Published in final edited form as:

Am J Addict. 2018 April ; 27(3): 231–237. doi:10.1111/ajad.12714.

Generalizability of the Findings from a Randomized Controlled Trial of a Web-based Substance Use Disorder Intervention

Ryoko Susukida, PhD^{1,2}, Rosa M. Crum, MD, MHS^{1,3,4}, Elizabeth A. Stuart, PhD^{1,5,6}, and Ramin Mojtabai, MD, PhD, MPH^{1,4}

¹Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, Baltimore, MD, 21205, USA

²Department of Mental Health Policy and Evaluation, National Institute of Mental Health, National Center of Neurology and Psychiatry, 4-1-1 Ogawa-Higashi, Kodaira, Tokyo 187-8551, Japan

³Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD, 21205, USA

⁴Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, 600 N. Wolfe St. Baltimore, MD 21287

⁵Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Broadway, Baltimore, MD, 21205, USA

⁶Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, Baltimore, MD, 21205, USA

Abstract

Background and Objectives: There is growing concern regarding the generalizability of findings from randomized controlled trials (RCTs) of interventions for substance use disorders (SUDs). This study used a selection model approach to assess and improve the generalizability of an evaluation for a web-based SUD intervention by making the trial sample resemble the target population.

Methods: The sample of the web-based SUD intervention (Therapeutic Education System vs. Treatment-as-usual; n=507) was compared with the target population of SUD treatment-seeking individuals from the Treatment Episodes Data Set-Admissions (TEDS-A). Using weights based on the probabilities of RCT participation, we computed weighted treatment effects on retention and abstinence.

Results: Substantial differences between the RCT sample and the target population was demonstrated in significant difference in the mean propensity scores (1.62 standard deviations at $P<.001$). The population effect on abstinence (12 weeks and 6 months) was statistically insignificant after weighting the data with the generalizability weight.

Discussions and Conclusions: Generalizability of the findings from the RCT could be limited when the RCT sample does not well represent the target population.

Scientific Significance: Application of generalizability weights can be a potentially useful tool to improve generalizability of RCT findings.

Introduction

Substance use disorders (SUDs) impose significant societal and economic burdens. In spite of high prevalence of SUDs and its various negative health consequences,¹ many individuals with SUDs do not receive treatment.² Blanco et al.³ estimated that only 13% of those with drug dependence and 5% of those with alcohol dependence sought treatment within the first year of the disorder onset in 2004–2005 in the United States. Although there are a number of effective evidence-based interventions for SUD,⁴ strong stigma toward these conditions⁵ and the limited access to SUD specialty treatment⁶ often prevent those with SUDs from receiving effective treatment. There is a clear need for SUD treatments with greater acceptability and accessibility.

Web-based SUD treatment is a promising behavioral intervention to treat individuals with SUDs who may not be willing or able to receive traditional face-to-face interventions.⁷ A systematic review by Moore et al.⁷ found that web-based SUD interventions decreased substance use, increased retention in treatments, raised patients' motivation to change, and improved knowledge of SUDs as compared with standard treatments. A web-based SUD treatment can offer potential benefits including lower implementation cost, greater scalability, greater accessibility in remote or rural areas that have limited options for SUD specialty treatment, higher confidentiality, 24-hour accessibility, opportunities for more frequent and longer intervention duration, and greater convenience and flexibility of access from patients' homes without a need for appointments.^{7,8} A growing number of randomized controlled trials (RCTs) demonstrated that web-based SUD interventions had higher treatment retention and resulted in greater motivation to change, decreased substance use, and greater knowledge about SUD as compared with treatment-as-usual.⁷ Moreover, some studies have shown that web-based cognitive-behavioral SUD interventions combined with Contingency Management interventions are effective in treating SUDs.^{9–11}

The promising treatment effects shown by the RCTs of web-based SUD interventions, however, do not necessarily guarantee the external validity of the findings to different populations. Particularly, limited external validity of the findings from RCTs is a concern when the characteristics of RCT participants are different from those of the target population for whom an intervention is intended and those characteristics moderate treatment effects. There are a growing number of studies suggesting that the participants of RCTs may not represent the target populations well, especially in the context of SUD treatments.^{12–14} A recent review by Moberg and Humphreys¹² found that commonly used exclusion criteria in RCTs of SUD treatments would exclude between 64% and 95% of potential participants. A study by Susukida et al.¹⁵ directly compared characteristics of participants in ten SUD RCTs with the target populations of individuals receiving SUD treatment in usual care settings and found that a significantly higher proportion of the RCT participants had higher educational

attainment and full-time jobs than those in the target populations. Another study by Susukida et al.¹⁶ examined how lack of representativeness of the SUD RCT samples affected the findings of the RCTs. The investigators¹⁶ used statistical weighting techniques to make the SUD RCT samples resemble the target populations and estimate population treatment effects, and showed that significant sample treatment effects often became insignificant after weighting.

Web-based SUD interventions potentially have greater scalability to broader target populations than clinic-based interventions. However, very few previous studies assessed the representativeness of the RCT participants of web-based SUD interventions as compared with the intended target populations, and whether and how the sample representativeness impacts the generalizability of RCT findings. A recent study by Blanco et al.¹⁷ directly compared the RCT sample of the web-based intervention, the Therapeutic Education System (TES), with two target populations: individuals seeking SUD treatment and individuals with SUD regardless of their treatment seeking behavior, both drawn from the Wave 1 of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), conducted between 2001 and 2002. These authors found substantial differences in characteristics between the RCT sample and the two target populations and also demonstrated that the treatment effect of TES became insignificant after the sample was weighted to resemble these target populations. However, illicit drug use in the US has been increasing in the past 10 to 15 years and the NESARC target populations used in Blanco et al.¹⁷ may not appropriately represent the current population with SUD. Approximately 9.4 percent of the US population used an illicit drug during the past month in 2013 compared with 8.3 percent in 2002.¹⁸ In addition to the higher prevalence of illicit drug use in recent years, the composition of those with illicit drug use has also changed. According to the report of the Substance Abuse and Mental Health Services Administration,¹⁹ the higher prevalence of illicit drug use among those 12 or older in recent years was primarily due to increases among those aged 18 years or older. Moreover, among various illicit drugs, in recent years marijuana and heroin have particularly increased in use by those aged 18 or older.¹⁹ Given these recent changes in nationwide trends in substance use behaviors, it is important to assess the representativeness of the SUD RCTs compared with more current target population. Furthermore, in Blanco et al.,¹⁷ both the treatment-seeking target population and the general target population were drawn from the same source, a general population survey that did not capture marginalized population groups such as homeless and/or incarcerated individuals.

The aims of this current study were (1) to assess the sample representativeness of an RCT of a web-based TES with the target population of individuals with recent drug use admitted into SUD treatment in usual care settings, and (2) to estimate sample treatment effects and the population effects of the web-based intervention. Unlike the study by Blanco et al.,¹⁷ this study used more recent target population, which was drawn from administrative data on individuals in treatment settings, including patients referred to treatment through the judicial system and marginalized population groups.

3.3. Methods

Data sources

Data for the TES RCT sample were drawn from the National Institute of Drug Abuse (NIDA) Clinical Trials Network (CTN) Data share Website.²⁰ This TES trial was part of the CTN studies, which are multi-site nationwide RCTs actively engaging diverse populations in community settings instead of conducting trials in specialized research settings. TES is a computerized psychosocial intervention of skill building modules with modest incentives of opportunities to draw vouchers for congratulatory messages or monetary prizes (mostly \$1, occasionally \$20, and rarely \$80-\$100), which are provided upon completion of the modules and abstinence from substance use.²¹ The 12-week trial successfully included a relatively larger group of study participants (N=507, recruited between June 2010 and August 2011) than previous web-based SUD interventions.¹⁰ Participants seeking outpatient treatment for SUD treatment who reported any illicit drug use in the past 30 days were randomized into two arms. Individuals were excluded from the trial if they were receiving opioid replacement medication, planning to move to different areas within the next three months, unable to provide informed consent for trial participation, and/or lacking English proficiency. Out of 1781 patients who initially consented to eligibility screening, 850 patients were excluded and 424 patients declined to participate in the trial.¹⁰ One arm received treatment as usual (TAU), comprised of standard SUD outpatient treatment, and the other arm was TAU plus TES. Illicit drugs assessed in TES included cocaine, amphetamine, cannabinoids (THC), methamphetamine, non-prescribed benzodiazepines, opioids (morphine, codeine, heroin, oxycodone, methadone), barbiturates, and MDMA.

The RCT sample was compared with the target population drawn from the 2012 Treatment Episodes Data Set-Admissions (TEDS-A)—the most recent data available at the time of this writing. The TEDS-A is an administrative database maintained by the Center for Behavioral Health Statistics and Quality (CBHSQ) of the Substance Abuse and Mental Health Services Administration (SAMHSA). The TEDS-A includes annual data on more than 1.5 million admissions of individuals aged 12 years or older to publicly funded SUD treatment facilities across the US. To make the sample comparable to the RCT sample we limited the TEDS-A sample to patients who were 18 years or older, received treatment in outpatient settings, and reported illicit drug use in the past 30 days. Since only a small portion of the TEDS-A were missing variables used in the analysis (n = 14,712, 2.4% out of 610,766), we conducted statistical analyses with complete cases (n = 596,054).

Measures

There were eight common variables assessed in the RCT sample and the target population: sex, race-ethnicity (white, black, Hispanic and other), age (recoded into 18–34, 35–49, 50 or older), educational attainment (less than high school, high school, more than high school), employment status (full-time, part-time, out of labor, unemployed), marital status (Never married, married, separated/divorced/widowed), intravenous drug use, and the history of past treatments for SUD.

Two outcomes from the RCT were generalized to the target population: successful retention in the study and abstinence from substance use. Successful retention in the study was defined by remaining in the RCT at each assessment point, which occurred at the end of trial (12 weeks), and at the 3-months and 6-month follow-ups. Study participants were considered “abstinent” if they submitted a negative urine toxicology sample and reported no drug use or heavy alcohol drinking (i.e., daily consumption of 5 or more drinks for men and 4 or more drinks for women) in the prior 4 days of each assessment point. Both outcomes were binary measures.

Statistical analysis

We first compared the RCT sample with the target population with regard to the eight observed characteristics noted above. Pearson’s χ^2 tests were conducted to compare the composition of the RCT sample and the target population. For variables with three or more categories, we conducted Pearson’s χ^2 tests by making binary variables for each category (e.g., White patients vs. Non-White patients). The cutoff for statistically significant differences between RCT and TEDS-A samples was set at $p < 0.0028$, which was adjusted at $0.05/18 = 0.0028$ for multiple testing correction.

Next, we estimated probabilities of trial participation based on the eight variables, for every individual both in the RCT sample and the target population. The participation probabilities (related to propensity scores for non-experimental studies) were estimated using the R package ‘randomForest’.²² While the random forests approach has the ability to reduce outlying propensity scores,²³ we still encountered some extreme values of propensity scores. We used weight trimming, also referred as truncation, in which we replaced extreme large values at the 95th percentile values following the method introduced in a study by Lee et al²⁴ in order to improve the performance of the propensity score based weighting.

After estimating the participation probabilities for each individual, we calculated a difference of mean participation probabilities between the RCT sample as a group and the target population. This mean participation probability difference, p , was introduced by Stuart et al.²⁵ as a measure to evaluate the representativeness of the RCT sample as compared with the target population. Standardized p , which is p divided by the pooled standard deviation of the propensity scores, is a summary index representing the difference between the RCT sample and its target population. In observational studies, values of p larger than 0.25 are generally considered as an indication of substantial difference between two samples.²⁶ Mamdani et al.²⁷ suggested a more conservative cutoff of p larger than 0.10 as indicating a meaningful difference between two samples.

We calculated the participation weight for each trial participant as $(1 - p)/p$, with which weighted outcome regression analyses were conducted using STATA *pweight* command (version 13) to estimate the population treatment effects. This weighting-based method allows researchers to estimate the treatment effects in the target population. This method is similar to inverse probability weighting method for non-experimental studies, with which the causal effect is estimated by making the unexposed group similar to the exposed group based on the commonly observed characteristics between these two groups.²⁸ This study weighed the RCT sample to resemble the target population. We used inverse odds weights,

$(1 - p)/p$, which are considered to be more appropriate when the study sample and its target population do not overlap,²⁹ such as in the case of this study where the RCT sample was not part of the TEDS-A sample. We used logistic regression models for two types of binary outcomes (retention and abstinence). We did not control for baseline characteristics in regression analyses, assuming that randomization was successful in each trial. We then statistically compared the effect sizes of unweighted and weighted models, with the STATA *suest* (seemingly unrelated estimation) command.³⁰

3.4. Results

Comparison of characteristics of RCT sample and target population

Table 1 presents the characteristics of the RCT participants and the target population. As compared with the TEDS-A target population, the RCT sample had significantly lower proportions of individuals with intravenous drug use, and those with a history of prior SUD treatment. On the other hand, the RCT sample had significantly higher proportions of individuals with 12 years or longer educational attainment and with full time jobs, as compared with the TEDS-A target population.

Comparison of propensity scores of RCT sample and target population

The estimated mean propensity score for the RCT sample (0.77) was significantly larger than for the target population (0.26). The standardized p (p divided by the pooled standard deviation) was approximately 1.62, which was substantially larger than the 0.25 standardized p cut-off suggested in the literature.^{26,31}

Comparison of unweighted outcomes and outcomes weighted by propensity scores

Table 2 presents the unweighted and weighted treatment effects on trial retention and abstinence. Odds ratios (ORs) from both unweighted and weighted logistic regression analyses with the 95% confidence intervals are presented. The unweighted treatment effects represent the effects in the RCT sample while the weighted treatment effects represent the effects that would be expected if the RCT sample was made to resemble the target population. In unweighted analyses, treatment was associated with significantly greater odds of abstinence at 12 weeks and 6-month follow-up. Significant treatment effects on abstinence at 12 weeks and 6-month follow-up became statistically non-significant after weighting with the generalizability weight. The treatment effect on retention was insignificant both in the unweighted and weighted analyses. For both abstinence and retention outcomes, comparison of treatment effect sizes did not reveal significant differences between unweighted and weighted models with the generalizability weight.

3.5. Discussion

This study demonstrated significant differences between the RCT sample of a web-based SUD intervention and target population of potential recipients of the intervention. The composition of the RCT sample substantially differed from those of the target population. Furthermore, the summary index of these differences, p , far exceeded the standardized p cutoff 0.25²⁶ or 0.10²⁷, indicating substantial differences between the RCT sample and the

target population. Standardized p can be interpreted similarly to *Cohen's d* effect size.³² Cohen's d value of 1.62 is equivalent to a 42% degree of overlap between the RCT sample and the target population. This study also demonstrated that the observed promising findings of the TES intervention may not be directly applicable to the potential target population. We showed that significant treatment effects on abstinence at 12 weeks and 6-month follow-up became insignificant after weighting. We also found some suggestive evidence that these differences between the unweighted effects and the weighted effects could be partially explained by the treatment effect heterogeneity across under- and over-represented subgroups of RCT participants, although the RCT sample was not sufficiently powered because exploration of treatment effect heterogeneity was not the primary purpose of the RCT (results available upon request).

The findings of this study are consistent with previous studies by Susukida et al.,¹⁵ which found statistically significant differences between the samples from ten SUD RCTs and the corresponding target populations. The findings are also consistent with Susukida et al.,¹⁶ which found the significance of estimated sample treatment effects was different from that of the population effects when the distribution of characteristics of RCT samples were made to resemble the distribution of the target populations by using the same statistical weighting techniques that this study used. Especially in the context of the generalizability of the findings of a web-based SUD intervention, this study's findings echo the findings of the recent study by Blanco et al.,¹⁷ which found that the significant treatment effect of TES estimated through RCT was insignificant after weighting the sample to resemble the target populations drawn from the 2001–2002 NESARC data. Our study confirms that their findings hold when the generalizability of the TES RCT was assessed with the target population from more recent years and with a population obtained from administrative data from usual treatment settings across the US. More than 70 % of the patients who were initially invited to the trial did not participate in the TES RCT due to either exclusion criteria or refusal to participate, which might have affected the representativeness of the trial sample and generalizability of the RCT findings to the target population.

Some limitations should be taken into consideration when interpreting this study's findings. First, only a limited number of characteristics were assessed in both the RCT sample and the target population. The presence of co-existing mental and/or physical disorders and other medications that patients received could have potentially affected the probability of RCT participation, and impacted treatment results. Therefore, the estimated propensity scores did not reflect such characteristics that may have differed between the RCT and the target population. Future studies should explore more potential treatment effect modifiers that over- or under-represent in RCT samples as compared to the target populations because these variables can compromise generalizability of the RCT findings.³³ Also, future studies would benefit from a method recently introduced by Nguyen et al.,³⁴ which allows researchers to estimate the population treatment effects in cases where treatment effect modifiers are not observed in target populations, but only in RCTs. Second, although the TEDS-A is one of the largest administrative data sources that covers patients with SUD in the US, some states exclude patients whose treatment is not covered by the state substance use agency funds such as the Federal Block Grant fund.³⁵ Patients who received treatment at private hospitals are usually excluded from the TEDS-A unless these treatment settings are

licensed by the state SUD treatment agency. Therefore, the TEDS-A may not completely represent the entire population of patients with SUD in the US.

Notwithstanding these limitations, findings from this study provide insights into differences between the RCT participants of a web-based SUD intervention and the target population from recent years. The results of this study also indicate how poor sample representativeness of the RCT compared with the target population impacted the observed findings of the web-based SUD intervention. Given the great potential for scalability of web-based SUD interventions,⁷ the representativeness of the sample with regard to the target population of potential users for this intervention should be carefully considered. Also, with a careful consideration of treatment effect modifiers which are over- or under-representing in RCTs, application of generalizability weights could be a potentially useful tool for assessing and improving the generalizability of the findings from RCTs when the RCT sample does not well-represent the target population.

Acknowledgement

This work was supported by grant R01 DA036520 from the National Institute on Drug Abuse (NIDA). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest

Dr. Susukida has nothing to disclose. Drs. Crum, Stuart and Mojtabai report grants from the National Institute on Drug Abuse, National Institute of Mental Health, and/or the National Institute on Alcohol Abuse and Alcoholism during the conduct of the study. Dr. Mojtabai has received research funding and consulting fees from Bristol-Myers Squibb and Lundbeck Pharmaceuticals.

References

1. Degenhardt L, Hall W. Extent of illicit drug use and dependence, and their contribution to the global burden of disease. *Lancet*. 2012;379(9810):55–70. doi:10.1016/S0140-6736(11)61138-0. [PubMed: 2225671]
2. Compton WM, Thomas YF, Stinson FS, et al. Prevalence, correlates, disability, and comorbidity of DSM-IV alcohol abuse and dependence in the United States: results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Arch Gen Psychiatry*. 2007;64(7):830–842. doi:10.1001/archpsyc.64.5.566. [PubMed: 17606817]
3. Blanco C, Iza M, Schwartz RP, Rafful C, Wang S, Olfson M. Probability and predictors of treatment-seeking for prescription opioid use disorders: A National Study. *Drug Alcohol Depend*. 2013;131(1–3):143–148. doi:10.1016/j.drugalcdep.2012.12.013. [PubMed: 23306097]
4. Jhanjee S. Evidence based psychosocial interventions in substance use. *Indian J Psychol Med*. 2014;36(2):112–118. doi:10.4103/0253-7176.130960. [PubMed: 24860208]
5. Keyes KM, Hatzenbuehler ML, McLaughlin KA, et al. Stigma and treatment for alcohol disorders in the United States. *Am J Epidemiol*. 2010;172(12):1364–1372. doi:kwq304 [pii]r10.1093/aje/kwq304. [PubMed: 21044992]
6. Substance Abuse and Mental Health Services Administration. Results from the 2012 National Survey on Drug Use and Health: Summary of National Findings.; 2012 <http://store.samhsa.gov/home>.
7. Moore BA, Fazzino T, Garnet B, Cutter CJ, Barry DT. Computer-based interventions for drug use disorders: A systematic review. *J Subst Abuse Treat*. 2011;40(3):215–223. doi:10.1016/j.jsat.2010.11.002. [PubMed: 21185683]

8. Budman SH. Behavioral health care dot-com and beyond: Computer-mediated communications in mental health and substance abuse treatment. *Am Psychol.* 2000;55(November):1290–1300. doi: 10.1037/0003-066X.55.11.1290. [PubMed: 11280939]
9. Higgins ST, Sigmon SC, Wong CJ, et al. Community Reinforcement Therapy for Cocaine-Dependent Outpatients. *Arch Gen Psychiatry.* 2003;60:1043–1052. doi:10.1001/archpsyc.60.9.1043. [PubMed: 14557150]
10. Campbell ANC, Nunes EV, Matthews AG, et al. Internet-delivered treatment for substance abuse: a multisite randomized controlled trial. *Am J Psychiatry.* 2014;171(6):683–690. doi:10.1176/appi.ajp.2014.13081055. [PubMed: 24700332]
11. Bickel WK, Marsch LA, Buchhalter AR, Badger GJ. Computerized behavior therapy for opioid-dependent outpatients: A randomized controlled trial. *Exp Clin Psychopharmacol.* 2008;16(2):132–143. doi:10.1037/1064-1297.16.2.132. [PubMed: 18489017]
12. Moberg CA, Humphreys K. Exclusion criteria in treatment research on alcohol, tobacco and illicit drug use disorders: A review and critical analysis. *Drug Alcohol Rev.* 2016. doi:10.1111/dar.12438.
13. Storbjörk J, Garfield JBB, Lerner A. Implications of Eligibility Criteria on the Generalizability of Alcohol and Drug Treatment Outcome Research: A Study of Real-World Treatment Seekers in Sweden and in Australia. *Subst Use Misuse.* 2017;52(4):439–450. doi: 10.1080/10826084.2016.1240695. [PubMed: 27849416]
14. Dennis BB, Roshanov PS, Naji L, et al. Opioid substitution and antagonist therapy trials exclude the common addiction patient: a systematic review and analysis of eligibility criteria. *Trials.* 2015;16(1):475. doi:10.1186/s13063-015-0942-4. [PubMed: 26489415]
15. Susukida R, Crum RM, Stuart EA, Ebnesajjad C, Mojtabai R. Assessing Sample Representativeness in Randomized Control Trials: Application to the National Institute of Drug Abuse Clinical Trials Network. *Addiction.* 2016.
16. Susukida R, Crum RM, Ebnesajjad C, Stuart EA, Mojtabai R. Generalizability of findings from randomized controlled trials: application to the National Institute of Drug Abuse Clinical Trials Network. *Addiction.* 2017;112(7):1210–1219. doi:10.1111/add.13789. [PubMed: 28191694]
17. Blanco C, Campbell AN, Wall MM, Olfson M, Wang S, Edward VN. Toward National Estimates of Effectiveness of Treatment for Substance Use. *J Clin Psychiatry.* 2017;78(1):e64–e70. [PubMed: 28129499]
18. National Institute on Drug Abuse. DrugFacts: Nationwide trends. <http://www.drugabuse.gov/publications/drugfacts/nationwide-trends>. Accessed November 6, 2016.
19. Substance Abuse and Mental Health Services Administration. Behavioral Health Trends in the United States: Results from the 2014 National Survey on Drug Use and Health.; 2015 <https://www.samhsa.gov/data/sites/default/files/NSDUH-FRR1-2014/NSDUH-FRR1-2014.pdf>. Accessed November 6, 2016.
20. National Institute on Drug Abuse Clinical Trials Network. Clinical Trials Network (CTN): Research Studies. <http://www.drugabuse.gov/about-nida/organization/cctn/ctn/research-studies>. Accessed November 6, 2016.
21. Budney AJ, Higgins ST. Therapy Manuals for Drug Addiction. Manual 2, a Community Reinforcement plus Vouchers Approach: Treating Cocaine Addiction.; 1998 <http://www.nida.nih.gov/TXManuals/CRA/CRA1.html>. Accessed November 6, 2016.
22. Liaw A, Wiener M. Classification and Regression by randomForest. *R news.* 2002;2(December):18–22. doi:10.1177/154405910408300516.
23. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods.* 2009;14(4):323–348. doi:10.1037/a0016973. [PubMed: 19968396]
24. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One.* 2011;6(3). doi:10.1371/journal.pone.0018174.
25. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. *J R Stat Soc Ser A.* 2011;174(2):369–386. doi: 10.1111/j.1467-985X.2010.00673.x.

26. Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci.* 2010;25(1):1–21. doi:10.1214/09-STS313. [PubMed: 20871802]
27. Mamdani M, Sykora K, Li P, et al. Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. *BMJ.* 2005;330(7497):960–962. doi:10.1136/bmj.330.7497.960. [PubMed: 15845982]
28. Mansournia MA, Altman DG. Inverse probability weighting. *BMJ Br Med J.* 2016. doi:10.1136/bmj.i189
29. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of Trial Results Using Inverse Odds of Sampling Weights. *Am J Epidemiol.* 2017;186(8):1010–1014. doi:10.1093/aje/kwx164. [PubMed: 28535275]
30. Weesie J. Seemingly unrelated estimation and the cluster-adjusted sandwich estimator. *Stata Tech Bull.* 1999;52:34–47. papers3://publication/uuid/90F5400C-1AAC-4ED5-A779-90373636BB8D. Accessed November 6, 2016.
31. Cochran WG, Rubin DB. Controlling bias in observational studies: A review. *Sankhy Indian J Stat Ser A.* 1973;35(4):417–446.
32. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* 2nd Edition Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
33. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prev Sci.* 2015;16(3):475–485. [PubMed: 25307417]
34. Nguyen TQ, Ebnesajjad C, Cole SR, Stuart EA. Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Ann Appl Stat.* 2017;11(1):225–247. doi:10.1214/16-AOAS1001.
35. Substance Abuse & Mental Health Services and Quality. Quick Statistics from the Drug and Alcohol Services Information System. <http://www.dasis.samhsa.gov/webt/information.htm>. Accessed November 6, 2016.

Table 1.

Comparison of baseline characteristics (%) of the samples in the Therapeutic Education System RCT and the target population (TEDS-A)

	Therapeutic Education System RCT (n=507)	TEDS-A (2012) target population (n=596,054)	p-value for differences between RCT and TEDS-A
	Percent	Percent	
Sex			
Female	37.9	36.6	0.584
Male	62.1	63.4	
Race			
White	58.2	59.0	0.702
Black	21.9	22.3	0.927
Hispanic	10.8	14.1	0.028
Other	9.1	4.7	<0.001
Age			
18–34	51.1	58.3	0.001
35–49	34.3	29.4	0.014
50 and over	14.6	12.3	0.115
Education			
<12 years	23.3	32.2	<0.001
12 years	76.7	67.8	
Employment			
Full-time	40.0	16.3	<0.001
Part-time	23.1	9.8	<0.001
Out of labor	10.7	31.1	<0.001
Unemployed	26.2	42.8	<0.001
Marital Status			
Married	14.2	15.0	0.636
Never married	60.8	64.8	0.054
Separated/divorced/widowed	25.0	20.2	0.007
IV drug use			
Yes	7.5	19.0	<0.001
No	92.5	81.0	
Prior treatments			
Yes	7.7	58.3	<0.001
No	92.3	41.7	

Notes: Pearson's χ^2 test was conducted. For variables with three or more categories, we conducted Pearson's χ^2 tests by making binary variables for each category (e.g., White patients vs. Non-White patients). Numbers written in bold letters indicate statistically significant differences between RCT and TEDS-A samples at $p < 0.0028$ (cutoff of statistical significance was adjusted at $0.05/18 = 0.0028$ for multiple testing correction).

Table 2.

Comparison of unweighted (RCT sample effect) and weighted (population effect) odds ratios of treatment effect on retention and abstinence

	Unweighted			Weighted			p	95%CI	p	Comparison (unweighted vs. weighted)
	OR	95%CI	p	OR	95%CI	p				
Retention 12 weeks	0.92	0.55, 1.52	0.74	0.80	0.30, 2.18	0.67				$\chi^2 = 0.05$, $p = 0.82$
Retention 3 months	0.60	0.33, 1.08	0.09	1.09	0.35, 3.35	0.88				$\chi^2 = 0.85$, $p = 0.36$
Retention 6 months	1.02	0.72, 1.45	0.90	1.16	0.60, 2.25	0.66				$\chi^2 = 0.11$, $p = 0.74$
Abstinence 12 weeks	1.67 ^{**}	1.15, 2.44	<0.01	1.62	0.80, 3.26	0.18				$\chi^2 = 0.01$, $p = 0.94$
Abstinence 3 months	1.18	0.82, 1.71	0.37	0.74	0.37, 1.48	0.39				$\chi^2 = 1.38$, $p = 0.24$
Abstinence 6 months	2.04 [*]	1.09, 3.85	0.03	1.65	0.50, 5.49	0.41				$\chi^2 = 0.09$, $p = 0.76$

¹. Weight is calculated as $(1-p)/p$, where p is a propensity score of being in a trial sample.

². Weight was truncated at 95 percentiles to eliminate extreme weights.

³. ** $p < 0.01$, * $p < 0.05$.

⁴. Individuals were considered “abstinent” if they submitted negative urine toxicology sample and they reported no drug use or heavy alcohol drinking (i.e., daily consumption of 5 or more drinks for men and 4 or more drinks for women) in the last 4 days of the assessment.