Published in final edited form as:

Health Serv Outcomes Res Methodol. 2012 June 1; 12(2-3): 84–103. doi:10.1007/s10742-012-0089-7.

Assessing the Sensitivity of Treatment Effect Estimates to Differential Follow-Up Rates: Implications for Translational Research

Beth Ann Griffin,

RAND Corporation, 1200 South Hayes Street, Arlington, VA 22202, Phone: 1-703-413-1100 x 5188, Fax: 1-703-413-8111

Daniel McCaffrey,

RAND Corporation, 4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213

Rajeev Ramchand,

RAND Corporation, 1200 South Hayes Street, Arlington, VA 22202

Sarah B. Hunter, and

RAND Corporation, 1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407

Marika Suttorp

RAND Corporation, 1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407

Beth Ann Griffin: bethg@rand.org

Abstract

We develop a new tool for assessing the sensitivity of findings on treatment effectiveness to differential follow-up rates in the two treatment conditions being compared. The method censors the group with the higher response rate to create a synthetic respondent group that is then compared with the observed cases in the other condition to estimate a treatment effect. Censoring is done under various assumptions about the strength of the relationship between follow-up and outcomes to determine how informative differential dropout can alter inferences relative to estimates from models that assume the data are missing at random. The method provides an intuitive measure for understanding the strength of the association between outcomes and dropout that would be required to alter inferences about treatment effects. Our approach is motivated by translational research in which treatments found to be effective under experimental conditions are tested in standard treatment conditions. In such applications, follow-up rates in the experimental setting are likely to be substantially higher than in the standard setting, especially when observational data are used in the evaluation. We test the method on a case study evaluation of the effectiveness of an evidence-supported adolescent substance abuse treatment program (Motivational Enhancement Therapy/Cognitive Behavioral Therapy-5 [MET/CBT-5]) delivered by community-based treatment providers relative to its performance in a controlled research trial. In this case study, follow-up rates in the community based settings were extremely low (54%) compared to the experimental setting (95%) giving raise to concerns about non-ignorable dropout.

Keywords

Non-ignorable drop-out; informative censoring; differential follow-up; nonresponse bias; substance abuse; evidence supported treatments

1. Introduction

Well-designed studies with high follow-up rates are the gold standard for drawing inferences about the effectiveness of interventions to treat addiction. Attrition of study participants threatens a study's internal validity and potentially impairs the conclusions that researchers can make about a given treatment. Achieving high follow-up rates with drug-abusing populations is notably challenging. Researchers, organizations, and treatment facilities interested in evaluating a treatment need to invest significant time, planning, and resources to achieve high rates of follow-up for study participants (e.g., Scott 2004).

Regardless of the follow-up efforts taken, attrition of at least some of the participants in studies of substance abuse treatment programs is common and sometimes attrition rates can be very high. A variety of statistical techniques exist for making inferences from incomplete follow-up data due to attrition of study participants. Under the assumption that the incomplete data is missing completely at random (MCAR), standard regression models can be used to provide unbiased estimates of the relative treatment effect between two programs (Rubin 1976b). If the incomplete data is assumed to be missing at random (MAR), numerous parametric approaches are available (Hedeker and Gibbons 2006; Little and Rubin 2002) for obtaining unbiased estimates of the treatment effects of interest. Additionally, under the MAR assumption, nonparametric modeling involving non-response weights are often utilized to correct for the nonresponse bias in treatment effect estimates (Fuller et al. 1994; McGuigan et al. 1997; Robins et al. 1995). In cases where incomplete data due to attrition are not likely to satisfy either the MCAR or MAR assumptions (commonly referred to as attrition due to *nonignorable missingness*, *nonignorable drop-out*, or *informative censoring*), analysts still have a variety of options available, many of which directly model the nonignorable missingness via a selection model and then jointly estimate the selection model and treatment effects of interest (Diggle and Kenward 1994; Glynn et al. 1986; Greenlees et al. 1982; Hedeker et al. 2007; Schluchter 1992; Troxel et al. 1998; Wu and Carroll 1988).

Unfortunately, methods assuming MAR and MCAR and methods that directly model the ignorable missingness mechanism all involve untestable (and sometimes unjustifiable) assumptions, since we can never have data on the participants lost to follow-up. In fact, methods that directly try to model the nonignorable missingness mechanism can be difficult to identify and estimate in many applications since the data to support the underlying model are never observed (Troxel et al. 2004).

In light of the difficulties in these methods, techniques for exploring the sensitivity of treatment effect estimates to the assumptions required in MAR or MCAR analyses may also be valuable. Sensitivity analyses allow policy makers and other consumers of the statistical analyses to take into account the uncertainty that exists in the estimated treatment effects, including the uncertainty in the assumptions about the outcomes of participants who are lost to follow-up. There have been a variety of analyses proposed for testing the sensitivity of results to nonignorable dropout (Scharfstein et al. 1999; Troxel 1998; Little and Wang 1996; Daniels and Hogan 2000; Copas and Eguchi 2001; Copas and Li 1997; Verbeke et al. 2001). The general idea is to fit models under a range of plausible assumptions about the strength of the nonignorable relationship between dropout and unobserved outcomes and then assess

whether findings from the MAR analysis are robust to these changes. Troxel et al. (2004) proposed use of a particularly promising index of sensitivity to nonignorability (ISNI) for assessing the impact that nonignorable missingness might have on inferences from cross-sectional data analyses assuming MAR when fitting parametric models to the complete data model. The most recent work of Xie and Qian (2012) develop a simple sensitivity index method for longitudinal data with an arbitrary pattern of missingness in follow-up visits and with potentially unknown missingness reasons.

Nonetheless, available sensitivity analyses typically require interpretation of parameters measuring the difference between respondents and nonrespondents that can be hard to interpret for most stakeholders with less quantitative backgrounds. The goal of this paper is to present a sensitivity analysis that can be simply applied in treatment evaluations where the two treatments being studied have very differential follow-up rates and for which the parameter measuring the degree of sensitivity is easily accessible by the average stakeholder.

We consider the problem of differential attrition in an example of translation research which tests whether a treatment found to be efficacious in a clinical trial is also effective when used in standard care settings. Translation of efficacious therapies is of great interest for adolescent substance abuse treatment because historically there has been little overlap between the treatments used in community practice and those with scientific evidence of efficacy (Miller et al. 2006) and because the efficacy of usual care in community settings for adolescents is not well established (Morral et al. 2004). Treatments found to be efficacious in clinical trials may not always be effective in real world practice settings (e.g., Blanco et al. 2008a; Blanco et al. 2008b; Dodge et al. 2005; Miller et al. 2005; Okuda et al. 2010) since clinical trials are conducted under controlled conditions that may hold little resemblance to typical delivery settings due to important differences in therapist characteristics (e.g., training, caseload) and in treatment settings (e.g., intensive supervision of counselors by researchers) that could affect client outcomes (Weisz et al. 1995).

The present case study tests the hypothesis that Motivational Enhancement Therapy/Cognitive Behavioral Therapy-5 (MET/CBT-5; Sampl and Kadden 2001) was less effective when implemented in community settings than it was under the experimental conditions in which it was first tested. We test this particular hypothesis because previous studies that have compared outcomes from experimental studies and practice based treatment studies have tended to show an advantage in the research based settings (Barkham et al. 2008). We focus on MET/CBT-5 because it was initially tested in a highly controlled randomized trial, the Cannabis Youth Trial (CYT), where it was shown to be as efficacious as four other promising treatments yet, due to its brevity, one of the more cost-effective of the treatments considered in the trial (Dennis et al. 2004). Youth who represent our experimental arm come from the group of youth randomized to receive MET/CBT-5 in the CYT. Youth in our community sample come from the Effective Adolescent Treatment (EAT) initiative through which the Substance Abuse and Mental Health Services Administration's (SAMHSA) Center for Substance Abuse Treatment (CSAT) funded 38 community based treatment programs to administer MET/CBT-5 and to monitor the clients who received the treatment for one year post intake.

A challenge to the comparison between the community and the experimental implementation of MET/CBT-5 was the large differential in follow-up rates between the experimental and community settings (95% and 54%, respectively). The major concern with the low follow-up rate exhibited in dissemination initiatives like EAT is the possibility that the likelihood of being followed-up is associated with the treatment outcomes, making the findings in community settings different from experimental settings. However, the direction

of that difference is unclear. If those youth with poor outcomes are more likely to be lost to follow-up, the community-based treatment programs will appear to perform better relative to the experimental condition than is actually the case; conversely, if those youth with poor outcomes are more likely to be followed (for example, due to re-admittance to drug treatment facility or incarceration) it will appear that the community-based setting performed worse relative to the experimental condition than is actually the case. The uncertainty of the outcomes of participants lost to follow-up and their potential impact on estimates of the relative efficacy of care must be taken into account when we interpret the treatment effect estimates comparing treatment delivered in two different settings. This paper highlights one method for evaluating the sensitivity of our results to the unobserved outcomes of participants lost-to-follow-up.

Low and differential response rates are not unique to our study or translation research. On the contrary they are common in studies of substance abuse and mental health treatment among both adolescent and adult populations. For example, Witbrodt et al. (2007) report 12-month follow up rates of 59% and about 80% for groups in community and hospital care for substance abuse, respectively. Studies by Bertelsen et al. (2008), Gruber et al. (2008) and MacKinnon et al. (2008) have follow-up rates that differ by 15 or more percentage points among groups and a review by Christensen (2009) reports on several studies for anxiety and depression with follow-up rates that differed most often by 10 to 20 percentage points with a maximum of 24 percentage points.

In the next section (2), we discuss a model for incomplete follow-up data and key assumptions that will be utilized in our approach. We also present our method for assessing the sensitivity of results to differential follow-up under a wide range of potential values for the strength of the relationship between outcomes and loss to follow-up. In Section 3, we describe the data used in our case study and our outcome regression models that are fit in both the analyses assuming MAR and the sensitivity analyses. In Section 4, we describe results from the analyses which assume MAR and the proposed sensitivity analyses using the MET/CBT-5 case study data, and finally, the paper concludes (Section 5) with a discussion of our findings and its implication for future translational research.

2. Model, assumptions, and proposed sensitivity analysis

2.1. A model for incomplete data in a treatment effects analysis

Following the potential outcomes approach to causal modeling (Holland 1986; Rubin 1976a), we assume there is a random variable, T , for the treatment assignment of every youth in the population such that $T = 1$ if the youth was assigned to receive MET/CTB5 in the experimental setting and $T = 0$ if treatment occurred in the community setting. For any outcome of interest, each youth has a potential value Y_1 that is observed when the youth is assigned to receive treatment in the experimental setting and a separate potential value Y_0 that is observed when the youth is assigned to treatment in the community setting. Treatment effects of interest equal the differences in the means of the two potential outcomes.

To account for attrition, we also define two potential response indicators for every youth: $R_1 = 1$ if the youth will respond when assigned to the experimental setting and equal to 0 if the youth will drop out of the study when assigned to that setting and $R_0 = 1$ or 0 depending on whether or not the youth will respond when assigned to the community setting. Like the potential outcomes, we observe only the response indicator corresponding to the treatment condition the youth received. For every youth we also observe a vector of baseline characteristics, $\mathbf{x}$.

Because we cannot observe the potential outcomes and response indicators for both treatments on the same youth, we must assume that the youth receiving one treatment can serve as the counterfactual for the youth receiving the other treatment. In translation research, the assignment of youth to experimental and community groups may not always be controlled by researchers (say, via randomization) but will instead be determined by the standard assignments of youth to treatment providers, as is true in our case study of MET/CTB5. Consequently, the youth in the groups may differ on both observed and unobserved baseline characteristics. We assume that those differences which are consequential for the outcomes of interest are captured through the observed baseline characteristics and thus that there is no unobserved confounding in our data (i.e., that treatment assignment is strongly ignorable given $\mathbf{x}$; Rosenbaum and Rubin 1983). This can be written as:

$$0 < \text{pr}(T=1|\mathbf{x}) < 1, \text{ and} \\ (Y_0, Y_1) \text{ are independent of } T \text{ conditional on } \mathbf{x}.$$

We will also assume strong ignorability for the response indicators so that (R_0, R_1) are independent of T conditional on $\mathbf{x}$. While these assumptions are quite strong and should be explored, for the current study we will accept both as reasonable assumptions in order to focus on assumptions concerning attrition in more detail. These strong ignorability assumptions allow us to make the treatment and control groups similar in terms of potential outcomes and response indicators by weighting or conditioning on the baseline variables. However, they do not imply that potential outcomes will be similar across the respondents in the two treatment groups as R_1 and R_0 are distinct potential variables and youth with $R_1 = 1$ may not have $R_0 = 1$.

To motivate our method for assessing sensitivity to differential follow-up rates between two treatment conditions in more detail, we will utilize the four cells in Table 1 to create a two-by-two classification of possible responses for all youth in our sample. The cells in Table 1 are defined by the two response indicators, R_0 and R_1 . Each youth in our sample belongs to exactly one of the four cells. Specifically, youth with $R_1 = R_0 = 0$ fall into the DD (dropout, dropout) cell and never respond to the follow-up regardless of the treatment they receive. Youth with $R_1 = R_0 = 1$ fall into the FF (follow-up, follow-up) cell and always respond to the follow-up regardless of treatment. On the other hand, youth with $R_1 = 0$ and $R_0 = 1$ fall into the DF (dropout, follow-up) cell only respond when $T = 0$ (i.e., when they are assigned to MET/CTB5 in the community setting) while youth with $R_1 = 1$ and $R_0 = 0$ fall into the FD (follow-up, dropout) cell only respond when $T = 1$ (i.e., when they are assigned to MET/CTB5 in the experimental setting). The four groups defined by Table 1 are often referred to as *principal strata* (Frangakis and Rubin 2002) because if we could observe which cell each youth would fall into we could estimate causal effects by comparing similar groups of youth within a given cell.

Unfortunately, we cannot fully classify youth into the four principal strata listed in Table 1 because only one response indicator is observed for each youth and the other must be inferred. Even if the two treatment groups are similar (i.e., Y_0, Y_1 are independent of T), the outcomes of the responding youth might not be similar because the followed-up youth in each treatment may end up being distributed across two cells (DF and FF for those with $T = 0$ and FD and FF for those with $T = 1$) in a way that might not guarantee balance. The only group that is common to both sets of respondents is the FF group. Consequently, a causal effect for treatment is only defined for youth in that cell. However we cannot identify the youth in the FF cell from either set of respondents without additional assumptions.

2.2 Assumptions concerning missing data mechanism

As noted above, it is common to assume the incomplete data are MAR so that the outcomes (Y_i) and response indicators (R_i) are independent conditional on the observed baseline covariates which implies that the distributions of Y given x are the same for youth in the DF and FD cells. Under the MAR assumption, analysts can create nonresponse weights to weight the respondents in each group back to their respective baseline samples or to directly balance baseline variables between the two groups of respondents and then use these when estimating treatment effects. Alternatively, various parametric methods could be used instead of weighting or in conjunction with weighting.

Relaxing the MAR assumptions generally requires alternative assumptions about how outcomes are distributed across the DF and FF cells for those with $T=0$ and across FD and FF cells for those with $T=1$. One common and plausible assumption in many studies like our current case study is *monotonicity* (Lee 2005). Monotonicity requires that every youth who would respond when assigned to the community setting would also respond if assigned to the experimental setting. In other words for our case study, the easy to track cases followed-up by the staff in the community setting would clearly be tracked by the aggressive follow-up of the experimental study which is able to track nearly all youth. Under monotonicity there are no youth with $R_1=0$ and $R_0=1$ and the DF cell is empty. This means all youth who dropout in the experimental condition ($R_1=0$) are in the DD cell or would dropout under either condition. Similarly, all youth who are followed-up in the community setting are in the FF cell and also would be followed-up in the experimental setting.

We utilized the assumption of monotonicity to design our sensitivity analysis so that we can estimate the population mean of the outcome following assignment to MET/CTB5 in the community setting for the FF group of youth and attempt to compute the treatment effect among youth like those in the FF cell using the responders from the experimental condition (who are still distributed across the FF and FD cells). Unfortunately, determining the mean outcome for youth in the FF cell following assignment to MET/CTB5 in the experimental condition can be tricky when we are uncertain about whether or not MAR holds because we cannot be certain who among these responders should fall into the FF cell or who should fall into the FD cell. Our simulation study aims to quantify this uncertainty in a meaningful way by examining the sensitivity of the estimated treatment effects to assumptions about how the outcomes in the experimental condition are distributed across these two cells. The simulation study allows the uncertainty of our estimated treatment effect resulting from missing data to be expressed by uncertainty about the distribution of outcomes in (and across) the FD or FF group.

2.3 The proposed approach: sensitivity analysis under varying assumptions about the relationship between outcomes and attrition

We developed a series of sensitivity analyses to determine how strong the relationship between the outcomes (Y_i) and response indicators (R_i) needs to be to alter inferences in standard statistical models which assume MAR. In developing the sensitivity analyses, we assumed that monotonicity holds so that the responders from the experimental group include youth in both the FF and FD groups in Table 1 while the responders in the community group include only youth in the FF group. Thus, under this assumption, the *proportion* of responders in the community group determines the *fraction* of the responders in the experimental group who are in the FF group. What remains unclear, however, is which responders in the experimental group would have fallen into the FF group and which would have fallen into the FD groups. The experimental-group responders in the FF group are those with the unobserved R_0 indicator equal one.

For our sensitivity analyses we create synthetic FF and FD groups for cases in the experimental group by positing a fixed magnitude for the relationship between outcomes and R_0 (the response indicator when assigned to the control group, here the community setting). This relationship is operationalized through the log-odds of R_0 and makes the youth in the synthetic FF group have different outcomes than the synthetic FD group, thereby allowing for informative nonresponse/nonignorable drop-out. Specifically, to create the synthetic FF group, we simulate a value of R_{0i} for $i = 1$ to n_1 (the number of responders in the experimental group) and then include those youth with $R_{0i} = 1$ in the synthetic FF group and repeat our outcomes analysis. We randomly sample values of R_{0i} based on the following assumed relationship between R_{0i} and a given 12-month outcome, denoted by Y_{1i} :

$$\begin{aligned} \text{logit}(\Pr(R_{0i}=1)) &= \alpha + \beta Y_{1i} \text{ for binary outcomes and} \\ \text{logit}(\Pr(R_{0i}=1)) &= \alpha + \beta Y_{1i} / \hat{\sigma}_y \text{ for continuous outcomes,} \end{aligned}$$

where $\hat{\sigma}_y$ denotes the estimated standard deviation of the given outcome in question. Our models for the response indicator do not include baseline characteristics measured in x because our goal is to create synthetic FF groups among treated youth that have a specified level of selection on the outcome. We use a simple model of attrition here so that it is straightforward to tune the selection in the FF group to test the possible bias in inferences from models that assume the data are MAR.

Values of β were set equal to $\log(OR)$ where OR was taken to range anywhere from 1 to 100 and values of α were solved for a given value of β in order to ensure that the overall censoring rate in the experimental condition would equal the attrition rate in the EAT-funded community sites. In total, 100 simulations were performed for each set of (α, β) with the following basic steps.

- i. Randomly sample values of R_{0i} for each youth in the experimental condition
- ii. Keep those youth with $R_{0i} = 1$
- iii. Re-estimate the treatment effect
- iv. Compute estimate of $\frac{E[Y_1|C=1] - E[Y_1|C=0]}{\sqrt{\text{Var}(Y_1)}}$, which we refer to as the Standardized Selection Index (SSI).

For each iteration, we recorded the (i) mean of the outcome among the synthetic FF and FD samples from the experimental condition group, (ii) the estimated treatment effect comparing the mean difference in outcomes of the propensity-score weighted regression models fit to the synthetic FF group and youth in the community setting group, and (iii) the standardized mean difference in outcomes among the responders in the experimental setting group with $R_{0i} = 0$ (the synthetic FD group) and those with $R_{0i} = 1$ (the synthetic FF group). Then across the iterations, we took the means for each of these measures, denoting the mean of the third measure by SSI, and noted the values for the 2.5th and 97.5th percentiles of the estimated treatment effects. SSI is our preferred measure for reporting the strength of the association between an outcome and attrition β for a given value of β when describing our results since it is similar in spirit to Cohen's d (Cohen 1992). Using this metric, values of 0.20, 0.50, and 0.80 are commonly considered to represent small, moderate, and large magnitudes, respectively.

Our results (see Figure 2) report two key values of SSI: (i) the value of SSI for which the mean treatment effect estimate for the FF group found by comparing MET/CBT-5 among the community responders and synthetic experimental FF group across the 100 simulations

was zero (i.e., gave treatment effect estimates indicating experimental and community settings were the same) and (ii) the value of SSI for which 95% of the simulations favored youth in the experimental setting (i.e., gave treatment effect estimates in favor of the experimental setting for the 2.5th and 97.5th percentiles).

It is important to note that we do not make any further adjustments to the synthetic FF group for the experimental setting and the observed FF group for the community setting other than controlling for the initial baseline differences using propensity score weights. This is because we assume that through our censoring we have identified the FF group in treatment so that the synthetic treatment and observed control groups are made up of youth from the FF sample with common backgrounds and common potential outcomes no matter the strength of the relationship between outcomes and follow-up. Thus, under the assumptions we are using to generate the synthetic group, treatment is strongly ignorable between the synthetic experimental and the observed community FF groups, provided we remove the initial baseline differences between the complete experimental and community groups successfully with the propensity score.

We also note that in the proposed analyses, it is not required that the outcome be the only missing variable. Indeed, in our case study we have missingness (generally less than 5% per variable) in our baseline covariates. We control for the impact of this missingness on inferences by balancing the groups on missingness indicators for each baseline characteristics in the propensity score model in addition to the variables themselves (see Section 3.2 and Appendix for details).

3. MET/CBT-5 case study

3.1 MET/CBT-5 experimental and dissemination studies

As described above, data for the current study are from two efforts supported by SAMHSA's CSAT. Youth who represent our experimental arm ($T=1$) come from 183 youth randomized to receive MET/CBT-5 at baseline in the CYT trial (Dennis et al. 2002), 95% of whom completed the follow-up assessment at 12-months post intake. Youth in our community sample ($T=0$) come from the EAT initiative in which all community based treatment sites administer MET/CBT-5. The 4,225 EAT youth who were eligible to have a 12-month follow-up interview during the initiative's funding period are included in this analysis as the community sample; 54% had 12-month follow-up data. As noted before, the follow-up rates across the 38 EAT sites ranged from 2% to 90%. In the current analysis, we retain all sites regardless of their follow-up rate as opposed to throwing away the particularly "bad" sites (e.g., those with follow-up rates less than 70%) in order to maintain the ability to generalize the results to all community-based treatment settings included in the EAT, not just those sites with good follow-up.

3.2. Adjustments to account for pre-treatment differences

The first step for comparing outcomes from youth in the experimental setting group to youth in the community setting group was to remove observed pre-treatment differences among the two samples. We decided that inferences about the effectiveness of disseminating MET/CBT-5 would have to be restricted to youth like those in the experimental group since the population in the experimental condition was more restrictive (based on requirements of the CYT study) than the population from the community-based treatment programs in our study. To do this, we first applied the CYT inclusion and exclusion criteria to the community setting sample (details on these criteria can be found in Dennis et al. 2004) and found that 69% ($n=2,908$) of youth in the EAT sample met the CYT study criteria at baseline. Next, we employed propensity score weights to make distributions of pretreatment variables ($\mathbf{x}$) for

the remaining youth in the community setting group match those of the youth in the experimental setting group. This means that youth in the community-setting group with pretreatment variables similar to those of the youth in the experimental setting group were given larger weights, and those who differ from experimental setting youth received lower weights. As discussed in Section 2.1, we assume that we have strong ignorability of the treatment (T) on both the potential outcomes (Y_i) and response indicators (R_i) with respect to the observed baseline characteristics included in the propensity score model. Details on the propensity score analysis can be found in the appendix (McCaffrey et al. 2004).

3.3 Outcomes and regression models for estimating treatment effects under MAR and in sensitivity analyses

We examined six treatment outcomes at 12-months post-intake from the widely used assessment tool, the Global Appraisal of Individual Needs (GAIN; Dennis 1999). The outcomes included two dichotomous variables (recovery and institutionalization) and four continuous change scores (measuring change in the substance problem scale [SPS], substance frequency scale [SFS], illegal activities scale [IAS], and emotional problems scale [EPS]). More details on these outcomes can be found in two related studies on MET/CBT-5 in the CYT and EAT programs (Hunter et al. In press; Ramchand et al. 2011) and were the outcomes used in the original CYT experiment (Dennis et al. 2004).

Our analytic approach when modeling outcomes in both the analyses that assume MAR and the sensitivity analyses was designed to answer the following question: *What is the relative efficacy of MET/CBT5 when implemented in the experimental setting compared to when implemented in community based treatment settings for youth like those who enrolled in the experimental study?* To answer this question, we used the propensity score weights described above to compare the outcomes of youth receiving MET/CBT5 in the experimental study (CYT) with those of similar clients receiving MET/CBT5 in the community-based settings (EAT). Specifically, we computed the average treatment effect on the treated population, or ATT (Rosenbaum and Rubin 1983; Hirano et al. 2003; Rosenbaum and Rubin 1984, 1985; Wooldridge 2002) for each of the six outcomes. In our application, the estimated ATT is the expected treatment effect of receiving MET/CBT5 in the experimental setting versus community-based treatment settings for the adolescent clients like those who actually received MET/CBT5 in the experimental setting.

An unbiased estimate of ATT can be obtained by computing the difference between the observed (unweighted) mean of the experimental group (which provides a consistent estimate for $E[Y_i / T=1]$) and the ATT propensity score weighted mean of the community based group (which provides a consistent estimate for $E[Y_0 / T=1]$, the expected value of the counterfactual outcomes for youth in the experimental group).

Since we had some lingering imbalances between the two treatment settings after using propensity score weights (see Table 2), we estimated the ATT's for each outcome in the analysis assuming MAR and the sensitivity analyses using propensity score weighted regression models that controlled for the treatment indicator of whether or not a youth was in the experimental study (CYT) and the three variables which were imbalanced after weighting (race and the illegal activities and substance frequency scales; see Table 2; Kang and Schafer 2007). We note that the multivariate models for changes in the illegal activities and substance frequency scales did not control for the baseline value of the scale in question because the baseline measures are already controlled for in the change score. Thus, for each continuous outcome, we let $Y_{i,obs} = Y_{1i}T_i + Y_{0i}(1-T_i) - Y_{B}$, where Y_{B} is the baseline value of the outcome, and estimated the coefficients of the following regression model

$$Y_{i,obs} = \mu + \delta T_i + \phi Z_i + \varepsilon_i$$

using weighted least squares where the weights are 1 for youth in the experimental group and $p(T=1 | \mathbf{x}) / \{1 - p(T=1 | \mathbf{x})\}$ for youth in the community group and $\mathbf{Z}$ is the vector of covariates controlled for in the model (race and the illegal activities and substance frequency scales or the appropriate subset).

For our dichotomous outcomes, we let $Y_{i,obs} = Y_{1i}T_i + Y_{0i}(1-T_i)$ and estimated the coefficients of the model

$$\Pr(Y_{i,obs}=1) = \frac{1}{1 + e^{-(\mu + \delta T_i + \phi Z_i)}}$$

using weighted logistic regression with the weights describe above to obtain $\hat{\mu}$, $\hat{\delta}$, and $\hat{\phi}$. The treatment effect was set equal to the difference in the predicted margins (Graubard and Korn 1999):

$$\sum_i^{n_T} \frac{1}{1 + e^{-(\hat{\mu} + \hat{\delta} + \hat{\phi}' z_i)}} - \sum_i^{n_T} \frac{1}{1 + e^{-(\hat{\mu} + \hat{\phi}' z_i)}}$$

where summation is over all n_T youth in the experimental group. Each estimated ATT was then divided by the standard deviation of the outcome among youth in the experimental setting in order to put all six treatment effects on an effect size scale (see Figure 1; Cohen 1992).

In our analysis, we first fit these models assuming MAR and statistical significance was assessed by checking if the 95% confidence interval for the standardized mean difference contained 0. We then conducted sensitivity analysis to our MAR assumptions using the methods of Section 2.3. As specified in Section 2.3, we censored the youth in the experimental (CYT) sample following steps (i) and (ii). Then, in step (iii), we used the same outcome model we used for the primary analysis assuming MAR but fit it to the reduced (simulated) dataset of all youth in the community (EAT) sample and simulated responders in the experimental sample. The same set of propensity score weights for balancing groups at baseline were used in all the regression models, because those weights allow us to ignore selection in the initial treatment assignment. We act as if the study is in essence a randomized trial design before youth in the sample were lost to follow-up. Conditional on the propensity scores correcting the imbalances between the two groups on baseline characteristics, we then use the simulation study to artificially censor youth in the experimental condition to have a loss to follow-up mechanism that results in the same (simulated) attrition rate that the community-based treatment settings had and only depends on the outcomes of those youth.

We note that we make no additional adjustments for attrition such as nonresponse weighting because there was no evidence of differential response on observed variables in the case study data after accounting for the baseline differences between the experimental and community samples. Typically analysts can adjust for differential nonresponse with respect to the observed covariates, if it exists, using nonresponse weights.

4. Results

4.1 Adjustment for pretreatment heterogeneity

Table 2 shows a comparison of our analytic samples before and after propensity score weighting across the 22 variables used in our weighting strategy. As shown, the two groups were imbalanced on seven pretreatment characteristics before weighting. After weighting, the majority of these differences were attenuated; however, the community sample (EAT) still had a greater percentage of Hispanic youth and had lower mean values for the Illegal Activities Scale and Substance Frequency Scale as compared to the experimental sample (CYT). Because of these sustained differences between the two samples, these three variables were controlled for in the regression models used in the outcomes analyses.

4.2 Results assuming MAR

Figure 1 shows our estimates of the relative effectiveness of MET/CBT-5 in the experimental setting versus the community setting in analyses which assume that after weighting on pretreatment covariates, those who are followed-up and those who are not do not differ on the outcomes (MAR). Effect size differences and 95% confidence intervals are shown for each outcome. At 12 months, youth in the experimental setting had significantly lower rates of recovery (effect size difference = -0.34 ; 95% CI = $-0.59, -0.10$) and higher rates of institutionalization (0.21 ; 95% CI = $0.01, 0.40$). There were no significant differences between the two groups when examining the four continuous change score outcomes.

4.3 Application of proposed sensitivity analysis

The dot plot in Figure 2 shows (i) the value of SSI for which the mean treatment effect estimate for the FF group found by comparing MET/CBT-5 among the community responders and synthetic experimental FF group across the 100 simulations was zero (i.e., gave treatment effect estimates indicating experimental and community settings were the same) and (ii) the value of SSI for which 95% of the simulations favored youth in the experimental setting (i.e., gave treatment effect estimates in favor of the experimental setting in 95% of the simulations). Only SSI's greater than zero (i.e., worse youth are more likely to be lost to follow-up) are considered since our analysis generally favors clients in the community setting over the experimental setting and we want to understand how strong the association would need to be *to change the direction of our findings* as opposed to strengthen our findings.

For the recovery outcome, youth who were not followed-up would have to have outcomes that were 0.77 standard deviation units (i.e., a large effect size) lower than those who were followed-up to make the significant mean difference observed in recovery rates between the two settings equal zero across the 100 simulations and 1.11 standard deviation units (i.e., a large effect size) lower to have 95% of our simulated data sets show differences in recovery rates that favor the experimental setting. These effect size differences (0.77 and 1.11) correspond to very large differences in recovery rates between the synthetic responders (the FF group) and the synthetic non-responders (the FD group); specifically, they correspond to a 37 percentage point difference (52% recovered in the FF group versus 15% in the FD group) and a 54 percentage point difference (60% recovered in the FF group versus 6% in the FD group), respectively (results not shown). For institutionalization, youth who were not followed-up would have to have outcomes that were 0.55 (moderately large) and 0.84 (large) standard deviation units different, respectively, to make the significant mean difference observed in institutionalization between the two settings equal zero across the 100 simulations and to have 95% of our simulated data sets show differences in institutionalization rates that favor the experimental setting. These effect size differences

also correspond to large differences in institutionalization rates between the synthetic responders (the FF group) and the synthetic non-responders (the FD group); specifically, they correspond to a 26 percentage point difference (14% institutionalized in the FF group versus 40% in the FD group) and a 37 percentage point difference (10% institutionalized in the FF group versus 47% in the FD group), respectively (results not shown).

The required values of SSI to yield a zero mean difference between the two settings for the Substance Frequency Scale, the Substance Problem Scale, and the Illegal Activities Scale are all very small (i.e., 0.02, 0.18, and 0.02, respectively) since these outcomes yielded observed mean differences in the analysis assuming MAR which were all close to zero (see Figure 1). The values of SSI that would be required to favor the experimental setting in 95% of our simulations for these outcomes, however, are much higher and would be categorized as moderate to moderately large effect sizes (0.43, 0.53, and 0.33, respectively) for these outcomes. Results for the Emotional Problems Scale are the most robust in terms of the continuous outcomes with values of SSI equal to 0.54 (moderately large) and 0.93 (large) being required to yield a zero mean difference between the two settings across the 100 simulations and to reverse our findings in favor of experimental setting 95% of the time, respectively.

5. Discussion

The proposed sensitivity analysis provides insight into the uncertainty about the relationship between outcomes and follow-up in analyses that compare two treatment groups with very disparate follow-up rates by determining how strong the association between outcomes and follow-up would need to be to change findings regarding the relative effectiveness of two treatment programs (or settings) when implementing analyses that assume MAR. The proposed approach offers a promising and intuitive metric for assessing how robust findings in a given comparison are to lost to follow-up that can be easily digested by stakeholders without extensive backgrounds in statistics.

In the case study of MET/CBT-5, analyses which ignored the follow-up issue and assumed MAR found that the outcomes achieved in the practice settings were generally similar if not better to those achieved in the experimental setting. Indeed, the clients who received the treatment in the community settings showed better recovery status and lower institutionalization rates than those receiving treatment in the more controlled research study. The sensitivity analyses demonstrated how susceptible these results are to assumptions about the relationship between follow-up and outcomes (i.e., about the strength of nonignorable drop-out in a given study). As shown, it would take small to moderate effect size differences between the outcomes of responders and non-responders to yield a 0 mean difference between the two settings in our case study. However, it would take a moderate to strong associations (i.e., effect size differences in outcomes greater than 0.33 to 1.11 between responders and nonresponders) to completely reverse our findings and favor outcomes in the experimental setting over the community setting. Overall, these findings taken together with a recent study showing that MET/CBT-5 is more effective in the community settings than other exemplary outpatient treatment programs (Hunter et al. In press) suggest that dissemination of MET/CBT-5 has led to improved outcomes in community adolescent treatment settings.

Clinical treatment translation studies will likely be faced with the same type of differential follow-up that is seen in our MET/CBT-5 case study, making the proposed sensitivity analyses very valuable. In addition, many other types of studies that compare different treatment approaches or treatments delivered in different settings will also likely see differential follow-up rates. The high follow-up rates usually achieved by experimental

studies such as CYT result from rigorous field work in which clients are tracked extensively by professionals who are experts in tracking substance using populations. However, following-up with substance abusers requires significant resources (e.g., see Scott 2004) and in research studies, is oftentimes conducted by staff independent from the treatment setting. Though the EAT initiative provided support to administer the GAIN and conduct follow-ups, the best response rate in the community at 12 months was 90% and the median was much lower (58%). These findings suggest that the resources in these settings were still insufficient.

The relatively low response rates in the community condition require that strong assumptions need to be made about the outcomes of nonresponders (e.g., that response is not related to outcomes controlling for baseline characteristics; MAR) in order for treatment estimates from standard statistical analyses to be unbiased, which creates a great deal of uncertainty about the size of the causal effects in the case study. Being able to quantify the uncertainty in treatment effect estimates, as done in the proposed sensitivity analysis, was especially useful here because attrition rates were so high in the community sample. When attrition rates are high, failure in the MAR assumption can have a greater impact on the treatment effect estimate since more data are missing and because it is more likely that few of the natural selection processes that make some youth harder to track have been overcome through aggressive field work.

We note that our case study only uses data from the baseline and 12-month follow-up visits of a prospective longitudinal study of youth enrolled in MET/CBT-5 via the EAT and CYT programs. We only utilize these two time points in our case study, because use of outcomes at a single follow-up for treatment evaluations is common in substance abuse treatment research. Thus, we have reduced the data to a simple pre-post design. However, the CYT and EAT programs also conducted 3 and 6 month follow-ups and we could use baseline and follow-up data together to develop more complex nonresponse weights and potentially reduce the risk of nonignorable nonresponse at 12-month follow-up where the outcomes of primary interest lie. Our censoring method could still be applied after developing such weights. If interest lies in estimating the effects of treatment on outcome trajectories, then alternative sensitivity analyses might be necessary, possibly building on the work of Xie and Qian (2012) who extend the ISNI of Troxel et al. (2004) to the longitudinal setting.

The proposed sensitivity analyses while promising are still limited because we do not know which youth are not being followed-up in the community-based setting. Without observing the outcomes for those lost to follow-up, it is infeasible to know what types of extrapolations are reasonable and thus it is easy for extrapolation assumptions to be wrong. For example, it is often presumed that participants who continue to use are the most difficult to follow-up (Walton et al. 1998) and under this assumption, one might readily impute the missing outcome data for participants not followed-up in such a way that their outcomes are on average worse than those who are followed-up. Unfortunately, the literature has also shown that it can be just as reasonable to assume that all youth who were not followed-up were the youth who were doing well in treatment since participants who are doing worse may end up in institutionalized settings (e.g., treatment or incarceration), making it easier to follow-up than non-using participants who develop new (lower risk) social networks that are unknown to researchers (Desmond et al. 1995).

SAMSHA's CSAT made a great investment in the EAT grantees who were included in this case study of MET/CBT-5 as the community-based sample by requiring client outcomes to be monitored using the same design and survey instrument across all settings. However, with the relatively low follow-up rates that were observed, the potential return on that investment was greatly limited since analyses assuming MAR are left with an uncomfortable

level of uncertainty concerning which group or treatment setting is really the most effective. The proposed sensitivity analyses provide researchers and providers who are faced with low follow-up rates an approach that can be utilized to draw inferences in a more meaningful and cautious way.

Acknowledgments

This research was supported by the National Institute on Drug Abuse (5R01DA017507-05; PI: Morral and 1R01DA015697; PI: McCaffrey). The authors would like to thank Andrew Morral for comments provided on earlier drafts of this research.

Abbreviations

ASMD	Absolute Standardized Mean Difference
CSAT	Center for Substance Abuse Treatment
CYT	Cannabis Youth Trial
EAT	Effective Adolescent Treatment
EPS	Emotional Problems Scale
GAIN	Global Appraisal of Individual Needs
GBM	Generalized Boosted Model
IAS	Illegal Activities Scale
ISNI	Index of Sensitivity to NonIgnorability
MAR	Missing at Random
MCAR	Missing Completely at Random
MET/CBT-5	Motivational Enhancement Therapy/Cognitive Behavioral Therapy-5
SAMHSA	Substance Abuse and Mental Health Services Administration
SFS	Substance Frequency Scale
SPS	Substance Problem Scale
SSI	Standardized Selection Index

References

- Alford GS, Koehler RA, Leonard J. Alcoholics Anonymous-Narcotics Anonymous model inpatient treatment of chemically dependent adolescents: A 2-year outcome study. *J Stud Alcohol*. 1991; 52:118–126. [PubMed: 2016871]
- American Society of Addiction Medicine. ASAM Patient Placement Criteria for the Treatment of Substance-related Disorders. 2. Lippincott Williams & Wilkins; Chevy Chase: 2001. Revised ed
- Barkham M, Stiles WB, Connell J, Twigg E, Leach C, Lucock M, Mellor-Clark J, Bower P, King M, Shapiro DA, Hardy GE, Greenberg L, Angus L. Effects of psychological therapies in randomized trials and practice-based studies. *Br J Clin Psychol*. 2008; 47:397–415. [PubMed: 18625084]
- Bertelsen M, Jeppesen P, Petersen L, Thorup A, Øhlenschlaeger J, le Quach P, Christensen TØ, Krarup G, Jørgensen P, Nordentoft M. Five-year follow-up of a randomized multicenter trial of intensive early intervention vs standard treatment for patients with a first episode of psychotic illness: the OPUS trial. *Arch Gen Psychiatry*. 2008; 65(7):762–771. [PubMed: 18606949]
- Blanco C, Olfson M, Goodwin RD, Ogburn E, Liebowitz MR, Nunes EV, Hasin DS. Generalizability of clinical trial results for major depression to community samples: results from the National epidemiologic survey on alcohol and related conditions. *J Clin Psychiat*. 2008a; 69:1276–1280.

- Blanco C, Olfson M, Okuda M, Nunes EV, Liu SM, Hasin DS. Generalizability of clinical trials for alcohol dependence to community samples. *Drug Alcohol Depend.* 2008b; 98:123–128. [PubMed: 18579319]
- Brown SA, Gleghorn A, Schuckit MA, Myers MG, Mott MA. Conduct disorder among adolescent alcohol and drug abusers. *J Stud Alcohol.* 1996; 57:314–324. [PubMed: 8709590]
- Brown SA, Tapert SF, Tate SR, Abrantes AM. The role of alcohol in adolescent relapse and outcome. *J Psychoactive Drugs.* 2000; 32:107–115. [PubMed: 10801072]
- Cady ME, Winters KC, Jordan DA, Solberg KB, Stinchfield RD. Motivation to change as a predictor of treatment outcome for adolescent substance abusers. *J Child Adolesc Subst Abuse.* 1996; 5:73–91.10.1300/J029v05n01_04
- Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. *J Med Internet Res.* 2009; 11(2):e13. [PubMed: 19403466]
- Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics.* 1968; 24(2):295–313. [PubMed: 5683871]
- Cohen J. A power primer. *Psychol Bull.* 1992; 112:155–159. [PubMed: 19565683]
- Copas JB, Eguchi S. Local sensitivity approximations for selectivity bias. *J Roy Statist Soc Ser B.* 2001; 63:871–895.
- Copas JB, Li HG. Inference for non-random samples (with discussion). *J Roy Statist Soc Ser B.* 1997; 59:55–95.
- Daniels MJ, Hogan JW. Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics.* 2000; 65:1241–1248. [PubMed: 11129486]
- Dennis ML. Global Appraisal of Individual Needs (GAIN): Administration guide for the GAIN and related measures (Version: 1299). Chestnut Health Systems. 1999
- Dennis ML, Godley SH, Diamond G, Tims FM, Babor T, Donaldson J, Liddle H, Titus JC, Kaminer Y, Webb C, Hamilton N, Funk R. The Cannabis Youth Treatment, CYT Study: Main findings from two randomized trials. *J Subst Abuse Treat.* 2004; 27(3):197–213. [PubMed: 15501373]
- Dennis ML, Titus JC, Diamond G, Donaldson J, Godley SH, Tims FM, Webb C, Kaminer Y, Babor T, Roebuck MC, Godley MD, Hamilton N, Liddle H, Scott CK, CYT. Steering Committee: The Cannabis Youth Treatment (CYT) experiment: rationale, study design and analysis plans. *Addiction.* 2002; 97(Suppl 1):16–34. [PubMed: 12460126]
- Desmond DP, Maddux JF, Johnson TH, Confer BA. Obtaining follow-up interviews for treatment evaluation. *J Subst Abuse Treat.* 1995; 12(2):95–102. [PubMed: 7623396]
- Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis (with discussion). *Appl Statist.* 1994; 43:49–73.
- Dodge R, Sindelar J, Sinha R. The role of depression symptoms in predicting drug abstinence in outpatient substance abuse treatment. *J Subst Abuse Treat.* 2005; 28:189–196. [PubMed: 15780549]
- Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics.* 2002; 58:21–29. [PubMed: 11890317]
- Friedman AS, Glickman NW. Program characteristics for successful treatment of adolescent drug abuse. *J Nerv Ment Dis.* 1986; 174(11):669–679. [PubMed: 3772356]
- Friedman AS, Glickman NW, Morrissey MR. Prediction to successful treatment outcome by client characteristics and retention in treatment in adolescent drug treatment programs: a large-scale cross validation study. *J Drug Educ.* 1986; 16:149–165. [PubMed: 3735024]
- Fuller WA, McLoughlin MM, Baker HD. Regression weighting in the presence of nonresponse with application to the 1987–1988 Nationwide Food Consumption Survey. *Surv Methodol.* 1994; 20:75–85.
- Glynn, RJ.; Laird, N.; Rubin, D. Selection Modeling vs. Mixture Modeling with Nonignorable Response. In: Wainer, H., editor. *Drawing Inferences From Self-Selected Samples.* Springer-Verlag; New York: 1986.
- Graubard BI, Korn EL. Predictive margins with survey data. *Biometrics.* 1999; 55(2):652–659. [PubMed: 11318229]

- Greenlees JS, Reece WS, Zieschang KD. Imputation of Missing Values When the Probability of Response Depends On the Variable Being Imputed. *J Am Stat Assoc.* 1982; 77 (378):251–261.
- Gruber VA, Delucchi KL, Kielstein A, Batki SL. A randomized trial of 6-month methadone maintenance with standard or minimal counseling versus 21-day methadone detoxification. *Drug Alcohol Depend.* 2008; 94(1–3):199–206. [PubMed: 18243585]
- Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychol Methods.* 2010; 15(3):234–249. [PubMed: 20822250]
- Hedeker, D.; Gibbons, RD. *Longitudinal Data Analysis.* John Wiley & Sons, Inc; Hoboken, New Jersey: 2006.
- Hedeker D, Mermelstein RJ, Demirtas H. Analysis of binary outcomes with missing data: missing = smoking, last observation carried forward, and a little multiple imputation. *Soc St Add.* 2007; 102:1564–1573.
- Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica.* 2003; 71:1161–1189.
- Holland PW. Statistics and causal inference (with discussion). *J Am Stat Assoc.* 1986; 81:945–960.
- Hunter S, Ramchand R, Griffin BA, Suttrop MJ, McCaffrey DF, Morral AR. The effectiveness of community-based delivery of an evidence-based treatment for adolescent substance use. *J Subst Abuse Treat.* (In press).
- Jenson JM, Wells EA, Plotnick RD, Hawkins JD, Catalano RF. The effects of skills and intentions to use drugs on post-treatment drug use of adolescents. *Am J Drug Alcohol Abuse.* 1993; 19:1–18. [PubMed: 8438825]
- Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci.* 2007; 22(4):523–539.
- Kennedy BP, Minami M. The Beech Hill Hospital/Outward Bound Adolescent Chemical Dependency Treatment Program. *J Subst Abuse Treat.* 1993; 10:395–406. [PubMed: 8411298]
- Lee, DS. NBER Working Paper. National Bureau of Economic Research; 2005. Training, wages, and sample selection: Estimating sharp bounds on treatment effects; p. 11721
- Lee DS. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Rev Econ Stud.* 2009; 76(3):1071–1102.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data.* 2. John Wiley; New York: 2002.
- Little RJA, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics.* 1996; 52:98–111. [PubMed: 8934587]
- MacKinnon A, Griffiths KM, Christensen H. Comparative randomised trial of online cognitive-behavioural therapy and an information website for depression: 12-month outcomes. *BJP.* 2008; 192:130–134.10.1192/bjp.bp.106.032078
- McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods.* 2004; 9:403–425. [PubMed: 15598095]
- McGuigan KA, Ellickson PL, Hays RD, Bell RM. Adjusting for attrition in school-based samples: Bias, precision, and cost trade-offs of three methods. *Evaluation Rev.* 1997; 21:554–567.
- Miller WR, Sorensen J, Selzer J, Brigham G. Disseminating evidence-based practices in substance abuse treatment: A review with suggestions. *J Subst Abuse Treat.* 2006; 31(1):25–39. [PubMed: 16814008]
- Miller WR, Zweben J, Johnson WR. Evidence-based treatment: Why, what, where, when, and how? *J Subst Abuse Treat.* 2005; 29:267–276. [PubMed: 16311179]
- Morral AR, McCaffrey DF, Ridgeway G. Effectiveness of community-based treatment for substance-abusing adolescents: 12-month outcomes of youths entering Phoenix Academy or alternative probation dispositions. *Psychol Addict Behav.* 2004; 18:257–268. [PubMed: 15482081]
- Myers MG, Brown SA, Mott MA. Preadolescent conduct disorder behaviors predict relapse and progression of addiction for adolescent alcohol and drug abusers. *Alcohol Clin Exp Res.* 1995; 19:1528–1536. [PubMed: 8749822]

- Neugebauer R, Van der Laan MJ. Why prefer double robust estimates? *J Stat Plan Infer.* 2005; 129(1–2):405–426.
- Okuda M, Hasin DS, Olfson M, Khan SS, Nunes EV, Montoya I, Liu SM, Grant BF, Blanco C. Generalizability of clinical trials for cannabis dependence to community samples. *Drug Alcohol Depend.* 2010; 111(1–2):177–181. [PubMed: 20537813]
- Ramchand R, Griffin BA, Harris KM, Morral AR. Using a cross-study design to assess the efficacy of MET/CBT-5 in treating adolescents with cannabis-related disorders. *J Stud Alcohol Drugs.* 2011; 72(3):380–389. [PubMed: 21513674]
- Richter SS, Brown SA, Mott MA. The impact of social support and self-esteem on adolescent substance abuse treatment outcome. *J Subst Abuse.* 1991; 3:371–385. [PubMed: 1668229]
- Ridgeway, G.; McCaffrey, DF.; Morral, AR. Computer software and reference manual. 2006. twang: Toolkit for weighting and analysis of nonequivalent groups.
- Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Amer Statist Assoc.* 1995; 90:106–121.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983; 70:41–55.
- Rosenbaum PR, Rubin DB. Estimating the effects caused by treatments - comment. *J Am Stat Assoc.* 1984; 79:26–28.
- Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics.* 1985; 41:103–116. [PubMed: 4005368]
- Rubin DB. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *J Educ Psychol.* 1976a; 66:688–701.
- Rubin DB. Inference and Missing Data. *Biometrika.* 1976b; 63:581–592.
- Rush, TV. Predicting treatment outcome for juvenile and young adult clients in the Pennsylvania substance abuse system. In: Beschner, GM.; Friedman, AS., editors. *Youth drug abuse: Problems, issues, and treatment.* Lexington Books; Lexington: 1979. p. 629–656.
- Sampl, S.; Kadden, R. CYT Cannabis Youth Treatment Series. Vol. 1. Center for Substance Abuse Treatment, Substance Abuse and Mental Health Services Administration; Rockville, MD: 2001. Motivational enhancement therapy and cognitive behavioral therapy for adolescent cannabis users: 5 sessions.
- Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable dropouts using semiparametric nonresponse models. *J Amer Statist Assoc.* 1999; 94:1096–1120.
- Schluchter MD. Methods for the analysis of informatively censored longitudinal data. *Statist Medicine.* 1992; 11:1861–1870.
- Scott CK. A replicable model for achieving over 90% follow-up rates in longitudinal studies of substance abusers. *Drug Alcohol Depend.* 2004; 7:21–36. [PubMed: 15072804]
- Shoemaker RH, Sherry P. Posttreatment factors influencing outcome of adolescent chemical dependency treatment. *J Child Adolesc Subst Abuse.* 1991; 2:89–106.
- Troxel AB. A comparative analysis of quality of life data from a Southwest Oncology Group randomized trial of advanced colorectal cancer. *Statist Medicine.* 1998; 17:767–779.
- Troxel AB, Harrington DP, Lipsitz SR. Analysis of longitudinal measurements with non-ignorable non-monotone missing values. *Appl Statist.* 1998; 47:425–438.
- Troxel AB, Ma G, Heitjan DF. An index of local sensitivity to nonignorability. *Stat Sinica.* 2004; 14:1221–1237.
- Verbeke G, Molenberghs G, Thijs H, Lesaffre E, Kenward MG. Sensitivity analysis for nonrandom dropout: a local influence approach. *Biometrics.* 2001; 57:7–14. [PubMed: 11252620]
- Walton MA, Ramanathan CS, Reischil TM. Tracking substance abusers in longitudinal research: understanding follow-up contact difficulty. *Am J Community Psychol.* 1998; 26:233–253. [PubMed: 9693691]
- Weisz J, Donenberg G, Han S, Weiss B. Bridging the gap between laboratory and clinic in child and adolescent psychotherapy. *J Consult Clin Psychol.* 1995; 63(5):688–701. [PubMed: 7593861]

- Witbrodt J, Bond J, Kaskutas LA, Weisner C, Jaeger G, Pating D, Moore C. Day hospital and residential addiction treatment: randomized and nonrandomized managed care clients. *J Consult Clin Psychol*. 2007; 75(6):947–959. [PubMed: 18085911]
- Wooldridge, JM. *Econometric Analysis of Cross Section and Panel Data*. 1. The MIT Press; Cambridge: 2001.
- Wooldridge, JM. *Econometric analysis of cross section and panel data*. MIT Press; Cambridge: 2002.
- Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*. 1988; 44:175–188.
- Xie H, Qian Y. Measuring the impact of nonignorability in panel data with non-monotone nonresponse. *J Appl Econ*. 2012; 27(1):129–159.

Appendix: Details on propensity score weighting

The propensity score weights were estimated using a Generalized Boosted Model (GBM) which is a non-parametric estimation technique that can account for a large number of pre-treatment covariates. GBM adaptively captures the functional form of the relationship between the covariates and treatment selection with less bias than traditional approaches (McCaffrey et al. 2004; Harder et al. 2010; Lee 2009). The “twang” package in R was used to conduct the weight estimation (Ridgeway et al. 2006).

Following the weighting procedure, we calculated the standardized mean difference of each of the pretreatment variables to measure the similarity of the CYT and propensity-score weighted comparison EAT samples. The standardized mean difference is defined using the following formula:

$$\frac{\widehat{\mu}_t - \widehat{\mu}_c}{\widehat{\sigma}_t},$$

where $\widehat{\mu}_j$ denotes the estimated mean value for the treatment and comparison conditions ($j = t$ and c , respectively); $\widehat{\sigma}_t$ denotes the estimated standard deviation in the treatment condition (the CYT/experimental youth) for a given pretreatment variable. Values of 0 for a standardized mean difference thus represent no difference in means while values of + or –1 represent one standard deviation difference between the two groups. Absolute standardized differences greater than 0.25 are considered to be ‘moderate effect size differences’ (Cochran 1968). To control for the possible confounding effects of pretreatment group differences that remain after weighting (Neugebauer and Van der Laan 2005), any variables for which the absolute standardized mean difference (ASMD) was greater than 0.25 were included in our regression models.

We began this analysis with 60 pre-treatment variables related to substance abuse patient placement criteria established by the American Society for Addiction Medicine (American Society of Addiction Medicine 2001); these 60 variables have been used in previous investigations examining the effectiveness of adolescent drug treatment (McCaffrey et al. 2004). Many of these characteristics have been shown in prior work to influence substance abuse treatment outcomes, including pre-treatment levels of substance use (Alford et al. 1991; Shoemaker and Sherry 1991; Jenson et al. 1993; Kennedy and Minami 1993), symptoms associated with emotional well-being and criminality (Alford et al. 1991; Brown et al. 2000; Brown et al. 1996; Myers et al. 1995), academic/scholastic attendance/performance (Rush 1979; Shoemaker and Sherry 1991), employment (Rush 1979), sociodemographics (Alford et al. 1991; Cady et al. 1996; Friedman and Glickman 1986; Friedman et al. 1986; Shoemaker and Sherry 1991), and social/familial substance use (Richter et al. 1991; Shoemaker and Sherry 1991).

Since controlling for variables unrelated to the outcomes may not improve estimates of treatment effects (Wooldridge 2001), we selected from this initial set of 60 pre-treatment covariates only those pretreatment covariates that explained at least one percent of the variance in our outcomes for inclusion our propensity score model. This resulted in a total of 22 pretreatment variables which are listed in Table 1. Missing values on the 22 pretreatment variables were low (mean = 0.60% and max = 4.87%) and were controlled for in the propensity score model by balancing the two groups on missing value indicators for each variable in addition to balancing on the variables themselves.

Figure 1.

Standardized mean differences (plus 95% confidence intervals) in outcomes among youth in the experimental setting (CYT) and propensity score weighted youth in the community-based sample (EAT).

Figure 2.

Sensitivity analysis showing (i) the value of SSI for which the mean treatment effect estimate for the FF group found by comparing MET/CBT-5 among the community responders and synthetic experimental FF group across the 100 simulations was zero (i.e., gave treatment effect estimates indicating experimental and community settings were the same) and (ii) the value of SSI for which 95% of the simulations favored youth in the experimental setting (i.e., gave treatment effect estimates in favor of the experimental setting for the 2.5th and 97.5th percentiles).

Note: SPS = Substance Problem Scale; SFS = Substance Frequency Scale; EPS = Emotional Problems Scale; IAS = Illegal Activities Scale

Table 1

Classification of Youth by Potential Response Indicators

R_1	R_0	
	0	1
0	DD	DF
1	FD	FF

D = dropout; F = followed-up

Table 2

Mean values on pre-treatment characteristics of the experimental (CYT) sample and the weighted and unweighted community-based treatment (EAT) samples at baseline

	CYT	EAT unweighted	EAT weighted
Pre-Treatment Characteristic (<i>range</i>)	n=183	n=2908	ESS=224
Demographics			
Race/Ethnicity (<i>0-1</i>)			
White/Caucasian	0.67	0.51 [*]	0.62
Black/African-American	0.27	0.11 [*]	0.23
Hispanic	0.03	0.22 [*]	0.09 [*]
Other	0.03	0.16 [*]	0.06
Current Living Situation (<i>0-1</i>)			
House/apartment	0.91	0.87	0.90
Friend/relatives house	0.07	0.05	0.05
Institutionalized Setting	0.00	0.02	0.02
Other	0.01	0.06	0.03
Training Activity Scale (<i>0-1</i>)	0.61	0.68 [*]	0.65
Crime and Justice			
Crime Violence Scale (<i>0-31</i>)	6.14	6.16	6.23
Illegal Activity Scale (<i>0-1</i>)	0.23	0.09 [*]	0.16 [*]
Criminal Justice System Index (<i>0-1</i>)	0.35	0.32	0.35
Environment			
Controlled Environment Index (<i>0-1</i>)	0.06	0.07	0.05
Environment Risk Scale (<i>0-84</i>)	31.49	35.93 [*]	33.41
Max. days in a controlled environment (<i>0-90</i>)	5.26	5.64	4.54
Social Risk Index (<i>0-28</i>)	14.58	13.17 [*]	14.02
Mental Health			
Any Prior MH Treatment (<i>0-1</i>)	0.27	0.36	0.33
Behavior Complexity Scale (<i>0-33</i>)	10.27	9.80	10.53
Emotional Problems Scale (<i>0-1</i>)	0.22	0.20	0.22
Internal Mental Distress Scale (<i>0-43</i>)	6.76	7.07	6.65
Mental Health Treatment Index (<i>0-1</i>)	0.01	0.01	0.01
Substance Use			
Days drunk or high most of day-past 90 days	13.75	12.59	11.88
Problem Orientation Scale (<i>0-5</i>)	0.72	0.65	0.74
Substance Dependence Scale-past month (<i>0-7</i>)	1.27	0.87	0.97
Substance Dependence Scale-past year (<i>0-7</i>)	2.66	2.28	2.38
Substance Frequency Scale (<i>0-1</i>)	0.18	0.11 [*]	0.14 [*]
Substance Problem Scale-past month (<i>0-16</i>)	3.72	2.81 [*]	3.10
Substance Problem Scale-past year (<i>0-16</i>)	7.09	6.35	6.61

Notes: CYT= Cannabis Youth Trial; EAT= Effective Adolescent Treatment sample; ESS = Effective Sample Size;

*
= Absolute standardized mean difference between value and CYT comparison value > 0.25