Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Addiction. Author manuscript; available in PMC 2018 July 01.

Published in final edited form as:

Addiction. 2017 July ; 112(7): 1210–1219. doi:10.1111/add.13789.

Generalizability of Findings from Randomized Controlled Trials: Application to the National Institute of Drug Abuse Clinical Trials Network

Ryoko Susukida, PhD[1], **Rosa M. Crum, MD**[1,2,3], **Cyrus Ebnesajjad, MA**[1,4], **Elizabeth A. Stuart, PhD**[1,5,6], and **Ramin Mojtabai, MD**[1,3]

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

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

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

[4]The Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, WA 98109

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

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

Abstract

Aims—To compare randomized controlled trial (RCT) sample treatment effects with the population effects of substance use disorder (SUD) treatment.

Design—Statistical weighting was used to re-compute the effects from ten RCTs such that the participants in the trials had characteristics that resembled those of patients in the target populations.

Settings—Multi-site RCTs and usual SUD treatment settings in the USA.

Participants—A total of 3,592 patients in ten RCTs and 1,602,226 patients from usual SUD treatment settings between 2001 and 2009.

Measurements—Three outcomes of SUD treatment were examined: retention, urine toxicology, and abstinence. We weighted the RCT sample treatment effects using propensity scores representing the conditional probability of participating in RCTs.

Corresponding author: Ryoko Susukida, 624 N. Broadway Room 897, Baltimore, MD, 21205, USA, Telephone: 443-703-9707, Fax: 410-614-7469, rsusuki1@jhu.edu.

Declarations of competing interest: Dr. Susukida and Mr. Ebnesajjad have nothing to disclose. Drs. Crum, Stuart and Mojtabai report grants from National Institute on Drug Abuse and National Institute of Mental Health during the conduct of the study. Dr. Mojtabai has received research funding and consulting fees from Bristol-Myers Squibb and Lundbeck Pharmaceuticals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Findings—Weighting the samples changed the significance of estimated sample treatment effects. Most commonly, positive effects of trials became statistically non-significant after weighting (three trials for retention and urine toxicology, and one trial for abstinence); but also, non-significant effects became significantly positive (one trial for abstinence), and significantly negative effects became non-significant (two trials for abstinence). There was suggestive evidence of treatment effect heterogeneity in subgroups that are under- or over-represented in the trials, some of which were consistent with the differences in average treatment effects between weighted and unweighted results.

Conclusions—The findings of randomized controlled trials (RCTs) for substance use disorder treatment do not appear to be directly generalizable to target populations when the RCT samples do not adequately reflect the target populations and there is treatment effect heterogeneity across patient subgroups.

Introduction

There is growing concern that the results from randomized controlled trials (RCTs) may not generalize to real world settings (1–6). Perhaps due to this, many interventions with strong efficacy evidence either cannot be replicated or produce smaller effects in different settings (7,8). Limitations in generalizability of the findings from RCTs are pose major clinical and policy concerns because RCTs are considered the most accepted study design for choosing evidence-based practices. The randomized study design does not necessarily ensure external validity, which means that the findings of an RCT may not be applicable to all individuals for whom treatment or intervention is intended. Individuals who volunteer to participate in RCTs are typically different from those who refuse to participate. Furthermore, strict eligibility criteria are likely to make the findings less applicable to subgroups who are excluded from trials.

Particularly in the context of RCTs of treatments for substance use disorders (SUD), there is a growing body of research indicating that the samples recruited to the RCTs are substantially different from target populations (2,9–11). It is also known that women, especially pregnant women, African-Americans, low-income individuals, and individuals with more severe alcohol, drug, and psychiatric problems are disproportionately under-represented in SUD treatment RCTs (11,12). Furthermore, commonly used eligibility criteria in SUD treatment RCTs exclude substantial portions of the target population. However, the prevalence of such exclusions varies across studies. For example, Humphreys et al. (12) found that 20% to 33% of patients with alcohol use disorders would be excluded by the eligibility criteria commonly used in RCTs of alcohol use disorders, whereas, Okuda et al. (2) found that as many as 80% of patients with cannabis dependence would be excluded by the commonly used eligibility criteria for cannabis treatment RCTs. A recent review study by Moberg and Humphreys (13) estimated that commonly used exclusion criteria in SUD trials would exclude between 64% and 95% of potential participants.

A study by Susukida et al. (14) compared the characteristics of participants in ten RCTs from the National Institute of Drug Abuse Clinical Trials Network and the intended target populations and found substantial differences in sociodemographic characteristics. The

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

proportion of individuals with more than 12 years of education and those who had full-time jobs were significantly higher among the RCT samples than among target populations (11).

While improving the representativeness of RCTs participants may be a reasonable solution to this problem, logistical considerations including concerns about safety, non-adherence with treatment, and drop-out from the study often limit investigators' ability to expand eligibility criteria. There is some evidence that the exclusion criteria of SUD treatment trials have become increasingly more restrictive over the years (15). Government-funded SUD treatment trials are particularly likely to use such restrictive exclusion criteria (15). Assessing how well the study samples represent potential target populations with regard to various sociodemographic and clinical characteristics, and how deviations from representativeness may have impacted the results of the study are important to examine the real-world relevance of RCTs (16). While previous studies have examined how well RCT samples represent target populations (2,11,12,14), few studies have assessed how representativeness of the RCT sample may affect the findings of the RCTs when generalized to a target population (17). Furthermore, there is little understanding of how heterogeneity of treatment effects among various subgroups that are differentially represented in RCTs may explain the generalizability of results. Generalizability of the findings for the RCTs is compromised when there are treatment effect modifiers that differ between the RCT samples and the target populations. If treatment effects among under- or over-represented subgroups in RCTs are heterogeneous, the findings from the RCT may not directly carry over to a population of interest (18).

The main aims of this study were (1) to estimate sample treatment effects and the population effects of RCTs of SUD treatment, and (2) to examine the treatment effect heterogeneity by subgroups that are under- or over-represented in the trials. To weight the results to a target population, we applied a weighting-based approach, which weights the RCT samples to resemble the target populations (17,19), and is similar to inverse probability weighting for non-experimental studies (20). This method was used by Stuart et al. (18) to examine the generalizability of the results of a randomized behavioral intervention trial in schools. This current study extends the analysis by Susukida et al. (14), which compared differences in characteristics of individuals who participated in ten SUD RCTs with individuals from target populations for whom these treatments are intended. We hypothesized that the estimated effects could be different in the RCT samples and the target populations of interest, which would be partially explained by differences in treatment effect by subgroups of individuals recruited into the RCTs.

Methods

Data source

The RCTs used in this study were the same RCTs used in our prior analyses (14). Briefly, a total of 3,592 individuals from ten RCTs from the National Institute of Drug Abuse (NIDA) Clinical Trials Network (CTN) and 1,602,226 individuals from the Treatment Episodes Data Set-Admissions (TEDS-A) between 2001 and 2009 were included. The NIDA CTN studies are multisite RCTs conducted in various settings in the United States to assess the effectiveness of treatments for SUD (21). For each RCT sample, we drew a separate

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

corresponding target sample from TEDS-A. The TEDS-A includes data on approximately 1.5 million patients (≥ 12 years old) admitted every year to SUD treatment facilities nationally. Every state that receives public funding for SUD treatment programs is mandated to provide records of all patients to the TEDS-A. Although the TEDS-A is one of the largest data set that covers patients with SUD in the US, some states limit the data to individuals whose treatment is covered by the state substance use agency funds (such as Federal Block Grant funds) (22). Treatment facilities that are managed by private agencies and hospitals are usually excluded from the TEDS-A unless they are licensed by the state substance abuse treatment agency.

The main criteria for defining target populations were the SUD that each RCT targeted, inclusion age criteria of RCT, treatment settings (outpatient vs. inpatient), and the years when the RCT was conducted. For example, the target population for CTN0001, an RCT of Buprenorphine/Naloxone Detoxification for individuals aged 18 years or older seeking treatment for opioid dependence in inpatient treatment settings, enrolled into the study between February 2001 and August 2002, was drawn from the population of patients in TEDS-A between 2001–2002 who were 18 years or older who received treatment for opioid dependence in inpatient treatment settings. For an RCT that targeted a more specific population such as pregnant women, we used the additional criteria to identify the target population. At the time of this study, target populations could be identified for a total of ten CTN studies included in the NIDA CTN database. eTable 1 (online supplement) in Susukida et al. (14) describes the definitions of the target populations for each RCT.

Table 1 describes characteristics of each CTN trial. Five trials (CTN0001(23), CTN0002(23), CTN0003(24), CTN0010(25), CTN0030(26)) examined the effectiveness of Buprenorphine/Naloxone detoxification (Bup/Nx-Detox) for opioid dependence. Three trials (CTN0004(27), CTN0005(28), CTN0013(29)) examined the effectiveness of motivational enhancement/interviewing (MEI) on SUD, and two trials (CTN0006(30), CTN0007(31)) examined the effectiveness of motivational incentives (Incentives) for cocaine, methamphetamine or amphetamine use.

Measures

There were nine comparable variables between the CTN and TEDS-A datasets: sex, race-ethnicity, age, educational attainment, employment status, marital status, admission through criminal justice, intravenous drug use, and the number of prior treatments for SUD. These nine variables were used to model the probabilities of trial participation, which were then used as weights to generalize the outcomes from the RCTs.

The following three outcomes from RCTs were generalized to the target populations: successful retention in the study, submission of a substance-free urine sample, and days of abstinence in the past 30 days. Remaining in the study until the end of the trial was considered successful retention in the study. Similarly, submitting a substance-free urine sample at the end of the trial was considered an indicator of successful detoxification. Study participants reported the number of days of use of the target substances in the past 30 days. Number of days abstinent was defined by the self-reported number of days free from the target substance in the past 30 days.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Statistical Analysis

This study used a weighting-based approach to estimate the treatment effects in the target populations. This approach is similar to inverse probability weighting for non-experimental studies, where researchers estimate the causal effect by making the exposed and unexposed samples in an observational studiy similar with respect to observed characteristics (20). In this study, we weighed both arms of the RCT samples to resemble the target populations (17,19). Unweighted and weighted analyses were conducted for all three outcomes. Thus, while the unweighted analyses estimate the effects in the trial samples, the weighted analyses estimate the population effects. The models used for the analyses were logistic regression for the binary outcomes of retention and urine toxicology, and linear regression models for days of abstinence in the past 30 days. Assuming that randomization was successful in each trial, we did not adjust for baseline variables within the trial samples.

To account for missing data, we performed multiple imputation with the STATA *ice* command (version 13) to generate 50 imputed data sets. eTable 2 in Susukida et al. (14) described the detailed patterns of missing data in each CTN sample and the corresponding target population, and the detailed procedures of multiple imputation.

Trial participation weights for each trial were calculated as $(1-p)/p$, where p was the mean propensity score across the 50 imputed data sets, defined as the probability of a patient participating in the RCT conditional on the nine variables described above. A non-parametric random forest, using the "randomForest"(32) package in R(33), was used to calculate the propensity scores for each patient (34,35). Weighted analyses with the weights for each trial, $(1-p)/p$, were conducted by using the STATA *pweights* command (version 13). In addition to comparing the statistical significance of the treatment effects from unweighted and weighted models, we statistically compared the treatment effect sizes of unweighted and weighted models, using the STATA *suest* (seemingly unrelated estimation) command (36).

We conducted subgroup analyses to examine the treatment effect heterogeneity by subgroups of RCT participants to help explain the differences between weighted and unweighted models. For example, if the statistical significance of the treatment effect of the RCT were different before and after weighting, and our analyses indicated that the RCT had enrolled a significantly larger proportion of patients with higher education, we examined heterogeneity of treatment effects between the low and high education subgroups in the RCT. We stratified RCT samples by subgroups based on variables used to model the probability of trial participation and performed *chi-squared* tests for binary outcomes and *t*-tests for continuous outcomes to explore treatment effects in different subgroups. We conducted subgroup analyses for the CTN studies that produced statistically significant results when weighted, but not when unweighted or vice versa. Furthermore, we only focused on the characteristics that significantly differed between RCT samples and the corresponding target populations. Our rationale for these further analyses was to identify the contribution of treatment effect heterogeneity to the biases in outcome produced as a result of the differences in the characteristics of the RCT samples and the target populations.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Role of the funding source

The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Results

Comparison of unweighted outcomes and outcomes weighted by propensity scores

Table 2 presents the results of the analyses for the effect of treatment on trial retention. Odds ratios (ORs) from both unweighted and weighted logistic regression models for all 10 trials are presented with the 95% confidence intervals. The unweighted models estimated the effects in the RCT samples while the weighted models estimated the effects that would be expected if the RCT sample had the same characteristics as the target populations. In unweighted analyses, treatment was associated with significantly greater odds of retention in 5 trials (CTN0001, CTN0002, CTN0003, CTN0006, and CTN0010). A significantly positive effect on retention in CTN0006, CTN0003, and in CTN0010 became statistically non-significant after weighting. Furthermore, there was a significant difference in estimated effects between unweighted and weighted models for CTN0002.

Table 3 presents comparisons of unweighted and weighted results of the studies for urine toxicology. Odds ratios (ORs) from both unweighted and weighted logistic regression models for all 10 trials are presented. In unweighted analyses, treatment was associated with significantly greater odds of drug-free urine samples in 5 trials (CTN0001, CTN0002, CTN0003, CTN0006, and CTN0010). Significantly positive effects on urine toxicology in CTN0006, CTN0003, and in CTN0010 became statistically non-significant after weighting. In all 10 trials, however, there was no statistically significant difference between unweighted and weighted models with regard to the estimated effects from the unweighted and weighted models.

Table 4 presents comparisons of unweighted and weighted linear regression results for the effect of treatment on days of abstinence in the past 30 days. Results from both unweighted and weighted linear regression models for all 10 trials are presented. In unweighted analyses, treatment was associated with significantly higher number of days of abstinence in one trial (CTN0001) and a significantly smaller number in 2 trials (CTN0004 and CTN0030). The significant positive effect in CTN0001 became non-significant after weighting. Similarly, the significant negative effects in CTN0004 and CTN0030 became statistically non-significant after weighting. Furthermore, the statistically non-significant positive effect in CTN0002 became statistically significant after weighting and, a statistically non-significant negative effect in CTN0010 became significant after weighting. There was a significant difference between unweighted and weighted effect estimates for CTN0002 but not for any of the other trials.

Subgroup analysis for treatment effect heterogeneity

As the results of our prior analyses (14) indicated, the composition of the CTN samples deviated significantly from the composition of the target populations with regard to the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

socio-demographic characteristics on which these samples were compared. Appendix eTable 1 presents the results of comparisons of the characteristics of RCT samples and target populations. To simplify the interpretation of the results, we presented the comparison using dichotomized variables in this study.

The proportion of those with 12 years or higher education was significantly larger among patients who participated in RCTs than among the target populations in seven of the ten trials (CTN0001, CTN0002, CTN0003, CTN0004, CTN0005, CTN0010, and CTN0030). The proportion of those with full-time jobs was also significantly larger among patients who participated in RCTs than among patients in target populations in all nine trials in which information on employment status was collected (CTN0001, CTN0002, CTN0003, CTN0004, CTN0005, CTN0006, CTN0007, CTN0013, and CTN0030). Furthermore, each RCT and its target population differed in terms of other characteristics although the patterns varied across trials. There were statistically significant differences in the proportions of female patients, certain race-ethnicity groups, age groups, married patients, patients who were admitted through the criminal justice system, patients with IV drug use, and patients with more than 5 prior treatments, between individual RCTs and the corresponding target populations.

We conducted subgroup analyses for outcomes of RCTs that showed a difference between the sample treatment effects and the population treatment effects subsequent to weighting. To limit the number of tests, these analyses were restricted to subgroups that met criteria for a statistically significantly difference in composition between the RCT samples and the corresponding target populations. Thus, we conducted 76 subgroup analyses (see eTable 2).

Results of subgroup analysis of treatment effects are presented in eTable 3. There were some consistent patterns in the directions of change in outcomes from weighting and examination of treatment effect heterogeneity by subgroups. As an example, in the case of CTN0006, some subgroups that were overrepresented in the RCT samples (e.g, females, married patients, those with full time jobs, and those not using IV drugs) also showed evidence of larger treatment effects on retention as compared with underrepresented subgroups. As another example, in the case of CTN0003, some subgroups that were overrepresented in the RCT samples (e.g, White patients, those with ≥12 years of education, those with full time jobs, and patients not admitted through criminal justice) also showed evidence of larger treatment effects on retention as compared with underrepresented subgroups. Weighting this RCT sample to be more similar in composition to the target sample increased the weights for subsamples with smaller effect sizes, leading to statistically non-significant estimates of the population effects.

Discussion

This study demonstrated that the observed outcomes of some RCTs may not carry over directly to potential target populations. In most cases, statistically significant results seen in the RCT samples became non-significant when weighted to the target population. These differences in effect estimates between the RCT samples and the target populations could be partially explained by the patterns in treatment effect heterogeneity across subgroups.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

A recent study by Stuart et al. (18) that applied the same weighting-based method to generalize the results of a behavioral intervention trial in school settings found that the weighted effect of intervention was just slightly attenuated compared to the effect seen in the trial. To our knowledge, the present study is the first to use this weighting approach to estimate target population effects using the results of SUD RCTs. Previous studies showing substantial differences between SUD RCT samples and target populations implied that the difference might affect generalizability of the results from RCTs (2,11,12,14); however, those studies did not attempt to estimates the population effects from trial results.

Our study findings have implications for the external validity of results from SUD RCTs. Susukida et al. (14) showed substantial variability in the likelihood of being in RCT samples across patient subgroups and indicated that poor representation of target populations might impact the generalizability of findings from RCTs. The results of the present study confirm this prediction by revealing differences in the statistical significance between the sample treatment effects and the population treatment effects. The present study also found suggestive evidence that treatment effect heterogeneity among under- or over-represented subgroups of patients in the RCTs could partially explain why the population treatment effects estimated by weighting the RCT samples differed from the sample effects.

The results of this study should be interpreted in light of several limitations. First, the number of characteristics measured in both the RCT samples and the target populations was limited. Therefore, it is likely that weights calculated in this study could not take into account other characteristics that may differ between the RCT samples and target populations and moderate treatment effects. Second, due to the significant differences between the RCT samples and their target populations, the weighting-based method may not have adequately made the RCT samples resemble the target populations to estimate the population treatment effects. In Susukida et al. (14), for all ten RCT studies, the difference in mean propensity scores between the RCT sample and its target population was much larger than the cut-off proposed by Stuart (37). Weighting the RCT samples to estimate the population treatment effects is more reliable when the RCT samples and the target populations are more similar to start with. Third, difference between the sample treatment effect and the population effect could be due to difficulties in equating the trial sample and population with respect to the covariates. For example, for the urine toxicology outcome in CTN0010, where a significant effect became non-significant after weighting, the distributions of educational attainment as well as marital status were significantly different between the RCT sample and its target population even after weighting. Furthermore, we did not find consistent patterns of the treatment effect heterogeneity of study participants by educational attainment and marital status. Fourth, the primary goal of the ten CTN studies was not to assess treatment effect heterogeneity. Hence, the subgroup analyses conducted for this study were not adequately powered and the findings only provide suggestive evidence of treatment effect heterogeneity across subgroups of patients. Fifth, TEDS-A data miss some groups of patients. Therefore, the population treatment effects estimated by this study may not represent treatment effects among all recipients of SUD treatment in the US. Furthermore, patients in TEDS-A represent treatment-seeking individuals and do not necessarily represent the whole population of individuals who need treatment and are potential recipients of such treatments (34). Results may differ if future studies use broader

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

definition of target populations, including non-treatment-seeking individuals. Finally, our estimates of the RCT results do not necessarily correspond to the published reports by primary investigators. The primary investigators of the CTN RCTs operationalized outcomes differently (23–31). For example, some original outcome studies published by primary investigators reported treatment effects by trial sites (28); whereas, the site identifiers were not provided in the publically available NIDA data. Therefore, we were not able to replicate these site-specific results. In order to compare how weighting affects the findings across the studies, we chose to use the same measures across the studies based on the raw RCT data provided in the NIDA CTN repository. It should also be noted that the unweighted sample treatment effects were not always significantly positive. This may have been possibly due to receipt of standard care among patients in control arm.

Acknowledging these limitations, results from this study provide a first insight into whether and how deviations in RCT sample representativeness from target populations influence the observed outcomes of SUD RCTs. It is critical for future CTN studies to place greater emphasis on external validity of RCTs, particularly because a primary goal of the NIDA CTN was to provide data on SUD treatments that can be disseminated in usual care settings. As interest in comparative effectiveness research in real-world treatment settings increases, RCTs for mental health treatments increasingly use less stringent eligibility criteria for participation, which may improve generalizability of the findings of RCTs (38). However, relaxing eligibility criteria may not be feasible for all RCTs, especially when there are safety concerns for patients such as allergic reactions to certain medications. In such cases, the weighting-based method that this study employed might be useful to examine to what extent the findings of RCTs are applicable to target populations. As attention to large-scale dissemination and implementation of evidence-based treatments and interventions increases (39), it becomes increasingly important to understand the applicability of the findings of RCTs in different populations with varying characteristics, contexts, and locations. It is also important to consider the change in the nature of target populations especially in the context of the United States, where more people are eligible for health insurance as a part of Affordable Care Act legislation (40), which may affect profiles of patient groups who seek and access treatments.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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.

References

1. Blanco C, Olfson M, Goodwin RD, Ogburn E, Liebowitz MR, Nunes EV, et al. Generalizability of clinical trial results for major depression to community samples: results from the National Epidemiologic Survey on Alcohol and Related Conditions. J Clin Psychiatry. 2008; 69(8):1276–80. [PubMed: 18557666]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2. Okuda M, Hasin DS, Olfson M, Khan SS, Nunes EV, Montoya I, et al. Generalizability of clinical trials for cannabis dependence to community samples. Drug Alcohol Depend. 2010; 111(1–2):177–81. [PubMed: 20537813]

3. Hoertel N, Le Strat Y, Blanco C, Lavaud P, Dubertret C. Generalizability of clinical trial results for generalized anxiety disorder to community samples. Depress Anxiety. 29(7):614–20. 2012. [PubMed: 22495990]

4. Hoertel N, Santiago H, Wang S, González-pinto A, Blanco C. Generalizability of Pharmacological and Psychotherapy Clinical Trial Results for Borderline Personality Disorder to Community Samples. Personal Disord. 2015; 6(1):81–7. [PubMed: 25580674]

5. Hoertel N, Le Strat Y, Lavaud P, Dubertret C, Limosin F. Generalizability of clinical trial results for bipolar disorder to community samples: findings from the National Epidemiologic Survey on Alcohol and Related Conditions. J Clin Psychiatry. 2013; 74(3):265–70. [PubMed: 23561233]

6. Rothwell PM. External validity of randomised controlled trials: “to whom do the results of this trial apply?”. Lancet. 2005:82–93.

7. Ozonoff S. Editorial: The first cut is the deepest: why do the reported effects of treatments decline over trials? J Child Psychol Psychiatry. 2011; 52(7):729–30. [PubMed: 21644983]

8. Ioannidis JP. Evolution and translation of research findings: from bench to where? PLoS Clin Trials. 2006; 1(7):e36. [PubMed: 17111044]

9. Blanco C, Olfson M, Goodwin RD, Ogburn E, Liebowitz MR, Nunes EV, et al. Generalizability of clinical trial results for major depression to community samples: results from the National Epidemiologic Survey on Alcohol and Related Conditions. J Clin Psychiatry. 2008; 69(8):1276–80. [PubMed: 18557666]

10. Humphreys K, Weingardt KR, Harris AH. Influence of subject eligibility criteria on compliance with National Institutes of Health guidelines for inclusion of women, minorities, and children in treatment research. Alcohol Clin Exp Res. 2007; 31(6):988–95. [PubMed: 17428295]

11. Humphreys K, Weisner C. Use of exclusion criteria in selecting research subjects and its effect on the generalizability of alcohol treatment outcome studies. Am J Psychiatry. 2000; 157(4):588–94. [PubMed: 10739418]

12. Humphreys K, Weingardt KR, Harris AH. Influence of subject eligibility criteria on compliance with National Institutes of Health guidelines for inclusion of women, minorities, and children in treatment research. Alcohol Clin Exp Res. 2007; 31(6):988–95. [PubMed: 17428295]

13. 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

14. 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

15. Humphreys K, Weingardt KR, Horst D, Joshi AA, Finney JW. Prevalence and predictors of research participant eligibility criteria in alcohol treatment outcome studies, 1970–98. Addiction. 2005; 100(9):1249–57. [PubMed: 16128714]

16. Humphreys K, Harris AH, Weingardt KR. Subject eligibility criteria can substantially influence the results of alcohol-treatment outcome research. J Stud Alcohol Drugs. 2008; 69(5):757–64. [PubMed: 18781251]

17. 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 Stat Soc. 2011; 174(2):369–86.

18. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the Generalizability of Randomized Trial Results to Target Populations. Prev Sci. 2015; 16(3):475–85. [PubMed: 25307417]

19. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. Am J Epidemiol. 2010; 172(1):107–15. [PubMed: 20547574]

20. Mansournia MA, Altman DG. Inverse probability weighting. BMJ Br Med J. 2016

21. National Institute on Drug Abuse Clinical Trials Network. Clinical Trials Network (CTN): Research Studies [Internet]. Available from: http://www.drugabuse.gov/about-nida/organization/cctn/ctn/research-studies

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

22. Substance Abuse & Mental Health Services and Quality. Quick Statistics from the Drug and Alcohol Services Information System [Internet]. [cited 2016 Nov 6]. Available from: http://wwwdasis.samhsa.gov/webt/information.htm

23. Ling W, Amass L, Shoptaw S, Annon JJ, Hillhouse M, Babcock D, et al. A multi-center randomized trial of buprenorphine-naloxone versus clonidine for opioid detoxification: findings from the National Institute on Drug Abuse Clinical Trials Network. Addiction. 2005; 100(8):1090–100. [PubMed: 16042639]

24. Ling W, Hillhouse M, Domier C, Doraimani G, Hunter J, Thomas C, et al. Buprenorphine tapering schedule and illicit opioid use. Addiction. 2009; 104(2):256–65. [PubMed: 19149822]

25. Woody GE, Poole SA, Subramaniam G, Dugosh K, Bogenschutz M, Abbott P, et al. Extended vs short-term buprenorphine-naloxone for treatment of opioid-addicted youth: a randomized trial. JAMA. 2008; 300(17):2003–11. [PubMed: 18984887]

26. Weiss RD, Potter JS, Fiellin DA, Byrne M, Connery HS, Dickinson W, et al. Adjunctive counseling during brief and extended buprenorphine-naloxone treatment for prescription opioid dependence: a 2-phase randomized controlled trial. Arch Gen Psychiatry. 2011; 68(12):1238–46. [PubMed: 22065255]

27. Ball SA, Martino S, Nich C, Frankforter TL, Van Horn D, Crits-Christoph P, et al. Site matters: multisite randomized trial of motivational enhancement therapy in community drug abuse clinics. J Consult Clin Psychol. 2007; 75(4):556–67. [PubMed: 17663610]

28. Carroll KM, Ball SA, Nich C, Martino S, Frankforter TL, Farentinos C, et al. Motivational interviewing to improve treatment engagement and outcome in individuals seeking treatment for substance abuse: a multisite effectiveness study. Drug Alcohol Depend. 2006; 81(3):301–12. [PubMed: 16169159]

29. Winhusen T, Kropp F, Babcock D, Hague D, Erickson SJ, Renz C, et al. Motivational enhancement therapy to improve treatment utilization and outcome in pregnant substance users. J Subst Abus Treat. 2008; 35(2):161–73.

30. Petry NM, Peirce JM, Stitzer ML, Blaine J, Roll JM, Cohen A, et al. Effect of prize-based incentives on outcomes in stimulant abusers in outpatient psychosocial treatment programs: a national drug abuse treatment clinical trials network study. Arch Gen Psychiatry. 2005; 62(10): 1148–56. [PubMed: 16203960]

31. Peirce JM, Petry NM, Stitzer ML, Blaine J, Kellogg S, Satterfield F, et al. Effects of lower-cost incentives on stimulant abstinence in methadone maintenance treatment: a National Drug Abuse Treatment Clinical Trials Network study. Arch Gen Psychiatry. 2006; 63(2):201–8. [PubMed: 16461864]

32. Liaw A, Wiener M. Classification and Regression by randomForest. R news. 2002 Dec.2:18–22.

33. R Development Core Team R. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2016:409.

34. Breiman L. Random Forests. Mach Learn. 2001; 45(1):5–32.

35. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med. 2010; 29(3):337–46. [PubMed: 19960510]

36. Weesie J. Seemingly unrelated estimation and the cluster-adjusted sandwich estimator. Stata Tech Bull. 1999; 52:34–47.

37. Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci. 2010; 25(1):1–21. [PubMed: 20871802]

38. Wang PS, Ulbricht CM, Schoenbaum M. Improving mental health treatments through comparative effectiveness research. Heal Aff. 2009; 28(3):783–91.

39. Flay BR, Biglan A, Boruch RF, Castro FG, Gottfredson D, Kellam S, et al. Standards of evidence: Criteria for efficacy, effectiveness and dissemination. Prevention Science. 2005:151–75. [PubMed: 16365954]

40. The Center for Medicaid and CHIP Services. Affordable Care Act.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Addiction. Author manuscript; available in PMC 2018 July 01.

Table 1

Description of CTN studies

CTN Study Number	Study Title	Years	Sample Size	Arm (T vs. C)	Example of Eligibility Criteria
Buprenorphine/naloxone (Bup/Nx) detoxification					
CTN0001	Buprenorphine/Naloxone (Bup/Nx) versus Clonidine for Inpatient Opiate Detoxification	2001–2002	113	Bup/Nx vs.Clonidine	Inpatient treatment-seeking males and non-pregnant and non-lactating females, 15 years and older, with DSM-IV opiate dependence
CTN0002	Buprenorphine/Naloxone (Bup/Nx) versus Clonidine for Outpatient Opiate Detoxification	2001–2002	230	Bup/Nx vs.Clonidine	Outpatient treatment-seeking males and non-pregnant and non-lactating females, 15 years and older, with DSM-IV opiate dependence
CTN0003	Suboxone® (Bup/Nx) Taper: A Comparison of Two Schedules	2003–2005	516	7-day vs. 28-day Bup/Nx Taper	Outpatient treatment-seeking males and non-pregnant and non-lactating females, 15 years and older, with DSM-IV opiate dependence
CTN00010	Buprenorphine/Naloxone (Bup/Nx) Facilitated Rehabilitation for Heroin Addicted Adolescents/Young Adults	2003–2006	154	Bup/Nx vs. Detox	Outpatient treatment-seeking males and non-pregnant and non-lactating females, 14–21 years old, with DSM-IV-TR opiate dependence
CTN00030	Buprenorphine/Naloxone Treatment Plus Individual Drug Counseling for Opioid Analgesic Dependence	2006–2009	653	Bup/Nx + Counseling vs. Bup/Nx	Outpatient treatment-seeking males and non-pregnant and non-lactating females, 18 years and older, with DSM-IV opiate dependence
Motivational enhancement/interviewing (MEI)					
CTN0004	Motivational Enhancement Treatment (MET) To Improve Treatment Engagement and Outcome in Subjects Seeking Treatment for Substance Abuse	2001–2004	461	MET vs. Counseling as usual (CAU)	Outpatient treatment-seeking individuals for any substance use disorder with use of any substance in the past 28 days, 18 years and older
CTN0005	Motivational Interviewing (MI) To Improve Treatment Engagement and Outcome in Subjects Seeking Treatment for Substance Abuse	2001–2002	423	MI vs. treatment-as-usual (TAU)	Outpatient treatment-seeking individuals for any substance use disorder with use of any substance in the past 28 days, 18 years and older
CTN0013	Motivational Enhancement Therapy (MET) to Improve Treatment Utilization and Outcome in Pregnant Substance Users	2003–2006	200	MET vs. TAU	Pregnant women (Less than 32 weeks), identified as needing substance abuse treatment, 18 years and older
Motivational incentives (Incentives)					
CTN0006	Motivational Incentives for Enhanced Drug Abuse Recovery: Drug Free Clinics	2001–2003	454	Incentives vs. TAU	Outpatient treatment-seeking individuals with evidence of cocaine or methamphetamine use, without gambling problems
CTN0007	Motivational Incentives for Enhanced Drug Abuse Recovery: Methadone Clinics	2001–2003	388	Incentives vs. TAU	Outpatient treatment-seeking individuals with evidence of cocaine or methamphetamine use, without gambling problems

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Addiction. Author manuscript; available in PMC 2018 July 01.

Table 2

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

Retention					
	OR	**95%CI**		**p**	**Comparison of the effect estimates from the unweighted and weighted models**
CTN1					
Unweighted	13.34	5.11	34.83	<.01	
Weighted	9.10	1.54	53.99	.02	F(1, 225)=0.14, p=.71
CTN2					
Unweighted	3.49	1.91	6.38	<.01	
Weighted	17.78	6.38	49.58	<.01	F(1, 459)=7.19, p=.01
CTN3					
Unweighted	2.06	1.39	3.05	<.01	
Weighted	1.16	0.42	3.25	.77	F(1, 1031)=1.04, p=.31
CTN4					
Unweighted	1.24	0.85	1.81	.26	
Weighted	1.26	0.50	3.21	.63	F(1, 921)=0.00, p=.98
CTN5					
Unweighted	1.26	0.80	1.98	.31	
Weighted	1.08	0.47	2.47	.85	F(1, 845)=0.10, p=.75
CTN6					
Unweighted	1.63	1.11	2.39	.01	
Weighted	1.26	0.62	2.53	.52	F(1, 907)=0.41, p=.52
CTN7					
Unweighted	1.21	0.81	1.80	.36	
Weighted	0.55	0.17	1.80	.32	F(1, 771)=1.51, p=.22
CTN10					
Unweighted	2.68	1.32	5.44	<.01	
Weighted	1.46	0.08	26.07	.80	F(1, 307)=0.16, p=.69
CTN13					
Unweighted	0.54	0.28	1.05	.07	
Weighted	0.31	0.08	1.19	.09	F(1, 399)=.52, p=.47

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Retention

	OR	95%CI		p	Comparison of the effect estimates from the unweighted and weighted models
CTN30					
Unweighted	0.91	0.67	1.24	.55	
Weighted	0.95	0.30	2.99	.93	F(1, 1305)=.00, p=.95

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Addiction. Author manuscript; available in PMC 2018 July 01.

Table 3

Comparison of unweighted (RCT sample effect) and weighted (population effect) odds ratios of treatment effect on urine toxicology

Urine toxicology					
	OR	**95%CI**		**p**	**Comparison of the effect estimates from the unweighted and weighted models**
CTN1					
Unweighted	8.22	3.26	20.72	<.01	
Weighted	8.26	1.43	47.76	.02	F(1, 225)=0.00, p=.99
CTN2					
Unweighted	10.80	2.52	46.21	<.01	
Weighted	59.76	10.89	327.87	<.01	F(1, 459)=2.24, p=.13
CTN3					
Unweighted	1.84	1.28	2.64	<.01	
Weighted	1.36	0.54	3.43	.52	F(1, 1031)=0.36, p=.55
CTN4					
Unweighted	1.11	0.77	1.60	.59	
Weighted	1.32	0.54	3.26	.54	F(1, 921)=0.13, p=.72
CTN5					
Unweighted	1.18	0.80	1.72	.40	
Weighted	1.79	0.80	3.99	.15	F(1, 845)=0.87, p=.35
CTN6					
Unweighted	1.48	0.99	2.20	.05	
Weighted	1.13	0.56	2.28	.74	F(1, 907)=0.44, p=.51
CTN7					
Unweighted	0.87	0.51	1.49	.62	
Weighted	0.48	0.13	1.82	.28	F(1, 771)=0.65, p=.42
CTN10					
Unweighted	5.55	2.71	11.36	<.01	
Weighted	4.71	0.29	76.54	.28	F(1, 307)=0.01, p=.91
CTN13					
Unweighted	0.72	0.41	1.25	.24	
Weighted	1.38	0.36	5.21	.64	F(1, 399)=0.78, p=.38

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Urine toxicology

	OR	95%CI		p	Comparison of the effect estimates from the unweighted and weighted models
CTN30					
Unweighted	0.72	0.41	1.25	.24	
Weighted	1.38	0.36	5.21	.64	F(1, 1305)=0.00, p=.97

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Comparison of unweighted (RCT sample effect) and weighted (population effect) regression coefficients of treatment effect on self-reported days of abstinence in the past 30 day

Abstinence					
	OR	**95%CI**		**p**	**Comparison of the effect estimates from the unweighted and weighted models**
CTN1					
Unweighted	6.47	1.60	11.35	.01	
Weighted	0.58	−3.82	4.98	.79	F(1, 225)=2.67, p=.10
CTN2					
Unweighted	3.07	−1.77	7.90	.21	
Weighted	13.10	5.82	20.37	<.01	F(1, 459)=4.95, p=.03
CTN3					
Unweighted	0.63	−1.75	3.00	.61	
Weighted	3.92	−1.31	9.15	.14	F(1, 1031)=1.28, p=.26
CTN4					
Unweighted	−2.52	−4.26	−0.79	<.01	
Weighted	−3.02	−6.98	0.94	.14	F(1, 921)=0.05, p=.82
CTN5					
Unweighted	−0.84	−2.88	1.20	.42	
Weighted	1.31	−5.57	8.20	.71	F(1, 845)=0.35, p=.55
CTN6					
Unweighted	0.16	−1.36	1.68	.83	
Weighted	2.53	−0.34	5.41	.08	F(1, 907)=2.07, p=.15
CTN7					
Unweighted	0.26	−1.34	1.87	.75	
Weighted	−0.12	−1.89	1.66	.90	F(1, 788)=0.10, p=.75
CTN10					
Unweighted	−0.94	−5.44	3.57	.68	
Weighted	−3.38	−5.57	−1.19	<.01	F(1, 307)=0.96, p=.33
CTN13					
Unweighted	0.72	−2.35	3.78	.64	

Addiction. Author manuscript; available in PMC 2018 July 01.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Abstinence

	OR	95%CI		p	Comparison of the effect estimates from the unweighted and weighted models
Weighted	1.70	−4.06	7.46	.56	F(1, 399)=0.09, p=.77
CTN30					
Unweighted	−1.79	−3.37	−0.20	.03	
Weighted	0.85	−4.08	5.78	.74	F(1, 1305)=1.00, p=.32