That method corrects for measurement error in dosage as well. (Such error might exist in services research because of problems with key sources of services data [eg, data entry] or because individuals access services not captured by available data.) Where such problems seem unlikely, the benefits of instrumental variables estimation might be outweighed by the costs stemming from the additional assumptions required.

### Acknowledgments

The author acknowledges the research assistance of Jess Pohl Taylor, who reviewed the adult literature for the article. The author also thanks Guido Imbens, Pamela Farley Short, Damon Jones, Robert Saunders, Beth Gifford, Jennifer Hill and Allison Olchowski for their helpful comments. The author is responsible for any remaining errors.

### References

1. **Davidson R, MacKinnon JG.** Estimation and Inference in Econometrics. New York: Oxford University Press, 1993.
2. **Mechanic D, McAlpine DD.** Mission unfulfilled: potholes on the road to mental health parity. *Health Aff* 1999;18:7-21.
3. **Mechanic DS.** Emerging trends in mental health policy and practice. *Health Aff* 1998;17:82-98.
4. **Dehejia RH, Wahba S.** Propensity Score Matching Methods for Non-experimental Causal Studies. Cambridge, MA: National Bureau of Economic Research, 1998.
5. **Heckman JJ, Ichimura H, Todd PE.** Matching as an econometric evaluation estimator: evidence from evaluating a job training program. *Review of Economic Studies* 1997;64:605-654.
6. **Heckman JJ, Ichimura H, Todd PE.** Matching as an econometric evaluation estimator. *Review Economic Studies* 1998;65:261-294.
7. **Rosenbaum PR, Rubin DB.** The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41-55.
8. **Wooldridge JM.** Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press, 2002.
9. **Imbens GW.** The role of the propensity score in estimating dose-response functions. *Biometrika* 2000;87:706-710.
10. **Mental Health: A Report of the Surgeon General.** Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health, 1999.
11. **Angrist JD, Hahn J.** When to control for covariates? Panel—asymptotic results for estimates of treatment effects (NBER Working Paper 241). Cambridge, MA: National Bureau of Economic Research, 1999.
12. **Hill JL, Brooks-Gunn J, Waldfogel J.** Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Dev Psychol* 2003;39:730-744.
13. **Foster EM.** Is more better than less? An analysis of children's mental health services. *Health Serv Res* 2000;35:1135-1158.
14. **Angold A.** Effectiveness of nonresidential specialty mental health services for children and adolescents in the 'real world.' *J Am Acad Child Adolesc Psychiatry* 2000;39:161-168.

15. **Bickman L, Andrade AR, Lambert EW.** Dose response in child and adolescent mental health services. *Ment Health Serv Res* 2002;4:57-70.
16. **Salzer MS, Bickman L, Lambert EW.** Dose-effect relationship in children's psychotherapy. *J Consult Clin Psychol* 1999;67:228-238.
17. **Andrade AR, Lambert EW, Bickman L.** Dose effect in child psychotherapy: outcomes associated with negligible treatment. *J Am Acad Child Adolesc Psychiatry* 2000;39:161-166.
18. **Foster EM, McLanahan S.** An illustration of the use of instrumental variables: do neighborhood conditions affect a young person's chance of finishing high school? *Psychol Methods* 1996;3:249-260.
19. **Bickman L, Guthrie PR, Foster EM, et al.** Evaluating Managed Mental Health Services: The Fort Bragg Experiment. New York: Plenum Press, 1995.
20. **Foster EM, Summerfelt TW, Saunders R.** The costs of a continuum of care: the lessons of the Fort Bragg demonstration. *J Ment Health Adm* 1996;23:92-106.
21. **Achenbach TM.** Manual for the Child Behavior Checklist and 1991 Profile. Burlington, VA: University of Vermont, 1991.
22. **Hodges K, Gust J.** Measures of impairment for children and adolescents. *J Ment Health Adm* 1995;22:403-413.
23. **Hodges K, Wong MM.** Use of the child and adolescent functional assessment scale to predict service utilization and cost. *J Ment Health Adm* 1997;24:278-290.
24. **Lutz W, Lowry JL, Kopta SM, et al.** Prediction of dose response relations. *J Clin Psychiatry* 2001;57:889-900.
25. **Kopta SM, Howard KI, Lowry JL, et al.** Patterns of symptomatic recovery in psychotherapy. *J Consult Clin Psychol* 1994;62:1009-1016.
26. **Kadera SW, Lambert MJ, Andrews AA.** How much therapy is really enough? A session-by-session analysis of the psychotherapy dose-effect relationship. *J Psychother Pract Res* 1996;5:132-151.
27. **Lueger RJ, Martinovich Z, Anderson EE, et al.** Assessing treatment progress of individual patients using expected treatment response models. *J Consult Clin Psychol* 2001;69:150-158.
28. **Howard KI, Kopta SM, Krause MS, et al.** The dose-effect relationship in psychotherapy. *Am Psychol* 1986;41:159-164.